

華中科技大學

本科毕业设计[论文]

基于分位回归的分布式大数据 的实证建模分析

院 系 数学与统计学院

专业班级 统计 1801 班

姓 名 周宇杰

学 号 U201810112

指导教师 李楚进

2022 年 05 月 20 日

学位论文原创性声明

本人郑重声明：所呈交的论文是本人在导师的指导下独立进行研究所取得的科研成果。除了文中特别加以标注引用的内容外，本论文不包括任何其他个人或集体已经发表或撰写的成果作品。本人完全意识到本声明的法律后果由本人承担。

作者签名：周家杰 2022年05月21日

学位论文版权使用授权书

本学位论文作者完全了解学校有关保障、使用学位论文的规定，同意学校保留并向有关学位论文管理部门或机构送交论文的复印件和电子版，允许论文被查阅和借阅。本人授权省级优秀学士论文评选机构将本学位论文的全部或部分内 容编入有关数据进行检索，可以采用影印、缩印或扫描等复制手段保存和汇编本学位论文。

本学位论文属于 1、保密 ☐，在 年解密后适用本授权书。

2、不保密 ☒

(请在以上相应方框内打“√”)

作者签名：周家杰 2022年05月21日

导师签名：李赞 2022年5月21日

摘 要

随着信息技术的发展,数据的体量越来越大,人们也着眼于将分位回归模型应用于大规模数据。针对大规模数据分布式储存的问题,为了提高模型通信效率,各种应用于分布式大规模数据的分位回归模型被提出。

本文基于一种已提出的分布式数据通信高效的惩罚分位回归模型,对多个专业领域数据进行建模拟合,并运用 ADMM 算法进行模型求解,进行参数估计,进一步验证分布式数据通信高效的惩罚分位回归模型的效率和准确性,并进一步比较该方法与传统方法之间的差异,展示出该方法在参数估计及高维分位回归中变量筛选的优良表现。

考虑到分布式数据不同机器间的差异性,本文也对模型做出了一些改进,改进模型在对来自不同机器的数据用于计算损失函数时赋予了不同权重。我们对同一实际数据分别使用改进模型与原模型进行拟合,验证了改进后的模型预测误差精度相较于原模型有显著提高。

关键词: 惩罚分位回归; 分布式高维数据; 通信高效的替代似然; ADMM 算法

Abstract

With the development of information technology, the volume of data is getting larger and larger. Researchers tend to apply quantile regression model to large-scale data. To solve the problem of communication efficiency, various quantile regression models applied to distributed large-scale data are proposed.

In this article, based on the proposed communication-efficient modeling with penalized quantile regression for distributed data, we use this model to simulate some pieces of specific field data and apply ADMM algorithm to solve the model and estimate parameters. Then we further verify of the efficiency and accuracy of communication-efficient modeling with penalized quantile regression for distributed data and compare the difference between this model and traditional models to show the excellent performance of this model in parameter estimation and variable selection in high-dimensional quantile regression.

Considering the differences between different machines of distributed data, we also make some improvements to the original model. The improved model assigns different weights to the data from different machines used to calculate the loss function. Then we use the improved model and the original model respectively to fit the same real data, which verifies that the prediction error accuracy of the improved model is significantly improved compared with the original model.

Key Words: Penalized quantile regression; Distributed high dimensional data; Communication-efficiency surrogate likelihood; ADMM algorithm

目 录

摘 要	I
Abstract.....	II
1 绪论	1
1.1 研究背景.....	1
1.2 研究意义.....	1
1.3 国内外研究现状.....	2
1.4 研究内容.....	3
2 基础理论知识	5
2.1 惩罚分位回归模型.....	5
2.2 ADMM 算法	7
2.3 近端 ADMM 算法.....	8
2.4 模型算法流程.....	10
2.5 改进的惩罚分位回归模型.....	10
2.6 本章小节.....	11
3 实证建模分析	12
3.1 模型在低维数据中的应用——基于气体排放数据.....	12
3.1.1 数据来源及说明.....	12
3.1.2 模型设定.....	12
3.1.3 实验结果及分析.....	13
3.2 改进模型的实证分析——基于气体排放数据.....	15
3.2.1 模型设定.....	15
3.2.2 实验结果及分析.....	16
3.3 模型在高维稀疏数据中的应用——基于美国犯罪率数据.....	18
3.3.1 数据来源及说明.....	18
3.3.2 模型设定.....	19
3.3.3 实验结果及分析.....	19
3.4 本章小结.....	21
4 总结与展望	23

致谢	24
参考文献	26
附录 美国犯罪率数据变量被选择频率	28

1 绪论

1.1 研究背景

“数字化”时代的来临,让现代生活的方方面面都与数据息息相关,而随着科技的进步,日常生活中产生的数据的体量也越来越大,对于大规模数据的处理分析也成为当下的热点问题。人们也在运用分位回归[1],一种自被 Koenker 等提出以来就广泛应用于各行各业的回归方法,对大规模数据进行分析、拟合、预测。

大规模数据由于体量巨大、数据采集地分散等原因通常存储于不同机器甚至于不同地区[2],而分布式存储的数据通常会给分位回归模型的计算拟合带来如下两个问题: 1. 对于存储于不同机器上的数据,在用于计算分位回归模型的损失函数时,尤其是处理高维数据时,由于数据量巨大,在不同机器间传递数据会产生较高的通信成本,导致模型算法时间复杂度和空间复杂度均较高。且许多真实的高维数据是稀疏数据,在机器间传输这些数据,通信效率也很低,导致分位回归模型在高维数据中应用效果并不理想[3]。 2. 不同机器上存储的数据存在差异,然而在实际应用中,大多数模型均忽略了这种差异性,计算损失函数时并不对来自不同机器的数据给予不同权重以体现不同机器存储的数据的差异,这也在某种程度上导致了模型预测误差较大,模型精度不高。

为了解决问题 1, Hu 等[4]提出了分布式数据通信高效的惩罚分位回归模型,大大提高了通信效率且模型的统计精度仍有保障。本文将基于这个模型,对多组数据,进行实证建模分析,以评估模型的性能。

1.2 研究意义

通过对不同领域数据进行建模拟合,进一步验证分布式数据通信高效的惩罚分位回归模型的效率和准确性,并比较该方法与传统方法之间的差异,展示出该方法在参数估计及高维分位回归中变量筛选的优良表现。同时,对于模型在实际数据拟合时出现的问题,加以分析和解释。

为了解决分位回归在分布式数据应用中不同机器间数据差异导致的预测精度问题,在实际应用时,对分布式数据通信高效的惩罚分位回归模型加以改进,考虑不同机器上存储数据的差异性,对来自不同机器的数据,计算损失函数时,结合数据自身特性,赋予不同权重。将改进后模型预测误差与原模型进行对比,分析相应结果,尽可能提升分位回归模型的预测精度。

1.3 国内外研究现状

Jordan 等[5]提出了一个通信高效的替代似然(CSL)框架来解决分布式统计推断问题。CSL 为全局似然提供了一种通信高效的替代方法,可用于低维估计、高维正则估计和贝叶斯推断。对于低维估计,可以证明 CSL 改进了朴素平均方法,并有助于置信区间的构造。对于高维正则估计,CSL 得到了通信代价可控的极小化极大最优估计。对于贝叶斯推断,CSL 可以用来形成一个通信高效的准后验分布,该分布收敛于真实的后验分布。这种准后验过程即使在非分布情形下也显著提高了马尔可夫蒙特卡罗(MCMC)算法的计算效率。因此,CSL 只需要交换大量的局部数据梯度,即可有效地降低分布式数据的传输成本。此外,该方法的估计具有与全局似然估计相同的收敛速度。

Chen 等[6]研究了高维分布分位回归,为了处理方差为无穷大的重尾噪声,采用分位回归损失函数代替常用的平方损失函数。在对分位数进行估计时,其思想是将分位数估计转换为转化后的响应变量和协变量之间的最小二乘估计。但是,计算需要估计误差的密度函数和协变量的协方差矩阵。

基于以上文献, Hu 等[4]提出了分布式数据通信高效的惩罚分位回归模型,用惩罚分位回归处理稀疏高维数据。在每一轮中,该方法只需要主机处理稀疏惩罚分位回归,该分位回归模型可以通过近端交替方向乘子法(ADMM)快速求解,而其他工作机器则只需要计算局部数据的次梯度。在通信效率方面,只要适当选择惩罚水平,该方法不牺牲任何统计精度,并可证明地改善了集中式方法获得的估计误差。

Di 等[7]主要研究了复合分位回归(CQR)的分布估计和统计推断。为了提高计算效率,他们将平滑思想应用于分布式数据的 CQR 损失函数,然后通过多轮聚合依次细化估计量。在 Bahadur 表示下,得到了所提出的多轮光滑 CQR 估计

的渐近正态性, 并通过同时分析整个数据集证明了它也能达到理想 CQR 估计的效率。此外, 为了提高 CQR 算法的效率, 他们提出了一种多轮平滑加权算法, 只要求初始值一致, 其余操作都是方便的矩阵操作。

Wang 等[8]提出了一种通信高效的替代分位回归框架, 克服了海量数据非随机分布的缺点。基于从不同机器上收集的小尺寸随机试点样本, 由主服务器上的一个代理函数近似全局分位回归目标函数。

Tan 等[9]提出并研究了一种新的基于卷积光滑和迭代加权的 l_1 -惩罚高维分位回归稀疏模型。为了处理非光滑问题, 其通过卷积来平滑分段线性分位数损失函数, 主要思想是对分位数损失函数的次梯度进行平滑处理, 然后对其进行积分得到光滑的、凸的损失函数。

Wang 等[10]考虑具有高维线性部分的分位回归中的半非参数学习和一个假定在再生核希尔伯特空间中的非参数函数。利用 Rademacher 复杂度作为主要工具, 建立条件分位数的总速率, 其等于非参数回归的最优速率与线性回归的最优速率之和, 因此一般无法提高。

基于多篇文献的研究成果, 不难发现, 研究趋势主要集中在以下几个方面: 一、尽可能地提高分位回归模型的精度, 寻找具有更好渐进性的统计量, 实现了最优的统计收敛速度, 改进估计误差。有的人[11]采用分治(DC)方法是将一个复杂的问题分解成若干个规模较小、相互独立, 但类型相同的子问题, 以便各个击破, 分而治之。也有通过通信效率替代似然(CSL)方法为全局似然性提供了一种通信高效的替代方法。二、改进模型求解算法, 采用更加先进快速的算法求解模型。随着机器学习与神经网络算法的发展, 不少研究也试图采用一些与机器学习相关的算法, 进行模型求解。三、拓展分位回归模型的应用领域。从经济到计算机再到生物医学领域, 各种不同背景的数据都试图使用分位回归理论建立相应模型, 测试分位回归模型对不同特性数据的适应性, 极大拓宽了分位回归的应用。

1.4 研究内容

本文对分布式数据通信高效的惩罚分位回归模型的提出背景进行了分析, 讨论了国内外将分位回归模型运用于大规模数据的研究现状, 并对分布式数据

通信高效的惩罚分位回归模型的相关基础知识进行总结。同时,考虑到分布式数据不同机器上存储数据的差异性,对原有模型加以改进,在使用来自不同机器的数据,计算损失函数时,根据数据的重要性不同,赋予不同权重。

本文余下部分结构如下:3.1 节中,我们将分布式数据通信高效的惩罚分位回归模型应用于低维实际数据,进行拟合预测。3.2 节中,我们对改进模型进行实证分析,并将结果与原模型进行了对比,讨论了改进模型在实际应用中需要注意的问题。3.3 节中,我们使用分布式数据通信高效的惩罚分位回归模型拟合高维稀疏的实际数据,检验其参数拟合与变量选择的性能。在第 4 章中,我们总结了第 3 章实证分析的结果。

2 基础理论知识

2.1 惩罚分位回归模型

本小节中, 先介绍通信高效的惩罚分位回归模型[4]:

设 $\{y_i, x_i\}_{i=1}^N$ 为 N 个样本观测值, 其中, y 为响应变量, $x = (x_1, \dots, x_p)^T$ 为 p 维协变量, 这里 p 和 N 都非常大。假设数据均匀存储于 k 台机器中, 用 $\{(x_{ji}, y_{ji}): i = 1, \dots, n\}$ 表示存储于第 j 个机器的子样本, 则 $N = nk$ 表示总样本观测数。

对于给定协变量 $x = (x_1, \dots, x_p)^T$, 考虑线性模型 $Y = x^T \beta + \varepsilon$, 响应变量 Y 的第 τ ($0 < \tau < 1$)个分位数是关于 $x = (x_1, \dots, x_p)^T$ 的线性函数, 因此有:

$$Q_\tau(Y|x) = x^T \beta_0(\tau),$$

其中 $P(\varepsilon \leq 0|x) = \tau$, 将 Y 的条件分布函数记为 $P(Y \leq y|x_i) = F_Y(y|x_i) = F_i(y)$, 则 $Q_\tau(Y|x_i) = F_i^{-1}(\tau|x_i) \equiv \xi_i(\tau)$ 。

因此, 分位回归模型就是求解以下优化问题:

$$\hat{\beta} = \arg \min_{\beta} \sum_{i=1}^N \rho_\tau(y_i - x_i^T \beta), \quad (2-1)$$

其中 $\rho_\tau(t) = t(\tau - 1_{\{t \leq 0\}})$, 是非对称绝对偏差函数[12]。这里将 $\beta_0(\tau)$ 和 $\xi_i(\tau)$ 中的 τ 省略, 分别记作 β_0 和 ξ_i 。

局部和全局分位回归损失函数定义如下:

$$L_j(\beta) = \frac{1}{n} \sum_{i=1}^n \rho_\tau(y_{ji} - x_{ji}^T \beta), \quad j = 1, \dots, k, \quad (2-2)$$

$$L_N(\beta) = \frac{1}{k} \sum_{j=1}^k L_j(\beta) = \frac{1}{N} \sum_{j=1}^k \sum_{i=1}^n \rho_\tau(y_{ji} - x_{ji}^T \beta), \quad (2-3)$$

其中, 使用存储于第 j 个机器 M_j 中的局部数据可以计算出 β 处的 $L_j(\beta)$ 。

我们现在采用分布式通信高效方法进行分位回归, 将以下函数作为替代损失函数:

$$\tilde{L}(\beta) := L_1(\beta) - \langle \beta, \nabla L_1(\beta^0) - \nabla L_N(\beta^0) \rangle,$$

这里, β^0 为 β 的任意初始估计, $\langle \cdot, \cdot \rangle$ 表示内积, $\nabla L_j(\beta)$ 表示 $L_j(\beta)$ 关于 β 的次梯度, 由式(2-2)和(2-3)有:

$$\nabla L_1(\beta) = -\frac{1}{n} \sum_{i=1}^n x_{1i} \psi_\tau(y_{1i} - x_{1i}^T \beta), \quad (2-4)$$

$$\nabla L_N(\beta) = -\frac{1}{N} \sum_{j=1}^k \sum_{i=1}^n x_{ji} \psi_\tau(y_{ji} - x_{ji}^T \beta), \quad (2-5)$$

其中, $\psi_\tau(\mu) = \nabla \rho_\tau(\mu) = \tau I_{(\mu > 0)} + (\tau - 1) I_{(\mu < 0)} + \xi I_{(\mu = 0)}$, $\xi \in [\tau - 1, \tau]$ 。

对于高维分位回归, 这里考虑如下分位回归的 L_1 -惩罚加权替代估计:

$$\min_{\beta} \tilde{L}(\beta) + \sum_{j=1}^p p_\lambda(|\beta_j|),$$

作为(2-1)中 $\hat{\beta}(\tau)$ 的替代估计, 其中 $\lambda > 0$ 是正则化参数。

下面, 本文使用凸的 ALasso(自适应 Lasso)罚函数[13]和非凸的 SCAD(平滑剪切绝对偏差)罚函数[14]来展示该模型的含义。对于 ALasso 罚函数, $p_\lambda(|\beta_j|) = w_\lambda(|\beta_j|)$, 其中 $w = (w_1, \dots, w_p)^T$ 是非负权向量, $w_j \geq 0$, $j = 1, \dots, p$, 在恰当选择的 $v > 0$ 后, 通常选取 w_j 为 $w_j = (|\hat{\beta}_j^{lasso}| + 1/n)^{-v}$, $j = 1, \dots, p$, 其中 $\hat{\beta}^{lasso} = (\hat{\beta}_j^{lasso}, j = 1, \dots, p)^T$ 表示分位数 Lasso 估计。

注意到, 上式为凸优化问题, 且计算较为简单。然而, 如果罚函数为非凸的, 例如 SCAD 型罚函数, 则有:

$$\begin{aligned} p_\lambda(|\beta|) = & \lambda(0 \leq |\beta| < \lambda) + \frac{a\lambda|\beta| - \frac{\beta^2 + \lambda^2}{2}}{a-1} \cdot I(\lambda \leq |\beta| < a\lambda) \\ & + \frac{(a+1)\lambda^2}{2} I(|\beta| > a\lambda), \end{aligned}$$

其中, 通常选取 $a = 3.7$ 。虽然非凸罚函数在理论上有很好的应用前景, 但罚函数的奇异性和非凸性带来了计算上的挑战。一些研究说明, 可以用局部线性近似(LLA)来替代非凸罚函数。

然后, 包含凸罚函数和非凸罚函数的惩罚过程转化为如下加权优化问题:

$$\min_{\beta} \tilde{L}(\beta) + \lambda \|w \circ \beta\|_1, \quad (2-6)$$

其中 $\|w \circ \beta\|_1 = \sum_{j=1}^p |w_j \beta_j| = \sum_{j=1}^p w_j |\beta_j|$, 在用 LLA 估计时, 通常取 $w_j = \lambda^{-1} p'_{\lambda}(|\hat{\beta}_j^{s-1}|), j = 1, \dots, p$, 其中 $\hat{\beta}^{s-1} = (\hat{\beta}_j^{s-1}, j = 1, \dots, p)^T$ 为 $(s-1)$ 次迭代的估计值。

2.2 ADMM 算法

许多研究者[15]都使用 ADMM 算法来解决大规模惩罚分位回归模型的计算问题, 这里先介绍原始的 ADMM 算法[16]。

一般优化问题:

$$\min f(x) + g(z) \quad s.t. Ax + Bz = c, \quad (2-7)$$

其中 $x \in R^s, z \in R^n, A \in R^{p \times q}, B \in R^{p \times n}, c \in R^p, f: R^s \rightarrow R, g: R^n \rightarrow R$. x 与 z 是优化变量; $f(x) + g(z)$ 是待最小化的目标函数, 它由与变量 x 相关的 $f(x)$ 和与变量 z 相关的 $g(z)$ 这两部分构成, 这种结构可以很容易地处理统计学习问题优化目标中的正则化项, $Ax + Bz = c$ 是 p 个等式约束条件的合写。优化问题 (2-7) 的增广拉格朗日函数为:

$$\begin{aligned} L_{\rho}(x, z, y) = & f(x) + g(z) + y^T(Ax + Bz - c) \\ & + (\rho/2) \|Ax + Bz - c\|_2^2, \end{aligned} \quad (2-8)$$

其中 y 是对偶变量 (或称为拉格朗日乘子), $\rho > 0$ 是惩罚参数, $\|\cdot\|_2$ 为欧氏空间的 L_2 -范数。 L_{ρ} 名称中的“增广”是指其中加入了二次惩罚项 $(\rho/2) \|Ax + Bz - c\|_2^2$ 。

则该优化问题的 ADMM 迭代求解方法为:

$$\begin{aligned} x^{k+1} &:= \arg \min_x L_{\rho}(x, z^k, y^k), \\ z^{k+1} &:= \arg \min_z L_{\rho}(x^{k+1}, z, y^k), \\ y^{k+1} &:= y^k + \rho(Ax^{k+1} + Bz^{k+1} - c). \end{aligned}$$

令 $u = (1/\rho)y$, 并对 $Ax + Bz - c$ 配方, 可得表示上更简洁的缩放形式:

$$\begin{aligned} x^{k+1} &:= \arg \min_x (f(x) + \left(\frac{\rho}{2}\right) \|Ax + Bz^k - c + u^k\|_2^2), \\ z^{k+1} &:= \arg \min_z (g(z) + \left(\frac{\rho}{2}\right) \|Ax^{k+1} + Bz - c + u^k\|_2^2), \\ u^{k+1} &:= u^k + Ax^{k+1} + Bz^{k+1} - c. \end{aligned}$$

2.3 近端 ADMM 算法

为了解决高维稀疏数据惩罚分位回归的求解问题, 需要对 ADMM 算法做一点改进[17]。

令 $z = y - X\beta$, 且 $Q_\tau(z) = \frac{1}{n} \sum_{i=1}^n \rho_\tau(z_i)$, 其中, $X = (x_{1i}, \dots, x_{1n})^T$, $y = (y_{1i}, \dots, y_{1n})^T$, 再令 $g = \nabla L_N(\beta^0) - \nabla L_1(\beta^0)$, β^0 为 β 的任意初始估计。则优化问题(2-6)转化为:

$$\min_{\beta, z} Q_\tau(z) + g^T \beta + \lambda \|w \circ \beta\|_1 \quad s.t. X\beta + z = y. \quad (2-9)$$

由式(2-8)得, 优化问题(2-9)的增广拉格朗日函数为:

$$\begin{aligned} L_\sigma(\beta, z, \theta) &= Q_\tau(z) + g^T \beta + \lambda \|w \circ \beta\|_1 - \langle \theta, X\beta + z - y \rangle \\ &\quad + (\sigma/2) \|X\beta + z - y\|_2^2. \end{aligned}$$

则优化问题(2-9)的 ADMM 迭代求解方法为:

$$\begin{aligned} \beta^{k+1} &= \arg \min_{\beta} L_\rho(\beta, z^k, \theta^k) \\ &= \arg \min_{\beta} g^T \beta + \lambda \|w \circ \beta\|_1 - \langle \theta^k, X\beta \rangle + \left(\frac{\sigma}{2}\right) \|X\beta + z^k - y\|_2^2, \\ z^{k+1} &= \arg \min_z L_\rho(x^{k+1}, z, y^k) \\ &= \arg \min_z Q_\tau(z) - \langle \theta^k, z \rangle + \left(\frac{\sigma}{2}\right) \|X\beta^{k+1} + z - y\|_2^2, \\ \theta^{k+1} &= \theta^k + \gamma \sigma (X\beta^{k+1} + z^{k+1} - y), \end{aligned}$$

其中, γ 为控制 θ 进程步长的常数。

注意到, 对于 z 进程, z^{k+1} 有非常容易计算的解析解, 该特性直接解决了由于分位回归损失函数的非平滑性引起的计算难度。

对于 z 进程, 可以分别计算 z^{k+1} 的每一个分量, 对于 $i = 1, \dots, n$, 有:

$$\begin{aligned} z_i^{k+1} &= \arg \min_{z_i} \frac{1}{n} \rho_\tau(z_i) - \theta_i^k z_i + \left(\frac{\sigma}{2}\right) (x_{1i}^T \beta^{k+1} + z_i - y_{1i})^2 \\ &= \arg \min_{z_i} \rho_\tau(z_i) + \left(\frac{n\sigma}{2}\right) \left[z_i - \left(y_{1i} - x_{1i}^T \beta^{k+1} + \frac{1}{\sigma} \theta_i^k \right) \right]^2. \end{aligned}$$

为了解决上述单变量极小化问题, 我们考虑更一般的形式, 定义 ρ_τ 的近端映射为:

$$Prox_{\rho_\tau}[\xi, \alpha] := \arg \min_{\mu \in \mathbb{R}} \rho_\tau(\mu) + \frac{\alpha}{2} (\mu - \xi)^2, \quad (2-10)$$

根据文献[15], (2-10)解为:

$$Prox_{\rho_\tau}[\xi, \alpha] = \begin{cases} \xi - \frac{\tau}{\alpha}, & \xi > \frac{\tau}{\alpha}, \\ 0, & \frac{\tau-1}{\alpha} \leq \xi \leq \frac{\tau}{\alpha}, \\ \xi - \frac{\tau-1}{\alpha}, & \xi < \frac{\tau-1}{\alpha}. \end{cases}$$

z 进程的计算转化为:

$$z_i^{k+1} = Prox_{\rho_\tau} \left[y_{1i} - x_{1i}^T \beta + \frac{1}{\sigma} \theta_i^k, n\sigma \right], \quad i = 1, \dots, n,$$

可以简便地计算 z_i^{k+1} 。

然而, 对于 β 进程, 并不能通过一个通用的设计矩阵 X 得到解析解。为了让 β 进程也有一个简单的解析解, 从而让算法更易于编码, 我们考虑在 β 进程的目标函数中添加一个近端项, 并用以下增广 β 进程替换标准 ADMM 中的 β 进程:

$$\begin{aligned} \beta^{k+1} &= \arg \min_{\beta} g^T \beta + \lambda \|w \circ \beta\|_1 - \langle \theta^k, X\beta \rangle \\ &\quad + (\sigma/2) \|X\beta + z^k - y\|_2^2 + \frac{1}{2} \|\beta - \beta^k\|_S^2, \end{aligned}$$

其中, S 为半正定矩阵, $S = \sigma(\eta I_p - X^T X)$, 其中 $\eta \geq \Lambda_{\max}(X^T X)$, $\Lambda_{\max}(\cdot)$ 为实对称矩阵的最大特征值。 $\|v\|_S^2 := \langle v, Sv \rangle$ 为 S 上半内积诱导的半范数。对于增广 β 进程:

$$\begin{aligned} \beta^{k+1} &= \arg \min_{\beta} g^T \beta + \lambda \|w \circ \beta\|_1 - \langle \theta^k, X\beta \rangle \\ &\quad + \left(\frac{\sigma\eta}{2}\right) \left\| \beta - \frac{\sigma\eta\beta^k - g + X^T(\theta^k + \sigma y - \sigma X\beta^k - \sigma z^k)}{\sigma\eta} \right\|_2^2 \\ &= \left(Shrink \left[\beta_j^k - \frac{g_j}{\sigma\eta} + \frac{1}{\sigma\eta} X_j^T (\theta^k + \sigma y - \sigma X\beta^k - \sigma z^k), \frac{\lambda\omega_j}{\sigma\eta} \right] \right)_{1 \leq j \leq p}, \end{aligned}$$

其中 $Shrink[\mu, \alpha] = sgn(\mu) \max(|\mu| - \alpha, 0)$ 为软收缩算子, $sgn(\cdot)$ 为符号函数, X_j 为矩阵 X 的第 j 列, β_j^k 和 g_j ($j = 1, \dots, p$) 分别为 β^k 和 g 的第 j 个分量。

2.4 模型算法流程

根据以上内容, 分布式通信高效的惩罚分位回归模型的主要算法流程如下 [4]:

1: 输入: 数据 $\{(x_{ji}, y_{ji}): i = 1, \dots, n\}$ 存储于机器 M_j 上, $j = 1, 2, \dots, k$, 常数 $\sigma > 0$ 、 $\tau > 0$ 、 $\gamma > 0$ 、和 $\lambda > 0$, 模型最大通信次数 $ncom$, 近端 ADMM 算法的最大迭代次数 $MaxIter$, 初始化 $(\beta^0, z^0, \theta^0) = (0, 0, 0)$, $g^0 = (0, \dots, 0)_{1 \times p}^T$ 。

2: For $t = 0, 1, \dots, ncom$ do.

3: 对于初值 (β^t, z^t, θ^t) , 令 $g = g^t$, 在主机 M_1 上基于数据 $\{(x_{1i}, y_{1i}): i = 1, \dots, n\}$, 利用近端 ADMM 算法, 计算 $(\beta^{t+1}, z^{t+1}, \theta^{t+1})$ 。

4: 主机 M_1 将 β^{t+1} 传输给其他机器 $M_j, j = 2, \dots, k$ 。

5: 在各机器 $M_j, j = 1, \dots, k$ 上, 代入 β^{t+1} 计算次梯度 $\nabla L_j(\beta^{t+1})$, 并将计算结果传输给主机 M_1 , 在主机 M_1 上计算 $g^{t+1} = \nabla \mathcal{L}_N(\beta^{t+1}) - \nabla \mathcal{L}_1(\beta^{t+1}) = \frac{1}{k} \sum_{j=1}^k \nabla L_j(\beta^{t+1}) - \nabla \mathcal{L}_1(\beta^{t+1})$ 。

6: $t = t + 1$ 。

7: 输出: β^{ncom+1} 。

2.5 改进的惩罚分位回归模型

在原模型中, 采用式(2-3)来计算全局损失函数, 其仅为通过各个机器计算出的局部损失函数的算数平均值。然而, 实际应用中, 由于不同机器上的数据存在差异, 储存于不同机器上的数据在用于预测分析时, 所占有的重要性可能并不相同, 为了提高模型的预测精度, 可以对来自不同机器的数据计算损失函数时加以不同的权重, 则改进后全局损失函数定义如下:

$$L_N(\beta) = \frac{1}{k} \sum_{j=1}^k m_j L_j(\beta) = \frac{1}{N} \sum_{j=1}^k m_j \sum_{i=1}^n \rho_\tau(y_{ji} - x_{ji}^T \beta), \quad (2-11)$$

其中, $m_j > 0, j = 1, \dots, k$ 为计算全局损失函数时, 对于机器 M_j 中储存数据赋予的权重, 满足 $\sum_{j=1}^k m_j = k$ 。在对实际数据处理, 运用分位数模型进行拟合时, 需要结合数据本身特性, 考虑各机器数据差异, 对重要级别更高的数据赋予更大的权重。

相应的, 式(2-5)中全局损失函数 $L_N(\beta)$ 关于 β 的次梯度也转变为:

$$\nabla L_N(\beta) = -\frac{1}{N} \sum_{j=1}^k m_j \sum_{i=1}^n x_{ji} \psi_{\tau}(y_{ji} - x_{ji}^T \beta).$$

2.6 本章小节

本章主要介绍了分布式数据通信高效的惩罚分位回归模型的相关基础知识, 详细介绍了惩罚分位回归模型的原理, 随后介绍了用于求解模型优化问题的 ADMM 算法以及改进的近端 ADMM 算法。为了更好地展现分布式数据通信高效的惩罚分位回归模型的具体工作机制, 本章梳理了模型的大致算法流程。

基于分布式数据通信高效的惩罚分位回归模型, 考虑到分布式数据机器间的差异, 本章最后对模型做了一些改进, 对来自不同机器的数据计算损失函数时加以不同的权重, 以此提高模型精度。

3 实证建模分析

本章将采用多组数据对分布式大数据的惩罚分位回归模型进行实证分析，并将模型预测结果与传统采用集中式算法的模型进行比较，检验分布式大数据的惩罚分位回归模型对低维数据的兼容性，以及在高维稀疏数据拟合时，在通信效率、预测精度、变量选择等各方面的优良性能。

同时，也对上述所提的改进模型进行实证分析，通过与原模型进行比较，检验改进模型是否在预测精度上对原模型有所优化。

3.1 模型在低维数据中的应用——基于气体排放数据

在本小节中，我们将分布式大数据的惩罚分位回归模型运用于低维数据，通过实际数据训练拟合，并进行预测分析，来研究模型在非高维稀疏数据情形下的表现性能。

3.1.1 数据来源及说明

本节所用燃气轮机碳氧化合物和氮氧化合物排放数据集来自于 UCI 机器学习数据存储库(<https://archive-beta.ics.uci.edu/ml/datasets/gas+turbine+co+and+nox+emission+data+set>)[18]，该数据集的数据由 11 个传感器监测来自土耳其西北地区的一台燃气轮机在一个小时内累计的烟气排放数据汇总得到。数据集共包含 2011 年至 2015 年五年的数据，在本小节中，选取采集自 2015 年的共计 6600 组数据进行实证研究，每组数据包含 11 个变量，我们将前 10 个变量(分别为 AT、AP、AH、AFDP、GTEP、TIT、TAT、TEY、CDP、CO，对于这些变量所代表的含义可在相关网站中获得，在此不再过多赘述)设为自变量，将 NO(氮氧化合物)作为响应变量，以研究分析燃气轮机氮氧化合物排放与自变量之间的相关性。

3.1.2 模型设定

对于气体排放数据，在本次实证中，选择使用 Alasso 罚函数的惩罚回归模

型, 将基于分布式大数据的惩罚分位回归模型(以下称为 E-ALasso 模型)与传统的采用集中式方法计算的模型(以下称为 C-ALasso 模型)进行对比。由于真实数据的具体参数 β 未知, 每次实验中, 我们将数据随机的分为 6000 个训练数据和 600 个测试数据, 并设定机器数 k 为 20, 将 6000 个训练数据随机存储于 20 个机器中, 每个机器包含 300 个数据。对于 E-ALasso 模型, 采用 20 倍的交叉验证对训练数据进行参数调整优化, 对于 C-ALasso 模型, 则选用 5 倍交叉验证来进行参数优化[4]。对于 E-ALasso 模型, 设定最大通信次数为 50 次, 每次拟合结果的最终预测误差为经过 50 轮通信后模型拟合结果的误差值。在环境科学的研究中, 更加注重对于上分位数的研究, 所以, 对于该数据, 我们分别取 $\tau = 0.5$ 和 $\tau = 0.75$ 进行拟合, 对数据随机分组, 分别进行 50 次随机实验, 对 50 次模型拟合的预测误差进行研究。

3.1.3 实验结果及分析

在 50 次随机实验中, E-ALasso 模型经过 50 轮通信后的预测误差与 C-ALasso 模型的预测误差的平均值与标准差如下表 3-1 所示。

表 3-1 基于气体排放数据 E-ALasso 与 C-ALasso 预测误差比较

模型	$\tau = 0.5$		$\tau = 0.75$	
	预测误差平均值	预测误差标准差	预测误差平均值	预测误差标准差
E-ALasso	0.0652	0.0243	0.0604	0.0245
C-ALasso	0.0684	0.0259	0.0647	0.0261

根据 50 次随机实验中 E-ALasso 模型经过 50 轮通信后的预测误差与 C-ALasso 模型的预测误差, 分别作出 $\tau = 0.5$ 和 $\tau = 0.75$ 时 E-ALasso 模型和 C-ALasso 模型的箱线图(分别为图 3-1(c)和(d))。再计算 50 次实验中经过不同通信次数后 E-ALasso 模型预测误差的平均值, 作出 τ 取不同值时预测误差相对于通信次数的曲线, 由于 C-ALasso 模型不需要进行通信, 所以其预测误差是一条水平线, $\tau = 0.5$ 和 $\tau = 0.75$ 分别对应图 3-1(a)和(b)。

通过图 3-1 中的箱线图[19]和表 3-1 的数据均可看出, 在 $\tau = 0.5$ 和 $\tau = 0.75$

时,在经过 50 轮通信后, E-ALasso 模型的预测误差比 C-ALasso 模型有一定的改进,且 50 次随机实验中, E-ALasso 模型预测误差的标准差更小,这就意味着运用 E-ALasso 模型的预测结果有更强的稳定性。

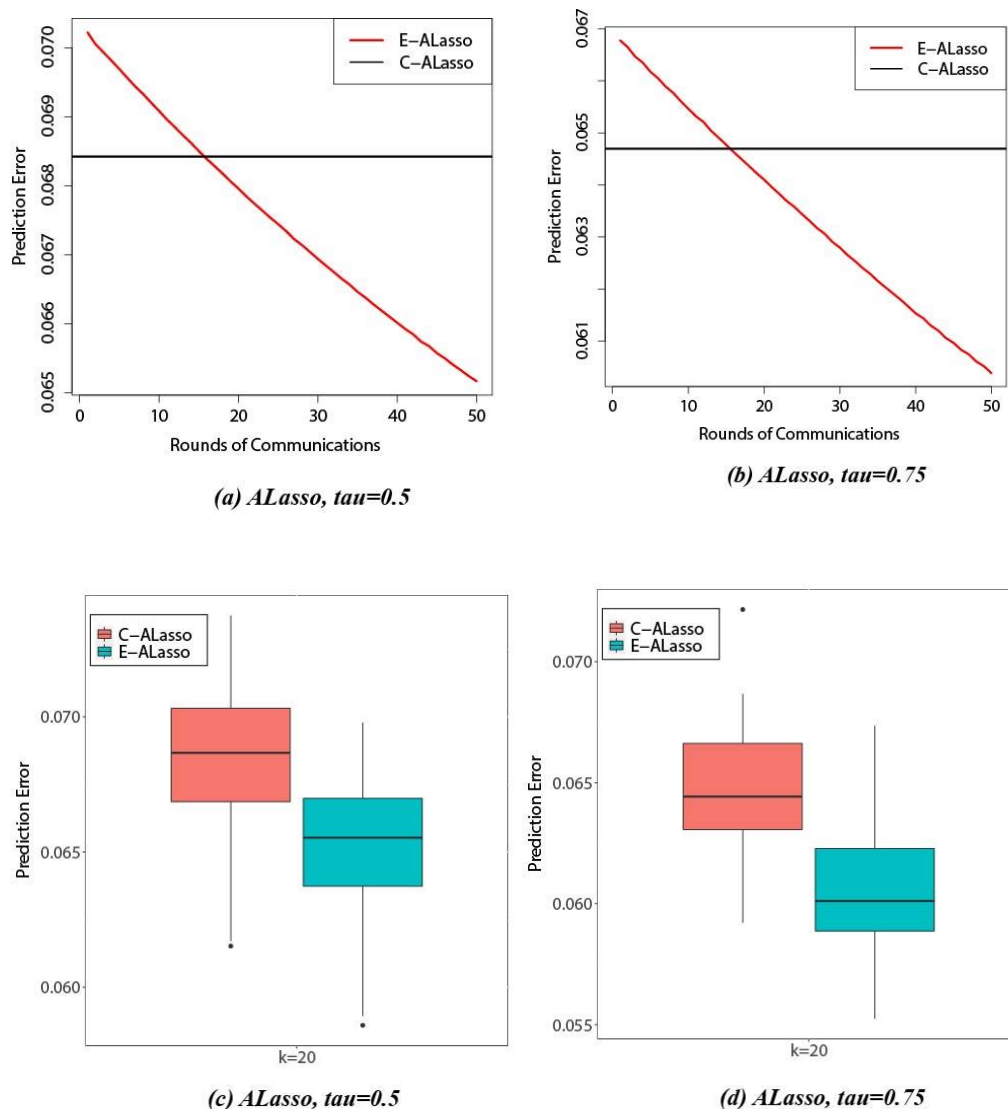


图 3-1 基于气体排放数据 E-ALasso 模型与 C-ALasso 模型预测误差比较

通过 E-ALasso 模型预测误差相对于通信次数的曲线图可以清晰地看出,虽然在较少轮通信后, E-ALasso 模型的预测误差要大于 C-ALasso 模型,但 E-ALasso 模型的预测误差随着通信次数的增加而急速下降,即使对于低维非稀疏的数据, E-ALasso 模型也可以在经过不到 20 轮通信后,预测误差达到与 C-ALasso 模型相媲美的精度,且在经过更多轮的通信后,预测误差仍在下降,从而达到显著低于 C-ALasso 模型预测误差的水平。

通过对气体排放数据的拟合预测分析可知,虽然该分布式数据通信高效的惩罚分位回归模型是针对高维稀疏数据而提出的,但是模型在低维非稀疏数据上仍有较好的性能,对低维数据使用替代损失函数来代替全局损失函数也不会造成较大的预测误差。

3.2 改进模型的实证分析——基于气体排放数据

在本小节中,我们将采用改进模型进一步对上述数据进行拟合,并对来自不同机器的数据计算损失函数时给予不同的权重,并将结果与原模型进行比较,以测试改进后模型的性能。

3.2.1 模型设定

对于本次实证分析,仍使用 Alasso 罚函数的惩罚回归模型进行拟合,对比改进的模型与原模型之间的差异。每次实验中,将不同年份采集的数据视为存储于不同机器中,分别从存储采集自 2011 年、2012 年、2013 年、2014 年、2015 年的五个机器(编号为 1-5)中数据容量均为 3000 的数据中各抽取 600 个数据,组成容量为 3000 的训练集,通过训练模型,拟合参数,然后从存储 2015 年数据的 5 号机器剩余的 2400 个数据中,随机抽取 100 个数据作为测试集,分别计算模型的预测误差。本次实验中,仅对 $\tau = 0.5$ 进行拟合,设定模型通信次数为 50 次。对数据进行 50 次随机实验,计算 50 次实验中经过不同通信次数后模型预测误差的平均值。为了研究主机(计算局部损失函数 L_1 时所用数据的存储机器)不同对实验结果的影响,在其他实验参数相同的情况下,分别将主机设为 1 和 5,比较两次实验结果的差异。

由于我们使用 2015 年采集的燃气轮机的气体排放数据(即机器 5 存储的数据)作为测试集,随着燃气轮机的使用,机器本身会产生相应的损耗,某种程度上会导致其所排放的气体与周边环境等的因素相关性会发生变化,这也就意味着,当我们使用 2011 年的数据(即机器 1 存储的数据)来预测 2015 年的气体排放数据时,产生的偏差会比使用 2015 年的数据要大,从某种程度上而言,这就体现了五个机器中所存储的数据的差异性。考虑到这种差异性的存在,对于不同机器上的数据,在用于分位回归模型计算损失函数时,不能一视同仁,需要对

来自不同机器的数据, 计算损失函数时加上不同的权重。由上述分析可知, 使用年份更接近 2015 年的数据, 预测应当更加准确, 所以在使用改进模型采用公式(2-11)对 1-5 号机器的数据计算局部损失函数时, 权重 $m_j, j = 1, \dots, k$ 分别赋予 0.5、0.5、1、1、2。对于原模型, 则不区分来自不同机器数据的差异, 仍使用公式(2-3)计算不同机器的局部损失函数, 不额外添加任何权重。

3.2.2 实验结果及分析

分别计算 50 次实验中经过不同通信次数后原模型与改进模型预测误差的平均值, 记为 $PE_{\text{原}}(t)$ 和 $PE_{\text{改}}(t), t = 1, \dots, ncom = 50$ 。定义 $DPE(t) = PE_{\text{原}}(t) - PE_{\text{改}}(t), t = 1, \dots, 50$, 为 50 次实验中, 经 t 次通信后原模型平均预测误差与改进模型平均预测误差之间的差值。 DPE 与通信次数之间的曲线如下图 3-2, 图 3-2(a)为使用机器 1(即 2011 年的数据)作为主机计算 L_1 时, DPE 与通信次数之间的关系; 图 3-2(b)为使用机器 5(即 2015 年的数据)作为主机计算 L_1 时, DPE 与通信次数之间的关系。

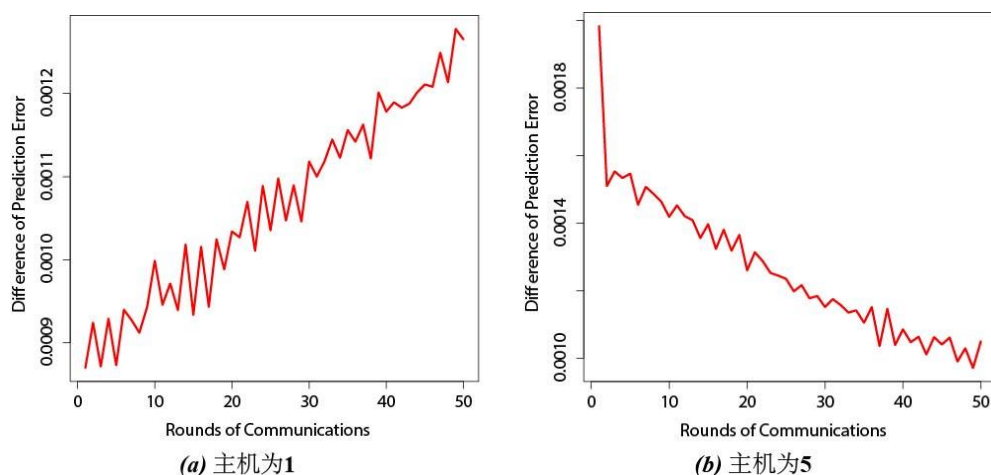


图 3-2 原模型与改进模型预测误差差值比较

从图 3-2 可以看出, 不论是选取哪一个机器作为主机, 不论经过多少轮通信, 改进后模型的预测误差总是小于原模型的预测误差。这说明, 在前文中我们关于燃气轮机随着使用年限的变化, 其氮氧化物排放的数据与自变量之间的相关关系会产生一定的差异, 也即不同机器储存的不同年份采集的燃气轮机气体排放数据之间存在差异性的假设成立的前提下, 考虑不同机器间数据的差

异性后,对不同机器的数据计算损失函数时,结合数据自身特性,选取适当的权重赋予各个机器,这样改进后的模型确实有效地提高了模型的预测精度,这就意味着本文对模型的改进是有意义的。

有意思的是,从图 3-2 不难发现,当选取不同机器作为主机时,DPE 与通信次数之间的相关性是完全相反的。当选取机器 1 作为主机时,随着通信次数的增加,DPE 有着上升的趋势,也即经过不断通信,改进模型的预测误差的精度相对于原模型在不断提高。然而,当使用机器 5 作为主机时,随着通信次数的增加,DPE 在不断减小,且经过 50 轮通信后,DPE 已经很小,接近于 0,这也就意味着,当通信次数足够多时,改进模型与原模型预测误差之间的差异会很小,改进模型相对于原模型的优势随着通信次数的增加逐渐被追平。

对于上述结果,接下来本文将给出一个可能的也较为合理的解释。根据前文的假设,不同机器存储数据之间存在差异性,当我们使用机器 1-5 的数据来做预测时,由于测试集的数据均来自于机器 5 所储存的数据,所以使用机器 5 的数据进行参数估计会比机器 1 更加精确,这里我们将机器 5 中储存的数据称为优良信息,而其他机器的数据称为不良信息,机器间的每一次通信,视作传递机器上的信息。在使用改进的模型时,我们考虑到了优良信息的重要性,从而对优良信息给与了更大的权重,而原模型不区分优良信息与不良信息。当使用机器 1 作为主机时,计算 L_1 的数据均为不良信息,而计算全局损失函数时,改进模型对来自机器 5 的优良数据赋予了更大的权重,当计算替代损失函数时,改进模型的主机收到的优良信息占比要比原模型大,且随着通信次数的增加,主机收到的优良信息在用于模型拟合的总信息中的占比在不断增加。尽管原模型的主机也会收到来自机器 5 的优良数据,但是由于机器 5 数据的权重与其他机器相同,所以随着通信次数的增加,原模型主机收到的优良信息在用于模型拟合的总信息中的占比相对于改进模型而言差距越来越大,这也就导致了随着通信次数的增加,改进模型的精度相对于原模型越来越高。当使用机器 5 作为主机时,计算 L_1 的数据均为优良信息,而用于计算全局损失函数的数据总是包含不良信息。由于改进模型中计算全局损失函数时不良信息占比较低,所以相较于原模型而言,需要通信更多的次数才能将等量的不良信息传递给主机。所以,在通信次数较少时,使用机器 5 作为主机的情形下,改进模型的预测误差

要小于原模型。但随着通信次数足够多,不良信息不断被主机所利用,这种优势也会逐渐消失殆尽,导致改进模型与原模型精度差异变小。

综上所述,在实际运用中,如果数据量非常庞大,通信成本会很高,期望通过较少通信次数使模型达到较高的精度时,可以使用改进模型,选取重要性等级高的数据所在的机器作为主机,按重要性给其他机器赋予合适的权重。如果因为某些不可抗力的因素,必须使用存储了较多重要性等级低的数据作为主机,那么在进行数据拟合时,需要在通信成本与预测误差之间权衡取舍,尽可能进行较多次通信,这样才能使改进模型的预测误差精度有明显的提升。

3.3 模型在高维稀疏数据中的应用——基于美国犯罪率数据

在本小节中,我们将通信高效的分布式大数据的惩罚分位回归模型运用于高维稀疏数据,对实际数据进行训练拟合,并进行预测分析,并将其与传统的采用集中式方法计算的模型进行对比,检验模型在通信效率、参数估计以及变量选择上的优良表现。

3.3.1 数据来源及说明

本节所用数据为社会科学领域数据,其数据条目纷繁复杂,综合了多个渠道的数据,社会经济方面的数据来自于美国 1990 年代的人口普查,执法数据取自 LEMAS 调查,犯罪数据则采取美国 FBI1995 年统计的数据,具体数据可从 UCI 机器学习数据存储器(<https://archive-beta.ics.uci.edu/ml/datasets/communities+and+crime>)[20]中找到。

数据以美国不同社区作为划分,统计了美国各个社区涉及人口组成、经济水平、警务信息、住房水平以及各种犯罪指标总计 128 个变量的数据。这些变量中包含一些非数值型变量以及某些变量含有大量的缺失值,不利于我们使用分位回归模型进行拟合实验。最终,在本节中,我们选择使用每十万人暴力犯罪总数作为响应变量,社区人口、家庭收入中位数、离婚率、全职警察数、每人口所承担警察预算开支等 100 个变量作为自变量,研究美国各社区暴力犯罪率与众多社区基本情况指标之间的相关关系,共从原始数据中选取 1800 组数据用于本次实证分析。

3.3.2 模型设定

在对美国犯罪率数据进行实证分析中,我们选择使用 SCAD 罚函数的惩罚回归模型,将基于分布式大数据的惩罚分位回归模型(以下称为 E-SCAD 模型)与传统的采用集中式计算的模型(以下称为 C-SCAD 模型)进行对比。由于真实数据的具体参数 β 未知,实验中我们将数据随机的分为 1500 个训练数据和 300 个测试数据,并设定机器数 k 为 10,将 1500 个训练数据随机存储于 10 个机器中,每个机器包含 150 个数据。对于 E-SCAD 模型,采用 10 倍的交叉验证对训练数据进行参数优化调整,对于 C-SCAD 模型,则选用 5 倍交叉验证来进行参数优化调整[4]。对于 E-SCAD 模型,设定最大通信次数为 50 次。由于本次数据选取的自变量很多,涉及社会生活等各个方面的指标,然而这些指标可能有许多与犯罪率的相关性不大,因此我们可以通过增加惩罚函数得到稀疏解。在社会科学领域的研究中,对某一社会群体或事件进行分析时,不仅需要考虑到上分位数的产生因素,也需要关注下分位数[21],所以,对于该犯罪率数据,我们分别取 $\tau = 0.25$ 、 $\tau = 0.5$ 和 $\tau = 0.75$ 进行拟合,对数据随机分组,分别进行 50 次随机实验,对 50 次数据拟合的预测误差取平均值进行研究。同时,为了研究模型在高维稀疏数据拟合时变量选择的能力,在每次实验中,记录经过 50 轮通信后模型拟合的不为 0 的参数所对应的变量,与 C-SCAD 模型的参数选择结果进行对比。

3.3.3 实验结果及分析

对实验结果进行计算整理,表 3-2 展示了在 50 次随机实验中,E-SCAD 模型经过 50 轮通信后的预测误差与 C-SCAD 模型的预测误差 PE 的平均值与标准差,其中括号内代表预测误差 PE 的标准差;E-SCAD 模型经过 50 轮通信后模型拟合的不为 0 的参数所对应的变量个数及 C-SCAD 模型拟合的不为 0 的参数所对应的变量个数 NUM 的平均值与标准差,其中括号内代表被选择的变量个数 NUM 的标准差。

计算 50 次实验中经过不同通信次数后 E-SCAD 模型预测误差的平均值,作出 τ 取不同值时,模型预测误差相对于通信次数的曲线,C-SCAD 模型的预测误差则是一条水平线, $\tau = 0.25$ 、 $\tau = 0.5$ 和 $\tau = 0.75$ 分别对应图 3-3 中(a)、(b)和(c)。

表 3-2 基于美国犯罪率数据 E-SCAD 与 C-SCAD 预测误差及变量选择比较

模型	$\tau = 0.25$		$\tau = 0.5$		$\tau = 0.75$	
	PE	NUM	PE	NUM	PE	NUM
E-SCAD	0.232 (0.0179)	63.9 (4.37)	0.239 (0.0143)	72 (2.896)	0.173 (0.01005)	74.3 (2.1)
C-SCAD	0.268 (0.0168)	64 (1.11)	0.299 (0.014)	70.7 (0.953)	0.197 (0.00993)	62.2 (0.37)

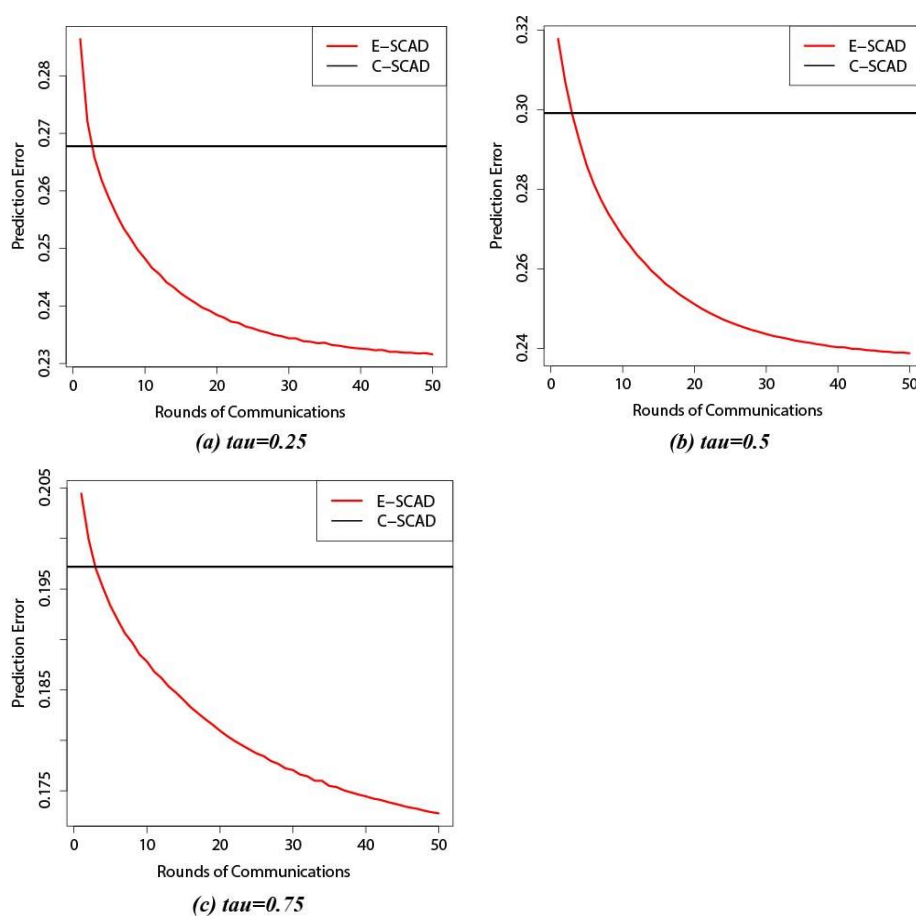


图 3-3 基于美国犯罪率数据 E-SCAD 模型与 C-SCAD 模型预测误差比较

统计 50 次实验中, 各自变量对应的参数经 50 轮通信后 E-SCAD 模型拟合不为 0 以及 E-SCAD 模型拟合不为 0 的次数, 由于自变量较多, 最终的表格见附录, 且因各自变量名字较长, 表格中各自变量名以数字 1-100 代替。

通过表 3-2 的数据均可看出, 在 $\tau = 0.25$ 、 $\tau = 0.5$ 和 $\tau = 0.75$ 时, 在经过 50 轮通信后, E-SCAD 模型的预测误差均比 C-SCAD 模型有明显改进, 且 50 次实验中, 预测误差的标准差虽然要较高与 C-SCAD 模型, 但与其差距非常小。

从模型预测误差与通信次数关系图可以看出, 对于高维稀疏数据, E-

SCAD 模型仅通过较少轮的通信, 就可以达到 C-SCAD 模型的预测误差, 且随着通信次数的增加, 预测误差仍可以降低。当 $\tau = 0.25$ 和 $\tau = 0.5$ 时, E-SCAD 模型的预测误差在通信次数达到 40 次后, 基本趋于稳定, 收敛速度很快; 当 $\tau = 0.75$ 时, 尽管已经通信了 50 次, E-SCAD 模型的预测误差仍存在较明显的下降趋势。

接下来对比两种模型变量选择的性能, τ 取不同值时, 两种模型所拟合的不为 0 的参数所对应的变量数量均有较大差异, 也就说明, 本节所用美国犯罪率的数据本身具有异质性[22]。从表 3-2 和附录中均可以看出, 当 $\tau = 0.25$ 和 $\tau = 0.5$ 时, E-SCAD 模型与 C-SCAD 模型变量选择的结果具有高度的一致性, E-SCAD 变量选择的性能相当不错。但当 $\tau = 0.75$ 时, E-SCAD 模型变量选择的结果与 C-SCAD 模型产生了较大出入。E-SCAD 模型拟合的不为 0 的参数对应的变量数量要明显高于 C-SCAD 模型的结果。根据附录, 找到被 E-SCAD 模型与 C-SCAD 模型挑选出的次数差异较大的几个变量, 通过查看 E-SCAD 模型在 50 轮通信后对其对应参数的估计值可以发现, 这些参数的绝对值相较于其他被选择的参数均非常小, 所以可能是因为数据的异质性, 导致模型对这些变量的参数拟合, 表现不是很好。但从总体上而言, 对于绝大多数变量, E-SCAD 模型在变量选择上与 C-SCAD 模型的性能相匹。

通过对该异质性的数据的拟合可以看出, E-SCAD 模型在经过几十轮的通信后, 无论是参数估计还是变量拟合, 均可以达到与 C-SCAD 模型相当的精度。且相比于前文中对低维数据的拟合, E-SCAD 模型在应用于高维数据时, 达到与 C-SCAD 模型预测误差相同的精度所需要的通信次数更少, 模型拟合的预测误差也更快收敛。前文中, 对气体排放数据的拟合, 模型在通信 50 轮后, 预测误差仍有较大下降趋势, 但本节对美国犯罪率数据的拟合中, 即使是对 $\tau = 0.75$ 的情形, 预测误差的下降趋势也远远不如气体排放数据的拟合结果。这也说明了, 对于高维稀疏数据, 由于数据的稀疏性, 使用替代函数的做法是非常明智的, 模型拥有很好的表现性能, 通信效率有很大提升[23]。

3.4 本章小结

本章使用低维的气体排放数据和高维稀疏的美国犯罪率数据对分布式大数

据的惩罚分位回归模型进行了实证分析, 通过将模型预测结果与传统采用集中式算法的模型进行比较, 验证了分布式大数据的惩罚分位回归模型在低维数据应用中有较好表现; 在高维稀疏数据拟合时, 在通信效率、预测精度、变量选择等各方面都展现了优良性能。

同时, 针对 2.5 节所提出的改进模型进行实证分析, 通过与原模型进行比较, 验证了改进模型在预测精度上对原模型有明显提升。

4 总结与展望

通过对两种不同行业背景、不同类型的数据的拟合,可以看出:1. 分布式数据通信高效的惩罚分位回归模型在参数估计方面具有良好的性能,对于高维稀疏数据,在变量选择上也有很好的表现。即使对异质性数据进行拟合,模型仍展示出了优良的变量选择能力。2. 对于一般的数据而言,模型可以在经过数十轮的通信后,预测误差达到与传统采用集中式算法的模型相媲美的精度,且经过更多次通信后,模型可以达到一个较为稳定的且显著低于采用集中式算法的模型的预测误差。

考虑到分布式数据不同机器上存储的数据的差异,在原模型的基础上,本文的改进模型在计算损失函数时,对来自不同机器的数据加以不同的权重。通过对实际数据拟合,可以发现:1. 不论经过多少次通信,改进模型在预测误差精度上比原模型有显著的提升。2. 根据使用不同机器作为主机用于计算 L_1 的拟合结果显示,当使用重要性等级高的数据作为主机时,改进模型在通信次数较少时,相比于原模型,预测误差精度有很大提升。

本文在应用改进模型和原模型对实际数据进行拟合时,对不同机器间数据的差异仅为猜想,且改进模型对于不同机器数据给予的权重也比较粗糙随意。这主要是因为本文所用的实际数据均为网上搜集所得,作者本人对于这些数据的本身特性并不是非常了解。且大多数即使是分布式存储的数据,在整理至网络后,均混杂在一起,无法区分原始存储数据的机器,所以对于这些数据,想要应用改进的模型会十分困难。如果有更好的数据,可以明确知道各个数据存储或采集自不同机器,且对于不同机器之间的数据也有着明确甚至可以量化的差异,那么可以使用改进模型对数据进行拟合,对不同机器赋予更加合理的权重,同时用不同机器作为主机,对结果进行对比,可以更加准确地评估改进模型。

同时,由于作者本人数学能力的不足,本文对于不同机器作为主机对改进模型与原模型之间预测误差差值与通信次数之间相关性的影响仅给出了一个通俗且符合直觉的解释,并未在数学上进行严格的证明。希望我学有所成后抑或其他感兴趣的人可以从数学上证明这种差异。

致谢

光阴似箭，岁月如梭，当我敲下这段文字的时候，也就意味着我的大学四年欢乐时光接近尾声。回想往昔，虽然疫情占据了我大学生涯中的大部分时光，让大学生活不那么完美，但这四年来，仍有许多美好值得慢慢回味。

我的母校——华中科技大学，作为一个以理工科见长的学校，却不失人文关怀。校园的环境虽然算不上美丽，但总可以为我提供静下心来学习的安静舒适场所，也可以让我偶有烦恼时四处漫步消磨时光。大学四年来每一个甜美的梦和我这如逻辑斯蒂曲线般的体重则可以让我对学校的食宿条件大加赞赏。当然，对于一所大学而言，最为重要的理应是教育质量，而华中大的教育质量则是毋庸置疑的，学在华中大的名号早已传遍五湖四海，在华中大的每一天都是遨游在知识的海洋的一天，都是充实自我的一天。前校长李培根院士曾说：“什么是母校？就是那个你一天骂他八遍却不许别人骂的地方。”可是对于华中科技大学这样的学校，每天夸上八遍都不够又怎么会舍得骂上一句呢？

上了大学后，有许多人会后悔自己所选择的专业，但数学与统计学院的老师、辅导员以及教学安排让我这个到大学后才意识到自己并不擅长数学的人也无悔于当初的选择。相比于华中大本就严谨治学的精神，作为数统学院的老师更是一丝不苟，每位老师在课堂上总是饱含激情，用粉笔在黑板上写下一个又一个定理的证明，力求让我们感受数学那最为纯粹的美。即使是疫情严重时期的线上课堂，各位老师也是排除万难，保证了极高的教学质量。老师们对待学术有多严谨对待学生就有多随和，无论是学习问题还是未来发展问题，每位老师总是乐于为我们提供帮助，知无不言，言无不尽，引导我们找到答案。我的辅导员——王玥，我们总是喜欢称呼她“玥姐”，则是一个亦师亦友的存在，学习上，作为我们与老师之间沟通的桥梁，将我们的问题及时反馈给老师；生活上，阴晴冷暖、无微不至。学院的教学安排同样也是精心设计，课程安排紧凑、循序渐进，充分尊重个体差异，在专业分流时，每个人都可以按照自己的意愿选择合适的专业。

我认为我的毕业设计导师就是学院老师的典范，李老师在毕业设计选题时，就十分尊重个人想法，允许我选择自己喜欢的方向进行研究。在得知我没有好

的选题后，又主动给出了多个建议选题让我从中选择。李老师的治学是相当严谨的，我的文献翻译、开题报告和论文，老师都十分认真地审阅并给出了改进意见。李老师也是一个非常和蔼的老师，脸上总是挂着笑容。和李老师的聊天总是轻松、愉悦的，即使是聊一些我一头雾水的问题，李老师也能娓娓道来，让人丝毫感受不到压力。还记得李老师在大学刚开学给我们说，大学快毕业了，趁着最后几个月要好好享受这段时光，然而，因为疫情反复，这学期我大部分时间都是呆在寝室。

大学里，和我相处最久的莫过于我的三个室友了。有的室友擅长数学，从大一到大四、从数学分析到毕业设计，他都为我解答了不计其数的数学问题，不厌其烦地纠正了我丰富多彩的数学错误，在考试前总会精准押题助我通过一门又一门考试。有的室友是老饕，带我吃遍校内与校外，何以解忧，唯有美食。有的室友则博古通今、学贯中西，政治、经济、历史、文化均能侃侃而谈，极大地丰富了我的知识面，让我对整个世界的了解豁然开朗。

此去一别，便将和大学中的人与物聚少离多。依依东望，愿我的母校、数学学院、各位老师、我和我的室友都有光明的未来，没有比脚更长的路，没有比人更高的山！

参考文献

- [1] Roger Koenker, and Gilbert Bassett Jr. Regression quantiles. *Econometrica*, vol. 46, no. 1, pp. 33–50, 1978.
- [2] Aijun Hu, Chujin Li, and Jing Wu. Expectile regression on distributed large-scale data. *IEEE Access*, vol. 8, pp. 270-280, 2020.
- [3] 蔡超. 基于大规模数据的分位数回归方法及应用[D]. 合肥: 合肥工业大学, 2017.
- [4] Aijun Hu, Chujin Li, and Jing Wu. Communication-efficient modeling with penalized quantile regression for distributed data. *Complexity (New York, N. Y.)*, vol. 2021, 2021.
- [5] Michael I. Jordan, Jason D. Lee, and Yun Yang. Communication efficient distributed statistical inference. *Journal of the American Statistical Association*, vol.114, no.526, pp.668-681, 2019.
- [6] Xi Chen, Weidong Liu, Xiaojun Mao, and Zhuoyi Yang. Distributed high-dimensional regression under a quantile loss function. *Journal of Machine Learning Research*, vol. 21, no. 182, pp. 1-43, 2020.
- [7] Fengrui Di, and Lei Wang. Multi-round smoothed composite quantile regression for distributed data. *Annals of the Institute of Statistical Mathematics*, Tokyo 2022.
- [8] Kangning Wang, Benle Zhang, Fayadh Alenezi, and Shaomin Li. Communication-efficient surrogate quantile regression for non-randomly distributed system. *Information Science*, vol. 588, pp. 425-441, 2022.
- [9] Kean Ming Tan, Lan Wang, and Wenxin Zhou. High-dimensional quantile regression: convolution smoothing and concave regularization. *Journal of the Royal Statistical Society*, vol. 84, no. 1, pp. 205-233, 2022.
- [10] Yue Wang, Yan Zhou, Rui Li, and Heng Lian. Sparse high-dimensional semi-nonparametric quantile regression in a reproducing kernel Hilbert space. *Computational Statistics and Data Analysis*, vol. 168, 2022.
- [11] Battey Heather, Jianqing Fan, Han Liu, Junwei Lu, and Ziwei Zhu. Distributed

testing and estimation under sparse high dimensional models. *The Annals of Statistics*, vol. 46, no. 3, pp. 1352-1382, 2018.

[12] Roger Koenker. *Quantile regression*. Cambridge University Press, 2005.

[13] 张翠霞. 基于自适应 LASSO 的二元选择分位回归应用分析[D]. 武汉: 华中科技大学, 2017.

[14] 刘惠篮. 基于复合分位数回归方法的统计模型的相关研究[D]. 重庆: 重庆大学, 2016.

[15] Liqun Yu, and Nan Lin, Admm for penalized quantile regression in big data. *International Statistical Review*, vol. 85, no. 3, pp. 494–518, 2017.

[16] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, vol. 3, no. 1, pp. 1–122, 2011.

[17] Yuwen Gu, Jun Fan, Lingchen Kong, Shiqian Ma, and Hui Zou. Admm for high-dimensional sparse penalized quantile regression. *Technometrics*, vol. 60, no. 3, pp. 319–331, 2018.

[18] Gas Turbine CO and NOx Emission Data Set[DB/OL]. UC Irvine Machine Learning Repository, <https://archive-beta.ics.uci.edu/ml/datasets/gas+turbine+co+and+nox+emission+data+set>. 2019.

[19] Huixia Judy Wang and Ian W. McKeague and Min Qian. Testing for marginal linear effects in quantile regression. *Journal of the Royal Statistical Society*, vol. 80, pp. 433-452, 2018.

[20] Communities and Crime[DB/OL]. UC Irvine Machine Learning Repository, <https://archive-beta.ics.uci.edu/ml/datasets/communities+and+crime>. 2009.

[21] Lingxin Hao, and Daniel Q. Naiman. *Quantile regression*. SAGE Publications, 2007.

[22] 梁亚坤. 内存约束下的高维变系数分位数回归模型[D]. 黑龙江: 哈尔滨工业大学, 2020.

[23] 王贺雨. 分布式分位回归算法及应用[D]. 杭州: 浙江大学, 2019.

附录 美国犯罪率数据变量被选择频率

变量	$\tau = 0.25$		$\tau = 0.5$		$\tau = 0.75$	
	E-SCAD	C-SCAD	E-SCAD	C-SCAD	E-SCAD	C-SCAD
1	0	0	0	0	0	0
2	34	50	37	50	45	50
3	23	49	37	32	45	4
4	50	50	50	50	50	50
5	25	50	1	50	0	0
6	23	50	0	0	2	0
7	50	50	50	50	50	50
8	50	50	50	50	50	50
9	48	50	50	50	37	50
10	50	50	50	50	50	50
11	0	0	0	0	0	0
12	50	50	50	50	50	50
13	27	0	30	0	41	0
14	50	50	50	50	50	50
15	38	13	39	50	35	50
16	50	50	50	50	50	50
17	50	50	50	50	50	50
18	38	0	50	1	50	50
19	38	50	48	50	50	50
20	35	0	44	0	50	0
21	28	0	28	0	43	0
22	27	0	26	0	46	0
23	11	0	21	0	23	0
24	0	0	6	0	8	0
25	25	0	43	0	50	0
26	15	0	31	0	44	0
27	35	31	49	42	45	5
28	0	0	0	0	0	0
29	40	50	50	50	50	50
30	39	50	50	50	50	50
31	50	50	50	50	50	50
32	45	34	50	50	50	50
33	48	50	46	50	50	50
34	50	50	50	50	50	50

35	50	50	50	50	29	50
36	50	50	50	50	50	50
37	50	50	50	50	50	50
38	50	50	50	50	50	50
39	49	50	43	50	50	50
40	49	50	45	50	35	50
41	50	0	50	50	50	50
42	50	41	45	50	50	50
43	47	50	50	50	50	50
44	50	50	50	50	50	50
45	50	50	50	50	50	50
46	50	50	50	50	50	50
47	50	50	50	50	50	50
48	34	50	50	50	50	50
49	50	50	48	50	30	50
50	0	0	0	0	0	0
51	10	25	41	1	50	49
52	0	0	0	0	0	0
53	16	0	50	50	33	50
54	27	0	50	50	50	50
55	35	0	48	50	28	50
56	44	0	50	50	50	50
57	3	50	0	50	0	0
58	6	50	0	50	0	0
59	6	50	0	50	0	0
60	8	50	0	50	0	0
61	50	50	50	50	50	50
62	2	50	0	2	0	0
63	2	0	24	0	47	0
64	0	0	16	0	38	0
65	44	4	50	50	28	50
66	50	50	50	50	50	50
67	18	0	50	0	42	50
68	39	50	50	50	48	50
69	2	50	1	47	2	0
70	50	50	50	50	50	50
71	45	50	49	50	50	50
72	0	0	0	0	0	0
73	50	50	50	50	50	50

74	50	50	50	50	50	50
75	41	50	50	50	50	50
76	50	50	50	50	50	50
77	48	0	37	49	50	50
78	50	50	50	50	50	50
79	15	46	48	50	50	50
80	50	50	50	50	50	50
81	50	50	50	50	50	50
82	50	50	50	50	39	50
83	1	48	6	17	14	0
84	10	3	20	48	46	0
85	20	50	28	50	50	0
86	16	0	27	46	50	0
87	49	0	50	0	50	50
88	25	6	50	0	50	0
89	50	50	50	50	50	50
90	0	0	0	0	0	0
91	0	0	0	0	0	0
92	3	50	0	50	0	0
93	50	50	50	50	50	50
94	50	50	50	50	50	50
95	45	50	43	50	50	50
96	49	50	50	50	50	50
97	0	0	0	0	0	0
98	4	0	24	0	34	0
99	41	50	49	50	11	0
100	0	0	23	0	46	0

华中科技大学

本科毕业设计（论文）任务书

题 目 基于分位回归的分布式大数据的实证建模分析

An empirical analysis based on modeling with quantile
regression for distributed big data

(任务起止日期: 2021 年 12 月 10 日~2022 年 05 月 31 日)

院 系 数学与统计学院

专业班级 统计 1801 班

姓 名 周宇杰

学 号 U201810112

指导教师 李楚进

教研室(系、所)负责人 刘建斌 2021年12月12日审查

院(系)负责人 王和军 2021年12月18日批准

课题内容:

1. 熟悉分布式大数据分位回归的基本理论与方法。
2. 尝试模型的创新改进、数值实现与实证分析等。
3. 熟悉常用统计软件的应用: Python、R、SAS 等。

课题任务要求:

数据分析与建模 统计软件应用。

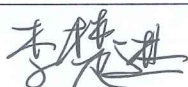
主要参考文献 (由指导教师选定):

- [1] M. I. Jordan, J. D. Lee, and Y. Yang. Communication efficient distributed statistical inference. *Journal of the American Statistical Association*, vol.114, no.526, pp.668-681, 2019.
- [2] X. Chen, W. D. Liu, X. J. Mao, and Z. Y. Yang. Distributed high-dimensional regression under a quantile loss function. *Journal of Machine Learning Research*, vol. 21, no. 182, pp. 1-43, 2020.
- [3] R. Koenker. *Quantile Regression*. Cambridge University Press, UK, 2005.

同组设计者:

蒋心怡

指导教师签名:



2022年5月20日

毕业设计（论文）成绩评定

班号： 统计1801班

学生姓名： 周俊杰

综合成绩： _____ 分

评分小组长（签名）： _____ 年 月 日

指导教师评定意见

一、对毕业设计（论文）的学术评语（应具体、确切、实事求是）

论文选题前沿，结构合理，思路清晰，行文流畅。内容充实，通过实证分析、比较分析研究分布式大数据回归建模改进效果。方法先进，结论有创新。这反映作者具有扎实的数据分析和统计应用能力；论文是一篇优秀的本科毕业论文。

二、对毕业设计（论文）评分

(1) 理工医科评分表

评分项目 (分值)	调研论证 (10分)	外文翻译 (5分)	设计(论文) 撰写质量 (10分)	学习态度 (10分)	基本理论和 基本技能 (50分)	创新 (15分)	合计 (100分)
得分	8	5	9	10	48	13	93

(2) 文科评分表

评分项目 (分值)	文献阅读与 文献综述 (10分)	外文翻译 (10分)	论文撰写质量 (10分)	学习态度 (10分)	学术水平、论证 能力和创新 (60分)	合计 (100分)
得分						

指导教师签字：

李赞进

2022年 5月 23日

答辩小组评定意见

一、评语（根据学生答辩情况及其设计（论文）质量综合评价）

周宇杰同学的论文通过比较和实证分析研究分布式大数据回归建模改进效果，方法恰当，结论充实有意义。

论文写作规范，行文顺畅，结构合理。在答辩过程中，周宇杰同学思路清晰，表述准确，能很好地讲述要点，回答答辩小组的问题。

答辩小组经评议，一致同意通过其论文答辩，并建议授予其理学学士学位。

二、评分

评分项目 (分值)	答 辩 情 况		论 文 质 量		合 计 (100 分)
	答辩情况 (15 分)	回答问题情况 (25 分)	规范要求与文字表达 (20 分)	学术水平 (40 分)	
得分	12	22	18	34	86

答辩小组长签字：李慧群

2022年 5月 28日