



双侧截断回归模型的变量选择

郑明, 林婵娟, 郁文*

复旦大学管理学院统计与数据科学系, 上海 200433

E-mail: mingzheng@fudan.edu.cn, chanjuan.lin@roche.com, wenyu@fudan.edu.cn

收稿日期: 2021-02-25; 接受日期: 2021-09-26; 网络出版日期: 2022-03-28; * 通信作者

国家自然科学基金 (批准号: 11771095 和 12071088) 资助项目

摘要 在回归分析中, 当因变量存在双侧截断时, 已有的统计方法会使得回归模型的系数估计与变量选择产生偏差. 本文提出一种适用于双侧截断回归模型的系数估计与变量选择方法, 且该方法允许回归模型中自变量的个数随着样本量增大并趋于无穷而趋于无穷. 该方法的主要思想是, 提出一种 Mann-Whitney 型的损失函数来进行纠偏, 随后加入自适应最小绝对收缩和选择算子 (least absolute shrinkage and selection operator, LASSO) 惩罚项来进行变量选择. 本文同时设计一种迭代算法来实现损失函数的优化; 且证明了所提出估计量的相合性与渐近正态性, 还给出所提出变量选择方法的神谕性 (oracle property). 本文通过随机模拟展示所提出方法在有限样本量下的表现, 并使用所提出方法分析一个天文学领域的实际数据集.

关键词 双侧截断 变量选择 维数发散 自适应 LASSO 最小绝对离差 神谕性

MSC (2020) 主题分类 62J07, 62N99

1 引言

截断数据常见于天文学、经济学以及临床医学等领域的数据分析中. 当被观测的数据发生截断时, 只有那些取值落在某个对应区间内的数据可以被观测到, 同时那个区间也可以被记录, 而那些取值处于相应区间之外的数据不可能被观测到, 因此, 对于数据收集者而言, 这些数据好似根本不存在. 当该区间只有下限时, 相应的截断称为左截断; 当该区间只有上限时, 相应的截断称为右截断; 而当该区间同时存在上下限时, 相应的截断称为双侧截断.

如果不加以适当处理, 截断会给数据分析的结果带来偏差. 现有文献中已有不少处理单侧截断的数据分析方法. 具体地, 文献 [1] 提出了一种算法, 在数据存在分组、删失或者截断时可以找到其分布的非参数极大似然估计. 文献 [2] 对单侧截断数据提出了一种与文献 [1] 类似的分布估计. 之后, 文献 [3-5] 引进了计数过程来处理截断数据. 在回归分析中, 文献 [6] 拓展了 Mann-Whitney 方法来处理单侧截断的因变量, 文献 [7] 发展了迭代纠偏方法, 文献 [8] 给出了基于 Kendall 相关系数的相关分析,

英文引用格式: Zheng M, Lin C J, Yu W. Variable selection in doubly truncated regression models (in Chinese). Sci Sin Math, 2022, 52: 1433–1448, doi: 10.1360/SCM-2021-0158

文献 [9] 提出了秩估计方法, 等等. 一些更近期的成果包括文献 [10] 在计量经济框架下的研究以及文献 [11,12] 在一类有偏抽样框架下提出的截断数据分析方法.

相比于单侧截断, 双侧截断在技术上更难处理. 就估计与检验问题而言, 文献 [13,14] 提出了数据存在双侧截断时其分布的极大似然估计, 文献 [15] 构造了密度函数的核估计, 文献 [16,17] 拓展了两样本的 Mann-Whitney 检验. 在回归分析中, 当因变量存在双侧截断时, 文献 [18] 提出了适用于低维自变量的核估计方法, 文献 [19] 考虑了基于一类半参数线性变换模型的估计方法. 最近, 文献 [20] 提出了一种拓展的 Mann-Whitney 型损失函数来估计固定维数线性模型中的回归系数, 该损失函数的最小值点被定义为回归系数的估计量.

双侧截断的存在不仅影响回归模型的估计, 也会给模型的变量选择带来困难. 据我们所知, 目前还没有文献研究过因变量存在双侧截断时线性回归模型的变量选择问题. 本文在文献 [20] 提出的 Mann-Whitney 型损失函数的基础上, 利用自适应最小绝对收缩和选择算子 (least absolute shrinkage and selection operator, LASSO) [21] 惩罚给出了一种可以同时完成回归模型系数估计与变量选择的方法. 该方法不仅适用于固定维数的回归模型, 也可以用于自变量维数随样本量增大并趋于无穷而趋于无穷的情形. 我们发现基于自适应 LASSO 的变量选择过程具有神谕性. 在计算层面, 本文设计了一种迭代算法来获取目标损失函数的最小值点. 在该算法的每一步迭代中, 只需完成一个最小绝对离差 (least absolute deviation, LAD) 最优化运算, 而 LAD 最优化是更容易高效实现的. 我们采用修正 Bayes 信息准则 (Bayesian information criterion, BIC) 方法来选择损失函数中的调节参数, 并设计一种基于随机加权技术的方法来估计非零系数估计量的标准误. 通过随机模拟与实际数据分析, 展示所提出方法在有限样本量下的表现.

本文余下内容的结构如下. 第 2 节给出一些本文所需的记号, 并介绍双侧截断回归模型的估计方法. 第 3 节详细描述所提出的变量选择方法, 包括基本思想、迭代算法、变量选择的神谕性以及调节参数的选择方法. 第 4 节给出一些随机模拟的结果, 以展示所提出方法在有限样本量下的表现, 同时也对一个实际数据集进行分析. 第 5 节总结全文. 所有的技术细节与证明过程均在附录之中.

2 记号、模型与估计

首先引入一些记号. 记 $\tilde{Y}$ 为因变量, $\tilde{X}$ 为 p 维自变量. 考虑如下线性模型:

$$\tilde{Y} = \beta^T \tilde{X} + \tilde{\varepsilon}, \quad (2.1)$$

其中, β 为 p 维回归系数向量, $\tilde{\varepsilon}$ 为独立于 $\tilde{X}$ 的随机误差. 因变量 $\tilde{Y}$ 存在双侧截断, 记 $\tilde{L}$ 与 $\tilde{R}$ 分别为左截断变量与右截断变量, 只有当 $\tilde{Y}$ 落于 $(\tilde{L}, \tilde{R})$ 时, $(\tilde{Y}, \tilde{L}, \tilde{R}, \tilde{X})$ 才能被观测到. 假设给定 $\tilde{X}$ 和 $\tilde{Y}$ 与 $(\tilde{L}, \tilde{R})$ 条件独立, 这是截断数据分析中常见的假设. 在模型 (2.1) 下, 该假设等价于假设 $\tilde{\varepsilon}$ 独立于 $(\tilde{L}, \tilde{R})$. 将 $\tilde{\varepsilon}$ 的分布函数与密度函数分别记为 F 与 f .

假设 $(\tilde{Y}_i, \tilde{L}_i, \tilde{R}_i, \tilde{X}_i)$ ($i = 1, \dots, \tilde{n}$) 为 $(\tilde{Y}, \tilde{L}, \tilde{R}, \tilde{X})$ 的 $\tilde{n}$ 个独立同分布复制. 如前所述, 只有当 $\tilde{L}_i < \tilde{Y}_i < \tilde{R}_i$ 时, $\tilde{Y}_i$ 以及 $(\tilde{L}_i, \tilde{R}_i, \tilde{X}_i)$ 才能被观测到. 将可观测到的数据的样本量记为 n , 即 $n = \#\{i : \tilde{L}_i < \tilde{Y}_i < \tilde{R}_i\}$; 将可观测到的数据记为 (Y_i, L_i, R_i, X_i) , $i = 1, \dots, n$. 此外, 对于每个 $i = 1, \dots, n$, 定义 $e_i(\beta) = Y_i - \beta^T X_i$, $L_i(\beta) = L_i - \beta^T X_i$ 以及 $R_i(\beta) = R_i - \beta^T X_i$.

为了纠正双侧截断给系数估计带来的偏差, 文献 [20] 提出了一种修正的 Mann-Whitney 型秩估计

方程

$$\sum_{i=1}^n \sum_{j=1}^n I\{L_j(\beta) < e_i(\beta) < R_j(\beta), L_i(\beta) < e_j(\beta) < R_i(\beta)\} (X_i - X_j) \operatorname{sgn}\{e_i(\beta) - e_j(\beta)\} = 0, \quad (2.2)$$

其中, $I\{\cdot\}$ 为示性函数, $\operatorname{sgn}\{\cdot\}$ 为符号函数. 不难看出, 估计函数是将样本中所有的可比较对进行求和, 因此该函数在 β 取真实值时期望为 0. 同时, 求解 (2.2) 等价于最小化损失函数

$$l_n(\beta) = \sum_{i=1}^n \sum_{j=1}^n |[(e_i(\beta) - e_j(\beta)) \wedge (R_j - Y_j) \wedge (Y_i - L_i)] \vee (L_j - Y_j) \vee (Y_i - R_i)|, \quad (2.3)$$

其中 $\wedge$ 与 $\vee$ 分别表示取小与取大. 将 $l_n(\beta)$ 的最小值点记为 $\hat{\beta}$, 取 $\hat{\beta}$ 作为回归系数 β 的估计量, 在一定的正则性条件下, 文献 [20] 证明了该估计量的相合性与渐近正态性.

3 变量选择

3.1 方法

本小节考虑用观测到的样本数据对模型 (2.1) 进行变量选择, 在损失函数 (2.3) 的基础上引入惩罚项. 注意到文献 [20] 考虑的是固定维数的自变量, 这里允许 X_i 的维数 p 可以随着样本量 n 趋于无穷而趋于无穷, 也就是说 p 可以随着 n 的变化而变化, 即 $p = p_n$. 因此, 自变量的函数都会依赖于 n . 同时, 还假设当 $\tilde{n}$ 趋于无穷时, $n/\tilde{n}$ 趋于一个小于 1 的常数.

由于损失函数 (2.3) 为 L_1 类型, 自然会想到引入 LASSO 类型的惩罚项. 具体地, 若在损失函数中引入通常的 LASSO 惩罚项, 可以得到损失函数 $l_n(\beta) + \lambda_n \sum_{j=1}^{p_n} |\beta_j|$, 其中 λ_n 是一个 (数据驱动的) 调节参数. 不过, 使用通常的 LASSO 惩罚项会带来两个问题.

第一个方面是计算层面的. 由于 $l_n(\beta)$ 关于 β 不是一个凸函数, 因此相应的优化问题也失去了凸性. 不过, 注意到 $l_n(\beta)$ 可以被等价地表达成

$$l_n(\beta) = \sum_{i=1}^n \sum_{j=1}^n I\{L_j(\beta) < e_i(\beta) < R_j(\beta), L_i(\beta) < e_j(\beta) < R_i(\beta)\} |e_i(\beta) - e_j(\beta)|.$$

上述损失函数的非凸性主要来源于示性函数, 而 $|e_i(\beta) - e_j(\beta)|$ 关于 β 是凸的. 因此, 可以考虑设计一种迭代算法, 在每一步迭代中, 将示性函数中的系数值确定, 对损失函数里剩余的部分进行优化. 关于计算的进一步细节在第 3.2 小节以及附录中进行展示.

第二个方面是理论层面的. 众所周知 LASSO 估计量是有偏的, 在某些情形下, 基于 LASSO 的变量选择并不相合. 为了克服这些问题, 文献 [21] 在 LASSO 的惩罚项中对不同的系数引入不同的权重, 由此提出了自适应 LASSO 惩罚. 文献 [21] 使用基于 β 收敛速度为 $n^{-1/2}$ 的相合估计构造的函数作为权重, 并证明了如果选用合适的调节参数 λ_n , 自适应 LASSO 对于固定维数的 β 进行选择时具有神谕性.

为了保持变量选择方法的神谕性, 这里也采用自适应 LASSO. 具体地, 我们提出如下损失函数:

$$l_n^{\text{AL}}(\beta) = l_n(\beta) + \lambda_n \sum_{j=1}^{p_n} \hat{w}_j |\beta_j|, \quad (3.1)$$

其中 $\hat{w}_j$ 为 β_j 的权重. 文献 [21] 选择 $\hat{w}_j = 1/|\tilde{\beta}_j|^\gamma$, 其中, $\tilde{\beta}_j$ 为 β_j 的收敛速度为 $n^{-1/2}$ 的相合估计, γ 是一个事先给定的正数. 在我们讨论的问题中, 即便 p_n 会随着 n 趋于无穷而趋于无穷, 但是只要 p_n 发散的速度不要太快, 我们仍然可以得到 $l_n(\beta)$ 的最小值点 $\hat{\beta}$. 基于 $\hat{\beta}$, 在 (3.1) 中定义 $\hat{w}_j = 1/|\hat{\beta}_j|^\gamma$, 其中 $\hat{\beta}_j$ 代表 $\hat{\beta}$ 的第 j 个元素. 我们将 $l_n^{\text{AL}}(\beta)$ 的最小值点作为推荐的估计量.

3.2 算法

第 3.1 小节中曾提到, $l_n(\beta)$ 关于 β 并非凸函数, $l_n^{\text{AL}}(\beta)$ 也是非凸的. 因此, 考虑用一种迭代算法来得到所推荐的估计量. 具体地, 将 $l_n^{\text{AL}}(\beta)$ 略作修改, 考虑

$$l_n^{\text{AL}}(\beta, b) = \sum_{i=1}^n \sum_{j=1}^n I\{L_j(b) < e_i(b) < R_j(b), L_i(b) < e_j(b) < R_i(b)\} |e_i(\beta) - e_j(\beta)| + \lambda_n \sum_{j=1}^{p_n} \hat{w}_j |\beta_j|.$$

寻找一个初始估计, 记为 $\hat{\beta}^{(0)}$ (例如, 可以将 $\hat{\beta}$ 作为初始估计). 随后, 在第 k 步迭代中, 定义

$$\hat{\beta}^{(k)} = \operatorname{argmin}_{\beta} l_n^{\text{AL}}(\beta, \hat{\beta}^{(k-1)}), \quad k \geq 1.$$

当 $k \rightarrow \infty$ 时, 如果 $\hat{\beta}^{(k)}$ 收敛于一个极限, 则将该极限作为最后的估计.

使用迭代算法的主要原因在于, 在第 k 步中, 目标函数可以被表达成

$$l_n^{\text{AL}}(\beta, \hat{\beta}^{(k-1)}) = \sum_{i,j \in \mathcal{S}_k} |e_i(\beta) - e_j(\beta)| + \lambda_n \sum_{j=1}^{p_n} \hat{w}_j |\beta_j|, \quad (3.2)$$

其中

$$\mathcal{S}_k = \{(i, j) \mid L_j(\hat{\beta}^{(k-1)}) < e_i(\hat{\beta}^{(k-1)}) < R_j(\hat{\beta}^{(k-1)}), L_i(\hat{\beta}^{(k-1)}) < e_j(\hat{\beta}^{(k-1)}) < R_i(\hat{\beta}^{(k-1)})\}.$$

由此可见, 这个最优化问题可以变成一个线性规划问题, 从而能够有效地求解. 此外, 我们还发现经过一些代数运算后, 第 k 步迭代中的这个最优化问题能够被表达成一个 LAD 问题, 可以使用一些标准的 LAD 算法进行求解, 如 R 中的 `quantreg` 包. 附录 A.1 提供了一些计算的细节.

3.3 大样本性质

接下来, 在一定的正则性条件下给出所提出自适应 LASSO 估计量的神谕性. 一般而言, 神谕性的建立需要 $\hat{\beta}$ 有相合性. 尽管文献 [20] 在一系列的条件下给出了 $\hat{\beta}$ 的相合性, 但该文献的结果仅限于固定维数的情形. 当 p_n 随着 n 发散时, $\hat{\beta}$ 在什么样的条件下有相合性, 收敛的速度是多少等问题, 均需要进一步的探讨. 将 β 的真实取值记为 β_0 . 下面的命题给出了维数发散时 $\hat{\beta}$ 的相合性结果, 该命题的证明在附录 A.2 中给出.

命题 3.1 假设附录 A.2 中的条件 A.1–A.6 均成立, 且 $p_n \log n / \sqrt{n} = o(1)$, 则

$$\|\hat{\beta} - \beta_0\| = O_p\left(\left(\frac{p_n}{n}\right)^{1/2}\right).$$

接下的定理给出自适应 LASSO 估计量的存在性, 并且证明其相合, 收敛速度为 $O_p((p_n/n)^{1/2})$. 该定理的证明在附录 A.2 中给出.

定理 3.1 假设附录 A.2 中的条件 A.1–A.6 均成立, 且 $\sqrt{n}\lambda_n = o(1)$, $p_n \log n / \sqrt{n} = o(1)$, 则 $l_n^{\text{AL}}(\beta)$ 存在一个局部最小值点的概率趋近于 1, 记该最小值点为 $\hat{\beta}^{\text{AL}}$, 该最小值点满足

$$\|\hat{\beta}^{\text{AL}} - \beta_0\| = O_p\left(\left(\frac{p_n}{n}\right)^{1/2}\right).$$

为了表述神谕性, 我们再引进一些记号. 记

$$\begin{aligned}\xi_{ij}(\beta) &= I\{L_j(\beta) < e_i(\beta) < R_j(\beta), L_i(\beta) < e_j(\beta) < R_i(\beta)\}(X_i - X_j)\text{sgn}\{e_i(\beta) - e_j(\beta)\}, \\ h_{ij}(\beta) &= |[(e_i(\beta) - e_j(\beta)) \wedge (R_j - Y_j) \wedge (Y_i - L_i)] \vee (L_j - Y_j) \vee (Y_i - R_i)|\end{aligned}$$

以及 $\bar{l}(\beta) = \mathbf{E}(h_{ij}(\beta))$, 注意到这里的期望是根据观测数据的分布来取的. 定义两个 $p_n \times p_n$ 的矩阵 $V = \mathbf{E}(\xi_{ij}(\beta_0)\xi_{ik}^T(\beta_0))$ 与 $A = \partial^2 \bar{l}(\beta) / \partial \beta \partial \beta^T|_{\beta=\beta_0}$. 将 β_0 分成 $\beta_{0,1}$ 与 $\beta_{0,2}$ 两部分, 其中, $\beta_{0,1}$ 表示 β_0 中的非零部分, $\beta_{0,2}$ 代表零的部分. 将 $\beta_{0,1}$ 的维数记为 $p_{n,1}$. 以 $\hat{\beta}_1^{\text{AL}}$ 与 $\hat{\beta}_2^{\text{AL}}$ 分别表示 $\beta_{0,1}$ 与 $\beta_{0,2}$ 的自适应 LASSO 估计量, 以 V_{11} 代表 V 中对应于 $\beta_{0,1}$ 的 $p_{n,1} \times p_{n,1}$ 维部分, 以 A_{11} 代表 A 中对应于 $\beta_{0,1}$ 的 $p_{n,1} \times p_{n,1}$ 维部分.

定理 3.2 假设附录 A.2 中的条件 A.1–A.6 均成立, 且 $\sqrt{n}\lambda_n = o(1)$, $p_n \log n / \sqrt{n} = o(1)$, 则 $\hat{\beta}^{\text{AL}}$ 有如下两条性质:

- (i) $\hat{\beta}_2^{\text{AL}} = 0$ 的概率趋近于 1;
- (ii) 对任意非零 $p_{n,1} \times 1$ 维常数向量 u_n (满足 $\|u_n\| = 1$), $\sqrt{n}u_n^T(\hat{\beta}_1^{\text{AL}} - \beta_{0,1}) / [2(u_n^T A_{11}^{-1} V_{11} Z_{11}^{-1} u_n)^{-1/2}]$ 的分布函数弱收敛于标准正态分布.

定理 3.2 的证明见附录 A.2. 由定理 3.2 可知, $\beta_{0,1}$ 每一维元素的估计量都是渐近正态的, 这个结果可用来对单个非零回归系数进行统计推断, 我们将在第 3.4 小节中讨论标准误的估计.

3.4 调节参数选择与标准误估计

调节参数比较典型的选择方法包括信息准则方法 (如 AIC (Akaike information criterion) 和 BIC) 与数据驱动方法 (如 K-折交叉验证). 考虑到计算负担, 我们采用信息准则方法. 众所周知, 在固定维数的情形下, BIC 方法通常具有变量选择的相合性. 然而, 当自变量维数发散时, 标准的 BIC 方法不能保证选择的相合性. 因此, 考虑如下的修正 BIC 准则:

$$\text{BIC}(\lambda) = l_n(\hat{\beta}_\lambda^{\text{AL}}) + \frac{\log n}{n} C_n \text{df}_\lambda,$$

其中, $\hat{\beta}_\lambda^{\text{AL}}$ 表示当调节参数取值为 λ 时的自适应 LASSO 估计量; C_n 为一列发散常数, 满足 $(\log n/n)C_n \rightarrow 0$, df_λ 表示当调节参数取值为 λ 时估计中非零估计量的维数. 在本文的数值计算中, 我们发现 $C_n = \log \log p_n$ 时有较好的表现. 将真实的模型与基于观测数据选择而得的模型分别记为 $\mathcal{M}^*$ 与 $\mathcal{M}_\lambda$. 下面的定理给出了修正 BIC 准则的选择相合性. 定理的证明在附录 A.2 中给出.

定理 3.3 假设附录 A.2 中的条件 A.1–A.6 均成立, 且 $\sqrt{n}\lambda_n = o(1)$, $\lambda_n n / p_n \rightarrow \infty$, $p_n \log n / \sqrt{n} = o(1)$, $\sqrt{n/\{C_n p_n \log n\}} \cdot \liminf_{n \rightarrow \infty} \{\min_{j \in \mathcal{M}^*} |\beta_{0,j}|\} \rightarrow \infty$, 则 $\mathbf{P}(\mathcal{M}_\lambda = \mathcal{M}^*) \rightarrow 1$, $n \rightarrow \infty$.

最后, 需要讨论 β_0 非零部分估计量的标准误估计问题. 注意到 $\hat{\beta}_1^{\text{AL}}$ 渐近分布的协方差矩阵与 ε 的密度函数 f 有关, 而直接估计 f 并不容易, 因此采用了一种随机加权的方法来估计标准误. 具体地, 定义一种置换惩罚估计 $\hat{\beta}^*$, 其为损失函数 $l_n^*(\beta) = \sum_{i=1}^n \sum_{j=1}^n (W_i + W_j) h_{ij}(\beta) + \hat{\lambda} \sum_{j=1}^{p_n} \hat{w}_j |\beta_j|$ 的最小值点, 其中, W_i ($i = 1, \dots, n$) 为独立同分布的非负有界随机变量, 且其期望为 0.5, 方差为 1, 与观测数

据相互独立. $l_n^*(\beta)$ 可被视为 $l_n^{\Delta^L}(\beta)$ 的一种置换版本. 记 $\hat{\beta}_1^*$ 为 $\hat{\beta}^*$ 中对应于 $\hat{\beta}_1^{\Delta^L}$ 的部分 (即 $\hat{\beta}^{\Delta^L}$ 的非零部分), 可以证明, 在一定条件下, 给定观测数据, $\hat{\beta}_1^*$ 的条件渐近分布与 $\hat{\beta}_1^{\Delta^L}$ 的渐近分布相同. 因此, 如果重复地生成随机变量序列 W_i ($i = 1, \dots, n$), 我们可以获得大量的 $\hat{\beta}_1^*$ 的实现值, 然后用这些实现值的样本标准差来近似 $\hat{\beta}_1^*$ 的标准误. 该结论的理论证明与文献 [20, 定理 2] 的证明类似.

4 数值结果

4.1 随机模拟

我们做了一些随机模拟来展示所提出方法在有限样本下的表现. 在模拟中, 取 $p_n = [7\tilde{n}^{1/5}]$, 这里 $\tilde{n}$ 如第 2 节中所定义, 代表全部数据的样本量. 回归系数中非零部分的维数 $p_{n,1}$ 取为 $[p_n/3]$. 分别考虑 $\tilde{n} = 300$ 与 $\tilde{n} = 500$ 两种情形. 在第一种情形中, $p_n = 21$, $p_{n,1} = 7$. 取 $\beta_{1,0} = (3.12, 2.20, -0.86, 0.92, -2.49, 1.95, -1.32)^T$. 对于自变量 $\tilde{X} = (\tilde{X}_1, \dots, \tilde{X}_{21})$, $\tilde{X}_1$ 服从成功概率为 0.25 的二项分布, $\tilde{X}_2$ 服从成功概率为 0.8 的二项分布, $\tilde{X}_8$ 服从 0 至 2 上的均匀分布, $\tilde{X}_9$ 服从成功概率为 0.5 的二项分布, $\tilde{X}_{10}$ 服从 -2 至 0 上的均匀分布, 这些自变量均相互独立. 其余的所有自变量均服从标准正态分布, 并且 $\tilde{X}_i$ 与 $\tilde{X}_j$ 之间的相关系数等于 $0.3^{|i-j|}$. 对于线性模型里的误差分布, 考虑两种选择, 分别为标准正态分布 ($N(0, 1)$) 与位置参数为 0、尺度参数为 1 的极小值分布 ($EV(0.1)$). 左截断变量 $\tilde{L}$ 服从 a 到 $b^T \tilde{X}$ 上的均匀分布, 右截断变量 $\tilde{R} = \tilde{L} + c$, 其中 a 、 b 与 c 为选取的常数, 使得截断比例能达到 30% 与 40%. 在这样的随机数生成机制下, 左右两侧截断的比例几乎相等.

在每种情形下, 随机数都被重复生成了 1,000 次. 我们定义了若干准则来考察变量选择方法的表现. 第一个是模型误差 (model error, ME), 其定义为 $ME = (\tilde{\beta} - \beta_0)^T E(X_i X_i^T) (\tilde{\beta} - \beta_0)$, 其中 $\tilde{\beta}$ 表示由某一种估计方法得到的对于 β_0 的估计量. 我们记录了 1,000 次数据生成的 ME 的中位数和中位绝对离差 (median absolute deviation, MAD) 来比较不同方法所给出的预测表现. 第二个是被估计为零系数的平均个数 (total number, TN), 即在 1,000 次重复中, 使用各种方法后被估计为零系数的个数平均值, 包括被正确估计为零系数的平均个数 (correct number, CN) 以及被错误估计为零系数的平均个数 (incorrect number, IN). 对于 CN 而言, 哪种方法使其更接近 $p_n - p_{n,1}$, 则哪种方法的表现更优; 对于 IN 而言, 哪种方法使其更接近 0, 则哪种方法的表现更优. 第三个是正确识别模型比率 (ratio of correct model, RCM), 即 1,000 次重复中各种方法能够完全选出正确模型的比例. 哪种方法使其更接近 100%, 则哪种方法的表现更优. 考虑 3 种变量选择方法, 包括本文所提出的选择方法 (推荐方法)、忽略截断的基于自适应 LASSO 惩罚的变量选择方法 (简稚方法) 以及事先给定真实非零子集模型估计 (真实子集). 模拟结果列于表 1 与 2.

从表 1 与 2 中可以看出, 基于 ME 准则, 本文所提出的方法表现要明显优于简稚方法. 这是可以预料的, 因为简稚方法忽略了截断, 对于回归系数的估计是有偏的. 基于 TN 与 RCM 准则, 推荐方法仍然比简稚方法有更好的表现, 不过超越的程度不是很大. TN 与 RCM 准则是衡量变量选择方法的准则, 尽管在我们的模拟机制下, 重要变量的选择受双侧截断的影响并不很明显, 但是如果可以纠正双侧截断所带来的偏差, 仍然可以改进变量选择的准确度.

4.2 实例分析

我们分析一个来自斯隆数字化天文调查 (Sloan Digital Sky Survey, SDSS) 团队编制的 46,420 颗类星体目录中的数据. 对于这个数据集, 主要的分析目标是使用一些相关的特征来预测类星体的

表 1 误差分布服从 $N(0, 1)$ 时变量选择方法的随机模拟结果

$\tilde{n}$	截断比例	方法	ME		TN		RCM (%)
			中位数	MAD	CN	IN	
300	0.3	推荐方法	0.220	0.137	13.921	0.000	92.836
		简稚方法	0.487	0.242	13.908	0.000	91.364
		真实子集	0.109	0.076	14.000	0.000	100.000
	0.4	推荐方法	0.385	0.232	13.862	0.000	88.679
		简稚方法	1.048	0.423	13.821	0.000	84.608
		真实子集	0.158	0.112	14.000	0.000	100.000
500	0.3	推荐方法	0.133	0.073	15.995	0.000	99.503
		简稚方法	0.289	0.135	15.966	0.000	96.723
		真实子集	0.062	0.043	16.000	0.000	100.000
	0.4	推荐方法	0.217	0.120	15.977	0.000	98.014
		简稚方法	0.611	0.229	15.938	0.000	94.048
		真实子集	0.084	0.057	16.000	0.000	100.000

表 2 误差分布服从 $EV(0, 1)$ 时变量选择方法的随机模拟结果

$\tilde{n}$	截断比例	方法	ME		TN		RCM (%)
			中位数	MAD	CN	IN	
300	0.3	推荐方法	0.226	0.162	13.85	0.00	86.905
		简稚方法	0.780	0.353	13.82	0.00	83.730
		真实子集	0.129	0.098	14.00	0.00	100.000
	0.4	推荐方法	0.380	0.293	13.81	0.00	83.765
		简稚方法	1.232	0.565	13.72	0.00	75.896
		真实子集	0.184	0.132	14.00	0.00	100.000
500	0.3	推荐方法	0.118	0.087	15.98	0.00	98.214
		简稚方法	0.489	0.208	15.96	0.00	95.933
		真实子集	0.075	0.048	16.00	0.00	100.000
	0.4	推荐方法	0.197	0.146	15.90	0.00	91.144
		简稚方法	0.989	0.369	15.88	0.00	89.154
		真实子集	0.107	0.072	16.00	0.00	100.000

红移. 考虑使用线性模型进行预测, 因变量取为类星体的红移. 较之于 SDSS 的完整目录, 这里分析的数据集在去掉了一些技术性变量后包含 23 个自变量. 关于该数据集更详细的信息可见 <https://astrostatistics.psu.edu/datasets>.

直观上并不容易意识到为何红移数据会存在双侧截断, 这主要是由于天文望远镜的观测技术限制. 由于天文望远镜的光圈, 类星体的红移观测存在一个上界, 对于此处分析的数据, 红移数值的最大上界, 也就是右截断变量取值为 5.4135, 任何大于该数值的红移数据会被右截断. 另一方面, 对于取值小的红移数据, 相应的测量并不可靠, 当红移测量值低于其估计误差时, 该测量值会被去除, 故估计误差会成为红移测量值的左截断变量. 因此, 类星体的红移测量值存在双侧截断.

在自变量层面, 23 个自变量中, 有些是 SDSS 团队根据其观测条件、探测背景以及其他技术条件所得到的关于亮度的测量误差, 这些自变量未被纳入在模型中, 我们使用了 13 个自变量 (见表 3) 来拟合模型. 考虑到计算能力的限制, 我们没有采用全部的数据来拟合, 而是抽取了一个容量为 929 的随机样本来展示所提出的方法. 所抽取的样本按照 7:3 的比例被随机分为训练集与测试集. 我们在训练集上使用所提出的变量选择方法选出了“重要”的自变量, 随后用选择后的模型在测试集上进行预测, 结果列于表 4.

基于我们所提出的变量选择方法, 得到了 5 个“重要”自变量, 对应的回归系数估计不为 0. 我们使用这 5 个自变量来对测试数据进行预测, 将预测结果展示在图 1 中. 在图中, 我们使用连线来比对

表 3 SDSS 数据中用于拟合线性的 13 个自变量的涵义描述

自变量名称	自变量涵义
R.A.	赤经
Dec.	赤纬
u_mag	紫外波段的亮度
g_mag	绿色波段的亮度
r_mag	红色波段的亮度
i_mag	深红波段的亮度
z_mag	极红波段的亮度
Radio	比值波段的亮度
X-ray	X 射线波段的亮度
J_mag	近红外 J 波段的亮度
H_mag	近红外 H 波段的亮度
K_mag	近红外 K 波段的亮度
M.i	i 波段的绝对亮度等级

表 4 SDSS 数据的变量选择与估计结果

	自变量	系数估计值	标准误估计值
1	R.A.	0	—
2	Dec.	0	—
3	u_mag	0.340	0.024
4	g_mag	−0.320	0.046
5	r_mag	0	—
6	i_mag	0.500	0.043
7	z_mag	0	—
8	Radio	0	—
9	X.ray	0	—
10	J_mag	0.017	0.007
11	H_mag	0	—
12	K_mag	0	—
13	M.i	−0.803	0.017

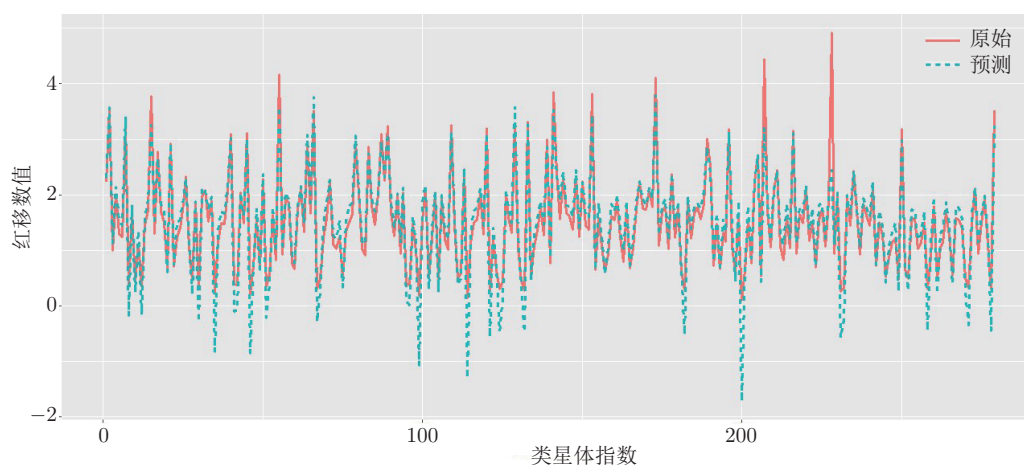


图 1 (网络版彩图) SDSS 数据测试集中红移数值的真实值与预测值

因变量的真实值与预测值, 可以发现预测值 (虚线表示) 与真实值 (实线表示) 较为接近, 表明所选的自变量给出了较好的预测表现.

5 总结

本文将 Mann-Whitney 型损失函数与自适应 LASSO 惩罚组合在一起, 设计了一种可以对双侧截断回归模型同时进行变量选择与系数估计的方法, 该方法的理论允许自变量维数随着样本量趋于无穷而趋于无穷. 本文设计了一种迭代算法来克服待优化目标函数的非凸性, 在每一步的迭代中, 目标函数的最优化可以转化成一个 LAD 问题. 本文证明了所提出选择方法的神谕性, 并给出了选择调节参数的修正 BIC 方法以及标准误估计的随机加权方法. 随机模拟表明所提出方法有较为合理的优先样本表现.

本文的主要贡献之一是, 在理论上允许自变量维数随着样本量趋于无穷而趋于无穷, 不过由于本文采用的是基于惩罚的变量选择方法, 故自变量维数趋于无穷的速度不能太快. 对于所谓超高维情形, 需要采用基于筛选的方法来进行变量选择, 这是未来值得进一步研究的方向之一.

致谢 作者感谢两位评审专家对论文初稿提出的建议, 依据这些建议所做的修改显著地提升了论文的质量.

参考文献

- 1 Turnbull B W. The empirical distribution function with arbitrarily grouped, censored and truncated data. *J R Stat Soc Ser B Stat Methodol*, 1976, 38: 290–295
- 2 Lynden-Bell D. A method of allowing for known observational selection in small samples applied to 3CR quasars. *Mon Not R Astron Soc*, 1971, 155: 95–118
- 3 Wang M C, Jewell N P, Tsai W Y. Asymptotic properties of the product limit estimate under random truncation. *Ann Statist*, 1986, 14: 1597–1605
- 4 Keiding N, Gill R D. Random truncation models and Markov processes. *Ann Statist*, 1990, 18: 582–602
- 5 Lai T L, Ying Z. Estimating a distribution function with truncated and censored data. *Ann Statist*, 1991, 19: 417–442
- 6 Bhattacharya P K, Chernoff H, Yang S S. Nonparametric estimation of the slope of a truncated regression. *Ann Statist*, 1983, 11: 505–514
- 7 Tsui K-L, Jewell N P, Wu C F J. A nonparametric approach to the truncated regression problem. *J Amer Statist Assoc*, 1988, 83: 785–792

- 8 Tsai W-Y. Testing the independence of truncation time and failure time. *Biometrika*, 1990, 77: 169–177
- 9 Lai T L, Ying Z. Rank regression methods for left-truncated and right-censored data. *Ann Statist*, 1991, 19: 531–556
- 10 Greene W H. *Econometric Analysis*, 7th ed. Upper Saddle River: Prentice Hall, 2012
- 11 Kim J P, Lu W, Sit T, et al. A unified approach to semiparametric transformation models under general biased sampling schemes. *J Amer Statist Assoc*, 2013, 108: 217–227
- 12 Liu H, Ning J, Qin J, et al. Semiparametric maximum likelihood inference for truncated or biased-sampling data. *Statist Sinica*, 2016, 26: 1087–1115
- 13 Efron B, Petrosian V. Nonparametric methods for doubly truncated data. *J Amer Statist Assoc*, 1999, 94: 824–834
- 14 Shen P-S. Nonparametric analysis of doubly truncated data. *Ann Inst Statist Math*, 2010, 62: 835–853
- 15 Moreira C, de Uña-Álvarez J. Kernel density estimation with doubly truncated data. *Electron J Stat*, 2012, 6: 501–521
- 16 Bilker W, Wang M-C. Generalized Wilcoxon statistics in semiparametric truncation models. *Biometrics*, 1996, 52: 10–20
- 17 Shen P-S. A class of rank-based tests for doubly-truncated data. *TEST*, 2013, 22: 83–102
- 18 Moreira C, de Uña-Álvarez J, Meira-Machado L. Nonparametric regression with doubly truncated data. *Comput Statist Data Anal*, 2016, 93: 294–307
- 19 Shen P-S. Regression analysis of interval censored and doubly truncated data with linear transformation models. *Comput Statist*, 2013, 28: 581–596
- 20 Ying Z, Yu W, Zhao Z, et al. Regression analysis of doubly truncated data. *J Amer Statist Assoc*, 2020, 115: 810–821
- 21 Zou H. The adaptive Lasso and its oracle properties. *J Amer Statist Assoc*, 2006, 101: 1418–1429
- 22 Wang Z, Lin Y, Liu W, et al. U-processes with increasing dimensional parameters. Working paper, 2019
- 23 Fan J, Peng H. Nonconcave penalized likelihood with a diverging number of parameters. *Ann Statist*, 2004, 32: 928–961
- 24 Cai J, Fan J, Li R, et al. Variable selection for multivariate failure time data. *Biometrika*, 2005, 92: 303–316
- 25 Cho H, Qu A. Model selection for correlated data with diverging number of parameters. *Statist Sinica*, 2013, 23: 901–927
- 26 Ni A, Cai J, Zeng D. Variable selection for case-cohort studies with failure time outcome. *Biometrika*, 2016, 103: 547–562
- 27 Pollard D. *Empirical Processes: Theory and Applications*. Hayward: Institute of Mathematical Statistics, 1990

附录 A

附录 A.1 算法细节

这里给出第 3.2 小节中所提出算法的一些细节. 对 $(i, j) \in \mathcal{S}_k$, 定义 $\check{Y} = Y_i - Y_j$, $\check{X} = X_i - X_j$, 记 m 为 $\mathcal{S}_k$ 的指标, 则

$$\sum_{i,j \in \mathcal{S}_k} |e_i(\beta) - e_j(\beta)| = \sum_{m=1}^N |\check{Y}_m - \beta^T \check{X}_m|.$$

进一步定义矩阵

$$\check{Y} = \begin{pmatrix} Y_{i_1} - Y_{j_1} \\ \vdots \\ Y_{i_N} - Y_{j_N} \end{pmatrix}, \quad \check{X}(\beta) = \begin{pmatrix} \beta^T(X_{i_1} - X_{j_1}) \\ \vdots \\ \beta^T(X_{i_N} - X_{j_N}) \end{pmatrix}$$

以及增广矩阵

$$\check{Y}_{\text{aug}} = \begin{pmatrix} \check{Y}/n(n-1) \\ \mathbf{0}_{p_n} \end{pmatrix}, \quad \check{X}_{\text{aug}}(\beta, \lambda_n) = \begin{pmatrix} \check{X}(\beta)/n(n-1) \\ \mathbf{diag}(\lambda_n \hat{w}) \end{pmatrix},$$

其中, $\mathbf{0}_{p_n}$ 为 p_n 维的零向量; $\mathbf{diag}(\lambda_n \hat{w})$ 为 $p_n \times p_n$ 维对角矩阵, 其对角线元素为 $(\lambda_n \hat{w}_1, \dots, \lambda_n \hat{w}_{p_n})$. 基于这些记号, 最小化 (3.2) 可转化成最小化 $\|\check{Y}_{\text{aug}} - \check{X}_{\text{aug}}(\beta, \lambda_n)\|_1$. 这是一个线性规划问题, 可以使用标准的无惩罚 LAD 算法来求解, 如 R 中的 quantreg 包. 因此, 获得估计的具体步骤如下.

- (i) 寻找初始估计, 例如, 最小化忽略双侧截断、无正则化项的损失函数而得到的估计.
- (ii) 用迭代算法最优化无惩罚的损失函数直至最小值点收敛, 记收敛值为 $\hat{\beta}$.
- (iii) 使用 $\hat{\beta}$ 构建损失函数 $l_n^{\text{AL}}(\beta)$ 和 $l_n^{\text{AL}}(\beta, b)$.
- (iv) 将 $\hat{\beta}^{(0)}$ 作为初始值, 用线性规划算法迭代地求解最优化问题 $\min_{\beta} \|\check{Y}_{\text{aug}} - \check{X}_{\text{aug}}(\beta, \lambda_n)\|_1$, 直至收敛, 将收敛值作为 $\hat{\beta}^{\text{AL}}$.

附录 A.2 命题与定理证明

为了证明第 3.3 小节中的命题与定理, 我们需要一些正则性条件. 记 $Z = (Y, R, L, X)$ 为观测数据的总体, $\mathcal{N}$ 为 β_0 的一个 $o(1)$ 领域, S 为 Z 的取值空间. 这里, 用 $h(Z_i, Z_j; \beta)$ 表示 $h_{ij}(\beta)$, 其中 $Z_i = (Y_i, R_i, L_i, X_i)$, $i = 1, \dots, n$. 定义 $\tau(z, \beta) = \mathbb{E}[h(z, Z; \beta)]$. 用 ∇_k 表示一个 β 的函数的 k 阶微分. 假设下列条件成立.

- 条件 A.1** 密度函数 f 有界, 且其导数有界.
- 条件 A.2** 参数空间 Θ 为紧空间, 且 β_0 为其内点.
- 条件 A.3** 每个自变量的二阶矩有界.
- 条件 A.4** 对于每个 $z \in S$, $\nabla_2 \tau(z; \beta)$ 在 $\mathcal{N}$ 上存在, 并且在 β_0 处连续.
- 条件 A.5** 记 $I(z) = \sup_{\alpha \in S_p} \alpha^T \nabla_2 \tau(z; \beta_0) \alpha$, 其中 $S_p = \{\alpha \in \mathbb{R}^{p_n} : \|\alpha\| = 1\}$, $\mathbb{E}[I^2(Z)] < \infty$.
- 条件 A.6** 存在正常数 C_1, C_2, C_3 和 C_4 使得对任意 n , 有

$$0 < C_1 < \lambda_{\min}(A) \leq \lambda_{\max}(A) < C_2 < \infty, \quad 0 < C_3 < \lambda_{\min}(V) \leq \lambda_{\max}(V) < C_4 < \infty,$$

其中 $\lambda_{\min}(\cdot)$ 与 $\lambda_{\max}(\cdot)$ 分别表示一个矩阵的最小与最大特征值.

条件 A.1 与 A.2 是较为普通的条件. 条件 A.2 被用来证明 $\hat{\beta}$ 的相合性, 由此可以证明 $\hat{\beta}^{\text{AL}}$ 的相合性. 由条件 A.3 可以推导出文献 [22, 条件 5], 由此可以获得高维情形下的 Hoeffding 分解并得到 $l_n(\beta) - l_n(\beta_0)$ 的二次型逼近. 条件 A.4 用来保证 $\tau(z; \beta)$ 可以在 β_0 处进行 Taylor 展开. 条件 A.5 用来对于 $l_n(\beta) - l_n(\beta_0)$ Hoeffding 分解中的经验项给出界限. 条件 A.6 用来对目标函数二次逼近的 Hesse 矩阵进行限制, 由此保证得分函数的协方差矩阵正定, 并且其所有的特征值对于任意的 n 都有一致的界限. 这一结果可以为 $\hat{\beta}^{\text{AL}}$ 中每一维的渐近正态性的证明提供支撑. 类似的条件可以在一些相关的文献中找到, 如文献 [23–26].

命题 3.1 的证明 考虑一个 2 维 U 过程, 核函数 $h \in \mathcal{H}$, 其中 $\mathcal{H} = \{h(\cdot; \beta) : \beta \in \Theta\}$ 表示 $S^2 \equiv S \otimes S$ 上一族可测对称并有非负包络的函数. 记一个长度为 $n(n-1)$ 的集合

$$\mathbf{h}_{\omega}(\cdot; \beta) = \{h(Z_1(\omega), Z_2(\omega); \beta), \dots, h(Z_{n-1}(\omega), Z_n(\omega); \beta)\}$$

以及函数族

$$\mathcal{H}_{\omega} = \{\mathbf{h}_{\omega}(\cdot; \beta) : \beta \in \Theta\}.$$

根据文献 [22, 例 1] 和 [27, 引理 4.4], 可以得到对于每个 ω , $\mathcal{H}_{\omega}$ 的伪维数至多等于 $100p_n$. 如果 $p_n \log n / \sqrt{n} = o(1)$, 则可以保证文献 [22, 命题 1] 里关于伪维数的条件成立. 因此, 可以得到如下二次逼近:

$$l_n(\beta) - l_n(\beta_0) = \frac{1}{2}(\beta - \beta_0)^T A(\beta - \beta_0) + (\beta - \beta_0)^T \hat{U}_n + o_p(\|\beta - \beta_0\|^2) + o_p\left(\frac{\sqrt{p_n} \|\beta - \beta_0\|}{\sqrt{n}}\right),$$

其中 $\hat{U}_n = 2 \sum_{i=1}^n \xi_{ij|i} / n$, $\xi_{ij|i} = \mathbb{E}[\xi_{ij}(\beta_0) | Z_i]$. p_n 的速度限制使得伪维数的增长速度为 $o(\sqrt{n} / \log(n))$. 由此, 文献 [22, 定理 3] 的条件均可满足, 由该定理可以证明 $\hat{\beta}$ 是相合的. $\square$

为了证明定理 3.1, 我们需要证明如下引理.

引理 A.1 设条件 A.1–A.6 均成立, 对于任意 $b \in \mathbb{R}^{p_n}$, 若 $\|b\|_2 = 1$, 则 $\sqrt{n}(b^T V b)^{-1/2} b^T \hat{U}_n / 2$ 的分布函数弱收敛于标准正态分布.

证明 不难看出

$$\frac{\sqrt{n}}{2} (b^T V b)^{-1/2} b^T \hat{U}_n = \frac{1}{\sqrt{n}} \sum_{i=1}^n (b^T V b)^{-1/2} b^T \xi_{ij|i} =: I_1.$$

注意到 $E(\xi_{ij|i}) = E(\xi(Z_i, Z_j, \beta_0)) = 0$ 以及 $\text{Var}(I_1) = 1$, 由条件 A.3 与 A.4 知, $b^T V b$ 有界. 由 Cauchy-Schwarz 不等式, 可得

$$\begin{aligned} \frac{1}{n^2} \sum_{i=1}^n E|(b^T V b)^{-1/2} \cdot b^T \xi_{ij|i}|^4 &= O(1) n^{-2} \sum_{i=1}^n E|b^T \xi_{ij|i}|^4 \leq O(1) n^{-2} \sum_{i=1}^n \|\xi_{ij|i}\|_2^4 \\ &= O_p(n^{-2} \cdot n \cdot p_n^2) = O_p\left(\frac{p_n^2}{n}\right) = o_p(1). \end{aligned}$$

因此, Lyapunov 条件对 I_1 成立. 由 Lyapunov 中心极限定理可知, I_1 的分布函数弱收敛于标准正态分布. $\square$

定理 3.1 的证明 设 $\alpha_n = (p_n/n)^{1/2}$, 可以证明对任意 $\epsilon > 0$, 存在一个足够大的常数 C , 使得任意常数向量 u_n 满足 $\|u_n\| = C$, 则有

$$P\left\{\inf_{\|u_n\|=C} l_n^{\text{AL}}(\beta_0 + \alpha_n u_n) > l_n^{\text{AL}}(\beta_0)\right\} \geq 1 - \epsilon.$$

由此可以推导出存在一个局部最小值点 $\hat{\beta}^{\text{AL}}$, 且满足 $\|\hat{\beta}^{\text{AL}} - \beta_0\| = O_p(\alpha_n)$. 随后可得

$$\begin{aligned} l_n^{\text{AL}}(\beta_0 + \alpha_n u_n) - l_n^{\text{AL}}(\beta_0) &= l_n(\beta_0 + \alpha_n u_n) - l_n(\beta_0) + \sum_{j=1}^{p_n} \{P_{\lambda_{nj}}(|\beta_{0j} + \alpha_n u_{nj}|) - P_{\lambda_{nj}}(|\beta_{0j}|)\} \\ &\geq \{l_n(\beta_0 + \alpha_n u_n) - l_n(\beta_0)\} + \sum_{j=1}^{p_{n,1}} \{P_{\lambda_{nj}}(|\beta_{0j} + \alpha_n u_{nj}|) - P_{\lambda_{nj}}(|\beta_{0j}|)\} \\ &=: U_1 - U_2. \end{aligned}$$

将 U_1 表达成

$$U_1 = \alpha_n u_n^T \hat{U}_n + \frac{1}{2} \alpha_n^2 u_n^T A u_n + (\|u_n\|^2 + \|u_n\|) o_p(\alpha^2) =: U_{11} + U_{12} + U_{13},$$

其中,

$$\begin{aligned} |U_{11}| &= |\alpha_n u_n^T \hat{U}_n| \leq \alpha_n \|u_n\| \|\hat{U}_n\| = \alpha_n \|u_n\| O_p\left(\sqrt{\frac{p_n}{n}}\right) = \|u_n\| O_p(\alpha_n^2), \\ |U_{12}| &= \left|\frac{1}{2} \alpha_n^2 u_n^T A u_n\right| \geq \frac{1}{2} \alpha_n^2 \|u_n\|^2 \lambda_{\min}(A) > \frac{1}{2} \alpha_n^2 \|u_n\|^2 C_1. \end{aligned}$$

因此, 在 U_1 中, 对于足够大的 $\|u_n\|$, $|U_{12}|$ 远大于 $|U_{11}|$ 与 $|U_{13}|$. 同时

$$|U_2| = \left| \sum_{j=1}^{p_{n,1}} \frac{\lambda_n}{|\hat{\beta}_j|} |\beta_{0j} + \alpha_n u_{nj}| - \frac{\lambda_n}{|\hat{\beta}_j|} |\beta_{0j}| \right|$$

$$\begin{aligned} &\leq \alpha_n \lambda_n \left| \sum_{j=1}^{p_{n,1}} \frac{|u_{nj}|}{|\hat{\beta}_j|} \right| \\ &= \alpha_n \lambda_n \sqrt{p_{n,1}} O_p(1) \|u_n\| \|u_n\| o_p(\alpha_n^2). \end{aligned}$$

对于足够大的 $C = \|u_n\|$, $|U_{12}|$ 远大于 $|U_2|$. 因为 A 是正定的, 故 $U_{12} > 0$. 由此, 当 $n \rightarrow \infty$ 且 C 足够大时, $l_n^{\text{AL}}(\beta_0 + \alpha_n u_n) - l_n^{\text{AL}}(\beta_0) > 0$ 的概率趋近于 1. $\square$

为了证明定理 3.2, 我们还需要一个引理.

引理 A.2 设条件 A.1–A.6 均成立, 如果 $\sqrt{n}\lambda_n = o(1)$, $\lambda_n n/p_n \rightarrow \infty$, $p_n \log n/\sqrt{n} = o(1)$, 则对任意满足 $\|\beta_1 - \beta_{0,1}\| = O(p_n^{1/2} n^{-1/2})$ 的 β_1 以及任意常数 C ,

$$l_n^{\text{AL}}((\beta_1^{\text{T}}, 0^{\text{T}})^{\text{T}}) = \min_{\|\beta_2\| \leq Cp_n^{1/2} n^{-1/2}} l_n^{\text{AL}}((\beta_1^{\text{T}}, \beta_2^{\text{T}})^{\text{T}})$$

的概率趋近于 1.

证明 为了得到引理结论, 只需证明对任意满足 $\|\beta_1 - \beta_{0,1}\| = O(p_n^{1/2} n^{-1/2})$ 的 β_1 以及 $\|\beta_2\| \leq Cp_n^{1/2} n^{-1/2}$, $\partial l_n^{\text{AL}}(\beta)/\partial \beta_j$ 与 β_j ($j = p_{n,1} + 1, \dots, p_n$) 同号的概率趋近于 1. 我们有

$$\begin{aligned} \frac{\partial l_n^{\text{AL}}(\beta_n)}{\partial \beta_j} &= \hat{U}_n(\beta_0)_j + \sum_{k=1}^{p_n} A_{jk}(\beta_k - \beta_{0k}) + o_p\left(\|\beta_n - \beta_0\| + \sqrt{\frac{p_n}{n}}\right) + P'_{\lambda_n j}(|\beta_j|)\text{sign}(\beta_j) \\ &=: V_1 + V_2 + V_3 + V_4. \end{aligned}$$

可以证明

$$\begin{aligned} V_1 &= O_p\left(\frac{1}{\sqrt{n}}\right) = o_p\left(\sqrt{\frac{p_n}{n}}\right), \\ |V_2| &= \left| \sum_{k=1}^{p_n} A_{jk}(\beta_k - \beta_{0k}) \right| \leq \|\beta - \beta_0\| \left(\sum_{k=1}^{p_n} A_{jk}^2 \right)^{\frac{1}{2}} = O_p\left(\sqrt{\frac{p_n}{n}}\right) O(1) = O_p\left(\sqrt{\frac{p_n}{n}}\right), \\ |V_3| &= o_p\left(\sqrt{\frac{p_n}{n}}\right). \end{aligned}$$

因此, $V_1 + V_2 + V_3 = O_p(\sqrt{p_n/n})$. 由此可得

$$\begin{aligned} \frac{\partial l_n^{\text{AL}}(\beta_n)}{\partial \beta_j} &= P'_{\lambda_n j}(|\beta_j|)\text{sign}(\beta_j) + O_p\left(\sqrt{\frac{p_n}{n}}\right) \\ &= \frac{\lambda_n}{|\hat{\beta}_{nj}|} \text{sign}(\beta_j) + O_p\left(\sqrt{\frac{p_n}{n}}\right) \\ &= \frac{\lambda_n}{|\hat{\beta}_{nj}|} \left\{ \text{sign}(\beta_j) + O_p\left(\frac{p_n}{n\lambda_n}\right) \right\}. \end{aligned}$$

因为 $\lambda_n n/p_n \rightarrow \infty$, 相比于 $\text{sign}(\beta_j)$, $O_p(p_n/(n\lambda_n))$ 可以被忽略. 由此, $\partial l_n^{\text{AL}}(\beta)/\partial \beta_j$ 与 β_j ($j = p_{n,1} + 1, \dots, p_n$) 同号的概率趋近于 1. $\square$

定理 3.2 的证明 第 (i) 部分的结论可以由引理 A.2 直接得到. 对于第 (ii) 部分, 先记

$$B_n = \{P'_{\lambda_{n1}}(|\beta_{01}|)\text{sign}(\beta_{01}), \dots, P'_{\lambda_{np_{n,1}}}(|\beta_{0k}|)\text{sign}(\beta_{0p_{n,1}})\}^{\text{T}}.$$

不难看出, 当 $n \rightarrow \infty$ 时, B_n 收敛至 0. 因为 $\hat{\beta}^{\text{AL}}$ 是 $l_n^{\text{AL}}(\beta)$ 的最小值点, 故 $\partial l_n^{\text{AL}}(\hat{\beta}^{\text{AL}})/\partial \beta_1 = 0$. 加上 $\hat{\beta}_2^{\text{AL}} = 0$, 可得

$$\hat{U}_1 + A_{11}(\hat{\beta}_1^{\text{AL}} - \beta_{0,1}) + o_p\left(\|\beta_n - \beta_0\| + \sqrt{\frac{p_n}{n}}\right) + B_n = 0, \quad (\text{A.1})$$

其中 $\hat{U}_1$ 包含 $\hat{U}_n$ 中最前面的 $p_{n,1}$ 个分量. 将 (A.1) 重新排布, 可得

$$\sqrt{n}(\hat{\beta}_1^{\text{AL}} - \beta_{0,1}) = -\sqrt{n}A_{11}^{-1}\hat{U}_1 - \sqrt{n}A_{11}^{-1}B_n + o_p(\sqrt{p_n}).$$

由条件 $\sqrt{n}\lambda_n = o(1)$, 可得 $\sqrt{n}A_{11}^{-1}B_n \rightarrow 0$. 再由条件 A.1 与 A.3, 可推导出 $E|\nabla_{1j}\tau(Z'_1\beta_0)|^2 < \infty$ 对 $j = 1, \dots, p_n$ 成立. 因此 $\|A_{11}^{-1/2}\hat{U}_1\| = O_p(\sqrt{p_n/n})$,

$$\sqrt{n}(\hat{\beta}_1^{\text{AL}} - \beta_{0,1}) = \sqrt{n}A_{11}^{-1}\hat{U}_1(1 + o_p(1)).$$

因为任意非零 $p_{n,1} \times 1$ 维常数向量 u_n 满足 $\|u_n\| = 1$, $\sqrt{n}u_n^T A_{11}^{-1}\hat{U}_1/[2[u_n^T A_{11}^{-1}V_{11}A_{11}^{-1}u_n]^{-1/2}]$ 的分布函数弱收敛于标准正态分布, 因此, 由 Slutsky 定理可以得到 $\sqrt{n}u_n^T(\hat{\beta}_1^{\text{AL}} - \beta_{0,1})/[2[u_n^T A_{11}^{-1}V_{11}A_{11}^{-1}u_n]^{-1/2}]$ 的分布函数弱收敛于标准正态分布. $\square$

定理 3.3 的证明 为了更清晰地陈述证明, 我们需要引入一些额外的记号. 将真实的模型记为 $\mathcal{M}^* = \{j : \beta_{0j} \neq 0, j = 1, \dots, p_n\}$. 如果我们用 $\hat{\beta}_\lambda$ 表示调节参数取值为 λ 时的回归系数估计, 则相应的模型记为 $\mathcal{M}_\lambda = \{j : \hat{\beta}_{\lambda,j} \neq 0, j = 1, \dots, p_n\}$. 对于任意大小为 m 的模型 $\mathcal{M} = \{j_1, \dots, j_m\}$, 其相应的系数估计记为 $\hat{\beta}_\mathcal{M} = (\beta_{j_1}, \dots, \beta_{j_m})$. 需要注意的是 $\hat{\beta}_\mathcal{M}$ 与 $\hat{\beta}_\lambda$ 定义不同, 我们需要仔细区分它们. 接下来, 将 $\mathbb{R}^p$ 空间划分成 3 个互不相交的区域: (i) $\Omega_- = \{\lambda > 0 : \mathcal{M}_\lambda \not\supset \mathcal{M}^*\}$; (ii) $\Omega_0 = \{\lambda > 0 : \mathcal{M}_\lambda = \mathcal{M}^*\}$; (iii) $\Omega_+ = \{\lambda > 0 : \mathcal{M}_\lambda \supset \mathcal{M}^*, \mathcal{M}_\lambda \neq \mathcal{M}^*\}$. 这一划分将 λ 的范围分为 3 段, 分别对应于拟合不足的模型、正确拟合的模型以及过度拟合的模型. 对于拟合不足的模型, 存在 $j \in \mathcal{M}^*$ 与 $j \notin \mathcal{M}_\lambda$. 过度拟合的模型是指那些如果 $j \in \mathcal{M}^*$ 则 $j \in \mathcal{M}_\lambda$ 但却存在 $k \in \mathcal{M}_\lambda$ 同时 $k \notin \mathcal{M}^*$ 的模型. 通过考虑拟合不足的模型与过度拟合的模型这两部分来证明定理.

对于拟合不足的模型, 当 $\lambda \in \Omega_-$ 时, 估计向量 $\hat{\beta}_S^{\text{AL}}$ 的长度与 β_0 不匹配. 为了使 $\hat{\beta}_S^{\text{AL}}$ 与 β_0 可比, 这里稍微修正 $\hat{\beta}_S$ 的定义. 对于任意模型 S , 定义

$$\hat{\beta}_S = \underset{\beta \in \mathbb{R}_{k_n} : \beta_j = 0, \forall j \notin S}{\operatorname{argmin}} l_n(\beta).$$

由定义可得

$$\begin{aligned} \inf_{\lambda \in \Omega_-} \text{BIC}_\lambda &= \inf_{\lambda \in \Omega_-} l_n(\hat{\beta}_\lambda) + \frac{\log n}{n} C_n \cdot \text{df}_\lambda \geq \inf_{\lambda \in \Omega_-} l_n(\hat{\beta}_\lambda) \\ &= \inf_{\lambda \in \Omega_-} l_n(\beta_0) + \frac{1}{2}(\hat{\beta}_\lambda^* - \beta_0)^T A(\hat{\beta}_\lambda^* - \beta_0) \\ &\geq l_n(\beta_0) + \frac{1}{2}\lambda_{\min}(A) \inf_{\lambda \in \Omega_-} \|\hat{\beta}_\lambda^* - \beta_0\|^2 \\ &\geq l_n(\beta_0) + \frac{1}{2}\lambda_{\min}(A) \inf_{S \not\supset \mathcal{M}^*} \|\hat{\beta}_S - \beta_0\|^2 \\ &\geq l_n(\beta_0) + \frac{1}{2}\lambda_{\min}(A) \min_{j \in \mathcal{M}^*} (\beta_{0,j}^2). \end{aligned}$$

由条件 A.6, 可得

$$\inf_{\lambda \in \Omega_-} \text{BIC}_\lambda - l_n(\beta_0) \geq \frac{1}{2}C_1 \min_{j \in \mathcal{M}^*} (\beta_{0,j}^2). \quad (\text{A.2})$$

再由条件

$$\sqrt{\frac{n}{\{C_n p_n \log n\}}} \times \liminf_{n \rightarrow \infty} \left\{ \min_{j \in S_T} |\beta_{0,j}| \right\} \rightarrow \infty,$$

可知 (A.2) 中右边项的阶数为 $O(C_n p_n \log n/n)$, 当取 $C_n = \log \log p_n$ 时, 该阶数等于 $O(\log \log p_n \times p_n \log n/n)$. 类似地, 可以推导出

$$\begin{aligned} \text{BIC}_{\hat{\lambda}} &= l_n(\hat{\beta}) + \frac{\log n}{n} C_n \cdot \text{df} \leq l_n(\hat{\beta}) + \frac{p_n \log n \log \log p_n}{n} \\ &= l_n(\beta_0) + \frac{1}{2}(\hat{\beta} - \beta_0)^T A(\hat{\beta} - \beta_0) + \frac{p_n \log n \log \log p_n}{n} \\ &\leq l_n(\beta_0) + \frac{1}{2} \lambda_{\max}(A) \|\hat{\beta} - \beta_0\|^2 + \frac{p_n \log n \log \log p_n}{n} \\ &= l_n(\beta_0) + O_p\left(\frac{p_n}{n}\right) + \frac{p_n \log n \log \log p_n}{n} \\ &= l_n(\beta_0) + o_p\left(\frac{p_n \log n \log \log p_n}{n}\right) + \frac{p_n \log n \log \log p_n}{n} \end{aligned}$$

与

$$\text{BIC}_{\hat{\lambda}} - l_n(\beta_0) = O_p\left(\frac{p_n \log n \log \log p_n}{n}\right).$$

综合这两个结果, 可得

$$\mathbb{P}\left(\inf_{\lambda \in \Omega_-} \text{BIC}_{\lambda} > \text{BIC}_{\hat{\lambda}}\right) \rightarrow 1. \quad (\text{A.3})$$

对于过度拟合的模型, 任取 $\lambda \in \Omega_+$, 有

$$\begin{aligned} \inf_{\lambda \in \Omega_+} (\text{BIC}_{\lambda} - \text{BIC}_{\hat{\lambda}}) &= \inf_{\lambda \in \Omega_+} l_n(\hat{\beta}_{\lambda}) - l_n(\hat{\beta}) + (\text{df}_{\lambda} - \text{df}_{\hat{\lambda}}) \frac{\log n}{n} C_n \\ &= \inf_{\lambda \in \Omega_+} \frac{1}{2}(\hat{\beta}_{\lambda} - \beta_0)^T A(\hat{\beta}_{\lambda} - \beta_0) - \frac{1}{2}(\hat{\beta} - \beta_0)^T A(\hat{\beta} - \beta_0) \\ &\quad + (\text{df}_{\lambda} - \text{df}_{\hat{\lambda}}) \frac{\log n}{n} C_n \\ &\geq -\frac{1}{2} \lambda_{\max}(A) \|\hat{\beta} - \beta_0\|^2 + (\text{df}_{\lambda} - \text{df}_{\hat{\lambda}}) \frac{\log n}{n} C_n \\ &\geq -\frac{1}{2} C_2 \|\hat{\beta} - \beta_0\|^2 + \frac{\log n}{n} C_n, \end{aligned}$$

其中最后一行成立是因为对任意 $\lambda \in \Omega_+$, 必有 $\text{df}_{\lambda} - \text{df}_{\hat{\lambda}} \geq 1$. 更进一步, 因为 $\|\hat{\beta}^{\text{AL}} - \beta_0\|^2 = O_p(p_n/n)$, 所以

$$\inf_{\lambda \in \Omega_+} (\text{BIC}_{\lambda} - \text{BIC}_{\hat{\lambda}}) \geq (1 + o_p(1)) \frac{p_n \log n \log \log p_n}{n} > 0.$$

这表明

$$\mathbb{P}\left(\inf_{\lambda \in \Omega_+} \text{BIC}_{\lambda} > \text{BIC}_{\hat{\lambda}}\right) \rightarrow 1. \quad (\text{A.4})$$

最后, 结合 (A.3) 与 (A.4), 可得

$$\mathbb{P}(\mathcal{M}_{\hat{\lambda}} = \mathcal{M}^*) \rightarrow 1, \quad n \rightarrow \infty.$$

证毕. □

Variable selection in doubly truncated regression models

Ming Zheng, Chanjuan Lin & Wen Yu

Abstract Doubly truncated data arise in many areas such as astronomy, econometrics, and medical studies. For the regression analysis with doubly truncated response variables, the existence of double truncation may bring bias for the estimation as well as affect the variable selection. We propose a simultaneous estimation and variable selection procedure for the doubly truncated regression, allowing a diverging number of regression parameters. To remove the bias introduced by the double truncation, a Mann-Whitney-type loss function is used. The adaptive LASSO (least absolute shrinkage and selection operator) penalty is added to the loss function to achieve simultaneous estimation and variable selection. An iterative algorithm is designed to optimize the resulting objective function. We establish the consistency and the asymptotic normality of the proposed estimator. The oracle property of the proposed selection procedure is also obtained. Some simulation studies are conducted to show the finite sample performance of the proposed approach. We also apply the method to analyze a real astronomical data set.

Keywords double truncation, variable selection, diverging number of parameters, adaptive LASSO, least absolute deviation, oracle property

MSC(2020) 62J07, 62N99

doi: 10.1360/SCM-2021-0158