山西大学学报(自然科学版)49(5):831—838,2026
Journal of Shanxi University(Nat. Sci. Ed.)

DOI:10.13451/j.sxu.ns.2024023

面板数据固定效应线性回归模型的变量选择

田果[1,2],杨宜平[1,2*]
(1.重庆工商大学 数学与统计学院,重庆 400067;
2.统计智能计算与监测重庆市重点实验室,重庆 400067)

摘　要:本文基于弹性网(Elastic net)惩罚为面板数据固定效应线性回归模型提出了一种新的变量选择方法,该方法能处理强相关变量的变量选择问题。先通过正向正交偏差变换消除固定效应,然后构造了基于Elastic net惩罚的最小二乘目标函数,进而同时进行回归系数的估计和变量选择,证明了所得估计的组效应性质。所提出的变量选择过程不仅消除了固定效应的影响,而且即使协变量存在共线性时,也能很好地识别出真实模型。其次,模拟研究了本文提出方法的有限样本性质。模拟研究表明:与最小绝对收缩和选择算子(Least Absolute Shrinkage and Selection Operator,Lasso),自适应最小绝对收缩和选择算子(Adaptive Least Absolute Shrinkage and Selection Operator,ALasso)以及平滑截断绝对偏差惩罚(Smoothly Clipped Absolute Deviation, SCAD)方法相比,提出的Elastic net变量选择方法能够很好地将强相关变量选入模型。最后,采用文中提出的方法分析了实际数据。

关键词:强相关变量;正向正交偏差变换;弹性网;组效应
中图分类号:O212.7　　**文献标志码**:A　　**文章编号**:0253-2395(2026)05-0831-08

Variable Selection of Linear Panel Data Models with Fixed Effects

Tian Guo[1,2], Yang Yiping[1,2*]
(1. *School of Mathematics and Statistics, Chongqing Technology and Business University, Chongqing* 400067*, China;*
2. *Chongqing Key Laboratory of Statistical Intelligent Computing and Monitoring, Chongqing* 400067*, China*)

Abstract: Based on Elastic net, a new variable selection method is proposed in this paper for linear panel data models with fixed effects. The variable selection problem with the strongly correlated variables can be dealt by the proposed variable selection method. Firstly, the fixed effect is eliminated by the forward orthogonal deviation transformation. Then, the penalized least squares objective function based on the Elastic net penalty is constructed, which can estimate regression coefficients and select important variables, simultaneously. The group effect property of the obtained estimation is proved. The influence of fixed effects is eliminated by the proposed variable selection method and the real model can be identified well even if the covariates are collinear. Secondly, the finite sample properties of the proposed method are accessed by the simulation. The simulation results show that compared with least absolute shrinkage and selection operator (Lasso), adaptive least absolute shrinkage and selection operator (ALasso) and smoothly clipped absolute deviation (SCAD) methods, strongly correlated variables can be better selected by the proposed Elastic net variable selection method. Finally, a real dataset is analyzed by the proposed method.

Keywords: strongly correlated variables; the forward orthogonal deviations transform; Elastic net; grouping effect

收稿日期:2023-09-13;**修回日期**:2024-03-07
基金项目:国家社会科学基金(18BTJ035);重庆市自然科学基金(cstc2021jcyj-msxmX0079);重庆市社科规划委托项目(2019WT58)
作者简介:田果(1999—),男,重庆酉阳人,硕士研究生,研究方向为大数据分析。E-mail:1148854212@qq.com
* **通信作者**:杨宜平(Yang Yiping),E-mail:yeepingyang@foxmail.com
引文格式:田果,杨宜平.面板数据固定效应线性回归模型的变量选择[J].山西大学学报(自然科学版),2026,49(5):831-838. DOI:10.13451/j.sxu.ns.2024023.

0 引言

在科学研究的过程中经常会产生面板数据。面板数据是多个个体在不同时间点上进行重复测度,得到每个个体在不同时间点上的观测值。因此,面板数据同时具有截面和时间双重的特性,是截面与时间的结合。面板数据广泛应用于经济学、生物学、医学、管理学等领域。由于信息技术不断发展,数据变得越来越复杂,对于面板数据而言,如何从众多变量中选取出重要的变量,成为当前统计学研究的热点问题之一。

目前,关于惩罚变量选择的方法已经有很多。Tibshirani[1]在惩罚函数的启发下,提出了针对普通线性模型下的Lasso(Least Absolute Shrinkage and Selection Operator)变量选择方法,它可以将无显著影响的变量系数压缩为0,并且同时进行变量选择和系数的估计。为了有效地解决Lasso的计算问题,Efron等[2]提出了最小角回归算法(Least Angle Regression,Lars),这种方法显著提高了Lasso方法的计算速度,使Lasso方法在模型估计和变量选择方面得到更广泛的应用。由于Lasso变量选择方法不具有Oracle性质,Zou[3]提出了ALasso(Adaptive Least Absolute Shrinkage and Selection Operator)变量选择方法,ALasso不仅具有Lasso方法在系数压缩时的相同力度,而且ALasso具有Oracle性质。一方面,Fan和Li[4]改进了Lasso方法,提出了具有Oracle性质的SCAD(Smoothly Clipped Absolute Deviation)惩罚变量选择方法。另一方面,Zou和Hastie[5]结合岭回归和Lasso变量选择方法,提出了Elastic net变量选择方法,该变量选择方法相比于Lasso,它的优点在于可以处理变量之间强相关的数据。因此,Elastic net变量选择方法被广泛应用。舒时克和李路[6]建立逻辑回归模型,采用Elastic net变量选择方法研究上证180指数成分股数据的选股问题。王文胜和宋家辉[7]基于Elastic net分位数回归对开放性基金绩效评价进行了研究。Huang等[8]将Elastic net变量选择方法推广到部分线性模型,并讨论了变量选择的组效应性质。唐礼智等[9]将Elastic net方法用于面板数据的贝叶斯分位数回归,并用所提出的方法分析了互联网金融类上市公司经济增加值数据。Shi等[10]将Elastic net变量选择方法应用到了广义线性模型。Li和Ding[11]针对Elastic net惩罚提出了广义条件梯度算法。邹璟婧等[12]结合QR分解和Elastic net惩罚方法提出了一种线性混合效应模型中固定效应的变量选择方法。由于面板数据从截面和时间两个维度分析实际问题,已有学者讨论了面板数据固定效应线性回归模型的变量筛选。

关于高维面板数据模型的研究,李扬和曾宪斌[13]针对面板数据模型的惩罚似然变量选择问题,给出了Lasso、ALasso、岭回归和SCAD四种惩罚函数的渐近性质。朱利荣等[14]针对面板数据模型提出了ALasso的复合分位数回归估计。Li等[15]采用SCAD惩罚变量选择方法研究了变系数固定效应模型的变量选择问题。已有针对高维面板数据模型的变量筛选,不能有效地处理相关性较强的变量。上述文献关于面板数据线性回归模型的变量选择并不能有效地处理相关性强的变量。为了解决该问题,本文运用Elastic net惩罚,消除固定效应的影响构造一种能处理高度相关变量的变量选择过程。首先引入正向正交偏差变换矩阵,消除固定效应项;然后结合Elastic net惩罚构造惩罚目标函数,进而同时估计回归系数并进行变量选择。进一步,证明了所提出估计的组效应性质。最后,模拟研究了所提出方法的有限样本性质。本文提出的方法具有两个优势:一、通过采用正向正交偏差变换的方法消除固定效应,避免了误差项序列相关。二、利用Elastic net惩罚有效地处理了强相关变量的变量筛选问题,所提出的方法具有组效应。

1 参数的变量选择

考虑如下面板数据固定效应模型

$$Y_{it}=\boldsymbol{X}_{it}^{\mathrm{T}}\boldsymbol{\beta}+c_i+\varepsilon_{it},\ i=1,2,\cdots,n;\ t=1,2,\cdots,T, \tag{1}$$

其中Y_{it}是响应变量,$\boldsymbol{\beta}$是p维未知参数,$\boldsymbol{X}_{it}$是p维协变量,ε_{it}是模型误差,c_i是不可观测的个体固定

效应项。为了模型可识别,假定 $\sum_{i=1}^{n} c_i = 0$。

令 $\boldsymbol{Y}_i = (Y_{i1}, Y_{i2}, \cdots, Y_{it})^{\mathrm{T}}$, $\boldsymbol{X}_i = (\boldsymbol{X}_{i1}, \boldsymbol{X}_{i2}, \cdots, \boldsymbol{X}_{iT})^{\mathrm{T}}$, $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \varepsilon_{i2}, \cdots, \varepsilon_{it})^{\mathrm{T}}$, $\boldsymbol{C}_i = \boldsymbol{1}_T c_i$, $\boldsymbol{1}_T$ 是全为 1 的 T 维向量。则(1)式可写为

$$\boldsymbol{Y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{C}_i + \boldsymbol{\varepsilon}_i, i = 1, 2, \cdots, n。 \tag{2}$$

由于 c_i 是个体固定效应,如果忽略 c_i 直接估计 $\boldsymbol{\beta}$ 获得的估计是有偏的。即如果直接估计 c_i,随着 n 的增加待估计参数 c_i 增加,会造成“维数祸根”问题。一个有效的办法就是消除 c_i。因此,本文借鉴了 Arellano[16] 的思想,通过正向正交偏差变换消除固定效应项。给出 $(T-1)\times T$ 维的滤子矩阵 $\boldsymbol{D}$,其形式为

$$\boldsymbol{D} = \begin{pmatrix} -1 & 1 & 0 & \cdots & 0 & 0 \\ 0 & -1 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & 0 \\ 0 & 0 & 0 & \cdots & -1 & 1 \end{pmatrix}。$$

在(2)式两边乘以矩阵 $(\boldsymbol{D}\boldsymbol{D}^{\mathrm{T}})^{-1/2}\boldsymbol{D}$,得到

$$\tilde{\boldsymbol{Y}}_i = \tilde{\boldsymbol{X}}_i\boldsymbol{\beta} + \tilde{\boldsymbol{\varepsilon}}_i, i = 1, 2, \cdots, n, \tag{3}$$

其中 $\tilde{\boldsymbol{Y}}_i = (\boldsymbol{D}\boldsymbol{D}^{\mathrm{T}})^{-1/2}\boldsymbol{D}\boldsymbol{Y}_i$, $\tilde{\boldsymbol{X}}_i = (\boldsymbol{D}\boldsymbol{D}^{\mathrm{T}})^{-1/2}\boldsymbol{D}\boldsymbol{X}_i$, $\tilde{\boldsymbol{\varepsilon}}_i = (\boldsymbol{D}\boldsymbol{D}^{\mathrm{T}})^{-1/2}\boldsymbol{D}\boldsymbol{\varepsilon}_i$。式(3)通过正向正交偏差变换消除了固定效应。避免了所转换的误差项的序列相关性,即

$$\mathrm{Var}(\boldsymbol{\varepsilon}_i) = \sigma^2\boldsymbol{I}_T \Rightarrow \mathrm{Var}(\tilde{\boldsymbol{\varepsilon}}_i) = \sigma^2\boldsymbol{I}_{T-1}。$$

通过变换后,样本量 nT 变成了 $n(T-1)$,我们记 $N = n(T-1)$。

令 $\tilde{\boldsymbol{Y}} = (\tilde{\boldsymbol{Y}}_1, \tilde{\boldsymbol{Y}}_2, \cdots, \tilde{\boldsymbol{Y}}_n)^{\mathrm{T}}$, $\tilde{\boldsymbol{X}} = (\tilde{\boldsymbol{X}}_1, \tilde{\boldsymbol{X}}_2, \cdots, \tilde{\boldsymbol{X}}_n)^{\mathrm{T}}$, $\tilde{\boldsymbol{\varepsilon}} = (\tilde{\boldsymbol{\varepsilon}}_1, \tilde{\boldsymbol{\varepsilon}}_2, \cdots, \tilde{\boldsymbol{\varepsilon}}_n)^{\mathrm{T}}$。对于模型(3),我们可以构造基于弹性网惩罚最小二乘估计,即

$$\hat{\boldsymbol{\beta}} = \arg\min\left\{\sum_{i=1}^{n}\left(\tilde{\boldsymbol{Y}}_i - \tilde{\boldsymbol{X}}_i\boldsymbol{\beta}\right)^{\mathrm{T}}\left(\tilde{\boldsymbol{Y}}_i - \tilde{\boldsymbol{X}}_i\boldsymbol{\beta}\right) + \lambda_2\left|\boldsymbol{\beta}\right|_2^2 + \lambda_1\left|\boldsymbol{\beta}\right|_1\right\}, \tag{4}$$

其中 λ_1, λ_2 是非负调节参数,$\left|\boldsymbol{\beta}\right|_2^2 = \sum_{j=1}^{p}\beta_j^2$, $\left|\boldsymbol{\beta}\right|_1 = \sum_{j=1}^{p}\left|\beta_j\right|$。

给出数据集 $(\tilde{\boldsymbol{Y}}, \tilde{\boldsymbol{X}})$ 和 (λ_1, λ_2),定义一个新的数据集 $(\tilde{\boldsymbol{Y}}^*, \tilde{\boldsymbol{X}}^*)$,其中

$$\tilde{\boldsymbol{X}}^*_{(N+p)\times p} = (1+\lambda_2)^{-1/2}\begin{pmatrix} \tilde{\boldsymbol{X}} \\ \sqrt{\lambda_2}\,\boldsymbol{I} \end{pmatrix}, \ \tilde{\boldsymbol{Y}}^*_{(N+p)} = \begin{pmatrix} \tilde{\boldsymbol{Y}} \\ \boldsymbol{0} \end{pmatrix},$$

令 $\gamma = \lambda_1/\sqrt{1+\lambda_2}$, $\boldsymbol{\beta}^* = \sqrt{1+\lambda_2}\,\boldsymbol{\beta}$,弹性网惩罚最小二乘的解等价于以下 Lasso 方法的解

$$\hat{\boldsymbol{\beta}}^* = \arg\min\left\{\left(\tilde{\boldsymbol{Y}}^* - \tilde{\boldsymbol{X}}^*\boldsymbol{\beta}^*\right)^{\mathrm{T}}\left(\tilde{\boldsymbol{Y}}^* - \tilde{\boldsymbol{X}}^*\boldsymbol{\beta}^*\right) + \gamma\left|\boldsymbol{\beta}^*\right|_1\right\}。 \tag{5}$$

从而

$$\hat{\boldsymbol{\beta}} = \frac{1}{\sqrt{1+\lambda_2}}\hat{\boldsymbol{\beta}}^*。$$

因此,目标函数(4)式的算法问题得到有效解决。注意到弹性网惩罚最小二乘目标函数涉及两个调节参数 λ_1, λ_2 的选取,为了避免二维最优化问题,类似 Li 等[17] 在模拟研究中,固定调节参数 $\lambda_2 = 0.01$,通过贝叶斯信息准则(Bayesian Information Criterion, BIC)选取 λ_1,即:

$$\mathrm{BIC}(\lambda_N) = \log(\hat{\sigma}_{\lambda_N}) + \frac{\log N}{N}\mathrm{d}f_{\lambda_N},$$

其中 $\hat{\sigma}_{\lambda_N} = \frac{1}{N}\sum_{i=1}^{n}\sum_{t=1}^{T-1}(\tilde{Y}_{it} - \tilde{\boldsymbol{X}}_{it}^{\mathrm{T}}\hat{\boldsymbol{\beta}})^2$,第一项控制模型拟合的好坏,第二项控制模型的稀疏性,df_{λ_N} 是 $\hat{\boldsymbol{\beta}}$ 中的非零系数个数,称为自由度。$\tilde{Y}_{it}$, $\tilde{\boldsymbol{X}}_{it}^{\mathrm{T}}$ 分别是经过正向正交偏差变换消除固定效应项后的响应

变量和协变量。

2 估计的组效应

现在给出弹性网方法的组效应性质,即将强相关变量全部选入模型或者全部剔出模型。

定理1 给出数据$(\tilde{\boldsymbol{X}},\tilde{\boldsymbol{Y}})$和参数$(\lambda_1,\lambda_2)$,响应变量$\tilde{\boldsymbol{Y}}$是经过中心化的,解释变量$\tilde{\boldsymbol{X}}$是经过标准化的。令$\hat{\boldsymbol{\beta}}(\lambda_1,\lambda_2)$是弹性网的估计。假设$\hat{\beta}_l(\lambda_1,\lambda_2)\hat{\beta}_k(\lambda_1,\lambda_2)>0$。定义

$$D_{\lambda_1,\lambda_2}(l,k)=\frac{1}{\left|\tilde{\boldsymbol{Y}}\right|_1}\left|\hat{\beta}_l(\lambda_1,\lambda_2)-\hat{\beta}_k(\lambda_1,\lambda_2)\right|,$$

则有

$$D_{\lambda_1,\lambda_2}(l,k)\leqslant\frac{1}{\lambda_2}\sqrt{2(1-\rho)},$$

其中$\rho=\tilde{\boldsymbol{X}}'^{\mathrm{T}}_l\tilde{\boldsymbol{X}}'_k$是样本相关系数,$\tilde{\boldsymbol{X}}'_l$和$\tilde{\boldsymbol{X}}'_k$是矩阵$\tilde{\boldsymbol{X}}$的第$l$列和第$k$列。

证明 令$\hat{\beta}_l(\lambda_1,\lambda_2)\hat{\beta}_k(\lambda_1,\lambda_2)>0$,则有$\operatorname{sgn}\left\{\hat{\beta}_l(\lambda_1,\lambda_2)\right\}=\operatorname{sgn}\left\{\hat{\beta}_k(\lambda_1,\lambda_2)\right\}$,且二者非零,其中$\operatorname{sgn}(\cdot)$是符号函数。由$\hat{\boldsymbol{\beta}}=\arg\min L(\lambda_1,\lambda_2,\boldsymbol{\beta})$,则$\hat{\beta}(\lambda_1,\lambda_2)$满足

$$\frac{\partial L(\lambda_1,\lambda_2,\boldsymbol{\beta})}{\partial\beta_k}=0。$$

所以有

$$\sum_{i=1}^{n}\sum_{t=1}^{T-1}\left\{(-2)\tilde{X}_{itk}(\tilde{Y}_{it}-\sum_{j=1}^{p}\tilde{X}_{itj}\hat{\beta}_j(\lambda_1,\lambda_2))\right\}+\lambda_1\operatorname{sgn}\left\{\hat{\beta}_k(\lambda_1,\lambda_2)\right\}+2\lambda_2\hat{\beta}_k(\lambda_1,\lambda_2)=0。\tag{6}$$

又有

$$\sum_{i=1}^{n}\sum_{t=1}^{T-1}\left\{(-2)\tilde{X}_{itl}(\tilde{Y}_{it}-\sum_{j=1}^{p}\tilde{X}_{itj}\hat{\beta}_j(\lambda_1,\lambda_2))\right\}+\lambda_1\operatorname{sgn}\left\{\hat{\beta}_l(\lambda_1,\lambda_2)\right\}+2\lambda_2\hat{\beta}_l(\lambda_1,\lambda_2)=0。\tag{7}$$

(7)减(6)得

$$\sum_{i=1}^{n}\sum_{t=1}^{T-1}\left\{(-2)(\tilde{X}_{itl}-\tilde{X}_{itk})(\tilde{Y}_{it}-\sum_{j=1}^{p}\tilde{X}_{itj}\hat{\beta}_j(\lambda_1,\lambda_2))\right\}+2\lambda_2(\hat{\beta}_l(\lambda_1,\lambda_2)-\hat{\beta}_k(\lambda_1,\lambda_2))=0。\tag{8}$$

整理(8)得

$$\hat{\beta}_l(\lambda_1,\lambda_2)-\hat{\beta}_k(\lambda_1,\lambda_2)=\frac{1}{\lambda_2}\sum_{i=1}^{n}\sum_{t=1}^{T-1}\left\{(\tilde{X}_{itl}-\tilde{X}_{itk})(\tilde{Y}_{it}-\sum_{j=1}^{p}\tilde{X}_{itj}\hat{\beta}_j(\lambda_1,\lambda_2))\right\}=\frac{1}{\lambda_2}\left(\tilde{\boldsymbol{X}}'^{\mathrm{T}}_l-\tilde{\boldsymbol{X}}'^{\mathrm{T}}_k\right)\hat{\gamma}(\lambda_1,\lambda_2),$$

其中$\hat{\gamma}(\lambda_1,\lambda_2)=\tilde{\boldsymbol{Y}}-\tilde{\boldsymbol{X}}\hat{\boldsymbol{\beta}}(\lambda_1,\lambda_2)$,$\tilde{\boldsymbol{X}}'_l$和$\tilde{\boldsymbol{X}}'_k$是矩阵$\tilde{\boldsymbol{X}}$的第$l$列和第$k$列。由$\hat{\beta}=\arg\min L(\lambda_1,\lambda_2,\boldsymbol{\beta})$知

$$L(\lambda_1,\lambda_2,\boldsymbol{\beta})\leqslant L(\lambda_1,\lambda_2,0),$$

等价于

$$\sum_{i=1}^{n}\sum_{t=1}^{T-1}(\tilde{Y}_{it}-\sum_{j=1}^{p}\tilde{X}_{itj}\beta_j)^2+\lambda_2\left|\beta\right|^2+\lambda_1\left|\beta\right|_1\leqslant\sum_{i=1}^{n}\sum_{t=1}^{T-1}\tilde{Y}_{it}^2,$$

即

$$\hat{\gamma}^2(\lambda_1,\lambda_2)+\lambda_2\left|\boldsymbol{\beta}\right|_2^2+\lambda_1\left|\boldsymbol{\beta}\right|_1\leqslant\sum_{i=1}^{n}\sum_{t=1}^{T-1}\tilde{Y}_{it}^2,$$

则可以得到$\hat{\gamma}^2(\lambda_1,\lambda_2)\leqslant\left|\tilde{\boldsymbol{Y}}\right|^2$,其中$\left|\tilde{\boldsymbol{Y}}\right|=\sqrt{\sum_{i=1}^{n}\sum_{t=1}^{T-1}\tilde{Y}_{it}^2}$,令$\left|\tilde{\boldsymbol{Y}}\right|_1=\sum_{i=1}^{n}\sum_{j=1}^{T-1}\left|\tilde{Y}_{it}\right|$,易得$\left|\tilde{\boldsymbol{Y}}\right|\leqslant\left|\tilde{\boldsymbol{Y}}\right|_1$。

即

$$\left|\hat{\gamma}(\lambda_1,\lambda_2)\right|\leqslant\left|\tilde{\boldsymbol{Y}}\right|。$$

则有

$$D_{\lambda_1,\lambda_2}(l,k)=\frac{1}{\left|\tilde{\boldsymbol{Y}}\right|_1}\left|\hat{\beta}_l(\lambda_1,\lambda_2)-\hat{\beta}_k(\lambda_1-\lambda_2)\right|\leqslant\frac{1}{\lambda_2}\frac{\left|\hat{\gamma}(\lambda_1,\lambda_2)\right|}{\left|\tilde{\boldsymbol{Y}}\right|}\left|\tilde{\boldsymbol{X}}_l'-\tilde{\boldsymbol{X}}_k'\right|\leqslant\frac{1}{\lambda_2}\sqrt{2(1-\rho)}\text{。}$$

因此,定理1得证。

注 定理1给出了弹性网的组效应性质,$D_{\lambda_1,\lambda_2}(l,k)$展示了两个变量$\tilde{\boldsymbol{X}}_l$和$\tilde{\boldsymbol{X}}_k$的系数路径之间的差异。如果$\tilde{\boldsymbol{X}}_l$与$\tilde{\boldsymbol{X}}_k$强相关,即若$\rho=1$,两者的系数路径之间的差异几乎为零。因此,可以看出弹性网变量选择方法可以有效处理强相关变量,而Lasso变量选择不能处理强相关变量。

3 模拟研究

本节通过模拟研究所提出方法的有限样本性质。模拟研究考虑了两种情形:协变量不存在完全相关变量和存在完全相关变量。为了度量变量选择和估计的精度,我们计算了非零变量错误估计为0的平均个数,记为"*IC*";零变量正确估计为0的平均个数,记为"*C*";以及均方误差(Mean Squared Error, MSE),MSE的定义为:

$$\mathrm{MSE}\left(\hat{\beta}\right)=\left(\hat{\beta}-\beta\right)^{\mathrm{T}}\left(\hat{\beta}-\beta\right)\text{。}$$

为了比较,模拟研究考虑了弹性网、Lasso、ALasso和SCAD四种变量选择方法,其中Lasso和ALasso利用R软件中的包"lar",SCAD采用包"ncvreg"。样本容量分别取$n=50$,100和150,每种情况重复运行1 000次。由于面板数据是不同个体不同时间上的观测,观测次数为$T=5$,个体效应满足$\sum_{i=1}^{n}c_i=0$。

模拟1 协变量不存在完全相关的情形。数据由以下模型产生

$$Y_{it}=\boldsymbol{X}_{it}^{\mathrm{T}}\beta+c_i+\varepsilon_{it},\ i=1,2,\cdots,n;\ t=1,\cdots,5,$$

其中$\beta=(3,1.5,1.5,0,\cdots,0)^{\mathrm{T}}$,维数$p=10$,协变量$\boldsymbol{X}=(\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_{10})$产生均值为零,协方差矩阵为$\mathrm{Cov}\left(\boldsymbol{X}_i,\boldsymbol{X}_j\right)=\rho^{|i-j|}$的多元正态分布,$\rho$的取值为0.5,$c_i=X_{i\cdot,1}+V_i,i=2,\cdots,n$,其中$V_i\sim N(0,1)$,$X_{i\cdot,1}=\frac{1}{5}\sum_{t=1}^{5}X_{it,1}$,且$c_1=-\sum_{i=2}^{n}c_i$,模型误差$\varepsilon_{it}\sim 0.5N(0,1)$。计算结果见表1。

表1 协变量不存在完全相关下不同样本容量变量选择和拟合结果

Table 1 Variable selection and fitting results of different sample sizes without complete correlation of covariates

n	方法	MSE	*IC*	*C*
n=50	Lasso	0.009 1	0	6.445
	Elastic net	0.009 1	0	6.424
	ALasso	0.006 4	0	6.445
	SCAD	0.008 0	0	6.445
n=100	Lasso	0.004 4	0	6.589
	Elastic net	0.004 4	0	6.570
	ALasso	0.003 0	0	6.589
	SCAD	0.004 1	0	6.396
n=150	Lasso	0.003 0	0	6.637
	Elastic net	0.003 0	0	6.630
	ALasso	0.002 0	0	6.637
	SCAD	0.002 6	0	6.459

从表1可以得到以下结论:

(1)随着样本量的增加,弹性网、Lasso、ALasso和SCAD四种变量选择的方法的MSE都减小。

(2)Elastic net和Lasso变量选择的方法的MSE相差不大,并且IC和C也相差不大,说明在协变

量不存在完全相关时,弹性网和Lasso两种方法效果相当。

(3)由于ALasso和SCAD两种方法具有Oracle性质,弹性网和Lasso两种方法不具有Oracle性质,所以ALasso和SCAD两种方法的MSE更小。

模拟2 协变量之间存在完全相关的情形。数据由以下模型产生

$$Y_{it}=\boldsymbol{X}_{it}^{\mathrm{T}}\boldsymbol{\beta}+c_i+\varepsilon_{it},\ i=1,2,\cdots,n;\ t=1,\cdots,5,$$

其中$\boldsymbol{\beta}=(3,1.5,1.5,0,\cdots,0)^{\mathrm{T}}$,维数$p=10$,协变量$\boldsymbol{X}=(\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_{10})$,其中产生$\boldsymbol{X}_2\sim N(0,1)$,变量$\boldsymbol{X}_2=\boldsymbol{X}_3$(协变量完全相关),其余变量$\boldsymbol{X}_i$产生均值为零,协方差矩阵为$\mathrm{Cov}(\boldsymbol{X}_i,\boldsymbol{X}_j)=\rho^{|i-j|}$的多元正态分布,$\rho$为0.5,$c_i=X_{i\cdot,1}+V_i,i=2,\cdots,n$,其中$V_i\sim N(0,1)$,$X_{i\cdot,1}=\frac{1}{5}\sum_{t=1}^{5}X_{it,1}$,且$c_1=-\sum_{i=2}^{n}c_i$,模型误差$\varepsilon_{it}\sim 0.5N(0,1)$。计算结果见表2。

表2 协变量存在完全相关下不同样本容量变量选择和拟合结果

Table 2 Variable selection and fitting results of different sample sizes under complete correlation of covariates

n	方法	MSE	IC	C
n=50	Lasso	4.379 5	1	6.495
	Elastic net	0.006 1	0	6.448
	ALasso	4.501 2	1	6.495
	SCAD	4.506 7	1	6.410
n=100	Lasso	4.406 6	1	6.590
	Elastic net	0.003 0	0	6.573
	ALasso	4.497 3	1	6.590
	SCAD	4.500 0	1	6.462
n=150	Lasso	4.423 2	1	6.644
	Elastic net	0.002 1	0	6.629
	ALasso	4.500 4	1	6.644
	SCAD	4.502 4	1	6.436

从表2可以看出:Lasso、ALasso和SCAD三种方法的IC值都为1,说明三种方法都将某个非零系数估计为零,并且导致了MSE过大。而弹性网方法的IC值为0,这表明了弹性网方法具有组效应,能够同时将强相关变量一起选入模型。

4 实例分析

本节分析一组来自北卡罗来纳州的犯罪数据。该数据可以从R软件包“plm”获得。该组数据采集了90个观测单位的数据,每个单位进行7次观测(每年观测一次),总共630个数据。本文构建如下面板数据模型分析该数据集,即

$$Y_{it}=\beta_1X_{it,1}+\beta_2X_{it,2}+\cdots+\beta_{16}X_{it,16}+\beta_{17}X_{it,17}+C_i+\varepsilon_{it},$$

其中Y_{it},X_{it}代表的变量在表3中给出。

类似模拟研究,采用四种方法分析了该数据集。为了比较四种方法的优劣,我们也计算了均方误差(MSE),其定义为:

$$\mathrm{MSE}=\frac{1}{630}\sum_{i=1}^{90}\sum_{t=1}^{7}(Y_{it}-\hat{\beta}_1X_{it,1}+\hat{\beta}_2X_{it,2}+\cdots+\hat{\beta}_{16}X_{it,16}+\hat{\beta}_{17}X_{it,17})^2,$$

其中$\hat{\beta}$为对应方法的系数估计值。计算结果见表4。

从表4可以看出,Lasso和SCAD从17个变量中选出了10个重要变量,ALasso和弹性网从17个变量中选出了7个重要变量。然而,与Lasso、ALasso和SCAD相比,弹性网给出了最小的MSE。

表3 变量含义及其符号

Table 3 Meanings and symbols of variables

变量	变量含义	符号	变量	变量含义	符号
lcrmrte	人均犯罪对数	Y_{it}	lwtrd	批发零售业周薪对数	$X_{it,9}$
lprbarr	被捕概率对数	$X_{it,1}$	lwfir	金融、保险和房地产周薪对数	$X_{it,10}$
lpolpc	人均警察对数	$X_{it,2}$	lwser	服务业周薪对数	$X_{it,11}$
lprbconv	定罪概率对数	$X_{it,3}$	lwmfg	制造业周薪对数	$X_{it,12}$
lprbpris	服刑概率对数	$X_{it,4}$	lwfed	联邦政府周薪对数	$X_{it,13}$
lavgsen	平均刑期对数	$X_{it,5}$	lwsta	州政府周薪对数	$X_{it,14}$
ldensity	每平方英里人口对数	$X_{it,6}$	lwloc	地方政府周薪对数	$X_{it,15}$
lwcon	建筑业周薪对数	$X_{it,7}$	lpctymle	年轻男性百分比对数	$X_{it,16}$
lwtuc	公共事业周薪对数	$X_{it,8}$	lpctmin	1980年少数民族百分比对数	$X_{it,17}$

表4 四种方法的系数估计值及MSE

Table 4 Coefficient estimates and MSE of the four methods

变量	$\hat{\beta}$(Lasso)	$\hat{\beta}$(ALasso)	$\hat{\beta}$(Elastic net)	$\hat{\beta}$(SCAD)
lprbarr	−0.292 5	−0.380 7	−0.394 9	−0.293 6
lpolpc	0.354 0	0.420 6	0.428 4	0.354 8
lprbconv	−0.235 0	−0.298 3	−0.308 3	−0.236 0
lprbpris	−0.125 9	−0.179 0	−0.197 4	−0.127 4
lavgsen	0	0	0	0
ldensity	0.100 1	0	0	0.091 7
lwcon	0	0	0	0
lwtuc	0.004 3	0	0	0.005 1
lwtrd	0	0	0	0
lwfir	0	0	0	0
lwser	0	0	0	0
lwmfg	−0.295 1	−0.326 9	−0.270 4	−0.279 3
lwfed	−0.456 8	−0.525 8	−0.450 9	−0.432 9
lwsta	0.049 9	0	0	0.053 5
lwloc	0.129 8	0.206 8	0.106 6	0.130 5
lpctymle	0	0	0	0
lpctmin	0	0	0	0
MSE	1.983 1	4.453 0	1.322 2	1.975 3

5 结束语

本文针对面板数据固定效应线性回归模型,提出了基于弹性网的变量选择过程。在构造弹性网惩罚目标函数时,为了消除固定效应,采用正向正交偏差变换的方法,避免了转换后的误差项存在序列相关性。通过最小化弹性网惩罚最小二次目标函数来实现系数压缩,将有些系数压缩为零。进一步,证明了所提出的变量选择过程的组效应性质。

参考文献:

[1] Tibshirani R. Regression Shrinkage and Selection via the Lasso[J]. *J R Stat Soc Ser B Methodol*, 1996, **58**(1): 267–288. DOI: 10.1111/j.2517-6161.1996.tb02080.x.

[2] Efron B, Hastie T, Johnstone I, *et al*. Least Angle Regression [J]. *Ann Statist*, 2004, **32**(2): 407–499. DOI: 10.1214/009053604000000067.

[3] Zou H. The Adaptive Lasso and Its Oracle Properties[J]. *J Am Stat Assoc*, 2006, **101**(476): 1418–1429. DOI: 10.1198/016214506000000735.

[4] Fan J Q, Li R Z. Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties[J]. *J Am Stat Assoc*, 2001, **96**(456): 1348–1360. DOI: 10.1198/016214501753382273.

[5] Zou H, Hastie T. Regularization and Variable Selection via the Elastic Net[J]. *J R Stat Soc Ser B Stat Methodol*, 2005, **67**(2): 301−320. DOI: 10.1111/j.1467-9868.2005.00503.x.

[6] 舒时克, 李路. 基于Elastic Net惩罚的多因子选股策略[J]. 统计与决策, 2021, **37**(16): 157−161. DOI: 10.13546/j.cnki.tjyjc.2021.16.036.
Shu S K, Li L. Multi-factor Stock Selection Strategy Based on Elastic Net Punishment[J]. *Stat Decis*, 2021, **37**(16): 157−161. DOI: 10.13546/j.cnki.tjyjc.2021.16.036.

[7] 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**(4): 721−733. DOI: 10.13860/j.cnki.sltj.20191021-003.
Wang W S, Song J H. Open-end Fund Investment Performance Based on Elastic Net Quantile Regression[J]. *J Appl Stat Manag*, 2020, **39**(4): 721−733. DOI: 10.13860/j.cnki.sltj.20191021-003.

[8] Huang D X, Li C H, Qin C Y, *et al*. The Application and Property of Elastic Net Procedure for Partially Linear Models[J]. *Chin Q J Math*, 2020, **35**(3): 290−301. DOI: 10.13371/j.cnki.chin.q.j.m.2020.03.004.

[9] 唐礼智, 李雨佳, 赵力静. 面板数据的贝叶斯Elastic Net分位数回归方法及其应用研究[J]. 统计研究, 2020, **37**(3): 94−113. DOI: 10.19343/j.cnki.11-1302/c.2020.03.008.
Tang L Z, Li Y J, Zhao L J. Study on the Bayesian Elastic Net Quantile Regression for Panel Data: Methods and Applications[J]. *Stat Res*, 2020, **37**(3): 94−113. DOI: 10.19343/j.cnki.11-1302/c.2020.03.008.

[10] Shi X Y, Liang B, Zhang Q. Post-selection Inference of Generalized Linear Models Based on the Lasso and the Elastic Net[J]. *Commun Stat Theory Meth*, 2022, **51**(14): 4739−4756. DOI: 10.1080/03610926.2020.1821892.

[11] Li H L, Ding L. Generalized Conditional Gradient Method for Elastic-net Regularization[J]. *J Comput Appl Math*, 2022, **403**: 113872. DOI: 10.1016/j.cam.2021.113872.

[12] 邹璟婧, 杨宜平, 赵培信. 基于正交投影的线性混合效应模型的弹性网变量选择[J]. 应用数学, 2024, **37**(2): 547−553. DOI: 10.13642/j.cnki.42-1184/o1.2024.02.004.
Zou J J, Yang Y P, Zhao P X. The Elastic Net Variable Selection for Linear Mixed-effects Models Based on the Orthogonal Projection[J]. *Mathematica Applicata*, 2024, **37**(2): 547−553. DOI: 10.13642/j. cnki. 42-1184/o1.2024.02.004.

[13] 李扬, 曾宪斌. 面板数据模型的惩罚似然变量选择方法研究[J]. 统计研究, 2014, **31**(3): 83−89. DOI: 10.19343/j.cnki.11-1302/c.2014.03.012.
Li Y, Zeng X B. Research on Variable Selection Method of Penalized Likelihood for Panel Data Model[J]. *Stat Res*, 2014, **31**(3): 83−89. DOI: 10.19343/j. cnki. 11-1302/c.2014.03.012.

[14] 朱利荣, 胡超竹, 罗幼喜. 面板数据模型的惩罚复合分位回归方法[J]. 统计与决策, 2022, **38**(13): 40−45. DOI: 10.13546/j.cnki.tjyjc.2022.13.008.
Zhu L R, Hu C Z, Luo Y X. A Penalized Composite Quantile Regression Method for Panel Data Model[J]. *Stat Decis*, 2022, **38**(13): 40−45. DOI: 10.13546/j. cnki. tjyjc.2022.13.008.

[15] Li G R, Lian H, Lai P, *et al*. Variable Selection for Fixed Effects Varying Coefficient Models[J]. *Acta Math Sin Engl Ser*, 2015, **31**(1): 91−110. DOI: 10.1007/s10114-015-3159-2.

[16] Arellano M. Panel Data Econometrics[M]. Oxford: Oxford University Press, 2003.

[17] Li N, Yang H, Yang J. Nonnegative Estimation and Variable Selection via Adaptive Elastic-net for High-dimensional Data[J]. *Commun Stat Simul Comput*, 2021, **50**(12): 4263−4279. DOI: 10.1080/03610918.2019.1642484.