标量对函数回归模型的分布式估计

Peilun He统计与应用概率系加利福尼亚大学圣巴巴拉分校 精算研究与商业分析系 麦考瑞大学

Han Lin Shang 精算研究与商业分析系 麦考瑞大学

Nan Zou 数学与物理科学学院 麦考瑞大学

arXiv:2601.04138v1 [stat.CO] 7 Jan 2026

摘要

{本文提出了一个用于标量对函数回归模型点推断和区间推断的分布式估计框架。}

本文针对三种标量对函数回归模型提出了分布式估计程序：函数型线性模型（FLM）、函数型非参数模型（FNPM）和函数型部分线性模型（FPLM）。该框架旨在应对函数型数据分析中的两项关键挑战，即大样本带来的高计算成本，以及机构之间共享原始数据的限制。蒙特卡洛模拟结果表明，对于这三种模型，分布式估计量在保持较高估计和预测准确度的同时，显著缩短了计算时间。当区块大小过小时，FPLM 会出现过拟合，导致预测区间变窄、经验覆盖概率降低。基于 *tecator* 数据集的实证研究示例进一步支持了这些发现。

关键词： 分布式学习、标量对函数回归、蒙特卡洛模拟

1 引言

函数型数据回归是函数型数据分析广泛研究领域中研究最为深入的主题之一。根据函数型数据在模型中所起的作用，回归模型可分为三类：标量响应和函数型预测变量（“标量对函数”回归）；函数型响应和标量预测变量（“函数对标量”回归）；以及函数型响应和函数型预测变量（“函数对函数”回归）。本文聚焦于第一种情形，并重新考察标量对函数回归中的线性模型、部分线性模型和非参数模型。标量对函数回归的应用领域包括化学计量学（Goutis，1998）、心脏病学（Ratcliffe 等人，2002a，b）、脑科学（Reiss 和 Ogden，2010）、气候科学（Ferraty 等人，2005）等。全面综述请参阅Morris(2015）、Reiss 等人(2017）以及Koner 和 Staicu(2023）。

传统上，标量对函数回归通常在单台计算机上使用相对较小的数据集进行。相比之下，现代应用日益涉及样本量庞大的函数型数据，这些数据分布在多个数据库中，例如由金融机构或医疗机构维护的数据库。随着样本量增长，拟合函数型回归模型的计算成本也会大幅增加。此外，在许多应用中，原始数据具有敏感性，因隐私或监管限制而无法传输至中央服务器。例如，患者医疗记录受隐私法规保护，不能在医院或研究机构之间共享。同样，银行持有的客户级交易记录高度机密，不能与其他金融机构共享。这些挑战使得将所有数据汇总到单台机器上进行分析变得不切实际。

一种常见的解决方案是在本地进行计算，然后将中间结果汇总至中央服务器。为应对这类分布式数据场景，人们开发了分布式学习方法，以便高效地传递并汇总来自本地数据库的信息。关于通用数据的分布式学习方法综述，参见 Mcdonald 等（2009）、Li 等（2013）、Zhang 等（2015）、Chang 等（2017）、Zhou 等（2024）和 Xiao 等（2024）。在函数型数据领域，Tong（2021）、Cai 等（2024）、Xue 等（2024）、Liu 等（2024）、Liu 和 Shi（2024）以及 Liu 和 Shi（2025）提出了相关理论进展。本文着重研究分布式学习方法在标量对函数回归中的实证表现。

本文针对三种标量对函数模型——函数型线性

模型、函数型非参数模型和函数型部分线性模型，提出了一种用于点估计、点预测和预测区间构造的分布式方法。该方法允许在数据划分后的各个子集上分别对每个模型进行本地拟合，仅将汇总后的中间结果传输至中央服务器，从而降低计算负担，并避免共享原始敏感数据。通过广泛的蒙特卡洛模拟以及对 Tecator 数据集的应用，考察了该方法的计算成本和统计准确性。结果表明，该分布式方法在保持较高预测准确度和数据隐私的同时，显著缩短了执行时间。尽管分而治之估计量已在函数型数据分析中得到广泛研究，但现有工作主要集中于线性情形（例如，参见 Tong，2021；Xue 等，2024；Liu 和 Shi，2025）。相比之下，本文在统一的分布式框架下，研究了线性、非参数和部分线性标量对函数模型。

本文结构如下。第2 节定义了三种模型：标量对函数线性模型、标量对函数非参数模型和标量对函数部分线性模型，并介绍了相应的参数估计方法。第3 节介绍了用于点估计和区间估计的分布式方法。第4 节和第5 节分别展示了模拟研究和 $tecator$ 数据分析的结果。最后，第6节总结全文，并探讨了扩展该方法的一些思路。

2 标量对函数回归模型

我们考虑三种类型的标量对函数回归模型：函数型线性模型（FLM）、函数型非参数模型（FNPM）和函数型部分线性模型（FPLM）。具体而言，对于 FLM，我们采用 B 样条展开和函数主成分分析（FPCA）两种估计方法。对于 FNPM 和 FPLM，我们采用一种基于核的方法，将 Nadaraya–Watson 估计量扩展到函数型设定中。

2.1 标量对函数线性模型

我们从标量对函数线性回归模型开始，其定义如下：

$$Y_i = \int_{\mathcal{T}} \beta(t)\mathcal{X}_i(t)dt + \epsilon_i, \tag{1}$$

其中，$Y_i \in \mathbb{R}$ 表示 $i = 1, \ldots, N$的标量响应，$\mathcal{X}_i \in \mathbb{H}$ 是取值于希尔伯特空间 $\mathbb{H}$ 、定义域为 $\mathcal{T}$的函数型协变量，$\beta \in \mathbb{H}$ 是相应的函数系数，$\epsilon_i \in \mathbb{R}$ 是与 $\mathcal{X}_i$独立的随机误差项，满足

$\mathbb{E}(\epsilon_i) = 0$ 和 $Var(\epsilon_i) = \sigma^2 < \infty$。

假设每个函数 $\mathcal{X}_i(\cdot)$ 仅在离散点 t_1、……、t_M处观测到。参数函数回归模型的目标是通过寻找使二次损失函数最小化的 β 来估计 β

$$L(\widehat{\beta}) = \frac{1}{N}\sum_{i=1}^{N}\left(Y_i - \int_{\mathcal{T}} \widehat{\beta}(t)\mathcal{X}_i(t)dt\right)^2.$$

Ramsay 和 Silverman (2005采用基展开方法估计函数系数 $\beta(\cdot)$，从而将无限维协变量转化为有限维表示。函数型协变量 $\mathcal{X}_i(t)$ 近似表示为

$$\mathcal{X}_i(t) \approx \sum_{p=1}^{P} c_{ip}\phi_p(t),$$

其中，$\phi_p(t)$ 表示已知的正交或非正交基函数，c_{ip} 是相应的系数。在离散时间点 t_1、……、t_M处，我们观测到 $\mathcal{X}_i(t_1)$ ……，$\mathcal{X}_i(t_M)$ 和 $\phi_p(t_1)$ ……，$\phi_p(t_M)$。通过最小化最小二乘准则估计系数 c_{ip}

$$\sum_{j=1}^{M}\left[\mathcal{X}_i(t_j) - \sum_{p=1}^{P} c_{ip}\phi_p(t_j)\right]^2.$$

类似地，函数系数 $\beta(t)$ 表示为

$$\beta(t) \approx \sum_{q=1}^{Q} b_q\psi_q(t),$$

其中，$\psi_q(t)$表示另一组带有系数 b_q的基函数。不失一般性，$\mathcal{X}_i(t)$和 $\beta(t)$可以采用不同的基组 $\phi_p(t)$和 $\psi_q(t)$，当然也可以共用同一组基函数。此外，我们假设 P 和 Q 均为固定且已知的。

将这些展开式代入（1）可得

$$\begin{aligned} Y_i &= \int_{\mathcal{T}} \beta(t)\mathcal{X}_i(t)dt + \epsilon_i \\ &= \int_{\mathcal{T}} \sum_{q=1}^{Q} b_q \psi_q(t) \sum_{p=1}^{P} c_{ip}\phi_p(t)dt + \epsilon_i \\ &= \sum_{q=1}^{Q} b_q \int_{\mathcal{T}} \psi_q(t) \sum_{p=1}^{P} c_{ip}\phi_p(t)dt + \epsilon_i \\ &= \sum_{q=1}^{Q} b_q U_{iq} + \epsilon_i, \end{aligned}$$

这对应于标准线性回归模型，其中

$$U_{iq} = \int_{\mathcal{T}} \psi_q(t) \sum_{p=1}^{P} c_{ip}\phi_p(t)dt$$

是确定性的。随后通过最小化来估计系数 b_q

$$\sum_{i=1}^{N} \left[Y_i - \sum_{q=1}^{Q} b_q U_{iq} \right]^2,$$

并得到估计的函数系数

$$\widehat{\beta}(t) = \sum_{q=1}^{Q} \widehat{b}_q \psi_q(t).$$

我们考虑两种基函数选择： $\phi_p(t)$ 和 $\psi_q(t)$。第一种选择是 B样条基，其中 $\phi_p(t)$ 和 $\psi_q(t)$ 均为 B样条函数，基元素的数量可能不同。在这种情况下，基函数是非正交且确定性的。第二种选择是在 $\mathcal{X}_i(t)$ 的经验协方差函数上通过函数主成分分析（FPCA）获得特征函数。此时， $\phi_p(t)$ 和 $\psi_q(t)$ 是正交归一特征函数 $Q = P$，且基是数据驱动的。我们分别将这两种方法称为 B样条展开和函数主成分分析方法。

2.2 标量对函数非参数模型

函数型非参数模型（FNPM）定义如下：

$$Y_i = m(\mathcal{X}_i(t)) + \epsilon_i, \tag{2}$$

其中，m: $\mathbb{H} \to \mathbb{R}$ 是一个未知的、可能为非线性的函数。假设误差项 ϵ_i 与 $\mathcal{X}_i$相互独立，且 $\mathbb{E}(\epsilon_i) = 0$ 、 $Var(\epsilon_i) = \sigma^2 < \infty$。这种表述使我们能够完全灵活地对函数型协变量与标量响应之间的关系进行建模，而无需施加任何特定的参数结构。

估计未知函数 $m(\cdot)$ 是 FNPM 中的一项关键挑战，主要因为 $m(\cdot)$ 的非线性，这使得第 2.1 节中的线性估计技术不再适用。我们采用 Ferraty 和 Vieu (2006) 提出的基于核的方法，将 Nadaraya–Watson 估计量推广到函数型设定。(·) 的核估计量定义如下：

$$\widehat{m}(\mathcal{X}) = \frac{\sum_{i=1}^{N} Y_i K(h^{-1} d(\mathcal{X}, \mathcal{X}_i))}{\sum_{i=1}^{N} K(h^{-1} d(\mathcal{X}, \mathcal{X}_i))}, \tag{3}$$

其中，K：$\mathbb{R} \to \mathbb{R}$ 是一个非对称核函数，$h > 0$ 是带宽参数，而d：$\mathbb{H} \times \mathbb{H} \to \mathbb{R}$ 是一个衡量函数型观测之间接近程度的半度量。

将归一化核权重定义为

$$w_h(\mathcal{X}, \mathcal{X}_i) = \frac{K(h^{-1} d(\mathcal{X}, \mathcal{X}_i))}{\sum_{i=1}^{N} K(h^{-1} d(\mathcal{X}, \mathcal{X}_i))},$$

估计量（3）可改写为

$$\widehat{m}(\mathcal{X}) = \sum_{i=1}^{N} w_h(\mathcal{X}, \mathcal{X}_i) Y_i,$$

它表示响应变量的加权平均 Y_i，其中权重满足$\sum_{i=1}^{N} w_h(\mathcal{X}, \mathcal{X}_i) = 1$。该估计量的渐近性质已在文献中得到广泛研究（例如参见，Masry，2005；Ferraty 和 Vieu，2006；Ferraty 等人，2007）。

我们采用如下定义的非对称正态核

$$K(x) = \begin{cases} \frac{2}{\sqrt{2\pi}} e^{-x^2/2} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases}$$

两个函数型观测之间的半度量定义为

$$d(\mathcal{X}, \mathcal{X}_i) = \|\mathcal{X} - \mathcal{X}_i\|_2 = \left(\int_{\mathcal{T}} |\mathcal{X}(t) - \mathcal{X}_i(t)|^2 \, dt \right)^{1/2}. \tag{4}$$

积分使用辛普森法则进行近似，并通过交叉验证选择带宽参数 h 。

2.3 标量—函数部分线性模型

函数型部分线性模型（FPLM）通过在标量响应 Y_i 与额外的标量协变量 Z_i 之间引入线性关系，扩展了 FNPM:

$$Y_i = \beta_{\mathrm{NF}} Z_i + m(\mathcal{X}_i(t)) + \epsilon_i, \tag{5}$$

其中， $Z_i \in \mathbb{R}$ 表示一个非函数型协变量，其对应系数为 $\beta_{\mathrm{NF}} \in \mathbb{R}$。在本文中，下标“NF”用于将与非函数型协变量相关的系数同与函数型预测变量相关的系数区分开来。在这种部分线性设定下，目标是同时估计 β_{NF} 和非线性函数型分量 $m(\cdot)$。

与 FNPM 不同，该公式包含额外的线性项 $\beta_{NF} Z_i$，这使得估计过程更加复杂。在这种情况下，既不能直接应用核估计量，也不能直接应用普通最小二乘法。

根据 Aneiros-P´erez 和 Vieu (2006)，定义 $\boldsymbol{Z} = [Z_1, \ldots, Z_N]^\top$， $\boldsymbol{Y} = [Y_1, \ldots, Y_N]^\top$，$\widetilde{\boldsymbol{Z}}_h = (\boldsymbol{I} - \boldsymbol{W}_h)\boldsymbol{Z}$， $\widetilde{\boldsymbol{Y}}_h = (\boldsymbol{I} - \boldsymbol{W}_h)\boldsymbol{Y}$，其中 $\boldsymbol{W}_h = \{w_h(\mathcal{X}_i, \mathcal{X}_j)\}_{N \times N}$是核权重矩阵。然后， β_{NF} 估计为

$$\widehat{\beta}_{NF} = (\widetilde{\boldsymbol{Z}}_h^\top \widetilde{\boldsymbol{Z}}_h)^{-1} \widetilde{\boldsymbol{Z}}_h^\top \widetilde{\boldsymbol{Y}}_h \tag{6}$$

且函数型分量 $m(\cdot)$ 的估计为

$$\widehat{m}(\mathcal{X}) = \sum_{i=1}^{N} w_h(\mathcal{X}, \mathcal{X}_i)(Y_i - \widehat{\beta}_{NF} Z_i) = \frac{\sum_{i=1}^{N} K(h^{-1} d(\mathcal{X}, \mathcal{X}_i))(Y_i - \widehat{\beta}_{NF} Z_i)}{\sum_{i=1}^{N} K(h^{-1} d(\mathcal{X}, \mathcal{X}_i))}. \tag{7}$$

拟合响应则为

$$\widehat{Y}_i = \widehat{\beta}_{NF} Z_i + \widehat{m}(\mathcal{X}_i).$$

这些部分线性估计量已在多个方向上得到进一步拓展。例如，

例如，Novo 等人（2021a）提出了一种快速且灵活的 k近邻（k近邻）核估计量扩展，以提高局部自适应性；而Novo 等人（2021b）则引入了式（6）的惩罚版本，以实现协变量选择。

对于 FPLM，核函数 $K(\cdot)$ 和半度量 $d(\cdot, \cdot)$ 的选择与 FNPM 中所用的相同。

3 方法

虽然第 2 节介绍了函数型回归模型的估计方法，但这些方法在实际应用中可能存在局限。首先，随着样本量增加，计算时间可能大幅增长。其次，在许多应用中，出于隐私或监管限制，机构之间无法共享原始数据。例如，医院可能各自收集患者数据，但禁止对外共享这些记录。因此，亟须开发相应方法，使模型估计能够在各站点利用本地数据完成，然后将所得估计值汇总至中央服务器。在第 3.1 节中，我们针对函数型回归模型（1）、（2）和（5）提出了分布式估计程序。这些分布式估计量在保护数据隐私的同时，大幅降低了计算成本。第 3.2 节介绍了用于区间估计和预测准确度的分布式程序。

3.1 分布式点估计

令 $\mathcal{D} = \{(\mathcal{X}_i, Y_i)\}_{i=1}^{N_{\text{train}}}$表示大小为 N_{train} 的训练数据集，并令 $S = \{1, \ldots, N_{\text{train}}\}$表示训练样本的索引集。该数据集被划分为 K 个互不重叠的子集 $S_k \subset S$，其中 S_k 对应于第 k^{th} 个数据块的索引。[1] 该划分满足$\cup_{k=1}^{K} S_k = S$，并且对于任意 $k_1 \neq k_2$都有 $S_{k_1} \cap S_{k_2} = \varnothing$。为简单起见，我们假设所有数据块大小相等，即对于所有 $k = 1, \ldots\ldots, K$，均有 $|S_k| = \frac{N_{\text{train}}}{K} = n$。 ——

在每个数据块内计算一个局部估计量。令 $\widehat{Y}^{(k)}(\cdot)$ 表示第 k^{th} 个数据块中响应变量的局部估计量。由于各数据块互不重叠，训练数据的全局（分布式）估计量可表示为

$$\widehat{Y}(\mathcal{X}_i) = \sum_{k=1}^{K} \widehat{Y}^{(k)}(\mathcal{X}_i) \cdot \mathbb{1}\left[i \in S_k\right],$$

[1]只有训练数据会被划分为数据块，而在所有实验中，测试数据均完整使用。

其中， $\mathbb{1}[\cdot]$ 表示示性函数。

对于函数线性模型，我们还估计函数系数 β。令 $\widehat{\beta}^{(k)}$ 表示由 k^{th} 数据块得到的局部估计量。然后，对所有局部估计值取平均，得到全局估计量：

$$\widehat{\beta} = \frac{1}{K}\sum_{k=1}^{K}\widehat{\beta}^{(k)}. \tag{8}$$

现在，我们转向对新观测值进行预测。令 $\mathcal{D}^* = \{(\mathcal{X}_\iota^*,\ Y_\iota^*)\}_{\iota=1}^{N_{\text{test}}}$表示大小为 N_{test} 的测试数据集。对于任意新的函数型协变量 $\mathcal{X}_\iota^*$，每个局部估计量 $\widehat{Y}_{(k)}(\cdot)$都会生成一个局部预测，最终的全局预测则通过对所有K 个区块取平均得到：

$$\widehat{Y}(\mathcal{X}_\iota^*) = \frac{1}{K}\sum_{k=1}^{K}\widehat{Y}^{(k)}(\mathcal{X}_\iota^*). \tag{9}$$

3.2 分布式区间估计

我们提出一种分布式方法，使用Y_i 最初由 Shang and Haberman(2025) 提出的基于标准差的方法，构造响应变量的预测区间。

令 $Y_{(k)\,i}$ 和 $\widehat{Y}_i^{(k)}$ 分别表示 k^{th}数据块中第 i^{th} 个观测值的观测值和拟合值。估计误差定义为

$$\widehat{\epsilon}_i^{(k)} = Y_i^{(k)} - \widehat{Y}_i^{(k)}.$$

然后计算此数据块内残差的标准差，计算公式为

$$\widehat{\sigma}^{(k)} = \text{sd}\left(\widehat{\epsilon}_i^{(k)}\right).$$

对于给定的显著性水平 α，我们寻求一个值 $\gamma^{(k)}$ ，使得

$$\Pr\left(-\gamma^{(k)}\widehat{\sigma}^{(k)} \le \widehat{\epsilon}_i^{(k)} \le \gamma^{(k)}\widehat{\sigma}^{(k)}\right)$$

近似名义覆盖概率（$1-\alpha$） $\times 100\%$。根据大数定律，该概率可通过经验方式近似为

$$\Pr\left(-\gamma^{(k)}\widehat{\sigma}^{(k)} \le \widehat{\epsilon}_i^{(k)} \le \gamma^{(k)}\widehat{\sigma}^{(k)}\right) \approx \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\left[-\gamma^{(k)}\widehat{\sigma}^{(k)} \le \widehat{\epsilon}_i^{(k)} \le \gamma^{(k)}\widehat{\sigma}^{(k)}\right].$$

该量 $\gamma^{(k)}$ 通过最小化覆盖概率差（CPD）得到：

$$\mathrm{CPD}^{(k)} = \left| \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\left[-\gamma^{(k)} \widehat{\sigma}^{(k)} \leq \widehat{\epsilon}_i^{(k)} \leq \gamma^{(k)} \widehat{\sigma}^{(k)} \right] - (1-\alpha) \right|.$$

然后对 K 个数据块取平均，得到全局估计值 γ 和 $\widehat{\sigma}$：

$$\gamma = \frac{1}{K} \sum_{k=1}^{K} \gamma^{(k)}, \tag{10}$$

$$\widehat{\sigma} = \frac{1}{K} \sum_{k=1}^{K} \widehat{\sigma}^{(k)}. \tag{11}$$

对于测试数据，采用相同的 γ 和 $\widehat{\sigma}$ 全局值。预测区间为

$$\widehat{Y}_{\iota}^{*} \pm \gamma \widehat{\sigma},$$

其中 $\widehat{Y}_{\iota}^{*}$是通过第3.1节所述的分布式点估计程序得到的全局预测。

4 蒙特卡洛模拟

本节介绍模拟研究的结果。数据生成过程（DGP）见第4.1节，其中针对每个模型独立生成训练集和测试数据集。用于评估样本内估计和样本外预测准确度的评估准则见第4.2节。第4.3 节分别报告函数线性模型（FLM）、FNPM和FPLM的结果。

4.1 数据生成过程

函数线性模型的数据生成过程。对于式（1）中的函数线性模型，我们采用与Beyaztas 和 Shang（2022）及Beyaztas 等人（2024）相同的数据生成过程。对于任意 $t \in [0,1]$，函数型协变量生成如下：

$$\mathcal{X}_i(t) = \sum_{j=1}^{5} k_{ij} v_j(t),$$

其中 $k_{ij} \sim N(0,\ 4j^{-3/2})$。基函数定义为

$$v_j(t) = \sin(j\pi t) - \cos(j\pi t),$$

且系数函数 $\beta(t)$ 指定为

$$\beta(t) = 2\sin(2\pi t).$$

响应变量 Y_i 按照以下方式生成

$$Y_i = \int_0^1 \beta(t)\mathcal{X}_i(t)dt + \epsilon_i,$$

其中 $\epsilon_i \sim N(0, 1)$。在离散情形下，定义域 $[0,1]$ 在步长为 0.01 的等间距网格上观测。

FNPM 的数据生成过程。对于式（2）中的 FNPM，我们采用Ferraty 等人（2007）提出的数据生成过程。对于任意 $t \in [-1,1]$，函数型协变量生成如下：

$$\mathcal{X}_i(t) = \cos(\omega_i t) + (a_i + 2\pi)t + b_i,$$

其中 a_i 和 b_i 服从 $[0,1]$, 上的均匀分布， ω_i 服从 $[0,2\pi]$上的均匀分布。函数 $m(\cdot)$ 定义为

$$m(\mathcal{X}(t)) = \int_{-1}^1 |\mathcal{X}'(t)|(1-\cos(\pi t))dt,$$

误差项 ϵ_i 服从 $N(0, 2)$。

FPLM 的数据生成过程。对于式（5）中的 FPLM，我们对 $\mathcal{X}_i(t)$ 、Y_i和 $m(\cdot)$ 采用与 FNPM 相同的模拟设置。非函数型协变量 Z_i 服从标准正态分布，非函数型系数设为 $\beta_{NF} = 0.5$。

4.2 评估准则

现在介绍用于评估模型性能的评价指标。为了评估训练数据中响应变量 Y_i 的样本内估计精度，我们使用均方根

平方误差 （RMSE）、相对误差（RE）和平均绝对误差（MAE），定义如下 s

$$\mathrm{RMSE} = \sqrt{\frac{1}{N_{\mathrm{train}}} \sum_{k=1}^{K} \sum_{i=1}^{n} \left(Y_i^{(k)} - \widehat{Y}_i^{(k)} \right)^2},$$

$$\mathrm{RE} = \frac{1}{N_{\mathrm{train}}} \sum_{k=1}^{K} \sum_{i=1}^{n} \left| \frac{Y_i^{(k)} - \widehat{Y}_i^{(k)}}{Y_i^{(k)}} \right|,$$

$$\mathrm{MAE} = \frac{1}{N_{\mathrm{train}}} \sum_{k=1}^{K} \sum_{i=1}^{n} \left| Y_i^{(k)} - \widehat{Y}_i^{(k)} \right|,$$

其中， $n = N_{\mathrm{train}}/K$ 表示每个数据块中的观测值数量， $Y_i^{(k)}$ 和 $\widehat{Y}_i^{(k)}$ 分别表示 k^{th} 数据块中第 i^{th} 个观测值的真实响应值和估计响应值。

对于函数线性模型，我们还使用重缩放弗罗贝尼乌斯范数、点偏差和点标准差来评估函数系数 $\beta(t)$的估计精度。重缩放弗罗贝尼乌斯范数（F.Norm）定义为

$$\mathrm{F.Norm} = \left\| \boldsymbol{\beta} - \widehat{\boldsymbol{\beta}} \right\|_F^2 \approx \frac{1}{M} \sum_{m=1}^{M} \left(\beta(t_m) - \widehat{\beta}(t_m) \right)^2,$$

其中， $\beta(\cdot)$在由 M个点 $t_1, \ldots, t_M$组成的密集网格上进行评估，并且

$$\boldsymbol{\beta} = [\beta(t_1), \ldots, \beta(t_M)]^\top, \qquad \widehat{\boldsymbol{\beta}} = [\widehat{\beta}(t_1), \ldots, \widehat{\beta}(t_M)]^\top.$$

$\widehat{\ }$全局估计量 $\boldsymbol{\beta}$ 使用式（8）得到。

蒙特卡罗重复试验中的点偏差和点标准差（ST.DEV）计算如下:

$$\mathrm{Bias}^2 = \frac{1}{M} \sum_{m=1}^{M} \left[\mathbb{E}\left(\widehat{\beta}(t_m) \right) - \beta(t_m) \right]^2,$$

$$\mathrm{ST.DEV} = \sqrt{\frac{1}{M} \sum_{m=1}^{M} \mathrm{Var}\left(\widehat{\beta}(t_m) \right)}.$$

为评估测试数据的样本外预测准确度，我们使用均方根预测误差（RMSFE）、相对预测误差（RFE）和平均绝对预测误差

（MAFE），定义为

$$\text{RMSFE} = \sqrt{\frac{1}{N_{\text{test}}} \sum_{\iota=1}^{N_{\text{test}}} \left(Y_\iota^* - \widehat{Y}_\iota^*\right)^2},$$

$$\text{RFE} = \frac{1}{N_{\text{test}}} \sum_{\iota=1}^{N_{\text{test}}} \left| \frac{Y_\iota^* - \widehat{Y}_\iota^*}{Y_\iota^*} \right|,$$

$$\text{MAFE} = \frac{1}{N_{\text{test}}} \sum_{\iota=1}^{N_{\text{test}}} \left| Y_\iota^* - \widehat{Y}_\iota^* \right|,$$

其中，全局预测 $\widehat{Y}_\iota^*$通过（9）计算得到。

对于区间估计，我们计算经验覆盖概率（ECP）和区间得分（IS）。对于训练数据，ECP 定义为

$$\text{ECP}_{\text{train}} = \frac{1}{N_{\text{train}}} \sum_{i=1}^{N_{\text{train}}} \mathbb{1}\left[\widehat{Y}_i - \gamma\widehat{\sigma} \le Y_i \le \widehat{Y}_i + \gamma\widehat{\sigma}\right],$$

其中 γ 和 $\widehat{\sigma}$ 由式（10）和（11）得到。

对于测试数据，采用相同的 γ 和 $\widehat{\sigma}$，相应的 ECP 计算如下

$$\text{ECP}_{\text{test}} = \frac{1}{N_{\text{test}}} \sum_{\iota=1}^{N_{\text{test}}} \mathbb{1}\left[\widehat{Y}_\iota^* - \gamma\widehat{\sigma} \le Y_\iota^* \le \widehat{Y}_\iota^* + \gamma\widehat{\sigma}\right].$$

对于一个中心（$1-\alpha$）$\times 100\%$ 预测区间，其下限为 $\widehat{Y}_\iota^* - \gamma\widehat{\sigma}$，上限为$\widehat{Y}_\iota^* + \gamma\widehat{\sigma}$，根据Gneiting 和 Raftery (2007）中式（43）的定义，区间得分为

$$\text{IS} = \frac{1}{N_{\text{test}}} \sum_{\iota=1}^{N_{\text{test}}} \Big\{ 2\gamma\widehat{\sigma} + \frac{2}{\alpha}(\widehat{Y}_\iota^* - \gamma\widehat{\sigma} - Y_\iota^*)\mathbb{1}\left[Y_\iota^* < \widehat{Y}_\iota^* - \gamma\widehat{\sigma}\right] + \frac{2}{\alpha}(Y_\iota^* - \widehat{Y}_\iota^* - \gamma\widehat{\sigma})\mathbb{1}\left[Y_\iota^* > \widehat{Y}_\iota^* + \gamma\widehat{\sigma}\right] \Big\}.$$

4.3 结果

本节介绍模拟研究的结果。对于每项研究，完整训练样本由 2000 个观测值组成，并被分为数量不同的数据块，区块数量范围为 1 至 40。测试样本采用随机抽取方式，样本量固定为 200、400 或 800。所有评估准则均基于 200 次蒙特卡罗重复试验取平均值。

4.3.1 标量对函数线性模型

表 1 报告了使用 B样条展开法的函数线性模型（FLM）的估计和预测性能。随着数据块数量 K 从 1 增加到 40，每个数据块的样本量从 2000 减少到 50。对于样本内估计，随着 K 增加，均方根误差（RMSE）和平均绝对误差（MAE）均略有下降，这表明响应变量的估计精度略有提高。相比之下，随着 K增大，函数系数的估计精度略有下降，这体现在弗罗贝尼乌斯范数、点偏差和标准差均有所增加。然而，这些增幅都非常小，表明即使进行大幅度的数据划分，对 $\beta(t)$ 的估计精度损失仍然有限。对于样本外预测，在每个固定的测试样本量下，均方根预测误差（RMSFE）和平均绝对预测误差（MAFE）均保持相对稳定。随着 K 增加，这些数值仅发生轻微变化，并且这一结论在所考察的所有测试样本量下均保持一致。

表1：在不同数据块数量和测试样本量下，使用 B 样条展开法的函数线性模型的估计和预测性能。所有结果均为 200 次蒙特卡洛模拟的平均值。

		训练数据						测试数据		
		响应变量			函数系数			响应变量		
K	Time	$N_{\text{训练}}$	RMSE	MAE	弗罗贝尼乌斯范数	偏差^2	标准差	N_{test}	均方根预测误差	MAFE
1	2.8880秒	2000	0.9997	0.7978	0.5515	0.4775	0.2728	200	0.9962	0.7971
								400	0.9975	0.7956
								800	1.0029	0.8010
2	0.3728秒	1000	0.9982	0.7966	0.5516	0.4769	0.2740	200	0.9962	0.7971
								400	0.9975	0.7956
								800	1.0029	0.8010
5	0.0332秒	400	0.9938	0.7932	0.5580	0.4823	0.2758	200	0.9962	0.7971
								400	0.9975	0.7957
								800	1.0030	0.8010
10	0.0092秒	200	0.9862	0.7870	0.5547	0.4769	0.2796	200	0.9961	0.7971
								400	0.9976	0.7957
								800	1.0030	0.8010
20	0.0051秒	100	0.9709	0.7748	0.5479	0.4683	0.2828	200	0.9963	0.7972
								400	0.9976	0.7957
								800	1.0030	0.8010
40	0.0041秒	50	0.9398	0.7498	0.5671	0.4780	0.2992	200	0.9964	0.7972
								400	0.9978	0.7958
								800	1.0032	0.8012

我们省略函数线性模型的相对误差和相对预测误差。在此模拟设计中，响应变量 Y_i 以零为中心。由于相对误差和相对预测误差都需要除以 Y_i，Y_i 的较小取值会使

相对误差大幅膨胀，从而产生具有误导性的较大统计量。

执行时间也作为衡量计算效率的指标进行报告。我们将其定义为在单个数据块上拟合模型所需的平均时间。假设有 K 台机器可用于并行运行计算，该平均时间可作为计算效率的合理代理指标。随着 K 增加以及区块大小减小，每个数据块的平均执行时间以及总执行时间[2] 都会大幅下降，体现了分布式方法在计算方面的优势。

表2 展示了 K取不同值时的经验覆盖概率和区间得分，其中 $\alpha = 0.05$ 以及 $\alpha = 0.20$。当$K = 1$时，训练数据的经验覆盖概率恰好等于名义覆盖概率$1 - \alpha$。随着 K 增加，经验覆盖概率略微下降至低于 $1 - \alpha$。测试数据也呈现出类似的模式，但与训练数据相比，测试数据的经验覆盖概率下降幅度更大。

表2：在不同数据块数量下，采用B样条展开法的函数线性模型的预测区间的平均经验覆盖概率（经验覆盖概率）和平均区间得分（区间得分）。所有结果均基于200次蒙特卡洛模拟取平均值。

		训练数据		$N_{\text{test}} = 200$		$N_{\text{test}} = 400$		$N_{\text{test}} = 800$	
判据	K	$\alpha = 0.05$	$\alpha = 0.2$	$\alpha = 0.05$	$\alpha = 0.2$	$\alpha = 0.05$	$\alpha = 0.2$	$\alpha = 0.05$	$\alpha = 0.2$
ECP	1	95.00%	80.00%	94.97%	80.23%	94.87%	80.16%	94.91%	79.84%
	2	94.95%	79.98%	94.94%	80.15%	94.82%	80.08%	94.86%	79.73%
	5	94.87%	79.91%	94.75%	79.85%	94.66%	79.83%	94.67%	79.47%
	10	94.76%	79.86%	94.53%	79.43%	94.42%	79.41%	94.40%	79.03%
	20	94.53%	79.71%	93.90%	78.43%	93.77%	78.57%	93.72%	78.18%
	40	94.06%	79.41%	92.53%	76.59%	92.37%	76.76%	92.27%	76.32%
IS	1	4.6707	3.5079	4.6678	3.4947	4.6896	3.5090	4.6911	3.5192
	2	4.6646	3.5030	4.6683	3.4948	4.6893	3.5088	4.6915	3.5193
	5	4.6428	3.4865	4.6690	3.4945	4.6907	3.5093	4.6926	3.5196
	10	4.6086	3.4601	4.6702	3.4950	4.6946	3.5098	4.6970	3.5205
	20	4.5433	3.4073	4.6862	3.4981	4.7140	3.5119	4.7164	3.5239
	40	4.4087	3.2992	4.7536	3.5102	4.7868	3.5242	4.7910	3.5374

对于区间得分，训练集与测试集的表现有所不同。当 $K = 1$ 时，两个数据集的区间得分值非常相似。随着 K 增大，训练数据的区间得分下降，表明预测区间变窄，精度提高。然而，对于所有测试样本量，测试数据的区间得分均随着 K 稳步上升。由于均方根预测误差和平均绝对预测误差保持

[2]这里，“总执行时间”指的是在单台机器上处理大型数据集、以缩短整体计算时间的情形。它可以近似表示为每个数据块的执行时间乘以数据块数量。

基本不变，经验覆盖概率的下降意味着预测区间正在变窄。因此，更多测试观测值落在区间之外，直接影响区间得分。

表3 报告了使用函数主成分分析方法的函数线性模型的估计和预测性能。与 B 样条方法一样，所有评价指标随 K的不同取值变化幅度都很小。然而，与 B 样条方法相比，函数主成分分析得到的函数系数估计值更准确，这体现在其弗罗贝尼乌斯范数、偏差和标准差均较低。这一改进源于 B 样条展开对曲线进行平滑近似，而函数主成分分析通过最优表示经验协方差结构，能够捕捉更多局部变异性。

表3：在不同数据块数量和测试样本量下，使用函数主成分分析方法的函数线性模型的估计和预测性能。所有结果均为200次蒙特卡洛模拟的平均值。

		训练数据						测试数据		
		响应变量			函数系数			响应变量		
K	Time	$N_{\text{训练}}$	RMSE	MAE	弗罗贝尼乌斯范数	偏差^2	标准差	N_{test}	均方根预测误差	MAFE
1	0.7949秒	2000	0.9997	0.7978	0.1550	0.1512	0.0619	200	0.9962	0.7971
								400	0.9975	0.7956
								800	1.0029	0.8010
2	0.3076秒	1000	0.9982	0.7966	0.1550	0.1512	0.0620	200	0.9962	0.7971
								400	0.9975	0.7956
								800	1.0029	0.8010
5	0.1533秒	400	0.9938	0.7932	0.1551	0.1512	0.0628	200	0.9962	0.7971
								400	0.9975	0.7956
								800	1.0030	0.8010
10	0.0896秒	200	0.9862	0.7870	0.1551	0.1512	0.0631	200	0.9961	0.7971
								400	0.9976	0.7957
								800	1.0030	0.8010
20	0.0561秒	100	0.9709	0.7748	0.1552	0.1512	0.0639	200	0.9963	0.7972
								400	0.9976	0.7957
								800	1.0030	0.8010
40	0.0299秒	50	0.9398	0.7498	0.1556	0.1512	0.0669	200	0.9964	0.7972
								400	0.9978	0.7958
								800	1.0032	0.8012

两种方法在执行时间上的差异更为显著。当$K = 1$ （区块大小为 2000）时，函数主成分分析明显更快（约 0.8 秒），而 B 样条方法约需 2.9 秒。随着 K 增加，B 样条方法的执行时间迅速缩短，当 K 翻倍时，执行时间约减少至原来的十分之一；相比之下，函数主成分分析的执行时间仅适度缩短，当 K 翻倍时，大约减半。

这一差异反映出，在我们的模拟研究中，与计算少量主成分相比，样条拟合的计算开销更大。

表 4报告了使用函数主成分分析的函数线性模型的平均经验覆盖概率和区间得分。与B样条方法一样，随着 K 增加，这两项指标均发生了小幅变化。经验覆盖概率略有下降，而区间得分略有上升，这与预测区间变窄、经验覆盖率降低的情况一致。

表4： 在不同数据块数量下，使用函数主成分分析方法的函数线性模型预测区间的平均经验覆盖概率和平均区间得分。所有结果均为200次蒙特卡洛模拟的平均值。

		训练数据		$N_{\text{test}} = 200$		$N_{\text{test}} = 400$		$N_{\text{test}} = 800$	
判据	K	$\alpha = 0.05$	$\alpha = 0.2$	$\alpha = 0.05$	$\alpha = 0.2$	$\alpha = 0.05$	$\alpha = 0.2$	$\alpha = 0.05$	$\alpha = 0.2$
ECP	1	95.00%	80.00%	94.97%	80.23%	94.87%	80.16%	94.91%	79.84%
	2	94.95%	79.98%	94.94%	80.15%	94.82%	80.08%	94.86%	79.73%
	5	94.87%	79.91%	94.75%	79.85%	94.66%	79.83%	94.67%	79.47%
	10	94.76%	79.86%	94.53%	79.43%	94.42%	79.41%	94.40%	79.03%
	20	94.53%	79.71%	93.90%	78.43%	93.77%	78.57%	93.72%	78.18%
	40	94.06%	79.41%	92.53%	76.60%	92.37%	76.76%	92.27%	76.32%
IS	1	4.6707	3.5079	4.6678	3.4947	4.6896	3.5090	4.6911	3.5192
	2	4.6646	3.5030	4.6683	3.4948	4.6893	3.5088	4.6915	3.5193
	5	4.6428	3.4865	4.6690	3.4945	4.6907	3.5093	4.6926	3.5196
	10	4.6086	3.4601	4.6702	3.4950	4.6946	3.5098	4.6970	3.5205
	20	4.5433	3.4073	4.6862	3.4981	4.7140	3.5119	4.7164	3.5239
	40	4.4087	3.2992	4.7536	3.5102	4.7868	3.5242	4.7910	3.5374

总体而言，对于函数线性模型，增加数据块的数量会使样本内估计准确性和样本外预测性能略有下降，同时预测区间也会略微变窄。不过，这些损失非常小，而计算时间却大幅缩短。因此，分布式估计在统计性能仅略有下降的情况下，显著提高了效率。

4.3.2 标量对函数非参数模型

表5 总结了 FNPM 的估计和预测性能。对于训练数据的样本内估计，三项评估准则（RMSE、RE 和 MAE）均显示，随着区块数量增加，性能略有提升。当 K 从 1 增至 40 时，RMSE 从 1.3971 降至 1.2776，RE 从 8.34% 降至 7.53%，MAE 从 1.1147 降至 1.0073。这些结果表明，将数据分成更小的数据块，可使训练样本的拟合准确性略有提高。

表5： 不同数据块数量和测试样本量下 FNPM 的估计和预测性能。所有结果均为 200 次蒙特卡洛模拟的平均值。

		响应变量							
		训练数据				测试数据			
K	Time	$N_{\text{训练}}$	RMSE	RE	MAE	N_{test}	均方根预测误差	RFE	MAFE
1	29.37秒	2000	1.3971	8.34%	1.1147	200	1.4182	8.49%	1.1341
						400	1.4365	8.60%	1.1478
						800	1.4335	8.56%	1.1429
2	7.61秒	1000	1.3757	8.21%	1.0977	200	1.4190	8.49%	1.1348
						400	1.4373	8.60%	1.1484
						800	1.4343	8.56%	1.1434
5	1.41秒	400	1.3494	8.05%	1.0762	200	1.4249	8.53%	1.1395
						400	1.4428	8.64%	1.1528
						800	1.4400	8.60%	1.1480
10	0.46秒	200	1.3324	7.94%	1.0616	200	1.4347	8.59%	1.1477
						400	1.4524	8.70%	1.1603
						800	1.4495	8.66%	1.1556
20	0.16秒	100	1.3170	7.82%	1.0456	200	1.4494	8.69%	1.1600
						400	1.4666	8.78%	1.1717
						800	1.4641	8.75%	1.1675
40	0.07秒	50	1.2776	7.53%	1.0073	200	1.4612	8.76%	1.1698
						400	1.4783	8.85%	1.1811
						800	1.4758	8.82%	1.1769

对于样本外预测，不同区块大小之间的差异要小得多。随着 K 增加，预测准确度仅略有下降。例如，在测试样本量为 800 时，将 K 从 1 增至 40，预测误差仅小幅增加：RMSFE 从 1.4335 升至 1.4758，RFE 从 8.56% 升至 8.82%，MAFE 从 1.1429 升至 1.1769。总体而言，分块后，FNPM 对训练数据的拟合准确性略有提高，但预测误差也略有增加。

分布式估计对 FNPM 最显著的优势体现在执行时间上。当 K 加倍时，执行时间会缩短至四分之一。这一现象与式（3）中核估计量的理论计算复杂度完全一致。对于单个函数型预测变量 $\mathcal{X}_j$，估计 $\widehat{m}(\mathcal{X}_j)$ 的计算分为三个步骤：

1) 对所有 d，$\mathcal{X}_j$，$\mathcal{X}_i$，计算（4）中的半度量 $i = 1, \ldots\ldots, n$，这需要 $\mathcal{O}(Mn)$ 次运算。

2）计算核函数值 $K(h^{-1}d(\mathcal{X}_j,\ \mathcal{X}_i))$，这需要 $\mathcal{O}(n)$ 次运算 s.

3) 计算加权平均，得到 $\widehat{m}(\mathcal{X}_j)$，这也需要 $\mathcal{O}(n)$ 次运算。

因此，估计一个函数的总复杂度为 $\mathcal{O}(Mn)$，而数据块中所有 n 个观测值的总复杂度为 $\mathcal{O}(Mn^2)$。如果 K 加倍，则数据块大小 n 减半，因此总计算负担将降至四分之一。表 5 中观察到的执行时间与这一理论预测高度一致。

表6 报告了不同数据块数量下的平均经验覆盖概率和区间得分。随着 K 增加，训练数据和测试数据的经验覆盖概率均有所下降，且测试数据的降幅更大。对于训练数据，随着 K 增加，区间得分略有下降，表明区间变窄，样本内精度有所提高。然而，对于测试数据，区间得分随 K增加而上升，这与函数线性模型中观察到的情况一致。由于预测误差在不同 K取值下保持相对稳定，经验覆盖概率下降意味着区间变窄，而区间得分升高仅反映出落在这些区间之外的测试观测值数量增加。

表6： 不同区块数量下，FNPM 预测区间的平均经验覆盖概率（ECP）和平均区间得分（IS）。所有结果均基于200次蒙特卡洛模拟取平均值。

		训练数据		$N_{\text{test}}=200$		$N_{\text{test}}=400$		$N_{\text{test}}=800$	
判据	K	$\alpha=0.05$	$\alpha=0.2$	$\alpha=0.05$	$\alpha=0.2$	$\alpha=0.05$	$\alpha=0.2$	$\alpha=0.05$	$\alpha=0.2$
ECP	1	95.00%	80.00%	94.57%	79.22%	94.34%	78.70%	94.30%	78.85%
	2	94.95%	79.97%	94.12%	78.43%	93.87%	77.91%	93.85%	78.10%
	5	94.87%	79.92%	93.49%	77.20%	93.15%	76.78%	93.17%	76.94%
	10	94.71%	79.80%	92.79%	76.24%	92.40%	75.61%	92.45%	75.92%
	20	94.32%	79.64%	91.83%	74.83%	91.42%	74.25%	91.51%	74.51%
	40	93.63%	79.29%	89.83%	72.40%	89.44%	71.81%	89.56%	72.11%
IS	1	6.5295	4.9036	6.6289	4.9832	6.7217	5.0442	6.7396	5.0402
	2	6.4300	4.8282	6.6488	4.9898	6.7420	5.0515	6.7646	5.0475
	5	6.3138	4.7379	6.7090	5.0203	6.8110	5.0832	6.8341	5.0778
	10	6.2558	4.6818	6.8050	5.0676	6.9166	5.1325	6.9354	5.1261
	20	6.2412	4.6403	6.9604	5.1411	7.0821	5.2077	7.1008	5.2012
	40	6.1514	4.5262	7.2625	5.2339	7.4034	5.3068	7.4193	5.2978

4.3.3 标量对函数部分线性模型

表 7 总结了 FPLM 的估计和预测性能。与函数线性模型和 FNPM 类似，增加数据块的数量 K 对样本外

预测误差仅有很小影响。对于所有测试样本量，随着 K 增加，RMSFE、RFE 和 MAFE 仅略有上升。相比之下，样本内估计误差的表现与前两个模型明显不同。当 K 从 1 增加到 40 时，三个评价指标（RMSE、RE 和 MAE）均急剧下降，降至原始值的大约一半。例如，RMSE 从 1.3842 降至 0.7094，RE 从 8.27% 降至 3.87%，MAE 从 1.1044 降至 0.5183。样本内拟合得到如此显著的改善，而样本外准确度仅发生轻微变化，这强烈表明当训练样本量较小时存在过拟合。

表7： 不同区块数量和测试样本量下 FPLM 的估计与预测性能。所有结果均为 200 次蒙特卡洛模拟的平均值。

		响应变量							
		训练数据				测试数据			
K	Time	$N_{\text{训练}}$	RMSE	RE	MAE	N_{test}	均方根预测误差	RFE	MAFE
1	185.71秒	2000	1.3842	8.27%	1.1044	200	1.4362	8.57%	1.1449
						400	1.4302	8.56%	1.1431
						800	1.4307	8.55%	1.1420
2	24.95秒	1000	1.3479	8.05%	1.0755	200	1.4363	8.57%	1.1450
						400	1.4305	8.56%	1.1435
						800	1.4307	8.55%	1.1420
5	2.65秒	400	1.2490	7.45%	0.9948	200	1.4368	8.58%	1.1453
						400	1.4308	8.56%	1.1436
						800	1.4310	8.55%	1.1423
10	0.61秒	200	1.1149	6.59%	0.8817	200	1.4375	8.58%	1.1457
						400	1.4315	8.57%	1.1442
						800	1.4318	8.56%	1.1429
20	0.19秒	100	0.9260	5.35%	0.7157	200	1.4389	8.59%	1.1466
						400	1.4328	8.57%	1.1453
						800	1.4332	8.57%	1.1439
40	0.07秒	50	0.7094	3.87%	0.5183	200	1.4430	8.62%	1.1503
						400	1.4367	8.60%	1.1486
						800	1.4370	8.59%	1.1469

与函数线性模型和 FNPM 一样，分布式学习在计算方面具有显著优势。执行时间大幅缩短，从使用单个数据块时的近 200 秒降至使用 40 个数据块时的仅 0.07 秒。这种经验性缩短与公式 (6) 和 (7) 中 FPLM 估计量的理论计算复杂度高度吻合。估计过程包括以下步骤:

1) 对所有 $d(\mathcal{X}_j, \mathcal{X}_i)$ 计算（4）中的半度量，其复杂度为 $i = 1, \ldots, n$，复杂度为 $\mathcal{O}(Mn)$。

2) 对所有 i计算 $w_h(\mathcal{X}_j,\ \mathcal{X}_i)$，复杂度为 $\mathcal{O}(n)$，因此步骤 1) 和 2) 的组合复杂度为 $\mathcal{O}(Mn)$。3) 构建 $n \times n$ 矩阵 $\boldsymbol{W}_h$。由于 $\boldsymbol{W}_h$ 是对称的，唯一距离的数量为 $n\,(n+1)\,2$，每个距离需要 $\mathcal{O}(M)$ 次运算。因此总复杂度为 $\mathcal{O}(Mn^3)$。~ ~4) 计算 $\boldsymbol{Z} = (\boldsymbol{I} - \boldsymbol{W}_h)\boldsymbol{Y}$和 $\boldsymbol{Y}_h = (\boldsymbol{I} - \boldsymbol{W}_h)\boldsymbol{Y}$，每项的复杂度为 $\mathcal{O}(n^2)$。^5) 估计 β_{NF}。由于 $\boldsymbol{Z}$ 是一维的，此步骤的复杂度为 $\mathcal{O}(n)$。6) 对所有函数计算 $\hat{m}(\mathcal{X})$，其复杂度为 $\mathcal{O}(Mn^2)$，与 FNPM 中相同。

主要的计算负担来自第 3) 步，其阶数为 $\mathcal{O}\left(Mn^2\right)$。因此，当 K 加倍时，数据块大小 n 减半，整体执行时间约缩短至原来的八分之一。表7中的经验结果与这一模式高度一致。

表8 报告了不同 K取值下的平均经验覆盖概率和区间得分。随着 K 增大，训练经验覆盖概率略有下降。然而，由于样本内估计误差大幅降低，训练数据的区间得分显著下降，表明预测区间窄了许多。区间宽度由 $\gamma\hat{\sigma}$给出，并相应减小。鉴于所有区块大小下的样本外预测误差都相近，区间变窄会使测试数据的经验覆盖概率大幅降低，因为更多测试观测值落在区间之外。随着 K 增大，经验覆盖概率下降，直接导致测试数据的区间得分降低。

表 8： 不同区块数量下 FPLM 预测区间的平均经验覆盖概率（ECP）和平均区间得分（IS）。所有结果均基于 200 次蒙特卡洛模拟取平均值。

		训练数据		$N_{\text{test}} = 200$		$N_{\text{test}} = 400$		$N_{\text{test}} = 800$	
判据	K	$\alpha = 0.05$	$\alpha = 0.2$	$\alpha = 0.05$	$\alpha = 0.2$	$\alpha = 0.05$	$\alpha = 0.2$	$\alpha = 0.05$	$\alpha = 0.2$
ECP	1	95.00%	80.00%	93.97%	78.33%	94.10%	78.40%	94.16%	78.43%
	2	94.94%	79.98%	93.26%	77.03%	93.40%	77.15%	93.45%	77.18%
	5	94.87%	79.92%	91.06%	73.28%	91.04%	73.35%	91.15%	73.45%
	10	94.71%	79.84%	86.91%	67.49%	87.15%	67.61%	87.15%	67.70%
	20	94.47%	79.70%	79.52%	57.93%	79.64%	58.05%	79.73%	58.02%
	40	94.03%	79.61%	67.25%	44.96%	67.42%	44.77%	67.52%	44.93%
IS	1	6.4686	4.8576	6.7802	5.0580	6.6952	5.0289	6.7099	5.0284
	2	6.3033	4.7306	6.8213	5.0690	6.7316	5.0406	6.7438	5.0391
	5	5.8656	4.3890	7.0411	5.1357	6.9527	5.1038	6.9530	5.1017
	10	5.3110	3.9386	7.6674	5.3106	7.5641	5.2763	7.5631	5.2729
	20	4.5729	3.3209	9.3085	5.7605	9.1775	5.7211	9.1830	5.7188
	40	3.7557	2.6144	12.8454	6.6344	12.7013	6.5964	12.696	6.5911

5 *Tecator* 数据分析

我们考虑一项食品质量控制应用，该应用由 Ferraty 和 Vieu（2006）以及 Aneiros-P´erez 和 Vieu（2006）研究。数据集来自 https://lib.stat.cmu.edu/datasets/tecator。对于每个样本单位 i （共 215 块切碎的肉），我们观测到一条光谱曲线，记为 $\mathcal{X}_i$，它对应于在 100 个波长网格上测得的吸光度（即，$\mathcal{X}_i = (\mathcal{X}_i(t_1) \ldots, \mathcal{X}_i(t_{100}))$）。对于每个样本单位 i，我们还观测到通过分析化学处理获得的脂肪/蛋白质/水分含量 $Y_i \in \mathbb{R}$ 。数据集包含这些配对观测值 $(\mathcal{X}_i, Y_i)_{i=1}^{215}$。给定一条新的光谱曲线 $\mathcal{X}$，我们的任务是预测相应的脂肪/蛋白质/水分含量。正如 Ferraty 和 Vieu（2006）所指出的，研究动机在于，获取光谱曲线所需的时间少于通过分析化学测定脂肪/蛋白质/水分含量所需的时间。图 1 展示了光谱曲线的图形。

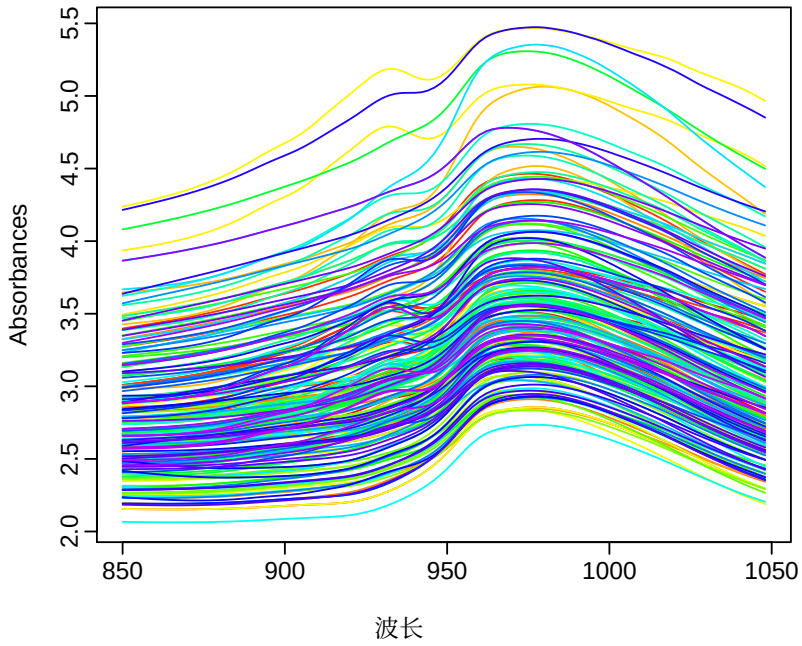

图 1：光谱曲线的图形展示。

我们分别研究光谱曲线与其对应的脂肪、蛋白质或水分含量之间的关系。我们采用非参数函数型 Nadaraya-Watson 估计量。为评估非参数函数型估计量的样本外预测准确度，我们将原始样本分为三个子样本。第一个称为训练样本，包含前 129 个个体 $\{(\mathcal{X}_i, y_i)_{i=1,\ldots,129}\}$。第二个称为验证样本，包含不重叠的 43 个个体 $\{(\mathcal{X}_i, y_i)_{i=130,\ldots,172}$。

训练样本使我们能够构建函数型 NW 估计量。为衡量预测准确度，我们使用验证样本评估函数型 NW 估计量，并据此预测响应变量 $(y_{130},\ldots,y_{172})$。我们计算残差 $\epsilon_\omega = y_\omega - \widehat{y}_\omega$, $\omega = 1,\ldots,43$。

在本节中，我们将函数线性模型、FNPM 和 FPLM 的分布式估计程序应用于*Tecator*数据集。完整数据集包含 215 个观测值。我们随机选取 150 个观测值作为训练集，并将剩余的 65 个观测值作为测试集。每个模型均使用训练数据进行拟合，并依据第4.2节所述的评价指标，使用测试数据进行评估。为获得稳定的结果，该抽样过程重复 200 次，所有报告的统计量均为这 200 个样本的平均值。

表9 报告了所有模型的估计和预测性能。对于 FPLM，尽管可以纳入两个非函数型协变量，但本分析仅考虑一个非函数型协变量。数据块数量逐步增加，最多增至五个，使每个数据块的样本量从 150 降至 30。在这三个模型中，增加数据块数量均能显著缩短执行时间，同时预测误差基本保持稳定。

表 9： 区块数量不同时所有模型的估计和预测性能。所有结果均为 200 次蒙特卡罗重复试验的平均值。

				训练数据			测试数据			
模型	K	时间$N_{\text{训练}}$		RMSE	RE	平均绝对误差	N_{test}	均方根预测误差	相对预测误差	平均绝对预测误差
响应变量：脂肪，非函数型协变量：水分										
函数线性模型（B样条）	1	0.0074秒	150	4.7541	50.13%	3.8388	65	5.0195	52.26%	4.0497
	2	0.0045秒	75	4.6323	48.40%	3.7307	65	5.0337	52.16%	4.0605
	3	0.0040秒	50	4.4984	46.62%	3.6181	65	5.0519	52.00%	4.0739
	5	0.0038s	30	4.2573	43.51%	3.4140	65	5.0983	52.03%	4.1015
函数线性模型（函数主成分分析）	1	0.1155s	150	3.2881	28.44%	2.5964	65	3.5582	30.25%	2.7960
	2	0.0393s	75	3.1820	27.43%	2.5194	65	3.5533	30.06%	2.7771
	3	0.0268s	50	3.0709	26.38%	2.4324	65	3.5571	29.97%	2.7671
	5	0.0195秒	30	2.8676	24.62%	2.2802	65	3.5701	29.87%	2.7551
FNPM	1	0.2918秒	150	5.5555	44.92%	4.3421	65	8.9223	65.56%	6.9614
	2	0.0880s	75	5.2692	40.55%	3.9620	65	9.2352	70.77%	7.3446
	3	0.0536s	50	5.9602	43.88%	4.3509	65	9.7325	76.19%	7.8627
	5	0.0363s	30	7.0819	51.06%	5.1269	65	10.3648	82.76%	8.4778
FPLM	1	0.3061s	150	1.1140	7.97%	0.7505	65	1.8035	12.89%	1.2856
	2	0.0978s	75	0.9102	6.28%	0.5927	65	1.8652	13.67%	1.3570
	3	0.0638s	50	0.7613	5.08%	0.4858	65	1.9107	14.06%	1.4010

			训练数据			测试数据				
模型	K	时间$N_{\text{训练}}$		RMSE	RE	平均绝对误差	N_{test}	均方根预测误差	相对预测误差	平均绝对预测误差
	5	0.0378秒	30	0.5904	3.72%	0.3590	65	2.0126	14.56%	1.4623
响应变量：水分，非函数型协变量：蛋白质										
函数线性模型（B样条）	1	0.0102秒	150	3.6828	4.81%	2.9670	65	3.8718	5.06%	3.1136
	2	0.0047秒	75	3.5916	4.68%	2.8848	65	3.8838	5.08%	3.1191
	3	0.0040秒	50	3.4917	4.54%	2.7999	65	3.9008	5.10%	3.1269
	5	0.0037秒	30	3.3063	4.29%	2.6432	65	3.9371	5.14%	3.1438
函数线性模型（函数主成分分析）	1	0.1192s	150	2.7573	3.79%	2.2249	65	2.9339	4.00%	2.3368
	2	0.0421s	75	2.6815	3.67%	2.1584	65	2.9413	3.99%	2.3247
	3	0.0270秒	50	2.6012	3.56%	2.0929	65	2.9551	3.98%	2.3170
	5	0.0187秒	30	2.4436	3.32%	1.9620	65	2.9783	3.97%	2.3016
FNPM	1	0.2644s	150	4.2591	5.70%	3.3938	65	6.7548	9.27%	5.3897
	2	0.0871s	75	3.9595	5.08%	3.0270	65	6.9662	9.78%	5.6628
	3	0.0576s	50	4.4698	5.61%	3.3168	65	7.3446	10.53%	6.0678
	5	0.0365s	30	5.3939	6.76%	3.9570	65	7.8667	11.46%	6.5735
FPLM	1	0.3563秒	150	2.2608	2.61%	1.6205	65	4.3559	5.11%	3.0819
	2	0.1001秒	75	1.8124	2.02%	1.2610	65	4.4720	5.31%	3.1927
	3	0.0598秒	50	1.5319	1.65%	1.0337	65	4.5291	5.41%	3.2452
	5	0.0352秒	30	1.1849	1.21%	0.7585	65	4.6300	5.54%	3.3245
响应变量：蛋白质，非函数型协变量：脂肪										
函数线性模型（B样条）	1	0.0078s	150	1.5895	7.27%	1.2519	65	1.6474	7.53%	1.2918
	2	0.0049秒	75	1.5464	7.10%	1.2245	65	1.6516	7.54%	1.2921
	3	0.0046秒	50	1.5014	6.91%	1.1930	65	1.6549	7.57%	1.2942
	5	0.0036秒	30	1.4203	6.55%	1.1325	65	1.6676	7.62%	1.3007
函数线性模型（函数主成分分析）	1	0.0679秒	150	1.3717	6.47%	1.0538	65	1.4506	6.83%	1.1088
	2	0.0575秒	75	1.3203	6.27%	1.0257	65	1.4374	6.76%	1.0968
	3	0.0321秒	50	1.2655	6.02%	0.9903	65	1.4272	6.72%	1.0893
	5	0.0291秒	30	1.1797	5.61%	0.9307	65	1.4200	6.68%	1.0810
FNPM	1	0.2375秒	150	1.5264	7.37%	1.2227	65	2.3202	11.66%	1.8972
	2	0.0817秒	75	1.4287	6.67%	1.1082	65	2.4228	12.43%	2.0100
	3	0.0605秒	50	1.6039	7.40%	1.2215	65	2.5539	13.29%	2.1386
	5	0.0378秒	30	1.8793	8.71%	1.4281	65	2.7117	14.31%	2.2919
FPLM	1	0.3257秒	150	0.7640	2.97%	0.5051	65	1.3525	5.57%	0.9089
	2	0.0818秒	75	0.6228	2.34%	0.3985	65	1.3894	5.85%	0.9503
	3	0.0467秒	50	0.5315	1.94%	0.3298	65	1.4045	5.98%	0.9670
	5	0.0367s	30	0.4195	1.44%	0.2455	65	1.4523	6.25%	1.0078

在这三个模型中，FPLM 在训练数据和测试数据上的误差均最低。随着 K 从 1 增至 5，执行时间从 0.3061 秒大幅缩短至 0.0378 秒，而预测误差仅略有增加。RMSFE 从 1.8035 增至 2.0126，RFE 从 12.89% 增至 14.56%，MAFE 从 1.2856 增至 1.4623。然而，当 $K = 5$时，训练误差与测试误差之间的差距变大。例如，训练数据的 RMSE 为 0.5904，而测试数据的 RMSFE 为 2.0126。这一显著差距表明，区块大小过小时会出现过拟合。

对于 FLM，FPCA 方法的预测误差低于 B 样条展开法，但需要更长的执行时间。随着 K 从 1 增至 5，执行时间从 0.1155 秒缩短至 0.0195 秒，而预测误差仅小幅增加。例如，RMSFE 从 3.5582 增至 3.5701。

最后，FNPM 的执行时间与 FPLM 相当，但 FNPM 产生了最大的估计误差和预测误差。样本内误差和样本外误差均显著高于 FLM 和 FPLM。

表 10报告了所有模型的平均经验覆盖概率。随着 K 从 1 增加到 5，训练数据的平均经验覆盖概率仅略有变化。对于 FLM 和 FNPM，不同 K取值下测试数据的平均经验覆盖概率变化小于 10%。相比之下，随着数据块数量增加，FPLM 的测试经验覆盖概率大幅下降。例如，在显著性水平为 $\alpha = 0.05$ 时，测试经验覆盖概率从一个数据块时的 86.32% 降至五个数据块时的 55.14%。与模拟研究结果一致，增加 K 会导致 FPLM 对训练数据过拟合，从而产生窄得多的预测区间。因此，落在这些区间内的测试观测值比例更低，导致测试经验覆盖概率下降。

表10：不同数据块数量下所有模型的预测区间平均经验覆盖概率（ECP）。所有结果均取200次蒙特卡罗重抽样的平均值。

		训练数据		测试数据	
判据	K	$\alpha = 0.05$	$\alpha = 0.2$	$\alpha = 0.05$	$\alpha = 0.2$
响应变量：脂肪，非函数型协变量：水分					
函数线性模型（B样条）	1	94.72%	80.10%	93.25%	77.41%
	2	94.36%	79.58%	91.98%	75.22%
	3	93.93%	79.12%	90.67%	73.58%
	5	93.19%	78.90%	87.20%	70.57%

		训练数据		测试数据	
判据	K	$\alpha = 0.05$	$\alpha = 0.2$	$\alpha = 0.05$	$\alpha = 0.2$
函数线性模型（函数主成分分析）	1	94.68%	80.03%	92.84%	76.69%
	2	94.14%	79.34%	91.68%	75.76%
	3	93.73%	79.18%	90.71%	74.79%
	5	93.00%	78.69%	88.53%	72.09%
FNPM	1	94.68%	80.02%	79.38%	60.28%
	2	94.09%	79.19%	74.05%	52.02%
	3	93.05%	79.25%	74.31%	52.56%
	5	91.86%	78.35%	77.22%	54.68%
FPLM	1	94.85%	80.06%	86.32%	59.35%
	2	94.01%	79.66%	77.76%	48.96%
	3	93.62%	79.97%	67.67%	42.37%
	5	93.38%	80.47%	55.14%	32.12%
响应变量：水分，非函数型协变量：蛋白质					
函数线性模型（B样条）	1	94.72%	80.06%	93.05%	77.95%
	2	94.21%	79.64%	91.64%	76.43%
	3	93.93%	79.23%	90.13%	75.18%
	5	93.34%	79.11%	87.58%	71.54%
函数线性模型（函数主成分分析）	1	94.73%	80.05%	92.82%	77.45%
	2	94.19%	79.49%	91.50%	76.51%
	3	93.78%	79.12%	90.18%	75.33%
	5	92.88%	78.32%	87.93%	72.88%
FNPM	1	94.71%	80.05%	78.60%	59.18%
	2	94.18%	79.26%	72.92%	50.24%
	3	93.20%	78.94%	73.15%	49.84%
	5	91.94%	77.78%	76.70%	53.42%
FPLM	1	94.67%	80.04%	77.46%	55.53%
	2	94.31%	79.79%	67.53%	45.74%
	3	93.89%	80.00%	59.63%	39.16%
	5	93.30%	79.79%	47.78%	29.13%

		训练数据		测试数据	
判据	K	$\alpha = 0.05$	$\alpha = 0.2$	$\alpha = 0.05$	$\alpha = 0.2$
响应变量：蛋白质，非函数型协变量：脂肪					
函数线性模型（B样条）	1	94.76%	80.05%	93.42%	78.35%
	2	94.28%	79.61%	92.35%	77.22%
	3	94.11%	79.65%	91.40%	75.78%
	5	93.11%	79.42%	88.32%	72.93%
函数线性模型（函数主成分分析）	1	94.71%	80.04%	93.41%	78.22%
	2	94.50%	79.83%	93.04%	77.31%
	3	94.18%	79.62%	92.08%	76.15%
	5	93.48%	79.31%	89.90%	73.24%
FNPM	1	94.74%	80.06%	80.40%	59.63%
	2	94.14%	79.24%	73.70%	50.24%
	3	93.04%	78.60%	74.40%	50.57%
	5	91.88%	77.09%	77.66%	53.49%
FPLM	1	94.72%	80.03%	84.70%	60.21%
	2	94.46%	79.90%	75.50%	49.41%
	3	94.27%	80.26%	68.58%	42.18%
	5	94.04%	80.67%	56.59%	31.94%

表 11 展示了所有模型的平均区间评分。与表 10中的结果一致，随着 K 增加，函数线性模型和 FNPM 的区间评分仅略有变化。具体而言，FNPM 的区间评分有所上升，而函数线性模型的区间评分有所下降。这种现象反映了潜在的估计误差。对于 FNPM，随着K增大，预测误差略有增加，导致预测区间变宽，区间评分值升高。然而，对于 FPLM，当 K 增加至五时，区间评分发生了显著变化。例如，当 $\alpha = 0.05$时，训练数据的区间评分从 6.5626 急剧下降至 3.7501，而测试数据的区间评分则从 12.6356 上升至 27.5887。这种模式再次表明，当区块大小过小时会出现过拟合。当 $K = 5$时，训练数据的估计误差非常小，因此产生较窄的预测区间，区间评分也较低。然而，测试数据对应的经验覆盖概率也较低，这意味着许多测试观测值落在这些过窄的区间之外。因此，测试数据的区间评分会大幅升高。

表11：不同数据块数量下所有模型预测区间的平均区间得分（IS）。所有结果均为 200 次蒙特卡罗重复试验的平均值。

		训练数据		测试数据	
判据	K	$\alpha = 0.05$	$\alpha = 0.2$	$\alpha = 0.05$	$\alpha = 0.2$
响应变量：脂肪，非函数型协变量：水分					
函数线性模型（B样条）	1	22.0441	16.4233	24.4219	17.6411
	2	21.4951	16.0297	24.7398	17.8333
	3	20.8970	15.6031	25.4121	18.0406
	5	19.9701	14.8478	27.4773	18.5540
函数线性模型（函数主成分分析）	1	16.3083	11.4966	18.9812	12.6204
	2	15.7430	11.1535	19.6721	12.6725
	3	15.0901	10.7836	20.4478	12.7205
	5	13.9522	10.0791	22.1615	12.9191
FNPM	1	26.6415	19.6655	63.3199	35.6100
	2	26.9487	19.0900	72.6947	39.5891
	3	32.3489	21.8528	73.7530	41.3284
	5	39.3029	26.1591	70.5169	42.6668
FPLM	1	6.5626	4.2027	12.6356	7.6570
	2	5.5417	3.4592	16.4605	8.6691
	3	4.6988	2.9016	20.8377	9.4773
	5	3.7501	2.2545	27.5887	10.7842
响应变量：水分，非函数型协变量：蛋白质					
函数线性模型（B样条）	1	16.6094	12.8201	18.5330	13.6903
	2	16.3121	12.5255	18.9859	13.7997
	3	15.9286	12.2026	19.6644	13.9612
	5	15.2902	11.6020	21.3095	14.3201
函数线性模型（函数主成分分析）	1	12.9184	9.5830	14.6998	10.3955
	2	12.6243	9.3573	15.4068	10.4891
	3	12.2062	9.0899	16.2291	10.6094
	5	11.5409	8.5659	18.0676	10.8925
FNPM	1	19.5157	14.8766	45.4947	26.4862
	2	19.4769	14.2343	52.7449	29.7252
	3	23.5492	16.2415	53.8459	31.0859
	5	29.2926	19.7284	51.4558	31.9695
FPLM	1	12.7646	8.1292	37.4739	18.1500
	2	10.3283	6.6241	47.4894	20.5748
	3	8.8720	5.6957	55.1092	22.1854
	5	7.1097	4.4895	68.5854	24.7925

		训练数据		测试数据	
判据	K	$\alpha = 0.05$	$\alpha = 0.2$	$\alpha = 0.05$	$\alpha = 0.2$
响应变量：蛋白质，非函数型协变量：脂肪					
函数线性模型（B样条）	1	7.4720	5.6147	8.0816	5.9077
	2	7.2149	5.4572	8.2794	5.9333
	3	6.9753	5.2937	8.4376	5.9593
	5	6.6466	5.0082	9.0431	6.0563
函数线性模型（函数主成分分析）	1	7.1004	4.8131	8.1230	5.1821
	2	6.6492	4.6053	8.0438	5.1286
	3	6.2708	4.4156	8.0646	5.0833
	5	5.7923	4.1331	8.2826	5.0933
FNPM	1	6.9003	5.3340	14.3843	8.9889
	2	6.8773	5.0921	17.5235	10.1068
	3	8.0607	5.7839	17.7427	10.5457
	5	9.5969	6.7964	17.0190	10.8129
FPLM	1	4.7066	2.8086	10.6579	5.4566
	2	3.8907	2.2947	13.1501	6.1707
	3	3.3728	1.9688	15.0646	6.6155
	5	2.7304	1.5531	19.0055	7.4875

6 结论与未来展望

本文提出了一个统一的分布式框架，用于对三种广泛应用的标量对函数回归模型进行点估计和区间估计，分别是函数型线性模型（FLM）、函数型非参数模型（FNPM）和函数型部分线性模型（FPLM）。该方法允许在分布式数据块上分别对各模型进行本地拟合，并仅将汇总后的中间结果传输至中央服务器。这一设计既保护了数据隐私，又显著提高了计算效率，因此该框架适用于大规模函数型数据和多机构场景。

模拟结果表明，分布式估计的表现非常出色。对于所有三种模型，数据划分在大幅缩短执行时间的同时，仅会轻微降低估计或预测准确度。然而，FPLM 对样本量更为敏感。当区块过小时，它往往会过拟合训练数据，导致点估计误差较大、预测区间过窄，且对新观测值的经验覆盖率下降。使用$tecator$ 数据集进行的实证分析进一步印证了这些发现。在所有模型中，随着区块数量

增加，执行时间会显著缩短，而预测误差总体保持稳定。在所考察的模型中，FPLM 的整体预测准确度最高，但它对过度分块也最为敏感。

本研究可从多个方向扩展，这里简要提及其中两个。首先，我们只考虑标量对函数回归模型，但也可以将其扩展到函数对标量回归和函数对函数回归。其次，我们发现理论计算时间与实证结果之间存在差异。或许有人会对解决这一差异感兴趣。

致谢

作者感谢麦考瑞大学 Data Horizon 项目提供的资助。

CRediT 作者贡献

Peilun He： 方法学、软件、验证、正式分析、调查研究、撰写初稿、撰写——审阅与编辑。

Nan Zou： 概念化、方法学、正式分析、调查研究、撰写初稿、撰写——审阅与编辑、指导。

Han Lin Shang: 概念构思、方法设计、软件、正式分析、调查研究、撰写初稿、审阅与编辑、可视化、监督、经费获取。

利益冲突声明

作者声明不存在需要披露的利益冲突。

数据可用性

蒙特卡洛模拟和实证分析的代码可在 GitHub 上获取，地址：
https://github.com/peilun-he/Distributed-Functional-Regression。

参考文献

Aneiros-P´erez, G. 和 Vieu, P. (2006). 半函数型偏线性回归。《统计与概率快报》，76(11):1102–1110. Beyaztas, U. 和 Shang, H. L. (2022). 一种用于标量对多函数回归的稳健函数型偏最小二乘。《化学计量学杂志》，36(4):e3394. Beyaztas, U., Tez, M. 和 Shang, H. L. (2024). 稳健标量对函数偏分位数回归。《应用统计学杂志》，51(7):1359–1377. Cai, T., Chakraborty, A. 和 Vuursteen, L. (2024). 异质隐私约束下函数均值估计的最优联邦学习。技术报告，arXiv。网址：https://arxiv.org/abs/2412.18992。C hang, X., Lin, S.-B. 和 Wang, Y. (2017). 分而治之局部平均回归。《电子统计学杂志》，11:1326–1350. Ferraty, F., Laksaci, A. 和 Vieu, P. (2005). 通过条件众数估计进行函数型时间序列预测。《数学报告》，340(5):389–392. Ferraty, F., Mas, A. 和 Vieu, P. (2007). 函数型数据的非参数回归：推断与实践方面。《澳大利亚与新西兰统计学杂志》，49(3):267– 286. Ferraty, F. 和 Vieu, P. (2006). 《非参数函数型数据分析：理论与实践》。施普林格，纽约。Gneiting, T. 和 Raftery, A. E. (2007). 严格适当评分规则、预测与估计。《美国统计学会杂志：评论文章》，102(477):359–378. Goutis, C. (1998). 应用于近红外光谱的二阶导数函数型回归。《皇家统计学会杂志 *B* 辑》，60(1):103–114. Koner, S. 和 Staicu, A.-M. (2023). 第二代函数型数据。《统计学及其应用年度评论》，10:547–572. Li, R., Lin, D. K. 和 Li, B. (2013). 海量数据集的统计推断。《商业与工业应用随机模型》，29(5):399–409. Liu, J., Li, R. 和 Lian, H. (2024). 函数型响应函数型线性回归的分布式估计。《*Metrika*》，87(1):21–30.

Liu, J. and Shi, L. (2024). 分而治之的基于核的函数型线性回归的统计最优性. 机器学习研究期刊, 25(155):1–56. Liu, J. and Shi, L. (2025). 基于离散观测函数型数据的分布式学习. 《反问题》, 41(4):045006. Masry, E. (2005). 相关函数型数据的非参数回归估计：渐近正态性. 随机过程及其应用, 115(1):155–177. Mcdonald, R., Mohri, M., Silberman, N., Walker, D., and Mann, G. (2009). 条件最大熵模型的高效大规模分布式训练. 神经信息处理系统进展, 22. Morris, J. S. (2015). 函数型回归. 《统计学及其应用年度评论》, 2:321–359. Novo, S., Aneiros, G., and Vieu, P. (2021a). 半参数函数型数据分析中的K近邻程序. 《统计与概率快报》, 171:109028. Novo, S., Aneiros, G., and Vieu, P. (2021b). 当预测变量由函数型变量和高维变量混合构成时的稀疏半参数回归. *Test*, 30:481–504. Ramsay, J. O. and Silverman, B. W. (2005). 函数型数据分析. 施普林格, 纽约. Ratcliffe, S. J., Heller, G. Z., and Leader, L. R. (2002a). 函数型数据分析及其在周期性刺激胎儿心率数据中的应用。II：函数型逻辑回归.医学统计学, 21(8):1115–1127. Ratcliffe, S. J., Leader, L. R., and Heller, G. Z. (2002b). 函数型数据分析及其在周期性刺激胎儿心率数据中的应用：I. 函数型回归. 医学统计学, 21(8):1103–1114. Reiss, P. T., Goldsmith, J., Shang, H. L., and Ogden, R. T. (2017). 标量对函数回归方法. 国际统计评论, 85(2):228–249. Reiss, P. T. and Ogden, R. T. (2010). 以图像为预测变量的函数型广义线性模型. 生物统计学, 66(1):61–69. Shang, H. L. and Haberman, S. (2025). 构造死亡年龄分布的预测区间. 斯堪的纳维亚精算期刊, pages 1–18.

Tong, H. (2021). 函数型线性回归的分布式最小二乘预测. 《反问题》, 38(2):025002. Xiao, P., Liu, X., Li, A., and Pan, G. (2024). 基于随机加权自助法的分位数回归模型分布式推断. 《信息科学》, 680:121172. Xue, G., Lin, Z., and Yu, Y. (2024). 隐私保护分布式函数型数据分析中的最优估计. 技术报告, arXiv。网址: https://arxiv.org/abs/2412.06582. Zhang, Y., Duchi, J., and Wainwright, M. (2015). 分治核岭回归：具有极小极大最优速率的分布式算法. 《机器学习研究杂志》, 16(1):3299–3340. Zhou, L., Gong, Z., and Xiang, P. (2024). 大数据的分布式计算与推断。《统计学及其应用年度评论》,11:533–551.