doi：10.13413/j.cnki.jdxblxb.2024107

# 分位回归基于最优去相关得分的子抽样算法

黄小峰，邹雨浩，袁晓惠
（长春工业大学 数学与统计学院，长春 130012）

**摘要**：针对海量数据下高维分位回归模型，首先，构造基于去相关得分函数的子抽样算法，以估计感兴趣的低维参数；其次，推导所提估计的极限分布，并根据渐近协方差矩阵求出L-最优准则下的子抽样概率，给出高效的两步算法. 模拟和实证分析结果表明，最优子抽样方法显著优于均匀子抽样方法.

**关键词**：去相关得分；高维；海量数据；分位回归；子抽样
**中图分类号**：O212.2 **文献标志码**：A **文章编号**：1671-5489(2024)05-1102-11

# Subsampling Algorithm for Quantile Regression Based on Optimal Decorrelation Score

HUANG Xiaofeng, ZOU Yuhao, YUAN Xiaohui
(*School of Mathematics and Statistics, Changchun University of Technology, Changchun* 130012, *China*)

**Abstract**: For the high-dimensional quantile regression model with massive data, firstly, a subsampling algorithm based on the decorrelation score function was constructed to estimate the low-dimensional parameters of interest. Secondly, we derived the limit distribution of the proposed estimates and calculated the subsampling probability under the L-optimal criterion according to the asymptotic covariance matrix, giving an efficient two-step algorithm. The simulation and empirical analysis results show that the optimal subsampling method is significantly superior to the uniform subsampling method.

**Keywords**: decorrelation score; high-dimensional; massive data; quantile regression; subsampling

目前海量数据的处理方式主要有三类方法：分布式计算[1-3]、子抽样算法[4-7]和数据流估计[8-9]，其中子抽样方法可减少资源消耗，提高处理速度，降低成本，保持数据代表性，因而受到广泛关注，并已取得了许多研究结果. 例如：Fithian 等[4]将子抽样方法推广到逻辑回归中；Ma 等[10]探讨了子抽样算法在线性回归中参数估计的统计特性；Ai 等[11]和 Fan 等[12]分别将子抽样算法应用到广义线性模型和线性分位回归中，并在一般抽样方法下建立了估计量渐近正态性的理论基础；袁晓惠等[13]基于D-最优准则构造了分位回归中信息阵的最优子抽样方法；Wang 等[14]构造了基于 L-最优准则下分位回归模型的最优子抽样方法. 虽然子抽样算法在研究低维参数估计问题方面取得了一些成果，但对高维海量数据分析方法的研究目前仍处于探索阶段. 例如，Gao 等[15]研究了广义线性模型中在干扰参数影响下对关注的低维参数实施最优子抽样估计及推断的统一框架，但其研究主要集中在广义线性模型

**收稿日期**：2024-03-22.

**第一作者简介**：黄小峰(1998—)，男，苗族，硕士研究生，从事数理统计的研究，E-mail：2829715347@qq.com. **通信作者简介**：袁晓惠(1983—)，女，汉族，博士，教授，从事数理统计的研究，E-mail：yuanxh@ccut.edu.cn.

**基金项目**：国家社会科学基金(批准号：22BTJ019)和吉林省教育厅科学研究项目(批准号：JJKH20230749KJ).

参数的估计，并未涉及其他类型的模型.

在众多数据分析模型中，分位回归[16]因其能揭示响应变量的全方位特征并从中获取丰富信息而备受关注. 它通常采用加权最小绝对差方法进行估计，因而对离群点不敏感，能提供更稳健的结果，从而得到广泛关注. 例如，Wang 等[17]分析了纵向数据中部分线性变系数模型的分位估计；袁晓惠等[18]在部分协变量随机缺失机制下的分位回归模型中，提出了回归参数的诱导光滑加权估计及其渐近协方差估计；Wang 等[19]针对删失分位回归提出了一种新的基于多重稳健倾向得分的估计方法；Cheng 等[20]提出了正则化的投影评分方法，以解决高维混杂协变量存在下分位回归的参数估计问题. 但在高维海量数据下进行分位回归模型参数估计的研究目前文献报道较少. 鉴于此，本文考虑将去相关得分方程推广到高维分位回归最优子抽样中，对感兴趣的低维参数进行估计，并利用子抽样方法提升计算效率，同时降低因干扰参数导致精度下降的问题.

# 1　方　法

## 1.1　高维分位回归模型的去相关得分估计

在高维回归模型中，参数的维度通常较高，但与响应变量相关的协变量可能很少. 那些非显著影响响应变量的协变量可视为混杂协变量. 如何在高维回归模型中有效地估计低维参数，是近年来统计学领域的研究热点. Zhang 等[21]提出了一种半参数有效得分方法，用于构建高维线性模型中低维系数的估计和置信区间；Ning 等[22]提出了一种可用于稀疏高维模型中假设检验和置信区间的去相关得分估计方法；Cheng 等[20]提出了一种正则化投影得分方法，在存在高维混杂协变量的情况下，用于估计高维分位回归中的低维感兴趣参数.

假设响应变量为 $\boldsymbol{y}$，协变量为 $\boldsymbol{x}=(\boldsymbol{u}^{\mathrm{T}},\boldsymbol{z}^{\mathrm{T}})^{\mathrm{T}}$，其中 $\boldsymbol{u}$ 是已知的维数为 $d$ 的低维感兴趣协变量，$\boldsymbol{z}$ 是维数为 $p$ 的在预测响应变量时可能产生干扰的高维混杂协变量. 观测数据为 $\mathbb{F}_n=\{y_i,\boldsymbol{u}_i,\boldsymbol{z}_i\}_{i=1}^n$. 本文考虑分位回归模型：

$$Q_\tau(y_i\,|\,\boldsymbol{u}_i,\boldsymbol{z}_i)=\boldsymbol{u}_i^{\mathrm{T}}\boldsymbol{\theta}+\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{\gamma},$$

其中 $Q_\tau(y_i\,|\,\boldsymbol{u}_i,\boldsymbol{z}_i)$表示在给定协变量 $\boldsymbol{u}_i$ 和 $\boldsymbol{z}_i$ 时 $y_i$ 的 $\tau$ 条件分位数，$\boldsymbol{\theta}$ 表示感兴趣的低维系数，$\boldsymbol{\gamma}$ 表示干扰参数. Cheng 等[20]基于投影法构造了 $\boldsymbol{\theta}$ 的去相关得分估计方程. 与经典的分位回归方程不同，去相关得分方法可有效处理高维干扰参数的影响，得分方程为

$$\Psi_n(\widetilde{\boldsymbol{H}}_{\mathbf{F}},\hat{\boldsymbol{\gamma}}_{\mathbf{F}},\boldsymbol{\theta})=\frac{1}{n}\sum_{i=1}^{n}\psi_\tau(y_i-\boldsymbol{u}_i^{\mathrm{T}}\boldsymbol{\theta}-\boldsymbol{z}_i^{\mathrm{T}}\hat{\boldsymbol{\gamma}}_{\mathbf{F}})(\boldsymbol{u}_i-\widetilde{\boldsymbol{H}}_{\mathbf{F}}\boldsymbol{z}_i),\tag{1}$$

其中函数 $\psi_\tau(t)=\tau-I(t<0)$为 $\rho_\tau(t)=t\{\tau-I(t<0)\}$关于 $t$ 的导数. 通过求解 $\Psi_n(\widetilde{\boldsymbol{H}}_{\mathbf{F}},\hat{\boldsymbol{\gamma}}_{\mathbf{F}},\boldsymbol{\theta})=0$，可得未知参数 $\boldsymbol{\theta}$ 的估计 $\hat{\boldsymbol{\theta}}_{\mathbf{F}}$.

当 $\boldsymbol{z}$ 的维数 $p$ 较小时，矩阵 $\widetilde{\boldsymbol{H}}_{\mathbf{F}}$ 可由下式得到：

$$\widetilde{\boldsymbol{H}}_{\mathbf{F}}=\left\{\frac{1}{n}\sum_{i=1}^{r}\boldsymbol{u}_i\boldsymbol{z}_i^{\mathrm{T}}\right\}\left\{\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{z}_i\boldsymbol{z}_i^{\mathrm{T}}\right\}^{-1},$$

未知参数 $\boldsymbol{\gamma}$ 的估计 $\hat{\boldsymbol{\gamma}}_{\mathbf{F}}$ 由下式得到：

$$(\hat{\boldsymbol{\theta}},\hat{\boldsymbol{\gamma}}_{\mathbf{F}})=\underset{\boldsymbol{\theta},\boldsymbol{\gamma}}{\arg\min}\,\frac{1}{n}\sum_{i=1}^{n}\rho_\tau(y_i-\boldsymbol{u}_i^{\mathrm{T}}\boldsymbol{\theta}-\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{\gamma}).$$

当 $\boldsymbol{z}$ 的维数 $p$ 非常大时，可使用 Lasso 拟合多响应线性回归得到矩阵 $\widetilde{\boldsymbol{H}}_{\mathbf{F}}$ 的估计：

$$\widetilde{\boldsymbol{H}}_{\mathbf{F}}=\underset{\boldsymbol{H}\in\mathbb{R}^{d\times p}}{\arg\min}\frac{1}{2n}\sum_{i=1}^{n}\|\boldsymbol{u}_i-\boldsymbol{H}\boldsymbol{z}_i\|^2+\lambda_2\|\boldsymbol{h}_j\|,$$

其中参数 $\boldsymbol{h}_j$ 表示矩阵 $\boldsymbol{H}\in\mathbb{R}^{d\times p}$的第 $j$ 列. 未知参数 $\boldsymbol{\gamma}$ 的估计 $\hat{\boldsymbol{\gamma}}_{\mathbf{F}}$ 由如下惩罚估计算法得到：

$$(\hat{\boldsymbol{\theta}},\hat{\boldsymbol{\gamma}}_{\mathbf{F}})=\underset{\boldsymbol{\theta},\boldsymbol{\gamma}}{\arg\min}\,\frac{1}{n}\sum_{i=1}^{n}\rho_\tau(y_i-\boldsymbol{u}_i^{\mathrm{T}}\boldsymbol{\theta}-\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{\gamma})+\lambda_1\|\boldsymbol{\gamma}\|_1.\tag{2}$$

计算过程的关键是求解式(1)中的 $\hat{\boldsymbol{\gamma}}_{\mathbf{F}}$ 和 $\widetilde{\boldsymbol{H}}_{\mathbf{F}}$. 在低维情形下，通过迭代求解 $\Psi_n(\widetilde{\boldsymbol{H}}_{\mathbf{F}},\hat{\boldsymbol{\gamma}}_{\mathbf{F}},\boldsymbol{\theta})=0$ 计算. 该算法仅在低维情形下有理想的估计效果，但对于高维情形，该方法性能欠佳. 针对高维情形，

Cheng 等[20]引入了一步估计法对式(1)进行修正，得到如下去相关得分函数：

$$\widetilde{\Psi}_n(\widetilde{\boldsymbol{H}}_{\mathbf{F}},\tilde{\boldsymbol{\gamma}}_{\mathbf{F}},\boldsymbol{\theta})=\frac{1}{n}\sum_{i=1}^{n}\psi(y_i-(\boldsymbol{u}_i^{\mathrm{T}}-\widetilde{\boldsymbol{H}}_{\mathbf{F}}\boldsymbol{z}_i)^{\mathrm{T}}\boldsymbol{\theta}-(\widetilde{\boldsymbol{H}}_{\mathbf{F}}\boldsymbol{z}_i)^{\mathrm{T}}\check{\boldsymbol{\theta}}-\boldsymbol{z}_i^{\mathrm{T}}\tilde{\boldsymbol{\gamma}}_{\mathbf{F}})(\boldsymbol{u}_i-\widetilde{\boldsymbol{H}}_{\mathbf{F}}\boldsymbol{z}_i),$$

其中 $\check{\boldsymbol{\theta}}$ 表示基于方程(2)得到的初始估计. 设 $\tilde{y}_i=y_i-(\widetilde{\boldsymbol{H}}_{\mathbf{F}}\boldsymbol{z}_i)^{\mathrm{T}}\check{\boldsymbol{\theta}}-\boldsymbol{z}_i^{\mathrm{T}}\tilde{\boldsymbol{\gamma}}_{\mathbf{F}}$，则求解关键问题 $\widetilde{\Psi}_n(\widetilde{\boldsymbol{H}}_{\mathbf{F}},\tilde{\boldsymbol{\gamma}}_{\mathbf{F}},\boldsymbol{\theta})=0$ 即等价于求解

$$\check{\boldsymbol{\theta}}_{\mathrm{one}}=\arg\min_{\boldsymbol{\theta}}\frac{1}{n}\sum_{i=1}^{n}\rho_\tau(\tilde{y}_i-(\boldsymbol{u}_i-\widetilde{\boldsymbol{H}}_{\mathbf{F}}\boldsymbol{z}_i)^{\mathrm{T}}\boldsymbol{\theta}).$$

估计 $\check{\boldsymbol{\theta}}_{\mathbf{F}}$ 的渐近正态分布为

$$\sqrt{n}(\check{\boldsymbol{\theta}}_{\mathbf{F}}-\boldsymbol{\theta}_0)\to N(\mathbf{0},\boldsymbol{Q}_1^{-1}\boldsymbol{D}_1\boldsymbol{Q}_1^{-\mathrm{T}}),$$

其中 $\boldsymbol{Q}_1=E[f(0|\boldsymbol{u},\boldsymbol{z})(\boldsymbol{u}-\boldsymbol{H}_0\boldsymbol{z})\boldsymbol{u}^{\mathrm{T}}]$，$f(\cdot\,|\boldsymbol{u},\boldsymbol{z})$ 是 $\varepsilon=y-\boldsymbol{u}^{\mathrm{T}}\boldsymbol{\theta}_0-\boldsymbol{z}^{\mathrm{T}}\boldsymbol{\gamma}_0$ 的密度函数，$\boldsymbol{D}_1=\tau(1-\tau)\times E[(\boldsymbol{u}-\boldsymbol{H}_0\boldsymbol{z})(\boldsymbol{u}-\boldsymbol{H}_0\boldsymbol{z})^{\mathrm{T}}]$. 修正得分函数后由一步算法得到的估计 $\check{\boldsymbol{\theta}}_{\mathrm{one}}$ 的渐近正态分布为

$$\sqrt{n}(\check{\boldsymbol{\theta}}_{\mathrm{one}}-\boldsymbol{\theta}_0)\to N(\mathbf{0},\boldsymbol{Q}_2^{-1}\boldsymbol{D}_1\boldsymbol{Q}_2^{-\mathrm{T}}),$$

其中 $\boldsymbol{Q}_2=E[f(0|\boldsymbol{u},\boldsymbol{z})(\boldsymbol{u}-\boldsymbol{H}_0\boldsymbol{z})(\boldsymbol{u}-\boldsymbol{H}_0\boldsymbol{z})^{\mathrm{T}}]$.

### 1.2 基于去相关得分函数的一般子抽样算法

在海量数据情形下，传统子抽样算法不能直接应用于得分方程中. Gao 等[15]将去相关得分方法推广到海量数据下的高维广义线性模型中，构造了关于感兴趣低维参数的最优子抽样估计，提出了基于去相关得分函数的子抽样算法. 受此启发，本文考虑海量数据下高维分位回归模型参数的估计问题，利用去相关得分函数降低不精确的干扰参数估计带来的影响，并通过子抽样算法提升收敛速率. 以概率 $\pi_i$ $\left(\text{满足}\sum_{i=1}^{n}\pi_i=1\right)$ 抽取样本容量为 $r$ 的子样本集合 $\{y_i^*,\boldsymbol{u}_i^*,\boldsymbol{z}_i^*\}_{i=1}^r$，相应的概率为 $\pi_i^*$. 在去相关得分函数构造中，如何寻找基于子样本的投影矩阵 $\boldsymbol{H}^*$ 是关键，从理论上保证基于子样本的参数估计的相合性和渐近正态性是一个难点. 对于 $\boldsymbol{z}$ 的维数 $p$ 较小的情形，$\boldsymbol{H}^*$ 的估计 $\hat{\boldsymbol{H}}^*$ 可由下式得到：

$$\hat{\boldsymbol{H}}^*=\left\{\frac{1}{nr}\sum_{i=1}^{r}\frac{1}{\pi_i^*}\boldsymbol{u}_i^*(\boldsymbol{z}_i^*)^{\mathrm{T}}\right\}\left\{\frac{1}{nr}\sum_{i=1}^{r}\frac{1}{\pi_i^*}\boldsymbol{z}_i^*(\boldsymbol{z}_i^*)^{\mathrm{T}}\right\}^{-1},\tag{3}$$

未知参数 $\boldsymbol{\gamma}$ 的估计 $\hat{\boldsymbol{\gamma}}^*$ 可由下式计算得到：

$$(\hat{\boldsymbol{\theta}}^*,\hat{\boldsymbol{\gamma}}^*)=\arg\min_{\boldsymbol{\theta},\boldsymbol{\gamma}}\frac{1}{r}\sum_{i=1}^{r}\frac{1}{n\pi_i^*}\rho_\tau(y_i^*-(\boldsymbol{u}_i^*)^{\mathrm{T}}\boldsymbol{\theta}-(\boldsymbol{z}_i^*)^{\mathrm{T}}\boldsymbol{\gamma}).$$

得到基于子样本的初始分位回归系数 $\hat{\boldsymbol{\gamma}}^*$ 和投影矩阵 $\hat{\boldsymbol{H}}^*$ 后，$\boldsymbol{\theta}$ 的子抽样去相关得分函数定义为

$$\hat{\Psi}_n^*(\hat{\boldsymbol{H}}^*,\hat{\boldsymbol{\gamma}}^*,\boldsymbol{\theta})=\frac{1}{r}\sum_{i=1}^{r}\frac{1}{n\pi_i^*}\psi_\tau(y_i^*-(\boldsymbol{u}_i^*)^{\mathrm{T}}\boldsymbol{\theta}-(\boldsymbol{z}_i^*)^{\mathrm{T}}\hat{\boldsymbol{\gamma}}^*)(\boldsymbol{u}_i^*-\hat{\boldsymbol{H}}^*\boldsymbol{z}_i^*).\tag{4}$$

最后，根据式(4)求解方程 $\hat{\Psi}_n^*(\hat{\boldsymbol{H}}^*,\hat{\boldsymbol{\gamma}}^*,\boldsymbol{\theta})=0$，得到未知参数 $\boldsymbol{\theta}$ 的估计 $\hat{\boldsymbol{\theta}}$.

假设：

($\mathrm{H}_1$) 存在 $L>0$，使得 $\|\boldsymbol{u}\|<L$，a. s.；

($\mathrm{H}_2$) 对所有的$(\boldsymbol{u},\boldsymbol{z})$，$0<f(0|\boldsymbol{u},\boldsymbol{z})<\infty$，在$(\boldsymbol{u},\boldsymbol{z})$的邻域内存在常数 $C$，使得

$$|f(\boldsymbol{w}|\boldsymbol{u},\boldsymbol{z})-f(0|\boldsymbol{u},\boldsymbol{z})|\leqslant C|\boldsymbol{w}|^{1/2};$$

($\mathrm{H}_3$) 令 $\max\limits_{1\leqslant j\leqslant p}E[\|(\boldsymbol{u}-\widetilde{\boldsymbol{H}}_{\mathbf{F}}\boldsymbol{z})\boldsymbol{z}_j\|]=O(1)$，$\max\limits_{1\leqslant j\leqslant d}E[\|(\boldsymbol{u}-\widetilde{\boldsymbol{H}}_{\mathbf{F}}\boldsymbol{z})\boldsymbol{u}_j\|]=O(1)$；

($\mathrm{H}_4$) $\max\limits_{1\leqslant i\leqslant n}(n\pi_i)^{-1}=O_p(1)$；

($\mathrm{H}_5$) 矩阵 $\frac{1}{n}\sum\limits_{i=1}^{n}f(0|\boldsymbol{u}_i,\boldsymbol{z}_i)(\boldsymbol{u}_i-\widetilde{\boldsymbol{H}}_{\mathbf{F}}\boldsymbol{z}_i)\boldsymbol{u}_i^{\mathrm{T}}$ 依概率收敛到正定矩阵；

($\mathrm{H}_6$) 矩阵 $\frac{1}{n}\sum\limits_{i=1}^{n}f(0|\boldsymbol{u}_i,\boldsymbol{z}_i)(\boldsymbol{u}_i-\widetilde{\boldsymbol{H}}_{\mathbf{F}}\boldsymbol{z}_i)(\boldsymbol{u}_i-\widetilde{\boldsymbol{H}}_{\mathbf{F}}\boldsymbol{z}_i)^{\mathrm{T}}$ 依概率收敛到正定矩阵.

上述算法得出的 $\hat{\boldsymbol{\theta}}$ 有以下渐近性质：

**定理 1** 假设条件($\mathrm{H}_1$)～($\mathrm{H}_5$)成立，则当 $n\to\infty$ 且 $r\to\infty$ 时，在给定数据 $\mathbb{F}_n$ 的条件下，有

$$(\boldsymbol{Q}^{-1}\boldsymbol{D}\boldsymbol{Q}^{-T})^{-1/2}\sqrt{r}(\hat{\boldsymbol{\theta}}-\check{\boldsymbol{\theta}}_{\mathbf{F}})\to N(\mathbf{0},\boldsymbol{I}),$$

其中

$$\boldsymbol{Q}=\frac{1}{n}\sum_{i=1}^{n}f(0\,|\,\boldsymbol{u}_i,\boldsymbol{z}_i)(\boldsymbol{u}_i-\widetilde{\boldsymbol{H}}_{\mathbf{F}}\boldsymbol{z}_i)\boldsymbol{u}_i^{\mathrm{T}},$$

$$\boldsymbol{D}=\frac{\tau(1-\tau)}{n}\sum_{i=1}^{n}\frac{1}{n\pi_i}(\boldsymbol{u}_i-\widetilde{\boldsymbol{H}}_{\mathbf{F}}\boldsymbol{z}_i)(\boldsymbol{u}_i-\widetilde{\boldsymbol{H}}_{\mathbf{F}}\boldsymbol{z}_i)^{\mathrm{T}}.$$

证明：首先证明给定完全数据 $\mathbb{F}_n$ 时，$\hat{\boldsymbol{H}}^*-\widetilde{\boldsymbol{H}}_{\mathbf{F}}=O_p(r^{-1/2})$. 令

$$\hat{\boldsymbol{H}}_1^*=\frac{1}{nr}\sum_{i=1}^{r}\frac{1}{\pi_i^*}\boldsymbol{u}_i^*(\boldsymbol{z}_i^*)^{\mathrm{T}},\qquad \widetilde{\boldsymbol{H}}_{\mathbf{F}1}=\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{u}_i\boldsymbol{z}_i^{\mathrm{T}},$$

$$\hat{\boldsymbol{H}}_2^*=\frac{1}{nr}\sum_{i=1}^{r}\frac{1}{\pi_i^*}\boldsymbol{z}_i^*(\boldsymbol{z}_i^*)^{\mathrm{T}},\qquad \widetilde{\boldsymbol{H}}_{\mathbf{F}2}=\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{z}_i\boldsymbol{z}_i^{\mathrm{T}},$$

则可得 $E(\hat{\boldsymbol{H}}_1^*\,|\,\mathbb{F}_n)=\widetilde{\boldsymbol{H}}_{\mathbf{F}1}$，$E(\hat{\boldsymbol{H}}_2^*\,|\,\mathbb{F}_n)=\widetilde{\boldsymbol{H}}_{\mathbf{F}2}$. 将矩阵 $\hat{\boldsymbol{H}}_1^*-\widetilde{\boldsymbol{H}}_{\mathbf{F}1}$ 的第 $j$ 行第 $k$ 列元素记为 $\Delta_{j,k}$，则 $E(\Delta_{j,k}\,|\,\mathbb{F}_n)=0$. 下面计算条件二阶矩：

$$E(\Delta_{j,k}^2\,|\,\mathbb{F}_n)=E\left[\left(\frac{1}{nr}\sum_{i=1}^{r}\frac{1}{\pi_i^*}u_{ij}^*z_{ik}^*-\frac{1}{n}\sum_{i=1}^{n}u_{ij}z_{ik}\right)^2\,\middle|\,\mathbb{F}_n\right]=\frac{1}{r}\sum_{i=1}^{n}\pi_i\left(\frac{1}{n\pi_i}u_{ij}z_{ik}-\frac{1}{n}\sum_{i=1}^{n}u_{ij}z_{ik}\right)^2=$$

$$\frac{1}{rn^2}\sum_{i=1}^{n}\frac{1}{\pi_i}(u_{ij}z_{ik})^2-\frac{1}{r}\left(\frac{1}{n}\sum_{i=1}^{n}u_{ij}z_{ik}\right)^2\leqslant$$

$$\frac{1}{r}\left(\max_{1\leqslant i\leqslant n}\frac{\|\boldsymbol{u}_i\|^2}{n\pi_i}\right)-\frac{1}{r}\left(\frac{1}{n}\sum_{i=1}^{n}u_{ij}z_{ik}\right)^2=O_p(r^{-1}).$$

根据 Chebyshev 不等式可知，$\Delta_{j,k}=O_p(r^{-1/2})$，从而 $\hat{\boldsymbol{H}}_1^*-\widetilde{\boldsymbol{H}}_{\mathbf{F}1}=O_p(r^{-1/2})$. 类似地，可证明 $\hat{\boldsymbol{H}}_2^*-\widetilde{\boldsymbol{H}}_{\mathbf{F}2}=O_p(r^{-1/2})$. 由于

$$(\hat{\boldsymbol{H}}_2^*)^{-1}-\widetilde{\boldsymbol{H}}_{\mathbf{F}2}^{-1}=-\widetilde{\boldsymbol{H}}_{\mathbf{F}2}^{-1}(\hat{\boldsymbol{H}}_2^*-\widetilde{\boldsymbol{H}}_{\mathbf{F}2})(\hat{\boldsymbol{H}}_2^*)^{-1}=O_p(r^{-1/2}),$$

因此可得

$$\hat{\boldsymbol{H}}^*-\widetilde{\boldsymbol{H}}_{\mathbf{F}}=(\hat{\boldsymbol{H}}_1^*)(\hat{\boldsymbol{H}}_2^*)^{-1}-\widetilde{\boldsymbol{H}}_{\mathbf{F}1}\widetilde{\boldsymbol{H}}_{\mathbf{F}2}^{-1}=$$

$$(\hat{\boldsymbol{H}}_1^*-\widetilde{\boldsymbol{H}}_{\mathbf{F}1})[(\hat{\boldsymbol{H}}_2^*)^{-1}-\widetilde{\boldsymbol{H}}_{\mathbf{F}2}^{-1}]+\widetilde{\boldsymbol{H}}_{\mathbf{F}1}[(\hat{\boldsymbol{H}}_2^*)^{-1}-\widetilde{\boldsymbol{H}}_{\mathbf{F}}^{-1}]+(\hat{\boldsymbol{H}}_1^*-\widetilde{\boldsymbol{H}}_{\mathbf{F}1})\widetilde{\boldsymbol{H}}_{\mathbf{F}2}^{-1}=O_p(r^{-1/2}).$$

根据文献[23]中定理 1 可知，$\hat{\boldsymbol{\theta}}^*-\check{\boldsymbol{\theta}}_{\mathbf{F}}=O_p(r^{-1/2})$，$\hat{\boldsymbol{\gamma}}^*-\boldsymbol{\gamma}_{\mathbf{F}}=O_p(r^{-1/2})$.

令

$$\hat{\Psi}_n^*(\widetilde{\boldsymbol{H}}_{\mathbf{F}},\boldsymbol{\gamma}_{\mathbf{F}},\check{\boldsymbol{\theta}}_{\mathbf{F}})=\frac{1}{r}\sum_{i=1}^{r}\frac{1}{n\pi_i^*}\psi_\tau(y_i^*-(\boldsymbol{u}_i^*)^{\mathrm{T}}\check{\boldsymbol{\theta}}_{\mathbf{F}}-(\boldsymbol{z}_i^*)^{\mathrm{T}}\boldsymbol{\gamma}_{\mathbf{F}})(\boldsymbol{u}_i^*-\widetilde{\boldsymbol{H}}_{\mathbf{F}}\boldsymbol{z}_i^*),$$

下面证明：给定全数据 $\mathbb{F}_n$ 时，$\hat{\Psi}_n^*(\widetilde{\boldsymbol{H}}_{\mathbf{F}},\boldsymbol{\gamma}_{\mathbf{F}},\check{\boldsymbol{\theta}}_{\mathbf{F}})=O_p(r^{-1/2})$. 计算可得

$$E(\hat{\Psi}_n^*(\widetilde{\boldsymbol{H}}_{\mathbf{F}},\boldsymbol{\gamma}_{\mathbf{F}},\check{\boldsymbol{\theta}}_{\mathbf{F}})\,|\,\mathbb{F}_n)=\frac{1}{n}\sum_{i=1}^{n}\psi_\tau(y_i-\boldsymbol{u}_i^{\mathrm{T}}\check{\boldsymbol{\theta}}_{\mathbf{F}}-\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{\gamma}_{\mathbf{F}})(\boldsymbol{u}_i-\widetilde{\boldsymbol{H}}_{\mathbf{F}}\boldsymbol{z}_i)=0,$$

$$\mathrm{Var}(\hat{\Psi}_n^*(\widetilde{\boldsymbol{H}}_{\mathbf{F}},\boldsymbol{\gamma}_{\mathbf{F}},\check{\boldsymbol{\theta}}_{\mathbf{F}})\,|\,\mathbb{F}_n)=\frac{1}{rn^2}\sum_{i=1}^{n}\frac{1}{\pi_i}[\psi_\tau(y_i-\boldsymbol{u}_i^{\mathrm{T}}\check{\boldsymbol{\theta}}_{\mathbf{F}}-\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{\gamma}_{\mathbf{F}})]^2(\boldsymbol{u}_i-\widetilde{\boldsymbol{H}}_{\mathbf{F}}\boldsymbol{z}_i)(\boldsymbol{u}_i-\widetilde{\boldsymbol{H}}_{\mathbf{F}}\boldsymbol{z}_i)^{\mathrm{T}}=O_p(r^{-1}).$$

根据 Chebyshev 不等式可知，$\hat{\Psi}_n^*(\widetilde{\boldsymbol{H}}_{\mathbf{F}},\boldsymbol{\gamma}_0,\check{\boldsymbol{\theta}}_{\mathbf{F}})=O_p(r^{-1/2})$. 同理可证明

$$\hat{\Psi}_n^*(\widetilde{\boldsymbol{H}}_{\mathbf{F}},\boldsymbol{\gamma}_{\mathbf{F}},\check{\boldsymbol{\theta}}_{\mathbf{F}})-\hat{\Psi}_n^*(\widetilde{\boldsymbol{H}}^*,\boldsymbol{\gamma}_{\mathbf{F}},\check{\boldsymbol{\theta}}_{\mathbf{F}})=\frac{1}{r}\sum_{i=1}^{r}\frac{1}{n\pi_i^*}\psi_\tau(y_i^*-(\boldsymbol{u}_i^*)^{\mathrm{T}}\check{\boldsymbol{\theta}}_{\mathbf{F}}-(\boldsymbol{z}_i^*)^{\mathrm{T}}\boldsymbol{\gamma}_{\mathbf{F}})(\widetilde{\boldsymbol{H}}^*-\widetilde{\boldsymbol{H}}_{\mathbf{F}})\boldsymbol{z}_i^*=$$

$$O_p(r^{-1}).\tag{5}$$

因此 $\hat{\Psi}_n^*(\widetilde{\boldsymbol{H}}^*,\boldsymbol{\gamma}_{\mathbf{F}},\check{\boldsymbol{\theta}}_{\mathbf{F}})=O_p(r^{-1/2})$. 下面证明 $\hat{\Psi}_n^*(\widetilde{\boldsymbol{H}}_{\mathbf{F}},\boldsymbol{\gamma}_{\mathbf{F}},\check{\boldsymbol{\theta}}_{\mathbf{F}})$的渐近正态性. 记

$$\boldsymbol{\eta}_i=\frac{1}{n\pi_i^*}\psi_\tau(y_i^*-(\boldsymbol{u}_i^*)^{\mathrm{T}}\check{\boldsymbol{\theta}}_{\mathbf{F}}-(\boldsymbol{z}_i^*)^{\mathrm{T}}\boldsymbol{\gamma}_{\mathbf{F}})(\boldsymbol{u}_i^*-\widetilde{\boldsymbol{H}}_{\mathbf{F}}\boldsymbol{z}_i^*),$$

则 $\hat{\Psi}_n^*(\widetilde{\boldsymbol{H}}^*,\boldsymbol{\gamma}_{\mathbf{F}},\check{\boldsymbol{\theta}}_{\mathbf{F}})=\frac{1}{r}\sum_{i=1}^{r}\boldsymbol{\eta}_i$. 在给定全数据 $\mathbb{F}_n$ 的条件下，$\boldsymbol{\eta}_i$ $(i=1,2,\cdots,r)$ 独立同分布，且

$E(\boldsymbol{\eta}_i \mid \mathbb{F}_n)=O_p(n^{-1/2})$，$\mathrm{Var}(\boldsymbol{\eta}_i \mid \mathbb{F}_n)=\boldsymbol{D}-o_p(1)$. 下面验证 Lindeberg-Feller 条件，对某个 $\delta>0$ 及任意的 $\epsilon>0$，有

$$\sum_{i=1}^{r} E[\| r^{-1/2}\boldsymbol{\eta}_i \|^2 I(\| \boldsymbol{\eta}_i \| > r^{1/2}\epsilon) \mid \mathbb{F}_n] \leqslant \frac{1}{r^{1+\delta/2}\epsilon^{\delta}} \sum_{i=1}^{r} E[\| \boldsymbol{\eta}_i \|^{2+\delta} I(\| \boldsymbol{\eta}_i \| > r^{1/2}\epsilon) \mid \mathbb{F}_n] \leqslant$$

$$\frac{1}{r^{1+\delta/2}\epsilon^{\delta}} \sum_{i=1}^{r} E(\| \boldsymbol{\eta}_i \|^{2+\delta} \mid \mathbb{F}_n) \leqslant$$

$$\frac{2}{r^{\delta/2} n^{2+\delta}\epsilon^{\delta}} \sum_{i=1}^{n} \frac{1}{\pi_i^{1+\delta}} \| \boldsymbol{u}_i - \widetilde{\boldsymbol{H}}_{\mathbf{F}} \boldsymbol{z}_i \|^{2+\delta} \leqslant$$

$$\frac{2}{r^{\delta/2}\epsilon^{\delta}} \max_{1\leqslant i\leqslant n} \frac{1}{(n\pi_i)^{1+\delta}} \left( \frac{1}{n} \sum_{i=1}^{n} \| \boldsymbol{u}_i - \widetilde{\boldsymbol{H}}_{\mathbf{F}} \boldsymbol{z}_i \|^{2+\delta} \right) = o_p(1).$$

由 Lindeberg-Feller 中心极限定理可知，$\boldsymbol{D}^{-1/2}\sqrt{r}\hat{\Psi}_n^*(\widetilde{\boldsymbol{H}}_{\mathbf{F}}, \boldsymbol{\gamma}_{\mathbf{F}}, \check{\boldsymbol{\theta}}_{\mathbf{F}}) \rightarrow N(\boldsymbol{0}, \boldsymbol{I})$. 根据式(5)和 Slutsky 定理可知，$\boldsymbol{D}^{-1/2}\sqrt{r}\hat{\Psi}_n^*(\widetilde{\boldsymbol{H}}^*, \boldsymbol{\gamma}_{\mathbf{F}}, \check{\boldsymbol{\theta}}_{\mathbf{F}}) \rightarrow N(\boldsymbol{0}, \boldsymbol{I})$. 用重期望公式可得

$$\hat{\Psi}_n^*(\widetilde{\boldsymbol{H}}^*, \boldsymbol{\gamma}_{\mathbf{F}}, \hat{\boldsymbol{\theta}}) - \hat{\Psi}_n^*(\widetilde{\boldsymbol{H}}^*, \boldsymbol{\gamma}_{\mathbf{F}}, \check{\boldsymbol{\theta}}_{\mathbf{F}}) = E[\hat{\Psi}_n^*(\widetilde{\boldsymbol{H}}^*, \boldsymbol{\gamma}_{\mathbf{F}}, \hat{\boldsymbol{\theta}})] - E[\hat{\Psi}_n^*(\widetilde{\boldsymbol{H}}^*, \boldsymbol{\gamma}_{\mathbf{F}}, \check{\boldsymbol{\theta}}_{\mathbf{F}})] + o_p(r^{-1/2}) =$$

$$\boldsymbol{Q}(\hat{\boldsymbol{\theta}} - \check{\boldsymbol{\theta}}_{\mathbf{F}}) + o_p(r^{-1/2}).$$

因此

$$\sqrt{r}(\hat{\boldsymbol{\theta}} - \check{\boldsymbol{\theta}}_{\mathbf{F}}) = -\boldsymbol{Q}^{-1}\boldsymbol{D}^{1/2}\boldsymbol{D}^{-1/2}\,\hat{\Psi}_n^*(\widetilde{\boldsymbol{H}}^*, \boldsymbol{\gamma}_{\mathbf{F}}, \check{\boldsymbol{\theta}}_{\mathbf{F}}) + o_p(1).$$

从而可得$(\boldsymbol{Q}^{-1}\boldsymbol{D}\boldsymbol{Q}^{-T})^{-1/2}\sqrt{r}(\hat{\boldsymbol{\theta}} - \check{\boldsymbol{\theta}}_{\mathbf{F}}) \rightarrow N(\boldsymbol{0}, \boldsymbol{I})$. 证毕.

当 $p$ 非常大时，$\hat{\boldsymbol{H}}^*$ 的估计效果可能会较差. 可使用 Lasso 拟合多响应线性回归模型，估计投影矩阵$\boldsymbol{H}^*$. 对任何 $\boldsymbol{H}^* \in \mathbb{R}^{d\times p}$，用 $\boldsymbol{h}_j^*$ 表示其第 $j$ 列，并通过下式估计 $\boldsymbol{H}^*$：

$$\hat{\boldsymbol{H}}^* = \underset{\boldsymbol{H}^* \in \mathbb{R}^{d\times p}}{\operatorname{argmin}} \frac{1}{r} \sum_{i=1}^{r} \frac{1}{\pi_i^*} \| \boldsymbol{u}_i^* - \boldsymbol{H}^* \boldsymbol{z}_i^* \|^2 + \lambda_1 \| \boldsymbol{h}_j \|. \tag{6}$$

类似地，可通过惩罚估计 $\boldsymbol{\gamma}$：

$$(\hat{\boldsymbol{\theta}}^*, \hat{\boldsymbol{\gamma}}^*) = \underset{\boldsymbol{\theta},\boldsymbol{\gamma}}{\operatorname{argmin}} \frac{1}{r} \sum_{i=1}^{r} \frac{1}{n\pi_i^*} \rho_\tau(y_i^* - (\boldsymbol{u}_i^*)^{\mathrm{T}}\boldsymbol{\theta} - (\boldsymbol{z}_i^*)^{\mathrm{T}}\boldsymbol{\gamma}) + \lambda_2 \| \boldsymbol{\gamma} \|_1.$$

在得到基于子样本的初始分位回归系数 $\hat{\boldsymbol{\gamma}}^*$ 和投影矩阵 $\hat{\boldsymbol{H}}^*$ 后，基于子样本的一步估计方法得到 $\hat{\boldsymbol{\theta}}_{\mathrm{one}}$，其渐近性质如下.

**定理 2** 假设条件$(\mathrm{H}_1)\sim(\mathrm{H}_4)$,$(\mathrm{H}_6)$成立，则当 $n\rightarrow\infty$且 $r\rightarrow\infty$，并给定数据 $\mathbb{F}_n$ 时，有

$$(\widetilde{\boldsymbol{Q}}^{-1}\boldsymbol{D}\widetilde{\boldsymbol{Q}}^{-\mathrm{T}})^{-1/2}(\hat{\boldsymbol{\theta}}_{\mathrm{one}} - \boldsymbol{\theta}_{\mathbf{F}}) \rightarrow N(\boldsymbol{0}, \boldsymbol{I}),$$

其中 $\widetilde{\boldsymbol{Q}} = \frac{1}{n}\sum_{i=1}^{n} f(0 \mid \boldsymbol{u}_i, \boldsymbol{z}_i)(\boldsymbol{u}_i - \widetilde{\boldsymbol{H}}_{\mathbf{F}}\boldsymbol{z}_i)(\boldsymbol{u}_i - \widetilde{\boldsymbol{H}}_{\mathbf{F}}\boldsymbol{z}_i)^{\mathrm{T}}$.

定理 2 的证明类似定理 1，故略.

由于去相关得分函数得到修正，所以定理 2 中的 $\widetilde{\boldsymbol{Q}}$ 与 $\boldsymbol{Q}$ 有差异. 可将修改去相关得分函数后得到的估计量 $\hat{\boldsymbol{\theta}}_{\mathrm{one}}$视为从初始估计 $\hat{\boldsymbol{\theta}}$ 的一步更新.

### 1.3 最优去相关得分子抽样概率

下面讨论最优子抽样概率的计算方法. 首先，基于 L-最优性准则提出一种子抽样概率的确定方法；其次，讨论该方法的实现策略；最后，总结一种两步算法.

由于定理 1 和定理 2 中的渐近协方差矩阵依赖于子抽样概率，因此下面通过选择最优子抽样概率，提出一种有效的子抽样方法. 通过最小化估计量 $\hat{\boldsymbol{\theta}}$ 的渐近均方误差获取最优子抽样概率，即在定理 1 中 $\min\limits_{\pi_i} \| \mathrm{Var}(\hat{\boldsymbol{\theta}}) \| = \min\limits_{\pi_i} \mathrm{tr}(\boldsymbol{Q}^{-1}\boldsymbol{D}\boldsymbol{Q}^{-\mathrm{T}})$，由于 $\boldsymbol{Q}^{-1}\boldsymbol{D}\boldsymbol{Q}^{-\mathrm{T}}$ 中只有 $\boldsymbol{D}$ 与抽样概率 $\pi_i$ 有关，所以 $\operatorname{argmin}\limits_{\pi_i} \mathrm{tr}(\boldsymbol{Q}^{-1}\boldsymbol{D}\boldsymbol{Q}^{-\mathrm{T}}) = \operatorname{argmin}\limits_{\pi_i} \mathrm{tr}(\boldsymbol{D})$. 又因为在定理 1 和定理 2 中，两个协方差矩阵的 $\boldsymbol{Q}$ 和 $\widetilde{\boldsymbol{Q}}$ 不相等，因此考虑通过最小化 $\mathrm{tr}(\boldsymbol{D})$寻求最优子抽样概率，即 L-最优性准则，旨在优化子抽样概率以提高估计效率. 下面根据 L-最优性准则确定最优子抽样概率.

**定理 3**　假设定理 1 的条件成立，则在 L-最优准则下，抽样概率形式为

$$\pi_i^{\mathrm{Lopt}}=\frac{\|\boldsymbol{u}_i-\widetilde{\boldsymbol{H}}_{\mathrm{F}}\boldsymbol{z}_i\|}{\sum_{i=1}^{n}\|\boldsymbol{u}_i-\widetilde{\boldsymbol{H}}_{\mathrm{F}}\boldsymbol{z}_i\|},\qquad i=1,2,\cdots,n. \tag{7}$$

证明：在 L-最优准则下，通过最小化 $\mathrm{tr}(\boldsymbol{D})$计算最优的子抽样概率，

$$\mathrm{tr}(\boldsymbol{D})=\tau(1-\tau)\mathrm{tr}\left[\frac{1}{r}\sum_{i=1}^{r}\frac{1}{\pi_i}(\boldsymbol{u}_i-\widetilde{\boldsymbol{H}}_{\mathrm{F}}\boldsymbol{z}_i)(\boldsymbol{u}_i-\widetilde{\boldsymbol{H}}_{\mathrm{F}}\boldsymbol{z}_i)^{\mathrm{T}}\right]=$$

$$\frac{\tau(1-\tau)}{r}\left(\sum_{i=1}^{r}\pi_i\right)\left(\sum_{i=1}^{r}\frac{\|\boldsymbol{u}_i-\widetilde{\boldsymbol{H}}_{\mathrm{F}}\boldsymbol{z}_i\|^2}{\pi_i}\right)\geqslant$$

$$\frac{\tau(1-\tau)}{r}\sum_{i=1}^{r}\|\boldsymbol{u}_i-\widetilde{\boldsymbol{H}}_{\mathrm{F}}\boldsymbol{z}_i\|,$$

其中，最后一步源于 Cauchy-Schwarz 不等式，当且仅当 $\pi_i\propto\|\boldsymbol{u}_i-\widetilde{\boldsymbol{H}}_{\mathrm{F}}\boldsymbol{z}_i\|$ 时等号成立. 证毕.

### 1.4　两步算法

根据定理 3 可知，最优子抽样概率 $\pi_i^{\mathrm{Lopt}}$ 是基于协变量的信息计算得出的，与响应变量 $y_i$ 无关. 该最优概率不仅适用于低维情形，也适用于高维情形. 式(7)中，最优抽样概率依赖于感兴趣协变量 $\boldsymbol{u}_i$、干扰协变量 $\boldsymbol{z}_i$及投影矩阵 $\widetilde{\boldsymbol{H}}_{\mathrm{F}}$. 由于该抽样概率不能直接得到，因此本文提出如下两步算法解决该问题.

**算法 1**　最优去相关得分子抽样算法.

步骤 1) 执行均匀子抽样以获取 $r_0$ 个子样本，基于该 $r_0$ 个样本估计式(7)中的子抽样概率. 对于子抽样概率中未知的 $\widetilde{\boldsymbol{H}}_{\mathrm{F}}$，由式(3)或式(6)计算得到. 替换原定理 3 中的 $\widetilde{\boldsymbol{H}}_{\mathrm{F}}$ 为 $\hat{\boldsymbol{H}}^*$，计算 L-最优性准则下的近似最优子抽样概率.

步骤 2) 根据步骤 1)中计算得到的子抽样概率，选取 $r$ 个子样本$\{y_i^*,\boldsymbol{u}_i^*,\boldsymbol{z}_i^*\}_{i=1}^r$. 基于上述子样本用式(4)求解方程 $\hat{\boldsymbol{\Psi}}_n^*(\hat{\boldsymbol{H}}^*,\hat{\boldsymbol{\gamma}}^*,\boldsymbol{\theta})=0$ 估计参数 $\boldsymbol{\theta}$.

## 2　模拟研究

下面利用数值模拟评估本文估计方法在有限样本容量下的性能，以验证去相关得分子抽样算法在实际应用中的可行性和准确性. 本文主要考察干扰参数的影响，分为低维($p=10$)和高维($p=700$)两种情形讨论.

由下式生成大小为 $n=10^5$ 的数据样本：

$$y_i=\sum_{j=1}^{3}u_{ij}\theta_j+\sum_{k=1}^{p-1}z_{ik}\gamma_k+\varepsilon_i,\qquad i=1,2,\cdots,n, \tag{8}$$

其中 $\boldsymbol{u}_i$ 表示低维感兴趣协变量，$\boldsymbol{z}_i$ 称为干扰协变量，二者皆源自多元正态分布，$(\theta_1,\theta_2,\theta_3)=(3,3,3)^{\mathrm{T}}$ 和 $\boldsymbol{\gamma}$ 分别为感兴趣低维参数和干扰参数，$p$ 表示干扰维数. 对式(8)的随机误差项 $\varepsilon_i$，考虑以下3 种分布类型：

误差 1) $\varepsilon_i$ 服从正态分布，$\varepsilon_i\sim N(0,1)$；

误差 2) $\varepsilon_i$ 服从自由度为 3 的 $t$ 分布，$\varepsilon_i\sim t(3)$；

误差 3) $\varepsilon_i$ 服从异方差正态分布，$\varepsilon_i=(1+2Z_{i2})Z_{i1}$，其中 $Z_{i1}\sim N(0,1)$，$Z_{i2}\sim$Bernoulli(0.5)，且 $Z_{i1}$和 $Z_{i2}$相互独立.

在产生随机数前，先对未知干扰参数向量 $\boldsymbol{\gamma}$ 设定一个真值，在低维情形下令$(\gamma_1,\gamma_2,\gamma_3,\cdots,\gamma_9)=(3,3,3,0,\cdots,0)$，在高维情形下令$(\gamma_1,\gamma_2,\gamma_3,\cdots,\gamma_{p-1})=(0,0,0,\cdots,0)$.

下面运行本文提出的两步算法，在两种干扰情形下算法 1 的步骤 1)中，均先选取子样本 $r_0=200$，以估计在 L-最优准则下的近似最优子抽样概率，再执行步骤 2)，分别考虑抽取 $r=200,400,600,800,1\,000$ 个样本. 完成抽样后，利用算法 1 对参数进行估计，并重复实验 $M=500$ 次，计算参数估计的平均值. 表 1 和表 2 分别列出了低维和高维情形下基于最优抽样所得估计参数的偏差(Bias)和标准差

(SD)(本文只列出了 $\tau=0.3$ 时的结果，且保留四位小数). 图1～图6分别为不同分位点处估计参数在两种干扰情形和3种不同误差下的总均方误差(MSE)，其中 $\mathrm{MSE}=\frac{1}{m}\sum_{m=1}^{M}\|\hat{\boldsymbol{\theta}}_U-\boldsymbol{\theta}_0\|^2$，$\hat{\boldsymbol{\theta}}_U$ 表示第 $m$ 个子抽样估计，$\boldsymbol{\theta}_0$ 表示参数的真值.

**表1 不同误差下低维分位回归估计参数的偏差与标准差**

**Table 1 Bias and standard deviations of low-dimensional quantile regression estimation parameters under different errors**

| $r$ | 参数 | 误差1) | | 误差2) | | 误差3) | |
|---|---|---|---|---|---|---|---|
| | | 偏差 | 标准差 | 偏差 | 标准差 | 偏差 | 标准差 |
| 200 | $\theta_1$ | 0.008 7 | 0.092 2 | −0.001 6 | 0.097 7 | 0.005 0 | 0.131 7 |
| | $\theta_2$ | −0.003 7 | 0.090 2 | −0.001 7 | 0.091 3 | −0.001 9 | 0.129 3 |
| | $\theta_3$ | −0.000 8 | 0.092 0 | 0.005 8 | 0.091 1 | −0.014 4 | 0.125 4 |
| 400 | $\theta_1$ | 0.000 9 | 0.062 2 | 0.001 3 | 0.065 4 | 0.005 3 | 0.090 5 |
| | $\theta_2$ | 0.004 1 | 0.060 6 | −0.002 4 | 0.063 7 | 0.000 9 | 0.087 5 |
| | $\theta_3$ | 0.005 3 | 0.059 5 | −0.003 9 | 0.064 7 | −0.000 7 | 0.091 0 |
| 600 | $\theta_1$ | 0.002 4 | 0.047 0 | 0.002 1 | 0.052 6 | 0.000 6 | 0.074 9 |
| | $\theta_2$ | −0.000 4 | 0.050 2 | −0.000 4 | 0.056 3 | 0.002 7 | 0.077 1 |
| | $\theta_3$ | −0.002 3 | 0.048 6 | 0.002 1 | 0.052 4 | −0.000 3 | 0.083 0 |
| 800 | $\theta_1$ | −0.000 6 | 0.043 4 | −0.001 7 | 0.044 8 | 0.000 5 | 0.064 6 |
| | $\theta_2$ | −0.000 1 | 0.043 4 | 0.004 3 | 0.045 0 | −0.001 0 | 0.064 2 |
| | $\theta_3$ | 0.003 5 | 0.041 6 | −0.001 9 | 0.047 5 | −0.003 6 | 0.063 5 |
| 1 000 | $\theta_1$ | −0.000 1 | 0.038 3 | 0.000 5 | 0.043 0 | 0.001 2 | 0.061 8 |
| | $\theta_2$ | 0.004 6 | 0.037 8 | 0.002 2 | 0.042 1 | 0.003 3 | 0.059 1 |
| | $\theta_3$ | 0.001 0 | 0.039 9 | 0.001 8 | 0.042 2 | −0.005 2 | 0.056 4 |

**表2 不同误差下高维分位回归估计参数的偏差与标准差**

**Table 2 Bias and standard deviations of high-dimensional quantile regression estimation parameters under different errors**

| $r$ | 参数 | 误差1) | | 误差2) | | 误差3) | |
|---|---|---|---|---|---|---|---|
| | | 偏差 | 标准差 | 偏差 | 标准差 | 偏差 | 标准差 |
| 200 | $\theta_1$ | −0.005 3 | 0.085 5 | 0.002 5 | 0.081 7 | 0.006 4 | 0.117 2 |
| | $\theta_2$ | −0.002 6 | 0.080 9 | −0.002 9 | 0.084 0 | 0.006 1 | 0.117 5 |
| | $\theta_3$ | 0.007 8 | 0.088 7 | 0.004 5 | 0.084 3 | 0.002 9 | 0.124 0 |
| 400 | $\theta_1$ | −0.001 4 | 0.057 1 | −0.002 0 | 0.056 7 | −0.010 7 | 0.080 9 |
| | $\theta_2$ | −0.001 6 | 0.058 3 | 0.001 2 | 0.059 4 | −0.003 9 | 0.089 6 |
| | $\theta_3$ | 0.004 8 | 0.060 8 | −0.001 6 | 0.059 0 | 0.004 7 | 0.084 2 |
| 600 | $\theta_1$ | 0.002 0 | 0.048 5 | 0.001 5 | 0.051 5 | −0.004 1 | 0.070 0 |
| | $\theta_2$ | −0.000 8 | 0.047 5 | 0.000 7 | 0.051 5 | 0.001 4 | 0.069 9 |
| | $\theta_3$ | −0.000 3 | 0.049 1 | 0.004 9 | 0.052 7 | −0.002 5 | 0.072 6 |
| 800 | $\theta_1$ | −0.000 1 | 0.041 6 | −0.002 3 | 0.045 7 | −0.002 2 | 0.063 7 |
| | $\theta_2$ | −0.001 7 | 0.039 5 | −0.000 9 | 0.045 9 | −0.006 3 | 0.063 1 |
| | $\theta_3$ | −0.000 5 | 0.042 6 | 0.000 4 | 0.046 3 | 0.005 3 | 0.066 4 |
| 1 000 | $\theta_1$ | 0.000 3 | 0.039 1 | 0.001 4 | 0.041 6 | −0.003 3 | 0.061 6 |
| | $\theta_2$ | −0.003 0 | 0.035 8 | −0.001 8 | 0.042 3 | 0.004 4 | 0.060 3 |
| | $\theta_3$ | −0.001 9 | 0.036 0 | 0.004 3 | 0.043 1 | −0.002 8 | 0.061 6 |

在两种不同干扰情形下，由表1和表2及图1～图6可见：本文提出的最优子抽样方法得到的每个估计参数的SD均随子样本量的增加而不断减小，说明该方法的估计性能随样本量的增加而变得更好，且估计结果是无偏的；在不同分位点 $\tau=0.3,0.5,0.7$ 时，所估计参数的MSE均随子样本的增加而逐渐减小，且本文提出的最优抽样方法得到估计的MSE均比基于均匀子抽样得到的MSE小，这与定理3最小化估计量 $\hat{\boldsymbol{\theta}}$ 的MSE理论结果一致. 模拟结果表明，本文提出的最优子抽样策略显著优于均匀子抽样.

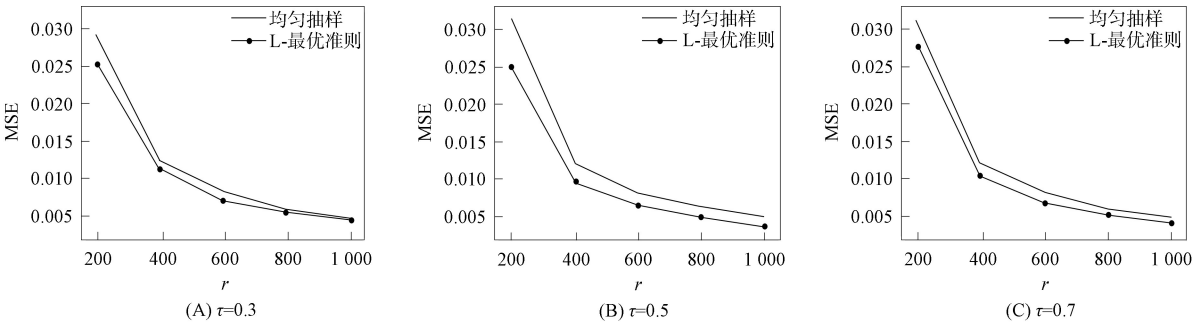


**图 1　当 $p=10$ 时，误差 1）下估计参数的 MSE**

**Fig. 1　MSE of estimated parameter under error 1）when $p=10$**

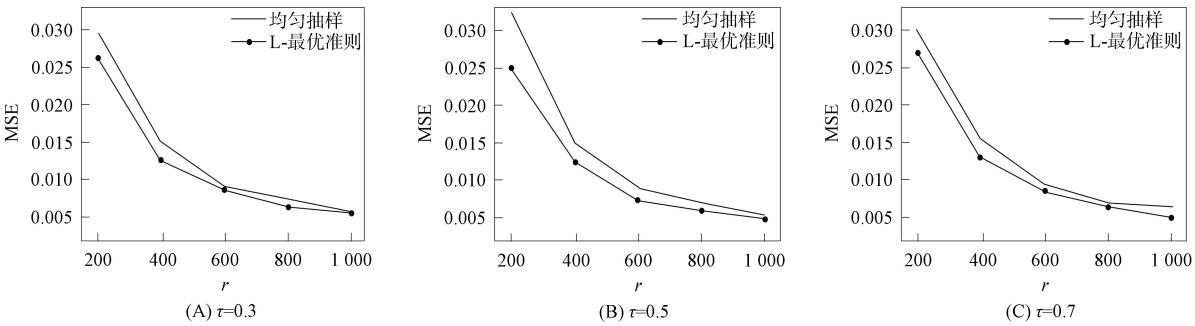


**图 2　当 $p=10$ 时，误差 2）下估计参数的 MSE**

**Fig. 2　MSE of estimated parameter under error 2）when $p=10$**

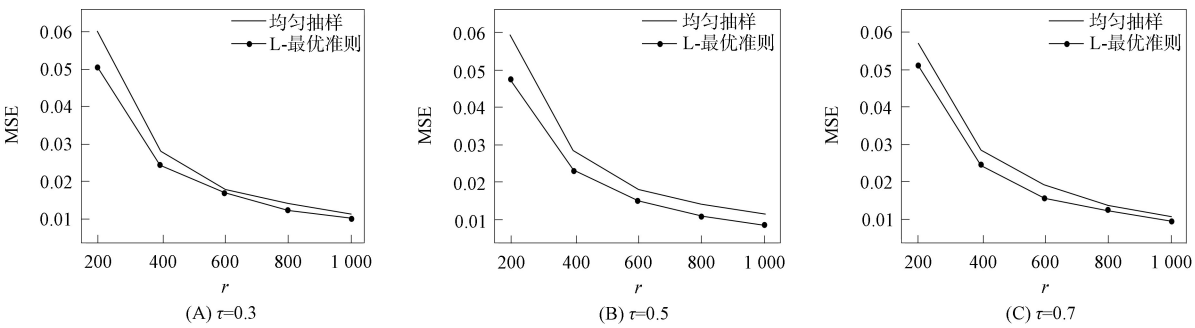


**图 3　当 $p=10$ 时，误差 3）下估计参数的 MSE**

**Fig. 3　MSE of estimated parameter under error 3）when $p=10$**

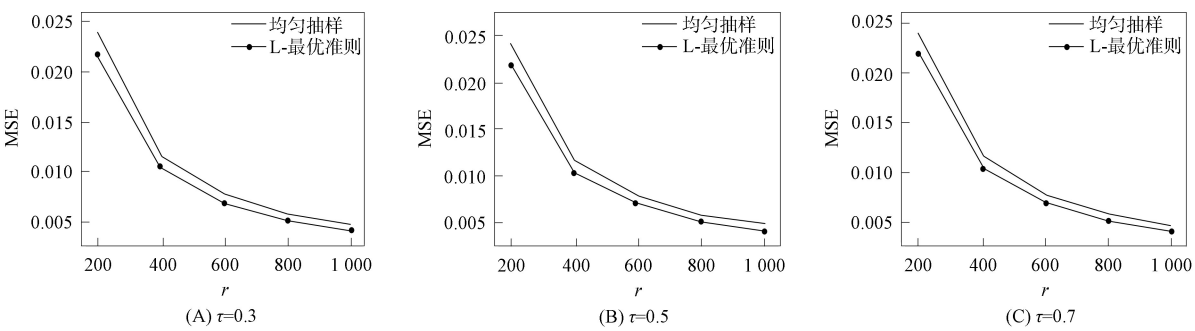


**图 4　当 $p=700$ 时，误差 1）下估计参数的 MSE**

**Fig. 4　MSE of estimated parameter under error 1）when $p=700$**

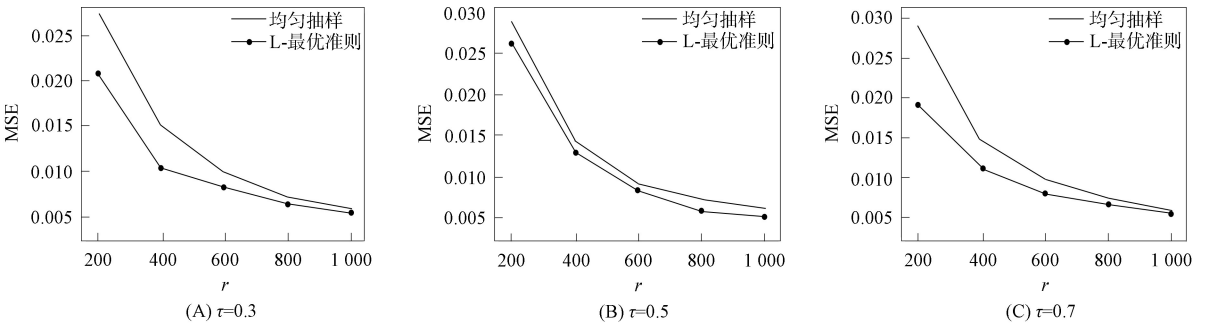


**图5 当 $p=700$ 时，误差2)下估计参数的MSE**

**Fig. 5 MSE of estimated parameter under error 2) when $p=700$**

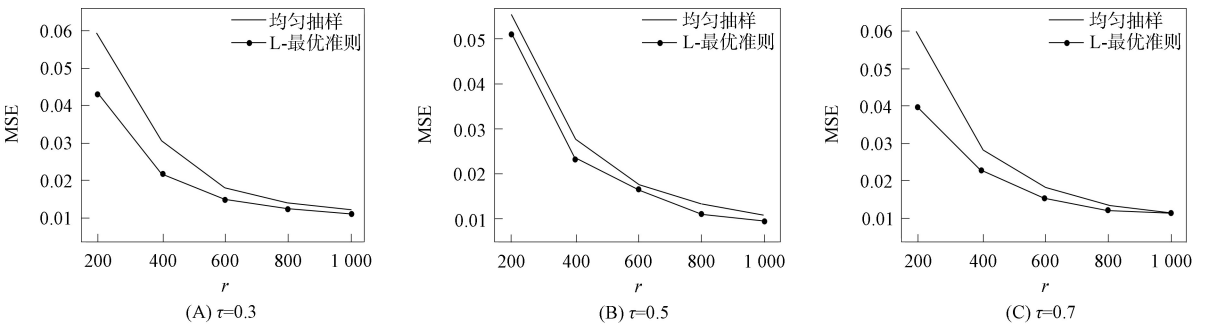


**图6 当 $p=700$ 时，误差3)下估计参数的MSE**

**Fig. 6 MSE of estimated parameter under error 3) when $p=700$**

## 3 实证分析

下面将本文提出的子抽样算法应用于来自UCI存储库的博客反馈数据集(https://archive.ics.uci.edu/ml/datasets/BlogFeedback). 该数据集收录了2010—2011年期间的博客数据，其中包含 $n=52\ 397$ 个样本和 $p=280$ 个协变量. 目标是预测给定博客的反馈数量与280个协变量之间的关系.

Wang等[23]分析该数据集的结果表明，博客的评论数($y$)主要受3个特定协变量(在过去24 h内对消息来源评论数量的中位数 $x_1$；消息来源在过去48～24 h内的评论数与在过去24 h内的评论数之间的平均差异 $x_2$；在过去24 h内对消息来源的评论数 $x_3$)的显著影响. 本文在Wang等[23]实证分析的基础上，在分位回归模型中添加23个对响应变量预测精准度较低的协变量作为干扰协变量，在进行数据分析前，先对响应变量和所有协变量进行标准化处理.

在 $\tau=0.5$ 分位点处，采用本文提出的算法对博客数据集进行建模分析. 设 $r_0=400$，$r=200,400,600,800$，重复计算500次并取均值. 由于在真实的数据场景中，通常无法直接获得模型参数的真实值，因此本文采取一种实用的替代方法：利用从全数据中得到的参数估计值替代未知的真实值. 考察上述3个特定的协变量，并在全数据的基础上对它们进行参数估计. 根据全数据下的分析，这3个协变量在全数据下的参数估计值分别为0.030 7,0.058 2,0.224 9. 该结果表明，响应变量 $\boldsymbol{y}$ 与这3个协变量之间均存在正向的关联性. 即这些协变量的增加倾向于与响应变量 $\boldsymbol{y}$ 的增加相关联，从而得到了对数据内在关系更深刻的理解. 表3列出了最优子抽样方法针对3个低维感兴趣协变量参数估计的Bias和SD值. 图7为这些协变量参数估计的MSE随子样本大小变化的趋势. 由表3可见，随着子样本量的增加，基于最优子抽样方法参数估计的标准差逐渐降低，该结果证实了所推导的渐近协方差矩阵在实际应用中的有效性. 由图7可见，无论哪种抽样方法，估计值的MSE均随子样本量的增加而减少. 此外，本文提出的最优子抽样策略得到的MSE始终低于均匀子抽样方法得到的MSE，该

结果进一步验证了最优子抽样策略在实际应用中的显著优势.

**表 3　当 $\tau=0.5$ 时，对博客数据集参数估计的偏差和标准差**

**Table 3　Bias and standard deviations of parameter estimates for blog dataset when $\tau=0.5$**

| 结果 | 参数 | $r$ | | | |
|---|---|---|---|---|---|
| | | 200 | 400 | 600 | 800 |
| 偏差 | $x_1$ | 0.014 5 | −0.035 5 | −0.044 1 | −0.051 3 |
| | $x_2$ | −0.008 5 | 0.000 7 | −0.007 0 | −0.013 0 |
| | $x_3$ | 0.095 8 | 0.116 4 | 0.124 5 | 0.126 9 |
| 标准差 | $x_1$ | 0.228 8 | 0.125 6 | 0.087 3 | 0.058 4 |
| | $x_2$ | 0.150 9 | 0.095 4 | 0.071 2 | 0.053 1 |
| | $x_3$ | 0.134 0 | 0.100 3 | 0.083 4 | 0.068 8 |

综上，本文将去相关得分方程推广到了高维分位回归模型的子抽样中，该方法可估计高维分位回归模型子抽样中的低维预测参数. 首先推导了一般去相关得分子样本估计量的渐近性质，然后根据 L-最优准则给出了最优子抽样概率，并提出了一种两步算法来近似最优的去相关得分子抽样概率. 为节约计算成本，在模拟实验中先固定算法第一步的较小子样本量 $r_0$，再逐步增加算法第二步的子样本量 $r$. 模拟研究结果表明，相比于均匀子抽样方法，本文方法优势显著. 最后，将本文方法应用于真实的博客数据集，实证结果表明，本文提出的最优子抽样策略可很好地在真实情形下估计感兴趣低维参数. 在实际应用中，推荐采用一步估计法，因为它能显著提高海量高维数据分析的计算效率，能更有效地处理大规模数据集.

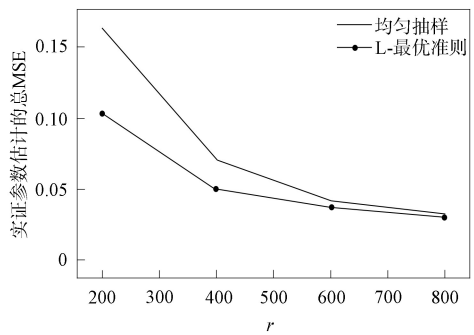


**图 7　当 $\tau=0.5$ 时，对博客数据集参数估计的 MSE**

**Fig. 7　MSE of parameter estimates for blog dataset when $\tau=0.5$**

## 参　考　文　献


[1] LIN N, XI R B. Aggregated Estimating Equation Estimation [J]. Statistics and Its Interface, 2011, 4(1): 73-83.

[2] CHEN X Y, XIE M G. A Split-and-Conquer Approach for Analysis of Extraordinarily Large Data [J]. Statistica Sinica, 2014, 24(4): 1655-1684.

[3] XU Q F, CAI C, JIANG C X, et al. Block Average Quantile Regression for Massive Dataset [J]. Statistical Papers, 2020, 61(1): 141-165.

[4] FITHIAN W, HASTIE T. Local Case-Control Sampling: Efficient Subsampling in Imbalanced Data Sets [J]. The Annals of Statistics, 2014, 42(5): 1693-1724.

[5] WANG Y, ZHU R, MA P. Optimal Subsampling for Large Sample Logistic Regression [J]. Journal of the American Statistical Association, 2018, 113: 829-844.

[6] YUAN X H, LI Y, DONG X G, et al. Optimal Subsampling for Composite Quantile Regression in Big Data [J]. Statistical Papers, 2022, 63(5): 1649-1676.

[7] JONES L H. Investigating the Properties of a Sample Mean by Employing Random Subsample Means [J]. Journal of the American Statistical Association, 1956, 51: 54-83.

[8] SCHIFANO E D, WU J, WANG C, et al. Online Updating of Statistical Inference in the Big Data Setting [J]. Technometrics, 2016, 58(3): 393-403.

[9] LUO L, ZHOU L, SONG P X K. Real-Time Regression Analysis of Streaming Clustered Data with Possible Abnormal Data Batches [J]. Journal of the American Statistical Association, 2023, 118: 2029-2044.

[10] MA P, MAHONEY W M, YU B. A Statistical Perspective on Algorithmic Leveraging [J]. Journal of Machine Learning Research, 2015, 16: 861-911.

[11] AI M, YU J, ZHANG H, et al. Optimal Subsampling Algorithms for Big Data Regessions [J]. Statistica Sinica,

2021, 31(2): 749-772.

[12] FAN Y, LIU Y K, ZHU L X. Optimal Subsampling for Linear Quantile Regression Models [J]. Canadian Journal of Statistics, 2021, 49(4): 1039-1057.

[13] 袁晓惠, 郭伟, 王纯杰. 大数据分位数回归下基于信息阵的最优子抽样 [J]. 东北师大学报(自然科学版), 2023, 55(3): 30-36. (YUAN X H, GUO W, WANG C J. Information Matrix Based Optimal Subsampling for Big Data Quantile Regression [J]. Journal of Northeast Normal University (Natural Science Edition), 2023, 55(3): 30-36.)

[14] WANG H Y, MA Y Y. Optimal Subsampling for Quantile Regression in Big Data [J]. Biometrika, 2021, 108(1): 99-112.

[15] GAO J Z, WANG L W, LIAN H. Optimal Decorrelated Score Subsampling for Generalized Linear Models with Massive Data [J]. Science China Mathematics, 2024, 67(2): 405-430.

[16] KOENKER R, BASSETT G, Jr. Regression Quantiles [J]. Econometrica, 1978, 46(1): 33-50.

[17] WANG J H, MENDEL F. Inference for Censored Quantile Regression Models in Longitudinal Studies [J]. The Annals of Statistics, 2009, 37(2): 756-781.

[18] 袁晓惠, 刘天庆. 协变量缺失下基于诱导光滑方法的加权分位数回归 [J]. 吉林大学学报(理学版), 2016, 54(6): 1314-1322. (YUAN X H, LIU T Q. Weighted Quantile Regression Based on Induced Smoothing Method with Missing Covariates [J]. Journal of Jilin University (Science Edition), 2016, 54(6): 1314-1322.)

[19] WANG X R, QIN G Y, SONG X Y, et al. Censored Quantile Regression Based on Multiply Robust Propensity Scores [J]. Statistical Methods in Medical Research, 2022, 31(3): 475-487.

[20] CHENG C, FENG X D, HUANG J, et al. Regularized Projection Score Estimation of Treatment Effects in High-Dimensional Quantile Regression [J]. Statistica Sinica, 2022, 32(1): 23-41.

[21] ZHANG C H, ZHANG S S. Confidence Intervals for Low Dimensional Parameters in High Dimensional Linear Models [J]. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2014, 76(1): 217-242.

[22] NING Y, LIU H. A General Theory of Hypothesis Tests and Confidence Regions for Sparse High Dimensional Models [J]. The Annals of Statistics, 2017, 45(1): 158-195.

[23] WANG L, ELMSTEDT J, WONG W K, et al. Orthogonal Subsampling for Big Data Linear Regression [J]. The Annals of Applied Statistics, 2021, 15(3): 1273-1290.

(责任编辑: 李 琦)