arXiv:2302.00755v1 [stat.ML] 1 Feb 2023

层次收缩高斯过程：在计算机代码仿真和动力系统恢复中的应用

Tao Tang[1], Simon Mak [2], David Dunson[2]

2023年2月3日

摘要

在科学与工程的许多领域，计算机模拟被广泛用作物理实验的替代方案，因为物理实验可能难以实施或不符合伦理。此类模拟往往计算成本高昂，因此可以训练仿真器，高效预测所需的响应面。高斯过程（高斯过程）是一种广泛使用的仿真器，它为高效预测和不确定性量化提供了灵活的框架。然而，标准高斯过程无法捕捉底层响应面上的结构化稀疏性，而这种稀疏性在许多应用中都存在，在物理科学领域尤为如此。因此，我们提出一种新的层次收缩高斯过程（HierGP），通过在高斯过程框架中采用累积收缩先验来纳入此类结构。我们证明，HierGP 隐式地融入了实验分析中著名的效应稀疏性、遗传性和层次性原则（Hamada 和 Wu，1992），使我们的模型能够利用有限数据从响应面中识别结构化稀疏特征。我们提出了用于模型训练和预测的高效后验抽样算法，并证明了 HierGP 具有理想的一致性性质。最后，我们通过一系列数值实验以及一项动力系统恢复应用，展示了 HierGP 相较于现有模型的性能提升。

关键词：计算机实验；动力恢复；仿真；高斯过程；收缩先验；不确定性量化。

[1]数学系，杜克大学[2]统计系，杜克大学

1 引言

科学计算在解决现代科学与工程问题中正发挥着日益重要的作用，这尤其得益于近年来数学建模与计算领域的突破。过去难以或无法通过物理实验观测的感兴趣量，如今可以通过计算机代码进行可靠的模拟。此类 计算机实验 通常涉及基于物理模型求解微分方程系统，并已对众多领域产生广泛影响，包括火箭发动机喷注器设Mak 等人，2018年）、个性化手术规划（Chen 等人，2021年）以及宇宙膨胀（Kaufman 等人，2011；Ji 等人，2021年）。然而，一个关键瓶颈在于，这些虚拟实验的计算资源消耗可能非常高。例如，对单个火箭发动机喷注器进行全规模模拟，可能需要数百万 CPU 小时（Mak 等人，2018年）。这进而对设计探索与优化所需的计算资源提出了巨大需求。

一种解决方法是训练一个仿真器模型，使其能够高效预测（或仿真）计算成本高昂的计算机代码 $f(\boldsymbol{x})$ 在不同参数 $\boldsymbol{x}$下的运行结果。其思路是在参数空间 $\mathcal{X}$中精心选取若干点运行计算机代码，然后将这些结果作为训练数据来拟合仿真模型。仿真模型的一种常见选择是高斯过程（GP；参见Rasmussen，2003），它是一种灵活的贝叶斯模型，可用于概率预测建模。高斯过程在计算机代码仿真方面有两大优势（Santner 等人，2003）：它能为仿真器的预测和不确定性量化提供闭式表达式，并为黑箱响应面的建模提供灵活的非参数框架。

尽管如此，标准高斯过程在仿真方面仍存在若干显著局限。一个问题是，当训练样本量 n 较小时，高斯过程高度灵活的形式可能弊大于利。具体而言，在样本量有限时，高斯过程后验分布可能会将概率分散到广泛的函数范围内，从而不仅导致预测性能不佳，还会带来较高的预测不确定性。一种颇有前景的解决方案是利用响应面 f 的先验知识（例如，基于对应用领域底层物理规律的了解），并将其纳入概率高斯过程模型中。这类知识可以体现为机理模型（Wheeler 等人，2014）、边界条件（Ding 等人，2019）或形状约束（Golchi 等人，2015；Wang 和 Berger，2016），并能在数据有限时显著提升高斯过程模型的预测性能。然而，对于高度复杂的物理系统，难以有把握地获取这类现有模型所需的特定先验知识，而且将其纳入概率建模也过于复杂。因此，在数据有限的情形下，需要其他仿真器来学习有助于预测的结构。在高斯过程的统计学文献中，一些相关的现有思路包括采用变量选择的高斯过程（Savitsky 等人，2011）和图拉普拉斯高斯过程（Dunson 等人，2020年），它们允许将输入限制在一个未知的低维流形上。

对于物理系统，复杂的响应面通常具有稀疏性，其中一部分

输入特征对应着系统中的主导物理机制，并驱动响应面的大部分变化。这种稀疏性得到了著名的 Buckingham-π 定理（Buckingham，1914）以及实验物理和理论物理领域大量文献的支持（例如，参见 Berkooz 等人，1993）。此外，这种稀疏性通常具有特定的结构，满足效应层级和遗传性（Hamada 和 Wu，1992）等统计学原理，以及边际性原则（Nelder，1977年），这些原理广泛用于分析物理系统的实验数据。此处，效应层级指的是这样一种假设：主效应通常比交互效应影响更大；效应遗传则指这样一种性质：只有当交互作用所包含的各主效应存在时，该交互作用才会出现。当前带有变量选择功能的高斯过程（Savitsky 等人，2011）无法体现这种结构化稀疏性。将主效应与成对交互作用分离的当前高斯过程（Ferrari 和 Dunson，2021年）也未纳入上述结构约束。因此，我们旨在将这些原理直接嵌入高斯过程先验设定中，从而构建一种灵活且由数据驱动的框架，以提高基于有限训练数据进行预测的效率。

我们将所提出的方法称为层次收缩高斯过程（HierGP）。该方法在高斯过程框架内通过精心设计的收缩先验，嵌入效应的结构化稀疏性（即效应层级和效应遗传性的原则）。HierGP 从高斯过程的基展开出发，然后为基函数系数指定层次收缩先验，以捕捉所需的层次稀疏结构。具体而言，我们改造了Legramanti 等人（2020年）提出的累积收缩先验（该先验曾用于因子分析），并证明嵌入此类先验的 HierGP 模型确实具有所需的效应层级与遗传性。随后，我们提出一种高效的吉布斯采样器，利用数据增广技巧实现高效的后验计算。在稀疏性满足温和条件时，我们进一步证明了 HierGP 模型在固定设计点和随机抽样设计点下的后验收缩结果。最后，我们通过一系列数值实验以及一项动力系统恢复应用，展示了所提出的 HierGP 相较于现有方法的有效性。

本文结构如下。第2 节介绍所提出的 HierGP。第3 节介绍用于后验抽样的数据增广吉布斯采样器。第4 节概述 HierGP 的后验一致性结果。第5 节报告数值实验以及动力系统恢复的应用。第6 节总结全文。

2 模型设定

本节介绍 HierGP，并讨论它与现有高斯过程模型的联系。我们首先回顾标准高斯过程及其表示形式，即由随机系数组成的无限基展开。然后，我们提出 HierGP，将其作为这种基表示的扩展，并采用精心选择的收缩先验，以捕捉所需的层次收缩特性。

2.1 高斯过程建模

在下文中，我们将 $f(\boldsymbol{x})$记为待模拟的昂贵黑箱函数，其中$\boldsymbol{x} \in \mathbb{R}^d$ 为其输入参数。高斯过程（Rasmussen，2003；Santner 等人，2003）对 $f(\boldsymbol{x})$采用如下概率先验：

$$f(\cdot) \sim \mathrm{GP}\{\mu(\cdot), k(\cdot,\cdot)\}. \tag{1}$$

这里，$\mu(\cdot)$ 是该过程的均值函数，而 $k(\cdot,\ \cdot)$ 是其协方差函数。高斯过程建模的一大优势在于，给定来自黑箱系统的观测数据 $f(\boldsymbol{x}_1)\ \cdots,\ f(\boldsymbol{x}_n)$ 后，后验预测分布 $[f(\cdot)|f(\boldsymbol{x}_1)\ \cdots,\ f(\boldsymbol{x}_n)]$仍然是一个高斯过程，且具有闭式的后验均值和协方差。这使得我们能够通过灵活的贝叶斯非参数模型进行预测和不确定性量化。

然而，为了融入所需的层级收缩结构，我们将采用高斯过程的另一种表示形式，即无限基展开。这一表示基于著名的 Karhunen–Lo`eve 定理（下文陈述），该定理表明，高斯过程可以表示为正交归一函数的无限基展开。

定理 *1*（*Karhunen–Lo`eve，Alexanderian*（*2015*）中的定理 *5.3*）。设 $\mathcal{X} \subseteq R^d$ 为所关注的输入空间。设 $f(\boldsymbol{x}) \sim$ 高斯过程$\{0, k(\cdot,\cdot)\}$ 为零均值高斯过程，其协方差核 k在 $\mathcal{X} \times \mathcal{X}$ 上连续，且 $k(\boldsymbol{x}, \boldsymbol{x}') \in L^2(\mathcal{X})$。那么，存在 $L^2(\mathcal{X})$ 的一个标准正交基 $\{\phi_k(\boldsymbol{x})\}_{k=1}^{\infty}$，使得：

$$f(\boldsymbol{x}) = \sum_{k=1}^{\infty} \lambda_k \phi_k(\boldsymbol{x}), \tag{2}$$

其中，系数 $\{\lambda_k\}_{k=1}^{\infty}$ 是相互独立的高斯随机变量，由 $\lambda_k = \int_{\mathcal{X}} f(\boldsymbol{x})\phi_k(\boldsymbol{x})d\boldsymbol{x}$ 给出，并满足 $\mathbb{E}(\lambda_k) = 0$ 以及 $\mathbb{E}(\lambda_j\lambda_k) = \mathbf{1}(j = k)\mathrm{Var}(\lambda_k)$。

这里，对核 k 的条件确保高斯过程 f 均方连续，且 $f \in L^2(\mathcal{X} \times \Omega)$ (Alexanderian，2015）。式（2）也称为 Karhunen–Lo`eve 展开，在统计学（Wang，2008）和不确定性量化（Xiu，2010；GhanemandSpanos，1991）中广泛使用。这种展开可视为经典傅里叶展开的一种随机类比。

作为一个直观示例，考虑维度为 $d = 1$ 且 $f:\ \mathbb{R} \to \mathbb{R}$的情形。假设 $f(x)$服从式（2）中的高斯过程，其均值函数为 $\mu(\cdot) \equiv 0$，协方差函数为平方指数型：

$$k(x, x') = \sigma^2 \exp\left\{-\frac{-(x - x')^2}{2l^2}\right\}. \tag{3}$$

然后可以证明（例如参见 Rasmussen，2003），相应的 Karhunen-Lo`eve 分解 $f\ (x)$ 可写成以下形式（2），其中

$$\lambda_k = \sqrt{\frac{2a}{A}} B^k \xi_k, \qquad \phi_k(x) = \exp\left\{-(c - a)x^2\right\} H_k(\sqrt{2c}x), \tag{4}$$

其中 $\xi_k \overset{i.i.d.}{\sim} N(0,\ 1)$，$H_k$ 是 k 阶厄米特多项式（例如参见 Xiu，2010），其中 $a^{-1} = 4\sigma_w^2$，$b^{-1} = 2l^2$，$c = \sqrt{a^2 + 2ab}$，$A = a + b + c$ 以及 $B = b/A$。

对于我们所关注的昂贵计算机代码仿真问题，式（2）中的分解揭示了标准高斯过程作为仿真模型的一项潜在局限。从先验设定的角度来看，对系数 λ_k（对应每个基函数 $\phi_k(\boldsymbol{x})$）使用相互独立的高斯先验，可能无法反映建模者的先验信念，即响应面中存在数量有限的稀疏且具有结构化特征的主导项。这是因为这些系数以概率 1 均为非零（因而处于激活状态）。因此，当响应面中确实存在这种稀疏结构时，使用标准高斯过程进行仿真可能会导致预测性能不佳且不确定性较高，这两者都是不理想的。接下来，我们将介绍对基表示（2）的一种新颖改进，它在式 $(\lambda_k)_{k=1}^{\infty}$ 的先验设定中嵌入了所需的层次稀疏结构，从而能够对响应面 $f(\boldsymbol{x})$ 中的主导特征进行结构化稀疏学习。

2.2 单变量 HierGP 模型

现在，我们给出 HierGP 的模型设定。该模型对式（2）中的高斯过程基表示进行了稀疏化改造。为简单起见，我们先在 $d = 1$ 维情形下介绍该模型。设设计空间为单位区间 $\mathcal{X} = [0, 1]$，并获得训练数据 $\{(x_i, y_i)\}_{i=1}^n$，其中 $x_1, \cdots, x_n \in \mathcal{X}$ 是训练输入参数，而 $y_1, \cdots, y_n$ 是其对应的输出。我们假设这些输出由以下模型生成：

$$y_i = f(x_i) + \epsilon_i, \quad \epsilon_i \overset{i.i.d.}{\sim} N(0, \theta^2), \quad i = 1, \cdots, n. \tag{5}$$

对于先前的计算机代码仿真问题，确定性模拟器的误差项 ϵ_i 会降为零，因为从 f 获得的观测值不存在噪声。为保持一般性，本文余下部分将采用上述含噪模型设定，并在适当情况下将误差项设为零。

按照式（2），HierGP 假设响应面 f 具有基展开模型：

$$f(x) = \sum_{k=1}^{\infty} \lambda_k \phi_k(x). \tag{6}$$

这里，$\{\phi_k(x)\}_{k=1}^{\infty}$ 是 $\mathcal{X}$ 上的固定 L^2-标准正交基，$\{\lambda_k\}_{k=1}^{\infty}$ 是其对应的系数。随后，我们对这些系数采用如下层次先验设定 λ_k：

$$\lambda_k \sim N(0, \sigma_k^2), \quad \sigma_k^2 \sim w_k \pi_k + (1 - w_k)\delta_{\sigma_\infty^2}, \quad k = 1, 2, \cdots \tag{7}$$

先验 $\sigma_k^2 \sim w_k \pi_k + (1 - w_k)\delta_{\sigma_\infty^2}$ 是尖峰与厚尾先验的简写（Ishwaran 和 Rao，2005），其中 π_k 是厚尾分布，而 $\delta_{\sigma_\infty^2}$（在 σ_∞^2 处的点质量）是尖峰

分布。以概率 w_k，此先验从 π_k中采样；否则，它会在 σ^2_∞处采样一个点质量。第一种情况可视为系数 λ_k 处于激活状态，即对响应面 f有影响；而后一种情况可视为 λ_k 处于非激活状态，即对响应面没有影响。厚尾分布的一种常见选择是 $\pi_k = IG(a_\sigma,\ b_\sigma)$；我们将在第 3 节中为所提出的吉布斯采样器采用这一选择。

对于尖峰与厚尾先验，众所周知，点质量 $\delta_{\sigma^2_\infty}$ 若取 $\sigma^2_\infty = 0$，可能导致计算不稳定性，并使所得的马尔可夫链蒙特卡罗链混合效果不佳（Scheipl 等人，2012）；我们在实现中也遇到了类似问题。常见的变通方法是将 σ^2_∞ 设为一个很小（但非零）的常数，使得 σ^2_∞ 远小于厚尾分布的均值 $E(\pi_k)$。超参数 a_σ、b_σ 和σ^2_∞ 的设定将在后续章节讨论。

当然，在观测数据之前，概率 w_k（λ_k 处于激活状态）通常是未知的。因此，我们为 累积先验分布指定如下形式 $\{w_k\}_{k=1}^\infty$，该形式改编自 Legramanti 等人（2020年）针对稀疏因子模型提出的累积收缩先验。

$$w_k = \prod_{j=1}^{k}(1-\nu_j), \quad \nu_j \overset{i.i.d.}{\sim} Beta(1, \alpha_j = \alpha), \quad w_0 = 1, \quad j = 1, 2, \cdots. \tag{8}$$

为完成设定，我们为噪声方差指定一个独立的逆伽马先验：$\theta^2 \sim IG(a_\theta, b_\theta)$。先验分布 (8) 具有一种理想的累积性质，可满足前文所述响应面 f 中对效应层级的需求。为了解其原因，请注意 $\mathbb{E}(w_k) = (\alpha/(1+\alpha))^k$。因此，对于较小的指标 k，可以看出该先验倾向于赋予 w_k 较大的值；类似地，对于较大的指标 k，它倾向于赋予 w_k 较小的值。这种累积性质恰当地体现了效应层级（Wu 和 Hamada，2009），这是实验分析中广泛采用的一项统计学原理，它认为低阶效应比高阶效应更为重要。先验分布 (7) 通过对低阶系数赋予较大的系数更高的权重、对高阶系数赋予较小的系数更高的权重，直接体现了这一性质。

2.3 多变量 HierGP 模型

在掌握了单变量情形后，我们现在介绍多变量 HierGP 模型。假设设计空间是单位超立方体 $\mathcal{X} = [0,1]^d$，并假设我们收集到训练数据 $\{(\boldsymbol{x}_i,\ y_i)\}_{i=1}^n$，其中 $\boldsymbol{x}_1,\ \cdots,\ \boldsymbol{x}_n \in \mathcal{X}$ 是训练输入参数，而 $y_1,\ \cdots,\ y_n$ 是其对应的输出。和之前一样，我们假设这些输出由以下模型得到：

$$y_i = f(\boldsymbol{x}_i) + \epsilon_i, \quad \epsilon_i \overset{i.i.d.}{\sim} N(0, \theta^2), \quad i = 1, \cdots, n, \tag{9}$$

其中 f 的形式为

$$f(\boldsymbol{x}) = \sum_{\boldsymbol{k}}^{\infty} \lambda_{\boldsymbol{k}} \phi_{\boldsymbol{k}}(\boldsymbol{x}), \quad \boldsymbol{k} = (k_1, \cdots, k_d), \tag{10}$$

其中 $\boldsymbol{k} \in \mathbb{N}_0^d$。这里，$\{\phi_{\boldsymbol{k}}(\boldsymbol{x})\}_{|\boldsymbol{k}|=1}^{\infty}$ 是 $\mathcal{X} = [0,1]^d$上的 L^2-标准正交基，$\{\lambda_{\boldsymbol{k}}\}_{|\boldsymbol{k}|=1}^{\infty}$是其系数。与之前一样，我们对这些系数采用相互独立的尖峰与厚尾先验：

$$\lambda_{\boldsymbol{k}} \overset{indep.}{\sim} N(0, \sigma_{\boldsymbol{k}}^2), \quad \sigma_{\boldsymbol{k}}^2 \sim w_{\boldsymbol{k}}\pi_{\boldsymbol{k}} + (1 - w_{\boldsymbol{k}})\delta_{\sigma_\infty^2}, \quad \boldsymbol{k} \in \mathbb{N}_0^d, \tag{11}$$

其中，$\pi_{\boldsymbol{k}}$ 和 $\delta_{\sigma_\infty^2}$ 分别再次表示厚板分布和尖峰分布。

在前面的单变量设定中，尖峰–厚板概率上的先验（8）体现了效应层级的第一个原则。现在，我们将这一先验推广到多变量设定，以捕捉第二个效应遗传原则：

$$w_{\boldsymbol{k}} = \prod_{m=1}^{d} w_{k_m,m}, \quad w_{k_m,m} = \prod_{j=1}^{k_m}(1 - \nu_{j,m}), \quad w_{0,m} = 1, \quad \nu_{j,m} \overset{i.i.d.}{\sim} Beta(1, \alpha_{j,m} = \alpha), \tag{12}$$

对于 $m = 1$、$\cdots$、d和 $j = 1$，$2, \cdots$。回顾一下，效应遗传原则（Hamada 和 Wu，1992）（或称边际性原则；参见 McCullagh 和 Nelder，1989）假定，只有当交互效应的所有组成变量也处于激活状态时，该交互效应才处于激活状态。在这里，如果其累积项w_{k_m}、$_m$ 较大，则第 j 个变量被视为“激活”，如果w_{k_m}、$_m$ 较小，则被视为“未激活”，因为较大的w_{k_m}、$_m$ 会导致其对应系数为非零的概率更大，反之亦然。要理解上述先验为何体现了这一原则，请注意，对于每个多重指标 $\boldsymbol{k}$，概率 $w_{\boldsymbol{k}}$ 被建模为各个 d 维度分别对应的累积项 w_{k_m}、$_m$ 的乘积。将 $w_{\boldsymbol{k}}$ 设为乘积形式（12）后，可以推出，只有当其所有组成变量（由其累积项表示）均处于激活状态时，项 $\boldsymbol{k}$ 才处于激活状态（即取值较大），这正是效应遗传。因而，可以将先验模型（12）视为一种针对系数的结构化收缩先验，它捕捉了响应面建模所需的层级效应原则。

层级收缩框架（12）可以通过多种方式轻松调整，以纳入建模者提供的其他先验信息。首先，在某些情况下，建模者可能偏好某种基 $\{\phi_{\boldsymbol{k}}(\boldsymbol{x})\}_{|\boldsymbol{k}|=1}^{\infty}$（该基可能不是 L_2-标准正交基），并且希望层级稀疏性在该基下成立。这种非标准正交基可以直接整合到此建模框架和后续的后验抽样器中（不过，我们对标准正交性的假设是后续理论结果所必需的）。其次，如果建模者预期系数具有更重的尾部效应，则可以在尖峰与厚尾先验框架（Carvalho 等人，2009）中轻松采用重尾分布（例如，11）。还可以通过谨慎设定超参数序列 $\alpha_{j,\ m}$ 来纳入更重的尾部。例如，将 $\alpha_{j,\ m}$ 设为递增序列（例如，$\mathcal{O}(m^2)$），便可以在展开的后续项上赋予更大的权重，从而产生更重的尾部行为。

2.4 收缩先验的其他选择：HierGP$_2$

对于某些问题，尽管已知系数 $\lambda_{\boldsymbol{k}}$ 很可能是稀疏的，但我们可能并不十分确信应对这些系数采用层级累积收缩。在这种

一种设定，另一种选 另一种关于 $\lambda_{\boldsymbol{k}}$ 的先验模型可以是以下可交换稀疏先验：

$$\lambda_{\boldsymbol{k}}|\sigma_{\boldsymbol{k}}^2 \sim N(0, \sigma_{\boldsymbol{k}}^2\tau^2), \quad \sigma_{\boldsymbol{k}} \sim \Psi, \quad \boldsymbol{k} \in \mathbb{N}_0^d, \tag{13}$$

其中， Ψ 是一个定义在 $(0, +\infty)$ 上的分布。这类先验称为全局-局部（GL）混合：$\{\sigma_{\boldsymbol{k}}\}_{\boldsymbol{k}}$ 通常称为局部收缩参数，而 τ 称为全局收缩参数。GL 先验为贝叶斯收缩和近似特征选择提供了灵活的框架： Ψ上的重尾分布有助于识别强信号，而其在零附近的集中性则提供了所需的（近似）稀疏行为。GL 先验的例子包括马蹄先验（Carvalho 等人，2010）、狄利克雷–拉普拉斯先验（Bhattacharya 等人，2015）以及广义双帕累托先验（Armagan 等人，2013）。我们将采用 GL 先验（13）的基展开模型（10）称为 HierGP$_2$ 模型；在后续实验中，HierGP$_2$ 采用 Carvalho 等人（2010）提出的马蹄先验作为 Φ。当底层响应面具有稀疏特征，且这些特征不严格遵循效应稀疏性或效应遗传性时，我们将在后文展示，HierGP$_2$ 也能比现有仿真模型取得更好的预测性能。

3 后验抽样

有了 HierGP 后，接下来我们将介绍一种高效的马尔可夫链蒙特卡罗算法，用于在给定数据的情况下对响应面 $f(\cdot)$ 进行后验抽样。我们首先介绍单变量 HierGP 的吉布斯采样器，然后将其扩展为多变量 HierGP 的吉布斯采样器。

3.1 单变量 HierGP 的吉布斯抽样

假设我们从模型（$\{(x_i, y_i)\}_{i=1}^n$（5）收集了训练数据，并令 $\Theta = \{(\lambda_k)_{k=1}^\infty, (v_j)_{j=1}^\infty, \theta^2\}$ 为用于后验推断的参数集。似然函数[1] 可写为：

$$L(\Theta|\{(x_i, y_i)\}_{i=1}^n) = [\{(x_i, y_i)\}_{i=1}^n|\Theta] = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\theta^2}} e^{-\frac{(y_i - f(x_i))^2}{2\theta^2}}, \tag{14}$$

其中，响应面 f 是参数（λ_k）$_{k=1}^\infty$ 的函数（见（6））。给定数据后，我们希望从后验分布中抽取样本 $\{\Theta_{[b]}\}_{b=1}^B$：

$$[\Theta|\{(x_i, y_i)\}_{i=1}^n] \propto L(\Theta|\{(x_i, y_i)\}_{i=1}^n)\,[\Theta]. \tag{15}$$

这里， $[\Theta] = [$（v_j）$_{j=1}^\infty$, （λ_k）$_{k=1}^\infty$, $\theta^2]$ 对于单变量 HierGP，可由式（6）和（7）中的先验模型得到。给定样本 $\{\Theta_{[b]}\}_{b=1}^B$，对 f（记作 $\{f_{[b]}(\cdot)\}_{b=1}^B$）的后验预测样本随后可通过将 $\{\Theta_{[b]}\}_{b=1}^B$ 代入式（6）获得。

[1]这里， $[X]$表示随机变量 X 的分布。

为了从（15）中抽样，我们将采用数据增广吉布斯采样器（Gelman 等人，1995），利用闭式全条件分布进行高效的后验抽样。该采样器的思路与 Legramanti 等人（2020年）针对因子模型提出的吉布斯采样器相似，但针对当前的高斯过程场景进行了调整。如前所述，我们采用 $\pi_k = \pi = IG(a_\sigma, b_\sigma)$ 作为（7）中厚尾分布的选择。我们还对 $f(\cdot)$ 采用以下截断：

$$f(x) = \sum_{k=1}^{K} \lambda_k \phi_k(x) \tag{16}$$

对于足够大的截断指标选择 K；更多细节请参见附录C，其中介绍了如何自适应地选择 K，参见 Legramanti 等人(2020年)。

现在我们推导所需的全条件分布。令 $\boldsymbol{X} = (\phi_1(x_i) \cdots, \phi_K(x_i))_{i=1}^n$为数据的设计矩阵。此外，对于 $k = 1, \cdots, K$，令 z_k为一个潜在分类随机变量，其条件定义为 $\mathbb{P}(z_k = l|\Theta) = \nu_l w_{l-1}$，其中 $l = 1, 2, \cdots, K$。可以证明[2]：

$$[\sigma_k^2|z_k] \sim \{1 - \mathbf{1}(z_k \leq k)\} IG(a_\sigma, b_\sigma) + \mathbf{1}(z_k \leq k)\delta_0, \tag{17}$$

其中**1(·)**为指示函数。利用这一数据增广技巧，可以轻松推导出 z_k 的全条件分布：

$$[z_k = l|-] \propto \begin{cases} \nu_l w_{l-1} \phi(\lambda_k; 0, \sigma_\infty^2), & l = 1, \cdots, k, \\ \nu_l w_{l-1} t_{2a_\sigma}(\lambda_k; 0, (b_\sigma/a_\sigma)), & l = k+1, \cdots, K, \end{cases} \tag{18}$$

其中，$\phi(\lambda; 0, \sigma_\infty^2)$和 $t_{2a}(\lambda; 0, (b_\sigma/a_\sigma))$分别是正态分布和 t分布（自由度为 $2a$ ）在 λ处的密度。

现在考虑参数集 Θ的全条件分布。令 $\boldsymbol{D} = diag\{\sigma_k^2\}_{k=1}^K$和 $\boldsymbol{y} = (y_1, \cdots, y_n)$。对于系数向量 $\boldsymbol{\Lambda} = (\lambda_k)_{k=1}^K$，可以证明其全条件分布为：

$$[\boldsymbol{\Lambda}|-] \sim N(\boldsymbol{V}\theta^{-1}\tilde{\boldsymbol{X}}\boldsymbol{y}, \boldsymbol{V}), \quad \boldsymbol{V} = (\boldsymbol{D}^{-1} + \theta^{-2}\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}})^{-1}. \tag{19}$$

同理，参数 v_j的全条件分布为：

$$[v_j|-] \sim \text{Beta}\left(1 + \sum_{k=1}^{K} \mathbf{1}(z_k = j), \alpha_j + \sum_{k=1}^{K} \mathbf{1}(z_k > j)\right), \quad j = 1, \cdots, K. \tag{20}$$

最后，对于噪声参数 θ^2，其全条件分布可表示为：

$$\left[\theta^2|-\right] \sim IG\left(a_\theta + \frac{N}{2}, \frac{b_\theta + \boldsymbol{S}^T\boldsymbol{S}}{2}\right), \quad \boldsymbol{S} = \boldsymbol{y} - \Lambda\tilde{\boldsymbol{X}}. \tag{21}$$

[2]为简洁起见，记号 $[\theta|-]$表示参数 θ的全条件分布，该分布以数据 $y_1, \cdots, y_n$ 以及 Θ 中除所考虑参数 θ 之外的所有参数为条件。

算法 1 单变量 HierGP 的吉布斯抽样

输入：超参数 $a_\sigma, b_\sigma, a_\theta, b_\theta, \sigma^2_\infty$，数据 $\{(x_i, y_i)\}_{i=1}^n$，迭代次数 B。

1：设置初始参数 $\Theta_{[0]} = \{(\lambda_k^{[0]})_{k=1}^K, (v_j^{[0]})_{j=1}^K, \theta^2_{[0]}\}$。**2：对于** $b = 1, \cdots, B$，**执行 3：对于** $k = 1, \cdots, K$，**执行 4：从式 (18) 中的全条件分布** $[z_k = l|-]$ **采样** $z_k^{[b]}$，**其中** $\lambda_k = \lambda_k^{[b-1]}$ 。**5：结束循环 6：对于** $k = 1, \cdots, K$，**执行 7：如果** $z_k^{[b]} \leq k$，**令** $\sigma_k^2 = \sigma^2_\infty$；**否则，从** $a_\sigma + 0.5, b_\sigma + 0.5(\lambda_k^{[b-1]})^2)$ **中采样** $\sigma_k^2 \sim IG$。**8：结束循环 9：对于** $j = 1, \cdots, K$，**执行 10：从贝塔分布 Beta(**$1 + \sum_{k=1}^K \mathbf{1}(z_k^{[b]} = j)$ $\alpha_j + \sum_{k=1}^K \mathbf{1}(z_k^{[b]} > j)$**)** **中采样** $v_j^{[b]} \sim$；**更新式 (12) 中的** w_j。**11：结束循环 12：采样** $\theta^2_{[b]} \sim IG(a_\theta + N/2,$ $b_\theta + \boldsymbol{S}^T\boldsymbol{S}/2)$，**其中** $\boldsymbol{S} = \boldsymbol{y} - \boldsymbol{\Lambda}_{[b-1]}\tilde{\boldsymbol{X}}$。（步骤 3）**13：采样** $\boldsymbol{\Lambda}_{[b]} = (\lambda_k^{[b]})_{k \leq K} \sim N($ (Step 1)
$\boldsymbol{V}\theta^{-2}_{[b]}\tilde{\boldsymbol{X}}\boldsymbol{y}, \boldsymbol{V})$，**其中** $\boldsymbol{V} = (\boldsymbol{D}^{-1} + \theta^{-2}_{[b]}\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}})^{-1}$ **和** $\boldsymbol{D} = diag(\sigma_k^2)_{k \leq K}$。（步骤 4）**14：结束循环** (Step 2)

这些全条件分布的完整推导见附录 B。

算法1 给出了吉布斯采样器的详细步骤，该采样器结合上述全条件步骤（18）—（21），用于对所需后验分布进行采样$[\Theta|\{(x_i, y_i)\}_{i=1}^n]$。这里，所有采样步骤都相当直接；对于（18），我们通过逆变换采样从该 K 点离散分布中采样，其概率由将（18）中的权重归一化至总和为 1 得到。

3.2 多变量 HierGP 的吉布斯采样

现在，我们将上述吉布斯采样器扩展至多变量 HierGP 模型[3]，该模型利用完整的累积收缩先验（12）。假设我们获得数据 $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$，并令$\Theta = \left\{(\lambda_{\boldsymbol{k}})_{|\boldsymbol{k}|=1}^\infty, (v_{j,m})_{j=1}^{\infty\ d}{}_{m=1}, \theta^2\right\}$为参数集。与之前一样，我们选择$\pi_k = \pi = IG(a_\sigma, b_\sigma)$作为厚尾分布。我们对 $f(\cdot)$ 采用以下截断：

$$f(\boldsymbol{x}) = \sum_{\boldsymbol{k} \leq \boldsymbol{K}} \lambda_{\boldsymbol{k}} \phi_{\boldsymbol{k}}(\boldsymbol{x}), \tag{22}$$

其中，$\boldsymbol{K} = (K_1, \cdots, K_d)$ 是每个维度对应的截断指标向量。同样，这些指标 $K_1, \cdots, K_d$ 应设置得足够大（后文将提供更多细节

[3]从此处起，我们将多变量 HierGP 模型简称为“HierGP 模型”。

算法 2 多变量 HierGP 的吉布斯抽样 输入：超参数 a_σ, b_σ, a_θ, b_θ, σ_∞^2, **数据** $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, **迭代次数** B。**设置初始参数** $\Theta_{[0]} = \{(\lambda_{\boldsymbol{k}}^{[0]})_{\boldsymbol{k}\le \boldsymbol{K}}, (v_j^{[0]}, _m)_{j\le K_m, m\le d}, \theta_{[0]}^2\}$。**对于** $b = 1, \cdots, B$，**执行 3：对于** $\boldsymbol{k}$，**取值范围为** $(1, \cdots, 1) : (K_1, \cdots, K_d)$，**执行：从 (24) 中按** $\lambda_{\boldsymbol{k}} = \lambda_{\boldsymbol{k}}^{[b-1]}$ **抽取** $\boldsymbol{z}_{\boldsymbol{k}}^{[b]}$。**结束循环。6：对于** $\boldsymbol{k} = (k_1, \cdots, k_d)$，**取值范围为** $(1, \cdots, 1) : (K_1, \cdots, K_d)$，**执行：如果** $\exists m$：$z_{\boldsymbol{k}}^{[b]}, _m \le k_m$，**则令** $\sigma_{\boldsymbol{k}}^2 = \sigma_\infty^2$；**否则** $\sigma_{\boldsymbol{k}}^2 \sim IG(a_\sigma + 0.5, b_\sigma + 0.5(\lambda_{\boldsymbol{k}}^{[b-1]})^2)$。**9：结束循环。对于** m，**取值范围为** $1, \cdots, d$，**执行：对于** j，**取值范围为** $1, \cdots, K_m$，**执行：12：抽取** $v_j^{[b]}, _m \sim \mathbf{Beta}(1 + \sum_{\boldsymbol{k}\le \boldsymbol{K}} \mathbf{1}(z_{\boldsymbol{k}}^{[b]}, _m = j)$ $\alpha_j + \sum_{\boldsymbol{k}\le \boldsymbol{K}} \mathbf{1}(z_{\boldsymbol{k}}^{[b]}, _m > j))$。**结束循环。结束循环。15：根据 (12) 更新** $w_j, _m$ **和** $w_{\boldsymbol{k}}$。**根据** $IG(a_\theta + N/2, b_\theta + S^T S/2)$ **更新** $\theta_{[b]}^2$，**其中** $S = \boldsymbol{y} - \Lambda_{[b-1]}\tilde{\boldsymbol{X}}$。**（步骤 3）抽取** $\boldsymbol{\Lambda}_{[b]} = (\lambda_{\boldsymbol{k}}^{[b]})_{\boldsymbol{k}\le \boldsymbol{K}} \sim N(\boldsymbol{V}\theta_{[b]}^{-2}\tilde{\boldsymbol{X}}\boldsymbol{y}, \boldsymbol{V})$，**其中** $\boldsymbol{V} = (\boldsymbol{D}^{-1} + \theta_{[b]}^{-2}\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}})^{-1}$ **和** $\boldsymbol{D} = diag(\sigma_{\boldsymbol{k}}^2)_{\boldsymbol{k}\le \boldsymbol{K}}$。**（步骤 4）18：结束循环**

(Step 1)

(Step 2)

见附录）。

现在，我们推导多变量 HierGP 中 Θ 的类似闭式全条件分布。设 $\tilde{\boldsymbol{X}} = (\phi_{\boldsymbol{k}}(\boldsymbol{x}_i))_{\boldsymbol{k}\le \boldsymbol{K},\ i=1,\ \cdots,\ n} \in \mathbb{R}^{n\times\|\boldsymbol{K}\|}$ 为数据的设计矩阵，其中 $\boldsymbol{k} \in \mathbb{N}_0^d$ 是遍历 $\boldsymbol{K}$ 和 $\|\boldsymbol{K}\| = \prod_{m=1}^d K_m$ 的多重指标。此外，令 $\boldsymbol{z}_{\boldsymbol{k}}$ 为潜在随机变量向量，其条件定义如下：

$$\mathbb{P}\{\boldsymbol{z}_{\boldsymbol{k}} = (l_1, \cdots, l_d) | w_{l_1,1}, \cdots, w_{l_d,d}\} = \prod_{m=1}^d v_{l_m,m} w_{l_m-1,m}, \tag{23}$$

其中 $w_{l_m,\ m}$ 的定义见（12）。据此，我们可以利用类似的数据增广技巧，推导 $\boldsymbol{z}_{\boldsymbol{k}} = [z_{\boldsymbol{k},1}, \cdots, z_{\boldsymbol{k},d}]$ 的全条件分布如下：

$$[\boldsymbol{z}_{\boldsymbol{k}} = \boldsymbol{l}|-] \propto \begin{cases} \left(\prod_{m=1}^d v_{l_m,m} w_{l_m-1,m}\right) \phi(\lambda_{\boldsymbol{k}}; 0, \sigma_\infty^2), & \text{otherwise,} \\ \left(\prod_{m=1}^d v_{l_m,m} w_{l_m-1,m}\right) t_{2a_\sigma}(\lambda_{\boldsymbol{k}}; 0, (b/a)), & l_1 > k_1, \cdots, l_d > k_d \end{cases} \tag{24}$$

通过对所有$\sigma_{\boldsymbol{k}}^2$进行边缘化。关于此边缘化的详细信息见附录B。

现在考虑参数集 Θ的全条件分布。令$\boldsymbol{D} = diag(\sigma^2_{\boldsymbol{k}})_{\boldsymbol{k}\leq\boldsymbol{K}}$。对于系数向量 $\boldsymbol{\Lambda} = (\lambda_{\boldsymbol{k}})_{\boldsymbol{k}\leq\boldsymbol{K}}$，可以证明其全条件分布与式（19）形式相同。参数 v_j^m的全条件分布如下：

$$[v_{j,m}|-] \sim \text{Beta}\left(1 + \sum_{\boldsymbol{k}\leq\boldsymbol{K}} \mathbf{1}(z_{\boldsymbol{k},m} = j), \alpha_j + \sum_{\boldsymbol{k}\leq\boldsymbol{K}} \mathbf{1}(z_{\boldsymbol{k},m} > j)\right),\ j = 1, \cdots, K_m,\ m = 1, \cdots, d. \tag{25}$$

最后， θ^2 的完全条件分布与式（21）中的形式相同。算法2 详细列出了该吉布斯抽样算法的步骤。与前文一样，可以选择足够大的截断指标 $\boldsymbol{K}$，也可以自适应地选择 $\boldsymbol{K}$；关于后一种方法的更多细节见附录C，遵循Legramanti 等人 (2020年）。

我们简要分析算法 2 中吉布斯采样器的计算复杂度。步骤 4 的计算成本可表示为 $\mathcal{O}(dn\|\boldsymbol{K}\| + (\|\boldsymbol{K}\|)^3)$，因为该步骤需要根据 n 个观测和 $\|\boldsymbol{K}\|$ 个基函数计算 $\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}}$，并通过矩阵求逆来计算 $\boldsymbol{V}$。步骤 3 的计算成本为 $\mathcal{O}(|\boldsymbol{K}|)$，因为我们要更新 $O(\|\boldsymbol{K}\|)$ 个变量，且每次更新的成本为 $O(1)$。步骤 2 的成本为 $\mathcal{O}(\|\boldsymbol{K}\|^2)$，因为它需要遍历 $\|\boldsymbol{K}\|$ 个变量，每次抽样都需要 $\mathcal{O}(\|\boldsymbol{K}\|)$ 的计算量。类似地，步骤 1 的成本可表示为 $O(\|\boldsymbol{K}\|^2)$。综上，该吉布斯采样器的计算复杂度为 $O\{B(dn\|\boldsymbol{K}\| + \|\boldsymbol{K}\|^3)\}$，其中 B 为吉布斯迭代次数。

3.3 HierGP 的吉布斯采样$_2$

在预期基函数系数具有稀疏性，但对效应层级或遗传性没有强先验信念的情况下，$_2$ 第 2.4 节中的另一种 HierGP 模型（采用 GL 收缩先验）可能是一个颇具吸引力的替代方案。接下来，我们针对这一替代模型给出一个类似的吉布斯采样器，其中 Φ采用 Carvalho 等人（2010）的马蹄先验。此处，参数集 $\Theta = \{(\lambda_k)_{k=1}^{\infty}, \theta^2\}$的后验抽样简化为与采用马蹄先验的贝叶斯线性回归相同的情形；针对后者已有现成的后验抽样算法。算法 3 将 Johndrow 等人（2020年）提出的分块 Metropolis-within-Gibbs 采样器直接扩展至 HierGP$_2$模型。

4 一致性结果

现在，我们给出若干理论结果，以确认（多元）HierGP 和 HierGP$_2$ 的一致性。这些结果扩展了现有关于高维线性回归一致性的理论，具体而言，包括 Song 和 Liang （2017）以及 Choi 和 Schervish （2007）的结果。如前所述，对于 HierGP，我们将采用点质量$\pi_k = \delta_{\sigma^2_{k,0}}$ 作为厚尾分布，其中$\{\sigma^2_{\boldsymbol{k}}\}$ 是预先指定的；这是一种常用的做法

算法 3 用于 HierGP$_2$ 的分块 Metropolis-within-Gibbs 采样器

输入：超参数 a_θ、b_θ、τ，数据 $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$，迭代次数 B。

1：设置初始参数 $\Theta_{[0]} = \{(\lambda_k^{[0]})_{k=1}^K, (\theta^{[0]})^2\}$。**2：对于** $b = 1, \cdots, B$，**执行 3：从** $N(\boldsymbol{V}(\theta^{[b-1]})^{-1}\tilde{\boldsymbol{X}}^T\boldsymbol{y}, \boldsymbol{V})$、$V = (\boldsymbol{D}^{-1} + (\theta^{[b-1]})^{-2}\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}})^{-1}$ **和** $\boldsymbol{D} = \mathrm{diag}((\sigma_1^{[b-1]})^2, \cdots, (\sigma_K^{[b-1]})^2)$ **中采样** $\boldsymbol{\Lambda}^{[b]} = (\lambda_1^{[b]}, \cdots, \lambda_K^{[b]})$。**（步骤 1）4：通过 Metropolis-Hastings 从与** $\prod_{k=1}^K \left(\frac{1}{1+\sigma_k}\right) \exp(-(\lambda_k^{[b]})^2 (\theta^{[b]})^2 / (2\tau\sigma_k))$ **成正比的密度中采样** $\{\sigma_k^{[b]}\}_{k=1}^K$。**5：根据** $IG(a_\theta + N/2, (b_\theta + \boldsymbol{S}^T\boldsymbol{S})/2)$ **更新** $(\theta^{[b]})^2$，**其中** $\boldsymbol{S} = \boldsymbol{y} - \boldsymbol{\Lambda}^{[b]}\tilde{\boldsymbol{X}}$。(Step 2)

（步骤 3）6：结束循环

在理论分析文献中。我们进一步证明了带有马蹄先验的 HierGP$_2$ 模型的形状近似满足伯恩斯坦-冯·米塞斯型Song 和 Liang (2017定理，将相关理论从 Song 和 Liang 扩展到了当前的 GP 设置。主要定理及相应的假设将在下文给出，证明则推迟至附录。

4.1 HierGP 模型

为了证明第 2.3节所提出的 HierGP 模型的一致性，我们需要以下一些假设：

(A1) 基表示：令 $f_0(\cdot)$ 表示我们希望预测的真实函数。我们假设 f_0 具有以下形式

$$f_0(\boldsymbol{x}) = \sum_{\boldsymbol{k} \in S} \lambda_{\boldsymbol{k}}^0 \phi_{\boldsymbol{k}}(\boldsymbol{x}), \tag{26}$$

其中，$S \subset \mathbb{N}_0^d$ 是一个大小为 $s = \mathrm{card}(S)$ 的有限索引集。这进一步意味着 $f_0 \in L^2(\mathcal{X}) \bigcap C^1(\mathcal{X})$。

(A2) 光滑性：我们进一步假设 $a_{\boldsymbol{k}} = \sup_{\boldsymbol{x}\in\mathcal{X}} |\phi_{\boldsymbol{k}}(\boldsymbol{x})| < \infty$、$b_{\boldsymbol{k}} = \sup_{\boldsymbol{x}\in\mathcal{X}} |\phi'_{\boldsymbol{k}}(\boldsymbol{x})| < \infty$、$\sum_{\boldsymbol{k}} a_{\boldsymbol{k}}\sigma_{\boldsymbol{k},0} < \infty$ 以及 $\sum_{\boldsymbol{k}} b_{\boldsymbol{k}}\sigma_{\boldsymbol{k},0} < \infty$。这些假设控制了底层函数 $f_0(\boldsymbol{x})$ 的光滑性。

(A3) 随机设计：设计点 $\{\boldsymbol{x}_i\}_{i=1}^\infty$ 是从某个概率分布 P_0 中独立同分布抽样得到的，该分布定义在 $\mathcal{X} = [0,1]^d$上。我们将在定理 3中使用这一假设。

（A4）固定设计：设计点 $\{\boldsymbol{x}_i\}_{i=1}^\infty$ 满足以下条件。对于 $\mathcal{X} = [0,1]^d$中的每个超立方体 $\mathcal{H}$，都存在一个常数 $0 < K_d \leq 1$，使得当其勒贝格测度 $\lambda(\mathcal{H}) \geq (K_d n)^{-1}$时，该超立方体 $\mathcal{H}$ 至少包含一个设计点。这可以看作是数字网平衡条件的扩展

准蒙特卡罗中的抽样（例如参见 Owen，1997；Dick 等人，2013），这确保设计点在 $\mathcal{X}$上均匀分布。该假设将在定理 2中使用。

(A5) 先验修正：在固定设计条件下，为证明 L_1-一致性，我们还需要对先验进行小幅修正。令 Π_1^* 表示第2.3节中 HierGP 模型的先验，并令 V为满足以下条件的常数： $V > \sup_{[0,1]^d} |f_0'|_\infty$ 定义 $\Omega = \{f : \ |f'|_\infty < V\}$。于是，修正先验模型可定义为$\Pi_2^*(\cdot) = \Pi_1^*(\cdot \bigcap \Omega)/\Pi_1^*(\Omega)$，其中 $\Pi_1^*(\Omega) > 0$。这是证明高维贝叶斯线性回归一致性时所采用的标准修正方法（例如参见 ChoiandSchervish (2007)）。

接下来，我们介绍几种用于证明一致性的邻域，其中两种用于随机设计，另外两种用于固定设计。这里，令 σ_0^2 表示观测值的真实噪声方差。

1. 海灵格邻域：

$$H_\epsilon = \{(f, \sigma) | d_H(p, p_0) < \epsilon\}$$

其中， d_H是海灵格距离 $d_H(p_1, \ p_2) = \int (\sqrt{p_1} - \sqrt{p_2})^2 d\xi$。这里， p是 $(\boldsymbol{x}, \ y)$ 关于 $\xi = P_0 \times \lambda$ 的概率密度函数，即 $p(x, \ y) = \phi([y - f(x)]/\sigma)/\sigma$ ，其中 ϕ 是标准正态密度， λ 是勒贝格测度。

2. 经验测度邻域：

$$W_{\epsilon,n} = \left\{ (f, \sigma) : \int |f(\boldsymbol{x}) - f_0(\boldsymbol{x})| \, dP_n(\boldsymbol{x}) < \epsilon, \left| \frac{\sigma}{\sigma_0} - 1 \right| < \epsilon \right\}$$

其中， $P_n(x) = n^{-1} \sum_{i=1}^n I_{x_i}(x)$ 是经验测度。

3. L_1-邻域：

$$L_\epsilon = \left\{ (f, \sigma) : \int |f(x) - f_0(x)| dx < \epsilon, \left| \frac{\sigma}{\sigma_0} - 1 \right| < \epsilon \right\}$$

4. P_0-邻域：

$$U_\epsilon = \left\{ (f, \sigma) : d_{P_0}(f_0, f) < \epsilon, \left| \frac{\sigma}{\sigma_0} - 1 \right| < \epsilon \right\}$$

其中 $d_{P_0}(f, g) =$下确界$\{\epsilon : \ P_0(\{\boldsymbol{x} : \ |f(\boldsymbol{x}) - g(\boldsymbol{x})|_\infty > \epsilon\}) < \epsilon\}$。

现在我们可以证明一致性结果，先考虑固定设计，再考虑随机设计。

定理 2（HierGP 的相合性，固定设计）。设 f 服从上述修正后的 *HierGP* 先验Π_2^*。设 Q_0 为给定固定设计点$\{\boldsymbol{x}_i\}_{i=1}^n$时数据 $\{y_i\}_{i=1}^n$ 的条件分布。假设（A1）、（A2）、（A4）和（A5）成立。那么：

(a) 对于任意 $\epsilon > 0$，我们有：

$$\Pi_2^*\left((f,\sigma) \in W_{\epsilon,n}^c | \{y_i, \boldsymbol{x}_i\}_{i=1}^n\right) \to 0, \quad [Q_0] \ almost \ surely. \tag{27}$$

(b) 我们还有：

$$\Pi_2^*(L_\epsilon^c | \{y_i, \boldsymbol{x}_i\}_{i=1}^n) \to 0, \quad [Q_0] \ almost \ surely. \tag{28}$$

定理 3（HierGP 的相合性，随机设计）。设 f 遵循第 Π_1^*节中的 *HierGP*先验 *2.3*。设 Q_0 表示 $\{\boldsymbol{x}_n, y_n\}_{n=1}^\infty$ 的联合分布。假设 (A1)、(A2) 和 (A3) 均成立。那么，对于遵循测度 P_0 的任意随机设计，我们有

$$\Pi_1^*(U_\epsilon^c | \{y_i, \boldsymbol{x}_i\}_{i=1}^n) \to 0, \quad [Q_0] \ almost \ surely, \tag{29}$$

and

$$\Pi_1^*(H_\epsilon^c | \{y_i, \boldsymbol{x}_i\}_{i=1}^n) \to 0, \quad [Q_0] \ almost \ surely. \tag{30}$$

定理 2 和3 表明，在正则性条件下，随着样本量趋于无穷，回归函数和方差参数的后验分布在适当的拓扑下确实收敛于真值。这证明了所提出模型的后验一致性，并保证所采用的收缩先验对假设 (A1) 中所述的函数类提供了足够的支持。这些定理的证明扩展了 Choiand Schervish (2007) 中的结果；详情见附录。

我们对为何 HierGP 模型仅证明了一致性作出一些解释。尽管关于标准收缩先验的收缩率已有大量研究，但 HierGP 中采用的层级累积收缩先验相当新颖。针对高维贝叶斯线性回归的标准分析工具（例如，Castillo 等人，2015，Song 和 Liang，2017，Jeong 和 Ghosal，2020年）并不适用于此情形，因为我们的收缩先验具有高度结构化的特性。因此，我们着重证明模型在这一新颖先验设定下的一致性，并将更复杂的收缩率问题留待未来研究。

4.2 TheHierGP$_2$ 模型

现在给出第$_2$ 2.4节和第3.3节中 HierGP 的一些结果。令（$\boldsymbol{\lambda^0}$，θ_0）=（（λ_1^0，$\cdots$，λ_K^0）θ_0）表示模型的真实系数和方差，并令 $\boldsymbol{\lambda}_{S'}$，$\tilde{\boldsymbol{X}}_{S'}$ 表示索引子集 $S' \in$（1，$\cdots$，K）对应的系数和设计矩阵。这里，我们需要以下假设：

(B1) 截断水平：设在样本量为 n 时采用的截断水平为 $K_{[n]}$。我们假设 $K_{[n]} \geq n$，即截断水平始终大于样本量。

(B2) 设计条件：存在常数整数 C（依赖于 n 和 $K_{[n]}$）以及标量 λ_0，使得 $C > s$，并且对于满足$|S'| \leq C$的任意子集 $S' \subset \{1, \cdots, K_{[n]}\}$， $\lambda_{\min}(\tilde{\boldsymbol{X}}_{S'}^T \tilde{\boldsymbol{X}}_{S'}) \geq n\lambda_0$，其中 $s = \text{card}(S)$ 是真实模型 S 的活跃指标集的大小。

(B3) 稀疏性：我们进一步假设 $s\log(K_{[n]}) \leq n$。这可以视为对真实函数 f_0的稀疏性假设。

(B4) 系数幅度：假设 $\max|\lambda_k^0/\theta_0| \leq \gamma_3 E_n$，其中某个固定的 $\gamma_3 \in (0, 1)$，且 E_n 关于 n 非递减。这为基函数系数的幅度提供了上界。

有了这些，我们便可以证明 HierGP$_2$ 模型的收缩和伯恩斯坦-冯·米塞斯（BvM）形状近似结果。

定理 4（HierGP 的收缩与 BvM 定理$_2$）。设 f 遵循第 $_2$ 节中的 *HierGP*Π 先验2.4。设 P^* 是在真实函数 $\{\boldsymbol{x}_i, y_i\}_{i=1}^n$ 下，$f_0(\cdot)$ 且噪声方差为 θ_0^2时的联合测度。假设 （B1）-（B5）成立，则：

（*a*）（后验收缩）有

$$P^*\left(\Pi\left(||f - f_0||_2 \geq c_1\theta_0\epsilon_n|\{\boldsymbol{x}_i, y_i\}_{i=1}^n\right) \geq e^{-c_2 n\epsilon_n^2}\right) \leq e^{-c3n\epsilon_n^2}, \tag{31}$$

其中某些正常数为 c_1、 c_2 和 c_3， $\epsilon_n = M\sqrt{s\log(K_{[n]})/n}$ 是收缩速率，而 M 是一个固定常数。

（*b*）（BvM 形状近似）后验分布 $\Pi([f - f_0](\boldsymbol{x}) \ \theta^2|\{\boldsymbol{x}_i, y_i\}_{i=1}^n)$ 在全变差意义下收敛到分布

$$\left[\sum_{k=1}^{K_{[n]}} \lambda_k\phi_k(\boldsymbol{x})\middle|\theta^2\right] [\theta^2], \tag{32}$$

其中

$$\begin{aligned}
[\{\lambda_k\}_{k=1}^{K_{[n]}}|\theta^2] &= \phi\left(0; \hat{\lambda}_S, \theta^2(\tilde{X}_S^T\tilde{X}_S)^{-1}\right)\prod_{k\notin S}\pi(\lambda_k|\theta^2), \\
[\theta^2] &\sim IG\left(\frac{n-s}{2}, \frac{\hat{\theta}^2(n-s)}{2}\right),
\end{aligned} \tag{33}$$

$\pi(\lambda_k|\theta^2)$ 是 (13)中的条件先验分布，且 $\hat{}\ \lambda_S$ 和 $\hat{}\theta^2$ 是基于数据λ_S 和 θ^2 得到的最大似然估计 $\{\boldsymbol{X}_S, y_i\}_{i=1}^n$。

定理 4表明，在满足正则性条件时，回归函数和方差参数的后验分布会以 ϵ_n 的速率集中于真值附近，并且随着样本量趋于无穷大，在适当的拓扑下渐近正态。后者称为“伯恩斯坦-冯·米塞斯”型定理。与 HierGP 先前的后验一致性结果相比，这类结果更强，因为它们不仅给出了明确的后验收缩速率，还阐明了随样本量变化的截断水平自适应选择。该定理的证明扩展了 Song 和 Liang (2017的结果；详细内容见附录。

5 数值实验

现在，我们通过一系列数值实验来探究所提出的 HierGP 和 HierGP$_2$ 。我们首先考察它们在计算机代码仿真中的表现，然后展示它们在动力系统恢复方面的有效性。

5.1 计算机代码仿真

在我们的计算机代码仿真实验中，我们将考虑一系列测试函数，并将所提出的模型（HierGP 和 HierGP$_2$）与几种流行的和/或相关的基于GP的仿真器进行比较。这包括采用 Mat´ern-3/2 核的标准 GP 仿真器（ Stein， 1999 ）；Lu 等人提出的加性 GP 模型 （ 2022），该模型基于近期利用加性低维结构的研究成果（Duvenaud 等人，2011）；“最小二乘”模型，它利用数据的最小二乘拟合 $\{(\boldsymbol{x}_i,\ y_i)\}_{i=1}^n$，使用预先˜指定的基矩阵 $\boldsymbol{X}$；以及稀疏最小二乘拟合，它在相同设置下使用 l_1- 正则化估计（通过交叉验证调整惩罚参数）。后两种方法类似于多项式混沌中使用的代理模型；全面综述请参见 Luthen 等人（2021）。这两种模型为 HierGP 提供了有用的基准，有助于凸显在高斯过程中嵌入效应稀疏性、层次性和遗传性的潜在优势。为确保比较公平，两种模型均使用与 HierGP 相同的基矩阵。

首先考虑一个简单情形：测试函数由第 3节中的 HierGP 先验模拟得到，适用于 $d = 2$ 和 $d = 3$ 维。该模拟采用参数 $\alpha = 6$、$a_\sigma = 1$、$b_\sigma = 1$、$\sigma_\infty = 0$，并且对于 $d = 2$ 和 $d = 3$ 分别使用截断上限 $\boldsymbol{K} = (8,\ 8)$ 和 $(4,\ 4,\ 4)$，以及正弦基 $\phi_{\boldsymbol{k}}(\boldsymbol{x}) = \Pi_{m=1}^d \sin(2\pi k_m x_m)$。因此，模拟函数体现了预设的效应稀疏性、遗传性和层次性原则。模型训练时，我们使用 $n = 70$ 个均匀采样的设计点。对于 HierGP、最小二乘和稀疏最小二乘，我们假设处于完全正确设定情形，即基矩阵和截断水平与模拟模型相同；接下来将探讨错误设定情形。随后重复该模拟 50 次，以衡量误差变异性。

图 1 和 2 展示了各方法在 $d = 2$ 维情形下，对 400 个均匀采样测试点的预测误差箱线图。可以看到，HierGP 的预测效果优于其他模型。这并不令人意外——当 f 中存在结构化稀疏性时，HierGP 能有效利用这种结构来改善预测。HierGP$_2$（它对基函数系数采用结构性较弱的马蹄先验）表现也相当不错，但误差略高，因为它没有纳入关于效应遗传性和层次性的先验信息。标准 Mat´ern 高斯过程和加性高斯过程的预测性能都显著较差；这是符合直觉的，因为这类模型没有在 f 中嵌入结构化稀疏性。最后，两种最小二乘拟合（即使基完全正确指定）表现也不佳，这在意料之中，因为尽管基设定完全正确，这类拟合仍无法捕捉所需的结构化稀疏性。

表1 通过报告 95% 后验预测可信区间的经验覆盖率及相应的平均预测区间宽度，比较了这些模型的不确定性量化表现。这里，我们仅比较了两种 HierGP 模型与标准高斯过程，因为其他方法并非完全概率化。对于 $d = 2$ 和 $d = 3$，我们看到，HierGP 模型的经验覆盖率明显高于标准高斯过程；标准高斯过程的覆盖率甚至可能低于名义上的 95%。此外，HierGP 模型的可信区间宽度显著小于标准高斯过程。这表明，当 f 具有预期的层次稀疏结构时，所提出的模型确实能够利用这种结构，提供更精确的概率预测，且其覆盖率优于标准高斯过程，正如预期。

接下来，考虑文献中两个合成测试函数的仿真，即 Branin 函数（Sobester 等人，2008）：

$$f(\mathbf{x}) = a(x_2 - bx_1 + cx_1 - r)^2 + s(1-t)\cos(x_1) + s,$$

其中 $a = 1$，$b = 5.1/(4\pi^2)$，$c = 5/\pi$，$r = 6$，$s = 10$ 以及 $t = 1/\pi$，以及 Cheng & Sandu 函数（Cheng 和 Sandu，2010）：

$$f(\mathbf{x}) = \cos(x_1 + x_2)\exp(x_1 x_2).$$

这两个函数在遵循效应遗传和层次原则方面的程度有所不同。在此，HierGP、最小二乘法和稀疏最小二乘法采用上述正弦基，截断上限为 $\boldsymbol{K} = (8, 8)$。这为检验所提出的模型在相对于所选基函数的效应原则存在轻微违反时的稳健性提供了良好的测试。与之前一样，使用 $n = 70$ 均匀采样的设计点进行训练。

图3 展示了每个测试函数在 400 个均匀采样的测试点上的预测误差箱线图。对于 Branin 函数，在所有考察的模型中，HierGP 的预测性能最佳，而 HierGP$_2$ 是一个实力相近的竞争者。进一步考察后，这并不令人意外，因为其函数形式表明存在某种形式的效应遗传和层次性。对于 Cheng & Sandu函数，我们看到

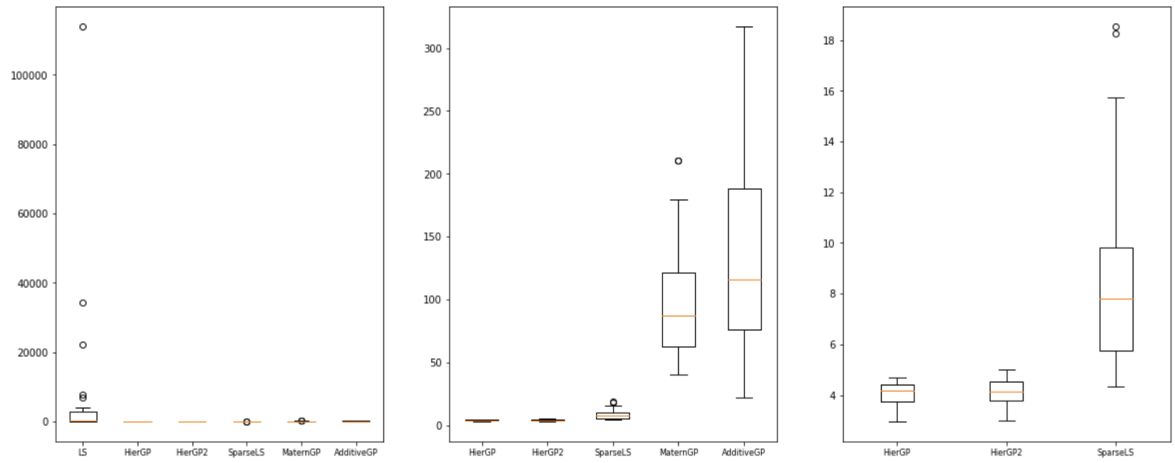


图 1：当 f 由 $d = 2$ 维的 HierGP 模拟生成时，各比较方法的预测误差箱线图。从左至右为箱线图的放大版本。

	HierGP	HierGP2	Matérn高斯过程
经验覆盖率（$d = 2$）	100.0%	100.0%	91.0%
经验覆盖率 ($d = 3$)	97.7%	99.1%	92.2%
平均可信区间宽度 ($d = 2$)	1.721	1.313	4.062
平均可信区间宽度 ($d = 3$)	1.277	1.798	7.250

表1：当 f 是从 d 维的 HierGP 中模拟生成时，95% 后验预测可信区间的经验覆盖率及相应的平均区间宽度。

HierGP_2 的预测性能最佳，HierGP 则是一个实力相近的竞争者。这可以用其函数形式中更复杂的交互结构来解释；这种结构可能难以通过 HierGP 中假定的结构化稀疏性识别出来。无论如何，上述实验表明，当 f 中存在效应层级和遗传性时（即使存在轻微违背），所提出的模型也能够学习并整合这种结构，从而提升预测性能。

5.2 动力系统恢复

现在，我们进一步探究 HierGP 在动力系统恢复与预测问题中的应用。这类问题广泛应用于气候学、生态学和金融学（例如，参见 Ghadami 和 Epureanu，2022，Luo 等人，2011，Mudelsee，2019）。我们首先参照 Brunton 等人 （2016）简要回顾这一问题。本文考虑如下动力系统

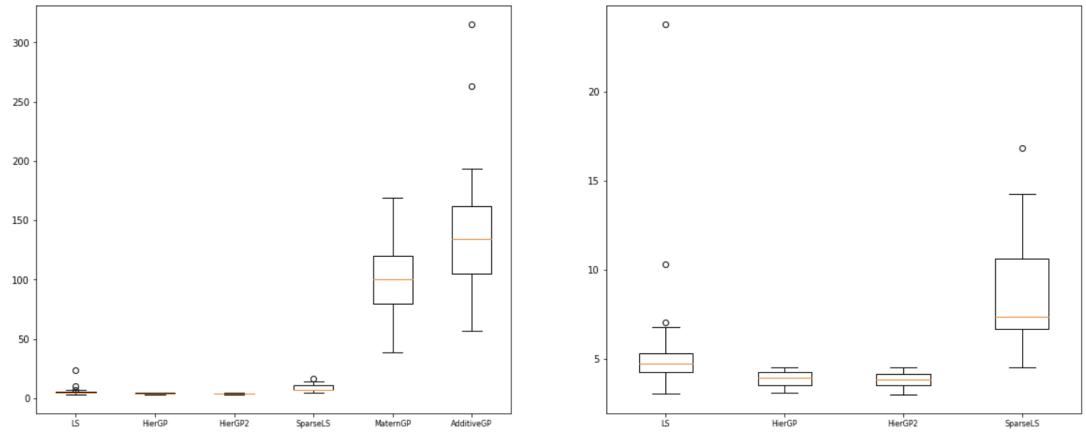


图2：当 f 在 $d=3$ 维度下由 HierGP 模拟时，各比较方法的预测误差箱线图。从左至右依次为箱线图的放大版本。

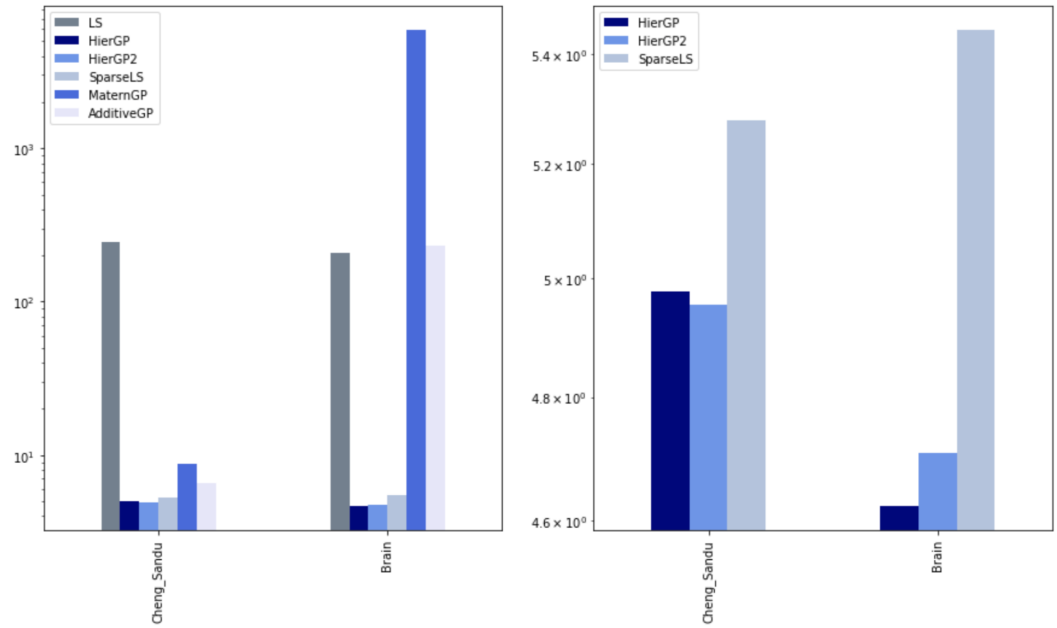


图3：当 f 取 Branin 函数和 Cheng & Sandu函数时，各比较方法的平均预测误差柱状图。从左至右依次为柱状图的放大版本。

(Guckenheimer 和 Holmes，2013），其形式为：

$$\frac{d}{dx}\boldsymbol{x}(t) := \dot{\boldsymbol{x}}(t) = \boldsymbol{f}(\boldsymbol{x}(t)), \tag{34}$$

这里，$\boldsymbol{x}(t) \in \mathbb{R}^q$ 表示系统在时间 t 的状态，而 $\boldsymbol{f}(\boldsymbol{x}(t))$ 表示支配系统状态运动方程的动力学约束。公式（34）涵盖了生态学、生物学及其他科学学科中使用的广泛动力系统（Delahunt 和 Kutz，2022）。

现在考虑这样的情形：在系统状态上观测到数据 $\{(\boldsymbol{x}(t_i)\ \dot{\boldsymbol{x}}(t_i)\}_{i=1}^n$，其中 t_1，$\cdots$，t_n 是采样时间点。我们可以将其重新排列为状态矩阵：

$$\boldsymbol{X} = (\boldsymbol{x}^T(t_1), \cdots, \boldsymbol{x}^T(t_n))^T = \begin{pmatrix} x_1(t_1) & x_2(t_1) & \cdots & x_d(t_1) \\ x_1(t_2) & x_2(t_2) & \cdots & x_d(t_2) \\ \cdots & \cdots & \ddots & \vdots \\ x_1(t_n) & x_2(t_n) & \cdots & x_d(t_n) \end{pmatrix} \in \mathbb{R}^{n\times d},$$

$$\dot{\boldsymbol{X}} = (\dot{\boldsymbol{x}}^T(t_1), \cdots, \dot{\boldsymbol{x}}^T(t_n))^T = \begin{pmatrix} \dot{x}_1(t_1) & \dot{x}_2(t_1) & \cdots & \dot{x}_d(t_1) \\ \dot{x}_1(t_2) & \dot{x}_2(t_2) & \cdots & \dot{x}_d(t_2) \\ \cdots & \cdots & \ddots & \vdots \\ \dot{x}_1(t_n) & \dot{x}_2(t_n) & \cdots & \dot{x}_d(t_n) \end{pmatrix} \in \mathbb{R}^{n\times d}.$$

接下来，我们将构建一个候选函数“函数库”，用于恢复 $\boldsymbol{f}$ （34）中的函数，从而恢复潜在的系统动力学。假设这些候选函数 $\mathcal{F} = \{\phi_{\boldsymbol{k}}(\boldsymbol{x})\}_{\boldsymbol{k}}$ 由多重指标 $\boldsymbol{k} = (k_1、\ \cdots、\ k_d)$ $\boldsymbol{k} \le \boldsymbol{K}$参数化。给定采样时间点后，该函数库可表示为模型矩阵：

$$\Phi(\boldsymbol{X}) = \begin{pmatrix} | & \cdots & | & \cdots & | \\ \Phi(\boldsymbol{X})^{[\boldsymbol{1}]} & \cdots & \Phi(\boldsymbol{X})^{[\boldsymbol{k}]} & \cdots & \Phi(\boldsymbol{X})^{[\boldsymbol{K}]} \\ | & \cdots & | & \cdots & | \end{pmatrix} \in \mathbb{R}^{n\times\|\boldsymbol{K}\|},$$

其中，$\|\boldsymbol{K}\| = \prod_{m=1}^d K_m$ 是 $\mathcal{F}$中的基函数总数，$\Phi(\boldsymbol{X})^{[\boldsymbol{k}]}$ 是由 $\mathcal{F}$ 中多重指标为 $\boldsymbol{k}$的基函数构成的模型矩阵。

动力系统（34）可表示为以下线性方程组：

$$\dot{\boldsymbol{X}} = \Theta(\boldsymbol{X})\boldsymbol{\Xi} \tag{35}$$

其中，$\boldsymbol{\Xi} = (\boldsymbol{\xi}_1,\ \cdots,\ \boldsymbol{\xi}_d) \in \mathbb{R}^{\|\boldsymbol{K}\|\times d}$ 是 $\boldsymbol{f}$的系数矩阵，而 $\boldsymbol{\xi}_m \in \mathbb{R}^{\|\boldsymbol{K}\|}$ 是 $\boldsymbol{f}$中第 m个分量的系数向量。

给定数据矩阵 $\boldsymbol{X}$ 和 $\dot{\boldsymbol{X}}$，恢复 $\boldsymbol{f}$ 可视为一个用于估计系数矩阵 $\boldsymbol{\Xi}$的回归问题。在一篇开创性论文中，Brunton 等人（2016）指出，由于许多物理系统的主导项只有少数几项，

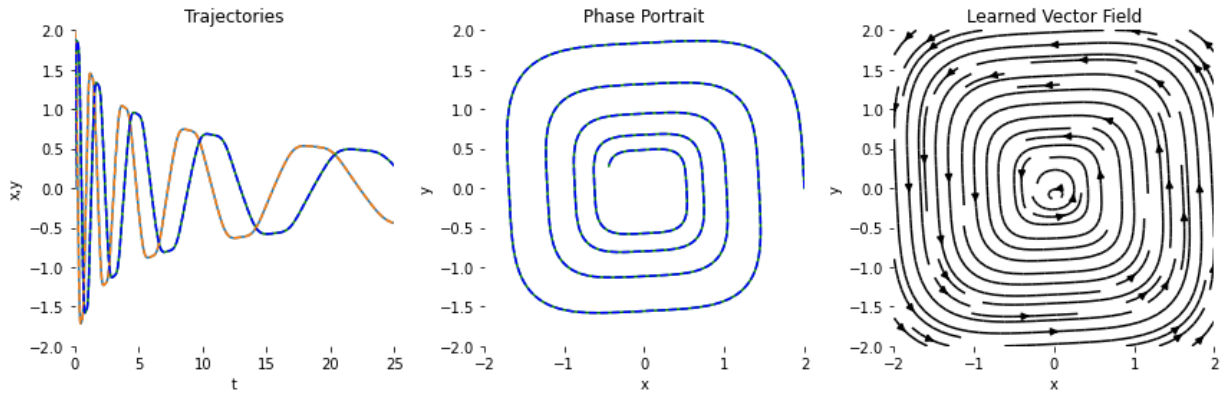


图 4：二维三次系统真实与预测动力系统轨迹的可视化。 [左] 使用 HierGP 得到的真实系统（实线）和预测系统（虚线）的 x-轨迹（橙色）与 y-轨迹（蓝色）。 [中] HierGP 得到的真实系统（实线）和预测系统（虚线）的二维轨迹。 [右] HierGP 学得的向量场。

支配着底层动力学，因此 $\boldsymbol{\Xi}$ 中的系数应以稀疏方式进行估计。为此，他们提出了一种称为非线性动力学稀疏辨识（SINDy）的方法，该方法利用压缩感知算法对 $\boldsymbol{\Xi}$进行稀疏估计，从而实现对系统 $\boldsymbol{f}$的稀疏识别。此后，SINDy 通过稀疏回归和深度学习得到了进一步发展；参见 Champion 等人（2020年）和Both 等人（2021年）。

上述基于 SINDy 的方法存在一个潜在局限：如第 1 节所述，物理系统中的稀疏性通常遵循效应层级与遗传性原则进行结构化（Hamada 和 Wu，1992）：主效应的影响通常大于交互作用，且只有在相应的组成部分主效应存在时，交互作用才会出现。一种刻画这种结构的方法是，在系数矩阵 $\boldsymbol{\Xi}$ 的每一行上分别赋予所提出的层级累积先验分布 (11) 和 (12)。借助这些先验分布，所得的线性系统 (35) 可视为拟合 m 个独立的 HierGP 模型，其基函数取自函数库 $\mathcal{F}$。因此，可以使用第 3.2 节中的吉布斯采样器，通过对给定数据 $\{(\boldsymbol{x}(t_i)\ \dot{\boldsymbol{x}}(t_i)\}_{i=1}^n$ 的系数矩阵 $\boldsymbol{\Xi}$ 进行后验抽样，来恢复控制方程。得到后验样本$\{\boldsymbol{\Xi}_1,\cdots,\boldsymbol{\Xi}_B\}$ 后，可以使用每个系数矩阵样本 $\boldsymbol{\Xi}_b$、 $b=1$、 $\cdots$、 B 对 (34) 进行前向求解，从而预测动力系统并量化其不确定性。

当控制方程中存在结构化稀疏性时，很自然会预期，将这种结构融入 HierGP 能够提升动力系统恢复的效果并增强其确定性，尤其是在数据有限的情况下。我们将在下文通过与现有方法在两个动力系统上的数值比较来探讨这一点。

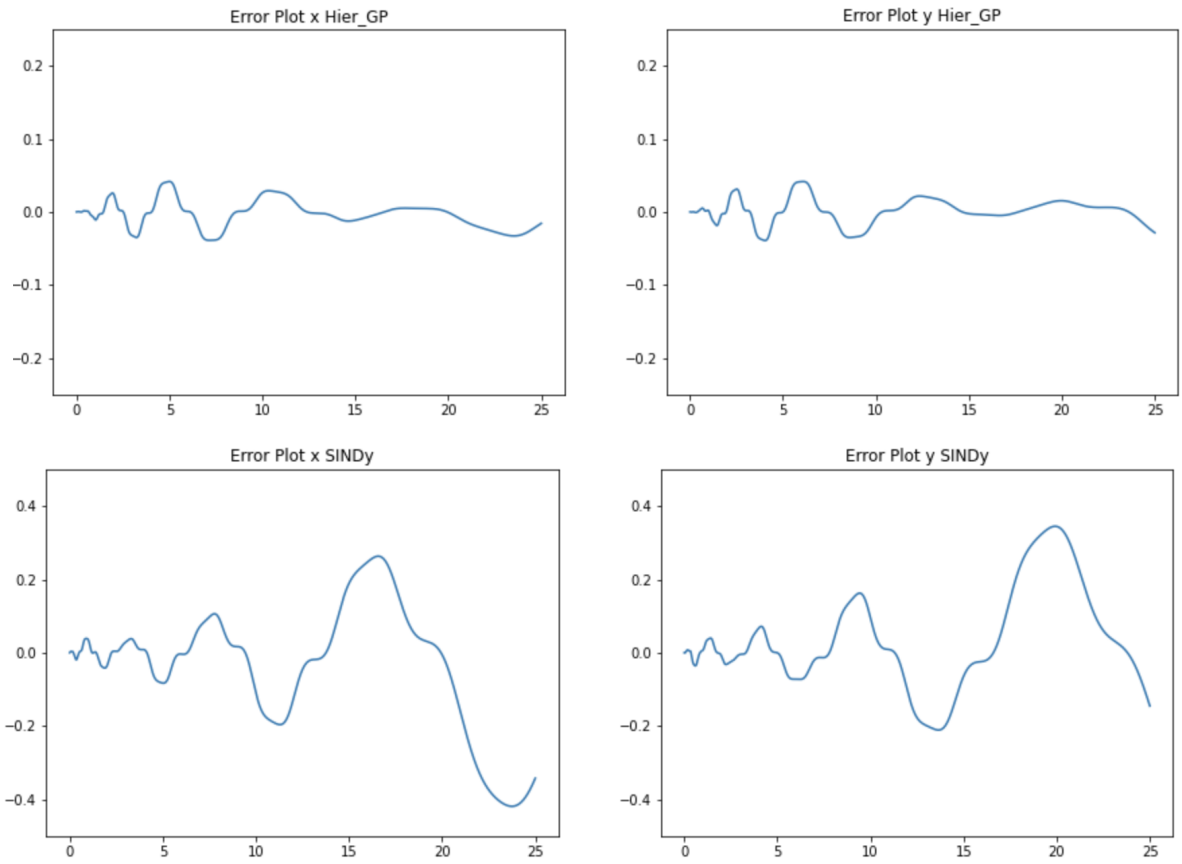


图5：二维三次系统运行500个时间步时，HierGP（上）和 SINDy（下）在 x 和 y 坐标上的预测误差。

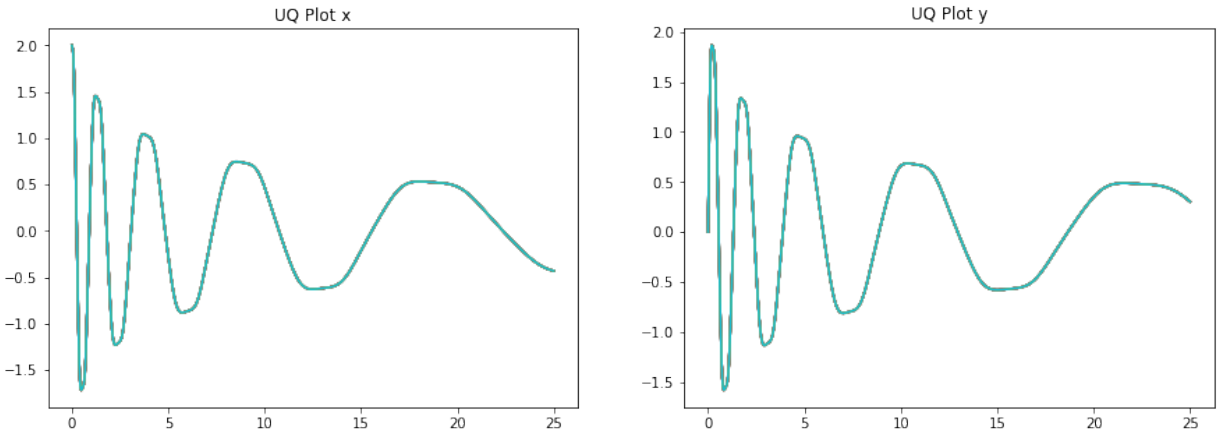


图6：以 x-y-z 坐标（从左到右）可视化二维洛伦兹系统中 HierGP 的50次后验样本抽取的正向运行。

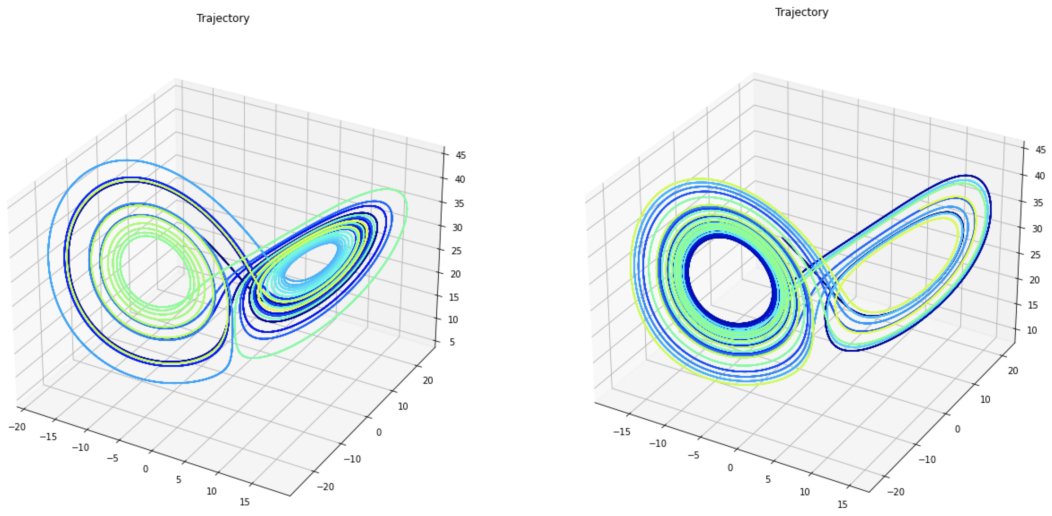


图 7：使用 HierGP 展示真实的（右）和恢复的（左）三维 Lorenz 系统。

5.2.1 二维三次方程

考虑以下二维平面动力系统（更多详情请参阅Brunton 等人(2016)）：

$$\begin{aligned} \frac{dx(t)}{dt} &= -ax(t)^3 + by(t)^3, \\ \frac{dy(t)}{dt} &= -bx(t)^3 + ax(t)^3, \end{aligned} \tag{36}$$

其中 a 和 b 是常数参数。尽管该系统十分简单，但它有两个引人注目的特征，可用于对恢复方法进行有趣的比较。首先，其导数函数体现了效应稀疏性和层次性，因为它们仅依赖于两个阶数相对较低的基函数。其次，可以证明，对于任意初始点（$x(0)$ y(0)，系统总会收敛到一个平衡点（0，0），并呈现准周期行为（Brunton 等人，2016），从而使数值解具有稳定性和可预测性。在以下实验中，我们将真实参数设为 $a = 0.1$ 和 $b = 2$，初始条件为（x_0，y_0）$= (2, 0)$。随后，我们通过对动力系统（36）进行数值求解，生成训练数据，再从一条受独立同分布高斯噪声（方差为 0.01）污染的轨迹中抽取 $n = 500$个观测值（时间步长为 0.04）。最后，我们为 HierGP 设定 $\boldsymbol{K} = (5, 5)$。

图 4展示了真实系统的轨迹，以及通过 SINDy 和所提出的 HierGP 恢复的系统的前向模拟。对于后者，我们首先对模型参数进行了后验抽样 Θ，然后将其后验均值用作参数

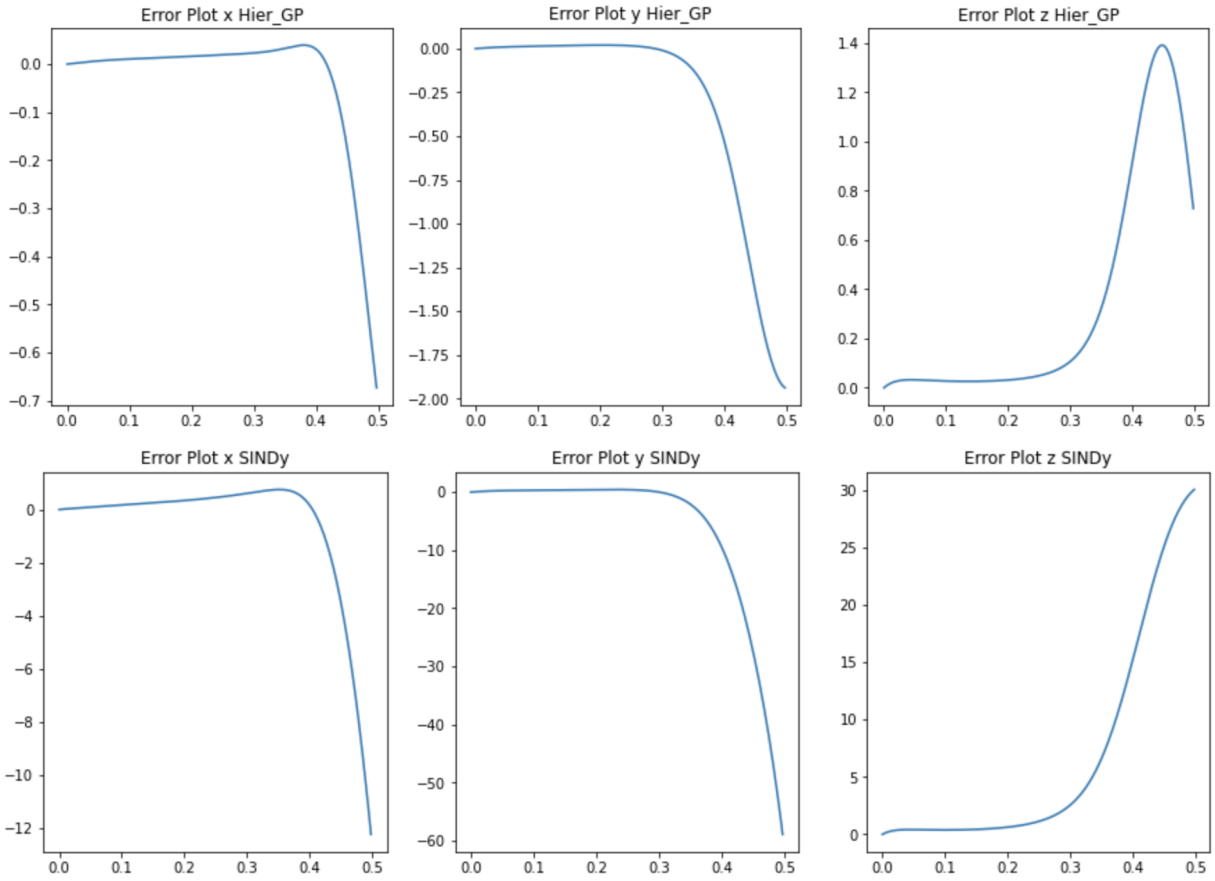


图8：HierGP（上）和 SINDy（下）对三维 Lorenz 系统的预测误差（ x-、 y-和 z-坐标）。

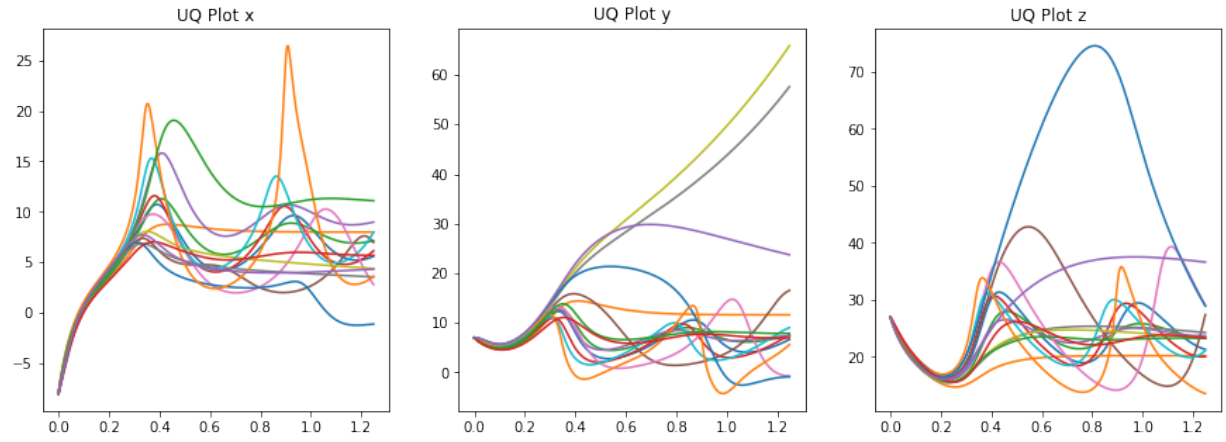


图9：以 x-y-z 坐标（从左到右）可视化 HierGP 对三维 Lorenz 系统的 50 次后验样本抽取的正向运行。

对系统进行一次正向求解（36）。从图中可以看出，HierGP 轨迹很好地捕捉了三次系统所需的周期性和渐近行为。图 5 展示了 HierGP 和 SINDy 恢复的系统在两个坐标上的相应预测误差。可以看到，HierGP 的预测确实显著优于 SINDy；这表明，当控制方程中存在此类结构化稀疏性时，将这种结构纳入先验设定确实有助于改善系统恢复。两种方法的误差都相对较小，这并不意外，因为真实的动力系统相当简单。误差会随时间略微增大；这也符合直觉，因为动力系统系数的估计误差会导致系统恢复误差随时间传播。

图 6 进一步探讨了所提出方法的不确定性量化：展示了 Θ 二维系统（36）在 x-和 y- 坐标上的 50 次后验样本抽取的正向运行。现有的 SINDy 方法（ Brunton 等人，2016）无法提供此类不确定性量化。可以看到，HierGP 恢复的轨迹不仅很好地恢复了真实系统，而且确定性也相对较高。我们还注意到，模型的后验不确定性会随时间逐渐增大；这并不令人意外，因为已知在这类系统中，模型参数的不确定性会随时间累积。

5.2.2 混沌 Lorenz 系统

接下来考虑以下三维混沌 Lorenz 系统（Lorenz，1963年），其表达式为：

$$\begin{aligned} \frac{dx(t)}{dt} &= \sigma(y(t) - x(t)), \\ \frac{dy(t)}{dt} &= x(t)(\rho - z(t)) - y(t), \\ \frac{dz(t)}{dt} &= x(t)y(t) - \beta z(t), \end{aligned} \tag{37}$$

其中，σ、β 和 ρ 为常数参数。尽管这些方程具有丰富且混沌的动力学，并在奇异吸引子上演化（Brunton 等人，2016），但每个导数对应的模型方程都是稀疏且阶数相对较低，因此满足三效应原则。具体而言，导数函数通常只受少数低阶项影响，这些项具有稀疏的层次结构（Brunton 等人，2016）。下文中，我们将真实参数设为 $\sigma = 10$、$\beta = 8/3$ 和 $\rho = 28$，初始条件为（x(0) y(0) z(0)）=（−8，7，27）。与前文一样，通过求解动力学系统（37）生成训练数据，然后从所得解中抽取 $n = 200$ （时间步长为0.05单位时间）个观测值，并为其添加高斯噪声（方差为0.01）；初始条件为 $\boldsymbol{K} =$（5，5，5）。

图 7 展示了真实 Lorenz 系统的轨迹，以及通过 SINDy 和所提出的 HierGP 恢复的系统轨迹。直观来看，两者

恢复的系统都呈现出预期的奇异吸引子行为和短时动力学：所有轨迹起初都在局部范围内移动且可预测，但随着时间推移，从更全局的角度来看则变得混沌，并受限于具有复杂几何结构的区域（Lorenz，1963年）。图 8展示了恢复的系统在三个坐标上的相应预测误差。我们再次看到，HierGP 的表现明显优于 SINDy；这也进一步表明，当存在结构化稀疏性时，整合这种结构确实有助于更准确地恢复系统。值得注意的是，误差会随着时间推移迅速增大，这说明系统恢复会变得越来越困难；鉴于该系统的混沌特性，这并不十分意外。

图9 进一步通过展示针对三维 Lorenz 系统的 50 个后验样本抽取的正向运行，深入考察了 HierGP 的不确定性量化： Θ 对于三维 Lorenz 系统，图中展示了其在 $x/y/z$坐标下的结果。起初，我们看到 HierGP 恢复的轨迹不确定性很小，所有样本路径都彼此非常接近。然而，随着时间推移，后验不确定性显著增大，样本路径逐渐分散，因为不确定性会随时间累积。这并不令人意外，因为该系统具有混沌特性，且误差会随时间传播；这再次表明，对此类系统进行长期预测是一项具有挑战性的任务。

6 结论

本文提出了一种新颖的层次收缩高斯过程（HierGP），在高斯过程模型中融入了效应稀疏性、遗传性和层次性这几项广为人知的原则（Hamada 和 Wu，1992）。与将这些原则用于经典实验分析的做法类似（例如，参见Wu 和 Hamada，2009），在贝叶斯非参数预测模型中融入这种结构化稀疏结构，能够在实验数据有限时提升预测性能。随后，我们推导出用于模型训练和预测的高效后验抽样算法，并证明了 HierGP 的理想一致性结果。数值实验证实，无论是计算机代码仿真还是动力系统恢复，HierGP 的性能均优于现有模型。

尽管取得了令人鼓舞的成果，未来仍有许多值得深入研究的方向。一个方向是为 HierGP 建立后验收缩率。在 f 具有假定的结构化稀疏性时，值得探究 HierGP 的收缩率是否优于高斯过程的标准收缩率；众所周知，后者会受到维度灾难的影响（van der Vaart 和 van Zanten，2008）。另一个方向是探索累积收缩先验，它可以捕捉较弱形式的效应遗传性（例如参见Wu 和 Hamada，2009；Mak 和 Wu，2019），即当至少一个分量效应处于激活状态时，交互作用效应也可以处于激活状态。在效应原则可能存在轻微偏离的情况下，这可以提供更灵活的模型。

参考文献

Alexanderian, A. (2015). 关于 Karhunen-Loeve 展开的一则简短说明。*arXiv* 预印本 *arXiv:1509.07526*。Armagan, A., Dunson, D. B., and Lee, J. (2013). 广义双帕累托收缩。《数理统计学报》, 23(1):119。Berkooz, G., Holmes, P., and Lumley, J. L. (1993). 正交适当分解在湍流分析中的应用。《流体力学年评》, 25(1):539–575。Bhattacharya, A., Pati, D., Pillai, N. S., and Dunson, D. B. (2015). 用于最优收缩的狄利克雷–拉普拉斯先验。*Journal of the American Statistical Association*, 110(512):1479–1490。Both, G.-J., Choudhury, S., Sens, P., and Kusters, R. (2021). DeepMoD：用于含噪数据模型发现的深度学习。《计算物理学杂志》, 428:109985。Brunton, S. L., Proctor, J. L., and Kutz, J. N. (2016). 通过非线性动力系统的稀疏识别，从数据中发现控制方程。《美国国家科学院院刊》, 113(15):3932–3937。Buckingham, E. (1914). 关于物理相似系统：量纲方程的应用示例。《物理评论》, 4(4):345。Carvalho, C. M., Polson, N. G., and Scott, J. G. (2009). 通过马蹄先验处理稀疏性。载于人工智能与统计，第 73–80 页。PMLR。Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010). 稀疏信号的马蹄估计量。*Biometrika*, 97(2):465–480。Castillo, I., Schmidt-Hieber, J., and Van der Vaart, A. (2015). 具有稀疏先验的贝叶斯线性回归。*The Annals of Statistics*, 43(5):1986–2018。Champion, K., Zheng, P., Aravkin, A. Y., Brunton, S. L., and Kutz, J. N. (2020). 一种统一的稀疏优化框架，用于从数据中学习简约型物理信息模型。*IEEE Access*, 8:169259–169271。Chen, J., Mak, S., Joseph, V. R., and Zhang, C. (2021). 函数对函数克里金法及其在主动脉组织三维打印中的应用。《技术计量学》, 63(3):384–395。Cheng, H. and Sandu, A. (2010). 配置点最小二乘多项式混沌方法。载于*2010* 年春季仿真多会议论文集，第 1–6 页。

Choi, T. and Schervish, M. J.（2007）。非参数回归问题中的后验一致性。*Journal of Multivariate Analysis*，98(10):1969–1987。Delahunt, C. B. and Kutz, J. N.（2022）。用于在高噪声情形下数据驱动发现控制方程的工具包。*IEEE Access*，10:31210–31234。Dick, J., Kuo, F. Y., and Sloan, I. H.（2013）。高维积分：准蒙特卡罗方法。*Acta Numerica*，22:133–288。Ding, L., Mak, S., and Wu, C. F. J.（2019）。BdryGP：一种融入边界信息的新型高斯过程模型。*arXiv preprint arXiv:1908.08868*。Dunson, D. B., Wu, H.-T., and Wu, N.（2020）。受限域上的基于图的高斯过程。*arXiv preprint arXiv:2010.07242*。Duvenaud, D. K., Nickisch, H., and Rasmussen, C.（2011）。加性高斯过程。*Advances in Neural Information Processing Systems*，24。Ferrari, F. and Dunson, D. B.（2021）。用于交互作用推断的贝叶斯因子分析。*Journal of the American Statistical Association*，116(535):1521–1532。Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B.（1995）。贝叶斯数据分析。Chapman and Hall/CRC。Ghadami, A. and Epureanu, B. I.（2022）。动力系统中的数据驱动预测：近期进展。*PhilosophicalTransactions of the RoyalSociety A*，380(2229):20210213。Ghanem, R. G. and Spanos, P. D.（1991）。随机有限元方法：响应统计量。载于随机有限元：谱方法，第101–119页。Springer。Golchi, S., Bingham, D. R., Chipman, H., and Campbell, D. A.（2015）。计算机实验的单调仿真。*SIAM/ASA Journal on Uncertainty Quantification*，3(1):370–392。Guckenheimer, J. and Holmes, P.（2013）。非线性振荡、动力系统与向量场分岔，第42卷。Springer Science & Business Media。Hamada, M. and Wu, C. F. J.（1992）。对具有复杂混杂的试验设计进行分析。*Journal of Quality Technology*，24(3):130–137。Ishwaran, H. and Rao, J. S.（2005）。尖峰与平板变量选择：频率学派与贝叶斯策略。*The Annals of Statistics*，33(2):730–773。Jeong, S. and Ghosal, S.（2020）。含干扰参数的稀疏线性回归统一贝叶斯理论。*arXiv preprint arXiv:2008.10230*。

Ji, Y., Mak, S., Soeder, D., Paquet, J., and Bass, S. A.（2021）。用于昂贵计算机代码多保真度仿真的图形高斯过程模型。*arXiv preprintarXiv:2108.00306*。Johndrow, J., Orenstein, P., and Bhattacharya, A.（2020）。针对马蹄先验的可扩展近似马尔可夫链蒙特卡罗算法。机器学习研究杂志，21(73):1–61。Kaufman, C. G., Bingham, D., Habib, S., Heitmann, K., and Frieman, J. A.（2011）。使用紧支撑相关函数的高效计算机实验模拟器及其在宇宙学中的应用。应用统计年鉴，5(4):2470–2492。Legramanti, S., Durante, D., and Dunson, D. B.（2020）。无限因子分解的贝叶斯累积收缩。*Biometrika*，107(3):745–752。Lorenz, E. N.（1963年）。确定性非周期流。大气科学杂志，20(2):130–141。Lu, X., Boukouvalas, A., and Hensman, J.（2022）。重新审视加性高斯过程。载于国际机器学习会议，第14358–14383页。PMLR。Luo, Y., Ogle, K., Tucker, C., Fei, S., Gao, C., LaDeau, S., Clark, J. S., and Schimel, D. S.（2011）。数据丰富时代的生态预测与数据同化。生态学应用，21(5):1429–1442。Luthen, N., Marelli, S., and Sudret, B.（2021）。稀疏多项式混沌展开：文献综述与基准测试。*SIAM/ASA Journal on Uncertainty Quantification*，9(2):593–649。Mak, S., Sung, C.-L., Wang, X., Yeh, S.-T., Chang, Y.-H., Joseph, V. R., Yang, V., and Wu, C. F. J.（2018）。用于大涡模拟仿真与物理提取的高效代理模型。*Journal of the American Statistical Association*，113(524):1443–1456。Mak, S. and Wu, C. F. J.（2019）。cmenet：一种用于条件主效应双层变量选择的新方法。*Journal of the American Statistical Association*，114(526):844–856。McCullagh, P. and Nelder, J. A.（1989）。广义线性模型，第37卷。Chapman & Hall。Mudelsee, M.（2019）。气候时间序列趋势分析：方法综述。*Earth-Science Reviews*，190:310–322。Nelder, J.（1977）。线性模型的重新表述。皇家统计学会杂志：*A*辑（综合），140(1):48–63。

Owen, A. B.（1997）。光滑函数积分的扰乱网方差。*The Annalsof Statistics*, 25(4):1541–1562。Rasmussen, C. E.（2003）。机器学习中的高斯过程。载于机器学习暑期学校，第63–71页。Springer。Santner, T. J., Williams, B. J., Notz, W. I., and Williams, B. J.（2003）。计算机实验的设计与分析，第1卷。Springer。Savitsky, T., Vannucci, M., and Sha, N.（2011）。非参数高斯过程先验的变量选择：模型与计算策略。统计科学，26(1):130。Scheipl, F., Fahrmeir, L., and Kneib, T.（2012）。结构化加性回归模型中用于函数选择的尖峰与厚尾先验。*Journal of the American Statistical Association*，107(500):1518–1532。Sobester, A., Forrester, A., and Keane, A.（2008）。通过代理建模进行工程设计：实用指南。约翰·威立父子出版公司。Song, Q. and Liang, F.（2017）。高维回归的近乎最优贝叶斯收缩。*arXiv:1712.08964*。Stein, M. L.（1999）。空间数据插值：克里金法的一些理论。Springer Science & Business Media。van der Vaart, A. W. and van Zanten, J. H.（2008）。基于高斯过程先验的后验分布收缩速率。*The Annals of Statistics*, 36(3):1435–1463。Wang, L.（2008）。*Karhunen-Loeve* 展开及其应用。伦敦政治经济学院。Wang, X. and Berger, J. O.（2016）。使用高斯过程估计形状约束函数。*SIAM/ASA Journal on Uncertainty Quantification*，4(1):1–25。Wheeler, M. W., Dunson, D. B., Pandalai, S. P., Baker, B. A., and Herring, A. H.（2014）。机制层次高斯过程。*Journal of the American Statistical Association*，109(507):894–904。Wu, C. F. J. and Hamada, M. S.（2009）。实验：规划、分析与优化。约翰·威立父子出版公司。Xiu, D.（2010）。随机计算的数值方法：谱方法途径。普林斯顿大学出版社。

A 技术证明

A.1 定理 2 的证明2

证明。为简化记号，我们先针对单变量HierGP证明这一结果，然后将这一论证推广到多变量 HierGP。下面使用以下记号： θ 表示（f， σ），其中 $\theta_0 =$（f_0， σ_0）。密度 f_i（x， σ）是均值为 f（x_i）、方差为 σ^2的正态密度。参数空间 Θ 是函数空间Θ_1 与 $\mathbb{R}^+$的乘积空间。 θ 上的先验是乘积测度 $\Pi = \Pi_1 \times \Pi_2$。

该定理与 ChoiandSchervish(2007) 中针对我们的先验所给出的定理 4 类似，只需针对先验 Π 检查同一论文中定理 1 的条件即可。根据该论文第4.2节给出的等价条件，只需检查:

(1) 先验正性: $\Pi(B_\delta) > 0$, $B_\delta = \{(f,\sigma) : |f - f_0|_\infty < \delta, |\frac{\sigma}{\sigma_0} - 1| < \epsilon\}$ 对于任意 $\delta > 0$(2) 概率为 $\Theta_n^c \quad Pr\ \Theta_n^c \leq C_2 e^{-c_2 n} \quad \Theta_n^c = \Theta_{n\,0}^C \bigcup \Theta_{n\,1}^C \quad \Theta_{n\,0}^C = \{f \quad |f|_\infty >:(),,,,,:$ $M_n\}\ \Theta_{n\,1}^C = \{f \quad |f'|_\infty > M_n\} \qquad M_n = O\ n^\alpha \qquad 1/2 < \alpha < 1$,,: 其中 ( ) 对于 . (3) 检验的存在性 : 参见 Choi andSchervish(2007) 中的公式 3 和定理 2。

我们首先证明（1）成立。由于 $\Pi_2(|\frac{\sigma}{\sigma_0} - 1| < \epsilon)$ 始终为正，只需证明 $\Pi_1(|f - f_0|_\infty < \delta) > 0$。注意，$\{|f - f_0|_\infty < \delta\} \supset \{|\sum_{k\in S}(\lambda_k\phi_k(x) - \lambda_k^0\phi_k(x)|_\infty < \delta/2\} \bigcap \{|\sum_{k\notin S}\lambda_k\phi_k(x)|_\infty < \delta/2\} = A \bigcap B$。这里我们知道，给定 w 后， A 和 B 相互独立，因为 λ_k 条件独立。因此， $P(A \bigcap B|w) = P(A|w)P(B|w)$。

对于集合 $C_\epsilon = \{w_k > \epsilon, k \in S\}$，其中 ϵ 足够小，我们断言 $\Pi_1(C_\epsilon) > 0$ 且 $P(A|w) > t(\epsilon, \lambda_0) > 0$，其中某个函数 t 对任意 $w \in C_\epsilon$ 均成立。第一个断言显然成立，因为 $w_k > 0$ 几乎必然成立且 S 有限；第二个断言成立，因为对于 $k \in S\}$， $D = \{\lambda_k \sim N(0, \sigma_k^0)$ 的概率大于 $\epsilon^{|n|}$（在 C 上，对所有协变量都选择厚尾部分的概率），且 $P(A|w, \{\sigma_k\}_k, k \in S) = P(A|\{\sigma_k\}_k = \{\sigma_{k,0}\}_k, k \in S) = P(|\sum_{k\in S}(\lambda_k\phi_k(x) - \lambda_k^0\phi_k(x)|_\infty < \delta/2|\{\sigma_k\}_k = \{\sigma_{k,0}\}_k, k \in S) \geq \Pi_{k\in S}P(a_k|\lambda_k - \lambda_k^0| < \delta/(2|S|)|\{\sigma_k\}_k = \{\sigma_{k,0}\}_k, k \in S) = s > 0$。随后对 σ_k 积分消去，即可证明第二个断言： $P(A|w) > s * \epsilon^{|S|} = t(\epsilon, \lambda_0) > 0$。

现在我们声称 $P(B|w) > 0$ 几乎必然成立。若此结论成立，则 $\Pi_1(A \bigcap B) = \mathbb{E}_{\Pi_1}(P(A \bigcap B|w)) = \mathbb{E}_{\Pi_1}(P(A|w)P(B|w)) \geq \mathbb{E}_{\Pi_1}(P(B|w) \cdot P(A|w)1_{C_\epsilon}) \geq \mathbb{E}_{\Pi_1}(P(B|w) \cdot t(\epsilon, \lambda^0)1_{C_\epsilon}) > 0$。最后一步是由 $P(B|w) > 0$ 几乎必然成立以及 $\Pi_1(C_\epsilon) > 0$ 得出的。

由此可见， $P(B|w) = \mathbb{E}(P(B|w, \{\sigma_k\}_{k=1}^\infty)|w)$。此外，根据 $B = \{|\sum_{k\notin S}\lambda_k\phi_k(x)|_\infty < \delta/2\}$ 的定义以及 $Y = \sum_{k\notin S}\lambda_k\phi_k(x)|w, \{\sigma_k\}_{k=1}^\infty$ 满足均值为 $\sum_{k\notin S}a_k\sigma_k$、方差为 0 的高斯过程这一事实，我们几乎必然有 $P(B|w, \{\sigma_k\}_{k=1}^\infty)$。（对于该高斯过程，我们总有 $P(|Y|_\infty < \delta) > 0$，参见 Choi 和 Schervish （ 2007 ）中概述的论证。）因此， $P(B|w) > 0$ 几乎必然成立。由此证明了该断言，且 (1) 成立。

接下来，我们证明 *(3)* 成立。这可由Choi 和 Schervish（2007）的定理 2 以及我们关于设计点抽样的假设推出。

最后，我们证明 *(2)* 成立。下面我们遵循与Choi andSchervish(2007)中例 6.1 类似的论证。

由于 $|\sum_{k=1}^{\infty}\lambda_k\phi_k(\boldsymbol{x})| < \sum_{k=1}^{\infty} a_k|\lambda_k|$，根据马尔可夫不等式、Chernoff 界以及 $|\lambda_k| \leq |N(0,\sigma_{k,0}^2)|$, $i > 0$ 在分布意义下这一事实，可得:

$$\begin{aligned}
\Pi_1\left(\sup|\sum_{k=1}^{\infty}\lambda_k\phi_k(\boldsymbol{x})| > M_n\right) &\leq \Pi_1\left(\sum_{k=1}^{\infty} a_k|\lambda_k| > M_n\right) \\
&\leq \exp(-tM_n)\mathbb{E}(\exp(\sum_{k=1}^{\infty} a_k|\lambda_k|)) \\
&= \exp(-tM_n)\mathbb{E}(\mathbb{E}(\exp(\sum_{k=1}^{\infty} a_k|\lambda_k|)|w)) \\
&\leq \exp(-tM_n)\mathbb{E}(\mathbb{E}(\exp(\frac{1}{2}t^2\sum_{k=1}^{\infty} a_k^2\sigma_{k,0}^2)2\Phi(a_j\sigma_{k,0}t)|w)) \\
&= \exp(-tM_n)\mathbb{E}(\exp(\frac{1}{2}t^2\sum_{k=1}^{\infty} a_k^2\sigma_{k,0}^2)2\Phi(a_j\sigma_{k,0}t)).
\end{aligned}$$

现在，左侧与Choiand Schervish(2007) 中的例 6.1 完全相同。因此，我们可以沿用其中的论证来证明

$$\Pi_1\left(\sup|\sum_{k=1}^{\infty}\lambda_k\phi_k(\boldsymbol{x})| > M_n\right) \leq \exp\left(-\frac{n}{4\sum_{k=1}^{\infty} a_k^2\sigma_{k,0}^2}\right).$$

同理，我们可以证明 $\Pi_1(\sup|\sum_{k=1}^{\infty}\lambda_k\phi_k'(\boldsymbol{x})| > M_n) \leq \exp(-\frac{n}{4\sum_{k=1}^{\infty} b_k^2\sigma_{k,0}^2})$，从而证明 (2)。

随后可将上述类似论证扩展至多元 HierGP。只需检查相同的三个条件 (1)-(3)，将正态随机变量替换为多元正态随机变量即可。可以采用类似的证明来验证这些条件。

□

A.2 定理的证明 3

证明。 该定理正是 Choi和 Schervish （2007）针对我们的先验提出的定理 6，因此只需检验该论文中定理 1 对先验 Π所给出的条件。根据该论文第4.2节给出的等价条件，只需检验:

(1) 先验正性： $\Pi(B_\delta) > 0$， $B_\delta = \{(f,\ \sigma)\colon\ |f - f_0|_\infty < \delta,\ |\frac{\sigma}{\sigma_0} = 1| < \epsilon\}$ 对于任意 $\delta > 0$

(2) 的概率 $\Theta_n^c\ \ Pr\ \Theta_n^c \le C_2 e^{-c_2 n}\ \ \Theta_n^c = \Theta_{n\,0}^C \bigcup \Theta_{n\,1}^C\ \ \Theta_{n\,0}^C = \{f\ \ \ |f|_\infty >: M_n\}\ \Theta_{n\,1}^C = \{f\ \ |f'|_\infty > M_n\}\ \ \ \ M_n = O\ n^\alpha\ \ \ \ 1/2 < \alpha < 1$ ( ) ,,,,: ,: 其中 ( ) 对于 。

(3) 检验的存在性：参见Choi andSchervish(2007) 的公式 3 和定理 2。

这些条件的验证与上一个定理相同。 □

A.3 定理 4 的证明

证明。 根据 Song 和 Liang (2017) 的定理 2.1、2.4、3.1，我们得到 $\{\lambda_k\}_{k=1}^{H_n}$ 的一致性和形状近似。由于 $\{\phi_k\}_{k=1}^{H_n}$ 是 L^2 中的标准正交基，根据 $l_2 - L_2$ 的等距映射，以及系数—函数映射下全变差的收缩性，结论得证。

□

B HierGP 的 Gibbs 采样器推导

这里给出的推导与Legramanti 等人（2020年中的推导类似，但为求完整起见，我们仍在下文给出推导过程。

对于单变量 HierGP，定义独立指示变量 z_k，其概率为 $p(z_k = l|\{\nu_k, w_k\}_{k=1}^K) = \nu_l w_{l-1}(k,\ l = 1,\ \cdots,\ K)$，以及由下式给出的 σ_k 的条件概率：

$$(\sigma_k|z_k) \sim (1 - 1_{z_k \le k}) IG(a_\sigma, b_\sigma) + 1_{z_k \le k}\delta_{\theta_\infty}. \tag{38}$$

注意，对 z_k 积分消去后即可得到原始先验分布。要获得全条件概率 $p(z_k|-)$，首先根据 $p(z_k = l|\{\nu_k, w_k\}_{k=1}^K)$ 的分布、(B.1) 和 $\lambda_k|\sigma_k$，求得联合条件分布 $p(z_k,\ \sigma_k|-)$。然后对 z_k 积分消去，即可得到 $p(z_k|-)$。具体而言，当 $z_k > k$ 时， $\sigma_k \sim IG(a_\sigma,\ b_\sigma)$；对 σ_k 积分消去后可得 $p(z_k = l|-) \propto \nu_l w_{l-1} \cdot \int p(\lambda_k|\sigma_k)p(\sigma_k|-)d\sigma_k \propto \nu_l w_{l-1} \cdot t_{2a_\sigma}(\lambda_k; 0,\ (b_\sigma/a_\sigma))$。这里用到的事实是，正态密度对逆伽马分布积分后会得到t-分布的密度。类似地，当 $z_k \le k$ 时，有 $p(z_k = l|-) \propto \nu_l w_{l-1} \cdot N(\lambda_k; 0,\ \sigma_\infty)$，因为此时 $\sigma_k = \theta_\infty$ 是一个常数。

对于多变量情形，我们定义多重指标独立指示变量 $p(z_{k_1 \cdots k_d} = (l_1 \cdots l_d)|\{\nu_{l_1}^1, \cdots, \nu_{l_d}^d\}\{w_{l_1}^1, \cdots, w_{l_d}^d\}) = \Pi_{m=1}^d \nu_{l_m}^m w_{l_m - 1}^m$，并定义条件分布

$$(\sigma_{\boldsymbol{k}}|z_{\boldsymbol{k}}) \sim (1 - 1_{z_{\boldsymbol{k}} \le \boldsymbol{k}}) IG(a_\sigma, b_\sigma) + 1_{z_{\boldsymbol{k}} \le \boldsymbol{k}}\delta_{\theta_\infty}. \tag{39}$$

再次注意，对 $z_{\boldsymbol{k}}$ 进行边缘化即可得到原始先验分布。如上所述，我们首先得到联合分布 $p(z_{\boldsymbol{k}},\ \sigma_{\boldsymbol{k}}|-)$，然后根据 (B.2) 对 $z_{\boldsymbol{k}}$ 进行边缘化，最终得到

$$p(z_{\boldsymbol{k}} = \boldsymbol{l}|-) \propto \begin{cases} (\Pi_{m=1}^{d} \nu_{l_m}^m w_{l_m-1}^m \cdot N(\lambda_{\boldsymbol{k}}; 0, \sigma_\infty), & \text{otherwise,} \\ (\Pi_{m=1}^{d} \nu_{l_m}^m w_{l_m-1}^m \cdot t_{2a_\sigma}(\lambda_{\boldsymbol{k}}; 0, (b_\sigma/a_\sigma)), & \text{for } l_1 > k_1, \cdots, l_d > k_d. \end{cases}$$

C HierGP 的自适应截断上限

这里，我们提供一种自适应实现的吉布斯采样器，用于拟合所提出的 HierGP 模型。这里的自适应性体现在：随着我们从 $\boldsymbol{K}$ 观测到更多数据，截断上限 f 会随之增大。我们采用 Legramanti 等人 (2020年) 提出的方法，以迭代方式（随时间推移，发生概率逐渐降低）移除未激活基函数的索引，并添加额外基函数的索引。这种概率方法使 HierGP 能够随着我们收集到更多响应面数据 f，自适应地更新截断上限；这与 Legramanti 等人 (2020年) 针对因子模型提出的方法类似。下面我们先针对单变量 HierGP，再针对一般的多变量 HierGP，介绍这种自适应吉布斯采样器。

算法 4 单变量 HierGP 的自适应吉布斯采样器的一个周期 输入：自适应开始前的迭代次数 B、初始截断 K、常数 α_0 和 α_1。对于 $b = 1, \cdots, B$ 执行 2：在算法 3 中执行一次吉布斯采样器迭代。结束循环 4：如果 $b \geq B$，则以概率 $p(b) = \exp(\alpha_0 + \alpha_1 b)$ ... 6：如果 $K^* = \sum_k 1_{z_k^{[b]} > k} \leq K$，则设置 $K \leftarrow K^* + 1$。8：从 Λ 中删除非活动列，同时删除$\Lambda, \sigma, \omega, \nu$中的相关参数。向 Λ, σ, w, ν 添加一个从相应先验分布中抽样得到的最终分量。10：否则设置 $K \leftarrow K + 1$。12：向 Λ添加一个从尖峰中抽样得到的最终列，同时向 σ, w, ν添加从相应先验分布中抽样得到的相关参数。结束条件 14：结束条件

算法 5 多变量 HierGP 的自适应吉布斯采样器的一个周期¯输入：自适应开始前的迭代次数 B**、初始截断** $\boldsymbol{K}$**、常数** α_0 **和** α_1。**对于** $b = 1, \cdots, B$ **执行 2：在算法 3.2 中执行一个吉布斯采样器周期。结束循环¯4：如果** $b \geq B$**，则以概率** $p(b) = \exp(\alpha_0 + \alpha_1 b)$ **... 6：对于** m**，在 1 :** d **中执行：如果** $K_m^* = \sum_{\boldsymbol{k}} 1_{z_{\boldsymbol{k}}^{[b]}, m > k_m} \leq K_m$**，则 8：设置** $K_m = K_m^* + 1$。**从** Λ **中删除非活动列，同时删除**σ, w, ν**中的相关参数。10：向** Λ, σ, w, ν**添加一个从相应先验分布中抽样得到的最终分量。否则 12：设置** $K_m = K_m + 1$。**向** Λ **添加一个从尖峰中抽样得到的最终列，同时向** σ, w, ν**添加从相应先验分布中抽样得到的相关参数 14：结束条件 结束循环 16：结束条件**