文章编号:1007-6735(2017)06-0542-07 DOI:10.13255/j.cnki.jusst.2017.06.007

基于 t 函数的稳健变量选择方法

钟先乐， 樊亚莉， 张探探
（上海理工大学 理学院,上海 200093）

摘要: 在已有研究的基础上,提出一种新的基于 t 函数的稳健变量选择方法.该方法通过惩罚估计方程中的惩罚函数达到变量选择的效果,方程中的权重矩阵和有界得分函数对自变量和因变量中的异常值有很好的限制作用,可同时达到稳健的变量选择和稳健估计.通过分析3种不同自由度的 t 函数性质,选取自由度为2的 t 函数,并与基于Huber函数的稳健变量选择方法进行比较.数值模拟结果表明,基于 t 函数的稳健变量选择方法在2种污染力度、3种污染方式的数据污染情况下,其稳健性均明显优于基于Huber函数的稳健变量选择方法.与参数估计效果相比,基于 t 函数的稳健变量选择方法优势更明显.

关键词: 稳健性;变量选择;惩罚函数;估计方程
中图分类号: O 212.1 **文献标志码**: A

Robust Variable Selection Method Based on t Function

ZHONG Xianle, FAN Yali, ZHANG Tantan
(*College of Science, University of Shanghai for Science and Technology, Shanghai 200093, China*)

Abstract: On the basis of existing research results, a new method for robust variable selection based on t function was proposed. The method achieves variable selection by the penalized function in the penalized estimating equation. Meanwhile, through the weight matrix and the bounded score function in the equation, the proposed method has a satisfactory resistance to outliers in independent variables and dependent variables, and can achieve the robust variable selection and estimation simultaneously. By analyzing the properties of t functions with 3 different kinds of degree of freedom, a t function with 2 degrees of freedom was used, and the method was compared with the Huber function based on robust variable selection method. The numerical simulation results show that, when data are contaminated with outliers by two kinds of contamination degree and three kinds of contamination mode, the proposed robust variable selection method based on t function performs better than that based on the Huber function, especially in variable selection.

Keywords: *robustness*; *variable selection*; *penalized function*; *estimating equation*

收稿日期: 2017-04-05
基金项目: 国家自然科学基金资助项目(11401383)
第一作者: 钟先乐(1993-),女,硕士研究生.研究方向:概率论与数理统计.E-mail:1040906542@qq.com
通信作者: 樊亚莉(1978-),女,讲师.研究方向:概率论与数理统计.E-mail:yalifan@usst.edu.cn

随着数据获取技术的迅猛发展，人们获取的数据结构越来越复杂，维数越来越高．统计学的主要任务就是对观测数据的因变量和自变量建立模型，进而对数据进行分析、预测以及一些统计推断．在现实问题中，因变量往往同时受多个自变量影响，但这些影响并不都很显著．人们通常希望在模型中只引进对因变量有重要影响的自变量，所以，变量选择就成了建模前的必要工作．但是，现实问题中，数据经常被污染，往往存在异常值，这时用普通的变量选择方法就会对模拟结果产生很大的偏差．

针对变量选择的问题，统计学家已经作出了大量研究．1996 年，统计学家 Tibshirani[1] 提出了一种变量选择方法 LASSO，基本思想是在最小二乘法的基础上施加 L_1 惩罚．2001 年，Fan 等[2] 提出了变量选择的 SCAD 方法，并研究了该方法的 Oracle 性质．在某些 LASSO 不相合的情况下，Zou[3] 又提出 Adaptive LASSO，该方法是对 LASSO 的一种改进，能够满足 Oracle 性质．为了克服 LASSO 的一些缺点，Zou 等[4] 提出了 Elastic Net 变量选择方法．针对高维数据，Candès 等[5] 提出了 Dantzig Selector 方法．

针对数据中可能存在异常值这一情况，有许多文献已经研究了稳健估计与稳健变量选择方法．文献[6－7]率先提出当正态分布被污染时，估计位置参数的渐进理论．文献[8－9]将最小一乘法用到稳健估计中，之后文献[10]进一步分析最小一乘法的优良性质．文献[11]提出了基于 t 函数的稳健估计方法，考察了基于 t 函数估计量的优良性．同时研究稳健估计方法的还有文献[12－14]．文献[15]提出了基于 Huber 函数的针对纵向数据的稳健变量选择方法．针对稳健估计中常用的 t 函数和 Huber 函数，文献[16]提出了基于 M 估计的稳健向前变量选择方法，并进一步考察了 t 函数和 Huber 函数在稳健向前变量选择中的性质．

本文在前人研究的基础上，提出一种新的基于 t 函数的稳健变量选择方法，并与文献[15]中基于 Huber 函数的稳健变量选择方法进行比较．模拟结果显示，t 函数方法对数据中的异常值有更好的限制作用，可以达到更好的变量选择效果．文章主要分为 5 个部分，第 1 部分介绍了稳健的惩罚估计方程．第 2 部分将 t 函数和 Huber 函数的性质进行比较分析，突出 t 函数在稳健变量选择方法中的优势．第 3 部分介绍本文中所用的算法．第 4 部分是数值模拟，通过 3 种污染方式来污染数据，比较本文方法与文献[15]中方法的模拟效果．第 5 部分为结论．

1　稳健的惩罚估计方程

考虑如下线性模型：

$$\boldsymbol{y} = \boldsymbol{x}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{1}$$

式中：$\boldsymbol{y} = (y_1, y_2, \cdots, y_n)^{\mathrm{T}}$；$\boldsymbol{x} = (x_1, x_2 \cdots, x_n)^{\mathrm{T}}$，$\boldsymbol{x}_i = (x_{i1}, x_{i2}, \cdots, x_{ip})$；$\boldsymbol{\beta} = (\beta_1, \beta_2, \cdots, \beta_p)^{\mathrm{T}}$；$\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n)^{\mathrm{T}}$，且 $i = 1, 2, \cdots, n$，ε_i 的期望值为 0，方差为 σ^2，$\boldsymbol{\varepsilon}$ 的各分量相互独立．

与文献[12]类似，考虑如下稳健估计方程

$$\boldsymbol{R}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \boldsymbol{x}_i^{\mathrm{T}} W_i \varphi\left(\frac{y_i - \boldsymbol{x}_i\boldsymbol{\beta}}{\sigma}\right) = 0 \tag{2}$$

式中，W_i 是权重矩阵 $\boldsymbol{W}$ 的第 i 个分量，权重 W_i 通过文献[13]得来，用来降低自变量中异常值的影响，定义如下：

$$W_i = \min\left\{1, \left[\frac{p_0}{(\boldsymbol{x}_i - m_x) S_x^{-1} (\boldsymbol{x}_i - m_x)^{\mathrm{T}}}\right]^{\frac{r}{2}}\right\} \tag{3}$$

式中：r 为大于 1 的常数；p_0 为自由度与 $\boldsymbol{x}_i$ 维数相同的卡方分布的 0.95 分位数；取 m_x 为 $\boldsymbol{x}_i$ 的中位数，则 m_x 的第 k 个分量取为 $\boldsymbol{x}$ 第 k 列的中位数；S_x 的第 k 个对角元取为 $1.483(\mathrm{median}|\boldsymbol{x}(k) - m_x(k) \otimes \boldsymbol{I}_n|)$，$\boldsymbol{x}(k)$表示 $\boldsymbol{x}$ 的第 k 列，$m_x(k)$表示 m_x 的第 k 个分量，$\otimes$表示 kronecker 乘积，$\boldsymbol{I}_n$ 表示 n 维元素全为 1 的列向量．

式(2)中，函数 $\varphi(\cdot)$是一个有界得分函数，用来限制因变量中异常值的影响，本文将此函数定义为自由度为 2 的 t 函数，记作 t_2 函数．当 $\varphi(x) = x$ 且 $W_i = 1$ 时，原稳健估计方程就退化成一般的估计方程，不再具有稳健性，即为非稳健的估计方程，此时估计方程(2)会对异常值有较大的敏感性．

通过求解式(2)，可以得到稳健参数估计．为了同时达到变量选择的效果，采用压缩估计方法，即在估计方程中再添加一个惩罚项．因此考虑惩罚稳健估计方程

$$\boldsymbol{R}^P(\boldsymbol{\beta}) = \sum_{i=1}^{n} \boldsymbol{x}_i{}^{\mathrm{T}} W_i \varphi\left(\frac{y_i - \boldsymbol{x}_i\boldsymbol{\beta}}{\sigma}\right) - n q_\lambda(|\boldsymbol{\beta}|)\mathrm{sgn}(\boldsymbol{\beta}) = 0 \tag{4}$$

式中，$q_\lambda(|\boldsymbol{\beta}|) = (q_{\lambda,1}(|\beta_1|), \cdots, q_{\lambda,p}(|\beta_p|))^{\mathrm{T}}$ 为某些惩罚函数的导数，即 $q_{\lambda,j}(\cdot) = p'_{\lambda,j}(\cdot)$，$j = 1, 2, \cdots, p$．$\lambda$ 是大于 0 的调节参数，其取值不同，惩罚的程度就不同．

本文所考虑的惩罚函数主要是 SCAD 惩罚函数[3]．取惩罚函数为

$$p_{\lambda,j}(|\beta_j|)=\lambda|\beta_j|\left\{\boldsymbol{I}(|\beta_j|<\lambda)+\frac{(a-|\beta_j|/2\lambda)}{a-1}\boldsymbol{I}(\lambda<|\beta_j|\leqslant a\lambda)+\frac{a^2\lambda}{(a-1)2|\beta_j|}\boldsymbol{I}(|\beta_j|\geqslant a\lambda)\right\},$$

$$j=1,2,\cdots,p,a>2$$

2 t 函数与 Huber 函数

t 分布的密度函数为

$$f_v(t)=\frac{\Gamma\left(\frac{v+1}{2}\right)}{\sqrt{v\pi}\Gamma\left(\frac{v}{2}\right)}\left(1+\frac{t^2}{v}\right)^{-\frac{v+1}{2}}$$

式中，v 是 t 分布的自由度，直接控制着 t 分布密度函数尾部的厚度，$\Gamma(a)=\int_0^{\infty}t^{\theta-1}\mathrm{e}^{-t}\mathrm{d}t$. 令 $\rho_v(t)=-\log f_v(t)$可以得出

$$\rho_v(t)=\log\left\{\frac{\Gamma\left(\frac{v+1}{2}\right)}{\sqrt{v\pi}\Gamma\left(\frac{v}{2}\right)}\right\}-\frac{v+1}{2}\log\left(1+\frac{t^2}{v}\right)$$

再令 $\varphi_v(t)=\rho'_v(t)$，得到 t 函数 $\varphi_v(t)=\frac{t}{t^2+v}$.

Huber 分布的密度函数为

$$f_c(x)=\frac{1-\zeta}{\sqrt{2\pi}\sigma}\mathrm{e}^{-\rho_c(x)}$$

其中

$$\rho_c(x)=\begin{cases}\frac{1}{2}x^2, & |x|<c\\ c|x|-\frac{1}{2}c^2, & |x|\geqslant c\end{cases}$$

且 ζ 满足$\frac{2\varphi(c)}{c}-2\Phi(-c)=\frac{\zeta}{1-\zeta}$，其中 φ 是标准正态分布的密度函数，Φ 是标准正态分布的分布函数. 令 $\varphi_c(x)=\rho'_c(x)$，得到 Huber 函数 $\varphi_c(x)=\min\{c,\max\{-c,x\}\}$，$c$ 是一个调节参数，平衡估计量的稳健性和效率. 在模拟中，c 越大，估计的效率越高，估计的稳健性就越差；反之，c 越小，估计的效率越低，但稳健性越好. 在本文的数值模拟中，模型的误差项服从标准正态分布，因此在模拟中，取$c=2$.

自由度不同，t 分布密度函数的尾部厚度不同，从而 t 函数对异常值的抑制效果不同. 图 1 是自由度分别为 2,6,10 的 t 分布密度函数，由图 1 可见，自由度越小，密度函数的尾部越厚. 文献[16]已经证明厚尾性对异常值有更好的抑制作用. 图 2 是自由度为 2 的 t 分布密度函数和 Huber 分布密度函数的图像比较，由图 2 显然可见，t 分布密度函数的尾部更厚. 由此可以初步推断，基于自由度为 2 的 t 函数的稳健压缩估计可以对异常值有更好的限制作用. 下面，进一步分析 t 函数和 Huber 函数的图像区别，以及通过图像显现出来的对异常值的作用效果，如图 3 和图 4 所示.

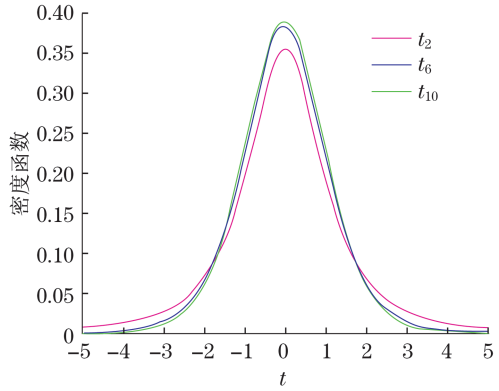

图 1 不同自由度的 t 分布密度函数

Fig.1 t distribution density function with different kinds of degree of freedom

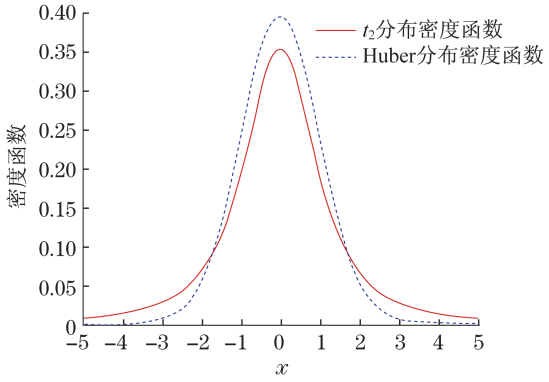

图 2 t 分布和 Huber 分布密度函数比较

Fig.2 Comparison density function between t distribution and Huber distribution

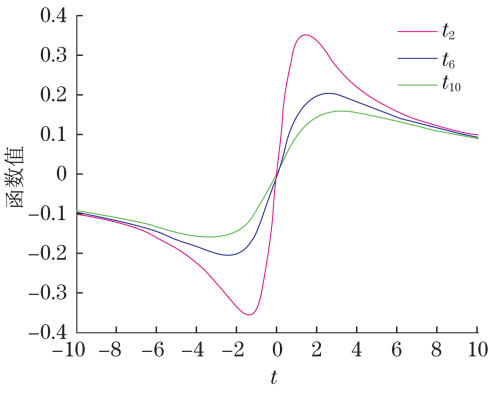

图 3 t 函数

Fig.3 t function

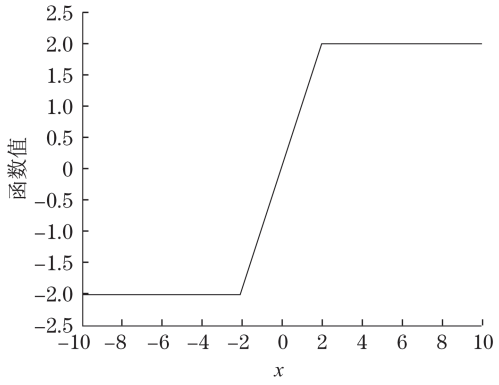

图 4　Huber 函数

Fig.4　Huber function

图 3 是自由度分别为 2,6,10 的 t 函数的图像，由图 3 可见，t 函数并不是单调的，而是一个回降函数，而且随着自由度的增大，在一定自变量范围内，函数的变化范围在变小. 可以看出，当变量 t 的绝对值变大时，t 函数将会对这些绝对值较大的变量产生作用，使其函数值接近于 0，因此 t 函数可以很好地抑制数值模拟中异常值的影响. 尽管自由度不同的 t 函数对较大异常值的抑制作用不相上下，但图 3 表明，当变量 t 处在正常值范围内时，自由度越大的 t 函数会对变量施加越大的抑制作用，使原本正常的数据也受到更大的影响，从而破坏了数据原有的真实性. 所以，综合而言，t_2 函数是最优的. 本文将选取自由度 $v=2$.

文献[11]也模拟分析了自由度分别为 1 和 4 的 t 函数在 M 估计中的稳健性，其模拟结果表明，自由度为 1 的 t 函数比自由度为 4 的 t 函数具有更好的稳健性. 本文也模拟分析了自由度分别为 1 和 2 的 t 函数在变量选择中的稳健性. 结果表明，自由度为 1 的 t 函数和自由度为 2 的 t 函数的模拟结果比较接近. 在模拟设置下，自由度为 2 的 t 函数比自由度为 1 的 t 函数在变量选择和参数估计方面稍好一些，因此本文只报告自由度为 2 的 t 函数的结果. 文献[16]也得出了与本文相类似的结论，它们的研究表明，在 M 估计中，取自由度较小的 t 函数对异常值有较好的限制作用.

通过图 3 和图 4 的比较可见，当自变量趋于正无穷时，Huber 函数值为 +2，当自变量趋于负无穷时，Huber 函数值为 −2. 而无论自变量趋于正无穷还是负无穷，t_2 函数值始终趋近于 0. 因此，t_2 函数的稳健方法能减小异常值在模型估计中的作用，更好地削弱异常值的影响[14]. 所以，t_2 函数在变量选择中比 Huber 函数具有更好的稳健性.

3　算法

本文算法与文献[15]类似，采用牛顿迭代法，具体算法如下：

a. 对给定的一个 λ 值，首先计算 $\boldsymbol{\beta}$ 的初始值 $\boldsymbol{\beta}^{(0)}$. 本文取最小二乘估计作为初始值.

b. 用 $\boldsymbol{\beta}^{(k)}$ 表示第 k 次迭代所得的估计值，$k\geqslant 0$，则

$$\boldsymbol{\beta}^{(k+1)}=\boldsymbol{\beta}^{(k)}-[\boldsymbol{D}\boldsymbol{\beta}^{(k)}-\Delta_\lambda(\boldsymbol{\beta}^{(k)})]^{-1}\boldsymbol{R}^p(\boldsymbol{\beta}^{(k)})$$

这里，$\boldsymbol{D}$ 为 p 阶导数阵 $\partial(\boldsymbol{R}^p(\boldsymbol{\beta}))/\partial\boldsymbol{\beta}$.

$$\Delta_\lambda(\boldsymbol{\beta})=\mathrm{diag}(q_\lambda(|\beta_1|)/(\delta+|\beta_1|),\cdots,q_\lambda(|\beta_p|)/(\delta+|\beta_p|))$$

式中，δ 是一个非常小的正数，文中取值为 10^{-4}.

c. 对当前给定的 λ 值，当 $\|\boldsymbol{\beta}^{(k+1)}-\boldsymbol{\beta}^{(k)}\|\leqslant 10^{-4}$ 时，停止迭代运算，记最终迭代的结果为 $\hat{\boldsymbol{\beta}}_\lambda$.

d. 将 $\hat{\boldsymbol{\beta}}_\lambda$ 代入到 CV 统计量中，来衡量模型和数据的拟合效果. $CV(\lambda)=\dfrac{RSS_R(\lambda)/n}{\{1-d(\lambda)/n\}^2}$，其中，$RSS_R(\lambda)$ 是稳健化的残差平方和，$d(\lambda)=tr[(\boldsymbol{D}+\Delta_\lambda(\hat{\boldsymbol{\beta}}_\lambda)^{-1})\boldsymbol{D}]$.

e. 取 $\lambda=\arg\min_\lambda CV(\lambda)$ 为最终的 λ，然后再进行步骤 a～d，算出最终的 $\hat{\boldsymbol{\beta}}$. SCAD 惩罚函数中的参数 a 采纳文献[3]的建议，令 $a=3.7$.

λ 的取值并没有特别规定，依据所用模型和数据将提前给出一个范围[1]. 文中设定 λ 为 $\dfrac{1}{\sqrt{n}}$～5 之间的均匀间隔的 50 个数.

4　数值模拟

通过式（1）产生真实数据，这里令 $\boldsymbol{\beta}=(3,2.5,0,0,3,0,0,0)^{\mathrm{T}}$，$\boldsymbol{x}_i$ 取自 p 维多元正态分布，即 $\boldsymbol{x}_i\sim N(0,\boldsymbol{I}_p)$，这里 $p=8$. 随机误差 ε_1，$\varepsilon_2,\cdots,\varepsilon_n$ 相互独立，且都服从标准正态分布.

为了全面考察本文所提方法的稳健性，这里将对数据进行 2 种污染力度、3 种污染方式来产生异常值. 通过 2 种污染力度在数据中分别产生较大异常值和较小异常值. 污染方式 1 是随机选取 5 个样本只对 $\boldsymbol{x}$ 进行污染，将原本 $\boldsymbol{x}_i$ 变为 $\boldsymbol{x}_i+5$（或 $\boldsymbol{x}_i+2.5$）；污染方式 2 是随机选取 5 个样本只对 $\boldsymbol{y}$ 污染，将原本 y_i 变为 y_i+10（或 y_i+3）；污染方式 3 是随机选取 3 个样本对 $\boldsymbol{x}$ 污染，再随机选取 2 个样本对

$\boldsymbol{y}$ 污染，且所选样本不重复，将选取的 $\boldsymbol{x}_i$ 变为 $\boldsymbol{x}_i+5$（或 $\boldsymbol{x}_i+2.5$），y_i 变为 y_i+10（或 y_i+3）.

为了排除随机性对该方法的影响，本文将重复模拟200次，采用8个指标来衡量各种方法的效果. 首先用估计的偏差、方差和均方误差来衡量估计的优劣；另一方面，用取对率、拟合不足率、拟合过度率、取对个数和取错个数来衡量变量选择的优劣. 其中，取对率就是将 $\boldsymbol{\beta}$ 中原本为0的分量估计为0，且将 $\boldsymbol{\beta}$ 中原本非0的分量估计为非0的总次数在200次模拟中的比率. 拟合过度率为将 $\boldsymbol{\beta}$ 中原本的非0估计为非0，原本为0的也估计为非0的总次数在200次中的比率. 拟合不足率即为200次模拟中，估计结果将 $\boldsymbol{\beta}$ 中原本不为0的估计为0的总次数在200次中的比率. 取对个数是200次模拟中正确估出0的个数的平均值，取错个数是200次模拟中将原本非0估计为0的个数的平均值，取对个数最好为5，取错个数最差为3.

SCAD方法是变量选择中比较常见的惩罚方法，由文献[2]可知，SCAD方法在稳健变量选择中具有较好的优良性，因此本文主要模拟不同SCAD方法的稳健效果. 表1～3分别表示污染1，2，3情况下各种方法变量选择的模拟结果，表中字体倾斜的数据表示污染力度较小的结果. SCAD-NR表示方程(4)中 $\varphi(x)=x$ 且 $w_i=1$ 时的非稳健方法；SCAD-R (Huber)表示文献[15]中所提的基于Huber函数的SCAD稳健方法；SCAD-R (t_2)表示本文所提的基于 t_2 函数的SCAD稳健方法. 通过与非稳健方法的比较，凸显出本文所提方法与文献[15]中方法是稳健的. 在数据中存在异常值时，稳健估计结果明显好于非稳健的. 而通过本文方法与文献[15]中方法比较，模拟结果证明本文所提的方法在变量选择上的效果明显优于基于Huber函数的方法.

表1 污染1情况下变量选择的模拟结果

Tab.1 Simulation results of variable selection under pollution 1

方法	拟合不足率	取对率	拟合过度率	取对个数	取错个数
SCAD-NR	0.475 0	0.000 0	0.525 0	2.605 0	0.540 0
	0.190 0	*0.000 0*	*0.810 0*	*2.885 0*	*0.195 0*
SCAD-R(Huber)	0.485 0	0.210 0	0.305 0	3.865 0	0.550 0
	0.190 0	*0.270 0*	*0.540 0*	*3.975 0*	*0.195 0*
SCAD-R (t_2)	0.475 0	0.485 0	0.040 0	4.735 0	0.540 0
	0.190 0	*0.795 0*	*0.015 0*	*4.905 0*	*0.195 0*

表2 污染2情况下变量选择的模拟结果

Tab.2 Simulation results of variable selection under pollution 2

方法	拟合不足率	取对率	拟合过度率	取对个数	取错个数
SCAD-NR	0.000 0	0.235 0	0.765 0	3.525 0	0.000 0
	0.000 0	*0.235 0*	*0.765 0*	*3.525 0*	*0.000 0*
SCAD-R(Huber)	0.000 0	0.425 0	0.575 0	3.900 0	0.000 0
	0.000 0	*0.425 0*	*0.575 0*	*3.900 0*	*0.000 0*
SCAD-R(t_2)	0.000 0	0.995 0	0.005 0	4.995 0	0.000 0
	0.000 0	*0.995 0*	*0.005 0*	*4.995 0*	*0.000 0*

由表1～3可见，在不同污染力度和不同污染方式下，非稳健方法选择的效果总体来说都没有稳健方法好. 非稳健方法拟合过度率比两种稳健的方法都高，也就是说，非稳健方法总会把 $\boldsymbol{\beta}$ 中原本为0的估计为非0. 由表中模拟结果可见，基于 t 函数的稳健方法取对率总是最高的. 无论是哪种污染方式，本文所提的方法与非稳健方法和文献[15]中的方法相比，取对率都远大于其他两种方法，而且取对个数也明显大于其他两种方法. 虽然在拟合不足率上看不出本文所提方法的优势，但是，这里几种方法的拟合不足率都非常接近，而在正确率、拟合过度率和取对率方面，却明显可以看出本文所提方法的优势. 可见，本文方法对异常值的抵抗力比文献[15]中基于Huber函数的稳健方法更强，大大减少异常值在模型估计中引起的偏差.

表 3　污染 3 情况下变量选择的模拟结果

Tab.3　Simulation results of variable selection under pollution 3

方法	拟合不足率	取对率	拟合过度率	取对个数	取错个数
SCAD-NR	0.410 0	0.000 0	0.590 0	2.655 0	0.475 0
	0.120 0	*0.000 0*	*0.880 0*	*3.040 0*	*0.120 0*
SCAD-R（Huber）	0.330 0	0.290 0	0.380 0	4.080 0	0.395 0
	0.110 0	*0.425 0*	*0.465 0*	*4.155 0*	*0.110 0*
SCAD-R（t_2）	0.400 0	0.595 0	0.005 0	4.830 0	0.465 0
	0.120 0	*0.880 0*	*0.000 0*	*4.910 0*	*0.120 0*

表 4 是在 2 种污染力度下第 1 种污染方式对 $\boldsymbol{\beta}$ 中非 0 分量的估计结果，统计了各非 0 分量的偏差、方差和均方误差. 从该结果中可见，两种稳健方法在估计方面也明显优于非稳健方法，本文方法此时较文献[15]并没有很明显的优势. 在偏差的第 1 个分量和方差的第 2,3 个分量甚至稍微变大，但均方误差却稍有优势. 相比较来说，基于 t_2 函数的均方误差比 Huber 函数的稍小.

表 4　污染 1 情况下估计的模拟结果

Tab.4　Simulation results of estimation under pollution 1

方法	偏差			方差			均方误差		
SCAD-NR	−1.045 3	−1.324 7	−1.195 7	0.680 1	1.008 5	1.121 6	1.555 3	2.771 9	2.687 6
	−0.955 5	*−0.875 4*	*−0.858 3*	*0.659 9*	*0.883 1*	*0.908 1*	*1.348 5*	*1.546 2*	*1.561 3*
SCAD-R（Huber）	−0.283 5	−1.115 2	−0.968 8	0.590 1	1.084 5	1.067 9	0.428 6	2.419 7	2.079 0
	−0.333 8	*−0.591 6*	*−0.614 6*	*0.449 7*	*0.839 5*	*0.712 2*	*0.313 6*	*1.054 6*	*0.885 0*
SCAD-R（t_2）	0.325 2	−0.855 1	−0.574 8	0.647 5	1.228 2	1.200 7	0.525 1	2.239 7	1.772 1
	0.076 2	*−0.354 6*	*−0.230 0*	*0.423 0*	*0.897 7*	*0.703 9*	*0.184 8*	*0.931 5*	*0.548 4*

综合表 1～4，在非零参数估计上本文方法与文献[15]中方法相比具有较小的均方误差，同时本文方法在变量选择方面的稳健性和优势较为突出，其对异常值的抵抗力比文献[15]中方法更强.

5　结　论

本文在前人研究的基础上提出基于 t_2 函数的稳健变量选择方法，并与文献[15]中稳健方法进行比较，通过数值模拟来验证本文方法的有效稳健性. 文中首先详细叙述了关于该方法的模型理论，在稳健估计方程中施加一个惩罚函数，来达到期望的稳健变量选择效果. 然后在第 2 部分进一步考察稳健方程中的有界得分函数，通过比较不同自由度的 t 函数与 Huber 函数的性质，初步判断方程(4)中的有界得分函数的稳健性. 第 3 部分介绍了本文所用的算法，采用牛顿迭代法. 最后通过第 4 部分的数值模拟来验证前文中预计的稳健性. 模拟结果体现了 t_2 函数在变量选择方面的明显优势，虽然参数估计的结果并不是明显好于 Huber 函数的结果，但是通过取对率、拟合不足率、拟合过度率、取对个数和取错个数体现的选择结果，说明 t_2 函数在变量选择上的稳健性优于 Huber 函数.

本文主要通过模拟来考察各种方法的优劣. 基于 t 函数的变量选择方法的大样本性质，以及将该方法应用到更复杂的纵向数据中，或应用到超高维的横截面数据中，这些问题还有待进一步研究.

参考文献：

[1] TIBSHIRANI R. Regression shrinkage and selection via the lasso: a retrospective[J]. Journal of the Royal Statistical Society, 2011, 73(3): 273－282.

[2] 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.

[3] ZOU H. The adaptive lasso and its oracle properties[J]. Journal of the American Statistical Association, 2006, 101(476): 1418－1429.

[4] ZOU H, HASTIE H. Regularization and variable selection via the Elastic Net[J]. Journal of the Royal

Statistical Society,2005,67(2):301－320.
[5] CANDÈS E,TAO T. Rejoinder: the Dantzig selector: statistical estimation when p is much larger than n [J]. The Annals of Statistics, 2007, 35 (6): 2392－2404.
[6] HUBER P J. Robust estimation of a location parameter [J]. The Annals of Mathematical Statistics, 1964, 35 (1):73－101.
[7] HUBER P J. Robust regression: asymptotics, conjectures and Monte Carlo [J]. The Annals of Statistics,1973,1(5):799－821.
[8] PORTNOY S,KOENKER R. The Gaussian hare and the Laplacian tortoise: computability of squared-error versus absolute-error estimators [J]. Statistical Science,1997,12(4):279－300.
[9] GILONI A,PADBERG M. Alternative methods of linear regression[J]. Mathematical and Computer Modelling, 2002,35(3/4):361－374.
[10] GILONI A,PADBERG M. The finite sample breakdown point of L_1-regression[J]. Journal on Optimization, 2004,14(4):1028－1042.
[11] HE X M, SIMPSON D G, WANG G Y. Breakdown points of t-type regression estimators[J]. Biometrika, 2000,87(3):675－687.
[12] HE X M,FUNG W K,ZHU Z Y. Robust estimation in generalized partial linear models for clustered data [J]. Journal of the American Statistical Association, 2005,100(472):1176－1184.
[13] SINHA S K. Robust inference in generalized linear models for longitudinal data[J]. The Canadian Journal of Statistics,2006,34(2):261－278.
[14] MARONNA R A, MARTIN R D, YOHAI V J. Robust statistics: theory and methods [M]. Chichester, England:John Wiley and Sons,2006.
[15] 樊亚莉,徐群芳.稳健的变量选择方法及其应用[J].上海理工大学学报,2013,35(3):256－260.
[16] SCHUMANN D H. Robust variable selection [D]. Carolina:North Carolina State University,2009.

(编辑:丁红艺)

(上接第541页)

[3] 林杰,林新棋.线性变换的张量积[J].福建广播电视大学学报,1998(2):37－40.
[4] ROTMAN J J. An introduction to homological algebra [M]. New York:Academic Press,1960.
[5] DUMMIT D S, FOOTE R M. Abstract algebra [M]. Hoboken:John Wiley and Sons,2004.
[6] GLASBY S P. On the tensor product of polynomials over a ring[J]. Journal of the Australian Mathematical Society,2001,71(3):307－324.
[7] 闫爱民,胡建华.型E7根系的结构[J].上海理工大学学报,2014.36(1):5－11.
[8] 胡建华,赵卫萍.一类特殊幂零李代数的结构[J].上海理工大学学报,2015.37(3):215－219.

(编辑:石 瑛)