基于弹性网惩罚的复合分位数回归估计

张国浩
重庆工商大学 数学与统计学院,重庆 400067

摘　要:针对高维数据的建模分析问题,提出一种基于弹性网络法和复合分位数回归相结合的稳健估计方法。在该估计方法中,所提出的模型能够有效进行变量选择与系数压缩,并处理数据间的多重共线性与群组效应问题,在大数据时代下具有较广的适应性。同时,与已有的惩罚最小二乘估计和惩罚分位数回归估计相比,该估计方法不仅放宽了对模型误差项的分布要求,而且综合考虑了多个分位点的损失,在面对离群值或呈现尖峰、厚尾分布数据时能够保持更强的稳健性和抗干扰性。在一定条件下,对所构建模型估计的相合性与稀疏性进行了理论分析,结果表明:所提出的模型能够将不相关的变量完全压缩至零,且估计量和真实系数以趋于 1 的概率相同。此外,在数值模拟方面,设置了 5 种误差项分布条件,根据设定的 4 项指标,通过与其他惩罚函数模型以及损失函数模型进行比较,结果表明新提出的方法具备更好的稳健性与有效性。

关键词:变量选择;稳健估计;弹性网;复合分位数回归
中图分类号:O212.7　　**文献标识码:**A　　doi:10.16055/j.issn.1672-058X.2023.0005.014

Compound Quantile Regression Model with Elastic Net Penalty

ZHANG Guohao
School of Mathematics and Statistics, Chongqing Technology and Business University, Chongqing 400067, China

Abstract: Aiming at the problem of modeling and analysis under high-dimensional data, a robust estimation method based on elastic network method and composite quantile regression was proposed. In this estimation method, the proposed model can effectively perform variable selection and coefficient compression, and deal with multicollinearity and group effects between data, and has wide adaptability in the era of big data. At the same time, compared with the existing penalized least squares estimation and penalized quantile regression estimation, this estimation method not only relaxes the distribution requirements of the model error term, but also comprehensively considers the loss of multiple quantiles, which can maintain stronger robustness and anti-interference in the face of outliers or data with spiky, thick-tailed distributions. Under certain conditions, a theoretical analysis of the consistency and sparsity of the constructed model estimates is carried out. The results show that the proposed model can completely compress uncorrelated variables to zero, and the estimate and the true coefficient have the same probability of tending to 1. In addition, in terms of numerical simulation, five kinds of error term distribution conditions are set. According to the four indicators set, the comparison with other penalty function models and loss function models is carried out. The results show that the newly proposed method has better robustness and effectiveness.

Keywords: variable selection; robust estimation; elastic net; composite quantile regression

收稿日期:2022-06-22　**修回日期:**2022-08-19　**文章编号:**1672-058X(2023)05-0104-09
基金项目:重庆市教委科学技术研究计划重大项目(KJZD-M202100801).
作者简介:张国浩(1998—),男,贵州遵义人,硕士生,从事高维数据分析及应用研究.
引用格式:张国浩.基于弹性网惩罚的复合分位数回归估计[J].重庆工商大学学报(自然科学版),2023,40(5):104—112.
ZHANG Guohao. Compound quantile regression model with elastic net penalty[J]. Journal of Chongqing Technology and Business University (Natural Science Edition), 2023, 40(5): 104—112.

1 引　言

高维数据的建模分析一直是学者们讨论的热门话题。针对其分析过程中的诸多难点,国内外学者一般围绕两个方面进行研究,一是如何从成千上万个变量中筛选出对预测因子真正起重要作用的变量以降低模型的复杂度,提高模型的解释性;二是如何利用更稳健的估计方法应对异方差结构数据并抵抗离群值和极端点影响。

第一个方面研究的便是变量选择问题。Akaike[1]与 Schwarz[2]最先提出 AIC 与 BIC 准则,他们均采用子集选择思想来减小模型维度以提高计算效率,但是两者过拟合与不稳定的缺点却使模型的预测能力变得很差。这也使后来的学者将大量的工作用于发展可以同时进行变量选择和系数估计的正则化方法,此类方法的特性是利用惩罚函数对待估系数进行限制,当采用岭回归[3]作为惩罚函数时,回归系数将会趋向于零收缩,但是永远无法收缩到零,因此岭回归并不会提供一个易于解释的模型。而当采用 LASSO[4]惩罚时,模型将通过使一些系数缩小到零来产生可解释的模型,以此实现变量选择,SCAD[5]惩罚则是对称且非凸性的,它虽然可以产生稀疏的模型,但迭代算法却极其缓慢,难以适应高维数据情形,MCP[6]惩罚与 SACD 同样因其非凸性,导致拟合模型的数值计算问题充满挑战。

此外,由 Zou 等[7]提出的弹性网络法的表现尤其出色,因其结合了 LASSO 惩罚和岭回归的优势,可以有效解决多重共线性和变量选择问题,特别是对于协变量数量远远大于观测值数量时,弹性网格外有效。卢[8]将弹性网推广到 Logistic 模型和 Poisson 模型,证明该方法具有将强相关性协变量全部选入或剔除模型的能力;黄[9]将该方法应用在部分线性模型中,证明了其具有 Oracle 性质;李[10]将该方法应用到平衡纵向数据模型的变量选择中,证明了该方法的相合性。

为了实现更稳健的估计,Koenker 等[11]提出更具鲁棒性的分位数回归。由于其更宽松的前提假设以及能够同时得到多个全面刻画响应变量整体条件分布特征的优势,逐渐被学者认为是最小二乘法的有利替代,并开始与变量选择方法结合。例如 Fan[12]提出自适应 LASSO 惩罚的分位数回归模型,该模型在变量选择与抵抗极端值影响方面都达到了极佳的效果,但并没有解决数据中的群组效应问题;Su 等[13]将弹性网与分位数回归结合,不仅在稳定估计的同时保留了变量选择的准确性和计算效率,还解决了数据中的群组效应问题,但是此方法的估计效率十分依赖分位点的选择且并未解决分位数回归中尾部的分位点过高或是过低造成的预测值左偏及右偏问题。

综上而言,各类变量选择方法虽然都在一定程度上降低了模型的维度,有的甚至能灵活处理多重共线性与群组效应问题,但绝大多数理论却受最小二乘法的制约,难以保持稳健的特性。同时,分位数回归估计因其稳健的特性被 Fan[12]和 Su[13]考虑与变量选择方法结合并取得了一定的成效,可是两者模型的估计效率却依赖于分位点的选择,这使得模型在实际使用中效果不佳。

本文在 Su 等[13]的弹性网与分位数回归结合及 Zou 等[14]的复合分位数估计的基础上,同时考虑多个分位点的损失来优化模型中的损失函数,进一步提出弹性网惩罚复合分位数回归模型。所提出的模型具有更强的稳健性与有效性,在理论上证明了该模型估计的稀疏性和相合性,并在数值模拟中也表明该模型优于其他估计方法。

2　模型简述

2.1　复合分位数回归

考虑一般线性回归模型:

$$\boldsymbol{y}_i=\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}+\varepsilon_i \tag{1}$$

其中,$\boldsymbol{x}_i^{\mathrm{T}}=(x_{i1},x_{i2},\cdots,x_{in})$为解释变量,$\boldsymbol{y}_i$ 为响应变量,$\boldsymbol{\beta}$ 为回归系数向量,ε_i 为随机误差项。

假设 $\boldsymbol{\varepsilon}$ 的 τ 分位数为 b_τ^*,其中 $\tau\in(0,1)$ 已知,Koenker 和 Bassett 提出的分位数回归可通过求解式(2)得到回归系数的解:

$$(\hat{b}_\tau,\hat{\boldsymbol{\beta}}^{QR})=\underset{b,\boldsymbol{\beta}}{\operatorname{argmin}}\frac{1}{n}\sum_i^n\rho_\tau(\boldsymbol{y}_i-b-\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}) \tag{2}$$

其中,$\rho_\tau(u)=u\cdot(\tau-I(u<0))$,$0<\tau<1$ 为分位数回归的损失函数。区别于最小二乘法只能估计响应变量的中心趋势,分位数回归能够得到响应变量各个位置的条件分布,有效避免了极端值的影响,其适应性和稳健性都得到了明显提高。然而,分位数回归在实际应用中,预测效率会受到分位点选择的制约,人们往往难以选择合适的分位点来构建模型。在这样的背景之下,Zou

和 Yuan 便以分位数回归为基础,提出了复合分位数回归的概念。

选取分位点 $0<\tau_1<\tau_2<\cdots<\tau_K<1$,在式(1)中,通过式(3)目标函数便可得到复合分位数回归系数的估计:

$$(\hat{b}_1,\hat{b}_2,\cdots,\hat{b}_K,\hat{\boldsymbol{\beta}}^{CQR}) = \underset{b_1,b_2,\cdots,b_K,\boldsymbol{\beta}}{\operatorname{argmin}} \sum_{k=1}^{K}\left\{\sum_{i=1}^{n}\rho_{\tau_k}(\boldsymbol{y}_i - b_k - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})\right\} \tag{3}$$

其中,$\tau_k=k/(K+1)$,$k=1,2,\cdots,K$。复合分位数的目标函数是多个分位数回归方程的加总最小和,且因其分位点的强制选择,包含的信息也更丰富,考虑不同分位点的变化,得到了一个不随 τ 值变化的估计系数,具备更强的稳健性。

2.2 弹性网惩罚复合分位数模型

考虑如下统计模型:

$$\boldsymbol{Y}=\boldsymbol{X\beta}+\boldsymbol{\varepsilon}$$

其中,$\boldsymbol{Y}=(\boldsymbol{y}_1,\boldsymbol{y}_2,\cdots,\boldsymbol{y}_n)^{\mathrm{T}}$ 为 n 维响应变量,$\boldsymbol{X}=(\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n)^{\mathrm{T}}=(\tilde{x}_1,\tilde{x}_2,\cdots,\tilde{x}_p)$ 为 $n\times p$ 维设计矩阵,$\boldsymbol{\beta}=(\beta_1,\beta_2,\cdots,\beta_p)^{\mathrm{T}}$ 为 p 维回归系数向量,$\boldsymbol{\varepsilon}=(\varepsilon_1,\varepsilon_2,\cdots,\varepsilon_n)^{\mathrm{T}}$ 为 n 维随机误差向量。

本文提出基于弹性网惩罚复合分位数回归模型(CQ-EN),即利用如下正则化问题来估计回归系数 $\boldsymbol{\beta}$:

$$\min_{b_1,b_2,\cdots,b_K,\boldsymbol{\beta}}\left\{\sum_{k=1}^{K}\sum_{i=1}^{n}\rho_{\tau_k}(\boldsymbol{y}_i-\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}-b_k)+n\lambda_n\|\boldsymbol{\beta}\|_1+\frac{n\mu_n}{2}\|\boldsymbol{\beta}\|_2^2\right\}$$

其中,$\|\boldsymbol{\beta}\|_1=\sum_{i=1}^{p}|\beta_i|$ 为 $\boldsymbol{\beta}$ 的 L_1 范数,$\|\boldsymbol{\beta}\|_2=\sqrt{\sum_{i=1}^{p}\beta_i^2}$ 为 $\boldsymbol{\beta}$ 的 L_2 范数,并且 $\lambda_n,\mu_n>0$ 为正则化系数,$n\lambda_n\|\boldsymbol{\beta}\|_1+\frac{n\mu_n}{2}\|\boldsymbol{\beta}\|_2^2$ 为弹性网惩罚项。

在高维线性回归问题中,为保证模型的可识别性,同时提升模型的拟合精度,真实的系数向量 $\boldsymbol{\beta}^*$ 往往被假定为稀疏的,即仅有一小部分非零。假设非零元素的数量为 s_n,不失一般性,假定真实的模型为 $\boldsymbol{M}_*=\mathrm{supp}(\boldsymbol{\beta}^*)=\{1,2,\cdots,s_n\}$,其补集 $\boldsymbol{M}_*^c=\{s_n+1,s_n+2,\cdots p_n\}$ 表示噪声协变量的指标集,同时记 $\boldsymbol{\beta}^*=((\boldsymbol{\beta}_1^*)^{\mathrm{T}},0^{\mathrm{T}})$,并将设计矩阵重写为 $\boldsymbol{X}=(\boldsymbol{S},\boldsymbol{Q})$,其中:

$$\boldsymbol{S}=(S_1,S_2,\cdots,S_n)^{\mathrm{T}}=(\tilde{\boldsymbol{x}}_1,\tilde{\boldsymbol{x}}_2,\cdots,\tilde{\boldsymbol{x}}_{s_n})$$

$$\boldsymbol{Q}=(Q_1,Q_2,\cdots,Q_n)^{\mathrm{T}}=(\tilde{\boldsymbol{x}}_{s_n+1},\tilde{\boldsymbol{x}}_{s_n+2},\cdots,\tilde{\boldsymbol{x}}_{p_n})$$

此处子矩阵 $\boldsymbol{S}$ 是非零系数的信号协变量矩阵,$\boldsymbol{Q}$ 则是剩余的噪声协变量矩阵,本文将设计矩阵标准化使得每一列的 L_2 范数均为 $\sqrt{n}$。

3 模型的统计性质

为了评估所提出的新方法,本节将建立 CQ-EN 模型的统计性质。首先,定义利用 Oracle 信息协助得到的正则化估计(ORE)为

$$(\hat{\boldsymbol{\beta}}^o,\hat{b}_k)=\arg\min_{\boldsymbol{\beta}\in\boldsymbol{M},\boldsymbol{b}}L_n(\boldsymbol{\beta},\boldsymbol{b})$$

其中,$\hat{\boldsymbol{\beta}}^o=((\hat{\boldsymbol{\beta}}_1^o)^{\mathrm{T}},0^{\mathrm{T}})^{\mathrm{T}}$。本节将表明 ORE 以接近于 1 的概率估计出真实系数向量,且当满足一些条件时,CQ-EN 估计量享有与 ORE 一样的性质。为此,参考 Fan[12],首先对 n 维误差向量 $\boldsymbol{\varepsilon}$ 的分布以及设计矩阵作出如下假定。

(**A1**) 一致存在大于零的常数 c_1 和 c_2,对任意满足 $|u|\leqslant c_1$ 的 u,有 $\{f_i(u)\}_{i=1}^n$ 一致有界且不为 0 和 ∞,并满足 $|F_i(u+b_k^*)-F_i(b_k^*)-uf_i(b_k^*)|\leqslant c_2u^2$,其中,$k=1,2,\cdots,K$,$f_i(u)$ 与 $F_i(u)$ 分别为 ε_i 的密度函数与分布函数。

(**A2**) 定义 $\boldsymbol{H}_k=\mathrm{diag}\{f_1(b_k^*),\cdots,f_n(b_k^*)\}$,则 $n^{-1}(1,\boldsymbol{S})^{\mathrm{T}}\boldsymbol{H}_k(1,\boldsymbol{S})$ 的特征值介于 c_0 与 c_0^{-1} 之间,且 $\kappa_n\equiv\max_{ij}|x_{ij}|=o(\sqrt{n}\,s_n^{-1})$。

(**A3**) 对后面定理 1 中定义的 γ_n,有 $\left\|\frac{1}{n}\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{H_kS}\right\|_{2,\infty}<\frac{\lambda_n}{2\gamma_n}$,其中,对于矩阵 $\boldsymbol{A}$ 与向量 $\boldsymbol{x}$,$\|\boldsymbol{A}\|_{2,\infty}=\sup_{x\neq0}\|\boldsymbol{Ax}\|_\infty/\|\boldsymbol{x}\|_2$,并且,$\lambda_n>\sqrt{(\log p_n)/n}$,其中,$\log p_n=o(n^b)$,$b\in(0,1)$,$\|\boldsymbol{x}\|_\infty=\max_{1\leqslant i\leqslant n}|\boldsymbol{x_i}|$。

(A1)是为了限制函数 $f_i(u)$ 的局部变动幅度,意味着任意 $f_i(u)$ 在 b_k^* 周围是 Lipschitz 连续的,当分布函数的二阶导数存在且有界时,(A1)成立,常用的分布,如 Laplace 和柯西分布均满足这一条件。(A2)是用来限制信号协变量矩阵 $\boldsymbol{S}$ 和设计矩阵 $\boldsymbol{X}$ 的规模,当设计矩阵 $\boldsymbol{X}$ 产生于某种分布式,上述条件中 κ_n 的界以渐近 1 的概率满足,例如 $\boldsymbol{X}$ 产生于次指数分布时,若 $s_n=(\sqrt{n}/\log p)$,则 κ_n 的界以渐近 1 的概率满足。(A3)则是为了控制设计矩阵 $\boldsymbol{X}$ 与信号变量和噪声变量矩阵列向量的相关性,从而进行必要限制,使得该估计在没有

Oracle 信息协助时也具备 ORE 的性质。

定理 1　若 $\lambda_n \sqrt{s_n}\kappa_n \to 0, \mu_n \sqrt{s_n}\kappa_n \to 0$,且条件(A1)(A2)成立,则存在常数 $c_1, c_2 > 0$ 使得

$$P(\max_{1 \leqslant k \leqslant K}(\|\hat{\boldsymbol{\beta}}_1^o - \boldsymbol{\beta}_1^*\|_2^2 + (\hat{b}_k - b_k^*)^2) \leqslant \gamma_n^2) \geqslant 1 - c_1 n^{-c_2 s_n}$$

其中,$\gamma_n = C\sqrt{s_n}\left(\lambda_n + \mu_n + \sqrt{(\log n)/n}\right)$,$C$ 为大于零的常数。

定理 1 是用来表明 $\hat{\boldsymbol{\beta}}^o$ 是真实向量 $\boldsymbol{\beta}^*$ 的相合估计,且是一致的,并且以趋于 1 的概率估计出真实系数向量的正确信号。但是,这是利用 Oracle 信息寻找信号协变量位置后得到的效果,真实情况下,信号协变量的位置是未知的。因此定理 2 将表明在没有 Oracle 信息的情况下,选取适当的 λ_n, μ_n 后,CQ-EN 模型估计量会以趋于 1 的概率享有和 ORE 相同的性质。

定理 2　设 $\max_j |\beta_j| = o(\lambda_n/\mu_n)$,$\gamma_n s_n^{3/2}\kappa_n^2(\log_2 n)^2 = o(n\lambda_n^2)$,$\kappa_n \gamma_n^2 = o(\lambda_n)$,且 $\lambda_n > 2\sqrt{(1+c)(\log p_n)/n}$,$c$ 为大于零的常数,则当条件(A1)—(A3)成立时,目标函数 $L_n(\boldsymbol{\beta}, \boldsymbol{b})$ 至少存在一个全局最小值 $\hat{\boldsymbol{\beta}} = ((\hat{\boldsymbol{\beta}}_1^o)^{\mathrm{T}}, \hat{\boldsymbol{\beta}}_2^{\mathrm{T}}, \hat{b}_k)$,且以大于 $1 - c_1 n^{-c_2 s_n}$ 的概率满足以下两个性质:

(1) $\hat{\boldsymbol{\beta}}_2 = 0$。

(2) $\max_{1 \leqslant k \leqslant K}[\|\hat{\boldsymbol{\beta}}_1^o - \boldsymbol{\beta}_1^*\|_2^2 + (\hat{b}_k - b_k^*)^2] \leqslant \gamma_n^2$。

4　数值模拟

本节将通过数值模拟以确定所提出估计方法的有限样本性质。考虑模拟数据来自高维线性回归模型:

$$y_i = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \varepsilon_i$$

其中,观测样本 $n = 100$,参数数量 $p = 400$,协变量 $\boldsymbol{x} \sim (0, \Sigma_x)$,其中 $(\Sigma_x)_{i,j} = 0.5^{|i-j|}$,真实的回归系数向量被固定为

$$\boldsymbol{\beta} = \{2, 1.5, 0, 0.8, 0, 1, 1.75, 0, 0.75, 0, 0.3, 0, \cdots, 0\}$$

考虑误差向量来自以下 5 个分布,标准正态分布 $\varepsilon \sim N(0,1)$;被少部分污染的正态分布 MN1:$\varepsilon \sim 0.9N(0,1) + 0.1N(0,25)$;Laplace 分布;自由度为 4 的 t 分布 $\varepsilon \sim t(4)$;Cauchy 分布。

本文将通过计算 4 个值来评估所提出的方法。

(1) L_1 损失:即 $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1$。

(2) L_2 损失:即 $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2$。

(3) FP:噪声协变量被选入模型的个数。

(4) FN:信号协变量未被选入模型的个数。

同时,为了评价与比较,本文将在每种误差分布条件下将 CQ-EN 模型分别与以下方法进行对比:

(1) Q-EN[13]:弹性网惩罚的分位数回归,其中 $\tau = 0.5$。

(2) CQ-LASSO:LASSO 惩罚复合分位数回归。

(3) CQ-Ridge:Ridge 惩罚复合分位数回归。

(4) EN[7]:弹性网惩罚最小二乘回归。

比较 CQ-EN,Q-EN,EN 可以反映在同样的惩罚函数下,不同的损失函数在处理各类误差分布情形时的有效性,而比较 CQ-EN,CQ-LASSO,CQ-Ridge 是用来分析复合分位数回归在不同惩罚函数下的性能,模拟结果如表 1 所示,其中各项数值均为模拟 100 次的均值。

表 1　相关协变量的模拟结果

Table 1　The simulation results of related covariates

误差项	评估项	CQ-EN	Q-EN	EN	CQ-LASSO	CQ-Ridge
$N(0,1)$	L_1 损失	6.045	5.851	7.719	11.803	9.236
	L_2 损失	3.928	3.678	16.267	25.967	5.378
	FP	152.380	140.815	32.420	10.093	393
	FN	0.280	0.267	0.356	1.269	0
MN1	L_1 损失	5.720	7.064	7.105	11.893	9.562
	L_2 损失	3.679	4.807	14.199	23.592	5.771
	FP	136.40	169.338	26.191	63.803	393
	FN	0.112	0.147	0.828	0.469	0
Laplace	L_1 损失	5.234	5.937	10.267	12.625	9.137
	L_2 损失	3.155	3.721	23.243	29.754	5.296
	FP	131.245	144.372	38.613	66.445	393
	FN	0.381	0.433	0.273	1.924	0
$t(4)$	L_1 损失	2.714	3.413	8.742	4.231	6.244
	L_2 损失	2.871	3.414	18.526	9.161	3.261
	FP	37.619	46.964	46.356	1.569	393
	FN	5.129	6.510	0.739	2.366	0
Cauchy	L_1 损失	4.607	5.736	12.184	93.604	6.138
	L_2 损失	6.032	8.323	57.749	1 778.998	5.413
	FP	38.685	56.637	19.466	130.26	393
	FN	6.253	6.592	7.296	5.04	0

首先比较 CQ-EN,Q-EN,EN 模型,从结果可以看出:对于 MN1 情形,CQ-EN 在所有 4 项指标中均优于

其余两者;对于 Cauchy 情形,本文所提出的模型也明显具有更低的 L_1,L_2 损失和更小的 FN;对于 $t(4)$ 情形,CQ-EN 具有更小的 L_1,L_2 损失和 FP;对于 Laplace 和 $N(0,1)$ 情形,CQ-EN 与 Q-EN 模型在 L_1 和 L_2 损失表现上平分秋色,其中 Q-EN 模型在 $N(0,1)$ 情形表现更好,而 EN 则在 FP 和 FN 的表现上优于 CQ-EN 与 Q-EN 模型。总体而言,在面对具有重尾分布和异常值,如 $t(4)$ 等情形时,CQ-EN 的表现明显优于 Q-EN 与 EN,这表明复合分位数回归比其余两者具有更稳健的特性。

此外,与 CQ-Ridge 和 CQ-LASSO 相比,本文所提出的模型除 Cauchy 情形外,L_1 和 L_2 损失都明显更低,对于 Cauchy 情形,CQ-Ridge 具有更小的 L_2 损失,这是因为 Ridge 惩罚项的良好系数压缩能力,但是它却选择了模型的所有变量,并未达到变量选择的目的。对于 $N(0,1)$,MN1,Laplace,$t(4)$ 情形,CQ-LASSO 具有明显更小的 FP,这表明 LASSO 作为惩罚项时,更倾向于选择真正的变量进入模型,从而严格进行筛选。同时,对于 $N(0,1)$,MN1,Laplace 情形,CQ-EN 具有更小的 FN,这表明该模型倾向于剔除大量不相关变量。对于 Cauchy 情形,CQ-EN 与 CQ-LASSO 在 FN 与 FP 上的表现各有突出,这体现了 CQ-EN 模型有能力选择更多重要的变量进入模型,而 CQ-LASSO 则更倾向于剔除不相关的变量。

总体而言,CQ-EN 模型不仅在各种误差分布情形下均保持了优异的表现,产生了稀疏的模型解,而且在重尾分布和离群值情形下,展现了优越的性能。

5 定理的证明

本节将给出定理 1 和定理 2 的证明过程。为了简单和方便,让 C 表示一个正常数,每次出现可能是不同的值,在开始证明前,先给出几个引理。

引理 1(Fan[12]) 若(A2)成立,则对任意 $t>0$,有

$$P\left(Z_{nk}(M_n) \geqslant 4M_n\sqrt{s_n/n}+t\right) \leqslant \exp\left(-nc_0t^2/(8M_n^{\,2})\right),k=1,2,\cdots,K$$

下面引理是来自 Bühlmann[15] 中的 Hoeffding 不等式。

引理 2 设 $Z_1,Z_2,\cdots,Z_n$ 是取值于某个空间 Γ 的独立随机变量,γ 为 Γ 上的实值函数,若存在正实数 $C_1,C_2,\cdots,C_n$,使得

$$E(\gamma(Z_i))=0,\ |\gamma(Z_i)|<C_i,i=1,2,\cdots,n$$

则对所有的 $t>0$,有

$$P\left(\left|\sum_{i=1}^n\gamma(Z_i)\right|\geqslant t\right)\leqslant 2\exp\left(-\frac{t^2}{2\sum_{i=1}^n C_i^2}\right)$$

引理 3 在以 $\boldsymbol{\beta}^*$ 为中心的球形邻域 R^s 内,存在某些序列 $\gamma_n\to 0$,记

$$\boldsymbol{N}=\{(\boldsymbol{\beta}_1^{\mathrm{T}},\boldsymbol{\beta}_2^{\mathrm{T}})^{\mathrm{T}}\in\boldsymbol{R}^p,b_k's\in\mathbf{R}:\boldsymbol{\beta}_2=0\in\boldsymbol{R}^{p-s},\max_{1\leqslant k\leqslant K}(\|\hat{\boldsymbol{\beta}}_1^{o}-\boldsymbol{\beta}_1^*\|_2^2+(\hat{b}_k-b_k^*)^2)\leqslant\gamma_n^2\}$$

若 $\sqrt{1+\gamma_n s_n^{3/2}}\ \kappa_n\ (\log_2 n)=o\ (\sqrt{n}\lambda_n)$,$n^{1/2}\lambda_n\times(\log p_n)^{-1/2}\to\infty$,$\kappa_n\gamma_n^2=o(\lambda_n)$,$\max_j|\boldsymbol{\beta}_j|=o(\lambda_n/\mu_n)$,则当(A1)—(A3)成立时,有

$$P\left(\sup_{(\boldsymbol{\beta},b_k)\in N}\left\|\sum_{k=1}^K\boldsymbol{Q}^{\mathrm{T}}\rho'_{\tau_k}(y-S\boldsymbol{\beta}_1-b_k)\right\|_\infty+n\mu_n\|\boldsymbol{\beta}_1\|_\infty\geqslant n\lambda_n\right)\leqslant c_1p_n^{-c_2}\tag{4}$$

其中,$\rho'_{\tau_k}(u)=\tau_k-I\{u\leqslant 0\}$,$c_1$,$c_2$ 为大于零的常数。

证 明 对于固定的 $j\in\{s_n+1,s_n+2,\cdots,p_n\}$,$\boldsymbol{\beta}=(\boldsymbol{\beta}_1^{\mathrm{T}},\boldsymbol{\beta}_2^{\mathrm{T}},b_k)^{\mathrm{T}}\in\boldsymbol{N}$,定义

$$\gamma_{\boldsymbol{\beta},j}(\boldsymbol{x}_i,\boldsymbol{y}_i)=\sum_{k=1}^K x_{ij}[\rho'_{\tau_k}(\boldsymbol{y}_i-\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}-b_k)-\rho'_{\tau_k}(\varepsilon_i-b_k)-E[\rho'_{\tau_k}(\boldsymbol{y}_i-\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}-b_k)-\rho'_{\tau}(\varepsilon_i-b_k)]]$$

其中,$\boldsymbol{x}_i^{\mathrm{T}}=(x_{i1},x_{i2},\cdots,x_{ip})$ 为设计矩阵的第 i 行。

为证此引理,需对式(4)进行分解:

$$\sup_{(\boldsymbol{\beta},b_k)\in N}\left\|\frac{1}{n}\sum_{k=1}^K\boldsymbol{Q}^{\mathrm{T}}\rho'_{\tau}(\boldsymbol{y}-S\boldsymbol{\beta}_1-b_k)\right\|_\infty\leqslant\sup_{(\boldsymbol{\beta},b_k)\in N}\left\|\frac{1}{n}\sum_{k=1}^K\boldsymbol{Q}^{\mathrm{T}}E[\rho'_{\tau}(\boldsymbol{y}-S\boldsymbol{\beta}_1-b_k)-\rho'_{\tau}(\boldsymbol{\varepsilon}-b_k)]\right\|_\infty+\left\|\frac{1}{n}\sum_{k=1}^K\boldsymbol{Q}^{\mathrm{T}}\rho'_{\tau_k}(\boldsymbol{\varepsilon}-b_k)\right\|_\infty+\max_{j>s}\sup_{(\boldsymbol{\beta},b_k)\in N}\frac{1}{n}\sum_i^n|\gamma_{\boldsymbol{\beta},j}(\boldsymbol{x}_i,\boldsymbol{y}_i)|\tag{5}$$

由式(4)及式(5)可知,引理 3 在以下各式至少存在 $1-c_1p_n^{-c_2}$ 可能性成立时得证。

$$I_1\equiv\sup_{(\boldsymbol{\beta},b_k)\in N}\left\|\frac{1}{n}\sum_{k=1}^K\boldsymbol{Q}^{\mathrm{T}}E[\rho'_{\tau_k}(\boldsymbol{y}-S\boldsymbol{\beta}_1-b_k)-\rho'_{\tau_k}(\boldsymbol{\varepsilon}-b_k)]\right\|_\infty=$$

$$o(\lambda_n) \tag{6}$$

$$I_2 \equiv \frac{1}{n}\left\| \sum_{k=1}^{K} \boldsymbol{Q}^{\mathrm{T}}\rho'_{\tau_k}(\boldsymbol{\varepsilon}-b_k) \right\|_{\infty} = o_p(\lambda_n) \tag{7}$$

$$I_3 \equiv \max_{j>s}\sup_{\beta\in N}\left| \frac{1}{n}\sum_{k=1}^{K}\sum_{i}^{n}\gamma_{\beta,j}(\boldsymbol{x}_i,\boldsymbol{y}_i) \right| = o_p(\lambda_n) \tag{8}$$

$$I_4 \equiv \mu_n \|\boldsymbol{\beta}_1\|_{\infty} = o(\lambda_n) \tag{9}$$

I_4 的成立是显而易见的，首先证明 I_1 的成立。式(6)可以改写为

$$I_1 = \max_{j>s}\sup_{(\beta,b_k)\in N}\left| \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K}x_{ij}E[\rho'_{\tau_k}(\varepsilon_i-b_k) - \rho'_{\tau_k}(\boldsymbol{y_i}-\boldsymbol{x_i}^{\mathrm{T}}\boldsymbol{\beta}-b_k)] \right| \tag{10}$$

由(A1)，有

$$E[\rho'_{\tau_k}(\varepsilon_i-b_k)-\rho'_{\tau_k}(\boldsymbol{y}_i-\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}-b_k)] = F_i(\boldsymbol{S}_i^{\mathrm{T}}(\boldsymbol{\beta}_1-\boldsymbol{\beta}_1^*))-F_i(b_k) = f_i(b_k)\boldsymbol{S}_i^{\mathrm{T}}(\boldsymbol{\beta}_1-\boldsymbol{\beta}_1^*)+\tilde{I}_i$$

其中，$F_i(t)$ 为 ε_i 的累积分布函数，

$$\tilde{I}_i = F_i(\boldsymbol{S}_i^{\mathrm{T}}(\boldsymbol{\beta}_1-\boldsymbol{\beta}_1^*))-F_i(b_k)-f_i(b_k)\boldsymbol{S}_i^{\mathrm{T}}(\boldsymbol{\beta}_1-\boldsymbol{\beta}_1^*)$$

对任意 $j>s$，有

$$\sum_{k=1}^{K}\sum_{i=1}^{n}x_{ij}E[\rho'_{\tau_k}(\varepsilon_i-b_k)-\rho'_{\tau_k}(\boldsymbol{y_i}-\boldsymbol{x_i}^{\mathrm{T}}\boldsymbol{\beta}_1-b_k)] = \sum_{k=1}^{K}\sum_{i=1}^{n}(f_i(b_k)\boldsymbol{S}_i^{\mathrm{T}}(\boldsymbol{\beta}_1-\boldsymbol{\beta}_1^*)) + \sum_{k=1}^{K}\sum_{i=1}^{n}x_{ij}\tilde{I}_i$$

由式(10)及柯西不等式，可推出

$$I_1 \leqslant \left\| \frac{1}{n}\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{H_k}\boldsymbol{S}(\boldsymbol{\beta}_1-\boldsymbol{\beta}_1^*) \right\|_{\infty} + \max_{j>s}\left| \frac{1}{n}\sum_{i=1}^{n}x_{ij}\tilde{I}_i \right| \tag{11}$$

对式(11)右侧第一项，由(A3)得

$$\left\| \frac{1}{n}\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{H_k}\boldsymbol{S}(\boldsymbol{\beta}_1-\boldsymbol{\beta}_1^*) \right\|_{\infty} \leqslant \left\| \frac{1}{n}\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{H_k}\boldsymbol{S} \right\|_{2,\infty} \|\boldsymbol{\beta}_1-\boldsymbol{\beta}_1^*\|_2 < \frac{\lambda_n}{2} \tag{12}$$

再由(A1)，$|\tilde{I}_i| \leqslant C((1,\boldsymbol{S}_i^{\mathrm{T}})(\boldsymbol{\beta}_1-\boldsymbol{\beta}_1^*))^2$，结合(A2)，可得

$$\max_{j>s}\left| \frac{1}{n}\sum_{i=1}^{n}x_{ij}\tilde{I}_i \right| \leqslant \frac{\kappa_n}{n}\sum_{i=1}^{n}|\tilde{I}_1| \leqslant C\frac{\kappa_n}{n}\sum_{i=1}^{n}((1,\boldsymbol{S}_i^{\mathrm{T}})(\boldsymbol{\beta}_1-\boldsymbol{\beta}_1^*))^2 \leqslant C\kappa_n(\|\boldsymbol{\beta}_1-\boldsymbol{\beta}_1^*\|_2^2+(b_k-b_k^*)^2)$$

又因 $(\boldsymbol{\beta},b_k)\in \boldsymbol{N}$，根据假设条件 $\lambda_n^{-1}\kappa_n\gamma_n^2=o(1)$，有

$$\max_{j>s}\left| \frac{1}{n}\sum_{i=1}^{n}x_{ij}\tilde{I}_i \right| \leqslant C\kappa_n\gamma_n^2 = o(\lambda_n)$$

结合不等式(11)(12)可知 I_1 成立。现证式(7)，由引理 2，若 $\lambda_n>2\sqrt{(1+c)(\log p_n)/n}$，则

$$P\left(\left\| \sum_{k=1}^{K}Q^{\mathrm{T}}\rho'_{\tau}(\boldsymbol{\varepsilon}-b_k) \right\|_{\infty} \geqslant n\lambda_n \right) \leqslant P\left(\sum_{k=1}^{K}\sum_{j=s+1}^{p}\sum_{i=1}^{n}\tilde{x}_{ij}\rho'_{\tau}(\varepsilon_i-b_k) \geqslant n\lambda_n \right) \leqslant \sum_{j=s+1}^{p}P\left(\sum_{k=1}^{K}\sum_{i=1}^{n}\tilde{x}_{ij}\rho'_{\tau}(\varepsilon_i-b_k) \geqslant n\lambda_n \right) \leqslant \sum_{j=s+1}^{p}2\exp\left[-\frac{n^2\lambda_n^2}{4\sum_{i=1}^{n}\tilde{x}_{ij}^2} \right] = 2\exp\left(\log(p_n-s_n)-\frac{n\lambda_n^2}{4} \right) \leqslant c_1p_n^{-c_2}$$

因此，I_2 至少存在有 $1-c_1p_n^{-c_2}$ 的可能性成立。

现证明式(8)，定义函数空间 $\Gamma_j=\{\gamma_{\beta,j}:(\boldsymbol{\beta},b_k)\in \boldsymbol{N}\}$，对任意的 $\gamma_{\beta,j}\in\Gamma_j$，有 $E[\gamma_{\beta,j}(\boldsymbol{x}_i,\boldsymbol{y}_i)]=0$ 且 ρ'_{τ} 有界，则由 Fan[12] 中的引理 2 知，当 $0\leqslant k\leqslant(\log_2 n)/2$ 时，

$$\|\gamma_{\boldsymbol{\beta},j}\|_n \equiv \left(\frac{1}{n}\sum_{i=1}^{n}\gamma_{\boldsymbol{\beta},j}^2(\boldsymbol{x}_i,\boldsymbol{y}_i)^2 \right)^{1/2} \leqslant 2$$

且有

$$\log(1+N(2^{2-k},\Gamma,\|\cdot\|_2)) \leqslant \log 6+s\log(1+C^{-1}2^k\gamma_ns_n^{1/2}\kappa_n^2) \leqslant \log 6+C^{-1}2^k\gamma_ns_n^{3/2}\kappa_n^2 \leqslant 4(1+C^{-1}\gamma_ns_n^{3/2}\kappa_n^2)2^{2k}$$

进而由 Bühlmann[15] 中的推论 14.4 知，对任意 $t>0$，有

$$P\left(\sup_{(\boldsymbol{\beta},b_k)\in N}\left| \frac{1}{n}\sum_{i=1}^{n}\gamma_{\boldsymbol{\beta},j}(\boldsymbol{x}_i,\boldsymbol{y}_i) \right| \geqslant \frac{8}{\sqrt{n}}\left(3C^{-1}\sqrt{1+\gamma_ns_n^{3/2}\boldsymbol{\kappa}_n^2}\log_2 n+4+4t\right) \right) \leqslant 4\exp\left(\frac{nt^2}{8}\right)$$

令 $t=\sqrt{C(\log p_n)/n}$，则

$$P\left(\max_{j>s}\sup_{(\boldsymbol{\beta},b_k)\in N}\left| \frac{1}{n}\sum_{i=1}^{n}\gamma_{\boldsymbol{\beta},j}(\boldsymbol{x}_i,\boldsymbol{y}_i) \right| \geqslant \frac{24}{\sqrt{n}}\left(3C^{-1}\sqrt{1+\gamma_ns_n^{3/2}\boldsymbol{\kappa}_n^2}\log_2 n\right) \right) \leqslant 4(p_n-s_n)\exp\left(\frac{C\log p_n}{8}\right) \to 0$$

因此,若$\sqrt{1+\gamma_n s_n}{}^{3/2}\kappa_n^2 \log_2 n = o(\sqrt{n}\lambda_n)$,则 I_3 至少有 $1-c_1 p_n^{-c_2}$ 的可能性成立。

定理 1 的证明 若引入记号

$$v_{nk}(\boldsymbol{\beta}, b_k) = \sum_{i=1}^{n} \rho_{\tau k}(\boldsymbol{y_i} - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta} - b_k)$$

则目标函数可以写作:

$$L_n(\boldsymbol{\beta}, b) = \sum_{k=1}^{K} v_{nk}(\boldsymbol{\beta}, b_k) + n\lambda_n \|\boldsymbol{\beta}\|_1 + n\mu_n \|\boldsymbol{\beta}\|_2$$

对于给定的数 $M_n>0$,定义集合 $B_0(M_n)$:

$$\{\boldsymbol{\beta} \in R^p, b_k^{'}s \in \mathbf{R}: \sup_{1\leqslant k\leqslant K} \|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_2^2 + (b_k - b_k^*)^2 \leqslant M_n{}^2, \mathrm{supp}(\boldsymbol{\beta}) \subset \mathrm{supp}(\boldsymbol{\beta}^*)\}$$

然后,定义函数

$$Z_{nk}(M) = \sup_{(\boldsymbol{\beta}, b_k)\in B_0(M)} \frac{1}{n} |(v_{nk}(\boldsymbol{\beta}, b_k) - v_{nk}(\boldsymbol{\beta}^*, b_k^*)) - E(v_{nk}(\boldsymbol{\beta}, b_k) - v_{nk}(\boldsymbol{\beta}^*, b_k^*))| \tag{13}$$

首先,证明对于 $B_0(M_n)$ 中的任意向量$(\boldsymbol{\beta}, b_k) = (\boldsymbol{\beta}_1^{\mathrm{T}}, 0^{\mathrm{T}}, b_k)$,当 n 充分大时,对 $k=1,2,\cdots,K$,有式(14)成立,其中 $M_n = o(\boldsymbol{\kappa}_n^{-1} s_n{}^{-1/2})$。

$$E[v_{nk}(\boldsymbol{\beta}, b_k) - v_{nk}(\boldsymbol{\beta}^*, b_k^*)] \geqslant \frac{1}{2} c_0 n(\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*\|_2^2 + (b_k - b_k^*)^2) \tag{14}$$

为证明式(14),令 $a_i = |\boldsymbol{S}_i^{\mathrm{T}}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*)|$,则对于$(\boldsymbol{\beta}, b_k) \in B_0(M)$,根据 Hölder 不等式有

$$|a_i + b_k - b_k^*| \leqslant \|(1, \boldsymbol{S}_i)\|_2 \|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*\|_2 \leqslant \sqrt{s_n}\boldsymbol{\kappa}_n M_n \to 0$$

此时,针对 $\boldsymbol{S}_i^{\mathrm{T}}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*)$ 的符号展开讨论,若 $\boldsymbol{S}_i^{\mathrm{T}}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*)>0$,则式(11)左侧有

$$E[v_{nk}(\boldsymbol{\beta}, b_k) - v_{nk}(\boldsymbol{\beta}^*, b_k^*)] = E\Big[\sum_{i=1}^{n}(\rho_{\tau_k}(\boldsymbol{y_i} - \boldsymbol{S}_i^{\mathrm{T}}\boldsymbol{\beta} - b_k) - \rho_{\tau_k}(\boldsymbol{y_i} - \boldsymbol{S}_i^{\mathrm{T}}\boldsymbol{\beta}^* - b_k^*))\Big] = E\Big[\sum_{i=1}^{n}(\rho_{\tau_k}(\varepsilon_i - b_k^* + b_k^* - b_k - a_i) - \rho_{\tau_k}(\varepsilon_i - b_k^*))\Big]$$

显然

$$|r-s| - |r| = -s(I(r>0) - I(r>0)) + \int_0^s [I(r\leqslant t) - I(r\leqslant 0)]\,\mathrm{d}t$$

从而

$$\rho_\tau(r-s) - \rho_\tau(r) = s(I(r<0) - \tau) + \int_0^s [I(r\leqslant t) - I(r\leqslant 0)]\,\mathrm{d}t$$

于是依次由 $P(\varepsilon_i < b_\tau^*) = \tau$、Fubini 定理、泰勒展开及(A1)得

$$\begin{aligned} &E[\rho_{\tau_k}(\varepsilon_i - b_k^* + b_k^* - b_k - a_i)) - \rho_{\tau_k}(\varepsilon_i - b_k^*)] = \\ &E\Big[\int_0^{b_k - b_k^* + a_i} I(0 \leqslant \varepsilon_i - b_k^* \leqslant t)\,\mathrm{d}t\Big] = \\ &\int_0^{b_k - b_k^* + a_i} (F_i(t + b_k^*) - F_i(b_k^*))\,\mathrm{d}t = \\ &\int_0^{b_k - b_k^* + a_i} (F_i(b_k^*) + tf_i(b_k^*) + o(t) - F_i(b_k^*))\,\mathrm{d}t = \\ &\frac{1}{2} f_i(b_k^*)(b_k - b_k^* + a_i)^2 + o(1)(b_k - b_k^* + a_i)^2 \end{aligned} \tag{15}$$

其中,$o(1)$ 对于 $i=1,2,\cdots,n$ 一致成立。当 $\boldsymbol{S}_i^{\mathrm{T}}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*)<0$,也可得到式(15)右侧同样的结果。

进一步,由(A2)得

$$\begin{aligned} &E[v_{nk}(\boldsymbol{\beta}, b_k) - v_{nk}(\boldsymbol{\beta}^*, b_k^*)] = \\ &\sum_{i=1}^{n} E[\rho_{\tau_k}(\varepsilon_i - b_k^* + b_k^* - b_k - a_i)) - \rho_{\tau_k}(\varepsilon_i - b_k^*)] \geqslant \\ &\frac{1}{2}\sum_{i=1}^{n} f_i(b_k^*)(b_k - b_k^* + a_i)^2 = \\ &\frac{1}{2}\boldsymbol{C}^{\mathrm{T}}(1, \boldsymbol{S})^{\mathrm{T}}\boldsymbol{H_k}(1, \boldsymbol{S})\boldsymbol{C} \geqslant \\ &\frac{1}{2} c_0 n(\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*\|_2^2 + (b_k - b_k^*)^2) \end{aligned}$$

其中,$\boldsymbol{C} = (b_k, \boldsymbol{\beta}_1) - (b_k^*, \boldsymbol{\beta}_1^*)$。

因此,当$(\boldsymbol{\beta}_1^{\mathrm{T}}, 0^{\mathrm{T}}, b_k)^{\mathrm{T}} \in \boldsymbol{B}_0(M_n)$时,式(14)成立。但是,前面所设定的 Oracle 估计器$\hat{\boldsymbol{\beta}}^o = ((\hat{\boldsymbol{\beta}}_1^o)^{\mathrm{T}}, 0^T, \hat{b}_k)^{\mathrm{T}}$可能并不在 $\boldsymbol{B}_0(M_n)$ 集合中,因此,令 $\tilde{\boldsymbol{\beta}} = (\tilde{\boldsymbol{\beta}}_1^{\mathrm{T}}, 0^{\mathrm{T}}, \tilde{b}_k)$,其中

$$\tilde{\boldsymbol{\beta}}_1 = \mu\hat{\boldsymbol{\beta}}_1^o + (1-\mu)\boldsymbol{\beta}_1^*$$

$$\tilde{\boldsymbol{b}}_k = \mu\hat{b}_k + (1-\mu)b_k^*$$

$$\mu = \frac{2M_n}{2M_n + \sqrt{\|\hat{\boldsymbol{\beta}}_1^o - \boldsymbol{\beta}_1^*\|_2^2 + \max\limits_{1\leqslant k\leqslant K}(\hat{b}_k - b_k^*)^2}}$$

此时$(\tilde{\boldsymbol{\beta}}, 0, \tilde{b}_k) \in \boldsymbol{B}_0(M)$,根据$(\hat{\boldsymbol{\beta}}_1^o, \hat{b}_k)$的定义以及目标函数为凸函数的性质,可得

$$L_n(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{b}}) \leqslant uL_n(\hat{\boldsymbol{\beta}}_1^o, 0, \hat{\boldsymbol{b}}) + (1-u)L_n(\boldsymbol{\beta}_1^*, 0, \boldsymbol{b}^*) \leqslant$$

$$L_n(\boldsymbol{\beta}_1^*,0,\boldsymbol{b}^*)=L_n(\boldsymbol{\beta}^*,\boldsymbol{b}^*)$$

其中,$\tilde{\boldsymbol{b}}=(\tilde{b}_1,\tilde{b}_2,\cdots,\tilde{b}_K)^{\mathrm{T}}$,结合式(13)及三角不等式得

$$\begin{aligned}E[v_{nk}(\tilde{\boldsymbol{\beta}},\tilde{b}_k)-v_{nk}(\boldsymbol{\beta}^*,b_k^*)]=&\\ \{v_{nk}(\boldsymbol{\beta}^*,b_k^*)-Ev_{nk}(\boldsymbol{\beta}^*,b_k^*)\}-&\\ \{v_{nk}(\tilde{\boldsymbol{\beta}},\tilde{b}_k)-Ev_{nk}(\tilde{\boldsymbol{\beta}},\tilde{b}_k)\}+&\\ L_n(\tilde{\boldsymbol{\beta}})-L_n(\boldsymbol{\beta}^*)+n\lambda_n\|\boldsymbol{\beta}_1^*\|_1-&\\ n\lambda_n\|\tilde{\boldsymbol{\beta}}_1\|_1+n\mu_n(\|\boldsymbol{\beta}_1^*\|_2^2-\|\tilde{\boldsymbol{\beta}}_1\|_2^2)\leqslant&\\ nZ_{nk}(M_n)+n\lambda_n\|(\boldsymbol{\beta}_1^*-\tilde{\boldsymbol{\beta}}_1)\|_1+&\\ n\mu_n(\|\boldsymbol{\beta}_1^*\|_2^2-\|\tilde{\boldsymbol{\beta}}_1\|_2^2)&\end{aligned}\tag{16}$$

定义事件

$$A_{nk}=\{Z_{nk}(M_n)\leqslant 2M_n n^{-1/2}\sqrt{s_n\log n}\}$$

根据引理 1,有

$$P(A_{nk})\geqslant 1-\exp(-c_0 s_n(\log n)/8)\tag{17}$$

令 $A_n=\{\max\limits_{1\leqslant k\leqslant K}Z_{nk}(M_n)\leqslant 2M_n n^{-1/2}\sqrt{s_n\log n}\}$,由式(17)可得

$$\begin{aligned}P(A_n)=1-P(\max_{1\leqslant k\leqslant K}Z_{nk}(M_n)\geqslant 2M_n n^{-1/2}\sqrt{s_n\log n})\geqslant&\\ 1-\exp(-c_0 s_n(\log n)/8)&\end{aligned}\tag{18}$$

又由 Cauchy - Schwarz 不等式,有

$$n\lambda_n\|(\boldsymbol{\beta}_1^*-\tilde{\boldsymbol{\beta}}_1)\|_1\leqslant n\lambda_n\sqrt{s_n}M_n$$

$$\begin{aligned}n\mu_n(\|\boldsymbol{\beta}_1^*\|_2^2-\|\tilde{\boldsymbol{\beta}}_1\|_2^2)\leqslant 2C_M n\mu_n\|\boldsymbol{\beta}_1^*-\tilde{\boldsymbol{\beta}}_1\|_1\leqslant&\\ 2C_M n\mu_n\sqrt{s_n}M_n&\end{aligned}$$

其中,$C_M=\max\{\boldsymbol{\beta}_1^*,\boldsymbol{\beta}_2^*,\cdots,\boldsymbol{\beta}_s^*,\tilde{\boldsymbol{\beta}}_1,\tilde{\boldsymbol{\beta}}_2,\cdots,\tilde{\boldsymbol{\beta}}_s\}$。因此,在事件 A_{nk} 上,对任意的 k,由式(16)可得

$$\begin{aligned}E[v_{nk}(\tilde{\boldsymbol{\beta}},\tilde{b}_k)-v_n(\boldsymbol{\beta}^*,b_k^*)]\leqslant&\\ (2\sqrt{s_n n\log n}+n\lambda_n\sqrt{s_n}+2C_M n\mu_n\sqrt{s_n})M_n&\end{aligned}$$

取 $M_n=2\sqrt{s_n/n}+\lambda_n\sqrt{s_n}+\mu_n\sqrt{s_n}$,由(A2)及假设 $\lambda_n\sqrt{s_n}\kappa_n\to 0$,$\mu_n\sqrt{s_n}\kappa_n\to 0$,可知 $M_n=o(\kappa_n^{-1}s_n^{-1/2})$,结合式(4),可得在事件 A_n 上,有

$$\begin{aligned}\frac{1}{2}c_0 n[\|\tilde{\boldsymbol{\beta}}_1-\boldsymbol{\beta}_1^*\|_2^2+(\hat{b}_k-b_k^*)^2]\leqslant&\\ E[v_{nk}(\tilde{\boldsymbol{\beta}},\tilde{b}_k)-v_n(\boldsymbol{\beta}^*,b_k^*)]\leqslant&\\ (2\sqrt{s_n n\log n}+n\lambda_n\sqrt{s_n}+2C_M n\mu_n\sqrt{s_n})M_n&\end{aligned}$$

即对任意的 k,有

$$\begin{aligned}\|\tilde{\boldsymbol{\beta}}_1-\boldsymbol{\beta}_1^*\|_2^2+(\tilde{b}_k-b_k^*)^2\leqslant&\\ C(\sqrt{s_n(\log n)/n}+\lambda_n\sqrt{s_n}+\mu_n\sqrt{s_n})^2&\end{aligned}$$

根据 $\sqrt{\|\tilde{\boldsymbol{\beta}}_1-\boldsymbol{\beta}_1^*\|_2^2+\max\limits_{1\leqslant k\leqslant K}(\tilde{b}_k-b_k^*)^2}\leqslant M_n$,可得 $\sqrt{\|\hat{\boldsymbol{\beta}}_1^o-\boldsymbol{\beta}_1^*\|_2^2+\max\limits_{1\leqslant k\leqslant K}(\hat{b}_k-b_k^*)^2}\leqslant 2M_n$。

所以在事件 A_n 上,有

$$\begin{aligned}\max_{1\leqslant k\leqslant K}(\|\hat{\boldsymbol{\beta}}_1^o-\boldsymbol{\beta}_1^*\|_2^2+(\hat{b}_k-b_k^*)^2)\leqslant&\\ C(\sqrt{s_n(\log n)/n}+\lambda_n\sqrt{s_n}+\mu_n\sqrt{s_n})^2&\end{aligned}$$

再结合式(18),可知事件 A_n 发生的概率大于等于 $1-c_1 n^{-c_2 s_n}$,定理 1 得证。

定理 2 的证明　令 $(\hat{\boldsymbol{\beta}}_1^o,\hat{b}_1,\hat{b}_2,\cdots,\hat{b}_k)$ 如定理 1 证明中所定义的目标函数 $L_n(\boldsymbol{\beta}_1,0,\boldsymbol{b})$ 的最小值点,根据 KKT 条件可得

$$\sum_{k=1}^{K}\boldsymbol{S}^{\mathrm{T}}\rho'_{\tau_k}(\boldsymbol{y}-\boldsymbol{S}\hat{\boldsymbol{\beta}}_1^o-\hat{b}_k)+n\lambda_n s(\hat{\boldsymbol{\beta}}_1^o)+n\mu_n\hat{\boldsymbol{\beta}}_1^o s(\hat{\boldsymbol{\beta}}_1^o)=0$$

其中,$s(\boldsymbol{\beta}_1)=(s(\beta_{11}),s(\beta_{12}),\cdots,s(\beta_{1s}))$,相应的函数形式为

$$s(x)=\begin{cases}1, & x>0\\ \alpha\in[-1,1], & x=0\\ -1, & x<0\end{cases}$$

对任意的 n 维向量 $\boldsymbol{u}=(u_1,u_1,\cdots,u_n)^{\mathrm{T}}$,有 $\boldsymbol{\rho}'_{\tau_k}(\boldsymbol{u})=(\boldsymbol{\rho}'_{\boldsymbol{\tau_k}}(u_1),\boldsymbol{\rho}'_{\boldsymbol{\tau_k}}(u_2),\cdots,\boldsymbol{\rho}'_{\boldsymbol{\tau_k}}(u_n))^{\mathrm{T}}$,其中 $\boldsymbol{\rho}'_{\tau_k}(u_i)=\tau_k-I(u_i\leqslant 0)$。根据 $L_n(\boldsymbol{\beta},\boldsymbol{b})$ 为凸函数的性质以及 KKT 条件,若要证明 $(\hat{\boldsymbol{\beta}}_1^o,\hat{b}_1,\hat{b}_2,\cdots,\hat{b}_k)$ 为目标函数 $L_n(\boldsymbol{\beta},\boldsymbol{b})$ 的全局最小值,只需检验下面的条件

$$\left\|\sum_{k=1}^{K}\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{\rho}'_{\boldsymbol{\tau}}(\boldsymbol{y}-\boldsymbol{S}\hat{\boldsymbol{\beta}}_1^o-\hat{b}_k)\right\|_\infty+n\mu_n\|\hat{\boldsymbol{\beta}}_1^o\|_\infty<n\lambda_n\tag{19}$$

定义事件

$$B_1=\{\max_{1\leqslant k\leqslant K}(\|\hat{\boldsymbol{\beta}}_1^o-\boldsymbol{\beta}_1^*\|_2^2+(b_k-b_k^*)^2\leqslant\gamma_n^2)\}$$

$$\begin{aligned}B_2=\{\sup_{\boldsymbol{\beta},b_1,b_2,\cdots,b_K\in N}\|\sum_{k=1}^{K}\boldsymbol{Q}^{\mathrm{T}}\rho'_{\tau_k}(\boldsymbol{y}-\boldsymbol{S}\boldsymbol{\beta}_1-b_k)\|_\infty+&\\ n\mu_n\|\boldsymbol{\beta}_1\|_\infty<n\lambda_n\}&\end{aligned}$$

其中,γ_n 同定理 1 中所定义,并且

$$\begin{aligned}\boldsymbol{N}=\{(\boldsymbol{\beta}_1^{\mathrm{T}},\boldsymbol{\beta}_2^{\mathrm{T}})^{\mathrm{T}}\in\boldsymbol{R}^p,b_k s\in\mathbf{R}:\sup_{1\leqslant k\leqslant K}\|\boldsymbol{\beta}_1-\boldsymbol{\beta}_1^*\|_2^2+&\\ (b_k-b_k^*)^2\leqslant\gamma_n^2,\boldsymbol{\beta}_2=0\in\boldsymbol{R}^{p-s}\}&\end{aligned}$$

由于在事件 B_1 上,$(\hat{\boldsymbol{\beta}},\hat{b}_k)\in\boldsymbol{N}$,故式(19)在事件 $B_1\cap B_2$ 上成立。结合定理 1 和引理 3,可得 $P(B_1\cap B_2)\geqslant 1-c_1 n^{-c_2 s_n}$,定理 2 得证。

6 结 论

本文提出了一种基于弹性网惩罚复合分位数回归的稳健估计模型。该模型不仅能够进行变量选择从而产生稀疏解,而且可以处理变量之间的高度多重共线性,具有较广的适应性和较强应用价值。此外,模型在面对具有极端值或是呈现尖峰、厚尾分布的数据时也能够保持很好的稳健特性。在一定条件下,理论分析证明了该模型估计所具有的相合性与稀疏性,表明所提出的模型能够将不相关的变量完全压缩至零,且估计值和真实系数以趋于 1 的概率相同。此外,数值模拟采用了 5 种误差项分布假设,根据设定的 4 项指标,通过与不同惩罚函数模型以及损失函数模型比较,证实了所提方法的优越性与稳健性。

参考文献(References):

[1] AKAIKE H. A new look at the statistical model identification[J]. Transactions on Automatic Control, 1974, 19(2): 716—723.

[2] SCHWARZ G. Estimating the dimension of a model[J]. Annals of Statistics, 1978, 6(1): 461—464.

[3] HOERL A, KENNARD R, HOERL R. Practical use of ridge regression: a challenge met[J]. Journal of the Royal Statistical Society: Series C, 1985, 34(2): 114—120.

[4] TIBSHIRANI R. Regression shrinkage and selection via the lasso[J]. Journal of the Royal Statistical Society: Series B, 1996, 58(1): 267—288.

[5] FAN J Q, LI R Z. Variable selection via nonconcave penalized likelihood and its oracle properties[J]. Journal of the American Statistical Association, 2001, 96(456): 1348—1360.

[6] ZHANG C H. Nearly unbiased variable selection under minimax concave penalty[J]. Annals of Statistics, 2010, 38(2): 894—942.

[7] ZOU H, HASTIE T. Regularization and variable selection via the clastic net[J]. Journal of the Royal Statistical Society, 2005, 67(2): 768—768.

[8] 卢颖. 广义线性模型基于 Elastic Net 的变量选择方法研究[D]. 北京: 北京交通大学, 2011.
LU Ying. Research on variable selection method of generalized linear model based on elastic net [D]. Beijing: Beijing Jiaotong University, 2011.

[9] 黄登香. Elastic Net 方法在几类模型变量选择中的应用[D]. 南宁: 广西大学, 2014.
HUANG Deng-xiang. Application of elastic net method in the selection of model variables [D]. Nanning: Guangxi University, 2014.

[10] 李洪选. 平衡纵向数据模型变量选择的 Elastic Net 方法研究[J]. 泰山学院学报, 2017, 39(3): 5—10.
LI Hong-xuan. Elastic net method for variable selection of balanced longitudinal data model [J]. Journal of Taishan University, 2017, 39(3): 5—10.

[11] KOENKER R. Quantiles regression [M]. Cambridge: Cambridge University, 2005.

[12] FAN J Q, FAN Y Y. Adaptive robust variable selection[J]. The Annals of Statistics, 2014, 42(1): 324—351.

[13] SU M H, WANG W J. Elastic net penalized quantile regression model[J]. Journal of Computational and Applied Mathematics, 2021, 392: 113462—113469.

[14] ZOU H, YUAN M. Composite quantile regression and the oracle model selection theory[J]. The Annals of Statistics, 2008, 36(4): 1108—1126.

[15] BÜHLMANN P, GEER S. Statistics for high-dimensional data: methods, theory and applications[M]. Berlin, Heidelberg: Springer, 2011.

责任编辑:李翠薇