文章编号: 1001-4543(2024)02-0208-10 DOI: 10.19570/j.cnki.jsspu.2024.02.013

流数据下的复合分位数回归

韩星敏[1], 姜 荣[2]

(1. 东华大学 理学院, 上海 201600; 2. 上海第二工业大学 数理与统计学院, 上海 201209)

摘 要: 随着互联网的发展, 数据规模急剧增长, 但有限的内存只能存储一小批数据, 因此在不访问历史数据的情况下进行分析是非常有必要的, 流数据分析也因而引起了广泛关注。同时复合分位数回归因其鲁棒性和全面性, 在许多领域得到应用, 但由于传统复合分位数回归是基于内存可容纳完整数据的条件, 因此在流数据环境中实现复合分位数回归是非常有挑战的。针对流数据提出了一种可更新的复合分位数回归方法, 可以随着数据的到达, 使用当前数据和历史数据的汇总统计量来更新估计量。在理论上证明提出的可更新估计量与使用完整数据得到的估计量是渐近等价的。最后通过模拟研究验证了所提出方法的有效性。

关键词: 复合分位数回归; 流数据; 在线可更新估计方程

中图分类号: O212.2 **文献标志码**: A

Renewable Composite Quantile Regression for Streaming Data Sets

HAN Xingmin[1], JIANG Rong[2]

(1. College of science, Donghua University, Shanghai 201600; 2. School of Mathematics, Physics and Statistics, Shanghai Polytechnic University, Shanghai 201209)

Abstract: With the development of the Internet, the scale of data has grown dramatically. However, due to limited memory capacity that can only store a small batch of data, it is essential to analyze data without accessing historical data. Consequently, streaming data analysis has attracted widespread attention. Meanwhile, composite quantile regression, known for its robustness and and comprehensiveness, has been applied in various fields. However, implementing composite quantile regression for streaming data is challenging since traditional methods are based on the condition that the entire dataset can fit into memory. An updating composite quantile regression method specifically is designed for streaming data. The estimates can be updated as the data arrives using both current data and the summary statistics of historical data. In theory, it is proven that the updatable estimator proposed is asymptotically equivalent to the estimator obtained using complete data. Finally, the effectiveness of the proposed method is verified through simulation research.

Keywords: composite quantile regression; streaming data; online updating estimating equation

0 引言

近些年, 随着网络科技的不断发展, 信息化进程加快演进, 网络计算技术、网络储存技术等逐渐补充完善。生活中越来越多的行为和属性被数字化, 与以往少量且固定的传统数据不同, 这些日常生活的数据被动态生成, 从而形成无限且不断增长的数据集, 即流数据。近些年流数据的概念也被各行各业广泛提起, 具有较大的研究价值和应用价值。与传统数据相比, 流数据主要具备以下几个特点 [1]: ① 实时性。流数据是实时发生的, 速度快且短暂易逝, 因此必须要实时处理和分析, 以便及时发现和处理问题。② 大量性。流数据通常是海量的, 其数据规模随着时间推移, 持续不断地增加, 没有上限且无

收稿日期: 2023-12-05

通信作者: 姜 荣 (1986–), 男, 上海人, 教授, 博士, 主要研究方向为大数据统计分析。E-mail: jiangrong@sspu.edu.cn

法预知其规模大小。③ 无序性。分批到达的数据之间相互独立, 不被应用系统所控制。④ 单遍历性。即数据一旦被经过处理, 如果没有存档, 则将会被抛弃。因此数据不能轻易找回再次计算, 只会经历单遍处理。

将统计模型应用到流数据中, 不仅会因为数据量大而对计算机内存造成压力, 而且会对计算效率造成压力。在文献 [2] 中, 把大数据的统计方法归纳为子采样算法 [3-5]、分治法 [6-7] 和在线更新方法 [8-11]。其中在线更新方法主要针对数据以流的形式到达的问题, 即可以使估计量随数据到达而更新。如何在不需要存储历史原始数据的情况下进行统计推断是在线更新方法解决的主要问题, 有大量的学者对在线更新方法进行了研究。为线性模型提出的在线迭代估计算法 [8], 可以使估计量随数据到达而更新, 通过检验新算法的拟合优度和所提出的估计量的理论性质, 证明了新算法的有效性, 但此方法需要满足一个强正则性条件, 即数据集批次 b 需满足 $b = O(n_j^c)$, 其中 $c < 1/3$, n_j 为第 j 批到达的数据集 D_j 的样本量。而广义的线性模型的可更新估计 (renewable least squares, RLS)[9] 克服了上述的严格限制。还有为高维线性模型提出的在线估计推理过程[10], 可以随着样本量的增加递归地进行模型选择, 并基于选择的变量建立模型。针对流数据的分位数回归参数估计和变量选择的可更新估计方法的提出 [11-12], 将在线更新方法进一步拓展, 并证明可更新估计量与利用全部数据进行分位数回归的估计结果具有相同的渐近正态性, 数值模拟实验结果也说明两者的精确性非常接近。

最小二乘回归 (ordinary least squares, OLS) 是一种拟合线性回归的常用方法, 又称为条件均值回归, 但均值不能反映一个分布的全部状况。另外, OLS 的估计结果只有在高斯噪声和轻尾噪声下才具有统计意义。与 OLS 不同, 分位数回归 [13] (quantile regression, QR) 根据因变量的条件分位数对自变量进行回归, 能够获得整个条件分布函数的特征, 对异常值具有鲁棒性, 但 QR 每次仅考虑一个分位数水平, 这可能会导致信息缺乏, 降低估计效率。基于此, 有学者提出了复合分位数回归 [14] (composite quantile regression, CQR), 通过对分位数回归损失函数进行优化以同时考虑多个分位数水平, 在保留分位数回归鲁棒性的同时, 又能够显著提高估计效率。该模型自建立以来受到许多学者的研究。譬如, 将 CQR 应用到变系数部分线性模型 [15]、单指标模型 [16] 和部分线性单指标模型 [17] 等。尽管 CQR 方法很受关注, 但其损失函数的不可微性也给直接处理大规模数据集带来了挑战。为了解决这一问题, 有学者使用取值范围为 $[0, 1]$ 的平滑核函数取代分位数损失函数的示性函数 [18], 使 CQR 损失函数转为平滑的。卷积平滑分位数回归模型 [19] (smoothing quantile regression, SQR) 的提出更是得到了一个二次可微的损失函数, 且被验证为是一种可靠的 QR 分析工具 [20], 被扩展到 CQR[21]。

CQR 在实际应用中也获得广泛关注, 例如利用 CQR 研究热轧带钢的抗拉强度和各生产参数与工艺参数之间的关系 [22], 预测新冠疫情背景下我国进出口贸易总额 [23], 对隧道围岩变形进行预测 [24] 等。但是目前的 CQR 都应用于固定数据集, 在流数据中的应用研究不足。

综上所述, 本文为 CQR 开发了一种可更新的估计算法, 该算法可以被应用到流数据中, 即可仅使用当前到达的原始数据和历史数据的汇总统计量来更新估计量, 大大减少数据存储规模。另外, 不仅从理论上证明本文所提出的估计量与使用全部数据得到的统计量是渐近等价, 也通过数值模拟和实例实验验证了本文提出的可更新估计量的可靠性。

1　复合平滑分位数回归

首先简要介绍用于分位数回归的卷积型平滑技术, 然后将该方法应用到 CQR 模型。

1.1　平滑分位数回归

假设 $Y \in R$ 是一个随机响应变量, $\boldsymbol{X} \in R^p$ 是一个 p 维协变量, $\{y_i, \boldsymbol{x}_i\}_{i=1}^n$ 是从以下线性回归模型中得到的 n 个独立同分布的观测值:

$$Y = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}_0 + \varepsilon \tag{1}$$

式中: $\boldsymbol{\beta}_0 \in R^p$ 为未知参数; ε 为随机误差项。假设 ε 与 $\boldsymbol{X}$ 无关, 且具有未知累计分布函数 $F_\varepsilon(\cdot)$。则对于任意 $\tau \in (0, 1)$, $Y|\boldsymbol{X}$ 的 τ 条件分位数为:

$$F^{-1}_{Y|\boldsymbol{X}}(\tau) = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}_0 + F^{-1}_{\varepsilon}(\tau) \tag{2}$$

则上述假设式 (1) 等价于:

$$\begin{aligned} &Y = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}_0 + \varepsilon, \\ &P(\varepsilon \leqslant 0|\boldsymbol{X}) = P(\varepsilon \leqslant 0) = \tau \end{aligned} \tag{3}$$

式中, 一般情况下假设 τ 条件分位数为 0, 即 $b_\tau := F_\varepsilon^{-1}(\tau) = 0$。

标准分位数回归估计量 [13] 为:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}\in R^p} \hat{Q}(\boldsymbol{\beta}) := \arg\min_{\boldsymbol{\beta}\in R^p} \frac{1}{n}\sum_{i=1}^{n} \rho_\tau(y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}) \tag{4}$$

式中: $\rho_\tau(u) = u\tau - uI(u<0)$ 为分位数损失函数; $I(u<0)$ 为示性函数。对于任意 $\boldsymbol{\beta}\in R^p$, 令 $F_n(;\boldsymbol{\beta})$ 为残差 $\{y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}\}_{i=1}^n$ 的经验累积分布函数 (cumulated distribution function, CDF), 即对于任意 $u\in R$, 都有 $F_n(u;\boldsymbol{\beta}) = \dfrac{1}{n}\sum_{i=1}^{n} I(y_i - x_i^{\mathrm{T}}\boldsymbol{\beta} \leqslant u)$。因此 $\hat{Q}(\cdot)$ 可以被等价写为 [13]:

$$\hat{Q}(\boldsymbol{\beta}) = \int_{-\infty}^{+\infty} \rho_\tau(u)\mathrm{d}F_n(u;\boldsymbol{\beta}) \tag{5}$$

值得注意的是, 由于 $\hat{Q}(\cdot)$ 实际上是在不同点处求损失函数的有限和, 因此 $\hat{Q}(\cdot)$ 是非光滑的。而卷积平滑核函数的引入解决了此问题, 即提出了一种核型 CDF 作为新的积分测度 [19]。具体来讲, 对于一系列的残差 $\{y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}\}_{i=1}^n$ 和带宽参数 h, 核型 CDF $\bar{F}_n(u;\boldsymbol{\beta})$ 可以被定义为:

$$\bar{F}_n(u;\boldsymbol{\beta}) = \frac{1}{nh}\sum_{i=1}^{n}\int_{-\infty}^{u} K\left(\frac{t + \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta} - y_i}{h}\right)\mathrm{d}t = \frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{u} K_h(t + \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta} - y_i)\mathrm{d}t \tag{6}$$

式中: $K(\cdot)$ 为对称非负核函数; $K_h(u) = K(u/h)/h$。

因此对 $\bar{F}_n(u;\boldsymbol{\beta})$ 积分后, 可以得到相应的分位数损失函数的经验风险表达式 $\hat{Q}_h$ 为:

$$\hat{Q}_h(\boldsymbol{\beta}) = \int_{-\infty}^{+\infty} \rho_\tau(u)\mathrm{d}\bar{F}_n(u;\boldsymbol{\beta}) = \frac{1}{nh}\sum_{i=1}^{n}\int_{-\infty}^{+\infty} \rho_\tau(u)K\left(\frac{u + \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta} - y_i}{h}\right)\mathrm{d}u = \frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{+\infty} \rho_\tau(u)K_h(u + \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta} - y_i)\mathrm{d}u \tag{7}$$

$\hat{Q}_h(\boldsymbol{\beta})$ 可以被证明是凸的, 且是二次可微的。因此平滑分位数回归估计量 $\hat{\boldsymbol{\beta}}_h$ 可以通过最小化以下函数得到:

$$\hat{\boldsymbol{\beta}}_h = \arg\min_{\boldsymbol{\beta}\in R^p} \hat{Q}_h(\boldsymbol{\beta}) \tag{8}$$

1.2 复合平滑分位数回归

受到卷积平滑技术的启发, 可以将此方法应用到估计效率更优的 CQR 中 [21]。固定 M 个分位数水平 $0<\tau_1<\cdots<\tau_M<1$, 为了计算方便, 选择等间隔分位数 $\tau_k = k/(M+1), k = 1,2,\cdots,M$。对于线性回归模型 (1), $Y|\boldsymbol{X}$ 的 τ_k 条件分位数为 $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}_0 + b_{0,\tau_k}$, 其中 $b_{0,\tau_k} = F_\varepsilon^{-1}(\tau_k)$ 为 ε 的 τ_k 条件分位数。利用 CQR 估计未知参数 $\boldsymbol{\beta}_0$:

$$(\boldsymbol{\beta}^C, \hat{b}_{\tau_1},\cdots,\hat{b}_{\tau_M}) = \arg\min_{\boldsymbol{\beta}, b_{\tau_1},\cdots,b_{\tau_M}} \sum_{k=1}^{M}\left\{\frac{1}{n}\sum_{i=1}^{n}\rho_{\tau_k}(y_i - b_{\tau_k} - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})\right\} \tag{9}$$

引入平滑核函数, 对于固定的分位数个数 M, 可以利用以下复合平滑分位数回归 (composite smoothing quantile regression, CSQR) 的损失函数估计未知参数 $\boldsymbol{\theta}_0 = (\boldsymbol{\beta}_0^{\mathrm{T}}, b_{0,\tau_1},\cdots,b_{0,\tau_M})^{\mathrm{T}}$:

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}=(\boldsymbol{\beta},b_{\tau_1},\cdots,b_{\tau_M})} \hat{Q}_h^C(\boldsymbol{\theta}) \tag{10}$$

式中,

$$\hat{Q}_h^C(\boldsymbol{\theta}) = \sum_{k=1}^{M}\hat{Q}_{h,k}^C(\boldsymbol{\beta}, b_{\tau_k}) = \sum_{k=1}^{M}\left\{\frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{+\infty}\rho_{\tau_k}(t)K_h(t - y_i + b_{\tau_k} + \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})\mathrm{d}t\right\}$$

此时, 由于 $\hat{Q}_h^C(\boldsymbol{\theta})$ 是二次可微的, 因此可求得其梯度和海森矩阵分别为:

$$\nabla\hat{Q}_h^C(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\begin{pmatrix} \sum_{k=1}^{M} x_i\left\{\bar{K}\left(\dfrac{b_{\tau_k} + \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta} - y_i}{h}\right) - \tau_k\right\} \\ \bar{K}\left(\dfrac{b_{\tau_1} + \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta} - y_i}{h}\right) - \tau_1 \\ \vdots \\ \bar{K}\left(\dfrac{b_{\tau_M} + \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta} - y_i}{h}\right) - \tau_M \end{pmatrix} \tag{11}$$

$$\nabla^2\hat{Q}_h^C(\boldsymbol{\theta})=\frac{1}{n}\sum_{i=1}^{n}\begin{pmatrix}\sum\limits_{k=1}^{M}\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}K_h(b_{\tau_k}+\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}-y_i) & \boldsymbol{x}_iK_h(b_{\tau_1}+\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}-y_i) & \cdots & \boldsymbol{x}_iK_h(b_{\tau_M}+\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}-y_i)\\ \boldsymbol{x}_iK_h(b_{\tau_1}+\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}-y_i) & K_h(b_{\tau_1}+\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}-y_i) & \cdots & 0\\ \vdots & \vdots & \vdots & \vdots\\ \boldsymbol{x}_iK_h(b_{\tau_M}+\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}-y_i) & 0 & \cdots & K_h(b_{\tau_M}+\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}-y_i)\end{pmatrix}$$

式中, $\bar{K}_h=\int_{-\infty}^{t}K(u)\mathrm{d}u$。

2　可更新参数估计方法

2.1　流数据下的复合平滑分位数回归

假设流数据集 $\{D_1,D_2,\cdots,D_b\}$ 由 b 批数据组成, $D_j=\{\boldsymbol{x}_j,\boldsymbol{y}_j\}$ 表示第 j 批到达的数据集, 若第 j 批数据集 D_j 的总样本量 n_j, 即 b 批流数据集的总样本量为 $N_b=\sum\limits_{j=1}^{b}n_j$。其中: $\boldsymbol{y}_j=(y_{j1},y_{j2},\cdots,y_{jn_j})^{\mathrm{T}};\boldsymbol{x}_j=(\boldsymbol{x}_{j1},\boldsymbol{x}_{j2},\cdots,\boldsymbol{x}_{jn_j})^{\mathrm{T}}$。

从两批数据 D_1 和 D_2 的简单场景开始, 其中 D_2 在 D_1 之后到达。在仅使用数据集 D_1 的汇总统计量而非 D_1 的原始数据的情况下, 通过最小化式 (10) 更新 CSQR 估计量 $\hat{\boldsymbol{\theta}}_1$ 为 $\hat{\boldsymbol{\theta}}_2^*$。通过式 (10) 、(11), CSQR 估计量 $\hat{\boldsymbol{\theta}}_1$ 满足:

$$\frac{1}{N_1}U(D_1;\hat{\boldsymbol{\theta}}_1;h_1)=0 \tag{12}$$

式中,

$$U(D_j;\boldsymbol{\theta};h)=\sum_{i\in D_j}\begin{pmatrix}\sum\limits_{k=1}^{M}\boldsymbol{x}_i\left\{\bar{K}\left(\dfrac{b_{\tau_k}+\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}-y_i}{h}\right)-\tau_k\right\}\\ \bar{K}\left(\dfrac{b_{\tau_1}+\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}-y_i}{h}\right)-\tau_1\\ \vdots\\ \bar{K}\left(\dfrac{b_{\tau_M}+\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}-y_i}{h}\right)-\tau_M\end{pmatrix}$$

D_1 的样本总量 $N_1=n_1$。则 $\hat{\boldsymbol{\theta}}_2^*$ 满足以下聚合方程:

$$\frac{1}{N_2}U(D_1;\hat{\boldsymbol{\theta}}_2^*;h_2)+\frac{1}{N_2}U(D_2;\hat{\boldsymbol{\theta}}_2^*;h_2)=0 \tag{13}$$

虽然通过式 (13) 可以求解 $\hat{\boldsymbol{\theta}}_2^*$, 但需要 D_1 和 D_2 的全部数据, 这与流数据的性质不相符。因此, 提出一种可更新的复合平滑分位数回归方法 (renewable composite smoothing quantile regression, RCSQR) 来解决这一问题。通过 2.2 节中定理 1 证明中的式 (30) 得到:

$$U(D_1;\hat{\boldsymbol{\theta}}_2^*;h_2)=U(D_1;\hat{\boldsymbol{\theta}}_2^*;h_1)+O_p(n_1h_1^2+n_1h_2^2) \tag{14}$$

式中, $O_p(\cdot)$ 为依概率有界。可以得到 $U(D_1;\hat{\boldsymbol{\theta}}_2^*;h_1)$ 在 $\hat{\boldsymbol{\theta}}_1$ 处的一阶泰勒展开式为:

$$\begin{aligned}U(D_1;\hat{\boldsymbol{\theta}}_2^*;h_1)=&U(D_1;\hat{\boldsymbol{\theta}}_1;h_1)+\\&J(D_1;\hat{\boldsymbol{\theta}}_1;h_1)(\hat{\boldsymbol{\theta}}_2^*-\hat{\boldsymbol{\theta}}_1)+\\&O_p(n_1\|\hat{\boldsymbol{\theta}}_2^*-\hat{\boldsymbol{\theta}}_1\|_2^2)\end{aligned} \tag{15}$$

式中: $J(D_j;\boldsymbol{\theta};h)=\partial U(D_j;\boldsymbol{\theta};h)/\partial\boldsymbol{\theta}$。将式 (12)、(14) 代入式 (15) 可得:

$$\begin{aligned}U(D_1;\hat{\boldsymbol{\theta}}_2^*;h_2)=&J(D_1;\hat{\boldsymbol{\theta}}_1;h_1)(\hat{\boldsymbol{\theta}}_2^*-\hat{\boldsymbol{\theta}}_1)+\\&O_p(n_1\|\hat{\boldsymbol{\theta}}_2^*-\hat{\boldsymbol{\theta}}_1\|_2^2+n_1h_1^2+n_1h_2^2)\hat{\boldsymbol{\theta}}\end{aligned} \tag{16}$$

将式 (16) 代入式 (13) 中, 可得:

$$\begin{aligned}&\frac{1}{N_2}J(D_1;\hat{\boldsymbol{\theta}}_1;h_1)(\hat{\boldsymbol{\theta}}_2^*-\hat{\boldsymbol{\theta}}_1)+\frac{1}{N_2}U(D_2;\hat{\boldsymbol{\theta}}_2^*;h_2)+\\&O_p\left(\frac{n_1}{N_2}\{\|\hat{\boldsymbol{\theta}}_2^*-\hat{\boldsymbol{\theta}}_1\|_2^2+h_1^2+h_2^2\}\right)=0\end{aligned} \tag{17}$$

当样本量 n_1 足够大时, 在一般的正则条件下, $\hat{\boldsymbol{\theta}}_1$ 和 $\hat{\boldsymbol{\theta}}_2^*$ 都是 $\hat{\boldsymbol{\theta}}_0$ 的相合估计量。另外, 当 h_1 和 h_2 足够小时, 误差项 $O_p\left(\dfrac{n_1}{N_2}\{\|\hat{\boldsymbol{\theta}}_2^*-\hat{\boldsymbol{\theta}}_1\|_2^2+h_1^2+h_2^2\}\right)$ 可以被忽略。去除误差项, 可以将式 (17) 修正为:

$$\begin{aligned}&\frac{1}{N_2}J(D_1;\hat{\boldsymbol{\theta}}_1;h_1)(\hat{\boldsymbol{\theta}}_2-\hat{\boldsymbol{\theta}}_1)+\\&\frac{1}{N_2}U(D_2;\hat{\boldsymbol{\theta}}_2;h_2)=0\end{aligned} \tag{18}$$

式 (18) 可以通过历史数据的汇总统计量, 即样本方差矩阵 $J(D_1;\hat{\boldsymbol{\theta}}_1;h_1)$ 和估计值 $\hat{\boldsymbol{\theta}}_1$ 来更新得到新的估计值 $\hat{\boldsymbol{\theta}}_2$。将式 (18) 推广到整个流数据集 $\{D_1,D_2,\cdots,D_b\}$, 从而得到 $\boldsymbol{\theta}_0$ 的可更新估计量 $\hat{\boldsymbol{\theta}}_b$ 为:

$$\begin{aligned}&\frac{1}{N_b}\sum_{j=1}^{b-1}J(D_j;\hat{\boldsymbol{\theta}}_j;h_j)(\hat{\boldsymbol{\theta}}_b-\hat{\boldsymbol{\theta}}_{b-1})+\\&\frac{1}{N_b}U(D_b;\hat{\boldsymbol{\theta}}_b;h_b)=0\end{aligned} \tag{19}$$

2.2 大样本性质

在证明本文所提出的估计量的渐近性质之前, 需要满足以下几个条件:

条件 1 $K(\cdot)$ 是非负对称核函数, 并且是二次可微的, 拥有有界的一阶和二阶导数, 则 $\int_{-\infty}^{+\infty} K(u)\mathrm{d}u = 1, 0 < \int_{0}^{\infty} K(u)\{1-K(u)\}\mathrm{d}u < \infty$。并有 $\int_{-\infty}^{+\infty} |u^2K(u)|\mathrm{d}u < \infty$, $\int_{-\infty}^{+\infty} uK(u)\mathrm{d}u = 0$ 和 $\int_{-\infty}^{+\infty} u^2K(u)\mathrm{d}u \neq 0$。

条件 2 误差项 ε 的密度函数 $f_\varepsilon(\cdot)$ 是 Lipschitz 连续的, 存在常数 $I_0 > 0$ 使得对于所有的 $u_1, u_2 \in R$, 都满足 $|f_\varepsilon(u_1) - f_\varepsilon(u_2)| \leqslant I_0|u_1 - u_2|$。更进一步假设存在两个常数 $\bar{f} > \underline{f} > 0$, 则 $\bar{f} \geqslant \max\limits_{1\leqslant k\leqslant M}\{f_\varepsilon(b_{0,\tau_1}), \cdots, f_\varepsilon(b_{0,\tau_M})\} \geqslant \min\limits_{1\leqslant k\leqslant M}\{f_\varepsilon(b_{0,\tau_1}), \cdots, f_\varepsilon(b_{0,\tau_M})\} \geqslant \underline{f}$。

条件 3 协变量 $\boldsymbol{X} \in R^p$ 的分量为正的且有界的随机变量, 且 $\boldsymbol{\Sigma} = E(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})$ 是 $p\times p$ 的正定矩阵。

条件 1 是核函数的标准条件, 高斯核函数、Epanechnikov 核函数等满足该条件。条件 2 对误差项密度函数施加了正则性条件, 规定了 $f_\varepsilon(\cdot)$ 的 Lipschitz 连续性和在 $\{b_{0,\tau_k}\}_{k=1}^{M}$ 计算的有界性。条件 3 用于估计量的渐近正态性。条件 1~3 是常用于 CSQR 的标准条件。

定理 1 在条件 1~3 都满足的条件下, 如果 $h_j = 0(N_j^{-1/4})$ 且 $h_j(N_j/\ln N_j)^{1/3} \to \infty$, 式中 $N_j = \sum\limits_{i=1}^{j} n_i, n_j \to \infty, i = 1, 2, \cdots, b$ 则有:

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_b - \boldsymbol{\beta}_0) \xrightarrow{d}$$
$$N\left(0, \boldsymbol{\Sigma}^{-1}\frac{\sum\limits_{k,k'=1}^{K}\min(\tau_k,\tau_{k'})(1-\max(\tau_k,\tau_{k'}))}{(\sum\limits_{k=1}^{K} f_\varepsilon(b_{0,\tau_k}))^2}\right)$$

其中, $\xrightarrow{d}$ 表示依分布收敛。

证明 定义函数 $G_b(\boldsymbol{\theta})$ 为:

$$G_b(\boldsymbol{\theta}) = \frac{1}{N_b}\sum_{j=1}^{b-1} J(D_j;\hat{\boldsymbol{\theta}}_j;h_j)(\theta - \hat{\boldsymbol{\theta}}_{b-1}) + \frac{1}{N_b}U(D_b;\boldsymbol{\theta};h_b) \tag{20}$$

根据式 (19) 可知可更新估计量 $\hat{\boldsymbol{\theta}}_b$ 满足 $G_b(\hat{\boldsymbol{\theta}}_b) = \mathbf{0}$。在满足条件 $n_1 \to \infty$ 的条件下, $\hat{\boldsymbol{\theta}}_1$ 是 $\sqrt{N_1}$ 相合的 [19]。如果 $\{\hat{\boldsymbol{\theta}}_j\}_{j=1}^{b-1}$ 是 $\sqrt{N_j}$ 相合的, 则有 $G_b(\boldsymbol{\theta}_0) = O_p(1)$。结合 Marcelo 等[19] 提出的引理 4 可得:

$$\begin{aligned} &G_b(\hat{\boldsymbol{\theta}}_b) - G_b(\boldsymbol{\theta}_0) = \\ &\frac{1}{N_b}\sum_{j=1}^{b-1} J(D_j;\hat{\boldsymbol{\theta}}_j;h_j)(\hat{\boldsymbol{\theta}}_b - \boldsymbol{\theta}_0) + \\ &\frac{1}{N_b}\{U(D_b;\hat{\boldsymbol{\theta}}_b;h_b) - U(D_b;\boldsymbol{\theta}_0;h_b)\} = \\ &\frac{1}{N_b}\Big\{\sum_{j=1}^{b-1} J(D_j;\hat{\boldsymbol{\theta}}_j;h_j) + J(D_b;\boldsymbol{\theta}_0;h_b)\Big\}(\hat{\boldsymbol{\theta}}_b - \boldsymbol{\theta}_0) + \\ &O_p\Big(\frac{n_b}{N_b}\|\hat{\boldsymbol{\theta}}_b - \boldsymbol{\theta}_0\|_2^2\Big) = O_p(1) \end{aligned} \tag{21}$$

结合式 (20)、(21) 和 $G_b(\hat{\boldsymbol{\theta}}_b) = \mathbf{0}$ 可得:

$$G_b(\boldsymbol{\theta}_0) = \frac{1}{N_b}\Big\{\sum_{j=1}^{b-1} J(D_j;\hat{\boldsymbol{\theta}}_j;h_j) + J(D_b;\boldsymbol{\theta}_0;h_b)\Big\} (\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}_b) + O_p\Big(\frac{n_b}{N_b}\|\hat{\boldsymbol{\theta}}_b - \boldsymbol{\theta}_0\|_2^2\Big)。$$

结合式 (20) 可得:

$$\begin{aligned} &-\frac{1}{N_b}\Big\{\sum_{j=1}^{b-1} J(D_j;\hat{\boldsymbol{\theta}}_j;h_j) + J(D_b;\boldsymbol{\theta}_0;h_b)\Big\}(\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}_b) + \\ &\frac{1}{N_b}\sum_{j=1}^{b-1} J(D_j;\hat{\boldsymbol{\theta}}_j;h_j)(\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}_{b-1}) + \\ &\frac{1}{N_b}U(D_b;\boldsymbol{\theta}_0;h_b) + O_p\Big(\frac{n_b}{N_b}\|\hat{\boldsymbol{\theta}}_b - \boldsymbol{\theta}_0\|_2^2\Big) = \mathbf{0} \end{aligned} \tag{22}$$

由 $U(D_1;\boldsymbol{\theta}_1;h_1) = \mathbf{0}$ 可得:

$$\begin{aligned} &U(D_1;\boldsymbol{\theta}_0;h_1) = U(D_1;\hat{\boldsymbol{\theta}}_1;h_1) + \\ &J(D_1;\hat{\boldsymbol{\theta}}_1;h_1)(\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}_1) + O_p(n_1\|\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_0\|_2^2) = \\ &J(D_1;\hat{\boldsymbol{\theta}}_1;h_1)(\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}_1) + O_p(n_1\|\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_0\|_2^2) \end{aligned} \tag{23}$$

通过式 (18) 可得:

$$\begin{aligned} &U(D_2;\boldsymbol{\theta}_0;h_2) = U(D_2;\hat{\boldsymbol{\theta}}_2;h_2) + \\ &J(D_2;\hat{\boldsymbol{\theta}}_2;h_2)(\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}_2) + O_p(n_2\|\hat{\boldsymbol{\theta}}_2 - \boldsymbol{\theta}_0\|_2^2) = \\ &J(D_1;\hat{\boldsymbol{\theta}}_1;h_1)(\hat{\boldsymbol{\theta}}_2 - \hat{\boldsymbol{\theta}}_1) + J(D_2;\hat{\boldsymbol{\theta}}_2;h_2)(\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}_2) + \\ &O_p(n_2\|\hat{\boldsymbol{\theta}}_2 - \boldsymbol{\theta}_0\|_2^2) \end{aligned} \tag{24}$$

因此, 结合式 (23)、(24) 可得:

$$\sum_{j=1}^{2} U(D_j;\boldsymbol{\theta}_0;h_j) = \sum_{j=1}^{2} J(D_j;\boldsymbol{\theta}_j;h_j)(\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}_{b-1}) + O_p\Big(\sum_{j=1}^{2} n_j\|\hat{\boldsymbol{\theta}}_j - \boldsymbol{\theta}_0\|_2^2\Big) \tag{25}$$

将式 (25) 推广到第 $b-1$ 批数据集, 可得:

$$\sum_{j=1}^{b-1}U(D_j;\boldsymbol{\theta}_0;h_j)=\sum_{j=1}^{b-1}J(D_j;\boldsymbol{\theta}_j;h_j)(\boldsymbol{\theta}_0-\hat{\boldsymbol{\theta}}_{b-1})+O_p\Big(\sum_{j=1}^{b-1}n_j\|\hat{\boldsymbol{\theta}}_j-\boldsymbol{\theta}_0\|_2^2\Big) \tag{26}$$

将式 (26) 代入到式 (22) 中, 可得:

$$-\frac{1}{N_b}\Big\{\sum_{j=1}^{b-1}J(D_j;\hat{\boldsymbol{\theta}}_j;h_j)+J(D_b;\boldsymbol{\theta}_0;h_b)\Big\}(\boldsymbol{\theta}_0-\hat{\boldsymbol{\theta}}_b)+\frac{1}{N_b}\sum_{j=1}^{b}U(D_j;\boldsymbol{\theta}_0;h_j)+O_p\Big(\sum_{j=1}^{b}\frac{n_j}{N_b}\|\hat{\boldsymbol{\theta}}_j-\boldsymbol{\theta}_0\|_2^2\Big)=0 \tag{27}$$

根据文献 [19] 中提出的引理 1 和引理 4, 在 $n_j\to\infty, j=1,2,\cdots,b$ 的条件下, 可得:

$$\|J(D_j;\hat{\boldsymbol{\theta}};h_j)-J(D_j;\boldsymbol{\theta}_0;h_j)\|=O_p(n_j\|\hat{\boldsymbol{\theta}}_j-\boldsymbol{\theta}_0\|_2),$$

$$\|J(D_j;\boldsymbol{\theta}_0;h_j)-E\{J(D_j;\boldsymbol{\theta}_0;h_j)\}\|=O_p(\sqrt{n_jh_j^{-1}\ln n_j}), \tag{28}$$

$$E\{J(D_j;\boldsymbol{\theta}_0;h_j)\}=\sum_{i\in D_j}\sum_{k=1}^{M}E\{f_\varepsilon(b_{0,\tau_k})\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}\}+O_p(n_jh_j)$$

因为 $\{\hat{\boldsymbol{\theta}}_j\}_{j=1}^{b-1}$ 是相合的, 在 $h_j(N_j/\ln N_j)^{1/3}\to\infty$ 和 $n_j\to\infty, j=1,2,\cdots,b$ 的条件下, 可以得到:

$$-\frac{1}{N_b}\Big\{\sum_{j=1}^{b-1}J(D_j;\hat{\boldsymbol{\theta}}_j;h_j)+J(D_b;\boldsymbol{\theta}_0;h_b)\Big\}=\sum_{k=1}^{M}f_\varepsilon(b_{0,\tau_k}\boldsymbol{\Sigma}+O_p(1) \tag{29}$$

根据文献 [19] 中提出的引理 1 和引理 5 可得:

$$U(D_j;\boldsymbol{\theta}_0;h_j)=U(D_j;\boldsymbol{\theta}_0;h_b)+O_p(n_jh_j^2+n_jh_b^2) \tag{30}$$

将式 (29)、(30) 代入到式 (28) 中, 可得:

$$(\sum_{k=1}^{M}f_\varepsilon(b_{0,\tau_k})\boldsymbol{\Sigma}+O_p(1))(\hat{\boldsymbol{\theta}}_b-\boldsymbol{\theta}_0)+\frac{1}{N_b}\sum_{j=1}^{b}U(D_j;\boldsymbol{\theta}_0;h_b)+\frac{n_b}{N_b}O_p(\|\hat{\boldsymbol{\theta}}_b-\boldsymbol{\theta}_0\|_2^2)+\frac{1}{N_b}\sum_{j=1}^{b-1}O_p(n_jh_j^2+n_jh_b^2+n_j\|\hat{\boldsymbol{\theta}}_j-\boldsymbol{\theta}_0\|_2^2)=0 \tag{31}$$

通过 Han 等[25] 提出的引理 12, 可得

$$\sum_{j=1}^{b}\frac{n_j}{N_j}\leqslant 1+\log\frac{N_b}{N_1}$$

$$\sum_{j=1}^{b}\frac{n_j}{\sqrt{N_j}}\leqslant 2\sqrt{N_b}$$

因此, 根据中心极限定理和条件 $h_j=o(N_j^{-1/4})$, 可以证明该定理。

2.3　算法

在数值上, 使用 Newton-Raphson 方法从式 (19) 的第 $r+1$ 次迭代中得到估计量 $\hat{\boldsymbol{\theta}}_b$:

$$\hat{\boldsymbol{\theta}}_b^{(r+1)}=\hat{\boldsymbol{\theta}}_b^{(r)}-\{\hat{J}_{b-1}+J(D_b;\hat{\boldsymbol{\theta}}_b^{(r)};h_b)\}^{-1}\hat{U}_b^{(r)}, \tag{32}$$

式中: $\hat{J}_{b-1}=\sum_{j=1}^{b-1}J(D_j;\hat{\boldsymbol{\theta}}_j;h_j)$; $\hat{U}_b^{(r)}=\hat{J}_{b-1}(\hat{\boldsymbol{\theta}}_b^{(r)}-\hat{\boldsymbol{\theta}}_{b-1})+U(D_b;\hat{\boldsymbol{\theta}}_b^{(r)};h_b)$, 当 p 很大时, 为了加快式 (32) 的计算速度, 可以避免更新 $J(D_b;\hat{\boldsymbol{\theta}}_b^{(r)};h_b)$, 用 $\hat{\boldsymbol{\theta}}_{b-1}$ 代替 $\hat{\boldsymbol{\theta}}_b^{(r)}$, 得到以下迭代算法:

$$\hat{\boldsymbol{\theta}}_b^{(r+1)}=\hat{\boldsymbol{\theta}}_b^{(r)}-\{\hat{J}_{b-1}+J(D_b;\hat{\boldsymbol{\theta}}_{b-1};h_b)\}^{-1}\hat{U}_b^{(r)} \tag{33}$$

在式 (33) 中, 仅利用截至第 $b-1$ 批数据集的历史汇总数据统计量 $\hat{J}_{b-1}$ 和 $\hat{\boldsymbol{\theta}}_{b-1}$, 而非 $\{D_1,D_2,\cdots,D_{b-1}\}$ 的所有原始数据信息。根据式 (33) 将提出的可更新 CSQR 算法归纳如下:

值得注意的是, 在以上算法的第 7 步中, 仅需要保存 $\hat{\boldsymbol{\theta}}_b$ 和 $\hat{J}_b$, 它们分别是 $(p+M)\times 1$ 维和 $(p+M)\times(p+M)$ 维。因此需要存储的数据规模从 N_bp 降到 $(p+M)\times(p+M+1)$。由于本文假设 p 是一个固定的数值, 因此该算法大大减少了数据存储量。

3　数值实验

3.1　数值模拟

使用两个蒙特卡洛数值模拟实验来评估所提出的可更新 CRSQR 方法的估计性能, 编程均由 R 语言实现。并与以下 3 种方法比较: ① 利用完整数据的 CQR 方法 [14], 通过 R 语言的 cqrReg 包实现; ② 可更新 RLS 方法 [9]; ③ RSQR 方法 [16]。

表 1 流数据的可更新 CSQR 估计算法

Tab. 1 Renewable CSQR estimation for streaming data sets

1. 输入: 流数据集 $D_1,\cdots,D_b,\cdots$, 分位数水平 τ_k, 核函数 $K(\cdot)$ 以及带宽 $h_b, b=1,2,\cdots$
2. 初始化: 利用数据集 D_1 通过最小化式 (10) 得到 $\hat{\boldsymbol{\theta}}_1$, 再计算 $J(D_1;\hat{\boldsymbol{\theta}}_1;h_1)$
3. for: $b=2,3,\cdots$
4. 读取数据集 D_b, 再计算 $\hat{J}_{b-1}+J(D_b;\hat{\boldsymbol{\theta}}_{b-1};h_b)$
5. 让初始的估计量 $\hat{\boldsymbol{\theta}}_b^{(0)}=\hat{\boldsymbol{\theta}}_{b-1}$, 然后进行以下迭代直至 $\hat{\boldsymbol{\theta}}_b$ 收敛:
   $\hat{\boldsymbol{\theta}}_b^{(r+1)}=\hat{\boldsymbol{\theta}}_b^{(r)}-\{\hat{J}_{b-1}+J(D_b;\hat{\boldsymbol{\theta}}_{b-1};h_b)\}^{-1}\hat{U}_b^{(r)}$
6. 用 $\hat{J}_b=\hat{J}_{b-1}+J(D_b;\hat{\boldsymbol{\theta}}_b;h_b)$ 更新 $\hat{J}_b$
7. 保存 $\hat{\boldsymbol{\theta}}_b$ 和 $\hat{J}_b$, 然后从内存中释放掉数据集 D_b
8. end
9. 输出: $\hat{\boldsymbol{\theta}}_b$

由于 SQR 对带宽 h 的选择不敏感, 因此依据定理 1, 为方便计算, 选取带宽 $h_j(N_j\ln N_j)^{-1/4}$ 和高斯核函数 $K(u)=(2\pi)^{-1/2}\exp(-u^2/2)$ 用于所有实验。CQR 研究中推荐分位数个数 $M=19$, 选取等间隔分位数 $\tau_k=k/20, k=1,2,\cdots,M$。每个批次有 $\bar{n}$ 个样本量, N_b 为 b 批数据的总样本量。样本数据由以下线性模型产生:

$$\boldsymbol{Y}=\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}_0+\varepsilon, \tag{34}$$

式中, p 维的协变量向量 $\boldsymbol{X}\in R^p$ 服从正态分布 $N(0,\bar{\Sigma})$, $\bar{\Sigma}$ 的组成为 $\bar{\Sigma}_{ij}=0.5^{i-j}, 1\leqslant i,j\leqslant p$, $p=10$。参数真值 $\boldsymbol{\beta}_0$ 是一个元素全为 1 的 p 维列向量。随机误差项 ε 将从以下 3 种分布中产生, Case 1 为轻尾分布, Case 2 和 Case 3 是重尾分布:

Case 1 $\varepsilon\sim N(0,1)$;

Case 2 $\varepsilon=\sigma(\boldsymbol{X})\varepsilon^*$, 式中, $\sigma(\boldsymbol{X})=\mathbf{1}_{N_b\times 1}+0.5\cos(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}_0)$, $\varepsilon^*\sim t(3)$;

Case 3 $\varepsilon\sim 0.8N(0,1)+0.2N(0,20)$。

为了评估每一种估计方法的性能, 选取均方根误差 RMSE 作为评价指标, 其中表示参数估计值与真实值的均方根误差, 该值越小, 则表示估计精确性越高。

$$\mathrm{RMSE}=\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|_2 \tag{35}$$

3.1.1 固定 $\bar{n}$ 改变 b

固定每批数据集样本量 $\bar{n}=200$, 改变数据集批次 $b=100,200,\cdots,1\,000$。RSQR 的分位数水平 $\tau=0.5$。表 2 所示结果均由 100 次重复模拟实验得到。

表 2 固定 $\bar{n}$ 时, 不同批次 b 下的均方根误差 (RMSE×100)

Tab. 2 RMSE×100 under different batches b with a fixed $\bar{n}$

	b	CQR	RLS	RSQR	RCSQR
Case 1	100	2.760	2.842	3.497	2.820
	200	1.886	1.928	2.375	1.921
	300	1.564	1.610	1.929	1.579
	400	1.386	1.412	1.693	1.392
	500	1.256	1.283	1.532	1.255
	600	1.135	1.150	1.450	1.156
	700	1.032	1.057	1.278	1.044
	800	0.997	1.101	1.265	1.009
	900	0.982	0.994	1.179	0.992
	1 000	0.862	0.875	1.023	0.856
Case 2	100	3.365	5.121	3.026	3.070
	200	2.552	3.671	2.345	2.328
	300	2.125	3.035	1.851	1.925
	400	1.710	2.465	1.595	1.557
	500	1.582	2.229	1.460	1.439
	600	1.430	2.100	1.308	1.287
	700	1.414	2.014	1.287	1.278
	800	1.289	1.808	1.160	1.156
	900	1.244	1.781	1.119	1.106
	1 000	1.108	1.608	1.104	1.009
Case 3	100	6.323	7.139	6.335	6.435
	200	4.146	5.354	4.233	4.165
	300	3.654	4.589	3.700	3.676
	400	2.899	3.681	2.948	2.947
	500	2.717	3.504	2.749	2.728
	600	2.383	3.305	2.427	2.415
	700	2.236	3.018	2.262	2.260
	800	2.123	2.705	2.175	2.154
	900	2.042	2.613	2.076	2.063
	1 000	1.917	2.402	1.956	1.929

由表 2 可以得出以下结论:

(1) 通过表 2 的均方根误差可以发现, 在其他条件不变的情况下, RCSQR 和 CQR 的估计精确性无明显差异且接近于真实值。并且随着批次数 b 的增加, RMSE 逐渐减小, 这是因为 b 增加导致样本量增加, 估计量的精度提高, 这与实际相符。

(2) 在其他条件不变的情况下, RCSQR 和 CQR 在重尾噪声下的表现更好 (Case 2 和 Case 3)。这表明当误差项不满足正态分布或包含离群值时, RCSQR 更有效, 不会受到偏态分布的影响。

3.1.2　固定 N_b 改变 b

固定全部数据样本量分别为 $N_b = 10^5$ 和 $N_b = 2 \times 10^6$, 改变数据批次为 $b = 100, 200, \cdots, 1\ 000$。依旧选择 CQR 和 RCSQR 的分位数水平为 $\tau_k = k/20, k = 1, 2, \cdots, 19$, RSQR 的分位数水平 $\tau = 0.5$。表 3 所示结果均由 100 次重复模拟实验得到。

表 3　固定 N_b, 不同批次 b 下的均方根误差 (RMSE×100)

Tab. 3　RMSE×100 under different batches b with fixed N_b

	b	$N_b = 10^5$				$N_b = 2 \times 10^6$			
		CQR	RLS	RSQR	RCSQR	CQR	RLS	RSQR	RCSQR
Case 1	100	1.261	1.295	1.511	1.284	0.269	0.261	0.306	0.246
	200	1.261	1.295	1.510	1.284	0.269	0.261	0.306	0.246
	300	1.261	1.300	1.519	1.291	0.269	0.261	0.306	0.246
	400	1.261	1.295	1.510	1.284	0.269	0.261	0.306	0.246
	500	1.261	1.295	1.512	1.285	0.269	0.261	0.306	0.246
	600	1.261	1.301	1.527	1.293	0.269	0.262	0.306	0.246
	700	1.261	1.306	1.526	1.296	0.269	0.261	0.306	0.246
	800	1.261	1.295	1.509	1.284	0.269	0.261	0.306	0.246
	900	1.261	1.301	1.516	1.292	0.269	0.262	0.306	0.246
	1 000	1.261	1.295	1.512	1.284	0.269	0.261	0.306	0.246
Case 2	100	1.432	2.141	1.441	1.341	0.405	0.526	0.315	0.305
	200	1.432	2.141	1.441	1.340	0.405	0.526	0.315	0.305
	300	1.432	2.150	1.446	1.349	0.405	0.526	0.314	0.305
	400	1.432	2.141	1.441	1.340	0.405	0.526	0.315	0.305
	500	1.432	2.141	1.44	1.341	0.405	0.526	0.315	0.305
	600	1.432	2.144	1.473	1.361	0.405	0.526	0.314	0.305
	700	1.432	2.161	1.434	1.353	0.405	0.526	0.314	0.305
	800	1.432	2.141	1.437	1.339	0.405	0.526	0.314	0.305
	900	1.432	2.151	1.444	1.347	0.405	0.526	0.314	0.305
	1 000	1.432	2.141	1.439	1.34	0.405	0.526	0.314	0.305
Case 3	100	3.037	3.495	3.152	3.017	0.599	0.745	0.617	0.605
	200	3.037	3.495	3.152	3.016	0.599	0.745	0.617	0.604
	300	3.037	3.489	3.156	3.015	0.599	0.745	0.615	0.603
	400	3.037	3.519	3.152	3.017	0.599	0.746	0.617	0.604
	500	3.037	3.510	3.152	3.019	0.599	0.745	0.617	0.605
	600	3.037	3.482	3.168	3.033	0.599	0.745	0.616	0.604
	700	3.037	3.519	3.182	3.051	0.599	0.744	0.616	0.604
	800	3.037	3.534	3.152	3.019	0.599	0.745	0.617	0.605
	900	3.037	3.495	3.155	3.019	0.599	0.745	0.616	0.604
	1 000	3.037	3.518	3.152	3.021	0.599	0.745	0.617	0.605

从表 3 得到结论如下:

(1) 通过表 3 各方法的均方根误差可以发现, 对于任何给定的总样本量 N_b、批次 b 和误差分布情况, RCSQR 的精确性都优于 RLS 和 RSQR。且 RCSQR 和 CQR 的精确性十分接近, 从数值模拟上说明 RCSQR 已达到最优估计效率。

(2) 当其他条件不变时, 在仅改变批次数量 b 的情况下, RCSQR 的 RMSE 无明显变化。说明其他条件不变时, RCSQR 的估计精确性受到批次数量 b 的影响较小。

3.2 实例分析

将本文提出的 RCSQR 应用于蛋白质 3 级结构数据集, 该数据集收集于 2013 年 3 月 31 日, 包括来自 CASP 5-9 的 45 730 个蛋白质结构实例, 可以通过蛋白质的理化性质预测 RMSD 的大小, 常被用于回归分析[26]。数据集可从 https://archive.ics.uci.edu/ dataset/265/ physicochemical+ properties+ of+ protein+ tertiary+ structure 网站中获取。

本文选用 RMSD 作为响应变量, RMSD 被定义为两个结构对应原子坐标差值的均方根值, 蛋白质的 RMSD 可以反映蛋白质在运动过程中的构象与初始构象之间的结构变化, RMSD 的值越小, 表示两个结构越相似。解释变量维度为 8, 解释变量及相关描述如表 4 所示。选取的数据集共包含 45 730 条样本数据, 数据中未含有缺失数据, 选择前 $N_b = 40\ 000$ 条数据作为训练集得到估计系数, 后 $N_c = 5\ 730$ 条数据作为测试集用于评价估计性能。由于解释变量数据的量级差距较大, 为确保不同量级的解释变量对模型的影响更加平衡, 首先对样本数据的解释变量进行标准化处理:

$$x' = \frac{x - \mathrm{mean}(x)}{\mathrm{sd}(x)} \tag{36}$$

式中: x' 是标准化后的样本数据; x 为原始样本数据; $\mathrm{mean}(x)$ 为变量 x 的均值; $\mathrm{sd}(x)$ 为变量 x 的标准差。另外, 设置批次数 $b = 200, 500, 800$, RSQR 的分位数水平 $\tau = 0.5$, RCSQR 和 CQR 的分位数水平 $\tau_k = k/20, k = 1, 2, \cdots, 19$。此外, 基于平均绝对误差 MAE, 将 RCSQR 和 CQR、RLS、RSQR 比较, MAE 表达式为:

$$\mathrm{MAE} = \frac{1}{N_c}\sum_{i=1}^{N_c}|Y_i - \hat{Y}_i| \tag{37}$$

表 4 实例数据集的变量信息

Tab. 4 Covariates and their descriptions for real data

变量名称	描述	均值	最小值	最大值	标准差
F_1	总表面积	9 873.682	2 783.150	43 340.200	4 011.808
F_2	非极性暴露区域	3 016.436	403.500	11 787.100	1 450.042
F_3	暴露的非极性残留物的分数面积	0.302	0.094	0.568	0.063
F_4	残留的非极性暴露部分的面积	103.404	10.689	343.239	54.939
F_5	分子量加权的暴露面积	1 369 092.965	374 315.516	4 467 324.700	558 385.282 3
F_6	与残留物标准暴露区域的平均偏差	145.545	33.646	460.897	69.305
F_7	欧几里得距离	3 987.146	1 108.900	83 153.57	1 880.514
F_8	二级结构惩罚	70.043	0.000	337.000	56.505 5

表 5 所示为 4 种方法应用于蛋白质三级结构数据集的 5 730 条测试集样本所得到的响应变量实际值与预测值的平均绝对误差 MAE, 可以发现 RCSQR 和 CQR 的估计精确性较为相似且 MAE 小于 RLS 和 RSQR。总而言之, 实例数据实验结果表明 RCSQR 的方法不仅在数值模拟实验中有效, 在实例实验中也表现优异, 具有较强的实际应用价值。

表 5 实例分析中不同方法 MAE 对比

Tab. 5 The MAE for different estimators for real data

b	CQR	RLS	RSQR	RCSQR
200	4.487	5.499	4.644	4.564
500	4.487	5.499	4.681	4.555
800	4.487	5.499	4.722	4.540

4 结 论

本文提出了一种可更新的复合分位数回归方法, 避免使用历史全部原始数据, 适用于流数据的特点。通过理论性质推导和数值实验, 可以得到提出的

RCSQR 具有以下优势: (1) 相比较于 CQR, RCSQR 将存储数据规模由 $N_b p$ 降到 $(p+M)\times(p+M+1)$。且在满足一些一般的温和条件下与使用完整原始数据的 CQR 是渐近等价的, 即在达到最优估计效率的同时, 又大大减少了数据存储量。(2) 相比较于 RLS 和 RSQR, RCSQR 的估计精确性表现更优, 尤其是在重尾噪声的情况下, 这是因为 RCSQR 能够同时考虑多个分位数水平的情况, 对存在异常值和重尾噪声的情况更具鲁棒性。

参考文献 :

[1] EFTEKHARI A, ONGIE G, BALZANO L, et al. Streaming principal component analysis from incomplete data [J]. Journal of Machine Learning Research, 2019, 20 (86): 1-62.

[2] WANG C, CHEN M H, SCHIFANO E, et al. Statistical methods and computing for big data [J]. Statistics and Its Interface, 2016, 9(4): 399-414.

[3] WANG H Y, MA Y Y. Optimal subsampling for quantile regression in big data [J]. Biometrika, 2021, 108(1): 99-112.

[4] 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.

[5] JIN J, LIU S Z, MA T F. Optimal subsampling algorithms for composite quantile regression in massive data [J]. Statistics, 2023, 57(4): 811-843.

[6] LIN N, XI R B. Aggregated estimating equation estimation [J]. Statistics and Its Interface, 2011, 4(1): 73-83.

[7] CHEN X Y, XIE M G. A split-and-conquer approach for analysis of extraordinarily large data [J]. Statistica Sinica, 2014, 24: 1655-1684.

[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, SONG X P. Renewable estimation and incremental inference in generalized linear models with streaming data sets [J]. Journal of the Royal Statistical Society: Series B, 2020, 82(1): 69-97.

[10] SHI C C, SONG R, LU W B, et al. Statistical inference for high-dimensional models via recursive online-score estimation [J]. Journal of the American Statistical Association, 2020, 116(535): 1307-1318.

[11] JIANG R, YU K M. Renewable quantile regression for streaming data sets [J]. Neurocomputing, 2022, 508: 208-224.

[12] WANG K N, WANG H W, LI S M. Renewable quantile regression for streaming datasets [J]. Knowledge-Based Systems, 2022, 235: 878-890.

[13] KOENKER R, BASSETT G. Regerssion quantiles [J]. Econometrical, 1978, 46(1): 33-50.

[14] ZOU H, YUAN M. Composite quantile regression and the oracle model selection theory [J]. Annals of Statistics, 2008, 36(3): 1108-1126.

[15] KAI B, LI R, ZOU H. New efficient estimation and variable selection methods for semiparametric varying-coefficient partially linear models [J]. Annals of Statistics, 2011, 39(1): 305-332.

[16] JIANG R, ZHOU Z, QIAN W, et al. Single-index composite quantile regression [J]. Journal of the Korean Statistical Society, 2012, 41(3): 323-332.

[17] 吕亚召, 张日权, 赵为华, 等. 部分线性单指标模型的复合分位数回归及变量选择 [J]. 中国科学: 数学, 2014, 44(12): 1299-1322.

[18] HOROWITZ J L. Bootstrap methods for median regression models [J]. Econometrica, 1998, 66(6): 1327-1351.

[19] MARCELO F, EMMANUEL G, EDUARDO H. Smoothing quantile regressions [J]. Journal of Business & Economic Statistics, 2021, 39(1): 338-357.

[20] TAN K M, BATTEY H, ZHOU W X. Communication-constrained distributed quantile regression with optimal statistical guarantees [J]. Journal of Machine Learning Research, 2022, 23(272): 1-61.

[21] DI F R, WANG L. Multi-round smoothed composite quantile regression for distributed data [J]. Annals of the Institute of Statistical Mathematics, 2022, 74(5): 869-893.

[22] 何晓霞, 魏茳越, 田彤, 等. 基于 LASSO 复合分位数回归的钢材力学性能预测 [J]. 数学的实践与认识, 2023, 53(1): 122-129.

[23] 单文. 基于复合分位数回归方法的我国对外贸易进出口总额预测研究 [D]. 烟台: 山东工商学院, 2022.

[24] 王江荣. 加权复合分位数自回归模型在隧道围岩变形预测中的应用 [J]. 大地测量与地球动力学, 2017, 37(5): 511-515.

[25] HAN R J, LUO L, LIN Y Y, et al. Online debiased lasso [J/OL]. https://doi.org/10.48550/arXiv.2106.05925.

[26] AHMED H A, MUHAMMAD P J, FAEQ A K, et al. An investigation on disparity responds of machine learning algorithms to data normalization method [J]. Aro-The Scientific Journal of Koya University, 2022, 10(2): 29-37.