正则稀疏化的多因子量化选股策略

舒时克,李 路
上海工程技术大学 数理与统计学院,上海 201620

摘 要:针对高维度数据集特征之间的复杂性,而传统的L1惩罚项不满足Oracle性质的无偏性,将逻辑回归弹性网(LR-Elastic Net)中的L1惩罚项替换为SCAD(Smoothly Clipped Absolute Deviation)和MCP(Minimax Concave Penalty)惩罚项,分别构建了LR-SCAD和LR-MCP模型,在保留稀疏性的同时满足了无偏性,并利用ADMM(Alternating Direction Method of Multipliers)算法进行求解。通过模拟实验发现,LR-Elastic Net模型能很好地处理特征存在相关性的小样本数据,而LR-SCAD和LR-MCP模型在特征存在相关性的大样本数据中表现较好;建立LR-Elastic Net、LR-SCAD和LR-MCP策略,并应用于沪深300指数成分股数据。回测结果显示,LR-SCAD和LR-MCP策略在股票相关性很强的数据中比LR-Elastic Net策略表现更好。

关键词:弹性网(Elastic Net);SCAD;MCP;ADMM算法;逻辑回归;多因子选股
文献标志码:A **中图分类号:**O212.1 doi:10.3778/j.issn.1002-8331.2002-0101

Multi-factor Quantitative Stock Selection Strategy Based on Sparsity Penalty

SHU Shike, LI Lu
School of Mathematics, Physics and Statistics, Shanghai University of Engineering Science, Shanghai 201620, China

Abstract: Aiming at the complexity between the characteristics of high-dimensional datasets. This paper proposes replace L1 penalty in LR-Elastic Net with SCAD(Smoothly Clipped Absolute Deviation) penalty and MCP(Minimax Concave Penalty), constructs LR-SCAD and LR-MCP models respectively, and uses ADMM(Alternating Direction Method of Multipliers)algorithm to solve. Simulation experiments show that LR-Elastic Net model is good at handling small sample data with correlation features, while LR-SCAD and LR-MCP models perform well in large sample data with correlation features. At the same time, the paper establishes LR-Elastic Net, LR-SCAD and LR-MCP strategies, and applies them to the data of the CSI 300 Index. Back-test results show that LR-SCAD and LR-MCP strategies perform better than LR-Elastic Net strategies in highly correlated data.

Key words: Elastic Net; Smoothly Clipped Absolute Deviation(SCAD); Minimax Concave Penalty(MCP); Alternating Direction Method of Multipliers(ADMM) algorithm; logistic regression; multi-factor stock selection

随着信息技术的发展,数据的规模越来越大,数据往往会出现维度较高而样本量较小的情况。因此,从众多的特征中选取出有效的特征就成为了一个难点。针对这种高维数据,目前的处理的方式大体分成两类:一类是从训练数据出发,通过特征工程等手段筛选特征,再通过模型进行预测;另一类是从模型本身出发,在模型中加入具有稀疏性质的惩罚项能够有效的筛选特征。经典的惩罚函数有L1惩罚项、L2惩罚项和Elastic Net惩罚项等。Jagnnathan[1]发现在线性回归中加入L1惩罚项的Lasso模型,从而建立更好的投资组合模型。但L1惩罚项存在过度稀疏的问题。针对L1惩罚函数的不足,Zou[2]在线性回归中同时加入L1和L2惩罚项,构建了弹性网模型(Elastic Net),并将其运用到高维数据上,该模型不仅能够克服了高维数据多重共线的问题,也克服了Lasso模型将特征压缩的过度稀疏的问题。文献[3]在对比了最小二乘(OLS)、Lasso和Elastic Net之后,应用于量化投资市场,发现Elastic Net模型能够比OLS模型和Lasso模型更有效的筛选因子,同时也能克服Lasso模型将系数矩阵过度压缩的缺点,并能构建出更加有效的投资组合。

文献[4]指出Lasso和Elastic Net的解虽然满足Oracle的稀疏性和连续性的假设,但是不满足无偏性的

基金项目:国家自然科学基金(11501055,11801362)。
作者简介:舒时克(1995—),男,硕士研究生,主要研究领域为机器学习与量化投资,E-mail:463927667@qq.com;李路(1968—),男,博士,副教授,主要研究领域为大数据计算。
收稿日期:2020-02-07 **修回日期:**2020-04-26 **文章编号:**1002-8331(2021)01-0110-08

性质,因此Fan等人提出了SCAD的惩罚函数,该惩罚项不仅满足Oracle的三个性质,并且也能对系数进行压缩。文献[5]提出了MCP惩罚函数,该惩罚项也满足Oracle的三个性质,而且能够很好的处理特征之间存在很高的相关性的数据。文献[6-7]表明Elastic Net、SCAD和MCP惩罚项在线性回归模型中取得很好的效果。

在分类问题中,逻辑回归作为一种统计分析方法,能够对分类问题进行有效的判别[8]。但是在高维数据中表现却不尽如人意。因此,为提高逻辑回归模型的分类性能,目前在逻辑回归模型中主要使用的惩罚函数有L1惩罚项、L2惩罚项和Elastic Net惩罚项[9]等。其中Elastic Net惩罚项结合了L1和L2惩罚项的优点,但不满足无偏性,即真实未知参数较大时,会产生较大的偏差。

因此,为处理特征之间复杂的关系,更好地筛选特征,本文在目前的逻辑回归弹性网(LR-Elastic Net)的基础上,将弹性网的L1惩罚项替换为SCAD和MCP惩罚项,分别构建LR-SCAD模型和LR-MCP模型。

1 逻辑回归弹性网

1.1 逻辑回归

逻辑回归作为一种统计分析方法,能够对分类的问题进行判别。设 $\boldsymbol{X}=(x_{ij})_{n\times p}\in R_{n\times p}$, x_{ij} 表示第 i 行数据的第 j 个特征的值,记 $\boldsymbol{x}_i=(x_{i1},x_{i2},\cdots,x_{ip})^{\mathrm{T}}$,表示第 i 行数据的全部特征值,则特征矩阵 $\boldsymbol{X}$ 为 $(x_1,x_2,\cdots,x_n)^{\mathrm{T}}$, $\boldsymbol{y}$ 为自变量,表示为 $(y_1,y_2,\cdots,y_n)^{\mathrm{T}}$,代表 $\boldsymbol{x}_i$ 的标签, $\boldsymbol{y}_i=1$ 或 0 ,则后验概率估计 $P(\boldsymbol{y}_i=1|\boldsymbol{x}_i)$ 和 $P(\boldsymbol{y}_i=0|\boldsymbol{x}_i)$ 可以表示为:

$$\begin{aligned}P(\boldsymbol{y}_i=1|\boldsymbol{x}_i)&=\frac{1}{1+\mathrm{e}^{-\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}_i}}\\P(\boldsymbol{y}_i=0|\boldsymbol{x}_i)&=1-\frac{1}{1+\mathrm{e}^{-\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}_i}}\end{aligned}\tag{1}$$

其中, $\boldsymbol{\beta}=(\beta_1,\beta_2,\cdots,\beta_p)^{\mathrm{T}}$ 是特征系数向量。则逻辑回归的目标函数可以表示为:

$$\begin{aligned}f(\boldsymbol{\beta})&=-\sum_{i=1}^{n}\left[\boldsymbol{y}_i\ln p(\boldsymbol{y}_i=1|\boldsymbol{x}_i)+(1-\boldsymbol{y}_i)\ln p(\boldsymbol{y}_i=0|\boldsymbol{x}_i)\right]=\\&-\sum_{i=1}^{n}\left[\boldsymbol{y}_i\ln p(\frac{1}{1+\mathrm{e}^{-\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}_i}})+(1-\boldsymbol{y}_i)\ln p(\frac{\mathrm{e}^{-\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}_i}}{1+\mathrm{e}^{-\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}_i}})\right]=\\&\sum_{i=1}^{n}\left[\ln(1+\mathrm{e}^{\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}_i})-y_i\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}_i\right]\end{aligned}\tag{2}$$

1.2 逻辑回归弹性网

在逻辑回归的交叉熵损失函数上加上弹性网惩罚项,构建为逻辑回归弹性网模型(LR-Elastic Net),该参数估计可以表示为:

$$\min f(\boldsymbol{\beta})+\lambda\left(\alpha\left\|\boldsymbol{\beta}\right\|_1+\frac{1}{2}(1-\alpha)\left\|\boldsymbol{\beta}\right\|_2^2\right)\tag{3}$$

其中, α 为惩罚项系数, $0\leqslant\lambda\leqslant1$ 。加入弹性网惩罚项之后,既能够筛选变量,将无关变量压缩到0,同时又能够避免特征系数向量过度稀疏。

2 正则稀疏化惩罚函数

Fan和Li [4]提出了Oracle性质来评判模型的优劣,主要包括三个性质:(1)稀疏性。模型中在估计参数时能将一些不重要的变量的系数压缩到零。(2)无偏性。模型中对估计的参数应该是无偏的或者是近似无偏的。(3)连续性。为了避免模型的不稳定性,参数估计与对应的系数应该是连续的。

而LR-Elastic Net中的惩罚项L1范数虽然满足Oracle的稀疏性和连续性,但是不满足无偏性[4],即当真实未知参数较大时,会产生较大的偏差。

2.1 SCAD惩罚函数

因此,Fan和Li[4]提出了SCAD惩罚函数来选择变量,并证明了该方法满足Oracle的三个性质。SCAD的惩罚函数为:

$$P_\lambda(\beta_j)=\begin{cases}\lambda|\beta_j|, & |\beta_j|\leqslant\lambda\\-\dfrac{|\beta_j|^2-2a\lambda|\beta_j|+\lambda^2}{2(a-1)}, & \lambda<|\beta_j|\leqslant a\lambda\\\dfrac{(a+1)\lambda^2}{2}, & |\beta_j|>a\lambda\end{cases}\tag{4}$$

其中 $a>2$,且Fan和Li[4]通过最小化贝叶斯风险值及蒙特卡洛模拟实验得出参数 a 的最优值约为3.7。SCAD惩罚函数的图像,如图1所示。

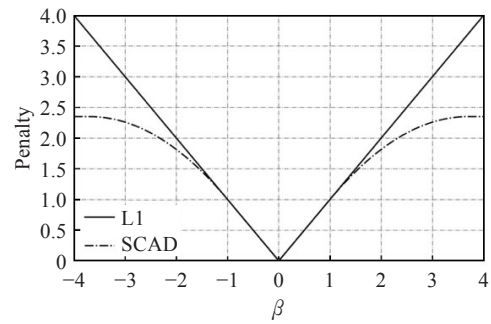

图1 SCAD惩罚函数

SCAD的惩罚函数导数为:

$$P_\lambda'(\beta_j)=\lambda\left(I(|\beta_j|\leqslant\lambda)+\frac{(a\lambda-\beta_j)_+}{(a-1)\lambda}I(|\beta_j|>\lambda)\right)\mathrm{sgn}(\beta_j)$$

即

$$P_\lambda'(\beta_j)=\begin{cases}\lambda, & |\beta_j|\leqslant\lambda\\\dfrac{2a\lambda-|\beta_j|}{(a-1)}, & \lambda<|\beta_j|\leqslant a\lambda\\0, & |\beta_j|>a\lambda\end{cases}\tag{5}$$

2.2 MCP惩罚函数

Zhang[6]提出了MCP惩罚函数,同样满足Oracle的三个性质,并且能够很好地处理特征之间存在很高的相关性的数据。

MCP惩罚函数为：

$$J(\beta_j,\lambda,\alpha)=\begin{cases}\lambda|\beta_j|-\dfrac{\beta_j^2}{2\alpha}, & |\beta_j|\leqslant\alpha\lambda\\ \dfrac{1}{2}\alpha\lambda, & |\beta_j|>\alpha\lambda\end{cases} \tag{6}$$

MCP的导数为：

$$J'(\beta_j,\lambda,\alpha)=\begin{cases}\lambda-\dfrac{\beta_j}{\alpha}, & 0<\beta_j\leqslant\alpha\lambda\\ -\lambda-\dfrac{\beta_j}{\alpha}, & -\alpha\lambda\leqslant\beta_j\leqslant 0\\ 0, & |\beta_j|>\alpha\lambda\end{cases} \tag{7}$$

MCP惩罚函数的图像，如图2所示。

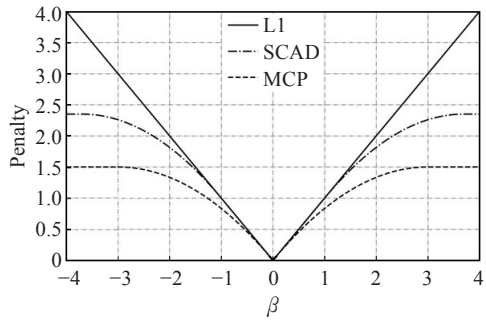

图2　MCP惩罚函数

如图2可见，MCP和SCAD惩罚函数相似，随着 β 的增加，惩罚力度逐渐减少，对回归系数采取有差别的惩罚，从而得到更加精确的估计[10]。

3　正则稀疏化逻辑回归

3.1　SCAD-逻辑回归

由于LR-ElasticNet中的L1惩罚项不满足Oracle无偏性的性质，为了能满足Oracle性质的稀疏性、无偏性和连续性，因此本文将逻辑回归弹性网模型中L1惩罚项替换为SCAD惩罚项，构建SCAD-逻辑回归模型（LR-SCAD），其目标函数可以表示为：

$$\min f(\boldsymbol{\beta})+\lambda_1\left(\alpha P_{\lambda_2}(\boldsymbol{\beta})+\frac{1}{2}(1-\alpha)\|\boldsymbol{\beta}\|_2^2\right) \tag{8}$$

其中，$f(\boldsymbol{\beta})$ 对 $\boldsymbol{\beta}$ 的一阶导数，二阶导数为：

$$\begin{cases}f'(\beta_0)=-\sum\limits_{i=1}^{n}\boldsymbol{x}_i\left(y_i-P(y_i=1|\boldsymbol{x}_i)\right)\\ f''(\beta_0)=\sum\limits_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}P(y_i=1|x_i)P(y_i=0|\boldsymbol{x}_i)\end{cases} \tag{9}$$

并且对 $f(\beta)$ 在 β_0 处泰勒展开：

$$f(\beta)=\sum_{i=1}^{n}\left[\ln(1+\mathrm{e}^{\beta^{\mathrm{T}}x_i})-y_i\beta^{\mathrm{T}}x_i\right]=\sum_{i=1}^{n}\left\{f(\beta_0)+f'(\beta_0)\|\beta-\beta_0\|+\frac{1}{2}f''(\beta_0)\|\beta-\beta_0\|_2^2+\varphi(\beta_0)\right\}=\frac{1}{2}\sum_{i=1}^{n}w_i\left(z_i-\beta^{\mathrm{T}}\boldsymbol{x}_i\right)^2+\varphi(\beta_0) \tag{10}$$

其中：

$$w_i=P(y_i=1|\boldsymbol{x}_i)P(y_i=0|\boldsymbol{x}_i)$$

$$z_i=\beta^{\mathrm{T}}\boldsymbol{x}_i+\frac{y_i-P(y_i=1|\boldsymbol{x}_i)}{P(y_i=1|\boldsymbol{x}_i)P(y_i=0|\boldsymbol{x}_i)}$$

令 $Z=(z_1,z_2,\cdots,z_n)^{\mathrm{T}}$，$\boldsymbol{W}=diag(w_1,w_2,\cdots,w_n)^{\mathrm{T}}$，则目标函数可以表示为：

$$\min\ \frac{1}{2}(\boldsymbol{X}\beta-Z)^{\mathrm{T}}W(\boldsymbol{X}\beta-Z)+\lambda_1\left(\alpha P_{\lambda_2}(\beta)+\frac{1}{2}(1-\alpha)\|\beta\|_2^2\right) \tag{11}$$

LR-SCAD的求解使用了交替方向乘子法ADMM算法[5]，ADMM算法结合了拉格朗日方法和对偶分解法的优点，通过增广拉格朗日函数构造，把原本复杂的高维问题分解成两个或者多个低维的更容易得到的全局解的交替极小化问题进行迭代求解，则LR-SCAD目标函数可以表示为：

$$\min\ \frac{1}{2}(\boldsymbol{X}\beta-Z)^{\mathrm{T}}W(\boldsymbol{X}\beta-Z)+\frac{1}{2}\lambda_1(1-\alpha)\|\beta\|_2^2+\lambda_1\alpha P_{\lambda_2}(\theta)\quad\text{subject to }\beta-\theta=0 \tag{12}$$

上式的增广拉格朗日方程为：

$$L(\beta,\theta,\mu)=\frac{1}{2}(\boldsymbol{X}\beta-Z)^{\mathrm{T}}W(\boldsymbol{X}\beta-Z)+\frac{1}{2}\lambda_1(1-\alpha)\|\beta\|_2^2+\lambda_1\alpha P_{\lambda_2}(\theta)+\mu(\beta-\theta)+\frac{\rho}{2}\|\beta-\theta\|_2^2 \tag{13}$$

其中，$\rho>0$ 为惩罚项系数，μ 是对偶变量，通过引入 θ 和 $\beta-\theta=0$ 的约束条件，简化了原问题的求解。变量迭代的规则如下：

$$\begin{cases}\beta^{k+1}=\arg\min\limits_{\beta}\left\{L(\beta,\theta^k,\mu^k)\right\}\\ \theta^{k+1}=\arg\min\limits_{\theta}\left\{L(\beta^{k+1},\theta,\mu^k)\right\}\\ \mu^{k+1}=\mu^k+\rho(\beta^{k+1}-\theta^{k+1})\end{cases} \tag{14}$$

更新 β：在第 $K+1$ 次的更新中，当 θ^k 和 μ^k 固定，需要通过求解 $\arg\min\limits_{\beta}\left\{L(\beta,\theta^k,\mu^k)\right\}$：

$$\min\ \frac{1}{2}(\boldsymbol{X}\beta-Z)^{\mathrm{T}}\boldsymbol{W}(\boldsymbol{X}\beta-Z)+\frac{1}{2}\lambda_1(1-\alpha)\|\beta\|_2^2+\mu(\beta-\theta)+\frac{\rho}{2}\|\beta-\theta\|_2^2 \tag{15}$$

对 β 求偏导并令其等于0，可以得到：

$$\frac{\partial}{\partial\beta}\frac{1}{2}(\boldsymbol{X}\beta-Z)^{\mathrm{T}}\boldsymbol{W}(\boldsymbol{X}\beta-Z)+\frac{1}{2}\lambda_1(1-\alpha)\|\beta\|_2^2+\mu(\beta-\theta)+\frac{\rho}{2}\|\beta-\theta\|_2^2=0 \tag{16}$$

化简可得：

$$\beta^{k+1}=\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{X}+\lambda_1(1-\alpha)+\rho I\right)^{-1}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}Z+\rho\theta-\mu\right) \tag{17}$$

更新 θ：在第 $K+1$ 次的更新中，当 β^{k+1} 和 μ^k 固定，需要通过求解 $\arg\min\limits_{\theta}\left\{L(\beta^{k+1},\theta,\mu^k)\right\}$：

$$\min\ \lambda_1\alpha P_{\lambda_2}(\theta)+\mu(\beta-\theta)+\frac{\rho}{2}\|\beta-\theta\|_2^2 \tag{18}$$

对 θ 求偏导并令其等于0,可以得到:

$$\frac{\partial}{\partial\theta}\lambda_1\alpha P_{\lambda_2}(\theta)+\mu(\beta-\theta)+\frac{\rho}{2}\|\beta-\theta\|_2^2=0 \tag{19}$$

化简可得:

$$\theta_j^{k+1}=\begin{cases}\frac{1}{\rho}(\rho\beta_j^{k+1}+\mu_j-\alpha\lambda_1\lambda_2), & |\theta_j^k|\leqslant\lambda_2\\ \frac{1}{\rho_1}\left[(a-1)\rho\beta_j^{k+1}+(a-1)\mu_j-2a\alpha\lambda_1\lambda_2\right], & \\ \qquad \lambda_2<|\theta_j^k|\leqslant a\lambda_2 & \\ \frac{1}{\rho}(\beta_j^{k+1}+\mu_j), & |\theta_j^k|>a\lambda_2\end{cases} \tag{20}$$

其中,$\rho_1=\frac{1}{(a-1)\rho-\alpha\lambda_1}$。

更新 μ:在第 $K+1$ 次的更新中,当 β^{k+1} 和 θ^{k+1} 固定,可以计算 μ^{k+1}:

$$\mu^{k+1}=\mu^k+\rho\left(\beta^{k+1}-\theta^{k+1}\right) \tag{21}$$

具体算法如下:

(1)随机初始化 β_{old},假设最终优化目标为 $F(\beta)$;

(2)在 β_{old} 处利用式(8)泰勒展开,得到 $f_{\text{old}}(\beta)$;

(3)利用式(17)(20)(21)迭代求得 $f_{\text{old}}(\beta)$ 的最优结果 β_{new};

(4)在 β_{new} 处利用式(10)继续泰勒展开,得到 $f_{\text{new}}(\beta)$;

(5)令 $\beta_{\text{old}}=\beta_{\text{new}}$,重复步骤(3)(4)直至收敛,最终得到解 β。

3.2 MCP-逻辑回归

同时,由于LR-Elastic Net中的L1惩罚项不满足Oracle无偏性的性质,为了能满足Oracle性质的稀疏性、无偏性和连续性,将逻辑回归弹性网模型中L1惩罚项替换为MCP惩罚项,构建MCP-逻辑回归模型(LR-MCP),其目标函数可以表示为:

$$\min f(\beta)+\lambda_1\left(\alpha_1 J(\beta,\lambda_2,\alpha_2)+\frac{1}{2}(1-\alpha)\|\beta\|_2^2\right) \tag{22}$$

通过泰勒展开LR-MCP模型为:

$$\min\ \frac{1}{2}(X\beta-Z)^{\mathrm{T}}W(X\beta-Z)+\alpha_1\lambda_1 J(\beta,\lambda_2,\alpha_2)+\frac{1}{2}\lambda_1(1-\alpha)\|\beta\|_2^2 \tag{23}$$

LR-MCP的求解同样使用ADMM算法,则LR-MCP模型的目标函数可以表示为:

$$\min\frac{1}{2}(X\beta-Z)^{\mathrm{T}}W(X\beta-Z)+\frac{1}{2}\lambda_1(1-\alpha)\|\beta\|_2^2+\lambda_1\alpha_1 J(\theta,\lambda_2,\alpha_2)$$
$$\text{subject to }\ \beta-\theta=0 \tag{24}$$

上式的增广拉格朗日方程为:

$$L(\beta,\theta,\mu)=\frac{1}{2}(X\beta-Z)^{\mathrm{T}}W(X\beta-Z)+\frac{1}{2}\lambda_1(1-\alpha)\|\beta\|_2^2+\lambda_1\alpha_1 J(\theta,\lambda_2,\alpha_2)+\mu(\beta-\theta)+\frac{\rho}{2}\|\beta-\theta\|_2^2 \tag{25}$$

参照LR-SCAD的求解方法,得出LR-MCP的迭代公式:

$$\begin{cases}\beta^{k+1}=\left(X^{\mathrm{T}}WX+\lambda_1(1-\alpha)+\rho I\right)^{-1}\left(X^{\mathrm{T}}WZ+\rho\theta-\mu\right)\\ \theta_j^{k+1}=\begin{cases}\frac{1}{\rho_1}(\rho\beta_j^{k+1}+\mu_j-\alpha_1\lambda_1\lambda_2), & 0<\theta_j^k\leqslant\alpha_2\lambda_2\\ \frac{1}{\rho_2}(\rho\beta_j^{k+1}+\mu_j-\alpha_1\lambda_1\lambda_2), & -\alpha_2\lambda_2\leqslant\theta_j^k\leqslant 0\\ \beta_j^{k+1}+\mu_j, & |\theta_j^k|>\alpha_2\lambda_2\end{cases}\\ \mu^{k+1}=\mu^k+\rho\left(\beta^{k+1}-\theta^{k+1}\right)\end{cases} \tag{26}$$

其中,$\rho_1=\frac{\alpha_2\rho-\lambda_1\alpha_1}{\alpha_2}$,$\rho_2=\frac{-\alpha_2\rho-\lambda_1\alpha_1}{\alpha_2}$。

4 模拟实验

为了探究在不同的数据结构下,不同惩罚函数的逻辑回归模型在参数估计、变量选择及模型准确度上的表现,因此设计了四组模拟实验,研究LR-Elastic Net、LR-SCAD和LR-MCP模型的优劣。

4.1 评价指标

Benjamini和Hochberg在1995年提出FDR(False Discovery Rate)和PSR(Positive Select Rate)指标,并广泛运用在高维数据的模型的评价中[11-15]。FDR指标代表估计为非零的系数中假阳性占的比例,PSR指标代表真实模型的非零系数中真阳性所占的比例。

$$FDR=\begin{cases}\frac{FP}{TP+FP}, & TP+FP>0\\ 0, & TP+FP=0\end{cases},$$
$$PSR=\frac{TP}{p} \tag{27}$$

其中,FP 代表真实系数为零,但估计成非零的系数个数;TP 代表真实系数为非零,但估计为零的系数的个数;p 为真实系数非零系数的个数。一般的,FDR越接近于0,PSR越接近于1,则模型表现越好。

RMSE(Root Mean Squared Error)则是用来评价估计系数与真实系数之间的差异大小的指标[16]。

$$RMSE=\sqrt{\frac{1}{p}\sum_{i=1}^{p}(\beta_i-\hat{\beta}_i)} \tag{28}$$

其中,β_i 为真实系数,$\hat{\beta}_i$ 为估计系数。一般的,RMSE越接近于0,则模型表现越好。

正确率指标(Accuracy)则表示最终模型的预测正确样本数量占总样本的比例。

4.2 实验结果

模拟实验1 随机生成小样本数据 $n=100$,$p=10$ 的二分类数据集,并且设定 p 个特征之间相关性系数 r 最大不能超过0.2,结果如表1所示。

表1 模拟实验1结果

模型	FDR	PSR	RMSE	Acc
LR-Elastic Net	0.13	0.88	0.30	0.94
LR-SCAD	0.11	1.00	0.38	0.91
LR-MCP	0.11	1.00	0.34	0.92

由表1可知,在小样本数据中,当特征之间的相关性系数 r 最大为0.2时,LR-Elastic Net模型对特征的压缩效果比较明显,其准确率Acc也最高。

模拟实验2 随机生成小样本数据 $n=100$, $p=10$ 的二分类数据集,并且设定 p 个特征之间相关性系数 r 最大不能超过0.8,结果如表2所示。

表2 模拟实验2结果

模型	FDR	PSR	RMSE	Acc
LR-Elastic Net	0.00	1.00	0.48	0.97
LR-SCAD	0.11	1.00	0.56	0.94
LR-MCP	0.11	1.00	0.56	0.95

由表2可知,在小样本数据中,当特征之间的相关性系数 r 最大为0.8时,LR-Elastic Net模型在FDR、PSR和Acc三个指标表现较好,模型分类效果最好,准确率达到了97%,且误选率FDR高于LR-SCAD和LR-MCP模型,同时系数估计准确率较低。

模拟实验3 随机生成大样本数据 $n=1000$, $p=20$ 的二分类数据集,并且设定 p 个特征之间相关性系数 r 最大不能超过0.2,结果如表3所示。

表3 模拟实验3结果

模型	FDR	PSR	RMSE	Acc
LR-Elastic Net	0.08	0.75	0.74	0.92
LR-SCAD	0.13	0.88	0.70	0.94
LR-MCP	0.13	0.88	0.60	0.94

由表3可知,在大样本数据中,当特征之间的相关性系数 r 最大为0.2时,LR-SCAD和LR-MCP模型在FDR、PSR和Acc三个指标相同且优于LR-Elastic Net模型,但LR-SCAD的系数估计准确率略差于LR-MCP。

模拟实验4 随机生成大样本数据 $n=1\,000$, $p=20$ 的二分类数据集,并且设定 p 个特征之间相关性系数 r 最大不能超过0.8,结果如表4所示。

表4 模拟实验4结果

模型	FDR	PSR	RMSE	Acc
LR-Elastic Net	0.23	0.63	0.74	0.92
LR-SCAD	0.14	0.75	0.70	0.93
LR-MCP	0.14	0.75	0.60	0.94

由表4可知,在大样本数据中,当特征之间的相关性系数 r 最大为0.8时,LR-MCP模型在FDR、PSR、RMSE和Acc四个方面表现最好,LR-SCAD模型次之,LR-Elastic Net模型相对较差。

综上所述,LR-Elastic Net模型在小样本数据中的表现优于LR-SCAD和LR-MCP模型;而在大样本数据集中,LR-SCAD和LR-MCP模型在特征相关性很强时,能够很好地保留重要的变量,从而取得较好的分类效果,而LR-Elastic Net具有较强的特征压缩的能力。

5 量化选股策略

优矿(http://uqer.io/)是研究量化投资的一个重要平台,在该平台上其因子数量超过400个。而不同的因子之间往往又互相存在着复杂的关系,故对因子的选择就成为了一个研究的难点。而LR-Elastic Net、LR-SCAD和LR-MCP模型对特征选择又有很好的表现。因此,本文考虑利用上述三种模型分别构建三种量化策略,应用于量化选股中。

5.1 策略构建

对沪深300指数成分股数据,基于上述LR-Elastic Net、LR-SCAD和LR-MCP模型,建立LR-Elastic Net、LR-SCAD和LR-MCP策略。首先构建LR-Elastic Net策略,过程如下。

5.1.1 数据处理

(1)沪深300指数成分股数据起始时间为 t_0 ,终止时间为 t_3 ,并取中间时间 t_1 和 t_2 ,满足 $t_0<t_1<t_2<t_3$ 。将 $[t_0,t_1]$ 作为训练集,记作 T_1 ; $(t_1,t_2]$ 作为测试集,记作 T_2 ; $(t_2,t_3]$ 作为回测区间,记作 T_3 。

(2)选取股票因子,并确定股票因子矩阵 X ,并计算股票月收益率,若收益率大于0,则标签 y_i 为1;若收益率小于0,则标签 y_i 为0。

(3)对因子矩阵 X 进行归一化处理,得到 X' :

$$x_j^{'}=\frac{x_j-x_{j\_\min}}{x_{j\_\max}-x_{j\_\min}} \tag{29}$$

根据上述的归一化得到的因子矩阵 X' 及股票标签 y ,通过式(3)建立LR-Elastic Net模型。

5.1.2 LR-Elastic Net模型

(1)利用上述ADMM方法求解LR-Elastic Net模型的方法得到因子估计系数 β 。

(2)每月月末利用式(1)计算每只股票的后验概率估计 $P(y_i=1|x_i)$ 和 $P(y_i=0|x_i)$,股票的得分用 s_i 表示,即 $s_i=P(y_i=1|x_i)$ 。

5.1.3 回测分析

(1)将 s_i 从大到小进行排序,取前10只股票,将这10只股票的得分记作 $S_1,S_2,\cdots,S_{10}$,计算买入股票的权重 q_i :

$$q_i=\frac{S_i}{\sum_{i=1}^{10}S_i},i=1,2,\cdots,10 \tag{30}$$

(2)计算每月月末买入股票的数量 Q 。

$$Q_i=\frac{q_i\times C}{p_i} \tag{31}$$

其中, C 为资金数, p_i 为月末股票 i 的价格。

通过上述步骤,得到LR-Elastic Net策略,将(2)中的LR-Elastic Net替换为LR-SCAD模型,可得到LR-SCAD策略;将(2)中的LR-Elastic Net替换为LR-MCP模型,可得到LR-MCP策略。

5.2 月交易策略结果

本文以沪深300指数成分股月度数据进行实证分析,取 t_0 为2010年1月1日, t_3 为2019年5月31日, t_1

表5　策略因子表

因子类型	因子名称
成长因子	净利润增长率(x_1)、净资产增长率(x_2)、八季度净利润变化趋势(x_3)
营运因子	八季度净利润变化趋势(x_4)、应付账款周转率(x_5)、应收账款周转天数(x_6)
交易因子	成交量比率(x_7)、成交量量指数平滑异同移动平均线(x_8)、5日平均换手率(x_9)、10日平均换手率(x_{10})、20日平均换手率(x_{11})、60日平均换手率(x_{12})、120日平均换手率(x_{13})、240日平均换手率(x_{14})、20日成交量标准差(x_{15})、20日成交金额的移动平均值(x_{16})、20日资金流量(x_{17})、OBV指标的20日平均(x_{18})、20日相对换手率(x_{19})、成交量的26日指数移动平均(x_{20})、成交量摆动指标(x_{21})
波动因子	当前价格处于过去1年股价的位置(x_{22})、历史贝塔(x_{23})、历史波动(x_{24})、股价偏度(x_{25})、5日价格动量(x_{26})、10日价格动量(x_{27})、20日价格动量(x_{28})、收益相对波动(x_{29})
盈利因子	流动比率(x_{30})、产权比率(x_{31})、市场杠杆(x_{32})、账面杠杆(x_{33})、股东权益比率(x_{34})、销售毛利率(x_{35})、资产回报率(x_{36})、权益回报率(x_{37})
估值因子	基本每股收益(x_{38})、稀释每股收益(x_{39})、对数流通市值(x_{40})、市盈率(x_{41})、市净率(x_{42})
均线因子	5日移动均线(x_{43})、10日移动均线(x_{44})、20日移动均线(x_{45})、60日移动均线(x_{46})、120日移动均线(x_{47})
趋势因子	赫斯特指数(x_{48})、动量指标(x_{49})、累计振动升降指标(x_{50})

和 t_2 分别为2014年1月1日和2015年12月31日,则 T_1 为2010年1月1日至2013年12月31日,T_2 为2014年1月1日至2015年12月31日,T_3 为2016年1月1日至2019年9月30日。

利用量化平台优矿网站,在考虑了成长因子、营运因子、交易因子、波动因子、盈利因子、估值因子、均线因子和趋势因子等因素后,共选取了50个因子[17-19],如表5所示。

按照上述时间区间确定股票因子矩阵 X 和股票的月收益率 y,同时去掉含有缺失值的股票,并将因子矩阵经过归一化处理。对上述处理完成的数据,利用上述ADMM算法分别求得LR-Elastic Net、LR-SCAD和LR-MCP模型的因子估计系数 β。因子估计系数 β 结果如图3所示。可见LR-Elastic Net对因子的压缩程度最大,能够很好地实现在保留重要因子的同时剔除不重要的因子;而LR-SCAD和LR-MCP只有在因子估计系数较小时,将估计系数压缩至零,当因子系数很大时,由于这部分系数是无偏的,则不进行压缩,当介于这二者之间时,则进行部分压缩,结果如图3所示。

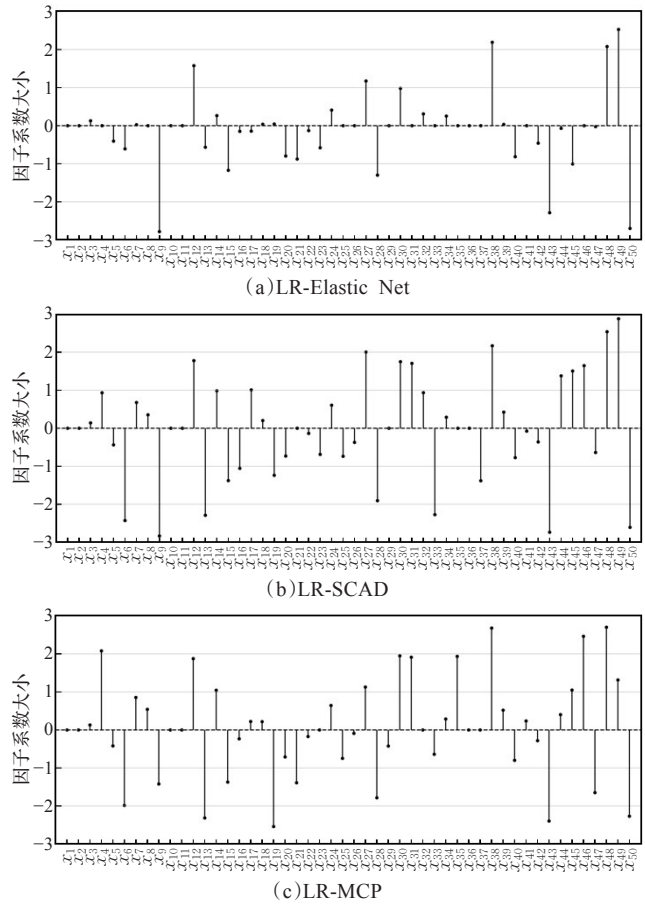

图3　因子系数结果

由于高频率的交易会带来过高的手续费,因此,本文采取月末策略进行调仓操作。此外,本文实验在优矿(http://uqer.io/)量化平台上进行,实验所设的初始资金为10 000万元,采用买入0.1%的税费,卖出0.2%的税费,印花税为0.1%,滑点为0。月交易回测结果,如表6及图4所示。

回测结果表明,同期以沪深300指数收益率为基准的年化收益率为0.61%,而LR-Elastic Net、LR-SCAD和LR-MCP策略均显著高于该水平,超额收益阿尔法值均在20%在以上。而LR-MCP策略不仅年化收益率高于LR-Elastic Net策略,而且其夏普比、最大回撤等主要评价指标均优于其他两种策略,这说明在相关性很强的股票数据中,LR-MCP模型比LR-SCAD和LR-Elastic Net

表6　月交易回测结果

策略	年化收益率/%	基准年化收益率/%	α/%	β	夏普比率	波动率/%	信息比率	最大回撤/%	年化换手率
LR-Elastic Net	22.63	0.61	21.01	0.65	1.09	17.54	1.39	27.61	6.41
LR-SCAD	23.90	0.61	22.38	0.68	1.12	18.29	1.47	27.14	4.86
LR-MCP	26.04	0.61	24.51	0.68	1.25	18.00	1.62	27.06	5.16

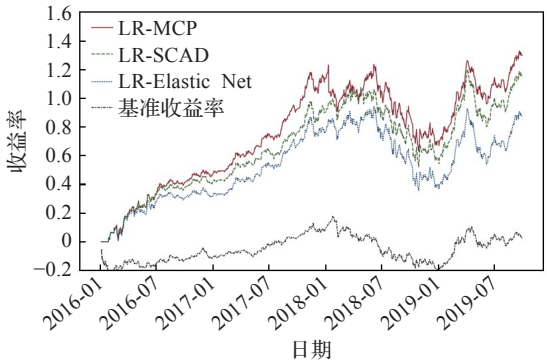

图4　月交易回测结果

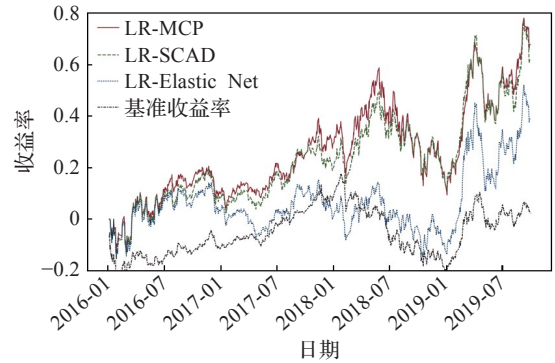

图5　周交易回测结果

表7　周交易回测结果

策略	年化收益率/%	基准年化收益率/%	α/%	β	夏普比率	波动率/%	信息比率	最大回撤/%	年化换手率
LR-Elastic Net	9.42	0.61	9.05	1.08	0.21	27.92	0.56	26.47	2.56
LR-SCAD	14.07	0.61	13.40	0.98	0.41	25.71	0.81	25.25	2.73
LR-MCP	15.23	0.61	14.59	0.99	0.45	26.24	0.85	31.13	3.33

表现更好。

5.3　周交易策略结果

多因子量化选股是采用数量化的方法进行股票组合的选择,将股票一系列的基本面因子作为选股标准,从而进行交易。

在现有的运用机器学习进行量化选股的研究中,普遍采用月度数据进行交易[3,18-20],交易频率较低,而为了验证本文提出的模型是否能够在高频交易中仍能取得较好的效果,因此使用周股票数据重复上述实验。实验过程与月交易策略相同,将月度股票数据替换为周股票数据,实验结果如图5及表7所示。

由回测结果可知,LR-SCAD和LR-MCP策略同样优于LR-Elastic Net策略,但实行周交易的回测收益的却低于月交易策略。考虑到提高交易频率后,税费、佣金等交易费用也会显著上升,为排除交易费用的影响,将月交易和周交易策略去除交易费用重新计算平均年化收益率,结果如表8所示。由表8可知,从月交易转变为周交易策略时,交易费用也会显著增长,并且交易费用的高低在一定程度上会对年化收益率造成较大影响,因此在确定交易频率时控制交易费用也是不可忽视的。而将所有的交易策略在去除交易费用后重新计算年化收益率,发现月交易策略仍然优于周交易策略,可见模型在捕捉股票数据的短期波动规律存在一定不足之处,后续可以针对模型这方面的不足继续展开研究,或选择反应股价短期波动的因子进行进一步的研究。

表8　交易费用

交易周期	策略	年化收益率/%	交易费用/万元	年化收益率(除交易费用)/%
月交易	LR-Elastic Net	21.71	556.74	22.86
	LR-SCAD	23.36	420.13	24.21
	LR-MCP	25.56	463.13	26.48
周交易	LR-Elastic Net	9.42	1 234.47	12.41
	LR-SCAD	14.07	1 495.16	17.47
	LR-MCP	15.23	1 911.47	19.52

5.4　日交易策略结果

同时为验证模型在日交易策略上的效果,使用日股票数据继续重复上述实验。实验过程与月交易和周交易策略相同,使用日股票数据进行实验,实验结果如图6及表9所示。

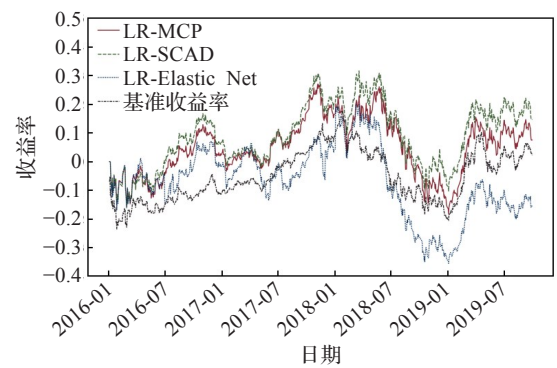

图6　日交易回测结果

表9　日交易回测结果

策略	年化收益率/%	基准年化收益率/%	α/%	β	夏普比率	波动率/%	信息比率	最大回撤/%	年化换手率
LR-Elastic Net	−4.50	0.61	−5.40	0.90	−0.32	25.19	−0.21	50.26	61.05
LR-SCAD	3.82	0.61	2.97	0.92	0.01	24.56	0.25	31.86	49.04
LR-MCP	2.02	0.61	1.20	0.93	−0.06	24.88	0.15	35.76	47.61

由回测结果可知，LR-SCAD和LR-MCP策略优于LR-Elastic Net策略，但相较于月和周交易策略，日交易策略各项主要评价指标显著降低，LR-Elastic Net策略甚至低于基准年化收益。而去除交易费用后重新计算平均年化收益率，结果如表10所示。

表10　交易费用

交易周期	策略	年化收益率/%	交易费用/万元	年化收益率（除交易费用）/%
日交易	LR-Elastic Net	−4.50	5 268.47	10.79
	LR-SCAD	3.82	4 699.64	16.05
	LR-MCP	2.02	4 394.13	13.72

由表10可知，日交易策略的交易费用较月交易策略增长率数10倍之多，极大程度地拉低了年化收益率，而在去除交易费用后，却能够取得较好的收益。因此，提高交易频率后，交易费用的存在很大程度上影响了策略的收益率。

6　结束语

针对高维度数据集特征之间的复杂性，本文将逻辑回归弹性网（LR-Elastic Net）中的L1惩罚项替换为SCAD和MCP惩罚，分别构建LR-SCAD和LR-MCP模型，并利用ADMM算法进行求解。

在模拟实验中发现LR-Elastic Net、LR-SCAD和LR-MCP模型在小样本低相关性数据集中均能取得很好的效果，分类准确率都在90%以上；而在小样本高相关性数据集中，LR-Elastic Net模型明显优于其他两种模型。在大样本数据集中，LR-SCAD和LR-MCP模型表现更好。

最后，将这三种模型运用到股票市场沪深300指数成分股数据中，构建相对应的月交易量化投资策略，发现LR-Elastic Net、LR-SCAD和LR-MCP策略均能显著优于大盘指数，其较高的超额收益均在20%以上，并且LR-SCAD和LR-MCP策略优于LR-Elastic Net策略。在此基础上，进行周交易和月交易策略，发现策略在实际执行时交易费用将是不可忽视的一项。下一步，在本文基础上，针对股票回测中的最大回撤等指标，研究在量化投资中如何利用惩罚函数有效控制风险；继续改进模型或选取有效的高频因子，以提高模型在高频交易上的效果。

参考文献：

[1] JAGANNATHAN R，MA T.Risk reduction in large portfolios：Why imposing the wrong constraints helps[J]. The Journal of Finance，2003，58(4)：1651-1684.

[2] ZOU H，HASTIE T.Regularization and Variable Selection via the Elastic net[J].Journal of the Royal Statistical Society，2005，67(2)：301-320.

[3] 谢合亮，胡迪.多因子量化模型在投资组合中的应用——基于LASSO与Elastic Net的比较研究[J].统计与信息论坛，2017，32(10)：36-42.

[4] FAN J，LI R.Variable selection via nonconcave penalized likelihood and its oracle properties[J].Journal of the American Statal Association，2001，96(9)：1348-1360.

[5] BOYD S，PARIKH N，CHU E，et al. Distributed optimization and statistical learning via the alternating direction method of multipliers[M].[S.l.]：Now Foundations and Trends，2011：1-122.

[6] ZHANG C H.Penalized linear unbiased selection[J]. Department of Statistics，2007(3)：1-22.

[7] 闫莉，陈夏.高维广义线性模型的惩罚拟似然SCAD估计[J].武汉大学学报(理学版)，2018，64(6)：533-539.

[8] 秦磊，谢邦昌.Logistic回归的ArctanLASSO惩罚似然估计及应用[J].数量经济技术经济研究，2015，32(6)：135-146.

[9] CANNARILE F，COMPARE M，BARALDI P，et al. Elastic net multinomial logistic regression for fault diagnostics of on-board aeronautical systems[J].Aerospace Science and Technology，2019，94(9)：1-15.

[10] 荣雯雯，张奇，刘艳.基于正则化回归的变量选择方法在高维数据中的应用[J].实用预防医学，2018，25(6)：645-648.

[11] SHERWOOD B. Variable selection for additive partial linear quantile regression with missing covariates[J]. Journal of Multivariate Analysis，2016，152(3)：206-223.

[12] 孙红卫，杨文越，王慧，等.惩罚logistic回归用于高维变量选择的模拟评价[J].中国卫生统计，2016，33(4)：607-611.

[13] 赵思雨.带惩罚的Logistic回归方法研究及其在企业财务预警中的应用[D].广州：暨南大学，2018.

[14] 刘乐平，张龙，蔡正高.多重假设检验及其在经济计量中的应用[J].统计研究，2007(4)：26-30.

[15] 李瑞.SNP定位的一种降维及变量选择方法[D].合肥：中国科技大学，2011.

[16] ALFONS A，CROUX C，GELPER S.Sparse least trimmed squares regression for analyzing high-dimensional large data sets[J].The Annals of Applied Statistics，2013，7(1)：226-248.

[17] 方匡南，杨阳.SGL-SVM方法研究及其在财务困境预测中的应用[J].统计研究，2018，35(8)：104-115.

[18] 李斌，林彦，唐闻轩.ML-TEA：一套基于机器学习和技术分析的量化投资算法[J].系统工程理论与实践，2017，37(5)：1089-1100.

[19] 韩杨.对技术分析在中国股市的有效性研究[J].经济科学，2001(3)：49-57.

[20] TAKEUCHI L.Applying deep learning to enhance momentum trading strategies in stocks[J].Expert Systems with Applications，2013，14：5501-5506.