

带有双层变量选择的logit模型在股价变动趋势预测中的应用

郭姝敏

河北工业大学理学院, 天津

收稿日期: 2025年12月15日; 录用日期: 2026年1月6日; 发布日期: 2026年1月19日

摘 要

股票价格受经济因素、投资者心理预期、股票市场走势、宏观政策等多种因素影响。因此, 准确预测股票价格的变动趋势是金融领域的重要难题。本文将技术指标分析与带有双层变量选择的惩罚三项logit模型相结合, 提出了Sparse Group Lasso/Group Bridge/Composite MCP/Group Exponential Lasso惩罚三项logit模型来预测股票价格的上涨、横盘和下跌趋势。首先, 选取58项重要技术指标, 将其分为13个互不重叠的组, 针对三只美国股票: 美源伯根(COR)、思科系统(CSCO)、麦当劳(MCD)分别构建模型; 接着利用训练集得到参数估计值, 利用测试集结合混淆矩阵、准确率、Kappa、HUM综合评估模型的预测性能; 最后引入Group Lasso/Group SCAD/Group MCP惩罚三项logit模型、支持向量机、随机森林、人工神经网络与本文提出的方法进行比较。结果表明, 综合各指标来看, 本文所提出的方法均优于六种对比方法。因此该方法可以有效提高预测准确率, 为投资者带来更高收益。

关键词

双层变量选择, 惩罚logit模型, 技术指标, 三分类预测性能, 机器学习方法

Application of Logit Models with Bi-Level Variable Selection in Predicting Stock Prices Movement Trends

Shumin Guo

College of Science, Hebei University of Technology, Tianjin

Received: December 15, 2025; accepted: January 6, 2026; published: January 19, 2026

Abstract

Stock prices are influenced by various factors, including economic conditions, investor psychological

文章引用: 郭姝敏. 带有双层变量选择的 logit 模型在股价变动趋势预测中的应用[J]. 统计学与应用, 2026, 15(1): 156-168. DOI: 10.12677/sa.2026.151016

expectations, market trends, and macroeconomic policies. Therefore, accurately predicting stock price movements remains a significant challenge in the field of finance. In this paper, we combine technical indicator analysis with a penalized trinomial logit model featuring bi-level variable selection, and propose Sparse Group Lasso/Group Bridge/Composite MCP/Group Exponential Lasso penalized multinomial logit models to forecast up trends, sideways trends and down trends in stock prices movement trends. Firstly, 58 important technical indicators are selected and divided into 13 mutually exclusive groups. Models are constructed for three U.S. stocks: Cencora (COR), Cisco Systems (CSCO) and McDonald's (MCD). Secondly, parameter estimates are obtained using the training set, and the predictive performance of the models is comprehensively evaluated on the test set using the confusion matrix, accuracy, Kappa and HUM. Finally, comparisons are made with Group Lasso/Group SCAD/Group MCP penalized trinomial logit models, SVM, RF and ANN. The results demonstrate that, across all evaluation metrics, the proposed methods outperform the other 6 approaches. Therefore, this method can effectively improve the prediction accuracy and provide investors with higher returns.

Keywords

Bi-Level Variable Selection, Penalized Logit Model, Technical Indicators, Three-Category Prediction Performance, Machine Learning Methods

Copyright © 2026 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

股票市场作为金融体系的重要组成部分，其价格波动受宏观经济环境、市场情绪及突发事件等综合影响，呈现出高度非线性与复杂性。同时，由于模型不确定性、参数不稳定性以及噪声的存在，传统模型难以捕捉其复杂的内在关联，导致股票收益运动方向预测准确率受限[1]。传统的股价预测研究主要建立在统计学模型的基础上，如自回归条件异方差模型[2]、自回归综合移动平均模型[3]等。

人工智能的快速发展为金融时间序列预测带来了很大的突破。Ballings 等人[4]比较了 RF、SVM、AdaBoost、神经网络和逻辑回归这五种机器学习算法在股价方向预测中的表现，证实了 RF 综合表现最优。Yun 等人[5]开发了混合 GA-XGBoost 预测框架，通过特征增强方法提升了股价趋势预测性能，证实了特征工程在金融预测中的关键作用。Long 等人[6]提出了双向 LSTM 与图神经网络的混合预测框架，并采用注意力机制实现多源信息的动态融合。Vuong 等人[7]提出了 XGBoost-LSTM 模型，通过两阶段建模实现股票价格预测。

随着数据维度急剧增长，稀疏正则化技术因能够同时实现特征选择与模型压缩，成为处理高维数据的重要方法。实际应用中，解释变量往往表现出组效应，需要引入带有组结构的惩罚方法，如 Yuan 和 Lin [8]提出的 Group Lasso 以及 Huang 等人[9]提出的 Group SCAD 和 Group MCP。在组内也稀疏的情况下，仅在组水平上选择变量会导致模型解释结果不合理，因此有学者提出可以进行双层变量选择的方法。Huang 等人[10]提出的 Group Bridge 方法是最早的双层变量选择方法。Breheny 和 Huang [11]提出了组内与组间惩罚都是 MCP 惩罚函数的 Composite MCP 方法。Wu 和 Lange [12]提出了 Sparse Group Lasso 方法，可以看作 Lasso 和 Group Lasso 的线性组合。Breheny [13]提出了 Group Exponential Lasso 方法，结合了指数惩罚函数与复合组 l_1 正则化框架。

多分类问题是机器学习的核心任务。Glonek 和 McCullagh [14]提出了多元 Logistic 模型，弥补了传统

二分类 Logistic 回归在分析多变量数据时的不足。Novoselova 等人[15]开发了 R 包 HUM, 为多类别受试者工作特征(Receiver Operating Characteristic, ROC)分析提供了开源工具。Li 等人[16]探讨了多类分类器性能评估指标, 包括通过 R 包 mcca 实现的 HUM、PDI 和 CCP 等。Hu 和 Yang [17]提出 G-LASSO/G-SCAD/G-MCP 惩罚三项 logit 动态模型, 同时结合技术指标, 解决了股票市场趋势三分类预测的问题。Hu 和 Yang [18]提出基于组惩罚(G-LASSO/G-SCAD/G-MCP)的多项 logit 模型, 采用准确率、Kappa、PDI 和 HUM 等多指标进行综合评估, 证实了该方法的有效性。

受上述方法的启发, 我们考虑用带有双层变量选择的 logit 模型来预测股票价格的上涨趋势、横盘趋势和下跌趋势, 以实现组间与组内变量的双重选择。在三个数据集上的实验结果表明, 该方法在准确率、Kappa、HUM 上均表现出良好的预测性能, 具有一定的实际应用价值。

2. 惩罚三项 Logit 模型

令 C_t 表示 t 交易日结束时的收盘价, $Z_{t+1} = C_{t+1} - C_t$ 表示第 $t + 1$ 个交易日的股票超额收益, $c(-\infty < c < \infty)$ 为收益率波动阈值常数, 则有股票收益运动方向指示函数:

$$Y_{t+1}(c) = \begin{cases} 1, & \text{if } Z_{t+1} < -c, \\ 2, & \text{if } -c \leq Z_{t+1} \leq c, \\ 3, & \text{if } Z_{t+1} > c. \end{cases} \quad (2.1)$$

$Y_{t+1}(c)=1$ 表示下跌趋势, $Y_{t+1}(c)=2$ 表示横盘趋势, $Y_{t+1}(c)=3$ 表示上涨趋势。当 $c=0$ 时, $Y_{t+1}(c)$ 为收益方向性指标; 当 $c=0.1C_t$ 时, $Y_{t+1}(c)$ 为识别显著正向收益的指标; 当 $c=-0.1C_t$ 时, $Y_{t+1}(c)$ 为识别显著下行风险的指标。这里我们取 $c=0.005C_t$ 。为简化分析, 假设未来的信息只取决于现在, 与过去无关。

下面, 我们用训练集 $\{X_t, Y_{t+1}(c)\}_{t=1}^N$ 来学习股票收益的三个运动方向, 对应的三项 Logit 模型可以表示为:

$$P(Y_{t+1}(c)=1 | x_t; \beta) = \frac{\exp(\beta_0^1 + \sum_{l=1}^L x_{t(l)} \beta_{(l)}^1)}{\sum_{k=1}^3 \exp(\beta_0^k + \sum_{l=1}^L x_{t(l)} \beta_{(l)}^k)}, \quad (2.2)$$

$$P(Y_{t+1}(c)=2 | x_t; \beta) = \frac{\exp(\beta_0^2 + \sum_{l=1}^L x_{t(l)} \beta_{(l)}^2)}{\sum_{k=1}^3 \exp(\beta_0^k + \sum_{l=1}^L x_{t(l)} \beta_{(l)}^k)}, \quad (2.3)$$

$$P(Y_{t+1}(c)=3 | x_t; \beta) = \frac{\exp(\beta_0^3 + \sum_{l=1}^L x_{t(l)} \beta_{(l)}^3)}{\sum_{k=1}^3 \exp(\beta_0^k + \sum_{l=1}^L x_{t(l)} \beta_{(l)}^k)} \quad (2.4)$$

其中, $\beta = (\beta_0^k, (\beta^k)^\top)^\top$ 为参数向量, β_0^k 为第 k 类截距项, β^k 为第 k 类系数向量, $\beta_{(l)}^k$ 为第 k 类中第 l 组对应的系数向量, L 为组的个数, $x_{t(l)}$ 为第 t 个样本的第 l 组技术指标向量, $y_{t+1}^k = I(Y_{t+1}(c)=k)$ 。可以得到如下三项 logit 负对数似然损失函数:

$$l(\beta) = \frac{1}{N} \sum_{t=1}^N \left\{ \log \left[\sum_{k=1}^3 \exp(\beta_0^k + x_t^\top \beta^k) \right] - \sum_{k=1}^3 y_{t+1}^k (\beta_0^k + x_t^\top \beta^k) \right\}. \quad (2.5)$$

直接求解上式会引发过拟合和多重共线性问题, 因此通过建立惩罚三项 logit 模型解决该问题。考虑到技术指标间存在组结构, 引入双层变量选择惩罚函数, 在组间筛选具有显著预测能力的指标组, 同时在组内进一步识别关键个体指标。带有双层变量选择的三项 logit 模型如下所示:

$$Q(\beta; \lambda) \equiv l(\beta) + \sum_{k=1}^3 \sum_{l=1}^L P\left(\|\beta_{(l)}^k\|; \lambda\right). \quad (2.6)$$

其中, $P(\|\beta_{(l)}^k\|; \lambda)$ 表示惩罚函数, λ 为控制惩罚程度的超参数。下面介绍本文用到的四种带有双层变量选择的三项 logit 模型:

1) Sparse Group Lasso (SGL) 惩罚三项 logit 模型

$$Q(\beta; \lambda) \equiv l(\beta) + \sum_{k=1}^3 \left[(1-\alpha)\lambda \sum_{l=1}^L \sqrt{p_l} \|\beta_{(l)}^k\|_2 + \alpha\lambda \|\beta^k\|_1 \right], \quad (2.7)$$

其中, $\alpha \in [0, 1]$, 当 $\alpha = 1$ 时为 Lasso 惩罚, 当 $\alpha = 0$ 时为 Group Lasso 惩罚; p_l 表示第 l 组技术指标的变量维度。

2) Group Bridge (GB) 惩罚三项 logit 模型

$$Q(\beta; \lambda) \equiv l(\beta) + \lambda \sum_{k=1}^3 \sum_{l=1}^L p_l^\gamma \|\beta_{(l)}^k\|_1^\gamma, \quad (2.8)$$

当 $\gamma \in (0, 1)$ 时, 可同时用于组水平及个体水平的变量选择。

3) Composite MCP (CMCP) 惩罚三项 logit 模型

$$Q(\beta; \lambda) \equiv l(\beta) + \sum_{k=1}^3 \sum_{l=1}^L P_{\lambda, b} \left(\sum_{j=1}^{p_l} p_{\lambda, a}(|\beta_{lj}^k|) \right), \quad (2.9)$$

其中, β_{lj}^k 为第 k 类对应的第 l 组的第 j 个指标系数, $P_{\lambda, a}(\cdot)$ 为 MCP 惩罚函数, 具体表达式为:

$$p_{\lambda, a}(|\beta_{lj}^k|) = \begin{cases} \lambda |\beta_{lj}^k| - \frac{(\beta_{lj}^k)^2}{2a}, & |\beta_{lj}^k| \leq a\lambda, \\ \frac{a\lambda^2}{2}, & |\beta_{lj}^k| > a\lambda. \end{cases} \quad (2.10)$$

4) Group Exponential Lasso (GEL) 惩罚三项 logit 模型

$$Q(\beta; \lambda) \equiv l(\beta) + \sum_{k=1}^3 \sum_{l=1}^L \frac{\lambda^2}{\tau} \left\{ 1 - \exp \left(-\frac{\tau \|\beta_{(l)}^k\|_1}{\lambda} \right) \right\}, \quad (2.11)$$

其中, τ 为耦合参数, 决定耦合的强度, 耦合指的是系数因与其他重要预测变量同组而使得惩罚强度减弱的现象。

3. 参数估计与概率估计

3.1. 坐标梯度下降算法

对于 SGL 惩罚三项 logit 模型, 基于 Vincent 和 Hansen [19] 提出的想法, 我们采用坐标梯度下降算法来进行求解。算法通过从最大正则化强度 $\lambda_{\max}$ 逐步递减至预设的最小值 $\lambda_{\min}$ 来进行求解。 $\lambda_{\max}$ 可通过下式进行计算:

$$\lambda_{\max} = \max_{l=1, \dots, L} \inf \left\{ \lambda > 0 \mid \sqrt{k \left(\lambda \alpha \xi_{(l)}, (\nabla l(0))_{(l)} \right)} \leq \lambda (1-\alpha) \gamma_l \right\}, \quad (3.1)$$

其中, K 为类别数, $\xi_{(l)}$ 的每个分量都为 1, $\gamma_l = \sqrt{p_l}$ 。在当前参数值 $\tilde{\beta}$ 处, 对式(2.5)中的负对数似然损失函数进行二次近似可得:

$$q^T(\beta - \tilde{\beta}) + \frac{1}{2}(\beta - \tilde{\beta})^T H(\beta - \tilde{\beta}), \quad (3.2)$$

其中, $q = \nabla l(\tilde{\beta})$ 为梯度, $H = \nabla^2 l(\tilde{\beta})$ 为 Hessian 矩阵。因为 H 为对称矩阵, 所以负对数似然损失函数 l 在 $\tilde{\beta}$ 处的二次近似可以化简为:

$$f(\beta) - q^T \tilde{\beta} + \frac{1}{2} \tilde{\beta}^T H \tilde{\beta}, \quad (3.3)$$

其中, $f(\beta) = (q - H \tilde{\beta})^T \beta + \frac{1}{2} \beta^T H \beta$ 。设 Sparse group lasso 惩罚函数为 $\phi(\beta)$, 因此可以将式(2.7)转化为如下惩罚二次优化问题:

$$\min_{\beta} \{Q(\beta) = f(\beta) + \phi(\beta)\}. \quad (3.4)$$

根据参考文献[19]中的介绍, 上述惩罚二次优化问题可通过块坐标下降算法, 将全局问题分解为对各特征组的局部子问题, 依次更新组内系数来逐步逼近最优解。对于组内系数, 采用改进的坐标下降算法来处理非光滑惩罚项, 确保组内稀疏性。

3.2. 局部坐标下降算法

对于其他三种惩罚函数对应的 logit 模型, 采用 One-vs-All (OvA) 策略进行多分类任务处理。该策略的核心思想是将多分类问题分解为多个独立的二分类任务, 为每个类别训练一个专门的二分类器, 其中当前类别作为正类, 其余所有类别统一作为负类, 也就是第 i 个分类器的任务是区分类别 i (正类) 和非类别 i (负类, 即其他所有类别的组合)。

针对本文的三分类场景, 首先将原始三分类标签转换为三个二分类标签, 并训练三个对应的二分类器。在模型预测阶段, 对于新的输入样本, 首先通过这三个二分类器分别获得线性预测值, 然后利用 softmax 函数将这些输出值转换为概率分布, 取概率最大的类别作为最终的输出结果。Rifkin 和 Klautau [20] 表明, 当基分类器性能良好时, OvA 策略的分类效果与其他复杂方法相当。此外, 考虑到本文使用的数据集不存在类别不平衡问题, 采用 OvA 策略具有实现简单、高效的优势, 能够有效满足本文的分类预测需求。

以 GB 惩罚三项 logit 模型为例, 采用 OvA 策略后需要优化的目标函数为:

$$\min_{\beta} \left\{ L(\beta) + \lambda \sum_{l=1}^L p_l^{\gamma} \|\beta_{(l)}\|_1^{\gamma} \right\}, \quad (3.5)$$

其中, $L(\beta) = \frac{1}{N} \left[\left(\sum_{t=1}^N \sum_{l=1}^L \log \left(1 + e^{X_{t(l)}^T \beta_{(l)}} \right) - Y_t X_{t(l)}^T \beta_{(l)} \right) \right]$ 为二分类问题对应的逻辑回归负对数似然损失函数, Y_t 为第 t 个样本对应的二分类标签。

上述问题可以用局部坐标下降算法来进行求解。当 $0 < \gamma < 1$ 时, group bridge 惩罚是非凸的, 直接优化上式是比较困难的, 因此将问题转化为更容易求解的凸问题, 首先定义:

$$S(\beta, \theta) = l(\beta) + \sum_{l=1}^L \theta_l^{1-\gamma} p_l \|\beta_{(l)}\|_1 + \tau \sum_{l=1}^L \theta_l, \quad (3.6)$$

其中 τ 为惩罚参数。根据 Huang 等人[11]的证明, 当 $\lambda = \tau^{1-\gamma} \gamma^{-\gamma} (1-\gamma)^{\gamma-1}$ 时, $\hat{\beta}$ 最小化式(3.5)当且仅当 $(\hat{\beta}, \hat{\theta})$ 在 $\theta_l \geq 0$ 的约束下最小化式(3.6)。基于上述转化, 可以得到下面两个关键迭代步骤, 首先计算:

$$\theta_l = \left[\frac{p_l (1-\gamma)}{\tau \gamma} \right]^{\gamma} \|\beta_{(l)}\|_1^{\gamma}. \quad (3.7)$$

然后按照下式更新 $\tilde{\beta}$:

$$\tilde{\beta} = \arg \min \left\{ l(\beta) + \sum_{l=1}^L \theta_l^{1-\frac{1}{\gamma}} p_l \|\beta_{(l)}\|_1 \right\}. \quad (3.8)$$

重复式(3.7)和式(3.8)直至收敛, 其中式(3.8)是主要的计算步骤, 这是一个 Lasso 问题, 可以用最小角回归算法有效求解。

CMCP 惩罚三项 logit 模型和 GEL 惩罚三项 logit 模型与上述求解过程相似, 在采用 OvA 策略进行转换后通过局部坐标下降算法进行求解。

4. 三分类预测性能

利用训练集 $\{x_t, y_{t+1}^k\}_{t=1}^N$ 拟合模型后, 基于测试集 $\{x_t, y_{t+1}^k\}_{t=N+1}^{N+n}$ 计算各类别的概率估计值:

$$\hat{P}(y_{t+1}^k = k | X_t; \hat{\beta}) = \frac{\exp(\hat{\beta}_0^k + x_t \hat{\beta}^k)}{\sum_{k=1}^3 \exp(\hat{\beta}_0^k + x_t \hat{\beta}^k)}, k = 1, 2, 3, \quad (4.1)$$

其中, $\hat{\beta} = (\hat{\beta}_0^1, (\hat{\beta}^1)^\top, \hat{\beta}_0^2, (\hat{\beta}^2)^\top, \hat{\beta}_0^3, (\hat{\beta}^3)^\top)^\top$ 为系数估计值。

贝叶斯分类器是风险最小的分类器, 因此针对三分类问题, 采用贝叶斯分类器来进行类别标签的预测:

$$\hat{Y}_{t+1}(c) = \arg \max_{k \in \{1, 2, 3\}} \hat{P}(y_{t+1} = k | X_t; \hat{\beta}). \quad (4.2)$$

表 1 为用于评估预测性能的三分类混淆矩阵, 其中 $V(i, j)$ 表示真实类别为 i 且预测类别为 j 的样本数量, 由此用正确预测的样本数除以总样本数可以计算得到准确率(Accuracy)。对于多分类问题, 一些传统的准确性度量指标如: 灵敏度(Sensitivity)、特异度(Specificity)、AUC (Area Under the Curve)等是不可用的, 因此这里采用多分类性能衡量指标。

Table 1. Three-class confusion matrices

表 1. 三分类混淆矩阵

	预测类 1: $\hat{Y}_{t+1} = 1$	预测类 2: $\hat{Y}_{t+1} = 2$	预测类 3: $\hat{Y}_{t+1} = 3$	合计
真实类 1: $Y_{t+1} = 1$	$V(1,1)$	$V(1,2)$	$V(1,3)$	$V(1,\cdot)$
真实类 2: $Y_{t+1} = 2$	$V(2,1)$	$V(2,2)$	$V(2,3)$	$V(2,\cdot)$
真实类 3: $Y_{t+1} = 3$	$V(3,1)$	$V(3,2)$	$V(3,3)$	$V(3,\cdot)$
合计	$V(\cdot,1)$	$V(\cdot,2)$	$V(\cdot,3)$	$V(\cdot,\cdot)$

Kappa 系数是一种用于评估分类任务中一致性的统计指标, 计算公式如下:

$$\frac{\sum_{k=1}^3 p_{kk} - \sum_{k=1}^3 p_{k+} p_{+k}}{1 - \sum_{k=1}^3 p_{k+} p_{+k}}, \quad (4.3)$$

其中, $p_{kk} = P(Y_{t+1} = k, \hat{Y}_{t+1} = k)$, $p_{k+} = P(Y_{t+1} = k)$, $p_{+k} = P(\hat{Y}_{t+1} = k)$ 。Kappa $\in (0,1)$ 且取值越大表示模型效果越好。

针对三分类问题, 可通过绘制三维 ROC 曲面来直观展示模型的分类性能, 并计算 ROC 曲面下的体积, 即 HUM (Hypervolume Under the Manifold)。HUM 是 AUC 在多分类问题中的推广, 能够综合反映模

型的内在准确性，计算公式为：

$$\text{HUM} = \int_0^1 \int_0^{f_1(t_1)} \dots \int_0^{f_{M-2}(t_1, \dots, t_{M-2})} f_{M-1}(t_1, \dots, t_{M-1}) dt_{M-1} \dots dt_2 dt_1, \quad (4.4)$$

其中， $t_i = f_{i-1}(t_1, \dots, t_{i-1})$, $i = 2, \dots, K$ ，表示第 i 类中的样本被正确分类的概率。具体实现可利用 R 包 HUM 来进行计算，该包专为多分类 ROC 分析设计，能够高效评估模型的整体分类准确性。一般来说，HUM 值越大，表示模型性能越优。

5. 实证分析

5.1. 数据集

我们选择了三只有代表性的美国股票：美源伯根(COR)、思科系统(CSCO)、麦当劳(MCD)。由于新冠疫情的影响，2019 年 12 月起股市波动很大，因此选取 2013 年 10 月 9 日至 2019 年 9 月 25 日期间的 1500 条完整交易数据来进行分析。股票数据通过 Tiingo 金融数据平台提供的 API 接口获取，采用 R 包 httr 实现自动化数据采集。交易数据包括每日开盘价、收盘价、最高价、最低价、调整后价格以及成交量。图 1 展示了这三只股票在该时期内的历史股价走势。

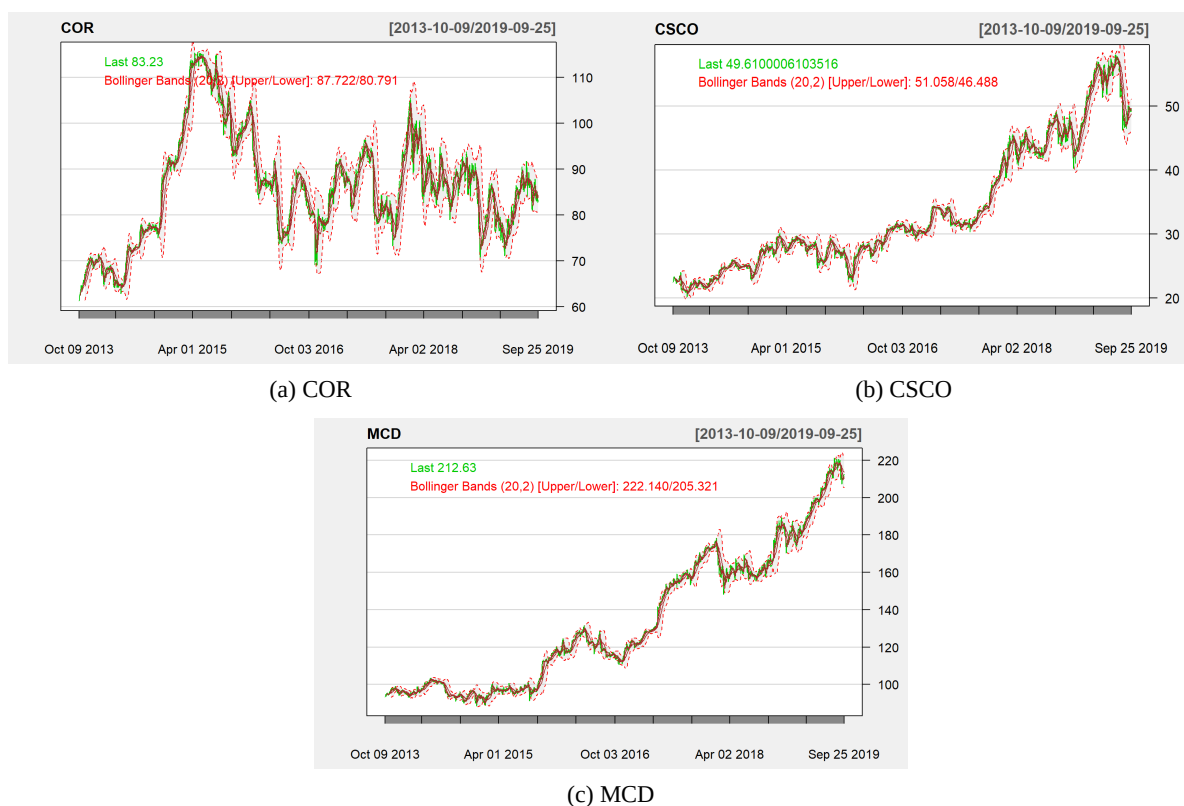


Figure 1. The stock price trends of COR, CSCO and MCD

图 1. COR、CSCO、MCD 的股价走势

接着将股票数据划分为训练集和测试集两部分。其中，训练集用于模型参数的估计，测试集用于检验模型的预测效果。本文采用常见的样本划分比例，将 70% 的数据作为训练集，30% 的数据作为测试集。具体而言，选取 2013 年 10 月 9 日至 2017 年 12 月 8 日的 1050 条观测数据作为训练集，2017 年 12 月 9 日至 2019 年 9 月 25 日的 450 条观测数据作为测试集。

5.2. 技术指标

在股票预测中,技术指标是通过对历史价格、成交量等市场数据进行数学运算得出的量化工具,用于分析市场趋势、动量、波动性和买卖信号,其数学本质是对市场行为的概率化表征,旨在提取市场微观结构中的有效信息。本文采用 R 包 TTR 计算得到 58 个技术指标,包括简单移动平均(SMA)、指数移动平均(EMA)、简易波动指标(EMV)、加权移动平均(WMA)、布林带下轨(BBands-L)、布林带上轨(BBands-U)、相对强弱指数(RSI)、商品通道指数(CCI)、资金流量指数(MFI)、去趋势价格震荡指标(DPO-C)、平衡交易量(OBV)、变动率(ROC)等。将这 58 个技术指标表示为 $X_{t,1}, \dots, X_{t,58}$, 并根据参考文献 34 中的分类方法,将技术指标分为 13 个互不重叠的组,得到组向量 $X_t = (X_{t(1)}, \dots, X_{t(13)})$, 其中 $X_{t(i)}$ 为第 i 组的技术指标向量。

为消除不同技术指标在量纲和取值范围上的差异,提升模型求解的数值稳定性与计算效率,并保证正则化项对各特征变量施加惩罚时的公平性,本文对所有特征变量进行标准化预处理。标准化后的特征组向量记为 $\tilde{X}_t = (\tilde{X}_{t(1)}, \dots, \tilde{X}_{t(13)})$, 后续模型的训练过程及预测性能评估均基于 $\tilde{X}_t$ 展开。

5.3. 超参数选择

正则化参数 λ 的选择取决于很多因素,包括数据维度、样本大小、分组个数、参数序列设定以及计算资源限制等。 λ 的取值过小会导致大量变量被保留,模型稳定性变差; λ 的取值过大可能会忽略对预测具有重要作用的特征变量,导致信息利用不足。本文采用网格搜索来寻找最优的超参数组合。对于 Sparse Group Lasso 惩罚,变量选择由 α 和 λ 决定;对于 Group Bridge 和 Composite MCP 惩罚,变量选择由 γ 和 λ 决定;对于 Group Exponential Lasso 惩罚,变量选择由 τ 和 λ 决定。

首先,给定超参数 $\alpha / \gamma / \tau$ 的取值范围,对于每一个取值,生成一个含有 100 个值的 λ 序列,算法遍历整个参数组合。为了选择出最优的参数组合,采用 10 折交叉验证计算分类错误率,最终选择使交叉验证错误率最低的参数组合来训练模型。使用交叉验证的目的是避免由于单一划分导致的偶然误差,提高模型的泛化能力,模拟模型在真实场景中的性能。表 2 为通过网格搜索和交叉验证确定的最优超参数组合,其中 $\lambda_i (i=1,2,3)$ 分别为三个类别对应的正则化参数。

Table 2. Optimal parameter settings for the four classification methods
表 2. 四种分类方法的最优参数设定

数据集	方法	最优参数设定
COR	SGL (α, λ)	(0.86, 0.0067)
	GB ($\gamma, \lambda_1, \lambda_2, \lambda_3$)	(0.80, 0.0079, 0.0029, 0.0080)
	CMCP ($\gamma, \lambda_1, \lambda_2, \lambda_3$)	(3.80, 0.0098, 0.0052, 0.0100)
	GEL ($\tau, \lambda_1, \lambda_2, \lambda_3$)	(0.20, 0.0152, 0.0212, 0.0519)
CSCO	SGL (α, λ)	(0.96, 0.0110)
	GB ($\gamma, \lambda_1, \lambda_2, \lambda_3$)	(0.90, 0.0100, 0.0095, 0.0098)
	CMCP ($\gamma, \lambda_1, \lambda_2, \lambda_3$)	(3.80, 0.0100, 0.0100, 0.0100)
	GEL ($\tau, \lambda_1, \lambda_2, \lambda_3$)	(0.20, 0.0228, 0.0383, 0.0435)
MCD	SGL (α, λ)	(0.60, 0.0088)
	GB ($\gamma, \lambda_1, \lambda_2, \lambda_3$)	(0.30, 0.0090, 0.0078, 0.0100)
	CMCP ($\gamma, \lambda_1, \lambda_2, \lambda_3$)	(4.30, 0.0100, 0.0100, 0.0080)
	GEL ($\tau, \lambda_1, \lambda_2, \lambda_3$)	(0.10, 0.0228, 0.0268, 0.0400)

5.4. 模型估计

基于训练集数据及表 2 中的最优超参数组合, 我们构建了四种惩罚 logit 模型。通过前面提到的算法分别求解各模型在三个数据集上的参数估计值, 得到截距项与组系数向量。由模型估计结果可得, 不同数据集所识别出的重要指标存在一定差异; 同时, 对于不同类别 k , 各指标对应的参数估计值也表现出明显不同。其中, 去趋势价格震荡指标(DPO-C)、简易波动指标(EMV)、变动率指标(ROC)在四种模型中通常系数绝对值都较大, 说明价格相对于其短期均值的偏离程度、价格变动与成交量之间的匹配关系以及价格在给定时间窗口内上涨或下跌的相对强度对预测股价运动趋势具有较大影响, 属于重要指标。

5.5. 预测性能

模型建立之后, 接着在测试集上评估其预测性能。为了证明本文所提出方法的有效性, 与参考文献 [19]中所提出的三种统计方法以及三种经典的机器学习方法(SVM、RF、ANN)进行对比分析。混淆矩阵可用于系统刻画模型预测结果与真实类别之间的对应关系, 不仅能评估整体性能, 也可以分析模型对于不同类别的识别能力差异。通过三个测试集上四种模型的三分类混淆矩阵, 可以分别计算得到下跌准确率(D-ACC)、横盘准确率(S-ACC)、上涨准确率(U-ACC)、灵敏度和特异度, 进而可以根据灵敏度和特异度这两个指标绘制得到三维 ROC 曲面, 具体见图 2~4。ROC 曲面的三个标签 1、2、3 分别表示正确分类三个类别的概率, 取值范围为[0,1]。ROC 曲面下的体积越大, HUM 值越高, 模型的分类预测性能越优。

表 3~5 为 10 种方法在三个不同数据集上的预测结果。结合 ROC 曲面以及预测结果可以看出, 四种带有双层变量选择的三项 logit 模型的预测性能优于其他六种模型, 特别是在 Kappa 系数这一评估指标上, 模型表现出显著优势, 表明双层变量选择策略能有效提升分类的一致性。

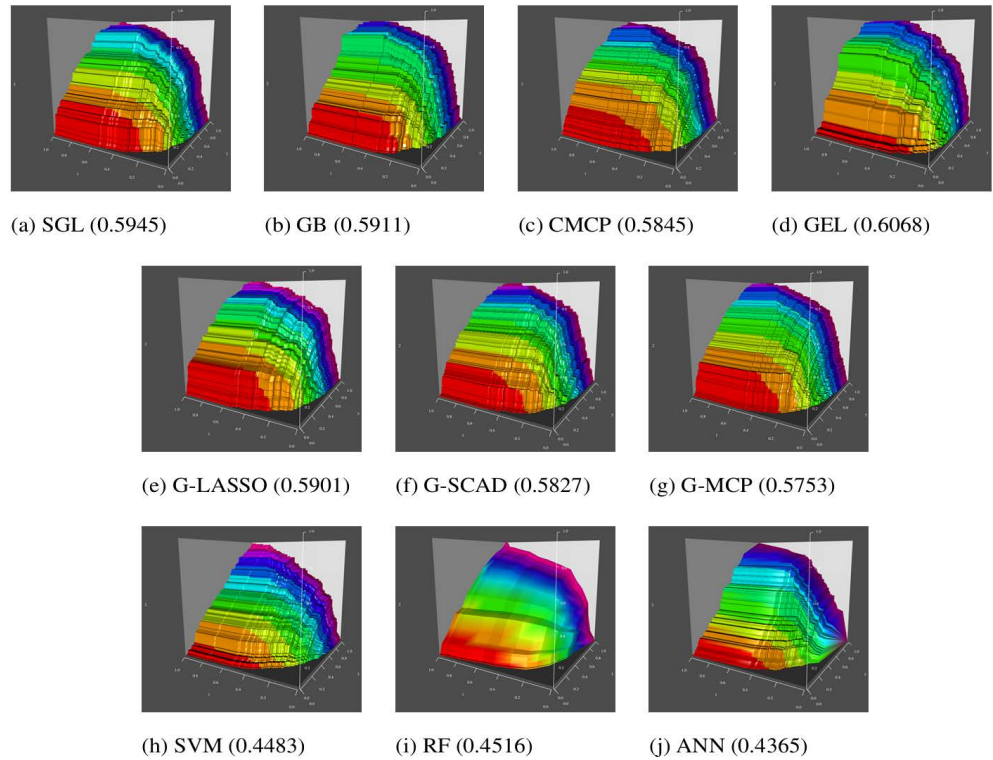


Figure 2. The ROC surfaces and the HUM values for COR
图 2. COR 数据集的 ROC 曲面及 HUM 值

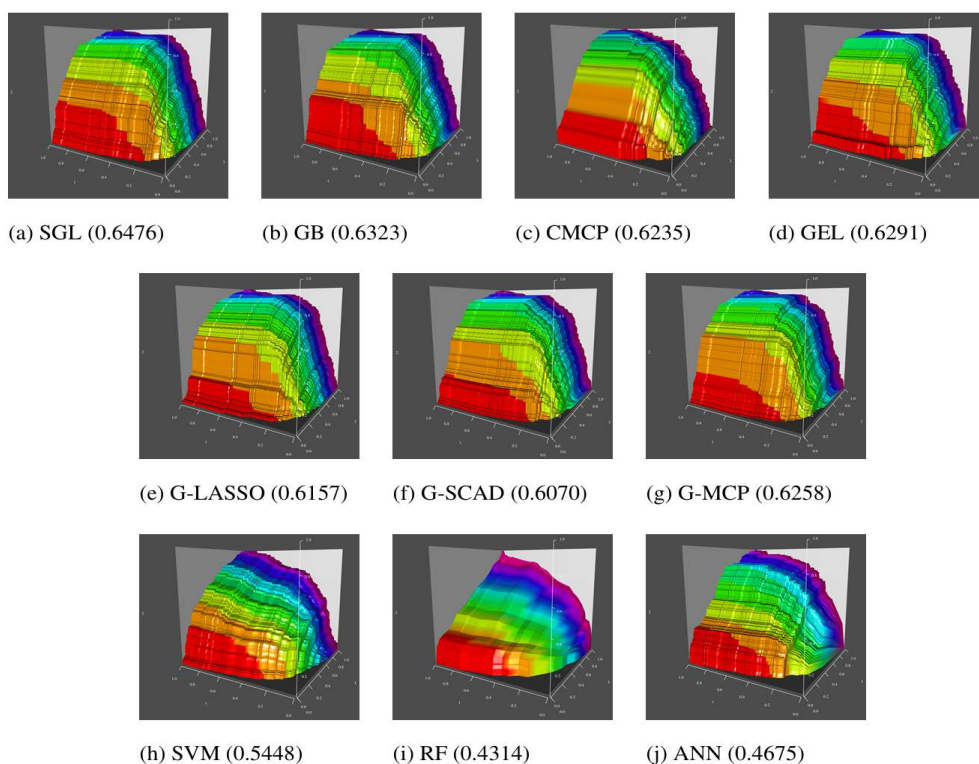


Figure 3. The ROC surfaces and the HUM values for CSCO

图 3. CSCO 数据集的 ROC 曲面及 HUM 值

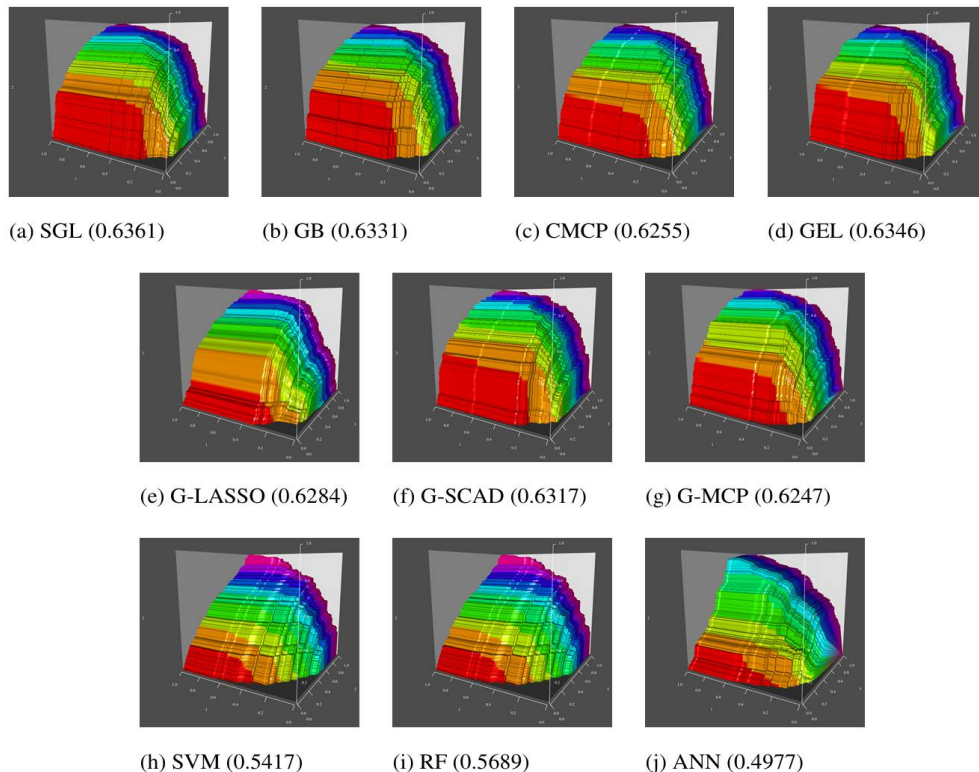


Figure 4. The ROC surfaces and the HUM values for MCD

图 4. MCD 数据集的 ROC 曲面及 HUM 值

Table 3. The prediction comparison to the 10 methods for COR**表 3.** COR 数据集上 10 种方法的预测性能比较

	D-ACC	S-ACC	U-ACC	ACC	Kappa	HUM
SGL	0.7342	0.4571	0.7112	0.6600	0.4815	0.5940
GB	0.7785	0.4095	0.7433	0.6778	0.5019	0.5911
CMCP	0.7658	0.4571	0.7166	0.6733	0.4997	0.5845
GEL	0.8291	0.3048	0.7807	0.6867	0.5063	0.6068
G-LASSO	0.6772	0.4857	0.7219	0.6511	0.4690	0.5901
G-SCAD	0.7532	0.4571	0.6791	0.6533	0.4697	0.5827
G-MCP	0.6899	0.5143	0.6631	0.6378	0.4518	0.5753
SVM	0.5823	0.5238	0.4813	0.5267	0.3071	0.4483
RF	0.7025	0.3810	0.6096	0.5889	0.3748	0.4516
ANN	0.7152	0.4190	0.5241	0.5667	0.3505	0.4365

Table 4. The prediction comparison to the 10 methods for CSCO**表 4.** CSCO 数据集上 10 种方法的预测性能比较

	D-ACC	S-ACC	U-ACC	ACC	Kappa	HUM
SGL	0.6525	0.7778	0.6264	0.6800	0.5230	0.6476
GB	0.6738	0.6963	0.6552	0.6733	0.5110	0.6323
CMCP	0.6596	0.7630	0.6149	0.6733	0.5132	0.6235
GEL	0.6454	0.7852	0.5862	0.6644	0.5015	0.6291
G-LASSO	0.5532	0.8370	0.5460	0.6356	0.4603	0.6157
G-SCAD	0.5603	0.8667	0.5402	0.6444	0.4743	0.6070
G-MCP	0.5674	0.8222	0.5862	0.6511	0.4820	0.6258
SVM	0.3121	0.3630	0.9023	0.5556	0.3066	0.5448
RF	0.4468	0.3407	0.8506	0.5711	0.3403	0.4314
ANN	0.4184	0.6741	0.5805	0.5578	0.3388	0.4675

Table 5. The prediction comparison to the 10 methods for MCD**表 5.** MCD 数据集上 10 种方法的预测性能比较

	D-ACC	S-ACC	U-ACC	ACC	Kappa	HUM
SGL	0.6545	0.7340	0.6058	0.6756	0.4874	0.6361
GB	0.6909	0.7438	0.5693	0.6778	0.4910	0.6331
CMCP	0.7182	0.7044	0.5693	0.6667	0.4769	0.6255
GEL	0.6909	0.6946	0.6058	0.6667	0.4778	0.6346
G-LASSO	0.6000	0.8227	0.4526	0.6556	0.4397	0.6284
G-SCAD	0.6455	0.7438	0.5620	0.6644	0.4667	0.6317
G-MCP	0.6182	0.7488	0.5693	0.6622	0.4624	0.6247
SVM	0.8091	0.5468	0.3577	0.5533	0.3243	0.5417
RF	0.5545	0.7340	0.4599	0.6067	0.3653	0.5689
ANN	0.6545	0.5271	0.5912	0.5778	0.3522	0.4977

6. 结论

本研究以三只美国股票为研究对象, 基于 58 个技术指标构建了四种带有双层变量选择的惩罚三项 logit 模型, 用于预测股票收益的上涨、横盘和下跌三种趋势。针对 SGL 惩罚三项 logit 模型, 研究采用坐标下降算法进行参数估计。为解决多分类问题中非凸目标函数优化困难的问题, 本研究创新性地提出采用 One-vs-All (OvA) 策略进行问题转换, 将原始三分类问题分解为三个独立的二分类子任务, 并运用局部坐标下降算法进行求解。模型建立之后, 在三个测试集上使用下跌准确率、横盘准确率、上涨准确率、总体准确率、Kappa 系数、三维 ROC 曲面和 HUM 值综合评估其预测性能。结果表明本文所提出的四种惩罚三项 logit 模型在预测股票收益问题中具有更好的表现, 优于只能进行组水平上特征选择的 G-LASSO/G-SCAD/G-MCP 惩罚三项 logit 模型, 同时显著优于 SVM、RF、ANN 这三种传统的机器学习方法。这表明选择重要的分组同时选择组内重要的特征可以提高预测的性能。同时, 对于不同的股票数据集, 模型选择的重要特征也存在差异。

参考文献

- [1] Biu, G.S. and Kusuma, P.K. (2023) Stock Market Volatility Analysis during the Global Financial Crisis: Literature Review. *Educational Journal of History and Humanities*, **6**, 2510-2520.
- [2] Engle, R.F. (1982) Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation. *Econometrica*, **50**, 987-1007. <https://doi.org/10.2307/1912773>
- [3] Afeef, M., Ihsan, A. and Zada, H. (2018) Forecasting Stock Prices through Univariate ARIMA Modeling. *NUML International Journal of Business & Management*, **13**, 130-143.
- [4] Ballings, M., Van den Poel, D., Hespeels, N. and Gryp, R. (2015) Evaluating Multiple Classifiers for Stock Price Direction Prediction. *Expert Systems with Applications*, **42**, 7046-7056. <https://doi.org/10.1016/j.eswa.2015.05.013>
- [5] Yun, K.K., Yoon, S.W. and Won, D. (2021) Prediction of Stock Price Direction Using a Hybrid GA-XGBoost Algorithm with a Three-Stage Feature Engineering Process. *Expert Systems with Applications*, **186**, Article 115716. <https://doi.org/10.1016/j.eswa.2021.115716>
- [6] Long, J., Chen, Z., He, W., Wu, T. and Ren, J. (2020) An Integrated Framework of Deep Learning and Knowledge Graph for Prediction of Stock Price Trend: An Application in Chinese Stock Exchange Market. *Applied Soft Computing*, **91**, Article 106205. <https://doi.org/10.1016/j.asoc.2020.106205>
- [7] Vuong, P.H., Dat, T.T., Mai, T.K., et al. (2022) Stock-Price Forecasting Based on XGBoost and LSTM. *Computer Systems Science and Engineering*, **40**, 237-246. <https://doi.org/10.32604/csse.2022.017685>
- [8] Yuan, M. and Lin, Y. (2006) Model Selection and Estimation in Regression with Grouped Variables. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, **68**, 49-67. <https://doi.org/10.1111/j.1467-9868.2005.00532.x>
- [9] Huang, J., Breheny, P. and Ma, S. (2012) A Selective Review of Group Selection in High-Dimensional Models. *Statistical Science*, **27**, 481-499. <https://doi.org/10.1214/12-sts392>
- [10] Huang, J., Ma, S., Xie, H. and Zhang, C. (2009) A Group Bridge Approach for Variable Selection. *Biometrika*, **96**, 339-355. <https://doi.org/10.1093/biomet/asp020>
- [11] Breheny, P. and Huang, J. (2009) Penalized Methods for Bi-Level Variable Selection. *Statistics and Its Interface*, **2**, 369-380. <https://doi.org/10.4310/sii.2009.v2.n3.a10>
- [12] Wu, T.T. and Lange, K. (2008) Coordinate Descent Algorithms for Lasso Penalized Regression. *The Annals of Applied Statistics*, **2**, 224-244. <https://doi.org/10.1214/07-aos147>
- [13] Breheny, P. (2015) The Group Exponential Lasso for Bi-Level Variable Selection. *Biometrics*, **71**, 731-740. <https://doi.org/10.1111/biom.12300>
- [14] Glonek, G.F.V. and McCullagh, P. (1995) Multivariate Logistic Models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, **57**, 533-546. <https://doi.org/10.1111/j.2517-6161.1995.tb02046.x>
- [15] Novoselova, N., Della Beffa, C., Wang, J., Li, J., Pessler, F. and Klawonn, F. (2014) HUM Calculator and HUM Package for R: Easy-to-Use Software Tools for Multicategory Receiver Operating Characteristic Analysis. *Bioinformatics*, **30**, 1635-1636. <https://doi.org/10.1093/bioinformatics/btu086>
- [16] Li, J., Gao, M. and D'Agostino, R. (2019) Evaluating Classification Accuracy for Modern Learning Approaches. *Statistics in Medicine*, **38**, 2477-2503. <https://doi.org/10.1002/sim.8103>
- [17] Hu, X. and Yang, J. (2024) G-LASSO/G-SCAD/G-MCP Penalized Trinomial Logit Dynamic Models Predict up Trends, Sideways Trends and down Trends for Stock Returns. *Expert Systems with Applications*, **249**, Article 123476.

- <https://doi.org/10.1016/j.eswa.2024.123476>
- [18] Hu, X. and Yang, J. (2024) Group Penalized Multinomial Logit Models and Stock Return Direction Prediction. *IEEE Transactions on Information Theory*, **70**, 4297-4318. <https://doi.org/10.1109/tit.2024.3376751>
 - [19] Vincent, M. and Hansen, N.R. (2014) Sparse Group Lasso and High Dimensional Multinomial Classification. *Computational Statistics & Data Analysis*, **71**, 771-786. <https://doi.org/10.1016/j.csda.2013.06.004>
 - [20] Rifkin, R. and Klautau, A. (2004) In Defense of One-vs-All Classification. *Journal of Machine Learning Research*, **5**, 101-141.