doi:10. 16055/j. issn. 1672-058X. 2022. 0002. 010

基于弹性网约束的稳健变量选择

魏双微
(重庆师范大学 数学科学学院,重庆 401331)

摘 要:大数据时代下收集到的数据常含有异常值或呈现尖峰厚尾以及变量之间具有较强的相关性,针对此问题,结合秩回归和自适应弹性网(Adaptive Elastic-net)提出了一种高效稳健的变量选择方法。此方法的最大优点在于不仅能够有效处理协变量之间的强相关性而且还能克服多重共线性问题,同时能抵抗厚尾分布或异常值的影响,实现稳健的变量选择。在数值计算方面,采用二次近似和牛顿迭代算法以获得新变量选择方法的稳定数值解,仿真实验表明:新提出的方法比现有方法表现更好,特别是对于厚尾分布或异常值的情况。最后,通过对中国重要的股票市场指数——中证 100 指数的跟踪,进一步表明该方法在有效样本下具有良好的表现。

关键词:秩回归;弹性网约束;稳健估计;变量选择;中证 100 指
中图分类号:O213. 9 **文献标志码**:A **文章编号**:1672-058X(2022)02-0068-07

0 引 言

当前社会数据海量、信息纷繁,如何从广大数据中寻找出有效信息已经成为学者们正在探讨的话题。因此,国内外学者先后提出了赤池信息准则、贝叶斯信息准则、广义交叉验证等方法来解决这一难题。这些方法不但缺乏稳定性,而且在自变量个数较多时还会耗费巨大的计算成本。寻找一种新的理论方法来解决高维数据的信息提取已然成为研究者们迫切需要解决的问题。Tibsniran[1]在惩罚函数的启发下提出了 LASSO(Least Absolute Shrinkage and Selection Operator)估计,该方法在惩罚参数的合理选择范围内可以压缩某些分量至零以实现变量选择,并进行参数估计;Fan 等[2]提出了惩罚似然函数的变量选择方法。然而,已有的文献大多是基于极大似然或最小平方进行研究和分析的,所得估计并不稳健。此外,这些方法不仅对于异常值很敏感,而且当误差为厚尾分布时估计效率会大大降低。因此,研究高维数据下估计方法更为稳健和有效就显得尤为重要。

Jaeckel[3]提出秩回归(Rank Regression)估计,其具备良好的稳健性和有效性;Wang 等[4]结合加权 SCAD(Smoothly Clipped Absolute Deviation)惩罚将秩估计推广到了固定维数参数模型下,并已证明该方法具有 Oracle 性质(即模型选择的相合性、参数估计渐近正态性);Wang 等[5]通过局部秩估计对 $\beta(\cdot)$ 的稳健推断问题进行了研究,结果表明:在误差是非正态分布情形下,此方法能够显著地改善经典局部最小二乘估计;Yang 等[6]基于 B

收稿日期:2021-03-05;**修回日期**:2021-05-18.
基金项目:国家社会科学基金(17CTJ015);重庆市基础科学与前沿研究技术专项项目(CSTC2018JCYJAX0659) .
作者简介:魏双微(1997—),女,重庆开州人,硕士研究生,从事应用统计研究.

样条基近似非参函数并利用 SCAD 罚函数惩罚秩回归,提出了一种新的稳健估计,此方法能够进行变量选择以及识别变系数与常系数;Kwessi[7]将秩估计引入半参数模型下,结合自适应 LASSO 惩罚表明在重尾分布下所得估计量是一致的,并给出了渐近正态性结果。

Zou 等[8]提出了弹性网方法,该方法可以处理协变量中出现的复共线性问题,其预测精度远远优于 Lasso;卢[9]将 Zou 等[8]的方法推广到了 Logistic 模型和 Poisson 模型中,证明该方法可将具有强相关性的变量全部选入模型或者剔除;黄[10]将 Zou 等[8]的方法推广到部分线性模型中,同时提出并证明其具有 Oracle 性质;在超高维数据下,Xiao 等[11]提出 MSA-Enet(Multi-step Adaptive Elastic Net)方法进行降维,其目的是让变量维数小于样本容量;李[12]将 Zou 等[8]的方法应用到平衡纵向数据模型的变量选择中,证明了该方法具有相合性和组效应性质;Li 等[13]将非负自适应弹性网估计推广到高维稀疏线性模型中,并在一些正则条件下证明了其 Oracle 性质和在有效样本下的有效性;王等[14]结合分位数回归和弹性网估计研究了基金绩效评价,且表明弹性网分位数回归比均值回归和 Lasso 分位数回归的评价更加准确。已有的研究已证明了弹性网约束良好的组效应性质,秩回归具有稳健性和有效性,因此如何将两者有效结合从而实现稳健变量选择是一个很有学术意义的问题。

Yang 等[15]结合秩回归与 SCAD 罚函数提出来一种稳健的变量选择方法,但当协变量中出现复共线性情形时,效果可能会受影响,因此如何在数据出现复共线性时,研究稳健的变量选择很有意义。在已有的研究成果中,弹性网估计方面的研究都是非稳健估计,秩回归方面的研究算法几乎都是采用 lars 算法,且从未与弹性网估计结合进行研究。本文将秩回归与弹性网相结合进行了研究,在响应变量含有异常值或重尾分布情况下,本文所提出的估计均具稳健性和有效性,且对强相关性数据的估计效果优于 Lasso 惩罚秩估计、惩罚分位数回归以及最小二乘估计。在算法上对损失函数和惩罚函数采用局部二次近似,使得目标函数能求出数值解,优化其迭代算法。

1　模型简述

考虑线性回归模型:

$$\boldsymbol{Y} = \alpha \boldsymbol{I}_n + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

其中,$\boldsymbol{Y} = (Y_1, Y_2, \cdots, Y_n)^{\mathrm{T}}$ 是 $n \times 1$ 维响应变量,α 是截距,$\boldsymbol{I}_n$ 是元素全是 1 的 $n \times 1$ 维向量,$\boldsymbol{X}$ 是 $n \times p$ 维协方差矩阵,且不丧失一般性,假设 $\boldsymbol{X}$ 中心化,$\boldsymbol{\beta}$ 是 $p \times 1$ 维未知参数,$\boldsymbol{\varepsilon}$ 是具有概率密度 $f(\cdot)$ 的独立同分布 $n \times 1$ 维随机误差向量。假设在真实模型中,$\boldsymbol{\beta}$ 的部分元素是零,本文的研究目标是实现零系数的识别和非零系数的稳健且有效估计。

1.1　秩回归

令 $e_i = y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}$,$i = 1, 2, \cdots, n$,初始估计量:

$$\tilde{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \frac{1}{n} \sum_{i<j} |e_i - e_j| \tag{1}$$

如式(1)所示,尽管可以得到参数估计的结果,但是不能把重要的协变量选择出来。Zou 等[8]提出了弹性网约束,能使部分参数压缩为零,实现变量选择。本文在式(1)基础上加入弹性网约束。

1.2　自适应弹性网秩回归

本文提出的自适应弹性约束秩回归指用自适应弹性网惩罚秩回归模型。Zou[16]对 L_1 惩罚部分进行加权,则惩罚函数的部分变为

$$\lambda_1 \sum_{j=1}^{p} v_j |\beta_j| + \lambda_2 \sum_{j=1}^{p} \beta_j^2$$

其中,$v_j = 1/|\tilde{\beta}_j|^{\gamma}$,$\gamma > 0$。记

$$\boldsymbol{C}(\boldsymbol{\beta}) = \frac{1}{n} \sum_{i<j} |e_i - e_j|$$

$$\boldsymbol{L}(\boldsymbol{\beta}) = n\lambda_1 \sum_{j=1}^{p} v_j |\beta_j| + n\lambda_2 \sum_{j=1}^{p} \beta_j^2$$

利用式(1)得到的 $\tilde{\boldsymbol{\beta}}$,最小化下面的目标函数,得到 $\boldsymbol{\beta}$ 的估计:

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\{\boldsymbol{C}(\boldsymbol{\beta}) + L(\boldsymbol{\beta})\} \tag{2}$$

式(2)称为自适应弹性约束秩回归(R-AEN)。

式(1)可以看作是 Jaeckel[17] 的 Wilcoxon 得分秩差分函数,基于文献[6],其中 $\boldsymbol{C}(\boldsymbol{\beta})$ 可有如下近似:

$$\frac{1}{n}\sum_{i<j}|e_i - e_j| \approx \sum_{i=1}^{n}\omega_i(e_i - \xi)^2 =$$

$$\sum_{i=1}^{n}\omega_i(y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta} - \xi)^2 \overset{\Delta}{=} \tilde{\boldsymbol{C}}(\boldsymbol{\beta})$$

其中,ξ 是 $\{e_i\}_{i=1}^{n}$ 的中位数,且

$$\omega_i = \begin{cases} \dfrac{R(e_i)}{n+1} - \dfrac{1}{2}, & e_i \neq \xi \\ \dfrac{e_i - \xi}{0,\text{其他}} \end{cases}$$

其中 $R(e_i)$ 是 e_i 的秩,$i = 1,2,\cdots,n$。

由此,目标函数式(2)可以变成如下形式:

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\{\tilde{\boldsymbol{C}}(\boldsymbol{\beta}) + \boldsymbol{L}(\boldsymbol{\beta})\} \tag{3}$$

用局部二次近似逼近罚函数的第一部分,得

$$|\beta_j| \approx \frac{|\beta_j|^2}{|\tilde{\beta}_j|} = \frac{\beta_j^2}{|\tilde{\beta}_j|}$$

其中,$\tilde{\beta}_j$ 是初始值的第 j 个元素。记

$$\boldsymbol{S} = \boldsymbol{Y} - \xi\boldsymbol{I}_{n\times 1}$$

$$\boldsymbol{W} = \operatorname{diag}(\omega_1,\omega_2,\cdots,\omega_n)$$

$$\boldsymbol{\Delta} = \operatorname{diag}\left(\frac{\lambda_1}{|\tilde{\beta}_1|^{\gamma+1}},\cdots,\frac{\lambda_1}{|\tilde{\beta}_p|^{\gamma+1}}\right)$$

$$\boldsymbol{D} = (\boldsymbol{S}-\boldsymbol{X}\boldsymbol{\beta})^{\mathrm{T}}\boldsymbol{W}(\boldsymbol{S}-\boldsymbol{X}\boldsymbol{\beta}) + n\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Delta}\boldsymbol{\beta} + n\lambda_2\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta}$$

如式(3)所示,可以近似成如下形式:

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\{\boldsymbol{D}\} \tag{4}$$

如式(4)所示,对 $\boldsymbol{\beta}$ 求导后令其为 0,得

$$-2\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}(\boldsymbol{S} - \boldsymbol{X}\boldsymbol{\beta}) + 2n(\boldsymbol{\Delta} + \lambda_2 I_{p\times p})\boldsymbol{\beta} = 0$$

则有

$$\hat{\boldsymbol{\beta}}_{\lambda_1} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{X} + n(\boldsymbol{\Delta} + \boldsymbol{\lambda}_2\boldsymbol{I}))^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{S}$$

1.3 调节参数选择

模型复杂度被参数 λ_1 与 λ_2 控制,λ_1 与 λ_2 在式(4)中起关键作用。参考 Li[13] 和 Zou[16],选取 $\gamma = 1$,$\lambda_2 = 0.01$,记 $\mu_i = y_i - \boldsymbol{x}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}_{\lambda_1}$,$i = 1,2,\cdots,n$,则通过最小化如下式子选择 λ_1:

$$\hat{\boldsymbol{\beta}}_{\lambda_1} = \log\left(n^{-2}\sum_{i<j}|\mu_i - \mu_j|\right) + df_{\lambda_1}\log n/n$$

其中,$\hat{\boldsymbol{\beta}}_{\lambda_1}$ 是在给定 λ_1 下弹性约束秩回归的估计值;df_{λ_1} 是 $\hat{\boldsymbol{\beta}}_{\lambda_1}$ 中非零元的个数。

1.4 算 法

基于以上讨论,可将 EN-R 估计的求解算法概括为以下几个步骤:

步骤 1 给定初始值 $\boldsymbol{\beta}^m(m=0)$,初始值可以由式(1)得到;

步骤 2 在当前估计值 $\boldsymbol{\beta}^{(m)}$ 下,利用 $\boldsymbol{\beta}^{(m+1)} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{X} + n\boldsymbol{\Delta} + \lambda_2\boldsymbol{I})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{S}|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(m)}}$,得到新估计值 $\boldsymbol{\beta}^{(m+1)}$;

步骤 3 迭代步骤 2 直至算法收敛。在实际操作过程中,当 $\|\boldsymbol{\beta}^{(m+1)} - \boldsymbol{\beta}^{(m)}\| < 10^{-6}$ 时,停止迭代。

2 模拟和数据分析

2.1 模拟研究

在 Tibshirani[1] 和 Fan 等[18] 文献中,数据来自于

$$y_i = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \varepsilon_i, i = 1,2,\cdots,n$$

其中,$\boldsymbol{\beta} = (3,1.5,3,0,0,0,0,0)$,$\boldsymbol{x}_i = (x_{i1},\cdots x_{in})^{\mathrm{T}} \sim N_n(0,\Omega)$。$\Omega$ 的第 (i,j) 个元素为 $\sigma^{|i-j|}$,$1 \leqslant i,j \leqslant n$。本文考虑 $N(0,1)$ 和自由度为 3 的 t 分布 $t(3)$,以及混合正态分布(MN):$0.9N(0,1) + 0.1N(0,25)$ 3 种误差分布。对于每种情况,分别进行 200 次模拟。为了度量估计精确度,通常采用计算 MSE 值:$F_{\mathrm{MSE}} = E(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^{\mathrm{T}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$。同时,变量选择性能由 N_{C},N_{IC} 及 N_{CF} 评判,其中,N_{C} 表示被正确估计的零系数数目,N_{IC} 表示系数中非零被估计为零的数目,N_{CF} 表示系数中零与非零同时被正确识别的比例。为方便,本文将提出的自适应弹性网秩回归估计记为 R-AEN,自适应弹性网最小二乘估计记为 LS-AEN,自适应弹性网分位数估计记为 QR-AEN,L_1 惩罚的秩回归估计记为 R-L_1,真实模型下所得估计记为 Oracle,考虑 $\sigma = 0.5$ 和 $\sigma = 0.7$ 两种情形。

表 1　$\sigma=0.5$ 下各方法的模拟结果

Table 1　Simulation results of all methods for $\sigma=0.5$

样本容量	误差分布	方　法	N_C	N_{IC}	N_{CF}	F_{WSE} (×100)
	$N(0,1)$	LS-AEN	4.760	0	0.910	5.104
		QR-AEN	4.860	0	0.865	4.601
		R-L_1	3.945	0	0.455	6.509
		R-AEN	4.995	0	0.995	2.753
		Oracle	5	0	1	2.663
200	t_3	LS-AEN	4.585	0	0.805	9.709
		QR-AEN	4.555	0	0.615	9.281
		R-L_1	3.865	0	0.445	12.098
		R-AEN	5	0	1	3.709
		Oracle	5	0	1	3.474
	MN	LS-AEN	4.73	0	0.860	7.546
		QR-AEN	4.845	0	0.860	6.739
		R-L_1	3.985	0	0.475	7.979
		R-AEN	4.995	0	0.995	3.698
		Oracle	5	0	1	2.716
	$N(0,1)$	LS-AEN	4.965	0	0.990	1.817
		QR-AEN	4.995	0	0.995	1.895
		R-L_1	4.020	0	0.500	3.449
		R-AEN	5	0	1	1.309
		Oracle	5	0	1	1.162
400	t_3	LS-AEN	4.786	0	0.910	4.815
		QR-AEN	4.820	0	0.820	6.665
		R-L_1	4.095	0	0.560	5.035
		R-AEN	5	0	1	2.050
		Oracle	5	0	1	1.876
	MN	LS-AEN	4.880	0	0.945	3.572
		QR-AEN	4.910	0	0.915	3.359
		R-L_1	4.085	0	0.520	4.194
		R-AEN	5	0	1	1.675
		Oracle	5	0	1	1.452

表 2　$\sigma=0.7$ 下各方法的模拟结果

Table 2　Simulation results of all methods for $\sigma=0.7$

样本容量	误差分布	方　法	N_C	N_{IC}	N_{CF}	F_{WSE} (×100)
	$N(0,1)$	LS-AEN	4.830	0	0.855	12.088
		QR-AEN	4.905	0.005	0.935	11.401
		R-L_1	2.880	0	0.160	9.740
		R-AEN	4.990	0	0.990	4.131
		Oracle	5	0	1	3.922
200	t_3	LS-AEN	4.555	0	0.670	35.405
		QR-AEN	4.690	0	0.745	22.682
		R-L_1	2.390	0	0.085	13.845
		R-AEN	5	0	1	6.327
		Oracle	5	0	1	5.811
	MN	LS-AEN	4.705	0	0.750	22.465
		QR-AEN	4.960	0	0.965	11.105
		R-L_1	2.690	0	0.140	9.898
		R-AEN	4.995	0	0.990	3.536
		Oracle	5	0	1	3.202
	$N(0,1)$	LS-AEN	4.920	0	0.945	6.273
		QR-AEN	4.990	0	0.990	3.257
		R-L_1	3.275	0	0.295	3.928
		R-AEN	5	0	1	1.919
		Oracle	5	0	1	1.849
400	t_3	LS-AEN	4.775	0	0.820	16.373
		QR-AEN	4.980	0	0.980	10.839
		R-L_1	3.140	0	0.275	6.817
		R-AEN	5	0	1	2.978
		Oracle	5	0	1	2.944
	MN	LS-AEN	4.710	0	0.770	12.037
		QR-AEN	4.940	0	0.942	6.365
		R-L_1	3.275	0	0.265	4.496
		R-AEN	5	0	1	1.971
		Oracle	5	0	1	1.949

由表 1 和表 2 可知,所提出的 R-AEN 估计相比其他 3 种方法表现更好,特别是对于厚尾(t_3)或异常值(混合正态)。从模型复杂度方面看,所提出的方法 N_C 很大,随着样本量的增加很快地接近 5,N_{IC} 接近 0,N_{CF} 接近 1,证实了所提方法能稳健有效地识别零和非零系数。Oracle 和 R-AEN 的 MSE 值很接近,并且随着样本量的增大越来越接近,说明所提方法的模型选择结果几乎接近于真实情况。随着样本量的增大,所有方法的 MSE 值越来越小,证明所有方法是相合估计。另外,R-AEN 处理强相关变量具有更好的稳健性和显著性。综上所述,新方法能同时实现高效、稳健的模型选择,并且处理强相关性数据的能力相对更好。

2.2 中证 100 指数据分析

本节重点讨论 R-AEN 在金融市场中的应用:追踪中证 100 指数的表现。

指数追踪是良好的资产配置方法,该方法利用部分成分股复制目标指数的表现。此外,由于成分股与目标指数之间存在复共线性,因此本文的指数追踪用自适应弹性网秩回归方法进行研究。

所用数据来自西南证券金点子财富管理终端,包含 2020-09-28—2020-12-22 的中证 100 指数以及所有成分股 30 min 线收盘价,共 919 个观测值,100 个协变量,能有效解决中证 100 指数的成分股半年更新导致变量发生改变以及样本量 $n < p$ 的问题。按指数与时间的关系将数据集分为两个部分,训练集为前 1/3 部分数据,测试集为余下 2/3 部分数据,由此建立高维模型。

分析过程中,令 x_{ij} 表示第 j 只成分股在第 i 次观测时的收盘价,y_i 表示第 i 次观测时的中证 100 指数。通常可用如下线性模型描述 x_{ij} 与 y_i 之间的关系:

$$y_i = \sum_{j=1}^{100} x_{ij}\hat{\beta}_j + \varepsilon_i, i = 1,2,\cdots,313$$

研究者在股指追踪过程中为使成本最低,希望通过成分股中较小的子集追踪指数的表现,因此需要进行变量选择。在使用自适应弹性网秩回归过程中,依然固定惩罚参数 $\gamma = 1$ 和 $\lambda_2 = 0.01$,然后利用 BIC 准则选定 λ_1,由此选出合适数量的成分股 44 只。根据选好的协变量在训练集上建立模型,分别计算训练集(U)和测试集(O)上的 MSE 值,内预测误差(ISPE):$F_{ISPE} = \sum_{t\in u}(y_t - \hat{y}_t)^2/|U|$;外预测误差(OSPE):$F_{OSPE} = \sum_{t\in O}(y_t - \hat{y}_t)^2/|O|$。$|\cdot|$ 表示模型指数个数。考虑同时应用 R-AEN,LS-AEN,QR-AEN 及 R-L_1 4 种方法在训练集和测试集上分别计算出 MSE 值,结果如表 3 和表 4 所示。图 1 给出了测试集上各方法的中证 100 指数收盘价的预测值。

表 3 各种方法的内预测误差(F_{ISPE})

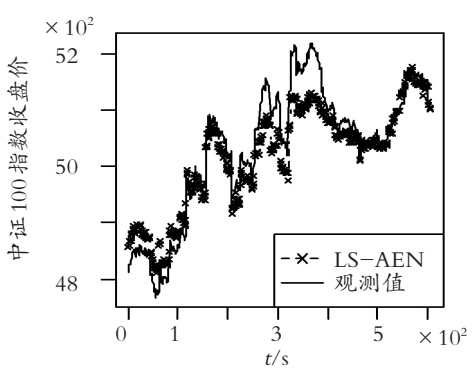
Table 3 Internal prediction error (F_{ISPE}) of various methods

LS-AEN	QR-AEN	R-L_1	R-AEN
714.955	644.520	983.508	611.042

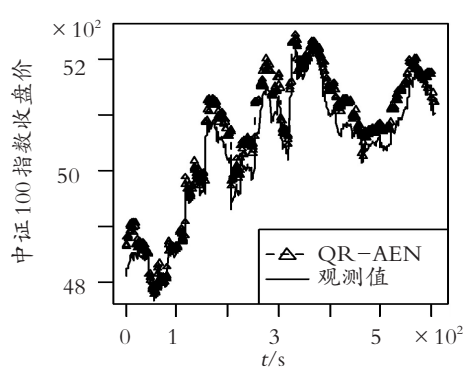
表 4 各种方法的外预测误差(F_{OSPE})

Table 4 External prediction error (F_{OSPE}) of various methods

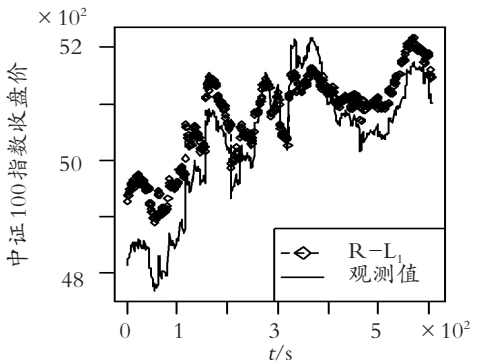
LS-AEN	QR-AEN	R-L_1	R-AEN
882 664.4	673 309.5	2 351 449.0	180 693.3



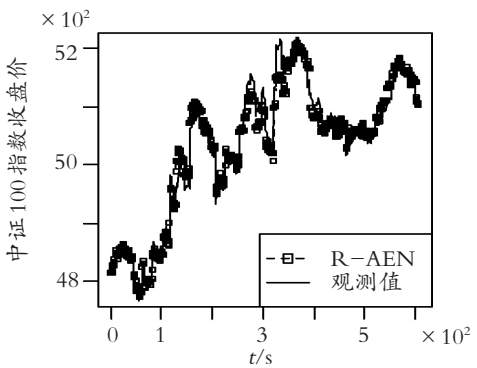
(a) LS-AEN 方法



(b) QR-AEN 方法



(c) R−L_1 方法



(d) R−AEN 方法

图 1　各种方法在测试集上得到的预测值

Fig. 1　The predicted values obtained by various methods on the test set

一种估计方法 F_{ISPE} 和 F_{OSPE} 越小,说明此方法的预测精度越高。由表 3 和表 4 知, R−AEN 方法的内预测误差为 611.042,明显小于其余 3 种方法;R−AEN 方法的外预测误差为 180 693.3,明显小于其余 3 种方法。说明 R−AEN 方法所选模型预测效果最佳。图 1 可以直观地看出: R−AEN 方法值曲线与观测值曲线更接近,说明预测效果最佳,同时也说明,R−AEN 方法对重要协变量的选择更加准确。

3　结　论

本文基于自适应弹性网和秩估计提出了稳健且有效的变量选择方法。通过数值模拟分析所得结论表明:当数据含有异常值或厚尾分布,或协变量具有强相关性时,所提 R−AEN 估计比现有方法更稳健和有效。本文仅从弹性网约束秩回归方面对变量选择进行了研究,关于弹性网约束秩回归中调节参数的选择还可进行研究。

参考文献(References):

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

[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): 1348—1360.

[3] JAECKEL L A. Estimating regression coefficients by minimizing the dispersion of residuals[J]. The Annals of Mathematical Statistics, 1972(43): 1449—1458.

[4] WANG L, LI R. Weight Wilcoxon-type smoothly clipped absolute deviation method[J]. Biometrics, 2009(65): 564—571.

[5] WANG H S, XIA Y C. Shrinkage estimation of the varying coefficient model[J]. Journal of the American Statistical Association, 2009(104): 747—757.

[6] YANG H, LV J, GUO C H. Robust variable selection and parametric component identification in varying coefficient models[J]. Communications in Statistics: Theory and Methods, 2016(45): 5533—5549.

[7] KWESSI E. Double penalized semi-parametric signed-rank regression with adaptive lasso[J]. Journal of Systems Science & Complexity, 2021(34): 381—401.

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

[9] 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.

[10] 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.

[11] XIAO N, XU Q. Multi-step adaptive elastic-net: reducing false positives in high-dimensional variable selection[J]. Journal of Statistical Computation and Simulation, 2015(85): 1—11.

[12] 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): 5—10.

LI Hong-xuan. Elastic net method for variable selection of balanced longitudinal data model [J]. Journal of Taishan University, 2017(39):5—10.

[13] LI N, YANG H, YANG J. Nonnegative estimation and variable selection via adaptive elastic-net for high-dimensional data [J]. Communications in Statistics: Simulation and Computation, 2019(7):1—17.

[14] 王文胜,宋家辉.基于弹性网分位数回归的开放型基金绩效研究[J].数理统计与管理,2020(39):721—733.
WANG Wen-sheng, SONG Jia-hui. Research on open fund performance based on elastic network quantile regression[J]. Mathematical Statistics and Management, 2020(39):721—733.

[15] YANG H, GUO C H, LYU J. SCAD penalized rank regression with a diverging number of parameters [J]. Journal of Multivariate Analysis,2015(133):321—333.

[16] ZOU H. The adaptive Lasso and its oracle properties[J]. Journal of the American Statistical Association, 2006(101):1419—1429.

[17] JAECKEL L A. Estimating regression coefficients by minimizing the dispersion of the residuals[J]. The Annals of Mathematical Statistics,1972(43):1449—1458.

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

Robust Variable Selection Based on Elastic Network Constraint

WEI Shuang-wei

(School of Mathematical Science, Chongqing Normal University, Chongqing 401331, China)

Abstract: In the era of big data, the collected data often contain outliers or present peak and thick-tails and strong correlations between variables. To solve this problem, an efficient and robust variable selection method combining rank regression and Adaptive Elastic Net is proposed. The greatest advantage of this method is that it can not only effectively deal with the strong correlation among concomitant variables but also overcome the multicollinearity problem, and it can resist the influence of thick-tailed distribution or outliers to achieve robust variable selection. In the aspect of numerical calculation, quadratic approximation and Newton iterative algorithm are used to obtain stable numerical solutions of the new variable selection method. Simulation results show that the proposed method performs better than the existing methods, especially for thick-tailed distributions or outliers. Finally, through the tracking of CSI 100, an important stock market index in China, it is further demonstrated that this method has a good performance under effective samples.

Key words: rank regression; elastic net constraint; robust estimation; variable selection; CSI 100

责任编辑:李翠薇

引用本文/Cite this paper:

魏双微.基于弹性网约束的稳健变量选择及在中证100指的应用研究[J].重庆工商大学学报(自然科学版),2022,39(2):68—74.

WEI Shuang-wei. Robust variable selection based on elastic network constraint[J]. Journal of Chongqing Technology and Business University (Natural Science Edition),2022,39(2):68—74.