稳健 L_q ($0 < q < 1$) 正则化理论: 解的渐近分布与变量选择一致性

常象宇^①, 徐宗本^{①*}, 张海^②, 王建军^③, 梁勇^④

① 西安交通大学理学院信息与系统科学研究所, 西安 710049;

② 西北大学数学系, 西安 710069;

③ 西南大学数学与统计学院, 重庆 400715;

④ 澳门科技大学资讯科技学院, 澳门 999078

E-mail: xiangyuchang@gmail.com, zbxu@mail.xjtu.edu.cn, zhanghai@nwu.edu.cn, wjj@swu.edu.cn, yliang@must.edu.mo

收稿日期: 2010-01-31; 接受日期: 2010-08-12; * 通信作者

国家重点基础研究计划 (批准号: 2007CB311002) 和国家自然科学基金 (批准号: 60975036, 11001227) 资助项目

摘要 本文提出稳健 L_q ($0 < q < \infty$) 正则化模型, 证明该模型解的全局渐近分布定理. 应用该定理可进一步证明当 $0 < q < 1$ 时所提出的模型具有变量选择一致性, 从而具有良好的变量选择能力. 基于所获得的理论结果, 提出一类求解该模型的无参量加权迭代算法, 并给出相应的正则化参数选择策略. 数值试验表明本文提出的模型与算法可行、有效, 有广泛的应用价值.

关键词 稳健回归 渐进分布 变量选择 Oracle 性质 LAD

MSC (2000) 主题分类 62F12, 62F35, 62J07

1 引言

随着信息获取技术的飞速发展, 人们所面临处理的数据常常呈现高维、海量和不确定等基本特征. 如何发展有效的数据处理技术以适应这样的高复杂性数据已成为科学与工程领域广泛关注的焦点. 线性回归模型作为一种最基本、最重要的数据分析技术已被广泛应用, 但对于海量、高维和不确定的高复杂性数据仍遭遇前所未有的挑战.

经典的线性回归模型假定回归变量 $y \in \mathbb{R}$ 与预测变量 $x \in \mathbb{R}^p$ 之间存在线性关系

$$y = x^T \beta + \varepsilon, \quad (1.1)$$

其中 $\beta = (\beta_1, \beta_2, \dots, \beta_p)^T \in \mathbb{R}^p$, $\varepsilon \sim N(0, \sigma^2)$ 服从正态分布. 为了估计系数 β , 一般假设已经采集了 n 个样本 $\{(x_i, y_i) | y_i \in \mathbb{R}, x_i \in \mathbb{R}^p, i = 1, 2, \dots, n\}$, 并采用最小二乘原理

$$\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^n (y_i - x_i^T \beta)^2$$

求解. 记 $Y = (y_1, y_2, \dots, y_n)^T \in \mathbb{R}^n$ 为回归样本, $X = (x_1, x_2, \dots, x_n)^T \in \mathbb{R}^{n \times p}$ 为预测样本, 则模型 (1.1) 的最小二乘解为

$$\hat{\beta}_{ols} = (X^T X)^{-1} X^T Y.$$

引用格式: Chang X Y, Xu Z B, Zhang H, et al. Robust regularization theory based on L_q ($0 < q < 1$) regularization: the asymptotic distribution and variable selection consistence of solutions (in Chinese). Sci Sin Math, 2010, 40(10): 985–998, doi: 10.1360/012010-77

最小二乘估计使用非常广泛, 且具有良好的理论性质 (如最优线性无偏估计性). 但问题是: 第一, 最小二乘估计对噪声敏感, 稳健性差, 尤其当误差为重尾分布时, 其预测效果不理想; 第二, 最小二乘估计虽然可以在一定条件下达到无偏, 但其估计值方差大. 这两个缺陷使得它在应用中受到局限. 为了克服这些局限, 人们提出了稳健回归方法与正则化方法.

稳健回归方法作为一种能够克服大噪声的回归估计方法而被广泛使用, 其中最为常用的是最小一乘方法, 亦被称之为 LAD (least absolute deviation) 方法:

$$\hat{\beta}_{\text{LAD}} = \arg \min \sum_{i=1}^n |y_i - x_i^T \beta|.$$

LAD 方法作为稳健回归估计方法具有很长的研究历史^[1], 自 Chanes 等^[2] 提出使用线性规划方法快速而高效地求解以来, 它特别引起了广泛的关注. 例如, 文献 [4–10] 详细分析了 LAD 的大样本统计性质 (如解的渐近分布性), 文献 [11–13] 揭示了它比最小二乘估计具有更好的稳健性.

正则化是针对如何均衡待估计统计量的偏差与方差, 以及解决病态问题而提出来的方法. 基于最小二乘的正则化方法具有如下形式:

$$\min \left\{ \sum_{i=1}^n (y_i - x_i^T \beta)^2 + \lambda_n \text{Pen}(\beta) \right\},$$

这里 $\lambda_n > 0$ 为正则化参数, Pen 是关于回归系数的函数, 通常取为 β 的某种范数形式. 例如, 取 $\text{Pen}(\beta) = \|\beta\|_2^2$, 该模型即为经典的岭回归^[14]; 取 $\text{Pen}(\beta) = \|\beta\|_1 = \sum_{i=1}^p |\beta_i|$, 该模型即为 Lasso 模型^[15]; 取 $\text{Pen}(\beta) = \sum_{j=1}^p |\beta_j|^q, 1 \leq q < \infty$ 时, 该模型为 Bridge 模型^[16]. Knight 和 Fu^[17] 最先研究了当 $\text{Pen}(\beta) = \sum_{j=1}^p |\beta_j|^q, 0 < q < \infty$ 时, 该模型解的渐近性质; Huang^[18] 等研究了 Bridge 模型的高维统计渐进性质. 为了同时克服最小二乘估计的两个缺陷, 文献 [19] 首先提出了结合稳健回归与正则化方法的 RLAD 模型, 文献 [20, 21] 又进一步将其推广到 LAD-Lasso 和 LAD-Adaptive Lasso 模型. 所有这些模型都具有如下统一的形式:

$$\min \left\{ \sum_{i=1}^n |y_i - x_i^T \beta| + \sum_{j=1}^p \lambda_{n,j} |\beta_j| \right\}.$$

显然, 这些模型属于稳健 L_1 正则化框架. 然而近期的研究发现, L_1 正则化框架有其明显的缺陷^[22–24], 而 L_q ($0 < q < 1$) 正则化框架有更好的理论性质与应用表现. 例如文献 [22] 通过实验研究了 L_1 框架下变量选择能力在高维情形下的不足; 而对于稀疏信号重建问题, 文献 [22–24] 均说明使用 L_q ($0 < q < 1$) 框架需要比 L_1 框架更少的采样; 进一步, 文献 [24] 使用 Donoho^[25] 提出的相变理论验证了变量选择能力上 L_q ($0 < q < 1$) 框架在很多情形下优于 L_1 框架. 基于此, 我们本文提出如下稳健 L_q 正则化框架

$$\hat{\beta}_{n,q} = \arg \min \left\{ \sum_{i=1}^n |y_i - x_i^T \beta| + \sum_{j=1}^p \lambda_{n,j} |\beta_j|^q \right\}, \quad (1.2)$$

并称之为 LAD- L_q 模型, 这里 $0 < q < \infty$. 显然, 当 $q = 1$ 时, LAD- L_q 模型退化为 RLAD^[19], LAD-Lasso^[20] 和 LAD-Adaptive Lasso^[21] 模型. 本文拟通过对以上 LAD- L_q 模型的研究, 试图从理论上揭示 LAD- L_q 模型当 $0 < q < 1$ 时比 $q = 1$ 时有更优良的性质.

本文第 2 节证明 LAD- L_q 模型解的全局渐近分布定理. 第 3 节应用所证明的定理分析 LAD- L_q 模型的变量选择能力, 证明基于 L_1 框架的 RLAD 模型在某些情况下不具有变量选择一致性, 而基于

L_q ($0 < q < 1$) 正则化框架的 LAD- L_q ($0 < q < 1$) 模型有良好的 Oracle 性质, 从而具有变量选择一致性. 基于所获得的理论结果, 第 4 节提出一类无参量的加权迭代求解算法, 以及与该算法配套的参数选择策略. 最后, 在第 5 节, 我们提供数值比较实验说明 LAD- L_q ($0 < q < 1$) 模型较其它模型有更突出的变量选择能力 (当误差分布为重尾分布时).

2 LAD- L_q 模型解的渐近分布

假定回归样本 $Y = (y_1, y_2, \dots, y_n)^T$, 预测样本 $X = (x_1, x_2, \dots, x_n)^T \in \mathbb{R}^{n \times p}$. 不失一般性 (必要时可对数据正规化), 我们假设

$$\sum_{i=1}^n y_i = 0, \quad \sum_{i=1}^n x_{ij} = 0, \quad \sum_{i=1}^n x_{ij}^2 = 1, \quad \forall j = 1, 2, \dots, p.$$

为了叙述方便, 我们还假设:

(i) 对任意的 y_i 成立 $y_i = x_i^T \beta + \varepsilon_i$, 其中 $\varepsilon_i, i = 1, 2, \dots, n$, 独立同分布 (其分布函数是 $F(x)$), 满足 $E(\varepsilon_i) = 0$ 和 $\text{Var}(\varepsilon_i) = \sigma^2$.

(ii)

$$\frac{1}{n} X^T X \rightarrow C \quad (n \rightarrow \infty), \quad (2.1)$$

这里 C 是正定矩阵.

(iii) 定义实值序列

$$\phi_n(t) = \int_0^t \sqrt{n}(F(s/\sqrt{n}) - F(0)) ds. \quad (2.2)$$

假设对于任意向量 $u \in \mathbb{R}^p$, 下述极限存在:

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \phi_n(u^T x_i) = \tau(u), \quad (2.3)$$

其中 $\tau(u)$ 为定义在 $[0, \infty]$ 上的凸的实值函数.

本节证明如下有关 LAD- L_q 模型解的渐近分布定理.

定理 1 假定 $0 < q \leq 1$ 且上述 (i)–(iii) 成立, C 和 $\tau(u)$ 如 (2.1) 与 (2.2) 所定义, $\hat{\beta}_{n,q}$ 为 LAD- L_q ($0 < q \leq 1$) 模型 (1.2) 的解. 则对于任意的 $j \in \{1, 2, \dots, p\}$, 当 $\lambda_{n,j}/\sqrt{n} \rightarrow \lambda_j \geq 0$ 时

$$\sqrt{n}(\hat{\beta}_{n,q} - \beta) \rightarrow_d \arg \min(V(u)),$$

其中, 当 $0 < q < 1$ 时

$$V(u) = -u^T W + 2\tau(u) + \sum_{j=1}^p \lambda_j q |u_j|^q I(\beta_j = 0); \quad (2.4)$$

当 $q = 1$ 时

$$V(u) = -u^T W + 2\tau(u) + \sum_{j=1}^p \lambda_j \{u_j \text{sgn}(\beta_j) I(\beta_j \neq 0) + |u_j| I(\beta_j = 0)\},$$

这里 $W \sim N(0, C)$. 进一步, 如果误差 ε 的分布函数 $F(x)$ 满足 $F'(0) \neq 0$ 且 $\lambda_j = 0, j = 1, 2, \dots, p$, 则

$$\sqrt{n}(\hat{\beta}_{n,q} - \beta) \rightarrow_d N\left(0, \frac{C^{-1}}{4F'^2(0)}\right).$$

证明 首先易证对于任意的 $x \neq 0$, 成立下述不等式

$$|x - y| - |x| = -y[I(x > 0) - I(x < 0)] + 2 \int_0^y [I(x \leq s) - I(x \leq 0)]ds. \quad (2.5)$$

记

$$V_n(u) = \sum_{i=1}^n \{|y_i - x_i^T(\beta + u/\sqrt{n})|\} + \sum_{j=1}^p \lambda_{n,j} |\beta_j + u_j/\sqrt{n}|^q - \sum_{i=1}^n \{|y_i - x_i^T \beta|\} + \sum_{j=1}^p \lambda_{n,j} |\beta_j|^q, \quad (2.6)$$

把 $V_n(u)$ 分解成三部分:

$$V_n(u) = V_{n1}(u) + V_{n2}(u) + \Phi(\lambda_n),$$

其中

$$\begin{aligned} V_{n1}(u) &= -\frac{1}{\sqrt{n}} \sum_{i=1}^n x_i^T u [I(\varepsilon_i > 0) - I(\varepsilon_i < 0)], \\ V_{n2}(u) &= 2 \sum_{i=1}^n \int_0^{x_i^T u/\sqrt{n}} [I(\varepsilon_i \leq s) - I(\varepsilon_i \leq 0)]ds = \sum_{i=1}^n V_{n2}^i(u), \\ \Phi(\lambda_n) &= \sum_{j=1}^p \lambda_{n,j} \{|\beta_j + u_j/\sqrt{n}|^q - |\beta_j|^q\}. \end{aligned}$$

我们将通过分别估计上述三部分的渐近分布来证明定理 1. 首先, 由 Lindeberg-Feller 中心极限定理, 对于任意的 u 易证, $V_{n1}(u) \rightarrow_d -u^T W$, 这里 $W \sim N(0, C)$. 而由假设 $\lambda_{n,j}/\sqrt{n} \rightarrow \lambda_j \geq 0$, 当 $0 < q < 1$ 时显然有

$$\Phi(\lambda_n) \rightarrow \sum_{j=1}^p \lambda_j q |u_j|^q I(\beta_j = 0),$$

而当 $q = 1$ 时有

$$\Phi(\lambda_n) \rightarrow \sum_{j=1}^p \lambda_j \{u_j \operatorname{sgn}(\beta_j) I(\beta_j \neq 0) + |u_j| I(\beta_j = 0)\}.$$

对于 $V_{n2}(u)$, 我们注意到, 它可改写为

$$V_{n2}(u) = \sum_{i=1}^n E(V_{n2}^i(u)) + \sum_{i=1}^n (V_{n2}^i(u) - E(V_{n2}^i(u))).$$

利用假设 (ii) 和 (iii) 易导出

$$\begin{aligned} \sum_{i=1}^n E(V_{n2}^i(u)) &= 2 \sum_{i=1}^n \int_0^{x_i^T u/\sqrt{n}} [F(s) - F(0)]ds \\ &= \frac{2}{\sqrt{n}} \sum_{i=1}^n \int_0^{x_i^T u/\sqrt{n}} [F(t/\sqrt{n}) - F(0)]dt \end{aligned}$$

$$\begin{aligned}
&= \frac{2}{n} \sum_{i=1}^n \int_0^{x_i^T u} \sqrt{n} [F(t/\sqrt{n}) - F(0)] dt \\
&= \frac{2}{n} \sum_{i=1}^n \phi_n(x_i^T u) \rightarrow 2\tau(u).
\end{aligned}$$

$$\begin{aligned}
\text{Var}(V_{n2}(u)) &= \sum_{i=1}^n E[(Z_{n2}^i(u) - E(V_{n2}^i(u)))^2] \\
&\leq \frac{2}{\sqrt{n}} \max_{1 \leq i \leq n} |x_i^T u| \sum_{i=1}^n E(V_{n2}^i(u)) \\
&= \frac{2}{\sqrt{n}} \max_{1 \leq i \leq n} |x_i^T u| E(V_{n2}(u)).
\end{aligned}$$

于是, 如果 $\tau(u) < \infty$, 我们有

$$V_{n2}(u) - E(V_{n2}(u)) \rightarrow_p 0, \text{ 当 } n \rightarrow \infty \text{ 时},$$

所以 $V_{n2}(u) \rightarrow_p 2\tau(u)$; 而如果 $\tau(u) = \infty$, 则由 Markov 不等式可导出

$$P(|V_{n2}(u) - E(V_{n2}(u))| > \epsilon E(V_{n2}(u))) \leq \frac{\text{Var}(V_{n2}(u))}{\epsilon^2 E(V_{n2}(u))^2} \leq \frac{2 \max_{1 \leq i \leq n} |x_i^T u|}{\sqrt{n} \epsilon^2 E(V_{n2}(u))} \rightarrow 0,$$

这推出 $V_{n2} \rightarrow_p \infty = \tau(u)$. 于是我们得到

$$V_n(u) \rightarrow_d V(u), \quad (2.7)$$

这里

$$V(u) = \begin{cases} -u^T W + 2\tau(u) + \sum_{j=1}^p \lambda_j q |u_j|^q I(\beta_j = 0), & 0 < q < 1, \\ -u^T W + 2\tau(u) + \sum_{j=1}^p \lambda_j \{u_j \text{sgn}(\beta_j) I(\beta_j \neq 0) + |u_j| I(\beta_j = 0)\}, & q = 1, \end{cases} \quad (2.8)$$

而根据 (2.6) 我们可以求得

$$\sqrt{n}(\hat{\beta}_{n,q} - \beta) = \arg \min(V_n(u)).$$

如果 $V(u)$ 具有唯一最小值, 通过 [26] 我们可以验证

$$\arg \min(V_n(u)) = \sqrt{n}(\hat{\beta}_{n,q} - \beta) \rightarrow_d \arg \min(V(u)).$$

进一步, 如果假设 $F(x)$ 是误差项 ε 的分布函数, 且 $F'(0) \neq 0, \lambda_j = 0, j = 1, 2, \dots, p$, 则

$$\begin{aligned}
\phi(t) &= \lim_{n \rightarrow \infty} \phi_n(t) \\
&= \lim_{n \rightarrow \infty} \int_0^t \sqrt{n} (F(s/\sqrt{n}) - F(0)) ds \\
&= \int_0^t \lim_{n \rightarrow \infty} \sqrt{n} (F(s/\sqrt{n}) - F(0)) ds \\
&= \int_0^t s F'(0) ds
\end{aligned}$$

$$= \frac{t^2}{2} F'(0).$$

而从 (2.3) 式知

$$\tau(u) = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \phi_n(x_i^T u) = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \frac{1}{2} F'(0) u^T x_i X^T u = \frac{F'(0)}{2} u^T C u. \quad (2.9)$$

结合 (2.7) 和 (2.8), 我们推出

$$V(u) = -u^T W + F'(0) u^T C u.$$

显然 $V(u)$ 是凸函数, 令

$$\frac{dV}{du} = -W + 2F'(0) C u = 0,$$

则立刻可得 $V(u)$ 的极值点 $u^* = \frac{C^{-1}W}{2F'(0)} \sim N(0, \frac{C^{-1}}{4F'^2(0)})$. 根据 (2.7), 这即说明

$$\sqrt{n}(\hat{\beta}_{n,q} - \beta) \rightarrow_d \arg \min(V(u)) = u^* \sim N\left(0, \frac{C^{-1}}{4F'^2(0)}\right).$$

证毕.

定理 2 假设 $q > 1$, 则在定理 1 相同的条件下

$$\sqrt{n}(\hat{\beta}_{n,q} - \beta) \rightarrow_d \arg \min(V(u)),$$

其中

$$V(u) = -u^T W + 2\tau(u) + \sum_{j=1}^p \lambda_j q u_j \operatorname{sgn}(\beta_j) |\beta_j|^{q-1}.$$

进一步, 如果 $F(x)$ 是误差项 ε 的分布函数, 且 $F'(0) \neq 0, \lambda_j = 0, j = 1, 2, \dots, p$, 则

$$\sqrt{n}(\hat{\beta}_{n,q} - \beta) \rightarrow_d N\left(0, \frac{C^{-1}}{4F'^2(0)}\right).$$

证明 证明过程类似于定理 1.

注 1 上述定理 1 和 2 刻画了 $LAD-L_q$ ($0 < q < \infty$) 模型解的渐近分布, 它对于理解 $LAD-L_q$ 模型的意义与价值有重要作用 (例如见第 3 节). 注意到, 文献 [20, 21] 曾经证明过 $LAD-L_1$ 模型的渐近分布定理. 虽然其证明的结论是全局性的, 但其证明方法却为一种局部的方法 (本质上利用了 $LAD-L_1$ 模型的凸性). 而本文所运用的证明方法是一种直接的全局性方法, 不仅在 $q = 1$ 时, $LAD-L_1$ 模型适用, 而且还可以用于证明非凸的 $LAD-L_q$ ($0 < q < 1$) 模型的全局渐进分布定理. 另外, 我们注意到, 导出回归模型解的渐近分布传统上总是假定误差服从正态分布 [6], 而上述定理 1 和 2 是在非常一般的误差分布 (特别的, 不假设正态分布) 条件下证明的. 这从一定意义上预示着: 本文所提出的 $LAD-L_q$ 模型较传统线性回归模型可能会更加稳健. 我们将在第 5 节给出应用实例进一步说明这一点.

3 $LAD-L_q$ ($0 < q \leq 1$) 模型的变量选择一致性

变量选择问题是在线性表示的意义下从繁多冗余的影响变量中确定真实主变量的问题. 传统的变量选择方法是比较待选变量的所有子集, 从中挑选特定意义下 (例如 AIC, BIC 和 C_p 准则) 的最优变

量. 但是, 这样的方法对于高维数问题无法应用 (事实上, 一个含有 p 个影响变量的问题, 其待选子集数为 $2^p - 1$ 个), 最小二乘正则化方法 (尤其是 Lasso^[15] 方法) 是解决高维数变量选择问题的有力工具. 但是正如文献 [27–29] 所指出的, Lasso 方法在很多条件下不具有变量选择一致性, 从而其变量选择能力并不理想. 克服这一缺陷陆续产生了 Adaptive Lasso^[28], Nonnegative Garotte^[29] 和 SCAD^[30] 等模型. 自然地, 人们也考虑运用稳健正则化方法来解决稳健回归中的变量选择问题, 例如先后提出 RLAD^[19], LAD-Lasso^[20] 和 LAD-Adaptive Lasso^[21] 等模型. 正如本文引言所指出的, 所有这些方法均属于稳健 L_1 正则化框架 (即 LAD- L_1 模型). 一个自然的问题是: 像 RLAD 这样的 LAD- L_1 模型是有效的变量选择方法吗? 对于变量选择, LAD- L_1 模型与 LAD- L_q ($0 < q < 1$) 模型有不同吗?

本节证明: RLAD 模型在某些情况下不具有变量选择一致性, 而 LAD- L_q ($0 < q < 1$) 模型具有变量选择一致性. 这说明 LAD- L_q ($0 < q < 1$) 模型比 LAD- L_1 模型具有更优良的变量选择能力.

为了以下叙述方便, 假设集合 $\mathcal{A} = \{j : \beta_j \neq 0\}$ 和 $\mathcal{A}_f = \{j : \hat{\beta}_f^j \neq 0\}$ 分别为真实模型中的非零变量指标集和某种估计方法 f 的非零变量指标集. 例如, 对于 RLAD 模型

$$\hat{\beta}_{\text{RLAD}} := \arg \min \left\{ \sum_{i=1}^n |y_i - x_i^T \beta| + \lambda_n \sum_{j=1}^p |\beta_j| \right\},$$

有 $\mathcal{A}_{\text{RLAD}} = \{j : \hat{\beta}_{\text{RLAD}}^j \neq 0\}$.

我们首先来说明 RLAD 模型的变量选择非一致性.

定理 3 在 RLAD 模型中, 如果 $\lambda_n/\sqrt{n} \rightarrow \lambda_0 \geq 0$, 则 $\hat{\beta}_{\text{RLAD}}$ 对变量选择不具有一致性, 换言之, 存在常数 $0 \leq c < 1$ 满足

$$\limsup_{n \rightarrow \infty} P\{\mathcal{A}_{\text{RLAD}} = \mathcal{A}\} \leq c < 1.$$

证明 我们简记 $\hat{\beta}_{\text{RLAD}}$ 的第 j 个分量为 $\hat{\beta}_j$. 如果 $\mathcal{A}_{\text{RLAD}} = \mathcal{A}$, 则显然 $\hat{\beta}_j = 0$ 对任意 $j \notin \mathcal{A}$ 成立, 从而, $P(\mathcal{A}_{\text{RLAD}} = \mathcal{A}) \leq P(\sqrt{n}\hat{\beta}_j = 0, \forall j \notin \mathcal{A})$. 于是由弱收敛的性质与定理 1, 有 $\limsup_{n \rightarrow \infty} P(\sqrt{n}\hat{\beta}_j = 0, \forall j \notin \mathcal{A}) \leq P(u_j^* = 0, \forall j \notin \mathcal{A})$. 因此我们只需证明 $c = P(u_j^* = 0, \forall j \notin \mathcal{A}) < 1$ 即可. 以下分两种情况讨论:

情况 1: $\lambda_0 = 0$, 此时 $u_j^* = \frac{1}{2F'(0)} C_j^{-1} W \sim N(0, (C_j^{-1})^T C C_j^{-1} / 4F'^2(0))$, 从而有 $c = 0$.

情况 2: $\lambda_0 > 0$, 此时由 Karush-Kuhn-Tucker 最优性条件推知

$$|-W_2 + 2F'(0)C_{21}u_1^*| \leq \lambda_0.$$

因此

$$c \leq P(|-W_2 + C_{21}C_{11}^{-1}(W_1 - \lambda_0 \text{sgn}(\hat{\beta}_{n1}))| \leq \lambda_0) < 1.$$

证毕.

然而, 下述定理说明, 当 $0 < q < 1$ 时, LAD- L_q 模型却具有变量选择一致性.

定理 4 设真实模型为 $y = \beta^T x$, $\beta = (\beta_1^T, \beta_2^T)^T$, 其中 $\beta_1 \neq 0$ 和 $\beta_2 = 0$. 而 $\hat{\beta}_{n,q} = (\hat{\beta}_{1,q}^T, \hat{\beta}_{2,q}^T)^T$ 是 LAD- L_q ($0 < q < 1$) 模型所定义的估计值, 其中 $\hat{\beta}_{1,q}^T, \hat{\beta}_{2,q}^T$ 分别是与 β_1^T, β_2^T 维数相同的分量. 则当 $\lambda_{n,j}/n^{q/2} \rightarrow \infty, \lambda_{n,j}/\sqrt{n} \rightarrow \lambda_j \geq 0$ 时, $\hat{\beta}_{n,q}$ 满足以下 Oracle 性质:

(i) 渐近正态性:

$$\sqrt{n}(\hat{\beta}_{1,q} - \beta_1) \rightarrow_d N\left(D, \frac{C_{11}^{-1}}{4F'^2(0)}\right),$$

其中 $D = (-q\lambda_1|\beta_1|^{q-1}\text{sgn}(\beta_1), \dots, -q\lambda_p|\beta_m|^{q-1}\text{sgn}(\beta_m))^T$, m 为 β_1 的维数, C_{11} 为正定阵 C 的前 $m \times m$ 主子矩阵.

(ii) 变量选择一致性:

$$\lim_{n \rightarrow \infty} P\{\mathcal{A}_{\text{LAD-}L_q(0 < q < 1)} = \mathcal{A}\} = 1.$$

证明 假设 $\beta_1 = (\beta_1, \beta_2, \dots, \beta_m)^T$ 对于任意的 $j = 1, 2, \dots, m, \beta_j \neq 0$ 且 $\beta_2 = (0, 0, \dots, 0)^T$. 则因 $\lambda_{n,j}/n^{q/2} \rightarrow \infty$ 和 $\lambda_{n,j}/\sqrt{n} \rightarrow \lambda_j \geq 0$, 所以有

$$\lim_{n \rightarrow \infty} \sum_{i=m+1}^p \lambda_{n,i} [|\beta_j + u_j/\sqrt{n}|^q - |\beta_j|^q] = \begin{cases} 0, & \text{当 } u_{m+1} = u_{m+2} = \dots = u_p = 0, \\ \infty, & \text{其他.} \end{cases}$$

于是, 依定理 1 和 (2.6) 式有

$$V(u) = \begin{cases} -u^T W + 2\tau(u) + \sum_{j=1}^m \lambda_j q u_j |\beta_j|^{q-1} \text{sgn}(\beta_j), & \text{当 } u_{m+1} = u_{m+2} = \dots = u_p = 0, \\ \infty, & \text{其他,} \end{cases}$$

且

$$\sqrt{n}(\hat{\beta}_{n,q} - \beta) \rightarrow_d \arg \min(V(u)) = u^* = (u_1^*, u_2^*),$$

所以

$$\hat{\beta}_{2,q} \rightarrow_d u_2^* = 0.$$

由于 $F(x)$ 在原点可导, 有

$$V(u_1) = -u_1^T W_1 + F'(0)u_1^T C_{11}u_1 + \sum_{j=1}^m \lambda_j q u_j |\beta_j|^{q-1} \text{sgn}(\beta_j).$$

由 $\frac{dV(u_1)}{du_1} = 0$, 于是

$$\sqrt{n}(\hat{\beta}_{1,q} - \beta_1) \rightarrow_d \arg \min V(u_1) = u_1^* = \frac{C_{11}}{2F'(0)}(W_1 + D) \sim N\left(D, \frac{C_{11}^{-1}}{4F'^2(0)}\right).$$

这推出 $\hat{\beta}_{1,q}$ 具有渐近正态性. 根据上式, 自然也推出 $\hat{\beta}_{n,q}$ 具有变量选择一致性. 证毕.

注 2 按照文献 [30] 的标准, 一个变量选择方法是有效的, 如果它具有 Oracle 性质. 所以, 上述定理 3 和 4 说明了: $\text{LAD-}L_q$ ($0 < q < 1$) 模型是一个有效的变量选择方法, 但是 $\text{LAD-}L_1$ (RLAD) 模型并不是. 这一结论展示了模型 $\text{LAD-}L_q$ ($0 < q < 1$) 模型的价值与重要性.

4 一个无参量加权迭代算法

$\text{LAD-}L_q$ ($0 < q < 1$) 模型是一类非凸优化问题, 对这类非凸优化问题近年来已发展出了有效的加权迭代算法 (例如见文献 [22, 31, 32]). 该类算法的基本思想是: 对于正则项 $|\beta|^q$ 用其在某一点处的 Taylor 展式一阶近似替代, 即

$$|\beta|^q \approx |\beta_0|^q + \frac{1}{|\beta_0|^{1-q}}(|\beta| - |\beta_0|).$$

于是 LAD- L_q ($0 < q < 1$) 模型的加权迭代算法可写为:

$$\min \left\{ \sum_{i=1}^n |y_i - x_i^T \beta| + \sum_{j=1}^p \frac{\lambda_{n,j}}{|\beta_{0,j}|^{1-q}} |\beta_j| \right\}.$$

而应用上述算法的关键是如何设计上述优化问题的正则化参数 ($\lambda_{n,j}, j = 1, 2, \dots, p$). 以下我们通过极大化后验概率的方法给出参数 ($\lambda_{n,j}, q$) 的选择方式. 假设回归系数有如下先验分布

$$f(\beta) = \frac{1}{\lambda} \exp\{-\lambda \|\beta\|^q\},$$

则可得 BIC 型后验概率的似然函数为

$$\sum_{i=1}^n |y_i - x_i^T \beta| + \sum_{j=1}^p (\lambda_{n,j} |\beta_j|^q - \log(0.5 \lambda_{n,j}) \log(n)).$$

极大化上述后验概率给出 $\lambda_{n,j} = \frac{\log(n)}{|\hat{\beta}_j|^q}$. 一个自然的想法是利用 β_j 的估计值来代替真实值, 而去获得 $\lambda_{n,j}$ 的估计值. 已经知道 [5], $\hat{\beta}_{\text{LAD}}$ 是 $\sqrt{n}$ 一致的估计 (相对于真实值), 所以我们建议使用 $\hat{\beta}_{\text{LAD}}$, 于是取

$$\hat{\lambda}_{n,j} = \frac{\log(n)}{|\hat{\beta}_{\text{LAD},j}|^q}.$$

容易验证, 以上选取的 $\hat{\lambda}_{n,j}$ 满足定理 4 之假设, 于是, 根据定理 4, 这样的选择使得该算法的解具有变量选择一致性.

总结以上结论, 我们提出求解 LAD- L_q ($0 < q < 1$) 模型的如下算法:

步骤 1 令 $t = 0$, 设定最大迭代次数 K , 初始化 $\beta_0 = \hat{\beta}_{\text{LAD}}$;

步骤 2 求解

$$\beta^{t+1} = \arg \min \left\{ \sum_{i=1}^n |y_i - x_i^T \beta| + \sum_{j=1}^p \frac{\log(n)}{|\beta_j^t|} |\beta_j| \right\},$$

令 $t = t + 1$;

步骤 3 当 $t < K$ 时, 转步骤 2; 否则终止.

注 3 (i) 上述步骤 2 中求解的是一个线性规划问题, 任何快速的线性规划求解方法都可以使用 (例如内点法等). (ii) 使用 [22, 定理 1] 或 [32, 定理 2] 可证明: 上述算法保证其目标函数单调下降, 并收敛到某个稳定值. (iii) 不同于已知的加权迭代算法, 上述算法不涉及参数设定问题 (既不要设定正则化参数 λ , 也不要选择 q), 是一类无参量算法, 应用非常方便.

5 数值比较实验

本节中, 我们提供一组实验, 比较各种已知变量选择方法的性能. 从而, 展示本文提出的 LAD- L_q ($0 < q < 1$) 模型的优良稳健性与变量选择能力.

实验 1 考虑如下模型 $y = x^T \beta + \alpha \varepsilon$, 其中 $\beta = (0.5, 1.0, 1.5, 2, 0, 0, 0, 0)$, $\varepsilon \sim t(3)$. 我们完成两组实验: 第一组实验中训练样本与测试样本均由标准正态分布 $N(0, I)$ 产生, 其中训练样本 50 个, 测试样本 100 个 (这样产生的样本间不具有相关性). 第二组实验中训练样本与测试样本均由非标准正态分布 $N(0, \Sigma)$, $\Sigma_{i,j} = 0.5^{|i-j|}$ 产生, 其中训练样本 50 个, 测试样本 100 个 (样本间具有相关性). 我们分别应用 OLS, Ridge^[14], Lasso^[15], Adaptive Lasso^[28], LAD^[2], RLAD^[19], LAD-Lasso^[20], LAD-Adaptive

Lasso^[21] 和 LAD- L_q ($0 < q < 1$) 等九种算法, 并比较了它们在三种指标 (变量选择中的零变量正确个数, 非零变量的不正确个数和测试集上的测试均方误差) 下 100 次试验的平均表现, 其中相应算法中所涉及的参数都按照对应文献所建议的执行. 计算与比较结果见表 1 和 2.

从表 1 和 2 可看出: 在具有重尾分布的噪声下, (i) 稳健正则化模型的解明显比最小二乘的解稳健, 测试误差较低; (ii) LAD- L_q ($0 < q < 1$) 模型在该组实验中的变量选择能力在大多数情况下都优于其它已知方法.

在实验 1 中, 我们比较了 LAD- L_q ($0 < q < 1$) 模型在低维数数据上的变量选择能力. 为了进一步研究该模型在高维数数据上的变量选择能力, 且由于试验 1 已经说明稳健正则化模型的解在稳健性方面明显优于最小二乘解, 以下在实验 2 中只比较稳健回归模型在重尾分布下的小样本高维数数据情形下的性态.

实验 2 考虑如下模型:

$$y = x^T \beta + \alpha \varepsilon,$$

其中

$$\beta = (\underbrace{3, 1.5, 0, 0, 2, 0, 0, \dots}_p), \quad \varepsilon \sim t(3).$$

训练样本与测试样本均由非标准正态分布 $N(0, \Sigma)$, 其中 $\Sigma_{i,j} = 0.5^{|i-j|}$ 产生. 训练样本 50 个, 测试样本 100 个, 而 $p = 10, 20, 50, 100$ 变化. 我们应用 LAD, LAD-Lasso, LAD-Adaptive Lasso 和 LAD- L_q ($0 < q < 1$) 等三种算法, 并比较它们在所述三种指标下 100 次试验的平均表现. 比较结果见表 3 和 4.

表 1 试验 1 比较结果 (样本间不相关情形)

α	方法	零系数正确个数	非零系数不正确个数	测试均方误差
0.5	OLS	-	-	0.9888
	Ridge	-	-	0.9887
	LAD	-	-	0.9577
	Lasso	1.65	0	0.9816
	Adaptive Lasso	2.96	0	0.9371
	RLAD	2.71	0	0.9564
	LAD-Lasso	3.97	0.03	0.9485
	LAD-Adaptive Lasso	3.48	0	0.9192
	LAD- L_q ($0 < q < 1$)	3.97	0	0.9191
1	OLS	-	-	3.1912
	Ridge	-	-	3.1853
	LAD	-	-	3.0084
	Lasso	1.86	0.09	3.1588
	Adaptive Lasso	2.88	0.19	3.0065
	RLAD	2.80	0.04	3.0658
	LAD-Lasso	3.68	0.22	2.9877
	LAD-Adaptive Lasso	3.32	0.11	2.9311
	LAD- L_q ($0 < q < 1$)	3.71	0.12	2.9330

表 2 实验 1 比较结果 (样本相关情形)

α	方法	零系数正确个数	非零系数不正确个数	测试均方误差
0.5	OLS	-	-	0.8586
	Ridge	-	-	0.8583
	LAD	-	-	0.8329
	Lasso	2.27	0	0.8351
	Adaptive Lasso	2.99	0	0.8080
	RLAD	2.88	0	0.8206
	LAD-Lasso	3.95	0.04	0.7989
	LAD-Adaptive Lasso	3.31	0	0.7883
	LAD- L_q ($0 < q < 1$)	3.96	0	0.7873
1	OLS	-	-	4.1136
	Ridge	-	-	4.1123
	LAD	-	-	4.0150
	Lasso	2.05	0.12	4.0638
	Adaptive Lasso	2.62	0.20	3.9749
	RLAD	2.88	0.01	3.9591
	LAD-Lasso	3.64	0.32	3.9264
	LAD-Adaptive Lasso	3.12	0.09	3.8766
	LAD- L_q ($0 < q < 1$)	3.66	0.14	3.8801

表 3 实验 2 比较结果 (参数 $\alpha = 1$)

p	方法	零系数正确个数	非零系数不正确个数	测试均方误差
10	LAD	-	-	1.3774
	LAD-Lasso	6.68	0	1.1235
	LAD-Adaptive Lasso	6.93	0	1.1147
	LAD- L_q ($0 < q < 1$)	6.94	0	1.0926
20	LAD	-	-	2.0080
	LAD-Lasso	15.87	0	1.1449
	LAD-Adaptive Lasso	18.58	0	1.1260
	LAD- L_q ($0 < q < 1$)	16.60	0	1.1047
50	LAD	-	-	8.7660
	LAD-Lasso	25.80	0.02	3.7163
	LAD-Adaptive Lasso	25.95	0.03	8.5037
	LAD- L_q ($0 < q < 1$)	46.01	0.01	1.6582
100	LAD	-	-	9.6330
	LAD-Lasso	92.58	0.01	1.2566
	LAD-Adaptive Lasso	95.72	0.01	1.2414
	LAD- L_q ($0 < q < 1$)	96.54	0	1.0895

表 4 实验 2 比较结果 (参数 $\alpha = 2$)

p	方法	零系数正确个数	非零系数不正确个数	测试均方误差
50	LAD	-	-	19.0200
	LAD-Lasso	16.40	0	18.4615
	LAD-Adaptive Lasso	25.00	0.05	18.0693
	LAD- L_q ($0 < q < 1$)	45.04	0.04	5.8146
100	LAD	-	-	16.0507
	LAD-Lasso	88.44	0.02	5.7928
	LAD-Adaptive Lasso	92.68	0.03	5.5179
	LAD- L_q ($0 < q < 1$)	95.32	0	4.7892

表 5 板材抗拉强度数据集比较结果

方法	测试均方误差	变量数	所选择变量
LAD	0.4977	24	全部
LAD-Lasso	0.5951	13	(2, 8, 9, 10, 11, 16, 17, 18, 19, 20, 22, 23, 24)
LAD-Adaptive Lasso	0.4683	12	(2, 3, 4, 8, 10, 11, 12, 14, 18, 20, 22, 23)
LAD- L_q ($0 < q < 1$)	0.4698	9	(2, 4, 8, 10, 11, 18, 20, 22, 23)

实验 3 (板材抗拉强度数据集) 抗拉强度 (MPa) 是衡量板材质量的一个重要指标. 在热连轧质量控制研究中, 提高板材的抗拉强度, 就能提高产品质量, 并增强产品在国际市场上的竞争力. 而热连轧生产是非常复杂的生产过程 (数据产生的过程中存在大量复杂噪声), 至今仍然没有完全了解影响其抗拉强度的主要控制因素. 本实验运用我们在 2009 年某钢铁企业所收集的 1549 毫米热连轧生产线生产中的板材抗拉强度数据集, 它主要由以下被认为影响抗拉强度的 24 个控制因素组成: 连铸坯的化学成分 (C、Mn、S、Si、Al 的含量)、5 个粗轧变形率、粗轧出口温度、7 个精轧变形率、精轧出口温度、卷取温度、穿带速度、加速度和成品厚度等工艺参数. 该数据集有 300 个训练样本, 90 个测试样本. 标准化处理该数据集, 使用稳健的回归方法 LAD, LAD-Lasso, LAD-Adaptive Lasso 和 LAD- L_q ($0 < q < 1$) 在该数据集上进行了测试. 测试结果见表 5.

由表 5 可看出: LAD-Adaptive Lasso 模型在该数据集上具有最好的预测误差, 而虽然 LAD- L_q ($0 < q < 1$) 模型预测误差稍低, 但其所选择的影响抗拉强度的主要控制变量个数最少. 这说明, 在该组实验中 LAD- L_q ($0 < q < 1$) 模型具有稳健的选择较少变量的能力. 我们已经把这一实验结论用于该钢厂的生产质量控制.

以上实验均验证了本文所提出的理论的正确性, 以及 LAD- L_q ($0 < q < 1$) 模型的广泛可用性与有效性. 然而, 值得注意的是, LAD- L_1 模型已有成熟的算法与应用软件方便使用. 相比较而言, LAD- L_q ($0 < q < 1$) 模型的求解更为复杂, 如何设计更高效的、更实用的优化解法是一个难点.

致谢 作者衷心的感谢审稿人提出的宝贵意见.

参考文献

- Portnoy S, Koenker R. The Gaussian hare and the Laplacian tortoise: computability of squared-error vs. absolute-error estimators, with discussion. *Statist Sci*, 1997, 12: 279-300

- 2 Charnes A, Cooper W, Ferguson O. Optimal estimation of executive compensation by linear programming. *Management Sciences*, 1955, 1: 138–151
- 3 Koenker R, D'orey V. Computing regression quantiles. *J Roy Statist Soc Ser C*, 1987, 36: 383–393
- 4 Bassett G, Koenker R. Asymptotic theory of least absolute error regression. *J Amer Statist Assoc*, 1978, 73: 618–622
- 5 Chen X R, Bai Z D, Zhao L C, et al. Asymptotic normality of minimum L_1 -norm estimates in linear models. *Sci China Ser A*, 1990, 33: 1311–1328
- 6 Pollard D. Asymptotics for least absolute deviations regression estimators. *Econometric Theory*, 1991, 7: 186–199
- 7 Bloomfield P, Steiger W L. *Least Absolute Deviation: Theory, Applications and Algorithms*. Boston: Birkhauser, 1983
- 8 Davis R A, Knight K, Liu J. M-estimation for autoregressions with infinite variance. *Stochastic Process Appl*, 1992, 40: 145–180
- 9 Knight K. Limiting distributions for L_1 regression estimators under general conditions. *Ann Statist*, 1998, 26: 755–770
- 10 Chen K, Ying Z, Zhang H, et al. Analysis of least absolute deviations. *Biometrika*, 2008, 95: 107–122
- 11 Giloni A, Padberg M. Alternative methods of linear regression. *Math Comput Modelling*, 2002, 35: 361–374
- 12 Giloni A, Padberg M. The finite sample breakdown point of L_1 -regression. *SIAM J Optim*, 2004, 14: 1028–1042
- 13 Peng L, Yao Q. Least absolute deviation estimation for ARCH and GARCH models. *Biometrika*, 2003, 90: 967–975
- 14 Hoerl A, Kennard R. Ridge regression. In: *Encyclopedia of Statistical Sciences*. New York: Wiley, 1998, 129–136
- 15 Tibshirani R J. Regression shrinkage and selection via the lasso. *J R Stat Soc Ser B Stat Methodol*, 1996, 58: 267–288
- 16 Fu W J. Penalized regressions: the bridge versus the lasso. *J Comput Graph Statist*, 1998, 7: 397–416
- 17 Knight K, Fu W. Asymptotics for lasso-type estimators. *Ann Statist*, 2000, 28: 1356–1378
- 18 Huang J, Horowitz J, Ma S. Asymptotic properties of bridge estimators in sparse high-dimensional regression models. *Ann Statist*, 2008, 36: 587–631
- 19 Li W, Michael D G, Ji Z. Regularized least absolute deviations regression and an efficient algorithm for parameter tuning. In: *Proceedings of the Six International Conference on Data Mining*. Washington: IEEE Computer Society, 2006, 690–700
- 20 Wang H, Li G, Jiang G. Robust regression shrinkage and consistent variable selection through the lad-lasso. *J Business Economic Statist*, 2007, 25: 347–355
- 21 Xu J F, Ying Z L. Simultaneous estimation and variable selection in median regression using lasso-type penalty. *Ann Inst Stat Math*, 2010, 62: 487–514
- 22 Xu Z B, Zhang H, Wang Y, et al. $L_{1/2}$ regularization. *Sci China Ser F*, 2010, 53: 1159–1169
- 23 Chartrand R, Staneva V. Restricted isometry properties and nonconvex compressive sensing. *Inverse Problems*, 2008, 24: 1–14
- 24 Xu Z B, Guo H L, Wang Y, et al. The representation of $L_{1/2}$ regularizer among L_q ($0 < q < 1$) regularizer: an experimental study based on phase diagram. Submitted to *Signal Processing*, 2009
- 25 Donoho D L, Tanner J. Observed universality of phase transitions in high-dimensional geometry, with applications in modern signal processing and data analysis. *Trans R Soc Lond Ser A Math Phys Eng Sci*, 2009, 367: 4273–4293
- 26 Geyer C J. On the asymptotics of convex stochastic optimization. Unpublished manuscript, 1996
- 27 Zhao P, Yu B. On model selection consistency of lasso. *J Mach Learn Res*, 2006, 8: 2701–2726
- 28 Zou H. The adaptive lasso and its oracle properties. *J Amer Statist Assoc*, 2006, 101: 1418–1429
- 29 Yuan M, Lin Y. On the nonnegative garotte estimator. *J R Stat Soc Ser B Stat Methodol*, 2007, 69: 143–161
- 30 Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. *J Amer Statist Assoc*, 2001, 96: 1348–1360
- 31 Candes E, Wakin M, Boyd S. Enhancing sparsity by reweighted L_1 minimization. *J Fourier Anal Appl*, 2008, 14: 877–905
- 32 Zou H, Li R Z. One-step sparse estimates in nonconcave penalized likelihood models. *Ann Statist*, 2008, 36: 1509–1533
- 33 Stamey T, Kabalin J, McNea I, et al. Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate II: radical prostatectomy treated patients. *J Urol*, 1989, 16: 1076–1083

Robust regularization theory based on L_q ($0 < q < 1$) regularization: the asymptotic distribution and variable selection consistency of solutions

CHANG XiangYu, XU ZongBen, ZHANG Hai, WANG JianJun & LIANG Yong

Abstract In this paper, we introduce the robust L_q ($0 < q < \infty$) regularization model, and then prove the global asymptotic distribution theorem for solutions of the model we propose. Applying the results, we can derive the model based on L_q ($0 < q < 1$) regularization satisfying the consistent property of variable selection; in other words, it has the capacity of variable selection. To solve this model, we develop an iterative weighted algorithm without extra parameters, and give the corresponding strategy of selecting regularization parameters. The experiment results reveal that the algorithm we introduce is available, efficient and widely valuable.

Keywords: robust regression, asymptotic distribution, variable selection, oracle property, LAD

MSC(2000): 62F12, 62F35, 62J07