arXiv:2301.01076v1 [math.ST] 3 Jan 2023

函数型线性乘法模型的最小乘积相对误差估计与最优子抽样

Qian Yan，Hanyu Li*

重庆大学

Abstract: 本文基于最小乘积相对误差准则研究函数型线性乘法模型。在一定的正则化条件下，我们建立了估计量的一致性和渐近正态性。此外，我们研究了该模型在海量数据下的最优子抽样问题。首先推导了子抽样估计量的一致性和渐近分布。随后，基于 A-最优性准则得到了最优子抽样概率。此外，我们还提出了无需计算海森矩阵逆矩阵、且更易于实际应用的有效替代子抽样概率。最后，我们通过数值研究和真实数据分析来评估该方法的性能

*通讯作者：Hanyu Li，重庆大学数学与统计学院，中国重庆 401331。电子邮箱：lihy.hy@gmail.com 或 hyli@cqu.edu.cn。

所提出的方法。

Key words and phrases: 渐近正态性、函数型乘法模型、最小乘积相对误差、海量数据、最优子抽样

1. 引言

在大数据时代，可以对时间和空间中密集采样的观测进行收集和记录。这些观测具有函数型特征，通常表现为曲线和图像。研究表明，函数型数据分析在处理这类数据集时表现出色。标量响应的函数型回归模型已得到广泛研究，其中最受欢迎的是函数型线性模型。

我们考虑一个标量对函数线性乘法模型

$$y = \exp\left(\int_0^1 X(t)\beta(t)\mathrm{d}t\right)\epsilon, \tag{1.1}$$

其中，协变量 $X(t)$ 和斜率 $\beta(t)$ 是定义在 $[0,1]$ 上的光滑且平方可积函数， y 是标量响应变量， ϵ 是随机误差。此外， y 和 ϵ 均严格为正。通过对数变换，模型（1.1）可转化为常规的函数型线性模型。不过，相比之下，乘法模型在处理收入、股票价格和生存时间等正响应时更实用、更灵活。

众所周知，在估计斜率时，绝对误差是设计损失函数时最常用的选择，例如最小二乘法（LS）和最小绝对偏差（LAD）。然而，在实际应用中，基于相对误差的损失函数可能更有效、更适用。相对误差有两种类型：相对于目标值 y 的误差，以及相对于预测值 y 的误差。Chen 等人（2010）将这两种相对误差相加，为标量线性乘法模型提出了最小绝对相对误差（LARE）准则。然而，LARE 准则并非光滑函数，因此计算起来有些复杂。后来，Chen 等人（2016）将这两种相对误差相乘，对该方法进行了改进，并提出了最小乘积相对误差（LPRE）准则。LPRE 准则具有无穷阶连续可微性且严格凸，因此能得到简单且唯一的估计量。此外，他们还证明，在某些条件下，LPRE估计比 LARE、LAD 和 LS 估计更有效。因此，该准则也已广泛应用于其他标量乘法模型（Chen 和 Liu（2021）；Chen、Liu 和 Ma（2022）；Ming、Liu 和 Yang（2022））。

据我们所知，关于函数型乘法模型的研究寥寥无几，且都聚焦于 LARE 准则。例如，Zhang、Zhang 和 Li（2016）首次将 LARE 准则拓展到

函数型模型。他们提出了函数二次乘法模型，并推导出估计量的渐近性质。随后，Zhang 等（2019）和 Fan、Zhang 和 Wu（2022）分别基于 LARE 准则，研究了部分稀疏和局部稀疏函数型线性乘法模型的变量选择。似乎尚无研究探讨针对函数型数据的 LPRE 准则。为填补这一空白，我们为函数型线性乘法模型提出 LPRE 准则，并推导出估计量的一致性和渐近正态性。

考虑到计算资源有限，传统技术已无法用于海量数据，因此一些研究人员致力于为海量数据的统计模型开发高效或最优的子抽样策略。例如，对于线性模型，Ma、Mahoney 和 Yu（2015）研究了算法杠杆估计量的偏差和方差。Ma 等（2020）进一步给出了 RandNLA 子抽样* 估计量的渐近分布。对于逻辑回归，Wang、Zhu 和 Ma（2018）提出了一种基于若干最优性准则（Atkinson、Donev 和 Tobias（2007））的最优子抽样方法。随后，Wang（2019）

*这类抽样的概率与杠杆值密切相关，

这些准则通常用于设计数值线性代数中的随机化算法。

提出了一种更高效的估计方法和泊松子抽样，以提高估计和计算效率。随后，Yao 和 Wang（2019）、Yu、Wang 和 Ai（2020）以及 Ai 等（2021b）分别将最优子抽样方法扩展至 Softmax 回归、准似然和广义线性模型。此外，考虑到重尾误差或响应变量中异常值的影响，一些学者研究了更稳健的模型。例如，Wang 和 Ma（2021）、Ai 等（2021a）、Fan、Liu 和 Zhu（2021）以及 Shao、Song 和 Zhou（2022）将最优子抽样方法用于普通分位数回归；Shao 和 Wang（2021）以及 Yuan 等（2022）则将子抽样方法扩展至复合分位数回归。最近，Ren、Zhao 和 Wang （ 2022 ）研究了基于 LARE 准则的线性乘法模型最优子抽样策略。他们推导了子抽样估计量的渐近分布，并证明在最优子抽样策略下，LARE 优于最小二乘法和 LAD。Wang 和 Zhang（2022）进一步将基于 LPRE 准则的最优子抽样扩展至线性乘法模型。

对于函数型回归模型，目前关于子抽样的研究还很少（Liu、You 和 Cao （ 2021 ）； He 和 Yan （ 2022 ）； Yan 、 Li 和 Niu （ 2022 ））。具体而言，He 和 Yan（2022）提出了一种函数型主子空间抽样概率，用于标量响应的函数型线性回归，

该方法消除了函数型主子空间内特征值的影响，并对残差进行了适当加权。Liu、You 和 Cao（2021）以及 Yan、Li 和 Niu（2022）分别将最优子抽样方法扩展到了函数型广义线性模型和标量响应的函数型分位数回归。

受上述研究启发，我们进一步研究基于 LPRE 准则的函数型线性乘法模型的最优子抽样，并首先证明子抽样估计量的一致性和渐近正态性。随后，在 A-最优性准则下，通过最小化渐近积分均方误差（IMSE）获得最优子抽样概率。此外，我们还提出了一种实用的替代最小化准则，以进一步降低计算成本。

本文其余部分安排如下。第 2 节介绍基于 LPRE 准则的函数型线性乘法模型，并研究估计量的渐近性质。第 3 节给出子抽样估计量的渐近性质及最优子抽样概率。本节还讨论这些概率的修正版。第 4 节和第 5 节分别通过数值模拟和真实数据展示我们的方法。

2. LPRE估计

2.1 估计

假设 $\{(x_i(t)\ y_i)\ i = 1、2、\ldots、\ n\}$ 是来自模型 (1.1) 的独立同分布样本。模型 (1.1) 的函数型 LPRE ˆ 估计量，即 $\beta(t)$，定义为

$$\arg\inf_{\beta}\sum_{i=1}^{n}\left\{\left|\frac{y_i-\exp\left(\int_0^1 x_i(t)\beta(t)\mathrm{d}t\right)}{y_i}\right|\ \times\ \left|\frac{y_i-\exp\left(\int_0^1 x_i(t)\beta(t)\mathrm{d}t\right)}{\exp\left(\int_0^1 x_i(t)\beta(t)\mathrm{d}t\right)}\right|\right\},$$

这等价于

$$\arg\inf_{\beta}\sum_{i=1}^{n}\left\{y_i\exp\left(-\int_0^1 x_i(t)\beta(t)\mathrm{d}t\right)+y_i^{-1}\exp\left(\int_0^1 x_i(t)\beta(t)\mathrm{d}t\right)-2\right\}.$$

我们旨在通过惩罚样条法估计斜率函数 $\beta(t)$。定义 K 个等距内部节点为 $0 = t_0 < t_1 < \ldots <t_K < t_{K+1} = 1$。令 $\boldsymbol{B}(t) = (B_1(t)\ B_2(t)\ \ldots, B_{K+p+1}(t))^T$ 为次数为 p 的归一化 B 样条基函数集合；这些基函数在每个子区间$[t_j, t_{j+1}], j = 0, 1, \ldots, K$上定义，并且在 $[0, 1]$ 上具有 $p-1$阶连续可微性。关于 B 样条函数的详细信息，参见 de Boor（2001）。因此，我们的 $\beta(t)$ 对 $\beta(t)$ 的函数型 LPRE 估计量定义如下：

$$\hat{\beta}(t) = \sum_{j=1}^{K+p+1}\hat{\boldsymbol{\theta}}_j B_j(t) = \boldsymbol{B}^T(t)\hat{\boldsymbol{\theta}}_{full},$$

其中 $\boldsymbol{\theta}_{full}$ 使惩罚函数型 LPRE 损失函数最小化

$$
\begin{aligned}
&L(\boldsymbol{\theta};\lambda,K)\\
=&\sum_{i=1}^{n}\left\{y_i\exp\left(-\int_0^1 x_i(t)\boldsymbol{B}^T(t)\boldsymbol{\theta}\mathrm{d}t\right)+y_i^{-1}\exp\left(\int_0^1 x_i(t)\boldsymbol{B}^T(t)\boldsymbol{\theta}\mathrm{d}t\right)-2\right\}\\
&+\frac{\lambda}{2}\int_0^1\left\{\left(\boldsymbol{B}^{(q)}(t)\right)^T\boldsymbol{\theta}\right\}^2\mathrm{d}t, \qquad (2.2)
\end{aligned}
$$

其中，$\lambda>0$ 是平滑参数，$\boldsymbol{B}^{(q)}(t)$是所有 B 样条函数的 q阶导数平方积分，其中 $q\leq p$为某个整数。为方便起见，令$\boldsymbol{B}_i=\int_0^1 x_i(t)\boldsymbol{B}(t)\mathrm{d}t$ 以及 $\boldsymbol{D}_q=\int_0^1\boldsymbol{B}^{(q)}(t)\{\boldsymbol{B}^{(q)}(t)\}^T\mathrm{d}t$，于是损失函数 (2.2) 可重写为

$$
L(\boldsymbol{\theta};\lambda,K)=\sum_{i=1}^{n}\left\{\omega_i(\boldsymbol{\theta})+\omega_i(\boldsymbol{\theta})^{-1}-2\right\}+\frac{\lambda}{2}\boldsymbol{\theta}^T\boldsymbol{D}_q\boldsymbol{\theta}, \qquad (2.3)
$$

其中 $\omega_i(\boldsymbol{\theta})=y_i\exp\left(-\boldsymbol{B}_i^T\boldsymbol{\theta}\right)$。值得注意的是，模型 (2.3) 无穷次可微且严格凸。由于函数 LPRE 估计量不存在通用的闭式解，因此将采用牛顿-拉夫森方法。也就是说，估计量 $\boldsymbol{\theta}_{full}$ 可通过迭代应用以下公式获得，直至 $\boldsymbol{\theta}_{t+1}$ 收敛。

$$
\begin{aligned}
\hat{\boldsymbol{\theta}}_{t+1}=&\hat{\boldsymbol{\theta}}_t-\left\{\sum_{i=1}^{n}\left(\omega_i(\hat{\boldsymbol{\theta}}_t)+\omega_i(\hat{\boldsymbol{\theta}}_t)^{-1}\right)\boldsymbol{B}_i\boldsymbol{B}_i^T+\lambda\boldsymbol{D}_q\right\}^{-1}\\
&\times\left\{\sum_{i=1}^{n}\left(-\omega_i(\hat{\boldsymbol{\theta}}_t)+\omega_i(\hat{\boldsymbol{\theta}}_t)^{-1}\right)\boldsymbol{B}_i+\lambda\boldsymbol{D}_q\hat{\boldsymbol{\theta}}_t\right\}.
\end{aligned}
$$

ˆ 注意，计算 $\boldsymbol{\theta}_{full}$ 的计算复杂度约为 O（ζn（$K+p+1$）^{2}），其中 ζ 是收敛前的迭代次数。

如我们所见，当完整数据规模 n 非常大时，计算成本很高。为解决这一问题，我们将在第 3 节提出一种子抽样算法，以降低计算成本。

ˆ 2.2 $\beta(t)$的理论性质

ˆ 我们将证明 $\beta(t)$的一致性和渐近正态性。为简洁起见，先给出以下记号。对于属于巴拿赫空间的函数 $f(t)$，$\|f\|_m = (\int_0^1 |f(t)|^m \mathrm{d}t)^{1/m}$，其中 $0 < m < \infty$。对于矩阵 $\boldsymbol{A} = (a_{ij})$，$\|\boldsymbol{A}\|_\infty = \max_{ij} |a_{ij}|$。此外，定义 $\boldsymbol{H} = \mathrm{E}\{\boldsymbol{B}\boldsymbol{B}^T(\epsilon + \epsilon^{-1})\} + \lambda/n\boldsymbol{D}_q$、$\boldsymbol{G} = \mathrm{E}\{\boldsymbol{B}\boldsymbol{B}^T(\epsilon - \epsilon^{-1})^2\}$，以及

$$\hat{\boldsymbol{G}} = \frac{1}{n}\sum_{i=1}^{n}\left\{-\omega_i(\hat{\boldsymbol{\theta}}_{full}) + \omega_i(\hat{\boldsymbol{\theta}}_{full})^{-1}\right\}^2 \boldsymbol{B}_i\boldsymbol{B}_i^T,$$

$$\hat{\boldsymbol{H}} = \frac{1}{n}\sum_{i=1}^{n}\left\{\omega_i(\hat{\boldsymbol{\theta}}_{full}) + \omega_i(\hat{\boldsymbol{\theta}}_{full})^{-1}\right\} \boldsymbol{B}_i\boldsymbol{B}_i^T + \lambda/n\boldsymbol{D}_q. \qquad (2.4)$$

此外，我们假设以下正则化条件成立。

(H.1)：对于函数型协变量 $X(t)$，假设$\mathrm{E}(\|X\|_8^8) < \infty$。

(H.2)：假设未知的函数型系数 $\beta(t)$足够光滑。

也就是说，$\beta(t)$ 具有 d' 阶导数 $\beta^{(d')}(t)$，使得

$$| \beta^{(d')}(t) - \beta^{(d')}(s) | \leq C_2 | t - s |^v, \quad t, s \in [0, 1],$$

其中常数为 $C_2 > 0$ 和 $v \in [0, 1]$。下文中，我们设

$d = d' + v \geq p + 1.$

（H.3）： E$\{(\epsilon - \epsilon^{-1}) \mid X\} = 0$。

(H.4)： E$\{(\epsilon + \epsilon^{-1})^6 \mid X\} < \infty$。

(H.5)：假设平滑参数 λ 满足 $\lambda = o(n^{1/2}K^{1/2-2q})$，其中 $q \leq p$。

(H.6)：假设节点数 $K = o(n^{1/2})$，且 $K/n^{1/(2d+1)} \to \infty$ as$n \to \infty$。

注 1. 假设 (H.1) 和 (H.2) 在函数型框架中相当常见（例如，Cardot、Ferraty 和 Sarda (2003)；Claeskens、Krivobokova 和 Opsomer (2009)）。假设 (H.3) 是对 $\beta(t)$ 进行 LPRE 估计的可识别性条件。假设 (H.3) 和 (H.4) 确保 LPRE估计量的一致性和渐近正态性。假设 (H.5) 和 (H.6) 主要用于获得 LPRE估计量的渐近无偏性。

ˆ 现在，我们给出 $\beta(t)$的一致性和渐近正态性。

定理1. 在假设 (H.1)–(H.6) 下，对于 $t \in [0, 1]$，当 $n \to \infty$ 时，我们有

ˆ (1)：（ 一致性）存在一个 LPRE估计量 $\beta(t)$，使得

$$\|\hat{\beta} - \beta\|_2 = O_P(n^{-1/2}K^{1/2});$$

(2)：（渐近正态性）

$$\{\boldsymbol{B}(t)^T \boldsymbol{V}_{full} \boldsymbol{B}(t)\}^{-1/2} \sqrt{n/K}(\hat{\beta}(t) - \beta(t)) \to N(0,1)$$

依分布，其中 $\boldsymbol{V}_{full} = K^{-1}\boldsymbol{H}^{-1}\boldsymbol{G}\boldsymbol{H}^{-1}$，其可由式（2.4）中定义的 $K^{-1}\hat{\boldsymbol{H}}^{-1}\hat{\boldsymbol{G}}\hat{\boldsymbol{H}}^{-1}$ 一致估计。

3. 最优子抽样

3.1 子抽样估计量及其理论性质

我们首先介绍一种适用于函数型线性乘法模型的一般随机子抽样算法，其中子样本根据某些抽样分布进行有放回随机抽取。

1. 抽样。给定较大的 K，我们生成 $\boldsymbol{B}_i = \int_0^1 x_i(t)\boldsymbol{B}(t)\mathrm{d}t$，新数据为 $\{(\boldsymbol{B}_i, y_i)\ i = 1, 2, ..., n\}$。为所有数据点分配子抽样概率 $\{\pi_i\}_{i=1}^n$，并根据 $\{\pi_i\}_{i=1}^n$ 从新数据中有放回地随机抽取一个大小为 $r(\ll n)$ 的子样本。将该子样本记为 $\{(\boldsymbol{B}_i, y_i, R_i)\ i = 1, 2, ..., n\}$，其中 R_i 表示全量数据中第 i 个数据点被选中的总次数，且 $\sum_{i=1}^n R_i = r$。

2. 估计。给定 λ，最小化以下损失函数以获得

基于子样本的˜估计 $\boldsymbol{\theta}$,

$$L^*(\boldsymbol{\theta};\lambda,K)=\frac{1}{r}\sum_{i=1}^{n}\frac{R_i}{\pi_i}\left(w_i(\tilde{\boldsymbol{\theta}}_t)+w_i(\tilde{\boldsymbol{\theta}}_t)^{-1}-2\right)+\frac{\lambda}{2}\boldsymbol{\theta}^T\boldsymbol{D}_q\boldsymbol{\theta}. \quad (3.5)$$

由于 $L^*(\boldsymbol{\theta};\lambda,K)$ 的凸性，采用牛顿-拉夫森方法，直到 $\boldsymbol{\theta}_{t+1}$ 和 $\boldsymbol{\theta}_t$ 足够接近,

$$\begin{aligned}\tilde{\boldsymbol{\theta}}_{t+1}=&\tilde{\boldsymbol{\theta}}_t-\left\{\sum_{i=1}^{n}\frac{R_i}{\pi_i}\left(\omega_i(\tilde{\boldsymbol{\theta}}_t)+\omega_i(\tilde{\boldsymbol{\theta}}_t)^{-1}\right)\boldsymbol{B}_i\boldsymbol{B}_i^T+\lambda D_q\right\}^{-1}\\&\times\left\{\sum_{i=1}^{n}\frac{R_i}{\pi_i}\left(-\omega_i(\tilde{\boldsymbol{\theta}}_t)+\omega_i(\tilde{\boldsymbol{\theta}}_t)^{-1}\right)\boldsymbol{B}_i+\lambda D_q\tilde{\boldsymbol{\theta}}_t\right\}. \quad (3.6)\end{aligned}$$

最后，我们可以得到子样本估计量 $\tilde{\beta}(t)=\boldsymbol{B}^T(t)\tilde{\boldsymbol{\theta}}$。

由于子抽样概率π_i可能依赖于全量数据 $\mathcal{F}_n=\{(x_i(t)\ y_i)\ i=1,2,\ldots,n,\ t\in[0,1]\}$，因此在采用逆概率加权技术时，损失函数 (3.5) 可保证无偏。˜下文将证明 $\beta(t)$ 朝向 $\beta(t)$ 的一致性和渐近正态性。还需要一个额外条件。

(H.7)：假设 $\max_{1\le i\le n} r(n\pi_i)^{-1}=O_P(1)$ 且 $r=o(K^2)$。

注2. 假设（H.7）常用于逆概率加权算法中，以限制权重，避免损失函数被子抽样概率极小的数据点过度放大（Ai 等（2021b）；Liu、You 和 Cao（2021）；Yan、Li 和 Niu（2022））。

定理2. 在假设（H.1）–（H.7）下，对于 $t \in [0,1]$，当 $r, n \to \infty$ 时，在给定 $\mathcal{F}_n$ 的条件下依概率有

$\tilde{}$(1)：（一致性）存在一个子抽样估计量 $\beta(t)$，使得

$$\|\tilde{\beta} - \hat{\beta}\|_2 = O_{P|\mathcal{F}_n}(r^{-1/2}K^{1/2});$$

(2)：（渐近正态性）

$$\left\{\boldsymbol{B}(t)^T \boldsymbol{V} \boldsymbol{B}(t)\right\}^{-1/2} \sqrt{r/K}(\tilde{\beta}(t) - \hat{\beta}(t)) \to N(0,1)$$

依分布，其中

$$\begin{aligned} \boldsymbol{V} &= \frac{1}{K}\hat{\boldsymbol{H}}^{-1}\boldsymbol{V}_{\pi}\hat{\boldsymbol{H}}^{-1}, \\ \boldsymbol{V}_{\pi} &= \frac{1}{n^2}\sum_{i=1}^{n}\frac{1}{\pi_i}\left\{-\omega_i(\hat{\boldsymbol{\theta}}_{full}) + \omega_i(\hat{\boldsymbol{\theta}}_{full})^{-1}\right\}^2 \boldsymbol{B}_i\boldsymbol{B}_i^T. \end{aligned} \quad (3.7)$$

3.2 最优子抽样概率

为了更好地逼近 $\beta(t)$，选择合适的子抽样概率至关重要。一种常用的准则是最小化 $\beta(t)$ 的渐近积分均方误差。根据定理2， $\beta(t)$ 的渐近积分均方误差如下

$$\mathrm{IMSE}(\tilde{\beta}(t) - \hat{\beta}(t)) = \frac{K}{r}\int_0^1 \boldsymbol{B}^T(t)\boldsymbol{V}\boldsymbol{B}(t)\mathrm{d}t. \quad (3.8)$$

注意，(3.7) 中定义的 $\boldsymbol{V}$ 是 $\sqrt{r/K}(\tilde{\boldsymbol{\theta}} - \hat{}\boldsymbol{\theta}_{full})$ 的渐近方差-协方差矩阵，并且积分不等式 $\int_0^1 \boldsymbol{B}^T(t)\boldsymbol{V}\boldsymbol{B}(t)\mathrm{d}t \leq$

$\int_0^1 \boldsymbol{B}^T(t)\boldsymbol{V}_1\boldsymbol{B}(t)\mathrm{d}t$ 当且仅当 $\boldsymbol{V} \leq \boldsymbol{V}_1$ 在 Loewner 序意义下成立。因此，我们专注于最小化 $\boldsymbol{\theta}$ 的渐近均方误差，并选择使 tr($\boldsymbol{V}$) 最小的子抽样概率。这在最优试验设计中称为 A-最优性准则；例如参见 Atkinson、Donev 和 Tobias（2007）。利用这一准则，我们能够推导出下述定理给出的最优子抽样概率。

定理3（A-最优性）。若子抽样概率 $\pi_i, i = 1, 2, \ldots, n$ 选取如下：

$$\pi_i^{FAopt} = \frac{|-y_i \exp(-\boldsymbol{B}_i^T\hat{\boldsymbol{\theta}}_{full}) + y_i^{-1}\exp(\boldsymbol{B}_i^T\hat{\boldsymbol{\theta}}_{full})|\|\hat{\boldsymbol{H}}^{-1}\boldsymbol{B}_i\|_2}{\sum_{i=1}^n |-y_i \exp(-\boldsymbol{B}_i^T\hat{\boldsymbol{\theta}}_{full}) + y_i^{-1}\exp(\boldsymbol{B}_i^T\hat{\boldsymbol{\theta}}_{full})|\|\hat{\boldsymbol{H}}^{-1}\boldsymbol{B}_i\|_2}, \quad (3.9)$$

则 $\sqrt{r/K}(\tilde{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_{full})$ 的总渐近均方误差 tr($\boldsymbol{V}$) 达到其最小值，$\beta(t)$ 的渐近积分均方误差也达到最小值。

然而，根据（2.4），我们知道 $\boldsymbol{H}$ 需要选定平滑参数 λ，而计算 $\|\hat{\boldsymbol{H}}^{-1}\boldsymbol{B}_i\|_2$ 的代价为 $O(n(K+p+1)^2)$，这会带来较高的计算开销。这些缺点使得这些最优子抽样概率不适合实际应用。因此，有必要寻找不需要 $\boldsymbol{H}$ 的替代概率，以降低计算复杂度。

请注意，正如式（3.7）所示，在渐近方差-协方差矩阵 $\boldsymbol{V}$ 中，只有 $\boldsymbol{V}_\pi$ 涉及 π_i。因此，根据 L¨ 下序关系，我们可以

只需关注 $\boldsymbol{V}_\pi$ 并使其迹最小化。鉴于其具有渐近无偏性，这可解释为使 $\sqrt{r/K}\,\hat{}\boldsymbol{H}(\tilde{\boldsymbol{\theta}} - \hat{}\boldsymbol{\theta}_{full})$ 的渐近均方误差最小化。在最优试验设计中，这称为 L-最优性准则（Atkinson、Donev 和 Tobias（2007））。因此，为缩短计算时间，我们考虑以下修正后的最优准则：使 tr($\boldsymbol{V}_\pi$) 最小化。

定理4（L 最优性）。若子抽样概率 $\pi_i, i = 1, 2, \ldots, n$ 选择为

$$\pi_i^{FLopt} = \frac{|-y_i \exp(-\boldsymbol{B}_i^T \hat{\boldsymbol{\theta}}_{full}) + y_i^{-1} \exp(\boldsymbol{B}_i^T \hat{\boldsymbol{\theta}}_{full})| \|\boldsymbol{B}_i\|_2}{\sum_{i=1}^n |-y_i \exp(-\boldsymbol{B}_i^T \hat{\boldsymbol{\theta}}_{full}) + y_i^{-1} \exp(\boldsymbol{B}_i^T \hat{\boldsymbol{\theta}}_{full})| \|\boldsymbol{B}_i\|_2}, \quad (3.10)$$

则 tr($\boldsymbol{V}_\pi$) 达到最小值。

由 (3.10) 可知，计算函数型 L 最优子抽样概率 π_i^{FLopt} 需要 $O(n($ $K + p + 1))$ 次浮点运算，这比计算 π_i^{FAopt} 的成本低得多，且随着 K 增大，这种差距会进一步扩大。

ˆ 注意到，子抽样概率 (3.10) 取决于 $\boldsymbol{\theta}_{full}$，而 $\boldsymbol{\theta}_{full}$ 是待估计的全数据估计，因此无法直接使用精确的概率分布。接下来，我们考虑一种近似分布，并提出一种两步算法。

1. 步骤 1：通过使用均匀抽样概率 r_0 ，运行一般子抽样算法，抽取大小为 r_0 的子样本，以获得初步估计量 $\boldsymbol{\theta}_{pilot}$ 。

$\pi_i^0 = 1/n$ 和 $\lambda = 0$。在（3.10）中用 $\boldsymbol{\theta}_{pilot}$ 替换$\hat{}$ $\boldsymbol{\theta}_{full}\tilde{}$，以推导最优子抽样概率的近似值。

2. 步骤 2：使用步骤 1 得到的近似最优概率抽取大小为 r 的子样本。给定 λ，使用式（3.6）基于该子样本获得估计量 $\boldsymbol{\theta}(\lambda)$，并根据下文讨论的贝叶斯信息准则$(\lambda)$，利用相应的子样本确定 λ。确定最优 λ 后，即可得到估计量 $\breve{\beta}(t) = \boldsymbol{B}^T(t)\breve{\boldsymbol{\theta}}$。

3.3 调优参数选择

对于次数 p 和导数阶数 q，我们根据经验选择次数为 3 的 B 样条和二阶惩罚项。节点数 K 并非关键参数，因为平滑程度由粗糙度惩罚参数 λ 控制（例如参见 Ruppert（2002）；Cardot、Ferraty 和 Sarda（2003））。对于参数 λ，我们选择贝叶斯信息准则来确定其取值：

$$\mathrm{BIC}(\lambda) = \log(\mathrm{RSS}) + \frac{\log(n)}{n}\mathrm{df},$$

其中，$\mathrm{RSS} = 1/n\sum_{i=1}^{n}\{\omega_i(\hat{\boldsymbol{\theta}}_{full}) + \omega_i(\hat{\boldsymbol{\theta}}_{full})^{-1} - 2\}$，df 表示有效自由度，即非零参数估计值的数量。然而，使用全量数据选择最优的 λ 在计算上代价高昂，因此我们采用最优子样本数据上的贝叶斯信息准则（BIC）对其进行近似。

4. 模拟研究

本节旨在通过合成数据研究所提出方法的有限样本性能。

4.1 LPRE 性能

在本实验中，我们将比较函数型最小二乘法（FLS）、函数型最小绝对偏差法（FLAD）和函数型最小乘积相对误差法（FLPRE）的性能。FLS 和 FLAD 估计值分别定义为使 $\sum_{i=1}^{n}[\log(y_i) - \int_0^1 x_i(t)\beta(t)\mathrm{d}t]^2$ and $\sum_{i=1}^{n}|\log(y_i) - \int_0^1 x_i(t)\beta(t)\mathrm{d}t|$ 最小化。模型 (1.1) 中的函数型协变量独立同分布地生成如下：$x_i(t) = \sum a_{ij}\boldsymbol{B}_j(t)\ i = 1, 2, \ldots, n$，其中 $\boldsymbol{B}_j(t)$ 是三次B样条基函数，在 0 到 1 之间的 100 个等间距点上取值。我们考虑基系数 $\boldsymbol{A} = (a_{ij})$ 的以下两种不同分布：

- C1. 多元正态分布 $N(0, \boldsymbol{\Sigma})$，其中 $\boldsymbol{\Sigma}_{ij} = 0.5^{|i-j|}$;
- C2. 具有 5 个自由度的多元 t 分布，$t_5(0, \boldsymbol{\Sigma}/10)$。

斜率函数 $\beta(t) = 7t^3 + 2\sin(4\pi t + 0.2)$ 和随机误差 ϵ_i 分为四种情况生成：

- R1. $\log(\epsilon) \sim N(0, 1)$;

- R2. $\log(\epsilon) \sim U(-2, 2)$;

- R3. ϵ 的概率密度函数为 $f(x) = c\exp(-x - x^{-1} - \log(x) + 2)I(x > 0)$, 其中 c 为归一化常数;

- R4. $\epsilon \sim U(0.5, b)$, 其中 b 的取值使得 $\mathrm{E}(\epsilon) = \mathrm{E}(1/\epsilon)$。

在具体模拟中，我们首先取 $n = 100$、500、1000 作为训练数据，然后取 $n = 300$、1500、3000 作为测试数据，并令节点数为$K = 10$。基于 500 次重复试验，我们分别使用均方积分误差平方根评估估计值的质量，并使用预测平方误差平方根（相对预测平方误差，RPSE）评估测试数据上的预测性能。定义如下:

$$\mathrm{IMSE} = \frac{1}{500}\sum_{k=1}^{500}\left[\int_0^1 \left\{\hat{\beta}^{(k)}(t) - \beta(t)\right\}^2 \mathrm{d}t\right]^{1/2},$$

and

$$\mathrm{RPSE} = \frac{1}{500}\sum_{k=1}^{500}\left[\frac{1}{n}\sum_{i=1}^{n}\left(\int_0^1 x_i(t)\beta(t)\mathrm{d}t - \int_0^1 x_i(t)\hat{\beta}^{(k)}(t)\mathrm{d}t\right)^2\right]^{1/2},$$

其中，$\tilde{\beta}^{(k)}(t)$是第 k 次运行得到的估计量。

模拟结果见表 1 和表 2。结果显示，除 C2-R1 这一种情况外，FLPRE 在所有情况下的表现都显著优于 FLS 和 LAD。在 C2-R1 情况下，FLPRE 始终优于 FLAD，而随着样本

量增大，LPRE 与 FLS 之间的差距逐渐缩小；当样本量达到 1000 时，LPRE 的表现略优于 FLS。此外，随着样本量增加，所有估计量的积分均方误差和相对预测平方误差均有所降低，这意味着样本量越大，所有估计量的表现越好。

4.2 子抽样性能

在本实验中，我们首先取 $n = 10^5$ 用于训练，然后取 $m = 1000$用于测试，以比较函数型 L 最优子抽样（FLopt）方法与均匀子抽样（Unif）方法的性能。模拟数据的分布与第 4.1 小节中的相同，并新增了一种关于基系数的情形。

- C3. 两个多元正态分布的混合 $0.5N(\mathbf{1}, \boldsymbol{\Sigma}) + 0.5N(-\mathbf{1}, \boldsymbol{\Sigma})$。

此外，根据假设 (H.6)，我们令节点数为 $K = \lceil n^{1/4} \rceil$。

为进行公平比较，所有情形均采用与全量数据相同的基函数和平滑参数。我们计算了子抽样大小分别为 2000、5000、8000、10000、15000 时相应子抽样估计量的均方积分误差平方根和相对预测平方误差，其中 $r_0 = 1000$

表 1：各估计量的积分均方误差。

	Dist	方法 R1		R2	R3	R4
$n = 100/300$	C1	FLPRE	1.4958	1.4588	1.2258	1.0766
		FLS	1.5023	1.6294	1.3061	1.1707
		FLAD	1.6797	2.0704	1.4343	1.2370
	C2	FLPRE	2.7053	2.5396	1.9177	1.4743
		FLS	2.4989	2.7184	1.9261	1.5298
		FLAD	2.8300	3.8908	2.2250	1.7391
$n = 500/1500$	C1	FLPRE	0.7550	0.7023	0.6086	0.5447
		FLS	0.8890	0.9575	0.7952	0.7321
		FLAD	1.1018	1.3702	0.9422	0.7867
	C2	FLPRE	1.5923	1.4627	1.1942	1.0242
		FLS	1.5522	1.6664	1.2809	1.1278
		FLAD	1.7679	2.2696	1.4302	1.2366
$n = 1000/3000$	C1	FLPRE	0.5508	0.4930	0.3879	0.3255
		FLS	0.6213	0.6600	0.5286	0.4834
		FLAD	0.8532	1.0904	0.6803	0.5346
	C2	FLPRE	1.2123	1.0902	0.9203	0.7943
		FLS	1.2552	1.3307	1.0612	0.9450
		FLAD	1.4475	1.8175	1.2096	1.0333

表 2：每种预测的相对预测平方误差。

	Dist	方法 R1		R2	R3	R4
$n = 100/300$	C1	FLPRE	0.2488	0.2416	0.2022	0.1770
		FLS	0.2516	0.2722	0.2170	0.1940
		FLAD	0.2828	0.3486	0.2394	0.2054
	C2	FLPRE	0.1869	0.1749	0.1323	0.1015
		FLS	0.1731	0.1882	0.1331	0.1053
		FLAD	0.1967	0.2696	0.1536	0.1201
$n = 500/1500$	C1	FLPRE	0.1240	0.1151	0.0983	0.0872
		FLS	0.1455	0.1569	0.1288	0.1181
		FLAD	0.1814	0.2276	0.1537	0.1274
	C2	FLPRE	0.1078	0.0989	0.0807	0.0686
		FLS	0.1055	0.1135	0.0872	0.0761
		FLAD	0.1208	0.1559	0.0978	0.0837
$n = 1000/3000$	C1	FLPRE	0.0902	0.0805	0.0625	0.0514
		FLS	0.1009	0.1083	0.0848	0.0767
		FLAD	0.1390	0.1792	0.1098	0.0855
	C2	FLPRE	0.0819	0.0734	0.0613	0.0525
		FLS	0.0847	0.0899	0.0710	0.0630
		FLAD	0.0985	0.1243	0.0815	0.0691

均方积分误差平方根和相对预测平方误差的定义如下

$$\begin{aligned}
\mathrm{IMSE} &= \frac{1}{1000}\sum_{k=1}^{1000}\left[\int_0^1 \left\{\tilde{\beta}^{(k)}(t)-\hat{\beta}(t)\right\}^2 \mathrm{d}t\right]^{1/2}, \\
\mathrm{RPSE} &= \frac{1}{1000}\sum_{k=1}^{1000}\left[\frac{1}{m}\sum_{i=1}^{m}\left(\int_0^1 x_i(t)\tilde{\beta}^{(k)}(t)\mathrm{d}t-\int_0^1 x_i(t)\hat{\beta}(t)\mathrm{d}t\right)^2\right]^{1/2}.
\end{aligned} \tag{4.11}$$

基于1000次重复试验，所有结果均经对数变换后展示于图1和图2中。

从图1可以清楚地看出，在所有情形下，FLopt子抽样方法的积分均方误差始终小于Unif子抽样方法，这与理论结果一致。也就是说，前者能够使子抽样估计量的渐近积分均方误差最小。尤其是当误差服从R1情形时，FLopt方法的表现要好得多。从图2可以看出，与积分均方误差的情况相同，在所有情形下，FLopt子抽样估计量的相对预测平方误差始终优于Unif子抽样估计量。此外，图1和图2还表明，FLopt方法同时取决于两类随机误差和协变量，且误差的影响大于协变量的影响。另外，正如预期的那样，随着子样本量增加，子抽样估计量的估计效率和预测效率均有所提高。

为了评估子抽样方法的计算效率，

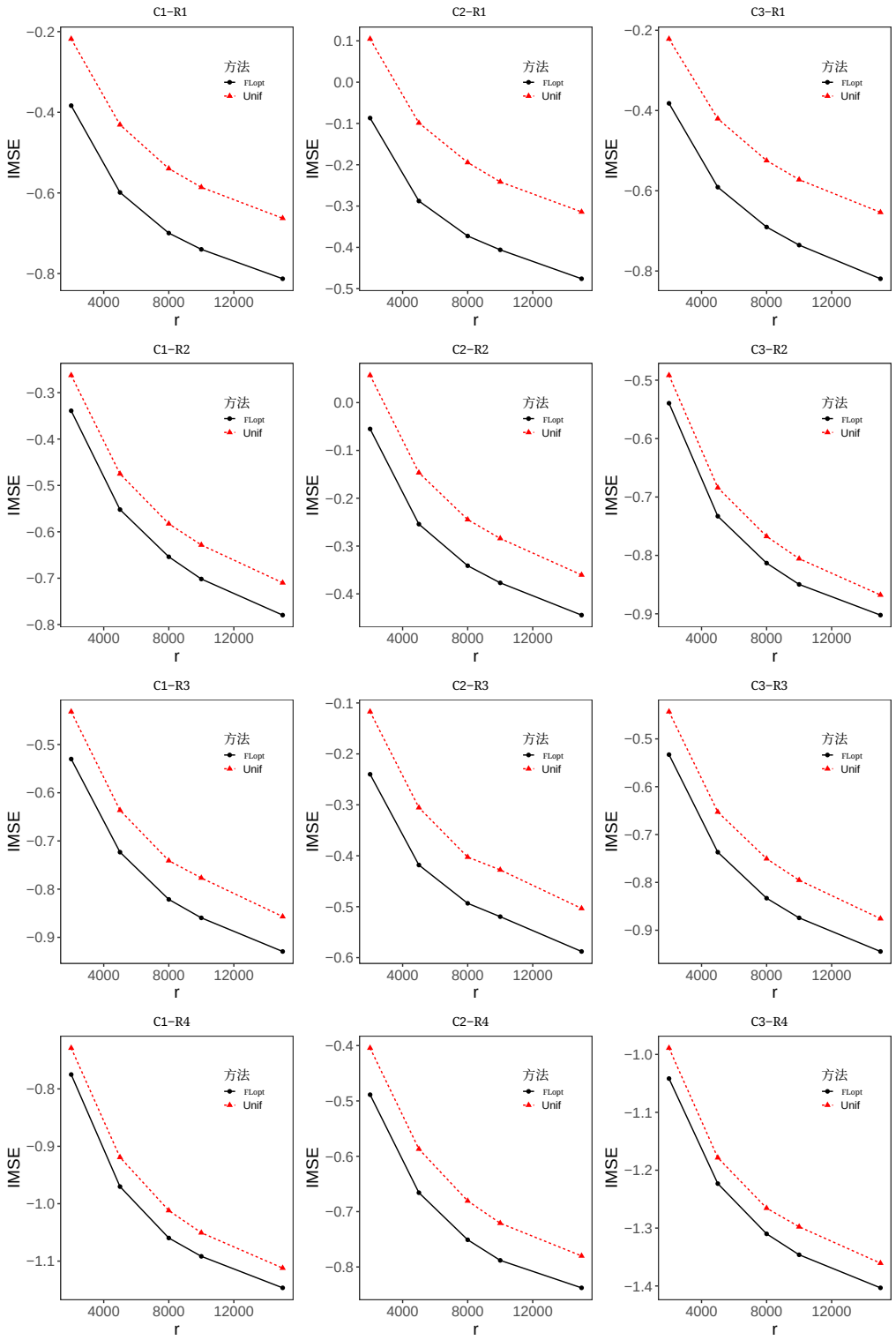


图1：不同子抽样大小下的积分均方误差 r 及固定的第一步子抽样 -

在 $n = 10^5$ 时，采用不同分布的抽样规模 $r_0 = 1000$ 。

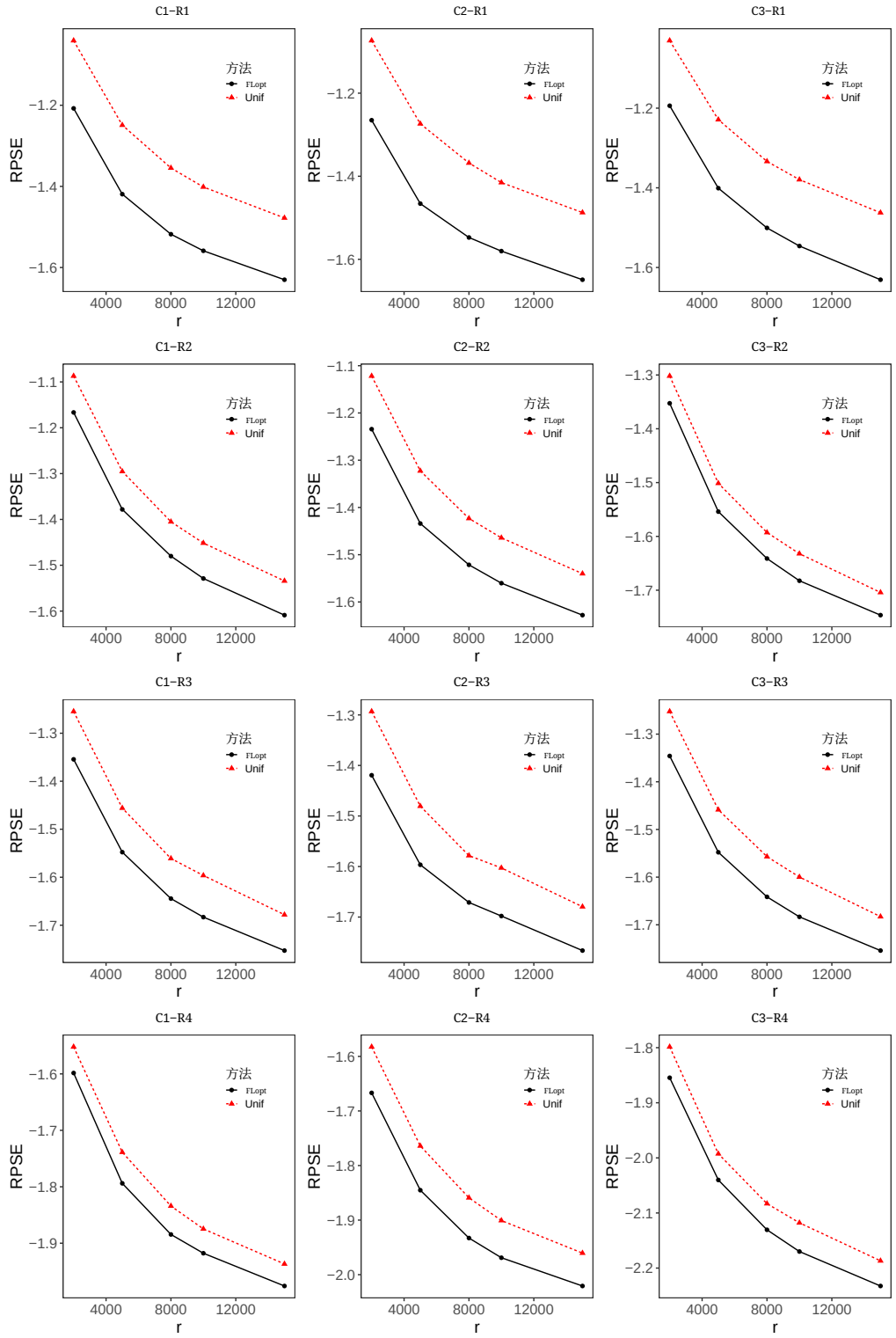


图2：不同子抽样大小下的相对预测平方误差 r 以及固定的第一步子抽样 -

当 $n = 10^5$ 时，不同分布下的抽样规模 $r_0 = 1000$ 。

表 3：在 $n = 10^6$、$r_0 = 200$、 $K = 50$ 和固定的 λ 条件下，不同子抽样大小 r 的 CPU 秒。所列时间为每种方法进行 100 次实现后计算出的平均时间。

方法	r				
	1000	2000	3000	4000	5000
FLopt	0.6036	0.6101	0.6230	0.6298	0.6393
Unif	0.0142	0.0250	0.0355	0.0455	0.0554
全量数据 CPU 秒：11.8155					

我们使用 R，在配备 Intel I5 处理器和 8GB 内存的 PC 上记录案例 C1-R1 中各方法的计算时间，不包括生成数据所需的时间。我们设定 $n = 10^6$, $r_0 = 200$，并将样条函数的节点数增加至 $K = 50$。每种子抽样策略均评估 100 次。表 3 给出了在固定 λ 的情况下，FLopt 和 Unif 子抽样方法在不同 r 上的结果。显然，与全量数据相比，子抽样可以显著提高计算效率；正如预期，FLopt方法的计算开销高于 Unif方法。

5. 真实数据分析

5.1 Tecator 数据

Tecator 数据可在 fda.usc软件包中获取，其中包含 215 个肉类样本。每个样本的数据包括 100 个通道的吸光度光谱，以及以百分比表示的脂肪、水和蛋白质含量。在 850-1050nm 波长范围内测得的 100 通道光谱以 2nm 为间隔进行密集测量，可视为函数型数据。图3展示了随机选取的 50 条吸光度光谱曲线，以及蛋白质含量的直方图。在本实验中，我们将使用所提出的 FLPRE 方法研究蛋白质含量。

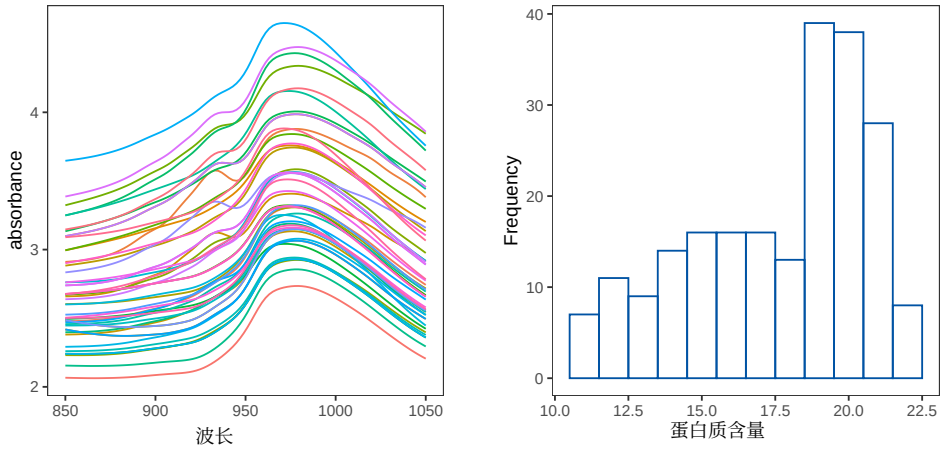


图3：左图：随机抽取的50条光谱曲线

右侧子图：蛋白质含量的直方图。

表 4: Tecator 数据的平均绝对百分比误差和平均乘积相对预测误差。

评价指标	FLS	FLAD	FLPRE
MAPE	3.5836	3.7452	3.5420
MPPE	0.0745	0.0739	0.0727

我们使用前160个观测值拟合模型，然后使用其余样本，通过平均绝对预测误差（平均绝对百分比误差，MAPE）和乘积相对预测误差（平均乘积相对预测误差，MPPE）评估预测效率。两个平均指标的计算方法如下:

$$\text{MAPE} = \frac{1}{55}\sum_{i=161}^{215}|y_i - \hat{y}_i|;$$

$$\text{MPPE} = \frac{1}{55}\sum_{i=161}^{215}(y_i - \hat{y}_i)^2/(y_i\hat{y}_i),$$

其中 $\hat{y}_i = \exp(\int_0^1 x_i(t)\hat{\beta}(t)\mathrm{d}t)$。所有结果见表 4，该表显示，所提出的 FLPRE 在预测蛋白质含量方面优于 FLS 和 FLAD。

5.2 北京多站点空气质量数据

此数据集可在 https://archive-beta.ics.uci.edu/ml/datasets/beijing+multi+site+ 中获取

数据由北京12个国家级空气质量监测站在2013年3月1日至2017年2月28日期间记录的逐小时空气污染物数据组成。我们在此主要关注利用前一天的PM_{10} 浓度预测每日最大值。

表5：空气质量数据的平均绝对百分比误差和平均乘积相对预测误差。

评价标准	FLS	FLAD	FLPRE
MAPE	7.9902	8.7580	7.6727
MPPE	0.5031	0.5223	0.4970

（$\mu g/m^3$），使用前一天的PM_{10} 轨迹（24小时）。我们删除所有缺失值，得到15573天的完整记录样本。取样本中前80%作为训练集，其余作为测试集。首先对原始观测值进行平方根变换，再使用15个傅里叶基函数将其转换为函数型数据。根据Sang和Cao（2020）的建议，可使用fda软件包中的Data2fd函数实现这一转换。图4左图展示了随机抽取的100条24小时PM_{10} 浓度曲线，其中时间尺度已转换为 $[0,1]$。图4右图展示了日内PM_{10} 浓度最大值的直方图。

我们采用平均绝对百分比误差和平均乘积相对预测误差作为评价标准，评估预测性能。表5显示，在预测 PM_{10} 浓度方面，所提出的 FLPRE 优于 FLS 和 FLAD。

此外，我们使用（4.11）计算积分均方误差和相对预测平方误差，并比较

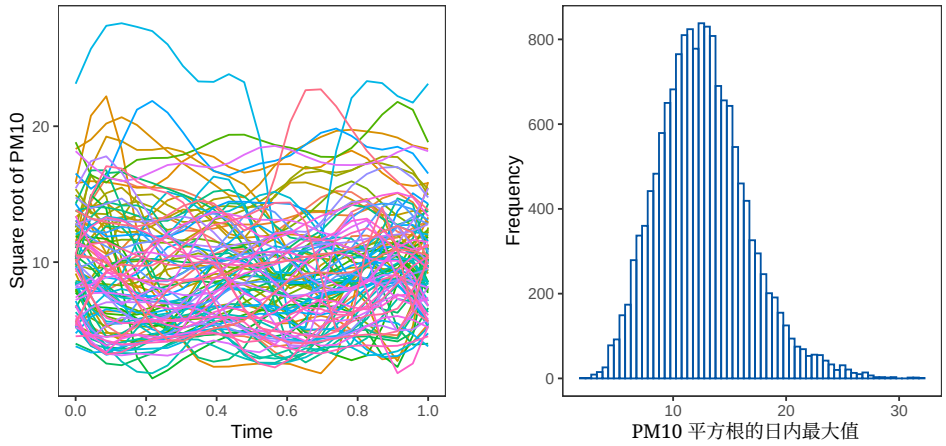


图4：左图：24小时 PM_{10}浓度曲线中随机选取的100条曲线。右图：日内 PM_{10} 浓度最大值的直方图。

FLopt方法与Unif方法。图5展示了抽样规模为 $r = 1000$、1500、2000、2500、3000时的结果，其中 $r_0 = 1000$。可以发现，与Unif方法相比，FLopt方法的积分均方误差和相对预测平方误差始终更小。此外，随着抽样规模 r 增加，积分均方误差和相对预测平方误差均逐渐减小，这表明子抽样方法具有估计一致性，并且其结果更接近基于全量数据的结果。

补充材料

所有技术证明均收录于在线补充材料中。

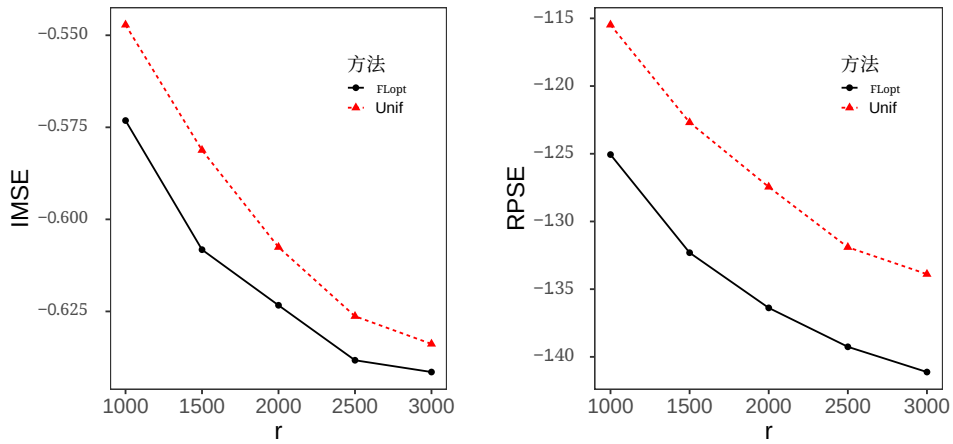


图5：不同子抽样大小 r 在 $r_0 = 1000$下重复1000次得到的积分均方误差和相对预测平方误差。

致谢

本研究得到国家自然科学基金（编号：11671060）和重庆市科技局自然科学基金项目（编号：cstc2019jcyj-msxmX0267）的资助。

参考文献

Ai M., Wang F., Yu J. 和 Zhang H.（2021年a）。大规模分位数回归的最优子抽样。*Journal of Complexity* **62**，101512。

Ai M., Yu J., Zhang H. and Wang H.（2021年b）。大数据回归的最优子抽样算法。*Statistica Sinica* **31**，749–772。

Atkinson A., Donev A. N. andTobias R. D.（2007）《最优实验设计》，含 SAS。牛津大学出版社，纽约。

Cardot H.、Ferraty F. 和 Sarda P.（2003）。函数型线性模型的样条估计量。*Statistica Sinica* **13**，571–591。

Chen, K.、Guo S.、Lin Y. 和 YingZ.（2010）。最小绝对相对误差估计。 *Journal of the American Statistical Association* **105**，1104–1112。

Chen K.、Lin Y.、Wang Z. 和 Ying Z.（2016）。最小乘积相对误差估计。 *Journal Of Multivariate Analysis* **144**，91–98。

Chen Y. 和 Liu H.（2021）。部分线性乘法模型的一种新相对误差估计。 *Communications in Statistics-Simulation and Computation*，1–19。

Chen Y.、Liu H. 和 Ma J.（2022）。针对正响应的单指标变系数乘法模型的局部最小乘积相对误差估计。 *Journal of Computationaland Applied Mathematics* **415**，114478。

Claeskens G.、Krivobokova T. 和 Opsomer J. D.（2009）。惩罚样条估计量的渐近性质。 *Biometrika* **96**，529–544。

de Boor C.（2001）。《样条实用指南》。施普林格出版社，柏林。

Fan, R.、Zhang S. 和 Wu Y.（2022）。具有局部稀疏性质的函数型乘法回归模型的惩罚相对误差估计。 *Journal of the Korean StatisticalSociety* **51**，666–691。

Fan Y.、Liu Y. 和 Zhu L.（2021）。线性分位数回归模型的最优子抽样。
Canadian Journal of Statistics **49**，1039–1057。

He S.、Yan X.（2022）。面向大规模函数型数据分析的函数主子空间抽样。
Electronic Journal of Statistics **16**，2621–2682。

Liu H.、You J. 和 Cao J.（2021）。面向海量数据的函数型 L-最优性子抽样。arXiv 预印本 arXiv:2104.03446。

Ma P.、Mahoney M. W. 和 YuB.（2015）。算法杠杆的统计视角。
Journal of Machine Learning Research 16，861–911。

Ma P.、Zhang X.、Xing X.、Ma J. 和 Mahoney M. W.（2020）。随机数值线性代数算法的抽样估计量渐近分析。载于 *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*，1026–1035。

Ming H.、Liu H. 和 YangH.（2022）。乘法加法模型中识别问题的最小乘积相对误差估计。
Journal of Computational and Applied Mathematics **404**，113886。

Ren M.、Zhao S. 和 Wang M.（2022）。海量数据下最小绝对相对误差估计量的最优子抽样。
Journal of Complexity **74**，101694。

Ruppert D.（2002）惩罚样条节点数量的选择。*Journal of Computational and Graphical Statistics* **11**，735–757。

Sang P. 和 Cao J.（2020）。函数型单指标分位数回归模型。*Statistics and*

Computing **30**，771–781。

Shao L.、Song S. 和 Zhou Y.（2022）。海量数据大样本分位数回归的最优子抽样。*Canadian Journal of Statistics*。https://doi.org/10.1002/cjs.11697

ShaoY. 和 Wang L.（2021）。海量数据复合分位数回归模型的最优子抽样。*Statistical Papers* **63**，1139–1161。

WangH.（2019）。采用最优子样本进行逻辑回归的更高效估计。机器学习研究期刊 20，1–59。

Wang H. 和 Ma Y.（2021）。大数据分位数回归的最优子抽样。*Biometrika* **108**，99–112。

Wang H.、Zhu R. 和 Ma P.（2018）。大样本逻辑回归的最优子抽样。*Journal of the American Statistical Association* **113**，829–844。

Wang T. 和 Zhang H.（2022）。海量数据乘法回归的最优子抽样。 *Statistica Neerlandica* **76**，418–449。

Yao Y. 和 Wang H.（2019）。Softmax 回归的最优子抽样。 *Statistical Papers* **60**，585–599。

Yan Q.、Li H. 和 Niu C.（2022）。函数型分位数回归的最优子抽样。*Statistical Papers*。https://doi.org/10.1007/s00362-022-01367-z。

Yu J.、Wang H.、Ai M. 和 Zhang H.（2020）。海量数据下最大准似然估计量的最优分布式子抽样。 *Journal of the American Statistical Associ-*

ation **117**，265–276。

Yuan X.、Li Y.、Dong X. 和 Liu T.（2022）。大数据中复合分位数回归的最优子抽样。*Statistical Papers* **63**，1649–1676。

Zhang T.、Huang Y.、Zhang Q.、Ma S. 和 Ahmed S. E.（2019）。部分函数线性乘法模型的惩罚相对误差估计。载于 AhmedS. E.、Carvalho F. 和 Puntanen S.（编），*Matrices, Statistics and Big Data*，施普林格，Cham。

ZhangT.、Zhang Q. 和 Li N.（2016）。函数二次乘法模型的最小绝对相对误差估计。*Communications in Statistics-Theory and Methods* **45**，5802–5817。

Qian Yan

重庆大学数学与统计学院，中国重庆 401331。

电子邮箱：qianyan@cqu.edu.cn

Hanyu Li

中国重庆市401331，重庆大学数学与统计学院。

电子邮箱：lihy.hy@gmail.com 或 hyli@cqu.edu.c n