文章编号: 1001-4543(2025)03-0323-07 DOI: 10.19570/j.cnki.jsspu.2025.03.011

分布式数据集的稳健统计诊断

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)

摘 要: 随着互联网、物联网、人工智能等领域的飞速发展, 分布式系统的应用场景正在不断拓宽。然而, 由于分布式系统中服务器来源的多样性, 可能存在异质性, 进而影响统计推断的准确性。因此, 在分布式系统中进行统计诊断具有重要意义。采用边际相关性作为诊断统计量, 并借助 Huber 回归增强对数据源多样性与重尾噪声影响下的稳健性。数值模拟结果验证了所提方法的有效性, 表明其在处理大规模高维数据集时, 在分布式计算环境中具有良好的适用性与优越性。

关键词: 统计诊断; 分布式数据; Huber 回归; 群组删除

中图分类号: O212.2 **文献标志码**: A

Robust Statistical Diagnostics for Distributed Datasets

HU Guanhao[1], JIANG Rong[2]

(1. School of Mathematics and Statistics, Donghua University, Shanghai 201600, China; 2. School of Statistics and Information, Shanghai University of International Business and Economics, Shanghai 201620, China)

Abstract: With the rapid development of fields such as the Internet, the Internet of Things, and Artificial Intelligence, the application scenarios of distributed systems are constantly expanding. However, due to the diversity of server sources in distributed systems, there may be heterogeneity that could affect the accuracy of statistical inference. Therefore, performing statistical diagnostics in distributed systems is of great significance. The marginal correlation is used as a diagnostic statistic and Huber regression is employed to enhance robustness when facing data source diversity and heavy-tailed noise. The results of numerical simulations are used to verify the effectiveness of the proposed method, demonstrating its good applicability and superiority in a distributed computing environment for handling large-scale, high-dimensional datasets.

Keywords: statistical diagnostics; distributed data; Huber regression; group deletion

0 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, 数据量逐年呈指数级增长。这些数据通过多种终端设备收集, 并存储在分布式环境中。传统集中范式的统计推断方法要求将所有数据存储在内存中, 并确保所有节点访问共享数据, 这导致了高内存需求和计算成本过大。因此, 面对大数据处理, 亟需探索更高效的分布式方法。而面对分布式系统的海量数据的统计分析, 通常需要借助计算机集群来进行处理。

在这种背景下, 分布式系统应运而生, 其能够有效应对海量数据的统计处理需求。分布式系统将庞大的计算任务拆分为多个小任务, 由多台通过网络连接的服务器并行处理, 最终将结果汇总。其中, 最经典的方法是单论型 (One-Shot) 方法, 这是一种简便有效的策略。该方法要求每台局部机器仅与中心机器通信一次, 具体步骤为: 局部机器计算本地统计

收稿日期: 2025-02-17

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@suibe.edu.cn

量并发送给中心机器, 后者整合这些统计量以得到最终的分布式估计。这不仅提高了计算效率, 还降低了对单一机器的内存需求。Lin 等 [1] 研究线性模型中未知参数的加权平均的单轮型估计法 (最小二乘方法)。Battey 等 [2] 提出基于 Wald 检验和 Rao 得分检验的高维单轮型估计量, 将估计量从低维广义线性模型扩展到高维稀疏的场景。Jiang 等 [3] 研究了加权平均的单轮型方法。然而, 单轮型方法的有效性受限于局部机器数量, 过多的局部机器可能导致估计精度下降。针对这个问题, Jordan 等 [4] 提出了分布式的通信有效方法。基于此, 研究利用 Jordan 的方法分析分布式数据。

同时, 因为数据集规模和维度的增加, 异常值的出现几乎不可避免。在高维回归分析中, 异常值的离群可能性会因维度的增加而放大。多个异常值的存在不可避免地会导致分析失真和预测不准确。为了解决异常值检测的问题, Cook[5] 在横截面数据的线性回归中提出了一种度量方法, 称为 Cook 距离。该方法通过比较完整数据上的最小二乘估计与删除特定样本子集后的最小二乘估计, 从而量化个别观测点对回归系数估计的影响。Cook 距离较大的观测点通常被视为具有显著影响力, 需进一步判断其是否为异常值或极端值。Cook 距离的计算不需要对每个被删除的观测值重新拟合模型, 这使得其在回归分析中得到了广泛应用。例如线性回归模型 [6−7]、分类数据分析 [8]、广义线性模型 [9]、广义估计方程 [10]、线性混合模型 [11]、广义线性混合模型 [12]、增长曲线模型 [13] 等。

对于高维数据的影响点诊断仍面临许多问题去解决, 当维度增加时, 影响点会误导变量选择的结果, 更新计算成本较大。因此, 考虑响应和自变量之间的边际相关系数。通过计算每个自变量与响应变量之间的边际相关系数, 然后根据这些相关系数的绝对值进行排序。然而, 当数据集中存在多个异常值时, 单一删除的方法可能会面临屏蔽或淹没的问题, 导致检测效力显著降低。具体而言, 如果一个有影响力的数据点未被识别, 就会产生掩盖现象; 反之, 如果一个无影响力的数据点被错误识别为有影响力, 就会导致淹没现象。为了解决这两种现象, 研究者提出了群组删除的方法。在高维数据的多点诊断研究中, Zhao 等 [14] 对多影响点检测进行了改进, 根据标准化残差的大小或类似标准进行删除。同时, Wang 等 [15] 提出了一种适用于稀疏高维线性回归模型的异常值检测方法。该方法通过降维和变量筛选来构建一个干净的子集, 然后对剩余观测值进行异常性测试。本研究采用分布式通信有效法进行参数估计, 采用边际相关性作为特征判定。结合 Zhao 和 Wang 的方法, 先构建干净子集, 然后借鉴 Cook 距离中的 “leave-one-out” 思想, 使用 “add-one-into” 测量方法进行分布式子机器上的全体数据的多点诊断。

在分布式系统中, 数据往往来自不同时间和地点的多种来源, 因此均方差假设通常不成立。此外, 重尾噪声在实际应用中也很常见, 且其方差可能是无限的。为了解决这一问题, 采用 Huber 回归方法, 它具有对异常值非常稳健的特性。这种方法最初由 Huber[16] 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。为了抵抗异常值或包含具有重尾分布的变量, Sun 等 [17] 研究了自适应 Huber 回归以实现稳健的估计和推理。为了减少噪声的影响, Gupta 等 [18] 通过使用不对称的 Huber 损失函数分析了稳健的正则化极限学习机框架。Luo 等 [19] 提出分布式正则化自适应矩阵回归, 在假设误差对称下, 研究了 Huber 损失下无截距项的高维线性模型的统计推断问题。潘莹丽等 [20] 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。Jiang 等 [21] 提出了一个新的平滑 Huber 的一阶导数, 可以用于快速和可扩展的算法来进行流式数据集的优化, 并且在不同版本中实现了最佳的 Huber 函数拟合。

与已有文献相比, 本研究考虑了分布式数据的稳健回归下的影响诊断, 主要目的是找出分布式中的异常子集。利用通信高效方法进行参数估计, 采用边际相关性作为特征判定, 预先构建干净子集, 然后借鉴 Cook 距离中的 “leave-one-out” 思想, 使用 “add-one-into” 测量方法进行子集诊断。这种研究方法计算效率高, 只需要一台高性能主机器, 其余机器限制低。通过数值模拟验证了研究提出基于通信有效 Huber 回归的多点影响诊断方法的可靠性。

1 分布式系统的 Huber 回归

1.1 分布式系统

考虑在分布式系统中, 研究假设数据集为 K

个子集, 即 $\{C_1, C_2, \cdots, C_K\}$, 其中 $C_k = \{(y_{ki}, x_{ki}), i = 1, 2, \cdots, n_k, k = 1, 2, \cdots, K\}$, $n_k = |C_k|$。观测值 $\{(y_{ki}, x_{ki}), i = 1, 2, \cdots, N = \sum_{k=1}^{K} n_k\}$ 独立同分布, 产生于模型 (1)。假定观测值来自以下线性回归模型:

$$Y = X'\beta_0 + \varepsilon \tag{1}$$

式中: ε 是随机误差; β_0 是 p 维未知回归系数变量; X' 表示 X 的转置。

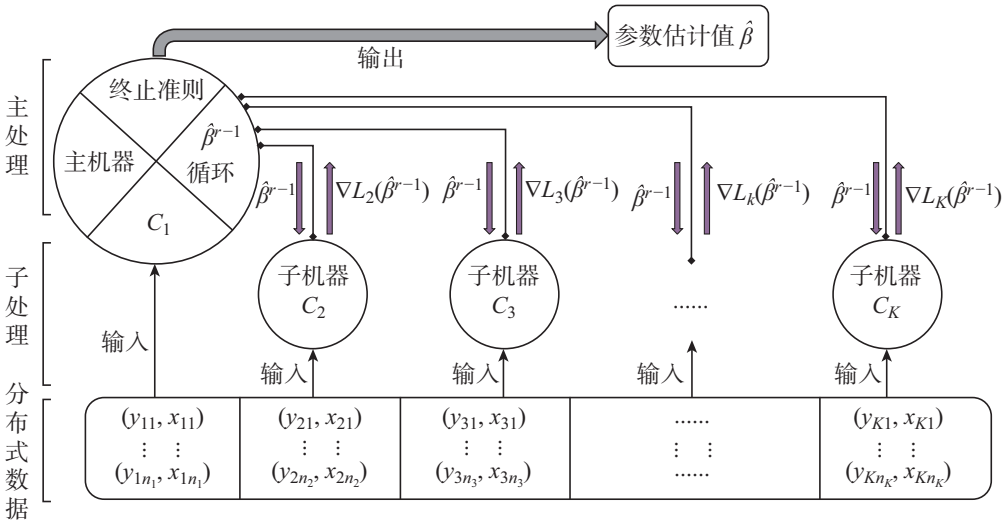

图 1　分布式迭代图

Fig. 1　Distributed iteration diagram

1.2　分布式系统下的 Huber 回归

传统的最小二乘估计, 对数据中的每个观测样本赋予相同的权重, 但如果数据存在异常值时, 容易产生估计偏差。面对高维海量数据, 数据来源的日益多元化会使得异常数据集几乎不可避免。为此, 考虑采用 Huber 回归方法以获取更稳健的估计结果。Huber 方法通过引入加权函数, 对较小残差赋予较大权重, 而对较大残差逐步降低权重, 从而有效降低异常值的影响。该方法结合了均方误差 (mean squared error, MSE) 和绝对误差 (mean absolute error, MAE) 特性, 使得参数估计量对离群点更具稳健。根据文献 [16], 该模型中真参数向量 β_0 在理论上可通过求解以下方程最小化的根:

$$\beta_0 = \arg\min_{\beta} E[L(Y - X'\beta_0)] \tag{2}$$

式中, $L(u) = u^2/2 I_{(|u|\leqslant\gamma)} + (\gamma|u| - \gamma^2/2)\, I_{(|u|>\gamma)}$ 为 Huber 损失函数。

在分布式系统中, 样本量 n 通常非常大。因此, 求解式 (2) 中的优化问题是不可行的。因此, 根据 Jordon[4] 等的研究, 使用通信有效的代替损失函数作为全局似然函数的代替, 可以得到了以下通信高效的估计量, 进而可以缓解分布式系统带来的统计困难。因此, 将通信高效的分布式方法扩展到 Huber 回归。

具体步骤如下:

步骤 1　根据 C_1 计算初始估计量

$$\hat{\beta}^0 = \arg\min_{\beta} \sum_{i=1}^{n_1} L(y_{1i} - x'_{1i}\beta) \tag{3}$$

步骤 2　计算对应其余 $K-1$ 个子机器的梯度 $\nabla L_k(\hat{\beta}^0)$, 并传输到主机器中。

在主机器中通过

$$\hat{\beta}^1 = \hat{\beta}^0 - [\nabla^2 L_1(\hat{\beta}^0)]^{-1} \sum_{k=1}^{K} \nabla L_k(\hat{\beta}^0) \tag{4}$$

式中:

$$\nabla L_1(\hat{\beta}^0) = -\sum_{i=1}^{n_1} \nabla L(y_{1i} - x'_{1i}\hat{\beta}^0) x_{1i},$$

$$\nabla L(y_{1i} - x'_{1i}\hat{\beta}^0) = (y_{1i} - x'_{1i}\hat{\beta}^0) 1_{(|y_{1i} - x'_{1i}\hat{\beta}|\leqslant\gamma)} + \gamma \mathrm{sign}(y_{1i} - x'_{1i}\hat{\beta}^0) 1_{(|y_{1i} - x'_{1i}\hat{\beta}^0|>\gamma)},$$

$$\nabla^2 L_1(\hat{\beta}^0) = \sum_{\mathrm{i}=1}^{n_1} 1_{(|y_{1i} - x'_{1i}\hat{\beta}^0|\leqslant\gamma)} x_{1i} x'_{1i},$$

$$L_k(\hat{\beta}^0) = \sum_{i=1}^{n_k} L(y_{ki} - x'_{ki}\hat{\beta}^0)。$$

步骤 3　迭代直到收敛。

综上, 现在引入一种估计 β_0 的分布式稳健

Huber 回归估计方法。该过程在表 1 的算法流程表中进行总结。

表 1 高效通信分布式 Huber 算法
Tab. 1 Communication-efficient distributed Huber algorithm

算法 1	
输入	子集 $C_k = \{(y_{ki}, x_{ki}), i = 1, 2, \cdots, n_k, k = 1, 2, \cdots, K\}$, γ, 循环次数 q
1	基于机器 C_1 计算初始的估计量 $\hat{\beta}^0$
2	for $r = 1 : q$ do

3	for $k = 1, 2, \cdots, K$ do
	将 C_1 中的估计量 $\hat{\beta}^{r-1}$ 传输到 C_k 中, 计算 C_k 局部梯度 $\nabla L_k(\hat{\beta}^{r-1})$ 并传输到主机器 C_1
	end for

4	C_1 中计算对应估计量 $\hat{\beta}^r = \hat{\beta}^{r-1} - \nabla^2 L_1(\hat{\beta}^{r-1})^{-1} \sum_{k=1}^{K} \nabla L_k(\hat{\beta}^{r-1})$, 再判断是否满足终止条件。
	end for
输出	最终估计量, 记作 $\hat{\beta}$

2 干净子集

2.1 干净子集的识别

随着分布式系统规模的不断扩大, 无异常子集或称干净子集成为了一个迫切需要解决的问题。为了提高识别准确性, 提出了一种新的方法, 该方法在克服了传统单点诊断方法在掩盖和淹没问题上的局限性, 并借鉴群体删除策略来识别一个无异常的服务器子集。这一策略不仅简化了异常判断过程, 还显著提高了计算效率。在此过程中, 结合了 Wang 等 [15] 的研究, 其主要工作依托 SIS 降维和 FAST-LTS 算法的 Component-step 步骤来寻找干净子集。在降维过程中, 通过稳健回归估计未知参数, 再计算并比较残差的大小, 区分服务器子集, 从而有效地提高诊断的准确性。

对每个服务器 C_k, $k = 1, 2, \cdots, K$, 利用子集的观测 (y_{ki}, x_{ki}), $i = 1, 2, \cdots, n_k$, $k = 1, 2, \cdots, K$ 计算对应残差, $e_{ki} = y_{ki} - x'_{ki}\hat{\beta}$, $i = 1, 2, \cdots, n_k$, $k = 1, 2, \cdots, K$, 进而可以求得服务器 C_k 的平均绝对残差:

$$\bar{e}_k = \frac{1}{n_k} \sum_{i=1}^{n_k} |e_{ki}| = \frac{1}{n_k} \sum_{i=1}^{n_k} \left| y_{ki} - x'_{ki}\hat{\beta} \right| \tag{5}$$

对 $\bar{e}_k, k = 1, 2, \cdots, K$ 进行递增排序, 表示为 $\bar{e}_{(1)}, \bar{e}_{(2)}, \cdots, \bar{e}_{(K)}$。

定义 1 所有服务器索引子集 (子集内元素数量规定为 h) 的集合定义为

$$H = \{H \subset \{1, 2, \cdots, K\}, |H| = h\} \tag{6}$$

式中, $|H|$ 为集合 H 元素的数量。

由于数据集中正常子集的数量通常多于异常子集, 在本文中选择了 $h = [\varsigma K]$ (其中 $[u]$ 表示 $[u]$ 整数部分)。Huber[16] 的研究侧重于高维线性回归模型中的异常值检测, 因此他们较为保守地选择了干净子集占据全集的比例为 $\varsigma = 0.55$。与之不同, 估计量的研究采用了稳健回归方法, 该方法通过赋予异常值更小的比例, 进而对估计量更加稳健, 那么对于干净子集的比例完全可以选择为 $\varsigma = 0.8$, 并通过数值研究进行确认。

2.2 干净子集的选取

设 $\bar{e}_i(H)$ 表示服务器子集 H 内第 i 个服务器的平均绝对残差。定义 K 个服务器干净服务器子集为 H, 即最小的 h 个 $\bar{e}_i(H)$ 的集合。具体过程如下:

(1) 基数 $h = [\varsigma K]$, 其中 $\varsigma = 0.8$ 为干净子集对全集的比例。

(2) 计算每个服务器的平均绝对残差 $\bar{e}_k$。

(3) 对所有服务器的平均绝对残差进行排序, 表示为 $\bar{e}_{\pi(1)}, \bar{e}_{\pi(2)}, \cdots, \bar{e}_{\pi(K)}$, 其中 π 是对 $\{1, 2, \cdots, K\}$ 的一个排序, 再选择其中的干净服务器集合 $H = \{\pi(1), \pi(2), \cdots, \pi(h)\}$。通过分析残差的大小, 可以筛选出在稳健回归下无异常值的服务器集合。

3 子集诊断检测方法

获得干净子集 H 之后, 以边际相关作为影响特征判断剩下的服务器是否异常。

3.1 边际相关的影响特征

对于服务器 $C_k, k = 1, 2, \cdots, K$ 上的观测值 $(x_{ki}, y_{ki}), i = 1, 2, \cdots, n_k$, 计算预测变量 X_j(表示 X 的第 j 个分量) 与响应变量 Y 的均值为 $\bar{X}_{kj} = 1/n_k \sum_{i=1}^{n_k} x_{kij}, \bar{Y}_k = 1/n_k \sum_{i=1}^{n_k} y_{ki}, j = 1, 2, \cdots, p$, 则基于干净服务器子集, 预测变量 X_j 与响应变量 Y 之间的边际相关性 H:

$$\omega_{j(H)} = E\{(\bar{X}_{j(H)} - \mu_{\bar{X}_{j(H)}})(\bar{y}_{(H)} - \mu_{\bar{y}_{(H)}})\} / (\sigma_{\bar{X}_{j(H)}} - \sigma_{\bar{y}_{(H)}}), \quad j = 1, 2, \cdots, p \tag{7}$$

式中: $\bar{X}_{j(H)}$ 为 H 集合中 X_j 的观测值的平均; $\bar{y}_{(H)}$

为 H 集合中 Y 的观测值的平均

$$\mu_{\bar{X}_{j(H)}} = E(\bar{X}_{j(H)}), \quad \mu_{y_{(H)}} = E(\bar{y}_{(H)})$$
$$\sigma^2_{\bar{X}_{(H)}} = \mathrm{var}(\bar{X}_{j(H)}) \quad \sigma^2_{\bar{y}_{(H)}} = \mathrm{var}(\bar{y}_{(H)})$$

令 $\hat{\omega}_{j(H)}$ 是 X_j 与 Y 基于干净服务器子集 H 的样本边际相关性:

$$\hat{\omega}_{j(H)} = \sum_{k\in H}(\bar{X}_{kj} - \bar{X}_j)(\bar{Y}_k - \bar{Y})/ \sqrt{\sum_{k\in H}(\bar{X}_{kj} - \bar{X}_j)^2 \sum_{k\in H}(\bar{Y}_k - \bar{Y})^2} \tag{8}$$

式中: $\bar{X}_j = \dfrac{1}{h}\sum\limits_{k\in H}\bar{X}_{kj}$; $\bar{Y} = \dfrac{1}{h}\sum\limits_{k\in H}\bar{Y}_k$。

参照 Cook 距离, 改进提出 “add-one-into” 方法。当第 $T(T \in \{1,2,\cdots,K\} - H)$ 个服务器被加到干净服务器子集中时, 预测变量 X_j 与响应变量 Y 之间的边际相关性的一致估计量表示为:

$$\hat{\omega}^{(T)}_{j(H)} = \frac{\sum\limits_{k\in H\cup\{T\}}(\bar{X}_{kj} - \bar{X}^{(T)}_j)(\bar{Y}_k - \bar{Y}^{(T)})}{\sqrt{\sum\limits_{k\in H\cup\{T\}}(\bar{X}_{kj} - \bar{X}^{(T)}_j)^2 \sum\limits_{k\in H\cup\{T\}}(\bar{Y}_k - \bar{Y}^{(T)})^2}} \tag{9}$$

式中: $\bar{X}^{(T)}_j = \dfrac{1}{h+1}\sum\limits_{k\in H\cup\{T\}}\bar{X}_{kj}$; $\bar{Y}^{(T)} = \dfrac{1}{h+1}\sum\limits_{k\in H\cup\{T\}}\bar{Y}_k$。

3.2　高维多异常点群检测

定义 2　高维异常值检测的统计量为:

$$D_T = \frac{1}{p}\sum_{j=1}^{p}(\hat{\omega}_{j(H)} - \hat{\omega}^{(T)}_{j(H)})^2 \tag{10}$$

当第 T 个服务器加入, $\hat{\omega}_{j(H)}$ 与 $\hat{\omega}^{(T)}_{j(H)}$ 之间的差异显著时, 就将第 T 个服务器标记为异常。为确定诊断统计量的临界值, 参照 Zhao 等 [14] 提出的 HIM 方法。鉴于大数据的样本量 N 和固定的维度 p 趋于无穷大, 进而将建立诊断统计量 D_T 的渐近分布。为此, 假设满足以下条件:

$C1$ 对于固定的 $j = 1,2,\cdots,p$, $\omega_{j(H)}$ 是一个不随 p 变化而变化的常数。

$C2$ 预测因子 X 服从多元正态分布, 随机噪声 ε 服从正态分布。

定理 1　在满足 $C1$ 和 $C2$ 情形下, 可以得到, 当没有异常且 $\min\{h,p\} \to \infty$ 时, 有

$$h^2 D_T \to \chi^2(1) \tag{11}$$

式中, $\chi^2(1)$ 是自由度为 1 的卡方分布。证明见 Zhao 等 [14] 的定理 1。

由此, 可以在多重假设检验的框架下识别多个异常服务器。具体以第 T 台服务器为例, 设定原假设和备择假设来制定 $K-h$ 个假设检验。

H_{0T} : 不是异常服务器; $\leftrightarrow$ H_{1T} : 是异常服务器。

在显著性水平 α 下, 如果 $h^2 D_T > \chi^2_{1-\alpha}(1)$, 则第 T 台服务器被识别为异常, 其中 $\chi^2_{1-\alpha}(1)$ 是卡方分布的上 $1-\alpha$ 分位数。

基于上述程序, 得到一个相对干净的服务器集合 H 后, 针对第 T 台服务器, 计算诊断统计量 D_T。

表 2　子集检测方法
Tab. 2　Subset detection method

算法 2	
输入	子集 $C_k = \{(y_{ki}, x_{ki}), i = 1,2,\cdots,n_k, k = 1,2,\cdots,K\}$, 估计 $\hat{\beta}$, $h = [\varsigma K]$
1	for $k = 1,2,\cdots,K$ do 将 $\hat{\beta}$ 传输到每个子机器 C_k
2	计算每个样本的残差 e_{ki}, 计算获得子机器 C_k 的 $\bar{e}_k$, $\bar{X}_{kj}$, $\bar{Y}_k$ 传输到主机器
3	在主机器上利用干净子集定义 h, 对 $\bar{e}_k$ 排序, 获得干净子集 H。
4	在主机器上, 计算干净子集的边际相关性 $\hat{\omega}_{j(H)}$

5	for $T = 1,2,\cdots,K-h$ do
	计算 $\bar{X}^{(T)}_j$, $\bar{Y}^{(T)}_j$ 以及 $\hat{\omega}^{(T)}_{j(H)}$, 再计算统计量 D_T
	end for

6	假设检验, 设定显著性水平 α
	for $T = 1,2,\cdots,K-h$ do
	若 $h^2 D_T > \chi^2_{1-\alpha}$ (1), 则第 T 个服务器被识别为异常
	end for

输出	异常服务器序号

4　数值研究

在本节中, 使用蒙特卡罗模拟研究来评估所提出在分布式稳健回归下的子集检测方法的有效性能。所有程序都是用 R 代码编写的。研究中, 由于缺少服务器, 无法实现真正的海量数据分布式存储。因此, 使用一台机器来分析模拟数据和真实数据的应用, 通过分块计算作为补偿替代, 但是研究结果仍然能说明所提方法的有效性。

4.1　数值设置

在本节中, 首先采用了 Huber 稳健回归法, 进行

前期参数估计计算, 为了模拟在分布式系统中处理海量数据的特点, 研究沿用了分而治之的方法, 将完整数据集分成多个小数据集, 每个小数据集的处理方法代表了局部服务器所需的计算方法。同时, 为了方便评估检测异常服务器的方法性能, 下面引入两个指标: 正确检测异常服务器的比例 (correct detection rate, CDR) 和将正常服务器误判为异常的比例 (false detection rate, FDR)。公式如下:

$$\mathrm{CDR}=\frac{\text{正确检测异常机器数}}{\text{真实异常机器数}} \tag{12}$$

$$\mathrm{FDR}=\frac{\text{正常机器误判数}}{\text{真实正常机器数}} \tag{13}$$

本节模拟数据皆由以下线性模型生成:

$$Y_i=X_i'\beta_0+\varepsilon_i,\quad i=1,2,\cdots,n \tag{14}$$

式中: $Y_i\in\mathbb{R}$ 是响应变量; β_0 为真实参数; $X_i=(1,X_{i1},X_{i2},\cdots,X_{ip})^{\mathrm{T}}$, 是一个 $p+1$ 维协变量; $(X_{i1},X_{i2},\cdots,X_{ip})$ 服从均值为 0, 方差为 1 的多元正态分布, 协方差矩阵为 $\sum$, 其中 $\sum\limits_{ij}=0.5^{|i-j|}$, $1\leqslant i,j\leqslant p$, ε_i 为误差; Huber 估计中的 $\gamma=\sqrt{n/p}$。

对数据进行均值扰动和方差扰动, 进而模拟设置异常服务器, 在不同的服务器 K 和对应 p 下, 检验本异常服务器检测法的准确率、适用性和稳健性。

4.2 数值案例研究

4.2.1 案例 1 固定异常服务器个数

设置总样本量 $N=4\times10^5$, 维度 $p=10$, 真参数 $\beta_0=(1,1,1,1,1,0,0,0,0,0,0)$, 异常服务器个数 $ff=10$ 和 $ff=0.1K$, 服务器个数依次设置为 $K=\{50,100,200\}$, 在对应个数下, 服务器随机均分总样本量。模型误差项服从标准正态分布。研究通过模拟实验评估分布式环境下异常服务器的检测性能, 重点考察了均值扰动与方差扰动两种异常类型的检测效果。对两种情形下的异常服务器由 Y_i 改到 $Y_i\times[1+N(0.5,0.5)]$ 和 $Y_i\times[1+N(1,2)]$, 模拟 500 次可得两类方差扰动见表 3。此外, 通过对两种情形下的异常服务器由 Y_i 改到 $Y_i+N(0.5,0.5)$ 和 $Y_i+N(1,2)$, 模拟 500 次可得两类均值扰动见表 4。

从表 3、4 可见, 实验结果表明, 针对均值扰动的检测方法表现出近乎完美的准确率: 在固定异常服务器数量 (10 个) 或比例 (10%) 的设置下, CDR 始终高于 0.975, 且 FDR 接近于零 ($\leqslant$0.001)。这一稳健性表明该方法对参数均值偏移类异常具有极强的识别能力。对于方差扰动, 尽管 CDR 在所有场景下均达到 1.000, 但 FDR 随服务器总数增加呈现上升趋势: 当服务器数从 50 增至 200 时, FDR 由 0.000 升至 0.158, 表明细粒度划分服务器可能引入误判风险。进一步分析显示, 扰动强度的增加可有效提升检测效率, 尤其在方差扰动场景中, 其 CDR 随扰动强度提升至理论最优值。综合而言, 在固定总样本数时, 所提方法在均值扰动检测中表现卓越, 而在方差扰动检测中需权衡服务器划分数量和 FDR。

表 3 案例1: 方差扰动
Tab. 3 Case 1: variance perturbation

ff	K	方差扰动 $Y_i\times[1+N(0.5,0.5)]$		方差扰动 $Y_i\times[1+N(1,2)]$	
		CDR	FDR	CDR	FDR
10	50	1.000	0.000	1.000	0.000
	100	1.000	0.111	1.000	0.148
	200	1.000	0.145	1.000	0.158
$0.1K$	50	1.000	0.076	1.000	0.000
	100	1.000	0.111	1.000	0.147
	200	1.000	0.157	1.000	0.158

表 4 案例1: 均值扰动
Tab. 4 Case 1: mean perturbation

ff	K	均值扰动 $Y_i+N(0.5,0.5)$		均值扰动 $Y_i+N(1,2)$	
		CDR	FDR	CDR	FDR
10	50	0.975	0.000	0.995	0.000
	100	0.994	0.000	1.000	0.000
	200	0.988	0.001	1.000	0.001
$0.1K$	50	0.988	0.001	0.998	0.001
	100	0.994	0.001	1.000	0.001
	200	0.987	0.000	1.000	0.000

4.2.2 案例 2 固定子机器样本量

设置子服务器样本量 $n=400$, 维度 $p=50$, 真参数 $\beta_0=(1,1,1,1,1,0,0,0,0,0,0)$, 服务器个数依次设置为 $K=\{200,300,400\}$, 在对应个数下, 总样本量依次为 $N=\{8\times10^4,1.2\times10^5,1.6\times10^5\}$。模型误差项服从标准正态分布。基于均值扰动和方差扰动来构造异常服务器。通过对 10 个和 $0.05K$ 个异常服务器由 Y_i 改到 $Y_i+N(1,1)$ 和由 Y_i 改到 $Y_i\times[1+N(1,1)]$, 模拟 500 次可得表 5。

从表 5 可见, 在两类扰动和不同子服务器数目下, 模型的分类检测准确率 CDR 接近 1, 误检率 FDR 接近 0, 表明模型在大多数情况下能够准确识别异常服务器且误判几率极低。在相同条件下, 均

值扰动下的检测效果优于方差扰动, 具体表现为均值扰动下的 CDR 更高, FDR 也显著降低。

表 5　案例 2
Tab. 5　Case 2

ff	K	均值扰动 $Y_i+N(1,1)$		方差扰动 $Y_i\times[1+N(1,1)]$	
		CDR	FDR	CDR	FDR
10	200	0.970	0.012	0.911	0.015
	300	0.999	0.011	0.937	0.012
	400	0.999	0.009	0.940	0.010
$0.05K$	200	0.971	0.013	0.864	0.014
	300	0.997	0.017	0.953	0.018
	400	0.999	0.018	0.980	0.020

总体而言, 随着子服务器数量的增加, 检测方法的稳健性和准确性得到了提升。尤其在均值扰动的情况下, 模型能够更好地适应数据分布的变化, 表现出较高的检测能力。在两类扰动和不同子服务器数目下, CDR 始终接近 1 且 FDR 接近 0, 说明模型在应对不同扰动类型时保持了较高的精确度和稳定性。

5　结　语

本文提出了一种基于边际相关性的分布式子集的统计诊断方法, 为了解决多个影响点的识别问题并避免淹没效应, 采用 “add-one-in” 方法进行子集诊断, 以处理高维数据中的异常点群。考虑到数据源多样性和重尾噪声的影响, 研究采用 Huber 回归方法, 以提升重尾分布下的稳健性。此方法为高维大数据下的分布式稳健分析和子集检测提供了有效解决方案。

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] BATTEY H, FAN J Q, LIU H, et al. Distributed testing and estimation under sparse high dimensional models [J]. The Annals of Statistics, 2018, 46(3): 1352-1382.

[3] JIANG R, YU K M. Smoothing quantile regression for a distributed system [J]. Neurocomputing, 2021, 466: 311-326.

[4] JORDAN M I, LEE J D, YANG Y. Communication-efficient distributed statistical inference [J]. Journal of the American Statistical Association, 2019, 114: 668-681.

[5] COOK R D. Influential observations in linear regression [J]. Journal of the American Statistical Association, 1979, 74(365): 169-174.

[6] COOK R D. Detection of influential observation in linear regression [J]. Technometrics, 2000, 42(1): 65-68.

[7] ATKINSON A C, COOK R D, WEISBERG S. Residuals and influence in regression [J]. Biometrics, 1983, 39(3): 818.

[8] ANDERSEN E B. Diagnostics in categorical data analysis [J]. Journal of the Royal Statistical Society: Series B: Statistical Methodology, 1992, 54(3): 781-791.

[9] THOMAS W, COOK R D. Assessing influence on regression coefficients in generalized linear models [J]. Biometrika, 1989, 76(4): 741-749.

[10] PREISSER J S, QAQAISH B F. Deletion diagnostics for generalised estimating equations [J]. Biometrika, 1996, 83(3): 551-562.

[11] BANERJEE M, FREES E W. Influence diagnostics for linear longitudinal models [J]. Journal of the American Statistical Association, 1997, 92(439): 999-1005.

[12] XIANG L M, TSE S K, LEE A H. Influence diagnostics for generalized linear mixed models: applications to clustered data [J]. Computational Statistics & Data Analysis, 2002, 40(4): 759-774.

[13] PAN J X, FANG K. Growth curve models and statistical diagnostics [M]. New York: Springer, 2002.

[14] ZHAO J L, LIU C, NIU L, et al. Multiple influential point detection in high dimensional regression spaces [J]. Journal of the Royal Statistical Society, 2019, 81(2): 385-408.

[15] WANG T, LI Q, CHEN B, et al. Multiple outliers detection in sparse high-dimensional regression [J]. Journal of Statistical Computation and Simulation, 2017, 88(1): 89-107.

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

[17] SUN Q, ZHOU W X, FAN J Q. Adaptive Huber regression [J]. Journal of the American Statistical Association, 2020, 115(529): 254-265.

[18] 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.

[19] LUO J, SUN Q, ZHOU W. Distributed adaptive Huber regression [J]. Computational Statistics & Data Analysis, 2022, 169: 107419.

[20] 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(4): 633-646.

[21] JIANG R, ZHAO Y B. Online updating mode learning for streaming datasets [J]. Journal of Statistical Computation and Simulation, 2024, 94(12): 2697-2709.