【统计理论与方法】

基于 Knockoff-Logistic 的多因子量化选股研究

王小燕，周 颖，唐婷婷，张中艳
（湖南大学 金融与统计学院，湖南 长沙 410079）

摘要：在数据驱动时代，如何挖掘金融资产的信息、挑选恰当的资产，对稳定收益、控制风险意义重大。多因子量化模型是选择股票的常用方法，选取最优解释力的因子集合是其主要目的之一。现有因子选择方法没有考虑到控制错误发现率（FDR），不利于构建稳健的投资策略。为此，在 Logistic 回归的基础上引入 Knockoff 方法进行因子选择，通过 Lasso 实现因子选择，利用 Knockoff 控制变量选择的 FDR 从而提高准确率。基于所选因子，在 Logistic 回归下进行股票预测，并与线性判别分析、支持向量机以及随机森林模型的预测结果进行对比。对沪深 300 指数和中证 500 指数成分股 2007—2020 年的数据进行实证研究，采用滑动回归法进行收益预测，并建立季度换仓的投资策略。研究表明，从变量选择上来看，基于 Knockoff 方法选出的因子所构造的选股模型具有更好的市场表现；从模型对比上来看，Logistic 回归预测的投资组合具备高收益、低风险的优势。综合来看，将 Knockoff 方法引入到多因子选股模型有利于提高因子选择的准确度，对优化资产配置具有参考意义。

关键词：监督型分类；错误发现率；因子选择；投资组合；变量选择

中图分类号：F830 **文献标志码**：A **文章编号**：1007－3116(2023)04－0019－14

引用格式：王小燕，周颖，唐婷婷，等. 基于 Knockoff-Logistic 的多因子量化选股研究[J]. 统计与信息论坛，2023，38(4)：19-32.

Citation Form：WANG X Y，ZHOU Y，TANG T T，et al. Multi-factor quantitative stock selection based on Knockoff-Logistic[J]. Journal of statistics and information，2023，38(4)：19-32.

一、引 言

2021 年，中国人民银行召开金融科技委员会会议，出台了新阶段金融科技发展规划。会议强调在科技与经济融合成果的基础上，以技术与数据双轮驱动金融创新为重点方向，深化金融数据应用。量化投资依托以机器学习为代表的大数据及人工智能技术，对蕴含在金融活动中的数据信息进行分析与决策，优化资产管理模式，在应用中充分发挥了“技术＋数据”的新动能。

针对如何从众多风险资产中建立最优资产组合的问题，学术界首先提出了资本资产定价模型（CAPM），它是一种单因子模型，重点关注风险资产收益与市场风险的数量关系[1]。随后，多因子模型被提出，并成为运用最广泛的选股模型之一，然而已有研究认为因子数量和因子类别都具有不确定性[2]。目前最有代表性的多因子模型是 Fama-French 三因子模型和 Fama-French 五因子模型[3-4]。虽然它们都从不同角度考虑了多个因子，但是考虑到金融市场的多变性与差异性，特别是大数据时代金融产品的数字特征跟随市场瞬息万

收稿日期：2022－02－18

基金项目：国家自然科学基金面上项目“多源数据融合的高维整合分析分类模型及其信用风险应用”（72271088）；教育部人文社科基金青年项目“基于多源数据的高维分类模型及其信用风险预警研究”（22YJC910012）；长沙市自然科学基金项目“大数据的整合分析分类模型及其违约风险管理应用研究”（kq2202180）

作者简介：王小燕（通讯作者），女，湖南娄底人，博士，副教授，博士生导师，研究方向：高维数据分析与数据挖掘；
周 颖，女，江苏淮安人，研究方向：大数据分析；
唐婷婷，女，贵州黔西人，研究方向：大数据分析；
张中艳，女，贵州毕节人，博士生，研究方向：高维数据分析。

变,有效因子的数量和类别也可能在不断变更,这对传统的多因子模型提出了新的挑战。因此,如何稳定地识别到能够解释超额收益来源的特征因子、构建稳健的多因子模型,值得进一步研究。

多因子选股模型的优劣主要取决于因子选择与模型设定,其中因子选择是统计学中典型的变量选择问题,代表性方法有子集选择法、降维法、正则化方法等。子集选择法的思想是从备选因子中选择最优的子集构建模型,具有代表性的是逐步回归[5]。子集法的思想简单,但是计算量较大,尤其是当因子维度很高时,计算成本可能呈指数级增长。降维法在保留原始变量的大部分信息的同时,将多个因子降维至少数几个综合变量,比如 Jiang 等使用的偏最小二乘(PLS)、PCA 和组合预测法(FC),Kelly 等提出的 IPCA 方法[6-7]。降维法尽管能够压缩因子维度,但是以损失原始变量的部分信息为代价,这可能影响模型预测性能。正则化方法是近 20 年来学术界广泛使用的方法。比如蒋翠侠等利用 Lasso 分位数回归挑选出对对冲基金收益有重要影响的风险因子[8]。李斌等利用 Lasso 回归、岭回归、弹性网等方法对 96 个异象因子系数进行压缩[9]。舒时克和李路对比了 Elastic Net、SCAD、MCP 惩罚项下的 Logistic 回归模型,发现它们能够很好地保留重要的变量[10]。在面对大量横截面因子时,Lasso、Elastic Net 等正则化方法被认为能够高效识别具有解释能力的因子指标,有利于提高股票收益的可预测性[8-9]。正则化具有良好的统计理论性质,且在选择因子的同时对因子的回归系数也进行估计,因此具有良好的稳健性。

上述方法均能实现因子筛选,然而在实际中,难以知晓所选因子是否为真正有价值和解释力的正确因子。特别在金融大数据背景下,数据的可变性和时效性都非常强,因子选择的结果可能不断变化,这就需要确保因子选择的准确性,控制因子选择的错误发现率(False Discovery Rate,FDR)(或假阳性)。已有研究中,Giglio 等提出了 Benjamini-Hochberg 法(BHq)对线性资产定价模型中基金选择的 FDR 进行控制[11-12]。然而相对于 BHq 方法,Barber 和 Candès 提出的 Knockoff 方法在理论上和应用上都具有更佳的表现,成为近年来控制变量选择 FDR 最受欢迎的方法[13]。目前关于 Knockoff 方法的研究,主要集中于“如何构造 Knockoff 变量”和“Knockoff 应用”等两个方面。在 Knockoff 变量构建方面,不少学者从变量结构、变量维度等方面进行了多项研究[14-15]。也有学者利用该方法解决一些变量选择的实际问题[16-17]。综合来看,Knockoff 方法在特征筛选方面具有控制 FDR、提高特征选择准确率的效果[13,18],因此,本文考虑将 Knockoff 方法纳入多因子选股模型用于控制因子选择的 FDR 问题,提高模型的预测能力和稳健性。

模型设定也是多因子选股模型的另一个重要研究内容,已有研究可分为统计模型和机器学习模型。统计模型中最具有代表性的为 Logistic 回归,代表性研究有舒时克和李路等[10,19-20]。机器学习模型中,神经网络、支持向量机、集成神经网络、深度神经网络(DNN)、随机森林等都被用于股票收益预测、制定短期套利策略等问题的研究[9,21-22]。机器学习模型能够训练大量样本数据,重点关注预测,但存在普遍公认的“黑盒问题”,即模型缺乏可解释性。统计模型弥补了这一缺点,它具有良好的解释能力和预测能力,所以被广泛应用。

综上所述,本文以提高因子筛选的正确率为目的,在 Logistic 回归下利用 Knockoff 方法构建有效因子体系,进而基于所选因子在统计模型和机器学习模型下进行预测。主要贡献在于:第一,引入正则化方法对因子进行选择,尽管在已有文献中有类似的研究,然而本文的特色在于构建 Knockoff 变量控制了因子识别的 FDR,确保因子选择的准确性。第二,首次将 Knockoff 方法应用于量化多因子模型,较对比方法能更好地兼顾投资组合的安全性和收益性,构建了相对稳健的选股策略。第三,以沪深 300 指数和中证 500 指数样本股为研究对象,利用 Logistic 回归构建多因子模型,对比线性判别分析、支持向量机及随机森林模型的预测性能,比较其构建的投资策略的市场表现,多角度验证了 Knockoff 方法在多因子量化选股中的有效性。

二、模型构建

本文选取沪深 300 指数和中证 500 指数的成分股作为混合股票池,所有股票的量化因子数据构成解释变量矩阵 X,股票收益是否高于相应的基准指数为响应变量 Y。图 1 展示了本文的总体设计。

首先,在因子选择阶段,本文采用 Knockoff-Logistic 回归模型;其中,Lasso 方法用于筛选有效因子,该变量选择方法在理论上具有良好的统计性质,且被广泛应用于基因筛选、投资组合、违约风险因素识别等方面。Knockoff 方法用于控制因子选择的 FDR,确保因子选择的准确性。

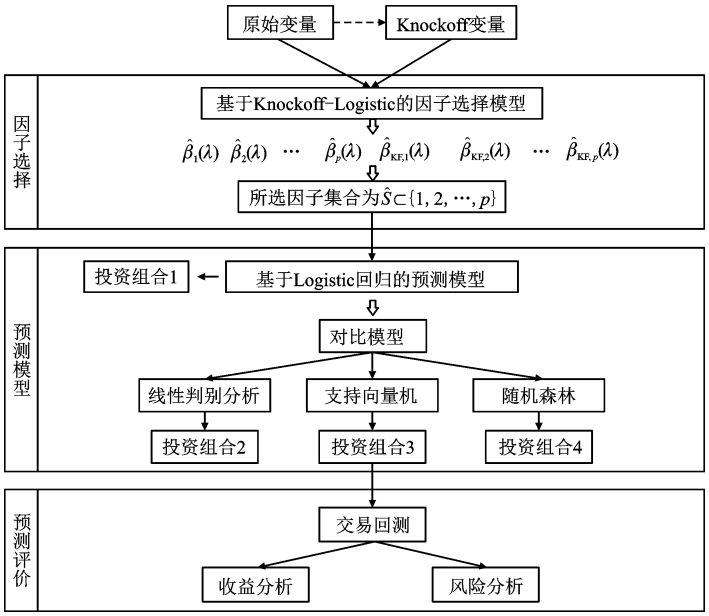

图 1　模型总体设计

其次，本文采用第一阶段选出的因子集合 $\hat{S}$ 拟合 Logistic 回归以及其他三种机器学习分类预测模型：线性判别分析、支持向量机和随机森林，构建多因子选股策略，并在测试集上进行回测，分析本文所选因子的合理性、有效性以及所构建模型的实际表现。由于中国股票市场做空机制不灵活，本文选取了最可能跑赢大盘的十只股票构建多头组合，实行买入并持有策略。利用上述四种模型分别构建四种投资策略，借助 Wind 系统的组合管理模块进行交易回测，以沪深 300 指数和中证 500 指数为业绩基准，评价各策略的绩效表现。

（一）因子选择模型

1. Lasso-Logistic 回归模型

设 Y 为二元响应变量，$Y_i=1$ 表示成分股 i 在 t 期的收益 R_u 大于相应的指数收益 R_t，反之 $Y_i=0$。$X=(X_1,X_2,\cdots,X_p)$ 表示 p 维因子指标，记第 i 只股票的观测值为 (x_i,y_i)。股票预期收益率大于市场基准的 Logistic 模型为：

$$\mathrm{Prob}(y_i=1\mid x_i)=\frac{\exp(x_i^{\mathrm{T}}\beta+\beta_0)}{1+\exp(x_i^{\mathrm{T}}\beta+\beta_0)} \tag{1}$$

其中，β_0 为截距项，β 为因子系数向量。采用极大似然法进行参数估计，对数似然函数表示为：

$$\ln L=\sum_{i=1}^{n}\{y_i\ln\mathrm{Prob}(y_i=1\mid x_i)+(1-y_i)\ln\mathrm{Prob}(y_i=0\mid x_i)\} \tag{2}$$

以最小化其负向对数似然函数为目标，求解可得各因子系数的极大似然估计值：

$$\hat{\beta}=\arg\min_{\beta}\left\{-\sum_{i=1}^{n}[y_i(x_i^{\mathrm{T}}\beta+\beta_0)-\ln(1+\exp(x_i^{\mathrm{T}}\beta+\beta_0))]\right\} \tag{3}$$

当因子维度较高时，往往存在对获取超额收益无显著作用的定价因子，此时需要对因子进行显著性识别，Lasso 是解决这类问题的常用方法，通过向目标函数引入 L_1 范数惩罚项，控制最终进入模型的因子数量，可以得到 Lasso-Logistic 回归的最优参数估计：

$$\hat{\beta}(\lambda)=\arg\min_{\beta}\left\{-\sum_{i=1}^{n}[y_i(x_i^{\mathrm{T}}\beta+\beta_0)-\ln(1+\exp(x_i^{\mathrm{T}}\beta+\beta_0))]+\lambda\sum_{j=1}^{p}\mid\beta_j\mid\right\} \tag{4}$$

其中，式(4)的第二部分为 Lasso 惩罚，λ 是调整参数，在损失函数和惩罚函数之间实现平衡。λ 越大，对参数估计值的压缩程度越大，部分参数被压缩为 0，从而达到因子选择的效果。

2. Knockoff 方法

Lasso 惩罚方法能够选择因子，具有稳健应对数据变动、计算方便等优点，但该方法在选择因子的同时并没有控制变量选择的 FDR，这很可能降低模型预测的准确性。为此 Barber 和 Candès 提出了一种控制假

阳性的方法——Knockoff 方法，它能在进行变量选择的同时控制错选变量所占的比例[13]。

该方法主要包含两大内容：Knockoff 变量构建和 Knockoff 模型估计（含 FDR 控制）。Knockoff 变量的构造是依据原始观测 X 来实现的。对每一个解释变量 X_j，通过模仿原始变量的相关性结构，构造 Knockoff 变量 $X_{KF,j}$，假定 $X_{KF}=(X_{KF,1},X_{KF,2},\cdots,X_{KF,p})$。

在标准化数据的基础上，计算出原始解释变量 X 的 Gram 矩阵 $\Sigma=X^{T}X$，为了确保 Knockoff 变量 X_{KF} 具备原始变量的相关结构，要求 X_{KF} 满足以下两个性质：

$$\Sigma=X^{T}X=X_{KF}^{T}X_{KF}$$

$$X^{T}X_{KF}=\Sigma-\mathrm{diag}\{s\} \tag{5}$$

因此，Barber 和 Candès 提出一种满足上述构造条件的 Knockoff 变量 X_{KF} 为[13]：

$$X_{KF}=X(1-\Sigma^{-1}\mathrm{diag}\{s\})+\tilde{U}C \tag{6}$$

其中，s 为 p 维非负向量，应选择尽可能大的 s，使 X_j 与 $X_{KF,j}$ 趋于正交，避免二者过于相似。s 的确定是一个使原始变量与 Knockoff 变量间平均相关性最小化的半正定规划问题：

$$\min\sum_{j}(1-s_j) \tag{7}$$

$$\text{s.t.}\quad 0\leqslant s_j\leqslant 1,\mathrm{diag}\{s\}\leq 2\Sigma$$

同时假设 Σ 是可逆的，保证模型能够被有效识别。$\tilde{U}$ 为 $n\times p$ 维矩阵，与 X 正交，即 $\tilde{U}^{T}X=0$。矩阵 C 满足 $C^{T}C=2\mathrm{diag}\{s\}-\mathrm{diag}\{s\}\Sigma^{-1}\mathrm{diag}\{s\}$。

除此之外，Gégout-Petit 等在 Barber 和 Candès 的基础上，为使 Knockoff 方法能广泛应用于各种回归且不受维度的限制，通过随机交换设计矩阵 X 的行构造 Knockoff 变量 X_{KF}，以确保 X_{KF} 之间的相关性与原始变量 X 相同，但 X_{KF} 与响应变量 Y 不相关。本文实证分析中采用了这种随机交换设计矩阵 X 的行的方法来构造 Knockoff 变量[13,15]。

3. 基于 Knockoff-Logistic 的因子选择

基于增广矩阵$[X\ X_{KF}]$与因变量 Y 进行建模，并设它们的回归系数为$[\beta\ \beta_{KF}]$。此时，用$[X\ X_{KF}]$构建如下 Knockoff-Logistic 回归：

$$(\hat{\beta}(\lambda),\hat{\beta}_{KF}(\lambda),\hat{\beta}_0)=\underset{\beta,\beta_{KF},\beta_0}{\arg\min}\Big\{-\sum_{i=1}^{n}[y_i(x_i^{T}\beta+x_{KF,i}^{T}\beta_{KF}+\beta_0)-\ln(1+\exp(x_i^{T}\beta+x_{KF,i}^{T}\beta_{KF}+\beta_0))]+\lambda\sum_{j=1}^{p}(|\beta_j|+|\beta_{KF,j}|)\Big\} \tag{8}$$

根据估计结果，可计算一个 $2p$ 维的向量$(Z_1,Z_2,\cdots,Z_p,Z_{KF,1},Z_{KF,2}\cdots,Z_{KF,p})$，其中 $Z_j=\sup\{\lambda:\hat{\beta}_j(\lambda)\neq 0\}$，$Z_{KF,j}=\sup\{\lambda:\hat{\beta}_{KF,j}(\lambda)\neq 0\}$。对每个 $j\in\{1,2,\cdots,p\}$，计算统计量为：

$$W_j=Z_j\vee Z_{KF,j}\begin{cases}+1, & Z_j>Z_{KF,j}\\ -1, & Z_j\leqslant Z_{KF,j}\end{cases} \tag{9}$$

W 统计量将起到控制 FDR 的作用，W_j 取正值表明原始变量比 Knockoff 变量更早地进入模型，W_j 取值越大对应变量 X_j 越可能是模型中的真实变量。但 W_j 的取值需要达到多大才能被选入模型，需要计算出一个明确的阈值 τ。为此，Gégout-Petit 等结合了 CUSUM 方法和 Auger 等人的分段邻域识别算法来确定阈值 τ[15,23-24]。该算法首先设 W 中有 ω 个正值，计算出这一部分的次序统计量 $0<W_{(1)}\leqslant W_{(2)}\leqslant\cdots\leqslant W_{(\omega)}$，定义 $e_j=W_{(j+1)}-W_{(j)}$，然后基于 e_j 计算阈值 τ。具体过程如下：

(1)基于 CUSUM 方法(Cumulative Sum)计算分割点 T_1。CUSUM 方法的主要思想是对样本信息进行累加，通过累计过程中每次的偏移量，达到放大的效果。当统计量的 CUSUM 明显高于或低于平稳条件 $\bar{e}=\frac{1}{\omega-1}\sum_{j=1}^{\omega-1}e_j$ 时，则意味着系统发生了变化。用 S_j 表示 e_j 在位置 j 处相对均值的偏移量，C_j 表示位置 j 处的累计偏移量：

$$S_j=e_j-\bar{e}\quad(j=1,2,\cdots,\omega-1) \tag{10}$$

$$C_j=\begin{cases}S_1, & j=1\\ C_{j-1}+S_j, & j=2,3,\cdots,\omega-1\end{cases} \tag{11}$$

分割点 T_1 是使得 e_j 相对于均值 $\bar{e}$ 的累计偏移最大化的位置，即 $T_1=\arg\max_j|C_j|$。

(2)基于 Auger 和 Lawrence 提出的分段邻域识别算法计算分割点 T_2。首先对每个可能的分段分别计算离差平方和，以衡量每段序列的分散程度：

$$R_{ij}=\begin{cases}\sum_{l=i+1}^{j}\left(e_l-\frac{1}{j-i}\sum_{k=i+1}^{j}e_k\right)^2, & i<j\\ +\infty, & i\geqslant j\end{cases} \tag{12}$$

其中，R_{ij} 表示$(i,j]$序列区间上的离差平方和。采用动态规划将一组相邻的分段序列拼凑为整个序列，根据组内平方和最小化原则选择将序列分为两段的最优分割点。令初始值 Testlik=$+\infty$，对 $k=1,2,\cdots,\omega-2$，如果 $R_{1,k}+R_{k,\omega-1}<$Testlik，则更新 Testlik$=R_{1,k}+R_{k,\omega-1}$，$T_2=k$，重复以上步骤，直至 $R_{1,k}+R_{k,\omega-1}>$Testlik，则最优分割点为 $T_2=\arg\min_k R_{1k}+R_{k,\omega-1}$。

(3)令 $e_j^*=-e_j$，$j=1,2,\cdots,\omega-1$，基于 e_j^* 重复上述过程(1)～(2)，可以计算出分割点 T_3 和 T_4。

(4)确定阈值 τ。比较四个分割点的大小，取最小值为最终分割点，记为 $T=\min(T_1,T_2,T_3,T_4)$，则阈值为 $\tau=W_{(T+2)}$，最终选择的变量集为 $\hat{S}:=\{j:W_j\geqslant\tau\}$，表示选入模型的变量的下标集合，则 FDR 定义为：

$$\mathrm{FDR}=E[\mathrm{FDP}]=E\left[\frac{\#\{j:\beta_j=0,j\in\hat{S}\}}{\#\{j:j\in\hat{S}\}\vee 1}\right] \tag{13}$$

(二)基于 Logistic 回归的预测模型

假定基于 Knockoff 方法选择的变量集 $\hat{S}=\{j:W_j\geqslant\tau\}$ 含有 m 个变量，记选出的因子集为 $X_{\hat{S}}\in\mathbb{R}^{n\times m}$，若无特殊说明，下文 $X_{\hat{S}}$ 均指 Knockoff 选择出的因子作为解释变量。根据 $X_{\hat{S}}$ 和 Y，构建基于 Logistic 回归的预测模型，用于量化投资决策分析。

首先，在训练集上，用 Knockoff-Logistic 选出的因子拟合 Logistic 回归，求解模型系数：

$$\hat{\beta}_{\hat{S}}=\arg\min_{\beta}\left\{-\sum_{i=1}^{n}\left[y_i(x_{i,\hat{S}}^{\mathrm{T}}\beta+\beta_0)-\ln(1+\exp(x_{i,\hat{S}}^{\mathrm{T}}\beta+\beta_0))\right]\right\} \tag{14}$$

然后，将测试集数据代入模型，预测股票跑赢业绩基准的概率为：

$$\hat{P}(y_i=1|x_{i,\hat{S}})=\frac{\exp(x_{i,\hat{S}}^{\mathrm{T}}\hat{\beta}_{\hat{S}}+\hat{\beta}_0)}{1+\exp(x_{i,\hat{S}}^{\mathrm{T}}\hat{\beta}_{\hat{S}}+\beta_0)} \tag{15}$$

最后，将预测概率降序排列，选取概率最大的十只股票构建投资组合。

本文的预测模型中，除 Logistic 回归外，采用了线性判别分析、支持向量机和随机森林等三种常用的模型作为对比，它们的相关理论不再赘述，可参考李斌、Krauss、王小燕等人的研究[21-22,25]。

三、数据说明和指标体系构建

(一)数据收集与处理

沪深 300 指数成分股由沪深两市具有代表性的 300 只股票构成，具备流动性强、规模大的特点；中证 500 指数成分股是 A 股市场一批中小市值公司的股票，每半年调整一次样本股，故本文选取 2021 年上半年入选沪深 300 指数和中证 500 指数的股票作为选股对象，避免股票池过于单一。以 Wind 数据库 190 个量化因子指标的季度数据与股票收益作为全部样本，选定 2007 年 3 月至 2020 年 12 月为研究区间，进行实证研究。

比较个股收益率与同期大盘收益率，若大于大盘则响应变量记为 1；否则为 0。模型以依据本期因子数据预测下期收益情况为目的，因此对因子指标做出相对于响应变量滞后一期的处理，即 T 期的因子数据对应 $T+1$ 期的响应变量。

原始数据存在部分缺失，缺失值的处理分为删除与插补两个步骤。第一步，先剔除收益率缺失的记录；再删除缺失比例大于 15%的因子指标以及因子值缺失 10%以上的记录。第二步，区分两种情形插补缺失值：若某只股票在某个因子上的数据不完全缺失，则使用时间序列中的线性插值法填补缺失值；若某只股票在某个因子上的数据全部缺失，则以与该股票属于同一行业的股票的同期数据的均值填补。第三步，对因子数据进行标准化处理，避免不同因子的取值量纲对预测效果和机器学习算法迭代收敛速度的影响。

(二)多因子指标体系构建

考虑到因子指标之间可能存在高度可替代性,本文参考《中国A股市场量化因子白皮书》等相关文献剔除了不具有代表性的冗余因子,最终保留了79个因子构成本文的多因子指标体系,按照因子属性可以划分为7大类,分别为财务流动性因子、成长因子、盈利因子、估值因子、规模因子、一致预期因子和技术因子。下文将选取部分因子对每个类别的含义及入选原因进行说明,详细的因子体系已省略,需要了解的可联系作者。

1.财务流动性因子

财务流动性指标常用于衡量企业资金的使用效率,包含16个因子。其中最具代表性的因子是流动比率与速动比率。这两项指标的分母都是流动负债,分子分别是流动资产与速动资产,反映了企业对于短期债务的偿还能力。与之对应,产权比率与股东权益比率能够度量企业的长期偿债能力。此外,本文还选取了一系列资产的周转率,以评价企业资产管理的效率,如应收账款、应付账款、流动资产、存货以及固定资产等。周转率指标可以视作对偿债能力的补充。

2.成长因子

成长因子反映了股票的成长性,含有7个因子。公司的基本活动可以划分为三类:筹资活动、投资活动与经营活动,分别可以使用筹资活动、投资活动产生的现金流量净额的增长率以及营业收入增长率这三项指标来描述三项活动的成长变化。此外,还使用了净现金流量增长率,起到综合描述的作用。

3.盈利因子

盈利因子,顾名思义就是描述公司赚取利润能力的因子,共有16个因子。首先,选取了杜邦分析体系中的核心指标,净资产收益率、资产回报率、资本报酬率,它们依次代表了股东权益、总资产和投入资金创造净利润的能力。其次,选取了"基本每股收益"反映普通股股东的当期净利润;选取"稀释每股收益",在基本每股收益的基础上考虑了稀释性潜在普通股转换为普通股的可能。二者描述普通股股东可能拥有的净利润。最后,也考虑了销售毛利率,粗略地反映了公司控制成本的能力,但该指标在不同行业间缺乏可比性。从持续经营角度考虑,还选取了5年平均权益回报率、5年平均资产回报率等长期指标,以衡量盈利水平的稳定程度。

4.估值因子

估值因子主要考虑了股票的价值,包括5个因子。企业价值评估中常用的三大指标为市盈率、市现率、市销率,三者分别是普通股每股市价与每股收益、每股现金流量、每股营业收入的比值,能够反映投资者对公司未来前景的预期。

5.规模因子

规模因子共包括18个因子。流通市值是规模因子中的基础指标,营收市值比、收益市值比、账面市值比都是以流通市值指标为分母构建的,分子依次是营业收入、净利润、流通股数除以总股数乘以所有者权益。由于流通市值量级较大,所以本文选取对数流通市值作为候选指标,同时也避免了与上述三项指标的高相关性。

6.一致预期因子

一致预期因子考虑了分析师对股票的预期是否会影响收益率或能否解释收益率,共计两个因子。分析师报告的指标主要有三类:盈利预测、目标价及评级。结合预测股票涨跌信息的目的,本文选取了盈利预测的指标。

7.技术因子

技术因子以价格、成交量等作为计算的基础,得到一些判断价格走势的数据或曲线,也就是技术分析的数学表达,共计15个因子。在参考文献中出现次数较多的此类指标有能量潮指标、心理线指标、随机指标KDJ的K值、D值和J值等,本文也将这些指标纳入了量化因子体系。

(三)描述性统计分析

以沪深300指数和中证500指数为业绩基准,当个股收益率高于相应大盘时,响应变量Y记为1,反之记为0。经过数据清理与因子指标体系构建,最终获得了79个量化因子的26 589条样本记录,即$n=$

26 589，$p=79$。图 2 展示了每一个季度的有效样本量和响应变量的分类占比情况，每个季度的有效观测最少有 295 条，最多为 630 条；收益率高于大盘的样本占比为 46.68%，样本分布较为均衡。

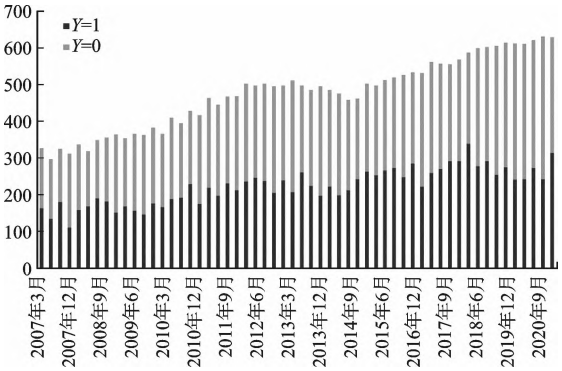

图 2　各个季度有效样本量

本文接下来对 7 类因子进行描述性统计分析。由于本文所用数据类型为面板数据，在进行描述统计分析时，先对各季度横截面数据计算描述统计量，再对其求时间序列均值。

首先分析因子间的相关性，考虑到因子维度较高，无法一一展示单因子间的相关性，因此按照类别展开相关性分析。用对应类别因子间两两相关系数的绝对值的均值表示类内或类间相关程度，计算结果如表 1 所示，其中非对角线元素表示两类因子间的平均相关程度，对角线元素表示每个类别内部不同因子间的平均相关水平。

表 1　各类因子间相关性描述

候选因子	财务流动性因子	成长因子	盈利因子	估值因子	规模因子	一致预期因子	技术因子
财务流动性因子	0.183	0.027	0.097	0.053	0.053	0.019	0.069
成长因子	0.027	0.097	0.060	0.018	0.033	0.014	0.029
盈利因子	0.097	0.060	0.195	0.030	0.053	0.031	0.095
估值因子	0.053	0.018	0.030	0.121	0.046	0.024	0.028
规模因子	0.053	0.033	0.053	0.046	0.233	0.013	0.075
一致预期因子	0.019	0.014	0.031	0.024	0.013	0.049	0.044
技术因子	0.069	0.029	0.095	0.028	0.075	0.044	0.241

从表 1 可以看出，各类因子内部的相关性都较弱，在对角线元素中，除规模因子的内部相关性达到了 0.233，技术因子的内部相关系数达到 0.241，其余五类因子内部相关水平均小于 0.2；一致预期因子的内部相关性仅有 0.049，小于 0.05，相关性极弱。类间相关性总体处于较低的水平，最高的是财务流动性因子与盈利因子的相关性，为 0.097，规模因子与一致预期因子的相关性最低，仅有 0.13。各类因子内部的相关性都明显高于与其他类别因子间的相关性，这说明各类因子间的信息重叠比较少，能从不同角度刻画个股信息。

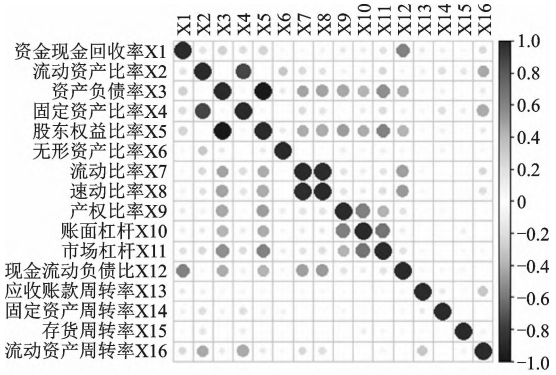

图 3　财务流动性因子间相关性

以 7 类因子中的财务流动性因子为例，对因子类内相关性进行具体分析。由图 3 可以看出，财务流动性因子之间的相关性总体并不显著，除速动比率与流动比率之间(0.987)、资产负债率与股东权益比率之间(0.999)存在高度相关关系，资金现金回收率与现金流动负债比之间(0.603)、流动资产比率与固定资产比率之间(−0.752)、资产负债率与市场杠杆(0.560)、股东权益比率与市场杠杆(−0.560)、速动比率与现金流动负债比之间(0.512)、账面杠杆与市场杠杆之间(0.671)呈现中度相关关系，其余因子之间仅存在微弱的相关性。

此外，经统计，单个因子两两之间相关性低于 0.2 的占 89.51%，相关性在 0.2 至 0.5 之间的因子占比 8.36%，相关性大于 0.5 的因子仅占 2.13%。详细的描述性统计已省略，需要了解的可联系作者。

四、多因子量化选股实证分析

(一)基于 Knockoff-Logistic 多因子选择结果

本文参考李斌等人研究，采用滑动窗口法选取训练集与测试集，窗口宽度固定为四个季度，窗口期作为

训练集，下一期为测试集。每次滑动完成后，窗口向后移动一期，依次进行模型的拟合与预测[9]。滑动窗口法尽可能地保留了时间序列的特征，并且具备交叉验证法的优点，充分利用了所有样本。

模型构建分为因子选择与模型拟合两个步骤。首先，进行因子选择，本文采用 Knockoff-Logistic 回归选择因子，同时为了验证 Knockoff 方法在选择因子方面的优势，以 Lasso-Logistic 回归作为对比方法来分析因子选择效果。每滑动一次，完成一次因子选择，记录被选择的因子。滑动结束后，统计各个因子指标被选入模型的频率。其次，利用选出的因子拟合 Logistic 回归与 3 个对比模型(线性判别分析、支持向量机以及随机森林)，进行对比分析。8 个模型的含义及简称具体见表 2。

表 2　八个模型含义及简称

预测模型	多因子选择模型	
	Knockoff-Logistic	Lasso-Logistic
Logistic	KF-L-L	Lasso-L-L
LDA	KF-L-LDA	Lasso-L-LDA
SVM	KF-L-SVM	Lasso-L-SVM
RF	KF-L-RF	Lasso-L-RF

表 3 分别展示了两种因子选择模型历经 52 次滑动累计选择次数排在前 15 名的因子指标。两种方法下，前 15 项指标中有 12 项重合，分别是对数流通市值、八个季度净利润变化趋势、市场杠杆、随机指标 KDJ 的 D 值与 J 值、固定资产比率、销售毛利率、12 月累计收益、流动资产比率、营收市值比、个股与市场相关系数、个股 120 日的 beta 值，其中有 5 项规模因子、3 项财务流动性因子、2 项技术因子、1 项成长因子和 1 项盈利因子。无论哪种因子选择模型，对数流通市值都是被选择的频率最高且唯一超过 50%的因子。由于取对数只会降低数据量级，对数流通市值与流通市值具备同等效力。李斌等通过单因子检验，验证了流通市值是影响中国股票截面收益的重要异象因子。可见，对数流通市值对股票收益胜率的影响是非常显著的[9]。

表 3　因子选择频率及秩

指标	Lasso-Logistic		Knockoff-Logistic	
	频率(%)	秩	频率(%)	秩
对数流通市值	67.31	1	63.46	1
八个季度净利润变化趋势	46.15	3	42.31	2
市场杠杆	40.38	8	40.38	3
随机指标 KDJ_D	42.31	6	40.38	3
随机指标 KDJ_J	44.23	5	40.38	3
固定资产比率	48.08	2	36.54	6
销售毛利率	38.46	12	36.54	6
12 月累计收益	42.31	6	34.62	8
流动资产比率	36.54	14	32.69	9
营收市值比	46.15	4	32.69	9
个股与市场相关系数	40.38	8	32.69	9
6 日能量潮指标	—	—	32.69	9
个股 120 日的 beta 值	40.38	8	30.77	13
资产负债率	—	—	28.85	14
个股 20 日的 beta 值	—	—	28.85	14
过去一个月收益率排名/股票总数的比值	40.38	8	—	—
资金现金回收率	38.46	12	—	—
基本每股收益	34.62	15	—	—

表 4 展示了两种回归模型滑动变量选择的数量序列的描述性统计。通过因子选择的结果来看，Lasso-Logistic 回归选择的变量个数极差为 46，波动范围较大，而 Knockoff-Logistic 选择的变量个数的极差仅为 16，模型更稳定，平均选择的变量个数更少，这是由于 Knockoff 方法在进行变量选择的同时控制了假阳性率，剔除更多的冗余变量，这与 Knockoff 方法为控制 FDR 而相对保守的特征吻合，Barber 和 Candès 通过模拟分析验证了此特征[13]。而 Lasso 倾向于选择更多的变量，尤其当变量相关性较强时，往往会选择不重要的变量进入模型，从而无法区分系数较小的变量和不重要变量，导致很高的假阳性[26]。

表 4　变量选择个数的描述统计

模型	Min	1st Qu	Median	Mean	3rd Qu	Max
Lasso-Logistic	4	11	18.5	19.2	24	50
Knockoff-Logistic	9	12	14	14.4	16	25

此外，本文还计算了两种方法之间的 MRV 系数，它表示变量选择所对应样本数据集之间的相似性，每次变量选择对应一个 MRV 值，MRV 越接近 1，表示两组样本数据的相似性越大[27-28]。52 次滚动回归下两种方法的 MRV 均值为 0.79，表明 Knockoff-Logistic 虽然选择的变量个数偏少，但与 Lasso-Logistic 所选变量具有较高的信息重叠。

（二）选股策略和回测分析

1. 选股策略

在 2017 年 3 月至 2020 年 12 月样本数据的基础上，使用滑动窗口法。每次滑动在训练集上用 Lasso 和 KF-Lasso 两种方法选择因子，再分别拟合 Logistic 回归、线性判别分析、支持向量以及随机森林模型进行参数估计。基于上述模型在测试集上对股票收益跑赢大盘的情况进行预测，将预测的概率值降序排列，选择前 10 只股票，构建季度换仓的投资策略。每季度初第一个交易日为调仓日，卖出当前持有的资产组合，将持有的资金等额分配给计划买入的每只股票并持有至季度末。将构建的 8 组选股策略分别按照持仓时间顺序导入 Wind 系统进行股票交易回测，持仓区间为 2018 年 4 月 1 日至 2021 年 3 月 31 日。

2. 基于 Logistic 的收益分析

基于 Logistic 模型构建的两种投资组合的累计回报变动和季度收益情况分别如图 4、图 5 所示。

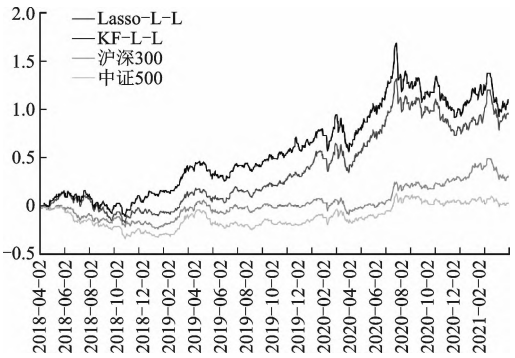

图 4　基于 Logistic 模型的累计回报

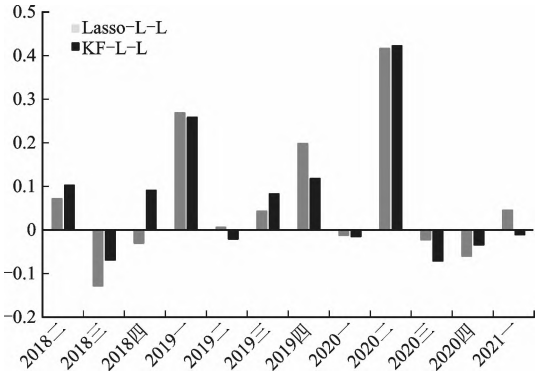

图 5　基于 Logistic 模型的季度收益

从图 4 可以看出，两种投资策略的收益变化趋势与沪深 300、中证 500 指数变动呈现相似的走势。回测区间内，沪深 300 指数的累计回报率始终高于中证 500 指数。2018 年第二季度初，Lasso-L-L 与 KF-L-L 策略呈上涨趋势，而基准指数处于下跌状态；KF-L-L 策略的累计回报最早由 2018 年第四季度末由负转正，此后领先基准收益的水平逐渐拉大，而 Lasso-L-L 策略则落后了一个季度才恢复正收益；KF-L-L 策略在 2020 年第三季度初，投资总回报达到了峰值，然后稍有波动，最终与 Lasso-L-L 策略的差距缩小，但仍远高于基准指数的累计回报。

图 5 所示的收益图，以季度为统计周期，展示投资策略每季度的盈亏状况，能够反映出相应模型对股票收益情况的预测能力。KF-L-L 策略有半数季度盈利，虽然比 Lasso-L-L 策略多一期亏损，但盈利水平较高，且最低盈利水平仍高于最大亏损程度，因此能够获得较高的累计总回报。

（三）回测对比分析

1. 与 LDA 模型对比

基于 LDA 预测模型构建的两种投资组合的累计回报变动和季度收益情况分别如图 6、图 7 所示。

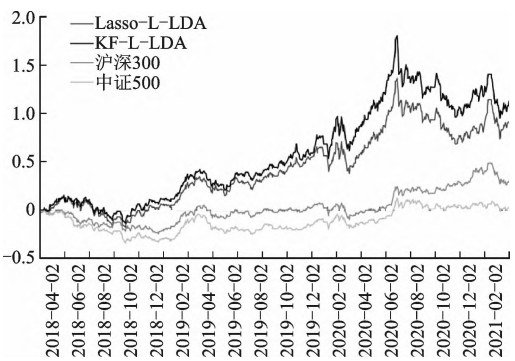

图 6　基于 LDA 模型的累计回报

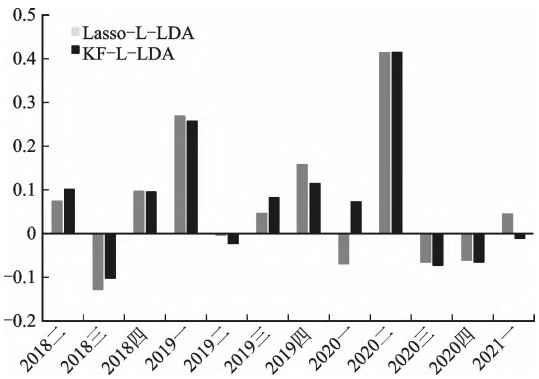

图 7　基于 LDA 模型的季度收益

如图 6 所示，基于 LDA 与 Logistic 模型的策略的累计回报曲线重合度较高。不同的是，直至 2019 年末，基于 LDA 模型的两种策略仍保持着较小的回报差异，且扭亏为盈的时间差更小。同样地，KF-L-LDA 和 Lasso-L-LDA 策略也是在 2020 年第三季度初达到最高累计回报。

从图 7 可以看出，基于 LDA 模型的策略都具有 7 个盈利季度，且盈亏情况仅有两个季度不一致。相较于 KF-L-L，KF-L-LDA 对 2020 年第一季度的预测更优越。但对比亏损期间可以发现，KF-L-L 策略的最大季度亏损不高于 10%，且仅有一个季度的亏损大于 5%，而 KF-L-LDA 策略有三个季度亏损 5%以上，最高亏损超过 10%。

2. 与 SVM 模型对比

基于 SVM 模型构建的两种投资组合的累计回报变动和季度收益情况分别如图 8、图 9 所示。

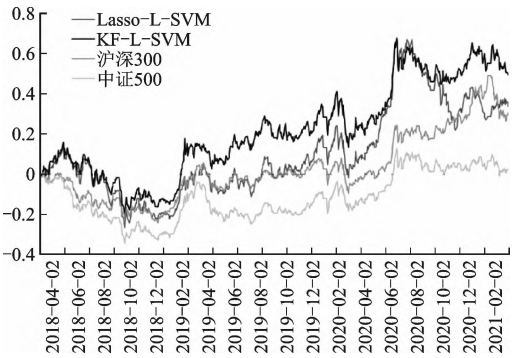

图 8　基于 SVM 模型的累计回报

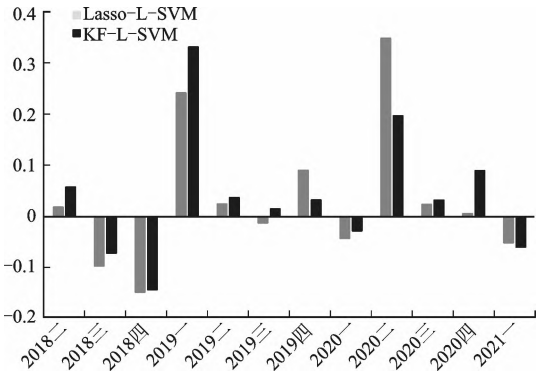

图 9　基于 SVM 模型的季度收益

从图 8 折线的走势可以看出，基于 SVM 模型的两种策略的累计回报在 2019 年第一季度与 2020 年第二季度发生两次较大幅度的持续增长。KF-L-SVM 策略与 Lasso-L-SVM 策略分别在这两个季度实现最大涨幅。在这两个季度之间的周期内，KF-L-SVM 策略的累计回报维持在高于 Lasso-L-SVM 策略的水平，Lasso-L-SVM 策略的收益围绕沪深 300 指数上下波动，但高于中证 500 指数。在最初两个季度及 2020 年第四季度，两种策略的回报率持平。在最后两个季度，Lasso-L-SVM 策略跌幅较大，两种策略的累计回报拉开差距。

在图 9 中，KF-L-SVM 策略与 Lasso-L-SVM 策略的最高季度收益分别位于 2019 年第一季度与 2020 年第二季度，这与图 9 的特征也是吻合的。KF-L-SVM 策略虽然有 8 个季度的正收益，但只有两个季度超过 10%，且这两种策略的最高季度收益率都不超过 35%。KF-L-L 策略有 4 个盈利周期高于 10%，最高达到了 42. 37%。

3. 与随机森林对比

基于随机森林(RF)构建的两种投资组合的累计回报变动各季度收益情况分别如图 10、图 11 所示。

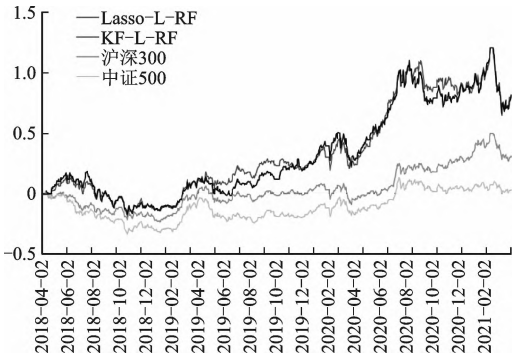

图 10　基于 RF 模型的累计回报

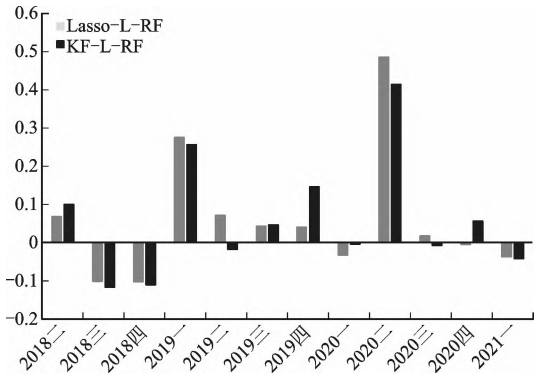

图 11　基于 RF 模型的季度收益

从图 10 可以看出，基于 RF 模型的两种策略均获得了稳定高于两项基准指数的超额回报。KF-L-RF 策

略与 Lasso-L-RF 策略在回测区间的多个周期内累计回报曲线几乎重合，甚至在 2019 年第三、四季度及 2020 年第四季度，Lasso-L-RF 策略的累计回报高于 KF-L-RF 策略。相较于 KF-L-L 策略，KF-L-RF 策略的亏损期更长，滞后两个季度恢复正收益。不同于基于 Logistic 的策略，基于 RF 的策略在最后一个季度内才实现最高累计回报。

如图 11 所示，基于 RF 模型的两种策略在 2018 年第三、四季度连续亏损超过 10%，导致期间累计回报持续为负。2019 年第二季度，LassoL-RF 策略盈利，而 KF-L-RF 亏损，Lasso-L-RF 的累计回报暂时超过 KF-L-RF 策略。Lasso-L-RF 策略的最高季度收益超过 50%，但由于最后两个季度的连续亏损导致 Lasso-L-RF 策略没有跑赢 KF-L-RF 策略。

（四）策略评价

综合比较上述 4 个模型的累计回报，Logistic 回归及 3 个对比预测模型构建的投资组合最终都获得了相对于两项基准指数的超额收益。除 RF 模型，在相同预测模型下，基于 Knockoff 方法的投资组合收益都明显优于对比方法的组合收益。比较季度收益可以看出，上述 8 种策略均在 2019 年第一季度和 2020 年第二季度迎来两次较大涨幅，且至少在半数周期内处于盈利状态。各策略的盈亏情况在大多数周期内是一致的，且符合大盘的波动特征，这说明模型的预测能力主要体现在盈亏的程度上。对比其他 3 个模型，Logistic 回归在持有期间内总体上保持着更高的收益和更低的亏损。

表 5 中各模型的投资组合季度平均收益的标准差体现了各投资策略收益的稳定程度。基于 Knockoff 方法的策略收益分布得更加集中，即盈亏波动的幅度相对较小。除 RF 模型外，Knockoff 方法对应的策略具有更高的季度平均收益。可以看出 LDA 模型的预测效果与 Logistic 模型最为相近，Lasso-L-L 策略比 Lasso-L-LDA 拥有更高的平均收益和更低的标准差；KF-L-L 策略的平均收益虽略低于 KF-L-LDA，但收益变动幅度更小。

表 5　各模型季度平均收益情况

模型	季度平均收益率	收益标准差
Lasso-L-L	0.066 6	0.154 8
KF-L-L	0.071 0	0.147 0
Lasso-L-LDA	0.064 9	0.156 0
KF-L-LDA	0.072 7	0.148 2
Lasso-L-SVM	0.033 5	0.139 3
KF-L-SVM	0.040 9	0.125 7
Lasso-L-RF	0.060 4	0.166 7
KF-L-RF	0.059 9	0.153 3

3 个对比预测模型中，LDA 模型的收益表现与 Logistic 回归相当，为更全面地比较基于不同变量选择方法及预测模型所构建的投资组合的收益高低与风险大小，本文还考虑了总回报、相对沪深 300 指数的总回报、年化平均回报、最大跌幅、最大跌幅天数、最大涨幅、最大涨幅天数、下行风险、年化波动率、Alpha、Beta 和夏普比率等评价指标，如表 6 所示。由于沪深 300 指数业绩表现优于中证 500 指数，因此选择沪深 300 指数作为相对指标的比较基准。

KF-L-L 和 KF-L-LDA 模型获得了大于 100%的总回报，分别是 107.22%和 110.40%，以及高于 25%的年化平均回报；全部策略相对于沪深 300 指数都获得了超额回报。4 组对比中，仅有 KF-L-L 相对于 Lasso-L-L 策略，提升了最大涨幅，降低了最大回撤。KF-L-RF 策略具有最长的最大涨幅天数和最短的最大跌幅天数，基于前 3 个模型的策略最大涨幅天数相同，除 KF-L-LDA 策略外，最大跌幅天数也相近。下行风险反映了随着市场环境变化，投资中出现最坏的情况时，投资者可能需要承担的损失。基于 LDA 模型的策略伴随着最高的下行风险，约 20%左右，Lasso-L-RF 策略的下行风险最低，为 17.22%，基于 Logistic 模型的策略位于二者之间。夏普比率是超额收益率与投资组合标准差的比值，反映了每承受一单位风险获得的超额收益。因此，投资者往往希望夏普比率越高越好。Alpha 表示投资组合相对于沪深 300 指数的超额收益，Beta 表示投资组合对基准指数的敏感程度。可以发现，除 RF 模型外，其他 3 个模型在引入 Knockoff 方法后都能够显著提升超额收益；相对于 KF-L-LDA，KF-L-L 策略对沪深 300 指数的业绩变动更不敏感，随着大盘变动的幅度更低。夏普比率最高的是 KF-L-LDA 模型，达到 0.91，其次是 KF-L-L 模型的 0.89，最低的是 Lasso-L-SVM 模型，仅有 0.31。比较两种变量选择方法，基于 Knockoff 方法的在收益方面的优势表现得更加明显，但在风险方面不太稳定，可以通过调整预测模型，提高 Knockoff 方法的稳定性。3 个对比模型中，LDA 和 RF 在部分指标上表现较为突出，SVM 的表现过于保守。在收益表现上，Logistic 模型明显优于 RF 模型，考虑收益波动、下行风险等风险指标时，Logistic 模型有效控制了部分损失。Logistic 回归相较

于其他 3 个模型达到了兼具提高超额收益与降低风险的作用。投资者对风险的偏好程度，也会影响其对投资组合的选择，风险偏好型的投资者可能更倾向于选择基于 LDA 模型的投资策略。综合考虑涉及的收益与风险指标，KF-L-L 策略表现出稳健的优势。

表 6　各模型投资策略的收益与风险评价

模型	总回报(%)	相对总回报(%)	年化平均回报(%)	年化波动率(%)	最大涨幅(%)	最大涨幅天数
Lasso-L-L	94.69	65.20	24.76	28.61	181.73	634
KF-L-L	107.22	77.73	27.37	29.05	202.53	634
Lasso-L-LDA	90.35	60.85	23.83	29.18	188.82	634
KF-L-LDA	110.40	80.90	28.02	29.28	225.39	634
Lasso-L-SVM	35.30	5.81	10.56	28.86	128.46	634
KF-L-SVM	49.84	20.34	14.37	25.85	108.36	634
Lasso-L-RF	79.16	49.66	21.36	25.58	172.39	846
KF-L-RF	80.53	51.04	21.67	27.15	166.27	846

模型	最大回撤(%)	最大跌幅天数	下行风险(%)	Alpha(%)	Beta	夏普比率
Lasso-L-L	−29.09	143	19.39	14.66	1.13	0.81
KF-L-L	−28.71	136	19.92	17.54	1.1	0.89
Lasso-L-LDA	−29.32	143	19.94	13.74	1.13	0.77
KF-L-LDA	−30.52	239	20.10	18.09	1.11	0.91
Lasso-L-SVM	−36.10	143	19.40	1.43	1.02	0.31
KF-L-SVM	−30.79	143	17.38	5.59	0.98	0.50
Lasso-L-RF	−28.76	134	17.22	12.48	0.99	0.78
KF-L-RF	−29.98	94	18.64	12.44	1.03	0.74

五、结　论

基于 2007 年 3 月至 2020 年 12 月沪深 300 指数和中证 500 指数成分股的季度数据，利用 Knockoff-Logistic 变量选择方法与 Logistic 回归、线性判别分析、支持向量机及随机森林分类预测方法构建组合模型，采用滑动窗口法划分训练集与测试集，从原始因子池中筛选出对股票收益影响相对显著的因子，对个股收益率能否跑赢大盘进行预测，并构建投资组合进行历史交易回测。研究得出以下结论：

第一，采用基于 Knockoff-Logistic 模型来筛选对收益率具有较高解释能力的量化因子指标。该方法与 Lasso-Logistic 回归选出的指标具有较高的重合度，但倾向于选择更少的变量，以达到控制假阳性的目的。本文采用了滑动窗口法进行建模，记录每次滑动时选择的因子，统计各因子出现频率，综合得出，对数流通市值对个股收益能力的解释程度最显著。回测结果表明，单从每个季度的收益来看，基于 Knockoff 的投资组合收益更加稳健。纳入对投资风险的考量，基于 Knockoff 方法能够构建出超额收益更高、风险更低的投资策略。

第二，选择 Logistic 回归作为预测模型，以及 3 种机器学习方法作为对比模型，在已选因子的基础上建立预测模型。从回测结果的角度综合评价，4 种模型的选股能力由高到低排序依次是 Logistic 回归、线性判别分析、随机森林、支持向量机。将 Logistic 回归用于股票收益情况预测，在引入 Knockoff 前后均能构建出高质量的投资组合，满足投资者适当规避风险的需求。

第三，着眼于中国股票市场，研究量化因子指标与股票收益情况之间的关系。事实证明，Knockoff 的思想应用于多因子选股有可取之处，能带来更好的投资绩效，更好地为投资者的决策服务，提供智能化投资决策方法。

综上所述，本文将 Knockoff 方法与 Logistic 回归以及支持向量机等机器学习的方法相结合。首先，在众多的候选因子中挑选出对获取股票超额收益具有真正显著影响的因子，Knockoff 方法具有显著的优越性。其次，Knockoff-Logistic 模型能够更好地挖掘金融资产的信息、挑选具有高回报的资产，这对稳定收益，控制市场风险具有重大意义。

本文的创新之处在于将 Knockoff 变量选择方法引入了量化选股模型的应用，并从多个角度收集了候选因子库，扩充了候选因子的数量。但由于因子数据与响应变量之间存在滞后期，对从实际意义上来解释模型选出的因子指标与股票收益情况之间的关系造成了困难。其次，本文构造的投资组合都是等权的，也曾考虑以概率赋权的形式构建投资策略，但由于在选股时以划分为 $Y=1$ 的概率值为依据，排在前 10 位的股票之间概率值相差甚微，与等权构造的投资组合收益不相上下。最后，模型所构建的投资组合 Beta 值均在 1 左右，表明其收益与基准指数同向变动，在市场整体下行时，难以规避系统性风险引发的损失。因此，合理优化投资组合中的股票权重以及控制投资组合对系统性风险的敏感程度将作为下一阶段的研究重点。

参考文献：

[1] SHARPE W F. Capital asset prices：a theory of market equilibrium under conditions of risk[J]. Journal of finance，1964，19(3)：425-442.

[2] ROSS S. The arbitrage theory of capital asset pricing[J]. Journal of economic theory，1976，13(3)：341-360.

[3] FAMA E F，FRENCH K R. Common risk factors in the returns on stocks and bonds[J]. Journal of financial economics，1993(1)：3-56.

[4] FAMA E F，FRENCH K R. A five-factor asset pricing model[J]. Journal of financial economics，2015(1)：1-22.

[5] LIU J，STAMBAUGH R F，YUAN Y. Size and value in China[J]. Journal of financial economics，2019(1)：48-69.

[6] JIANG F，TANG G，ZHOU G. Firm characteristics and Chinese stocks[J]. Journal of management science and engineering，2018，3(4)：259-283.

[7] KELLY B T，PRUITT S，SU Y. Characteristics are covariances：a unified model of risk and return[J]. Journal of financial economics，2019(3)：501-524.

[8] 蒋翠侠，刘玉叶，许启发. 基于 LASSO 分位数回归的对冲基金投资策略研究[J]. 管理科学学报，2016，19(3)：107-126.

[9] 李斌，邵新月，李玥阳. 机器学习驱动的基本面量化投资研究[J]. 中国工业经济，2019(8)：61-79.

[10] 舒时克，李路. 正则稀疏化的多因子量化选股策略[J]. 计算机工程与应用，2021，57(1)：110-117.

[11] GIGLIO S，LIAO Y，XIU D. Thousands of alpha tests[J]. The review of financial studies，2021，34(7)：3456-3496.

[12] BENJAMINI Y，HOCHBERG Y. Controlling the false discovery rate：a practical and powerful approach to multiple testing[J]. Journal of the royal statistical society：series B(satistical methodological)，1995，57(1)：289-300.

[13] BARBER R F，CANDÈS E J. Controlling the false discovery rate via Knockoffs[J]. Annals of statistics，2015，43(5)：2055-2085.

[14] CANDÈS E，FAN Y，JANSON L，et al. Panning for gold：'Model-X' knockoffs for high dimensional controlled variable selection[J]. Journal of the royal statistical society：series B(statistical methodology)，2018，80(3)：551-577.

[15] GÉGOUT-PETIT A，GUEUDIN-MULLER A，KARMANN C. The revisited knockoffs method for variable selection in L1-penalized regressions[J]. Communications in statistics-simulation and computation，2020(10)：1-14.

[16] SRINIVASAN A，XUE L，ZHAN X. Compositional knockoff filter for high-dimensional regression analysis of microbiome data[J]. Biometrics，2021，77(3)：984-995.

[17] ZHU G，ZHAO T. Deep-gknock：nonlinear group-feature selection with deep neural networks[J]. Neural networks，2021(135)：139-147.

[18] LIU W，KE Y，LIU J，et al. Model-free feature screening and FDR control with knockoff features[J]. Journal of the American statistical association，2020，117(537)：1-43.

[19] ZHONG Y，LUO L，WANG X，et al. Multi-factor stock selection model based on machine learning[J]. Engineering letters，2021，29(1)：177-182.

[20] 许林，林晓滢，肖万. 基于修正 MF-ADCCA 模型的全球股票市场非对称交叉相关性研究[J]. 统计与信息论坛，2021，36(10)：41-54.

[21] 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.

[22] KRAUSS C，DO X A，HUCK N. Deep neural networks，gradient-boosted trees，random forests：statistical arbitrage on the S&P 500[J]. European journal of operational research，2016，259(2)：689-702.

[23] DARKHOVSKII B S，BRODSKII B E. A nonparametric method for fastest detection of a change in the mean of a random

sequence[J]. Theory of probability and its applications,2006,32(4):640-648.
[24] AUGER I E, LAWRENCE C E. Algorithms for the optimal identification of segment neighborhoods[J]. Bulletin of mathematical biology,1989,51(1):39-54.
[25] 王小燕,张中艳. 含图结构的 GR-LDA 方法及其信用违约预警应用[J]. 统计研究,2021,38(7):140-152.
[26] HUANG J,ZHANG T. The benefit of group sparsity[J]. Annals of statistics,2010,38(4):1978-2004.
[27] SMILDE A K, KIERS H A L, BIJLSMA S, et al. Matrix correlations for high-dimensional data: the modified RV-coefficient[J]. Bioinformatics,2009,25(3):401-405.
[28] 贺毅岳,李萍,韩进博. 基于 CEEMDAN-LSTM 的股票市场指数预测建模研究[J]. 统计与信息论坛,2020,35(6):34-35.

Multi-factor Quantitative Stock Selection Based on Knockoff-Logistic

WANG Xiao-yan,ZHOU Ying,TANG Ting-ting,ZHANG Zhong-yan
(College of Finance and Statistics,Hunan University,Changsha 410079,China)

Abstract: In the current data-driven era, how to mine useful information of financial assets and select appropriate assets from numerous options available is of great importance for stabilizing returns and controlling risks in investment activities. As a widely used method of selecting stocks, multi-factor quantitative model mainly aims at selecting a set of factors with optimal explanatory power. The existing factor selection methods do not take into account the false discovery rate (FDR) of selection, which may be not conducive to the construction of a robust investment strategy. In this study, we adopt Lasso penalty to realize factor selection and uses Knockoff method to control the FDR of factor selection at a given level to ensure the accuracy. Based on the selected factors, Logistic regression is used to make prediction on returns, and linear discriminant analysis, support vector machine, and random forest are used as comparison models. An empirical study is carried out based on the quarterly frequency factor data of CSI 300 and CSI 500 index stocks from 2007 to 2020. 79 effective factors are selected to construct the factor database, which describes seven aspects of stocks, namely financial liquidity, growth, profitability, valuation, size, consensus expectations, and technology. The rolling regression is used to make investment return and other prediction performance. Finally, quarterly trading strategies are designed, and then applied to back test to compare the model performance.

The empirical results show that, Knockoff-Logistic and Lasso-Logistic model select different numbers of factors on average: the former identifies 14 variables, and the latter identifies 19 variables. Knockoff-Logistic model contains 79% of the variable information provided by Lasso-Logistic. Meanwhile, this method controls the FDR of variable selection and eliminates redundant variables. Among the selected variables, the logarithmic market value has the most significant impact on stock returns in the above two models. The quantitative trading strategy based on Knockoff-Logistic model achieves a Sharpe ratio of 89%, an annualized rate of return of 27. 37%, an excess return of 77. 73%, a maximum drawdown rate of 28. 71%, an alpha value of 17. 54%, and a beta value of 1. 1 during the back test period. Compared with other models, the proposed stock selection model achieves better back test performance and plays a vital role in increasing excess returns and reducing downside risks in general.

From the perspective of variable selection, the model based on factors selected by Knockoff method achieved better market performance than those based on Lasso. From the perspective of model comparison, the portfolio predicted by Logistic regression has the advantage of relatively higher returns and lower risks than linear discriminant analysis, support vector machine, and random forest. On the whole, Knockoff method is beneficial to improve the accuracy of factor selection, enhance the predictability of Chinese stock market, and has significance for optimizing the asset allocation.

Key words: supervised classification; false discovery rate; factor selection; portfolio; variable selection

(责任编辑:李　勤)