流数据下的在线可更新稳健期望分位数回归

胡冠浩[1]，姜荣[2]*

1.东华大学，数学与统计学院，上海 201600

2.上海对外经贸大学，统计与信息学院，上海 201620

DOI:10.61369/ASDS.12181

摘　　要： 随着实际数据分析中动态流式数据集的比例不断上升，流数据的应用场景正在不断拓宽。在海量数据下，尖峰厚尾数据分布占据重要比例，如何进行对应的稳健估计是非常有必要的。本研究提出基于Huber损失函数的稳健期望分位数回归方法，仅使用历史汇总统计量实现在重尾噪声下给出实时高效的稳健估计，并且在特定假设条件下建立估计量的渐近性质。模拟研究进一步验证，该方法在流式计算环境中处理大规模数据集时，具有稳健性和实时性。

关 键 词： 流数据；稳健期望回归；在线可更新估计；非对称损失函数；厚尾分布

Online Updatable Robust Expectile Regression with Streaming Data

Hu Guanhao[1], Jiang Rong[2]*

1. School of Mathematics and Statistics, Donghua University, Shanghai 201600

2. School of Statistics and Information, Shanghai University of International Business and Economics, Shanghai 201620

Abstract: With the continuous increase in the proportion of dynamic streaming datasets in practical data analysis, the application scenarios of streaming data are expanding. Under massive data environments, peaked and heavy-tailed data distributions occupy a significant proportion, making it necessary to develop corresponding robust estimation methods. This study proposes a robust expectile regression method based on the Huber loss function, which achieves real-time and efficient robust estimation under heavy-tailed noise through incremental parameter updates using only historical summary statistics, and establishes the asymptotic properties of the estimator under specific assumptions. Numerical experiments further verify that the proposed method exhibits robustness and real-time performance when processing large-scale datasets in streaming computation environments.

Keywords: streaming data; robust expectile regression; online renewable estimation; asymmetric loss function; heavy-tailed distribution

引言

随着数字基础设施的不断完善以及数据与各行业的深度融合，全球数据生成速率正以高速复合增长率突破物理存储边界，其中超过70%的新增数据以流式形态动态抵达。这种数据洪流在金融高频交易、工业物联网监测等场景中呈现出显著的重尾分布特征——据纽约证券交易所实测数据显示，约5%的极端值承载着超过80%的风险信息。同时，移动支付的吞吐量达到2Tbps，实时处理流数据成为关键问题。流数据（Streaming data）处理主要以在线更新算法为主。流数据作为动态时序数据序列，其本质为数据观测点按序到达且具有单次遍历性。此类数据可形式化表示为$\{D_t\}_{t=1}^{\infty}$，其中D_t表示第t个时刻到达的数据切片。在线更新算法的核心机制在于增量计算框架的构建——通过设计递推公式实现统计量的动态更新，例如均值估计量$\bar{D}_t=\frac{t-1}{t}\bar{D}_{t-1}+\frac{1}{t}D_t$，该公式仅需存储前序统计量而非完整历史数据[1]，有效满足流处理的有限存储约束。处理流数据的最严格的约束条件是，一旦处理了流数据中的元素，就必须将其丢弃或只有少量特定的数据元素才会被存储。同时流数据分析有时效性要求。针对流数据处理的上述特性，学界已发展出多类流式估计方法：Schifano等[2]基于线性模型提出滑动窗口参数更新策略，Wang等[3]则拓展至分位数回归框架，但二者均存在流的数量上限有所局限。Luo和Song[1]构建的广义线性模型迭代估计器，突破传统方法的流数量约束，其渐近相合性证明为实际工程应用提供了理论保障。因此，流数据的特性就是规模宏大、实时连续到达、极值不可预测，尤其是仅可处理一次的特性，在不访问历史数据的情况下，利用在线可更新框架进行统计推断。

作者简介：胡冠浩，东华大学数学与统计学院，硕士研究生，研究方向为大数据分析；

通讯作者：姜荣，上海对外经贸大学统计与信息学院，教授，硕士生导师，研究方向为大数据统计、在线数据统计分析、分布式计算、分位数回归、非参数/半参数模型。

在稳健回归方面，Huber回归方法具有对异常值非常稳健的特性[4]。这种方法最初由Huber在1964年提出，旨在解决传统最小二乘法对异常值敏感的问题。为了减少噪声的影响，Gupta 等[5]通过使用非对称 Huber 损失函数分析了稳健的正则化极限学习机框架。针对高维非光滑优化问题，Yi等[6]提出一种融合半光滑牛顿法与坐标下降策略的混合算法，适用于特定情况下的 Huber回归及分位数回归场景。在稳健回归方法研究中，为抵抗异常值或包含具有重尾分布的变量，Sun等[7]研究了自适应 Huber回归，揭示稳健性调节参数与样本规模、数据维度及高阶矩的关联，推导出具有维度自适应特性的误差上下界，实现高维数据的稳健估计与推理。俞搏[8]在 Huber损失函数的基础上，通过引入可调节的尾部衰减机制，实现对 Huber函数抗离群值能力的定向优化。Akkaya等[9]研究 Huber 损失函数极小化问题，研究控制全局最小值集稀疏性的正则化参数的选择巧妙应对高维数据噪声。潘莹丽等[10]结合分布式优化的思想，采用 Huber回归方法去消除异常值和厚尾变量的影响。具有无限方差的重尾噪声在实践中普遍存在。与分位数回归相比，期望回归涉及到最小化二次损失，并且可以对重尾响应敏感。为了解决这一限制，Man 等[11]提出了一种稳健回归方法，通过重构非对称最小二乘框架的权重分配机制，保留期望回归计算效率优势，引入基于残差分布自适应的调节参数，使模型对极端值的敏感度降低。其理论分析表明，该方法在响应变量仅需有限二阶矩存在的弱假设下，仍能保持估计量的 Oracle性质。该方法继承了期望回归的计算便捷性和统计效率，并且在重尾响应分布下几乎与分位数回归一样稳健。针对流数据处理，研究提出一种基于 Huber损失的在线可更新期望分位数估计方法，实现对分布尾部的非对称的探索。基于在线可更新机制，仅需历史数据的的历史统计量即可实现参数动态更新，满足流处理系统的实时性处理需求。

一、基于 Huber损失的稳健期望分位数回归

首先定义 $y\in\mathbb{R}$ 是响应变量，$x=\left(x_1,...,x_p\right)^T\in\mathbb{R}^p$ 是一个 p 维协变量，$\left\{D_1,D_2,...,D_b\right\}$ 为流数据集，其中 $D_t=\left\{\left(y_{ti},x_{ti}\right),i=1,...,n_t,t=1,...,b\right\},n_t=\left|D_t\right|$，其中 $N_b=\sum_{t=1}^{b}n_t;t=1,..,b$。观测值 $\left\{\left(y_i,x_i\right),i=1,...,n\right\}$ 独立同分布于总样本 (y,x)。对于给定的参数 $\tau\in(0,1)$,假定观测值来自以下线性回归模型：

$$y=x^T\beta_0+\varepsilon. \tag{1-1}$$

其中，T 表示转置，ε 是随机误差，$\beta_0=\left(\beta_1,...,\beta_p\right)^T$ 是 p 维回归变量，随 τ 的不同值而变化，从而提供了对给定 x 的 y 的整个条件分布的信息。

期望分位数回归的定义非对称平方损失函数 $L_\tau(u)$ 如下：

$$L_\tau(u)=\left|\tau-I(u\le0)\right|u^2=\begin{cases}\tau u^2, & u>0\\(\tau-1)u^2 & u\le0,\end{cases} \tag{1-2}$$

其中 τ 为取值介于0到 1 之间的非对称参数，依靠 τ 控制损失函数的非对称程度。与分位数回归相比，期望回归涉及到最小化二次损失，因此对重尾响应敏感。为了解决这一限制，Man等人提出了一种稳健期望分位数回归方法，该方法继承了期望分位数回归的计算便捷性和统计效率，并且在重尾响应分布下几乎与分位数回归一样稳健。

理论上，对于模型（1-1）中的真参数 β_0，采用稳健期望分位数回归方法求解真参数，对于观测值 $\left\{\left(y_i,x_i\right),i=1,...,n\right\}$，即求解如下方程最小化的根：

$$\hat{\beta}=\arg\min_{\beta}\sum_{i=1}^{n}L_{\tau,\gamma}\left(y_i-x_i^T\beta\right) \tag{1-3}$$

其中 $L_{\tau,\gamma}(u)=\left|\tau-I(u<0)\right|\cdot\ell_\gamma(u)$ 是损失函数，$I(v)$ 为示性函数，而 $\ell_\gamma(\cdot)$ 服从下文第2小节中的条件 C1。Huber损失 $\ell(u)=u^2/2\cdot I(|u|\le1)+|(u)-1/2|\cdot I(|u|>1)$ 也符合条件C1。

二、在线可更新方程

（一）全数据集 Oracle估计

首先给出 Oracle估计量的表达式。根据式（1-3），Oracle回归估计量 $\hat{\beta}_{N_b}$ 的表达式，模型中针对求解真参数向量 β_0 的估计，即求解如下方程最小化的根：

$$\hat{\beta}_{N_b}=\arg\min_{\beta}\frac{1}{N_b}\sum_{t=1}^{b}\sum_{i=1}^{n_t}L_{\tau,\gamma}\left(y_{ti}-x_{ti}^T\beta\right) \tag{2-1}$$

其中 $L_{\tau,\gamma}(u)=\left|\tau-I(u<0)\right|\cdot\ell_\gamma(u)$ 是损失函数，$I(v)$ 表示示性函数，而 $\ell_\gamma(\cdot)$ 服从C1。根据 Man等[11]，给出以下假设条件。

C1：损失函数 $\ell_\gamma(u)=\gamma^2\ell(u/\gamma)$ 和 $\ell(\bullet)$ 满足：$(i)\nabla\ell(0)=0$，对于任意的 $u\in\mathbb{R}$ 有 $\left|\nabla\ell(u)\right|\le\min\left(a_1,|u|\right)$；$(ii)\nabla^2\ell(0)=1$，对于任意的 $|u|\le a_3$ 有 $\nabla^2\ell(u)\ge a_2$；(iii) 对于任意 $u\in\mathbb{R}$，有 $|\nabla\ell(u)-u|\le u^2$，其中 a_1,a_2,a_3 都是正的常数。

C2：$\Sigma_t=E\left(x_{ti}x_{ti}^T\right)$ 是一个正定矩阵，有 $\lambda_u\ge\lambda_{\max}\left(\Sigma_t\right)\ge\lambda_{\min}\left(\Sigma_t\right)\ge\lambda_l\ge0$ 并存在 $c\ge1$，对所有的 $\delta\in\mathbb{R}$ 和 $t\ge0$，有 $P\left(\left|x^T\Sigma_t^{-\frac{1}{2}}\delta\right|\ge c\|\delta\|_2t\right)\ge2e^{-\frac{t^2}{2}}$。

C3：回归误差是独立的，$E\left(\varepsilon^2|x\right)\le\sigma_\varepsilon^2<\infty$，其中 $\sigma_\varepsilon^2=\max_{1\le j\le d}\sigma_{jj}$，并且 $E\left(w_\tau(\varepsilon)\cdot\varepsilon|x\right)=0$，其中 $w_\tau(u)=\left|\tau-I(u<0)\right|$。

假设条件C1、C2、C3成立可得：

$$\sqrt{N}\left(\hat{\beta}_n-\beta_0\right)\xrightarrow{d}N\left(0,J^{-1}E\left\{\xi^2(\varepsilon)xx^T\right\}J^{-1}\right) \tag{2-2}$$

其中 $J=E\left\{w_\tau(\varepsilon)xx^T\right\}$。

（二）在线可更新稳健期望分位数回归估计

在给出了流数据集的 Oracle估计的情况下，在流式数据集 $\{D_1,D_2,...,D_b\}$。根据（1-3）设定 $\hat{\beta}_t=\arg\min Q_{N_t}(\beta)$，其中

$$Q_{N_t}(\beta)=\frac{1}{n_t}\sum_{i=1}^{n_t}L_{\tau,\gamma}\left(y_{ti}-x_{ti}^T\beta\right) \tag{2-3}$$

其中，子集大小 n_t 远小于全集大小 N_b。因此，计算速度更加

快捷，内存占用更少。在$C1$、$C2$、$C3$下，局部估计$\hat{\beta}_t, t=1,...,b$也服从渐近正态分布

$$\sqrt{n_t}\left(\hat{\beta}_t-\beta_0\right)\xrightarrow{\mathrm{d}}N\left(0,J_t^{-1}E\left\{\xi^2(\varepsilon)x_tx_t^T\right\}J_t^{-1}\right),n_t\to\infty.\quad(2\text{-}4)$$

其中$J_t=E\left\{w_\tau(\varepsilon)x_tx_t^T\right\}$。由(2-4)式$\left\{\hat{\beta}_1,...,\beta_b\right\}$服从复合正态分布 $\left\{N\left(\beta_0,\frac{1}{n_1}J_1^{-1}E\left\{\xi^2(\varepsilon)x_{1i}x_{1i}^T\right\}J_1^{-1}\right),...,N\left(\beta_0,\frac{1}{n_b}J_b^{-1}E\left\{\xi^2(\varepsilon)x_{bi}x_{bi}^T\right\}J_b^{-1}\right)\right\}$.

则极大似然函数如下$L(\beta)=\prod_{t=1}^{b}\left(\frac{1}{\sqrt{2\pi}}\right)^p\left(\frac{1}{n_t}J_t^{-1}E\left\{\xi^2(\varepsilon)x_{ti}x_{ti}^T\right\}J_t^{-1}\right)^{\frac{1}{2}}\exp$

$$\left\{-\frac{n_t}{2}J_t^{-1}E\left\{\xi^2(\varepsilon)x_{ti}x_{ti}^T\right\}J_t^{-1}\left(\hat{\beta}-\beta\right)^T\left\{J_tE\left\{\xi^2(\varepsilon)x_{ti}x_{ti}^T\right\}^{-1}J_t\right\}\left(\beta_t-\beta\right)\right\}$$

由上式取对数可得

$\log\left(L(\beta)\right)=C-\frac{1}{2}\sum_{t=1}^{b}n_t\left(\hat{\beta}_t-\beta\right)^T\left\{J_tE\left\{\xi^2(\varepsilon)x_{ti}x_{ti}^T\right\}^{-1}J_t\right\}\left(\beta_t-\beta\right)$ 其中C为常数。

令上式最大化，可得到所需估计量，等价最小化下式损失

$$Q(\beta)=\sum_{t=1}^{b}n_t\left(\hat{\beta}_t-\beta\right)^T\left\{\hat{J}_t\hat{\Sigma}_t^{-1}J_t\right\}\left(\beta_t-\beta\right)\qquad(2\text{-}5)$$

令

$$\hat{\beta}_{N_b}^{OHER}=\left(\sum_{t=1}^{b}\hat{J}_t\hat{\Sigma}_t^{-1}J_t\right)^{-1}\left(\sum_{t=1}^{b}J_t\Sigma_t^{-1}J_t\beta_{t-1}\right),\ \mathrm{b}=2,3,...\qquad(2\text{-}6)$$

根据上式，只需要在每一个D_b中针对求解$J_t\Sigma_t^{-1}J_t$和$\hat{\beta}_t$。其中，$\hat{J}_t=\frac{1}{n_t}\sum_{i=1}^{n_t}w_\tau(\varepsilon_{ti})x_{ti}x_{ti}^T$、$\hat{\Sigma}_t=\frac{1}{n_t}\sum_{i=1}^{n_t}\hat{\xi}^2(\varepsilon_{ti})x_{ti}x_{ti}^T$，$\hat{\xi}(\varepsilon)=w_\tau(\varepsilon)\nabla\ell_\gamma(\varepsilon)$，$\ell_\gamma(u)=\gamma^2\ell(u/\gamma)$，$w_\tau(u)=\left|\tau-I(u<0)\right|$。取Huber损失函数，$\ell(u)=u^2/2\cdot I(|u|\le1)+\left(|u|-1/2\right)\cdot I(|u|>1)$。则$L_{\tau,\gamma}(u)=\left|\tau-I(u<0)\right|\cdot\ell_\gamma(u)$，则

$$L_{\tau,\gamma}(u)=\left|\tau-I(u<0)\right|L_\gamma(u)=\left|\tau-I(u<0)\right|\left\{\frac{u^2}{2}I(|u|\le\gamma)+\left(\gamma|u|-\frac{\gamma^2}{2}\right)I(|u|>\gamma)\right\}.$$

具体算法流程可见表1。

表1 在线更新算法

算法：在线更新算法	
1	输入：子集$D_t=\left\{\left(y_{ti},x_{ti}\right),i=1,...,n_t,t=1,...,b\right\},N,b,\tau,\gamma.$
2	初始：初始值 $\mathrm{U}_t=0_{P\times P},V_t=0$
3	for $t=1,2,\ldots,b$ do
4	读取数据集 D_t 取数据，并由（2-6）式计算$\hat{\beta}_t$，和
5	$\mathrm{U}_t=\mathrm{U}_t+\hat{J}_t\hat{\Sigma}_t^{-1}J_t$，$V_t=V_t+\hat{J}_t\hat{\Sigma}_t^{-1}J_t\hat{\beta}_t$
6	end
7	输出：$\hat{\beta}_{N_b}^{OHER}=U_b^{-1}V_b$

三、大样本性质

为了证明可更新在线稳健回归估计的理论性质，在$C1$、$C2$、$C3$的基础上，根据Wang等人[12]增加一些条件：

$C4$：设$n=\frac{N_b}{b}$为每个子集的平均大小。要求$\frac{n}{\sqrt{N_b}}\to\infty$，并且所有的子集大小$n_t$都以$\mathrm{o}(n)$阶发散，即存在正的常数$c_1$、$c_2$，有$c_1\le\min_t\frac{n_t}{n}\le\max_t\frac{n_t}{n}\le c_2$。

$C5$：存在正定矩阵$\Sigma_1,\Sigma_2\ \ldots\ \Sigma_b$满足$\frac{1}{n_t}\sum_{i=1}^{n_t}x_{ti}x_{ti}^T\xrightarrow{P}J_t^{-1}\Sigma_tJ_t^{-1}$。

定理1在$C1-C5$的条件下，在线可更新稳健期望分位数估计量的渐近分布如下：

$$\left[\Phi(f)\right]^{\frac{1}{2}}\sqrt{N_b}\left(\hat{\beta}_{N_b}^{OHER}-\beta_0\right)\xrightarrow{\mathrm{d}}N\left(0,I_p\right)$$

且$\Phi(f)=J_t^{-1}E\left\{\xi^2(\varepsilon)x_{ti}x_{ti}^T\right\}J_t^{-1}$。

证明：对于局部稳健的估计量$\hat{\beta}_t$，其服从渐近条件如下：

$$\hat{\beta}_t-\beta_0=J_t^{-1}\frac{1}{n_t}\sum_{i=1}^{n_t}\xi(\varepsilon_i)x_{ti}+O_p\left(\frac{1}{n_t}\right)$$

根据方程（2-6），我们可以得到$\hat{\beta}_{N_b}^{ohr}$和$\hat{\beta}_t$的关系是如下所示：$\hat{\beta}_{N_b}^{OHER}=\left(\sum_{t=1}^{b}\hat{J}_t\hat{\Sigma}_t^{-1}J_t\right)^{-1}\left(\sum_{t=1}^{b}J_t\Sigma_t^{-1}J_t\beta_t\right)=\left(\sum_{t=1}^{b}w_t\frac{\hat{J}_t\hat{\Sigma}_t^{-1}J_t}{n_t}\right)^{-1}\left(\sum_{t=1}^{b}w_t\frac{J_t\Sigma_t^{-1}J_t}{n_t}\beta_t\right)$，其中$w_t=\frac{n_t}{N_b}$，则直接计算可得：

$$\sqrt{N_b}\left(\hat{\beta}_{N_b}^{OHER}-\beta_0\right)=\left(\sum_{t=1}^{b}w_t\frac{\hat{J}_t\hat{\Sigma}_t^{-1}J_t}{n_t}\right)^{-1}\left(\sqrt{N_b}\sum_{t=1}^{b}w_t\frac{J_t\Sigma_t^{-1}J_t}{n_t}\left(\beta_t-\beta_0\right)\right)$$

且$\sqrt{N_b}\left(\sum_{t=1}^{b}w_t\left(\frac{\hat{J}_t\hat{\Sigma}_t^{-1}J_t}{n_t}-J_t\Sigma_t^{-1}J_t\right)\left(\hat{\beta}_t-\beta_0\right)\right)=O_p\left(\frac{b}{\sqrt{N_b}}\right)$和由正则性条件$C1-C5$，可以获得

$$\begin{aligned}&\sqrt{N_b}\left(\sum_{t=1}^{b}w_t\frac{\hat{J}_t\hat{\Sigma}_t^{-1}J_t}{n_t}\left(\hat{\beta}_t-\beta_0\right)\right)\\&=\sqrt{N_b}\left(\sum_{t=1}^{b}w_t\Sigma_t\left(\hat{\beta}_t-\beta_0\right)\right)+\sqrt{N_b}\left(\sum_{t=1}^{b}w_t\left(\frac{\hat{J}_t\hat{\Sigma}_t^{-1}J_t}{n_t}-\Sigma_t\right)\left(\beta_t-\beta_0\right)\right)\\&=\sqrt{N_b}\left(\sum_{t=1}^{b}w_t\hat{J}_t\hat{\Sigma}_t^{-1}J_t\left(\frac{1}{E\left[w_\tau(\varepsilon)\right]}\Sigma_t^{-1}\left\{\sum_{i=1}^{n_t}\xi(\varepsilon_i)x_{ti}+O_p\left(\frac{1}{n_t}\right)\right\}\right)\right)+O_p\left(\frac{b}{\sqrt{N_b}}\right)\\&=\frac{1}{\sqrt{N_b}}\sum_{t=1}^{b}\frac{1}{E\left[w_\tau(\varepsilon)\right]_t}\sum_{i=1}^{n_t}x_{ti}\xi(\varepsilon_i)+O_p\left(\frac{b}{\sqrt{N_b}}\right)\end{aligned}$$

当$\sum_{t=1}^{b}w_t=1$，则$\sum_{t=1}^{b}w_t\frac{\hat{J}_t\hat{\Sigma}_t^{-1}J_t}{n_t}-\sum_{t=1}^{b}w_tJ_t\Sigma_t^{-1}J_t=\sum_{t=1}^{b}w_t\left(\frac{J_t\Sigma_t^{-1}J_t}{n_t}-J_t\Sigma_t^{-1}J_t\right)=O_p(1)$

由此可得

$$\begin{aligned}&\left[\Phi(f)\right]^{\frac{1}{2}}\sqrt{N_b}\left(\hat{\beta}_{N_b}^{OHER}-\beta_0\right)\\&=\left[\Phi(f)\right]^{\frac{1}{2}}\left(\sum_{t=1}^{b}w_t\frac{\hat{J}_t\hat{\Sigma}_t^{-1}J_t}{n_t}\right)^{-1}\left(\sqrt{N_b}\sum_{t=1}^{b}w_t\frac{J_t\Sigma_t^{-1}J_t}{n_t}\left(\hat{\beta}_t-\beta_0\right)\right)\\&=\left[\Phi(f)\right]^{\frac{1}{2}}\left(\sum_{t=1}^{b}w_t\frac{J_t\Sigma_t^{-1}J_t}{n_t}\right)^{-1}\left(\frac{1}{\sqrt{N_b}}\sum_{t=1}^{b}\frac{1}{E\left[w_\tau(\varepsilon)\right]_t}\sum_{i=1}^{n_t}x_{ti}\xi(\varepsilon_i)+O_p\left(\frac{b}{\sqrt{N_b}}\right)\right)\\&\xrightarrow{d}N\left(0,I_p\right)\end{aligned}$$

如果每次数据流中的协变量是齐次的，即$\Sigma_1=\Sigma_2=...=\Sigma_b=\Sigma$，$\Phi(f)=J_t^{-1}E\left\{\xi^2(\varepsilon)x_{ti}x_{ti}^T\right\}$，这与方程(2-1)中Oracle估计量的渐近方差一致。可更新估计量$\hat{\beta}_{N_b}^{OHER}$与Oracle估计量$\hat{\beta}_{N_b}$渐近等价。

四、数值实验

在本节中，使用蒙特卡罗模拟研究来评估所提出方法的有效性能。所有程序都是用R代码编写的。实验围绕流数据处理展开，针对每一批到达数据，通过在线可更新稳健期望分位数回归估计与全数据集估计进行比较。从以下线性模型生成数据：

$$Y_i=X_i^T\beta_0+\left\{\varepsilon_i-e(\tau)\right\},i=1,\ldots,N_b,$$

其中$e(\tau)=min_e\sum_{i=1}^{n}L_{\tau,\gamma_t}\left(Y_i-e\right)$是为了消除估计器对$\tau$的影响，使得在不同的$\tau$值下真值为$\beta_0$。参数的真值为$\beta_0=(1,1,0,2)$ $X_i=[1,X_{i1},X_{i2},X_{i3},X_{i4}]^T$，$X_{i1},X_{i2},X_{i3},X_{i4}$服从多重正态分布$N[0,\Sigma]$，其中协方差矩阵$\Sigma$是由$\Sigma_{ij}=0.5^{i-j}$，$1\le i,j\le4$构成的。根据Man等人[11]的参数研究，取总样本参数$\gamma=\sqrt{N_b/\left(\mathrm{p}+\log(N_b)\right)}$和子集样本参数$\gamma_t=\sqrt{n_t/\left(\mathrm{p}+\log(n_t)\right)}$。固定样本子集大小$n_t=500$，并改变机器数量$K$，然后得到总样本大小$n=500K$.误差$\varepsilon_i$独立于以下三个分布之一产生：（i）$\varepsilon_i\sim N(0,1)$，（ii）$\varepsilon_i\sim0.85N(0,1)+0.15N(0,4)$，

（iii） $\varepsilon_i \sim 0.85N(0,1)+0.15N(4,8)$。

为了评估提出方法的性能，通过计算了均方差（MSE）和计算时间（以秒为单位）。表2和表3给出了基于100次仿真模拟的分位数水平τ为0.2，0.5和0.8的结果。

表2 不同K和τ下对应MSE均值

Error	τ	$D=10$	$D=20$	$D=50$	$D=100$	$D=200$
(i)	0.2	0.110	0.108	0.110	0.113	0.116
	0.5	0.107	0.102	0.098	0.102	0.101
	0.8	0.113	0.116	0.107	0.109	0.109
(ii)	0.2	0.117	0.115	0.110	0.114	0.121
	0.5	0.110	0.114	0.109	0.100	0.106
	0.8	0.112	0.108	0.115	0.115	0.105
(iii)	0.2	0.156	0.162	0.164	0.152	0.163
	0.5	0.148	0.158	0.150	0.141	0.149
	0.8	0.153	0.159	0.146	0.152	0.158

表3 不同K和τ下对应计算累计的时间（10^{-2}秒）

Error	τ	Method	$D=10$	$D=20$	$D=50$	$D=100$	$D=200$
(i)	0.2	All	3.158	5.83	17.468	41.65	80.968
		OHR	0.517	0.494	0.621	0.904	0.905
	0.5	All	4.519	7.775	18.267	33.121	63.018
		OHR	0.737	0.875	0.762	0.76	0.765
	0.8	All	4.749	9.496	22.363	41.797	80.626
		OHR	0.840	0.713	0.840	0.765	0.813
(ii)	0.2	All	5.240	8.645	22.612	39.706	76.955
		OHR	0.793	0.743	0.869	0.842	0.865
	0.5	All	4.506	7.495	17.795	32.365	62.381
		OHR	0.698	0.820	0.890	0.709	0.818
	0.8	ALL	5.269	9.400	21.728	40.563	78.291
		OHR	0.794	0.778	0.750	0.769	0.794
(iii)	0.2	All	3.035	5.842	15.422	36.817	72.718
		OHR	0.520	0.489	0.620	0.784	0.763
	0.5	All	4.204	8.072	17.581	32.545	62.067
		OHR	0.831	0.735	0.718	0.898	0.743
	0.8	All	4.933	8.534	9.592	37.112	72.652
		OHR	0.703	0.757	0.882	0.735	0.898

从表2和表3可以得出以下结论：

1.关于表2中的MSE，所有估计量MSE的结果都相对非常小，即是在重尾噪声下，所对应的均方误差依然相对较小，估计值都接近真值，因此提出的在线估计量具有稳健性。

2.关于表3中的计算时间，注意到，对于任意给定数量的机器和分位数τ，提出的估计量（OHER)比全数据集更快，当全数据集所花时间不断上升时，可更新处理的时间变化极低并维持在极短时间内，在单位10^{-2}秒下，流数据所耗费时间全部达到毫秒级满足实时处理要求，并且随着流的数目从$D=10$到$D=200$，处理时间差异远小于流的增加量，验证随着流数的增加，流数据实时处理的仍具有可行性。

五、结束语

本研究提出基于Huber损失函数的稳健期望分位数回归的在线可更新估计方法，仅使用历史汇总统计量实现参数增量更新，在重尾噪声下给出实时高效的稳健估计。在满足特定假设条件下，建立估计量的渐近性质。通过数值实验进一步验证，该方法在流式计算环境中处理大规模数据集时，该方法在保证统计效率的同时，显著提升了计算时效性。

参考文献

[1]Luo L, Song P. Renewable estimation and incremental inference in generalized linearmodels with streaming data sets[J]. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2020,82(1), 69 - 97.

[2]Schifano E D, Wu J, Wang C, Yan J, Chen M H. Online updating of statistical inference in the big data setting[J].Technometrics, 2016, 58(3):393-403.

[3]Wang K, Wang H, Li S. (2022). Renewable quantile regression for streaming datasets. Knowledge-Based Systems, 235:107675.

[4]Huber P J. Robust estimation of a location parameter[J]. Annals of Mathematical Statistics, 1964,35 (1): 73-101.

[5]Gupta D, Hazarika B B, Berlin M. Robust regularized extreme learning machine with asymmetric Huber loss function[J]. Neural Computing and Applications. 2020,32 (16): 12971 - 12998.

[6]Yi C, Huang J. Semismooth Newton coordinate descent algorithm for elastic-net penalized Huber loss regression and quantile regression[J]. Journal of Computational and Graphical Statistics, 2017, 26(3): 547-557.

[7]Sun Q, Zhou W X. Fan J. Adaptive Huber regression[J]. Journal of the American Statistical Association, 2020, 115(529): 254-265.

[8]俞搏天 . p-Huber损失函数及其鲁棒性研究 [D]; 浙江师范大学 , 2021.

[9]Akkaya D, Pınar M Ç. Minimizers of sparsity regularized huber loss function[J]. Journal of Optimization Theory and Applications. 2020,187(1): 205-233.

[10]潘莹丽 , 刘展 , 朱千慧子 . 大数据背景下基于 Huber回归模型的分布式优化方法研究 [J]. 数理统计与管理 , 2022, 41(04): 633-646.

[11]Man R, Tan K, Wang Z, Zhou W. Retire: Robust expectile regression in high dimensions[J]. Journal of Econometrics, 2023,239 (2).

[12]Wang K, Wang H, Li S. Online renewable quantile regression and penalized learning for streaming datasets[J]. IEEE Transactions on Neural Networks and Learning Systems,2021,34(5),2341 - 2352.

