

# 复杂数据背景下的高斯过程回归模型

邓旭生, 卢志义

天津商业大学理学院, 天津

收稿日期: 2022年9月19日; 录用日期: 2022年10月9日; 发布日期: 2022年10月25日

## 摘要

高斯过程回归模型是一种基于贝叶斯推断的非参数模型, 它以概率论为基础, 通过在模型中明确地引入随机性, 将研究者的先验知识与从观察数据中学习到的知识进行有机融合, 并通过贝叶斯推断来减小不确定性。高斯过程回归模型所具有的天然的可解释性、灵活性以及稳健性, 决定了其在统计学习领域发挥了重要且不可替代的作用, 被广泛应用于各个领域。近年来, 随着大数据时代的到来, 现实数据不断趋于复杂化、非结构化以及实时化, 催生了该模型在可扩展性以及模型结构更新等方面的快速发展。本文对近十年来高斯过程回归模型在大数据领域的拓展算法以及模型的改进方法进行分析总结, 概述了各个方法的优缺点, 并对高斯过程回归模型的未来研究方向进行展望。

## 关键词

高斯过程回归, 大数据, 统计学习, 稀疏近似, 核函数

# Gaussian Process Regression Model in the Context of Complex Data

Xusheng Deng, Zhiyi Lu

School of Science, Tianjin University of Commerce, Tianjin

Received: Sep. 19<sup>th</sup>, 2022; accepted: Oct. 9<sup>th</sup>, 2022; published: Oct. 25<sup>th</sup>, 2022

## Abstract

Gaussian process regression model is a nonparametric model based on Bayesian inference. It is based on probability theory. By explicitly introducing randomness into the model, it organically integrates researchers' prior knowledge and knowledge learned from observation data, and reduces uncertainty through Bayesian inference. The natural interpretability, flexibility and robustness of Gaussian process regression model determine that it plays an important and irreplaceable role in the field of statistical learning and is widely used in various fields. In recent years,

with the advent of the big data era, the real data has become increasingly complex, unstructured and real-time, which has led to the rapid development of the model in terms of scalability and model structure update. This paper analyzes and summarizes the expanding algorithms and improved methods of the Gaussian process regression model in the field of big data in the past decade, summarizes the advantages and disadvantages of each method, and looks forward to the future research direction of the Gaussian process regression model.

## Keywords

Gaussian Process Regression, Big Data, Statistical Learning, Sparse Approximations, Kernel Function

Copyright © 2022 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

## 1. 引言

高斯过程回归(GPR)是一种贝叶斯框架下的非参数模型,常用于非线性建模。该模型以概率论为基础,通过在模型中明确地引入随机性,将研究者的先验知识与从观察数据中学习到的知识进行有机融合,并通过贝叶斯推断来减小不确定性,从而得出具有概率意义的估计。高斯过程回归模型具有泛化能力强、模型训练简单、超参数自适应、可解释性以及稳健性等优点,尤其适于解决非线性、高维度、小样本数据的回归问题。由于以上优点,高斯过程回归已成为统计学习的热门研究和应用领域,以其为基础的模型层出不穷,并广泛应用于锂电池数据处理、装备剩余生命预测、损失准备金估计、风速预测、流场流动控制、新冠疫情预测等诸多领域[1]-[6]。

本文在对高斯过程回归模型的基本原理进行介绍的基础上,对相应的改进模型进行分析对比,并对目前存在的主要问题探讨,并对高斯过程回归模型的未来研究方向进行展望。

## 2. 高斯过程回归的基本原理

作为一种贝叶斯非参数模型,高斯过程回归以高斯过程为先验,直接在函数空间进行贝叶斯推断。高斯过程是多维高斯分布的推广,它完全由均值函数和协方差函数确定:

$$f(x) \sim \text{GP}(m(x), K(x, x')) \quad (1)$$

其中  $m(x)$  为均值函数,  $K(x, x')$  为协方差函数(核函数)。

考虑回归问题  $y = f(x) + \varepsilon$ , 假设噪声  $\varepsilon \sim N(0, \sigma_\varepsilon^2)$ 。可得到观测值  $y$  与在测试点  $\mathbf{x}_* = (x_{n+1}, \dots, x_{n+m})^T$  处预测值  $\mathbf{f}_*$  的联合分布

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{f}_* \end{bmatrix} \sim \mathcal{N} \left( \begin{bmatrix} m(\mathbf{x}) \\ m(\mathbf{x}_*) \end{bmatrix}, \begin{bmatrix} K(\mathbf{x}, \mathbf{x}) + \sigma_n^2 \mathbf{I} & K(\mathbf{x}, \mathbf{x}_*) \\ K(\mathbf{x}_*, \mathbf{x}) & K(\mathbf{x}_*, \mathbf{x}_*) \end{bmatrix} \right) \quad (2)$$

其中均值函数通常设为 0, 即  $m(\mathbf{x}) = m(\mathbf{x}_*) = 0$ ,  $K(\mathbf{x}, \mathbf{x})$  为  $n \times n$  阶对称半正定协方差矩阵;  $K(\mathbf{x}_*, \mathbf{x}) = K(\mathbf{x}, \mathbf{x}_*)^T$  为  $m \times n$  阶矩阵, 有  $[K(\mathbf{x}_*, \mathbf{x})]_{ij} = K(x_{n+i}, x_j)$ ;  $K(\mathbf{x}_*, \mathbf{x}_*)$  为  $m \times m$  阶矩阵, 且  $[K(\mathbf{x}_*, \mathbf{x}_*)]_{ij} = K(x_{n+i}, x_{n+j})$ 。由此可以得到  $\mathbf{f}_*$  的后验分布

$$\mathbf{f}_* | \mathbf{x}, \mathbf{x}_*, \mathbf{y} \sim \mathcal{N}(\bar{\mathbf{f}}_*, \text{cov}(\mathbf{f}_*)) \quad (3)$$

其中

$$\begin{aligned}\bar{f}_* &= \mathbb{E}[f_* | \mathbf{x}, \mathbf{x}_*, \mathbf{y}] \\ &= m(\mathbf{x}_*) + K(\mathbf{x}_*, \mathbf{x}) [K(\mathbf{x}, \mathbf{x}) + \sigma_\varepsilon^2 \mathbf{I}]^{-1} (\mathbf{y} - m(\mathbf{x}))\end{aligned}\quad (4)$$

$$\text{cov}(f_*) = K(\mathbf{x}_*, \mathbf{x}_*) - K(\mathbf{x}_*, \mathbf{x}) [K(\mathbf{x}, \mathbf{x}) + \sigma_\varepsilon^2 \mathbf{I}]^{-1} K(\mathbf{x}, \mathbf{x}_*) \quad (5)$$

如此便可得到  $f_*$  的均值和方差。

高斯过程回归可以通过选择不同的核函数来捕捉不同的统计特征，其中最常用的是平方指数核(SE):

$$K_{SE} = \eta^2 \exp\left(-\frac{|x_i - x_j|^2}{2l^2}\right) \quad (6)$$

其中  $\eta^2$  为信号方差,  $l$  为长度尺度,  $\theta = \{\eta^2, l, \sigma_n^2\}$  即为模型的超参数。一般情况下, 通过各种优化算法最大化对数边际似然函数来得到最优超参数, 其对数边际似然函数如下:

$$\log p(\mathbf{y}) = -\frac{1}{2} \log(2\pi) - \frac{1}{2} \log |K(\mathbf{x}, \mathbf{x}) + \sigma_\varepsilon^2 \mathbf{I}| - \frac{1}{2} \mathbf{y}^T (K(\mathbf{x}, \mathbf{x}) + \sigma_\varepsilon^2 \mathbf{I})^{-1} \mathbf{y} \quad (7)$$

在训练完成后即可利用式(4) (5)对任何测试点进行预测。

### 3. 复杂数据背景下高斯过程回归模型的改进

高斯过程回归虽然有许多优点, 但目前仍存在一些问题, 一是其时间复杂度达到样本数的三次方, 因此并不适用于大样本数据; 二是高斯过程回归模型的效果很大程度上取决于核函数的选择, 但当前对于如何针对具体问题构造组合核函数还没有统一的理论支撑。

何志昆等人曾对降低高斯过程回归模型时间复杂度的改进算法进行过详细的介绍[7], 但该领域研究十分活跃, 近十年间已提出更多有效方法, 故本文对此方面再次进行了总结, 并在第 4 节中详细描述。同时, 自 Duvenaud 等人提出组合核搜索法(CKS)后, 组合核函数的自动构造方法不断涌现, 本文将这些方法分为了两类, 在第 5 节中分别进行了介绍。

### 4. 适应大样本数据的改进方法

高斯过程回归作为一种贝叶斯非参数模型, 需要“记住”完整的数据集, 以便进行训练和预测。由于在训练过程中涉及到大量的矩阵求逆运算, 当对大小为  $n$  的数据集应用高斯过程回归时, 精确推断的计算复杂度为  $O(n^3)$ , 预测的计算复杂度为  $O(n^2)$  [8]。因此, 计算量会随着数据集的增大而增加。这限制了 GPR 模型在许多领域, 特别是在大型时空数据集、视频、大型社交网络数据、人口规模的医疗数据集等方面的应用。过去的 10 年来, 为了解决上述问题, 学者提出了许多有效的近似方法。这些方法大体上可以分为全局近似和局部近似两类。

#### 4.1. 全局近似

通过使用训练数据的子集; 利用稀疏核函数去除协方差矩阵  $K_{mm}$  中低相关的元素; 对核矩阵进行稀疏近似, 均可以实现核矩阵的稀疏性, 从而达到降低计算复杂度的目的。接下来, 我们将对上述方法进行详细描述。

##### 4.1.1. 数据子集近似法

数据子集(SD)近似法是通过使用完整  $n$  维数据集  $D$  的一个维数为  $m$  的子集  $SD$  作为训练集, 用于 GPR

的训练与预测。因此, SD 近似法在时间复杂度较低的  $O(m^3)$  的情况下进行标准 GP 推断, 因为它对由  $m$  个 ( $m \ll n$ ) 数据点产生的核矩阵  $K_{mm}$  进行求逆操作, 而  $K_{mm}$  的大小仅为  $m \times m$ 。应用 SD 近似法的关键是如何选取合适的子集, 目前常用的方法有: 从原数据集  $D$  中随机选择  $m$  个数据点; 使用聚类技术, 如  $k$  均值聚类和 KD 树, 将数据划分为  $m$  个子集, 并选择它们的质心作为子集; 使用主动学习准则, 如微分熵, 信息增益和匹配追求顺序查询数据点, 但计算成本较高[9] [10] [11] [12]。

#### 4.1.2. 稀疏核函数法

特别设计的稀疏核函数在  $|x_i - x_j|$  超过一定阈值时, 可令  $k(x_i, x_j) = 0$ , 相应的矩阵元素也为零[13]。如此便可以得到完整协方差矩阵  $K_{mm}$  的一个稀疏表示  $\tilde{K}_{mm}$ 。由于只有  $\tilde{K}_{mm}$  中的非零元素涉及到求逆计算。因此, 使用稀疏核的 GPR 的训练复杂度为  $O(\alpha n^3)$ ,  $0 < \alpha < 1$ 。构造有效稀疏核的主要挑战是需要确保  $\tilde{K}_{mm}$  半正定(PSD)。

#### 4.1.3. 降秩近似法

通过对协方差矩阵  $K_{mm}$  进行特征值分解, 然后保留  $m$  个特征值来近似协方差矩阵, 即  $K_{mm} = VV^T$ , 其中  $V$  为  $n \times m$  维矩阵。然后, 根据矩阵求逆引理可得

$$(K_{mm} + \sigma_n^2 I_n)^{-1} = \sigma_n^{-2} I_n - \sigma_n^{-2} V (\sigma_n^2 I_m + V^T V)^{-1} V^T \quad (8)$$

可以看出, 对协方差矩阵  $K_{mm}$  的求逆已经转变成对  $m \times m$  维矩阵求逆, 训练计算量由  $O(n^3)$  降至  $O(n^2 m)$ , 预测计算量则由  $O(n^2)$  降至  $O(m^2)$ 。

但对  $K_{mm}$  进行特征值分解操作的计算量为  $O(n^3)$ , 实际上并未降低计算量。在这种情况下, 可以用从原数据集中抽取的维数为  $m$  的子集近似  $K_{mm}$  的特征值, 该方法被称为 Nyström 近似。但该方法可能导致负的预测方差, 因此在此基础上人们提出了降秩近似。

降秩近似可大致分为两大类: 先验近似与后验近似。区别在于前者近似先验, 但执行精确后验推断, 后者保留精确的先验, 但对后验执行近似推断

##### (a) 先验近似

通常由“诱导点”法来达到稀疏目的。引入一组诱导点  $(\mathbf{X}_m, \mathbf{f}_m)$ , 诱导变量  $\mathbf{f}_m$  与  $f$  遵循相同的 GP 先验  $p(\mathbf{f}_m) = \mathcal{N}(0, K_{mm})$ 。此外, 假设  $\mathbf{f}_m$  是  $f$  的一个充分统计量, 给定独立性假设  $f \perp \mathbf{f}_* | \mathbf{f}_m$ 。联合先验分布可表示为

$$p(\mathbf{f}, \mathbf{f}_*) = \int p(\mathbf{f} | \mathbf{f}_m) p(\mathbf{f}_* | \mathbf{f}_m) p(\mathbf{f}_m) d\mathbf{f}_m \quad (9)$$

训练条件概率和测试条件概率分别为

$$p(\mathbf{f} | \mathbf{f}_m) = \mathcal{N}(\mathbf{f} | K_{mm} K_{mm}^{-1} \mathbf{f}_m, K_{mm} - \mathbf{Q}_{mm}) \quad (10)$$

$$p(\mathbf{f}_* | \mathbf{f}_m) = \mathcal{N}(\mathbf{f}_* | \mathbf{k}_{*m} K_{mm}^{-1} \mathbf{f}_m, K_{**} - \mathbf{Q}_{**}) \quad (11)$$

其中  $\mathbf{Q}_{mm}, \mathbf{Q}_{**}$  为 Nyström 记号, 有  $\mathbf{Q}_{ab} = K_{am} K_{mm}^{-1} K_{mb}$  [14]。

从上式可以看出,  $\mathbf{f}_m$  之所以被称为诱导变量, 是因为  $\mathbf{f}$  和  $\mathbf{f}_*$  之间的依赖关系是由  $\mathbf{f}_m$  诱导出来的。为减小时间复杂度, 训练条件概率和测试条件概率近似为

$$q(\mathbf{f} | \mathbf{f}_m) = \mathcal{N}(\mathbf{f} | K_{mm} K_{mm}^{-1} \mathbf{f}_m, \tilde{\mathbf{Q}}_{mm}) \quad (12)$$

$$q(\mathbf{f}_* | \mathbf{f}_m) = \mathcal{N}(\mathbf{f}_* | \mathbf{k}_{*m} K_{mm}^{-1} \mathbf{f}_m, \tilde{\mathbf{Q}}_{**}) \quad (13)$$

相应的, 式(7)中对数边际似然  $\log p(\mathbf{y})$  近似为

$$\log q(\mathbf{y}) = -\frac{1}{2} \log(2\pi) - \frac{1}{2} \log |\tilde{\mathbf{Q}}_{nn} + \mathbf{Q}_{nn} + \sigma_\varepsilon^2 \mathbf{I}| - \frac{1}{2} \mathbf{y}^\top (\tilde{\mathbf{Q}}_{nn} + \mathbf{Q}_{nn} + \sigma_\varepsilon^2 \mathbf{I})^{-1} \mathbf{y} \quad (14)$$

通过指定具体的  $\tilde{\mathbf{Q}}_{nn}$  与  $\tilde{\mathbf{Q}}_{**}$ , 就能以较低的计算复杂度  $O(n^2m)$  完成推断。

不同的诱导点方法给出了不同的假设以及不同的  $\tilde{\mathbf{Q}}_{nn}$  与  $\tilde{\mathbf{Q}}_{**}$ , 但目的都是将对  $K_{nn}$  求逆转换为对  $K_{mm}$  求逆, 其中比较著名的有回归量子集(SR)法、全独立条件(FIC)近似法、部分独立条件(PIC)近似法等。

(b) 后验近似

与先验近似不同, 后验近似保留精确先验, 对后验分布进行近似推断。通过引入变分分布  $q(\mathbf{f}, \mathbf{f}_m | \mathbf{y})$  近似后验分布  $p(\mathbf{f}, \mathbf{f}_m | \mathbf{y})$ , 它们之间的差异可以用 KL 散度表示

$$\begin{aligned} & KL(q(\mathbf{f}, \mathbf{f}_m | \mathbf{y}) \| p(\mathbf{f}, \mathbf{f}_m | \mathbf{y})) \\ &= \log p(\mathbf{y}) - \mathbb{E}_{q(\mathbf{f}, \mathbf{f}_m | \mathbf{y})} [\log p(\mathbf{f}, \mathbf{f}_m, \mathbf{y}) - \log q(\mathbf{f}, \mathbf{f}_m | \mathbf{y})] \\ &= \log p(\mathbf{y}) - \mathcal{L}(q) \end{aligned}$$

其中  $\mathcal{L}(q)$  为变分下界(EBLO)。  $\log p(\mathbf{y})$  对于  $q(\mathbf{f}, \mathbf{f}_m | \mathbf{y})$  是常数, 因此最大化变分下界等价于最小化 KL 散度。通过使用琴生不等式, 对变分下界进行优化, 便可以得到最优的变分分布  $q^*(\mathbf{f}_m)$ 。此时只需优化诱导点  $\mathbf{X}_m$  的位置便可得到最优后验分布。由此可见, 该方法对于模型的加速, 依旧体现在对诱导变量  $\mathbf{f}_m$  的使用上。另外, Huggins 等用 pF (preconditioned Fisher)散度替代 KL 散度, 以进一步提升性能[15]。Cheng 和 Boots 则从权重空间观点推导出了一个随机变分框架, 使推断更加灵活[16]。

## 4.2. 局部近似

局部近似通过将原始数据集划分为  $m$  个数据子集, 在每个数据子集上分别训练子 GPR 模型, 该方法可以通过并行训练各个子模型提升 GPR 模型的可扩展性。此外, 局部近似方法有易于捕获非平稳性、异方差性等局部特征。根据对数据子集的划分方式, 可分为简单局部专家模型和混合专家模型。

(a) 简单局部专家模型

众所周知, 距离较远的两个点之间的相关性较低。因此, 将预测点直接分配给距离预测点最近的数据子集训练出的子模型, 有望在低计算复杂度下产生合理预测[17]。简单局部专家模型通过聚类算法或树结构将原始数据划分  $M$  个静态区域, 并在各个区域训练相互独立的子模型, 则训练各个子模型的时间复杂度为  $O(nm_0^2)$ , 其中  $m_0 = n/M$  为每个子集的规模[18]。但此方法容易导致不同子集间的预测不连续。针对此问题, Park 等在相邻子集间添加了约束性条件; Urtasun 和 Darrell 则将距离预测点最近几个子模型的预测结果进行加权平均[19]。Chen 等将所有子模型的预测结果进行简单平均, 从而进一步提升预测精度, 同时预防过拟合问题[20]。

(b) 混合专家模型

不同于简单局部专家模型借助聚类算法对原始数据作静态划分, 混合专家模型通过概率模型对数据进行概率划分, 从而确保准确率和可靠性。混合专家模型通常表现为如下形式

$$p(\mathbf{y} | \mathbf{x}) = \sum_{i=1}^M g_i(\mathbf{x}) p_i(\mathbf{y} | \mathbf{x}) \quad (15)$$

其中  $g_i(\mathbf{x})$  为门控函数, 它对输入空间进行概率划分, 定义了各个子模型负责的子区域。  $p_i(\mathbf{y} | \mathbf{x})$  则由负责  $i$  区域的子模型  $M_i$  给出。通过梯度下降优化器或 EM 算法最大化对数似然, 可以同时学习门控函数和子模型[21]。  $\mathbf{x}_*$  的预测分布如下:

$$p(\mathbf{y}_* | D, \mathbf{x}_*) = \sum_{i=1}^M g_i(\mathbf{x}_* | D) p_i(\mathbf{y}_* | D, \mathbf{x}_*) \quad (16)$$

Bishop 和 Svenskn 采用贝叶斯方法代替最大似然方法, 在一定程度上消除了过拟合问题; Nguyen 等则使用了变分法进行模型推断, 进一步降低了时间复杂度[22] [23]。

## 5. 核函数的构造

核函数的选择决定了 GP 模型可以学习的结构类型, 并且通过将基本核函数进行相乘和相加可以构建出各种各样的组合核函数。但是, 对于具体问题, 构造合适的核函数仍然依赖于专家经验。通常情况下, 特别是对于多维数据, 核函数应该如何构造还很难。近些年来, 为了解决这个问题, 人们提出了许多不同的方法, 大体上可以分为以下两类。

### 5.1. 组合核搜索法

组合核搜索法将固定集合中的基本核函数相加或相乘来构建一个开放的组合核空间, 并使用各类算法来搜索这个开放的组合核空间, 以找到与数据结构匹配的核函数。搜索过程遵循一个简单的迭代过程, 其目的是在每一次迭代过程中改进上一次迭代中推断出的最佳协方差函数  $k_{best}$ 。该策略通过基本核函数来扩展之前的最佳协方差函数  $k_{best}$ , 从而产生一系列新的候选函数  $k_c$ , 并在其中比较出新的最佳协方差函数  $k_{best}$ 。不断迭代, 便可得到最契合数据结构的组合核函数  $k_{best}^*$ 。

Duvenaud 等人使用贪心算法对组合核空间进行搜索, 以固定集合中的基本核函数作为起点, 在每个阶段选择得分最高的核函数, 不断迭代直到进一步的复杂性不能改善模型质量或超出复杂性限制[24]。在模型比较阶段 Duvenaud 等人选择对数边际似然作为迭代过程中模型的比较标准, 并且引入贝叶斯信息准则(BIC)以避免过拟合。Steinruecken 则纳入 sigmoid 变点运算, 起到局部限制协方差函数的作用[25]。Kronberger 则通过语法定义有效协方差函数集, 并使用遗传规划算法来搜索该语法下可能的组合核空间[26]。该方法下, 核函数的超参数不受搜索的影响, 而是使用标准的梯度下降优化器进行优化。

经过该类方法学习的组合核函数通常可以分解成不同的、可解释的核函数组件, 因此拥有较强的可解释性。但该类算法多以对数边际似然评价模型质量, 计算成本过高, 这也是高斯过程框架本身固有的缺点[27]。这种模型的三次运行复杂度限制了此类算法在大规模数据集上的应用[28]。

### 5.2. 加权多核法

与自动搜索法不同, 加权多核法预设了具有加权和或加权乘积形式的组合核函数。通常表示为如下形式:

$$k_c(\mathbf{x}^i, \mathbf{x}^j, \Theta) = \prod_{l=1}^L w_l k_l(\mathbf{x}^i, \mathbf{x}^j, \Theta_l) \quad (17)$$

其中  $\boldsymbol{\omega} = (\omega_1, \dots, \omega_L)^\top$  为权重向量, 并且有  $\sum_l \omega_l = 1$ 。  $\Theta = (\theta_1, \dots, \theta_L)^\top$  是核函数的超参数。因此, 整体的超参数向量变为  $\mathbf{t} = (\Theta, \boldsymbol{\omega})^\top$ 。使用式(10)中定义的核函数可以直接计算协方差矩阵  $K$ 。

Chugh 使用多目标优化器来进行 GP 训练, 以调整各个核的超参数和权值, 并得到平衡数据拟合和模型复杂度[29]。Palar 等人则利用现有核的加权和组合构造一个新的组合核, 权重和核函数的超参数被同时训练[30]。

该类方法通过考虑核的加权组合, 更充分地利用了组合核的全部潜力。并且此类方法较自动搜索法所需训练成本更低。但组件核函数以及它们的超参数必须提前指定, 这在一定程度上限制了该类方法的能力。同时也导致模型的可解释性明显低于自动搜索法学习到的组合核函数。

## 6. 总结与展望

自 Williams 和 Rasmussen 第一次提出高斯过程回归至今已有 20 多年历史。近年来, 高斯过程回归方

法得到了快速发展。同时, 高斯过程回归的应用领域也愈加广泛。本文介绍了高斯过程回归的基本原理, 并综述了大数据背景下高斯过程回归模型的拓展形式, 以及针对复杂问题的组合核函数构造方法。

面向未来, 高斯过程回归无论在理论上还是在应用上, 仍有许多尚未解决但很有研究空间的问题。作为本文的结束, 我们提出高斯过程回归未来需要进一步研究的问题和方向。

#### 1) 应对在线数据

上述模型都基于完整数据集  $D$  可用来进行离线训练的假设。然而, 实践中普遍存在这样一种情况: 数据是连续到达的, 即在线数据或流数据, 以小批未知的方式到达。对于复杂的在线回归, 模型应做到:

(a) 对流数据具有实时适应性; (b) 能处理大规模的数据。

#### 2) 适应大数据的组合核函数构造法

本文中介绍的组合核函数构造方法都受限于高斯过程方法优化过程中较高的时间复杂度, 抑制了这些构造方法在大规模数据集中的应用。通过将本文介绍的近似方法与目前提出的组合核函数构造法相结合, 有望在一定程度上解决此问题。

## 参考文献

- [1] 张江帆. 基于高斯过程回归模型的锂电池数据处理[D]: [硕士学位论文]. 北京: 北京交通大学, 2017.
- [2] 郭忠义, 李永华, 李关辉, 彭志勇, 张宁, 于振中. 装备系统剩余使用寿命预测技术研究进展[J]. 南京航空航天大学学报, 2022, 54(3): 341-364.
- [3] Lally, N. and Hartman, B. (2018) Estimating Loss Reserves Using Hierarchical Bayesian Gaussian Process Regression with Input Warping. *Insurance: Mathematics and Economics*, **82**, 124-140.  
<https://doi.org/10.1016/j.insmatheco.2018.06.008>
- [4] 余柏杨, 吕宏强, 周岩, 罗振兵, 刘学军. 基于机器学习的高速复杂流场流动控制效果预测分析[J]. 实验流体力学, 2022, 36(3): 44-54.
- [5] 梁智, 孙国强, 俞娜燕, 倪晓宇, 沈海平, 卫志农. 基于高斯过程回归和粒子滤波的短期风速预测[J]. 太阳能学报, 2020, 41(3): 45-51.
- [6] Velásquez, R.M.A. and Lara, J.V.M. (2020) Forecast and Evaluation of COVID-19 Spreading in USA with Reduced-Space Gaussian Process Regression. *Chaos, Solitons & Fractals*, **136**, Article ID: 109924.  
<https://doi.org/10.1016/j.chaos.2020.109924>
- [7] 何志昆, 刘光斌, 赵曦晶, 王明昊. 高斯过程回归方法综述[J]. 控制与决策, 2013, 28(8): 1121-1129+1137.
- [8] 牛文杰. 薄板样条法和泛克里金法在理论和应用方面的比较[J]. 工程图学学报, 2010, 31(4): 123-129.
- [9] Preparata, F.P. and Shamos, M.I. (2012) Computational Geometry: An Introduction. *Springer Science & Business Media*, New York.
- [10] Lawrence, N., Seeger, M. and Herbrich, R. (2002) Fast Sparse Gaussian Process Methods: The Informative Vector Machine. *Advances in Neural Information Processing Systems* 15, Vancouver, 9-14 December 2002, 625-632.
- [11] Seeger, M. (2003) Bayesian Gaussian Process Models: PAC-Bayesian Generalisation Error Bounds and Sparse Approximations. University of Edinburgh, Edinburgh.
- [12] Keerthi, S. and Chu, W. (2005) A Matching Pursuit Approach to Sparse Gaussian Process Regression. *Proceedings of the 18th International Conference on Neural Information Processing Systems*, Vancouver, 5-8 December 2005, 643-650.
- [13] Melkumyan, A. and Ramos, F.T. (2009) A Sparse Covariance Function for Exact Gaussian Process Inference in Large Datasets. *21st International Joint Conference on Artificial Intelligence*, Pasadena, 11-17 July 2009, 1936-1942.
- [14] Williams, C.K.I. and Seeger, M. (2001) Using the Nystrom Method to Speed Up Kernel Machines. *Proceedings of the 13th International Conference on Neural Information Processing Systems*, Denver, January 2000, 661-667.
- [15] Huggins, J.H., Campbell, T., Kasprzak, M. and Broderick, T. (2019) Scalable Gaussian Process Inference with Finite-Data Mean and Variance Guarantees. *The 22nd International Conference on Artificial Intelligence and Statistics*, Naha, 16-18 April 2019, 796-805.
- [16] Cheng, C.A. and Boots, B. (2017) Variational Inference for Gaussian Process Models with Linear Complexity. *Annual Conference on Neural Information Processing Systems* 2017, Long Beach, 4-9 December 2017, 5190-5200.

- 
- [17] Choudhury, A., Nair, P.B. and Keane, A.J. (2002) A Data Parallel Approach for Large-Scale Gaussian Process Modeling. *Proceedings of the 2002 SIAM International Conference on Data Mining*, Arlington, 11-13 April 2002, 95-111. <https://doi.org/10.1137/1.9781611972726.6>
- [18] Gramacy, R.B. and Lee, H.K. (2008) Bayesian Treed Gaussian Process Models with an Application to Computer Modeling. *Journal of the American Statistical Association*, **103**, 1119-1130. <https://doi.org/10.1198/016214508000000689>
- [19] Urtasun, R. and Darrell, T. (2008) Sparse Probabilistic Regression for Activity-Independent Human Pose Inference. *2008 IEEE Conference on Computer Vision and Pattern Recognition*, Anchorage, 23-28 June 2008, 1-8. <https://doi.org/10.1109/CVPR.2008.4587360>
- [20] Chen, T. and Ren, J. (2009) Bagging for Gaussian Process Regression. *Neurocomputing*, **72**, 1605-1610. <https://doi.org/10.1016/j.neucom.2008.09.002>
- [21] Chen, K., Xu, L. and Chi, H. (1999) Improved Learning Algorithms for Mixture of Experts in Multiclass Classification. *Neural Networks*, **12**, 1229-1252. [https://doi.org/10.1016/S0893-6080\(99\)00043-X](https://doi.org/10.1016/S0893-6080(99)00043-X)
- [22] Bishop, C.M. and Svenskn, M. (2002) Bayesian Hierarchical Mixtures of Experts. In: D'Ambrosio, B., Smets, P. and Bonissone, P.P., *Uncertainty in Artificial Intelligence*, Morgan Kaufmann Publishers Inc., Burlington, MA, USA, 57-64.
- [23] Nguyen, T.N.A., Bouzardoum, A. and Phung, S.L. (2016) Variational Inference for Infinite Mixtures of Sparse Gaussian Processes through KL-Correction. *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, 20-25 March 2016, 2579-2583. <https://doi.org/10.1109/ICASSP.2016.7472143>
- [24] Duvenaud, D., Lloyd, J., Grosse, R., et al. (2013) Structure Discovery in Nonparametric Regression through Compositional Kernel Search. *International Conference on Machine Learning*, Atlanta, 16-21 June 2013, 1166-1174.
- [25] Steinruecken, C., Smith, E., Janz, D., Lloyd, J. and Ghahramani, Z. (2019) The Automatic Statistician. In: Hutter, F., Kotthoff, L. and Vanschoren, J., Eds., *Automated Machine Learning*, Springer, Cham, 161-173. [https://doi.org/10.1007/978-3-030-05318-5\\_9](https://doi.org/10.1007/978-3-030-05318-5_9)
- [26] Kronberger, G. and Kommenda, M. (2013) Evolution of Covariance Functions for Gaussian Process Regression Using Genetic Programming. *International Conference on Computer Aided Systems Theory*, Las Palmas de Gran Canaria, 10-15 February 2013, 308-315. [https://doi.org/10.1007/978-3-642-53856-8\\_39](https://doi.org/10.1007/978-3-642-53856-8_39)
- [27] Hensman, J., Fusi, N. and Lawrence, N.D. (2013) Gaussian Processes for Big Data. *Proceedings of the 29th Conference on Uncertainty in Artificial Intelligence*, 12-14 July 2013, arXiv preprint arXiv:1309.6835.
- [28] Kim, H. and The, Y.W. (2018) Scaling up the Automatic Statistician: Scalable Structure Discovery Using Gaussian Processes. *International Conference on Artificial Intelligence and Statistics*, Lanzarote, 9-11 April 2018, 575-584.
- [29] Chugh, T., Rahat, A. and Palar, P.S. (2019) Trading-Off Data Fit and Complexity in Training Gaussian Processes with Multiple Kernels. *International Conference on Machine Learning, Optimization, and Data Science*, Siena, 10-13 September 2019, 579-591. [https://doi.org/10.1007/978-3-030-37599-7\\_48](https://doi.org/10.1007/978-3-030-37599-7_48)
- [30] Palar, P.S. and Shimoyama, K. (2019) Kriging with Composite Kernel Learning for Surrogate Modeling in Computer Experiments. *AIAA Scitech 2019 Forum*, 7-11 January 2019, San Diego, 2209. <https://doi.org/10.2514/6.2019-2209>