**基于随机森林算法和拥挤交易模型的多因子选股策略研究**

**A Study on Multi-factor Stock Selection Strategy Based on Random Forest Algorithm and Crowded Trading Model**

分类号 密级

U D C 编号



硕 士 学 位 论 文

**学位论文题目**： 基于随机森林算法和拥挤交易模型

的多因子选股策略研究

**姓 名**  S2402108

**学 号**

**学 院**

**学位类别：□学术硕士 ☑专业硕士 □同等学力**

**学科专业**

**指导教师**

**第二导师**

**提交论文日期：2024年 2月 28日**

**独 创 性 声 明**

本人郑重声明：所呈交的学位论文是本人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得中央财经大学或其他教育机构的学位或证书所使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

**学位论文作者签名**：

年 月 日

**学位论文版权使用授权书**

本学位论文作者完全了解中央财经大学有关保留、使用学位论文的规定。特授权中央财经大学可以将学位论文的全部或部分内容编入有关数据库进行检索，并采用影印、缩印或扫描等复制手段保存、汇编以供查阅和借阅。同意学校按规定向国家有关部门或机构送交论文和磁盘。

（保密的学位论文在解密后适用本授权说明）

**学位论文作者签名： 导师签名：**

年 月 日 年 月 日

**摘要**

我国股票市场成立时间较短，与发达国家成熟的金融市场相比仍有较大差距，目前处于相对较弱的弱有效市场阶段。在A股市场中，行业轮动效应非常显著，在多方力量的共同作用下，短线资金在一个又一个热点板块之间来回切换，带来短期内的大起大落。尤其是在近几年，伴随着基金规模的快速增长，基金经理间的投资决策高度一致，出现大量特征类似的资金共同配置于相似行业的现象，即产生了行业板块的拥挤交易。对拥挤交易现象的研究始于2008年金融危机，近年来才走入国内研究者的视野，从拥挤交易的角度对A股行业轮动现象的研究是非常有必要的

基于此，本文梳理了大量文献，参考Kinlaw(2019)中提出的“资金集中度”因子，对国内行业板块的拥挤交易现象进行研究。在分析了国内外行业分类的差异性后，本文选择使用28个申万一级行业的划分作为分类依据，构建了这些一级行业的集中度因子，用以衡量行业板块的热度。由于资金集中度因子的升高无法区分目前是处于拥挤交易的泡沫生成阶段还是泡沫破裂阶段，所以加入“相对估值”指标用以规避泡沫破裂。在对集中度因子进行有效性检验后，本文结合相对估值指标，构建了一个基于拥挤交易泡沫阶段的行业轮动策略，在2019年9月至2024年1月的区间内进行回测，验证了策略的有效性。

在上述策略的基础上，本文使用了随机森林算法，尝试在优势行业中选取优质个股，实现更大的α收益。使用TA-Lib量化金融研究库根据个股的交易数据构建了近50个技术因子，而后使用相关系数矩阵和LightGBM模型从中筛选出30个重要性较高的因子用以训练随机森林模型。在使用GridSearchCV选出较优随机森林参数后，同样在上述区间内以3个月为学习窗口期对随机森林模型进行滚动训练，以保证模型的时效性。随机森林模型选出的个股进行等权重投资，回测结果在行业轮动策略的基础上又取得了非常理想的收益，且在2023年的熊市期下相对抗跌，证明该策略是一个能实现超额收益的稳健策略。

本文的研究成果在一定程度上验证了A股市场的拥挤交易现象，可以作为为投资者选取投资行业的策略参考，同时也可以作为监管部门的风险识别指标。最后，本文对当下A股的投资前景进行了展望。

**关键词：**拥挤交易、随机森林算法、行业轮动、多因子选股

**Abstract**

China's stock market has a relatively short history and still lags behind mature financial markets in developed countries, remaining in a relatively weak stage of market efficiency. In the A-share market, the industry rotation effect is very significant, with short-term funds switching back and forth between hot sectors under the combined influence of multiple forces, resulting in large fluctuations in the short term. Especially in recent years, with the rapid growth of fund size, there has been a phenomenon where a large amount of funds with similar characteristics are collectively allocated to similar industries by fund managers, leading to crowded trades in industry sectors. Research on crowded trades began with the 2008 financial crisis and has only recently entered the field of view of domestic researchers. Therefore, studying the industry rotation phenomenon from the perspective of crowded trades in the A-share market is very necessary.

Based on this, this paper reviews a large number of literature, and referring to the "Fund Concentration" factor proposed by Kinlaw (2019), it studies the phenomenon of crowded trades in domestic industry sectors. After analyzing the differences in domestic and foreign industry classifications, this paper selects the division of 28 SW One-level industries as the classification basis, constructs concentration factors for these one-level industries to measure the heat of industry sectors. Since the increase in concentration factors cannot distinguish whether the current stage is in the bubble generation or bubble burst stage of crowded trades, a "Relative Valuation" indicator is added to avoid bubble bursts. After testing the effectiveness of the concentration factors, this paper combines the relative valuation indicators to construct an industry rotation strategy based on the stage of crowded trades bubbles. A backtest is conducted from September 2019 to January 2024 to verify the effectiveness of the strategy.

Based on the above strategy, this paper uses the Random Forest algorithm to select high-quality individual stocks in dominant industries to achieve greater α returns. Using the TA-Lib quantitative finance research library, nearly 50 technical factors are constructed based on the trading data of individual stocks. Then, using correlation matrices and the LightGBM model, 30 factors with higher importance are selected to train the Random Forest model. After selecting the optimal Random Forest parameters using GridSearchCV, the Random Forest model is rolled trained with a learning window period of 3 months within the above interval to ensure the timeliness of the model. The stocks selected by the Random Forest model are equally weighted for investment. The backtest results achieve very ideal returns on top of the industry rotation strategy and relatively resist downturns during the bear market in 2023, proving that this strategy is a robust strategy that can achieve excess returns.

The research results of this paper verify the phenomenon of crowded trades in the A-share market to a certain extent, which can serve as a strategic reference for investors in selecting investment industries, and can also serve as a risk identification indicator for regulatory authorities. Finally, this paper provides an outlook on the current investment prospects of A-shares.

**Keywords**: crowded trades, Random Forest algorithm, industry rotation, multi-factor stock selection

[第1章 绪论 1](#_Toc160011026)

[1.1 选题背景及研究意义 1](#_Toc160011027)

[1.1.1 研究背景 1](#_Toc160011028)

[1.1.2 研究意义 2](#_Toc160011029)

[1.2 国内外研究现状 3](#_Toc160011030)

[1.2.1 多因子模型 3](#_Toc160011031)

[1.2.2 随机森林算法 4](#_Toc160011032)

[1.2.3 行业轮动效应与拥挤交易 5](#_Toc160011033)

[1.2.4 国内外研究评述 7](#_Toc160011034)

[1.3 研究内容和论文结构 8](#_Toc160011035)

[1.3.1 研究内容 8](#_Toc160011036)

[1.3.2 论文结构 8](#_Toc160011037)

[第2章 理论概述 10](#_Toc160011038)

[2.1 多因子选股模型 10](#_Toc160011039)

[2.2 决策树算法和随机森林 11](#_Toc160011040)

[2.2.1 决策树算法 11](#_Toc160011041)

[2.2.2 随机森林算法 12](#_Toc160011042)

[2.3 拥挤交易 14](#_Toc160011043)

[2.3.1 PCA主成分分解算法 14](#_Toc160011044)

[2.3.2 资金集中度 17](#_Toc160011045)

[第3章 基于拥挤交易的行业轮动策略 18](#_Toc160011046)

[3.1 样本选取与数据处理 18](#_Toc160011047)

[3.2 “资金集中度”因子的构建 19](#_Toc160011048)

[3.3 相对估值指标的构建 20](#_Toc160011049)

[3.4 因子有效性的检验 22](#_Toc160011050)

[3.4.1 排序打分法检验 22](#_Toc160011051)

[3.4.2 分层测试法检验 24](#_Toc160011052)

[3.5 基于集中度因子和相对估值指标的行业轮动策略 25](#_Toc160011053)

[3.5.1 评价标准 25](#_Toc160011054)

[3.5.2 回测结果 25](#_Toc160011055)

[第4章 基于随机森林的多因子模型构建 29](#_Toc160011056)

[4.1 数据源及因子提取 29](#_Toc160011057)

[4.2 因子筛选 31](#_Toc160011058)

[4.2.1 相关系数筛选 31](#_Toc160011059)

[4.2.2 LightGBM特征因子评分 33](#_Toc160011060)

[4.3 训练随机森林模型 34](#_Toc160011061)

[第5章 行业轮动策略和随机森林选股多因子模型的整体回测 36](#_Toc160011062)

[5.1 策略详细描述 36](#_Toc160011063)

[5.2 模型回测与评价 36](#_Toc160011064)

[第6章 研究结论与展望 41](#_Toc160011065)

[6.1 研究总结 41](#_Toc160011066)

[6.2 未来展望 41](#_Toc160011067)

[参考文献 43](#_Toc160011068)

# 绪论

## 选题背景及研究意义

### 研究背景

过去三十多年来，我国的股票市场在国民经济中发挥着越来越重要的地位。不仅为上市公司提供了越来越广泛的融资渠道，也为广大投资者提供了越来越多样化的投资渠道。把股票作为资产配置一部分的投资者比例不断增加。在过去三十多年的发展中，越来越多的机构和个人投资者借此实现了财富增值。自从股票市场出现之来，如何在股票交易中获益，成为了广大股民不懈追求的目标。

我国股票市场成立时间较短，早期受行政力量干预较多，与发达国家成熟的金融市场相比仍有较大差距：监管和处罚力度不足导致财务造假和内幕交易屡见不鲜、退市机制不完善导致股市中垃圾股票不断沉淀、投资者长期亏损导致投机氛围较浓……。中国股市目前处于，且未来大概率长期处于相对较弱的弱有效市场阶段。在这个市场下尽管往往会受到公开信息的影响，但仍然存在一些情况下，价格可能没有充分反映所有可用信息，从而给技术分析和基本面分析提供了空间。

早期的投资流派分为技术分析和基本面分析两大主流：技术分析以市场上的指标为研究对象，试图寻找出指标波动的规律；基本面分析是一种自上而下的系统分析法，侧重公司的实际经营情况，以公司实际表现为根据进行投资。二十世纪五十年代，伴随着马科维茨投资组合理论的提出，数理统计的方法真正被引入到了投资分析中，为量化投资的发展开拓了理论基础。在马科维茨投资组合理论的基础上，资本资产定价模型（CAPM）和套利定价模型（APT）的先后提出降低了马科维茨理论的使用难度，并使量化投资的研究由单因子模型扩展到多因子模型。

量化分析兼具技术分析和基本面分析的优势，在弱有效市场下，通过挖掘潜在因子使得投资者追求超额收益成为了可能。目前主流的TA-Lib库中已经有超过150种技术因子。依靠人类的信息处理能力难以对如此多因子进行归纳整理，更何况股票市场具有高维度、非线性的特点，因而传统的线性多因子模型仍不能很好地预测股票收益率。近年来，伴随着机器学习技术的进步和硬件性能的提升，机器学习算法对非线性、非参数化数据的有效的特征提取能力和较高的拟合准确率为股价预测开辟了一条新的道路，也逐渐成为量化分析师手中一件有力的工具。因此，将多因子模型与机器学习算法相结合进行股价预测是非常有必要的。

拥挤交易是指市场上出现大量投资者对某一特定投资机会或资产进行追涨或抛售的情况。这种情况下，投资者的共同行为导致市场上出现了高度集中的交易活动，使得某些投资品种或市场部分过度拥挤。拥挤交易往往会导致资产大幅偏离其内在价值，一旦市场情绪发生逆转，可能会引发大规模的资金流出，导致价格的急剧下跌，从而加剧了市场的不稳定性。拥挤交易现象在A股普遍存在。

我国股市受到市场情绪波动的影响较大。由于散户投资者的比例较高，外加投机资金的介入，他们更容易受到市场消息、传言以及媒体报道的影响，从而导致市场出现剧烈的波动，行业轮动效应显著。这些不理性因素为投资提供了一个很好的切入点：即选择市场热度大的板块进行交易，获得短期的超额收益。Kinlaw(2019)提出了将“资金集中度”与相对估值结合以选择出现拥挤交易行业，参考这篇文章的做法，本文试图构建适用于中国股市的拥挤交易测量指标，以选取资金集中的热点板块。

### 研究意义

基于以上背景，本文拟参照Kinlaw(2019)构建的“资金集中度”指标，用以筛选出A股中正在发生拥挤交易的行业板块，并与随机森林算法和多因子模型结合，筛选出行业板块内部的热门股票。同时，参考“相对估值”指标，用以指导止盈止损条件，以实现在各种市场状况下都能获得稳健收益的量化投资策略，其理论和现实意义如下：

#### 理论意义

多因子模型通过拟合出各种因子与收益率的关系，以预测股票的涨跌情况，该问题是机器学习中最经典的分类问题。过去一般采用线性模型进行拟合，本文将在对多种因子进行筛选排序后，选取最有影响力的因子，基于随机森林构建回归选股模型。随机森林算法有很好的高维数据处理能力，不易过拟合且可解释性强，非常适合此类问题。在择时策略方面，考虑到A股在震荡市时行业轮动效应较明显和散户投资者较多的特点，使用随机森林对行业指数进行回归学习预测行业轮动情况，综合考虑投资者情绪指标决定择时结果。

#### 现实意义

对选股策略的研究最终目的是为了帮助投资者获得稳健的超额收益。选股和择时是量化策略的两个重要部分。本文参考Kinlaw(2019)构建的“资金集中度”指标可以为投资者提供衡量A股拥挤交易程度的指标，由于基金披露持仓情况往往延后一个季度才公布，因此该指标可以在一定程度实时反应机构持仓集中程度。随机森林算法在拟合数据时可以给出因子重要程度的排序，可以据此筛选出相对重要的因子，为后续策略的构建提供参考。

本文所用到的数据和代码均已上传至GitHub，原始数据来自同花顺iFind客户端，数据处理的代码也一并上传以便交叉检验。本人保证所用数据的真实性。可访问<https://github.com/LongLiveMaxwell/Sector-rotation-and-Random-Forest>以获得更详细的信息。

## 国内外研究现状

### 多因子模型

1952年，Markowitz使用数理统计方法，建立了均值-方差模型，以组合收益率的方差来表示风险，开创了量化分析的先河。但是，均值方差模型计算复杂度较高，内部干扰严重，与实际投资仍有很大距离，难以在大规模的股票市场上广泛应用。Treynor(1962)和Sharpe(1964)等人为了降低Markowitz模型的使用难度，在Markowitz投资组合模型的基础上提出了资本资产定价模型（CAPM），CAPM模型大大降低了Markowitz投资组合理论的使用难度，为今后的多因子定价模型奠定了基础。与此同时，Fama和Samuelson(1965)提出了有效市场假说，其核心假设包括：投资者高度理性、市场是无摩擦完全竞争的有效市场、短期股价是随机游走不可预测的。虽然Fama对有效市场的设想和复杂的市场环境相距甚远，但作为一个描述市场的基础性理论，有效市场假说在市场研究中的定位和牛顿运动定律在经典力学中的定位类似，为日后的研究提供了基本的理论模型。

20世纪70年代，Black和Scholes提出了期权定价模型，其中体现的套利定价的思想成为了套利定价理论(APT)(Ross(1976))的基础。Ross认为：投资者的套利行为促使市场均衡价格的形成，根据无套利原则可以得出：风险资产的均衡收益和多个因素存在线性关系。在APT模型提出以后，学者们开始转向研究多因子模型。Fama (1993)发现β系数无法完全解释不同股票间的收益率差异，而另一些指标能够解释这些差异，这表明公司与公司之间经营状况的不同对股票收益率有影响，在此基础上，他们又提出了包含市场资产组合因子、市值规模因子（SMB）和账面市值比因子（HML）的三因子模型。和CAPM模型相比，三因子模型能更好地解释超额收益率的来源。但随着市场的变化，学者在研究的过程中发现了越来越多仅凭三因子难以解释的情况，这表明还有很多因素在三因子模型的范围外，于是学者们试图找到更多因子，丰富多因子模型。Carhart(1997)加入了动量因子，认为股票收益率存在趋势效应，在某些情况下受公司基本面的影响不太大，这也为后来的动量策略埋下了伏笔。除此之外，他还引入了公司财务基本面因子、年收益率因子、规模因子(SML)等。这些因子的提出很好地完善了多因子模型因子库的丰富程度。Joseph(2000)提出了使用股票打分制的方法：小于阈值不得分，大于则得一分，根据多个因子的最终得分选择分数高的形成投资组合，实证证明打分制也具有非常高的收益。Fama 和 French(2015)又对他们提出的三因子模型做出更新，新增了公司的投资和盈利水平两个因子，发表了五因子模型。目前主流的TA-Lib库中已经有超过150种技术因子，随着研究的不断深入，还会挖掘出更多因子。

中国的股票市场成立仅有不到四十年，我国学者在最开始对多因子模型的研究过程中往往采用照搬国外成熟因子做实证研究阶段。但是我国股市尚未成熟，有着浓厚的地域特色，照搬国外有效的多因子策略并不一定能获得较优的回报。黄兴旺(2002)将三因子模型应用于国内市场，发现规模因子对收益的影响非常显著，非常符合A股前十年中资金集中于大型国企的特点。王涛(2012)等利用2004至2011年A股的月度数据，从更精细的角度研究三因子模型在中国市场的普适性，结果表明账面市值因子对中账面市值比组合不显著，对大、小账面市值比组合显著；同时，文章还对A股在牛市、熊市和震荡市的情况进行分析，得到规模因子会在震荡市中失效的结论。吴荻(2011)等人通过打分的方式构建多因子选股模型，表明打分发设计的多因子模型能够得到较好的效果。李倩倩(2019)利用打分法构建八因子选股模型，回测结果表明新加入的因子能够很好地增加超额收益，同时具有更强的控制回撤的能力。

### 随机森林算法

随机森林算法起源于决策树算法，在随机森林算法被提出之前研究人员通过决策树模型研究选股问题。Sorensen(2000)选取了包括每股收益和股票价格动能在内的六项指标，构建决策树从美国科技板块中筛选股票，回测结果表明筛选出的股票组合可以获得较好收益。焦健等(2010)在利用Sorensen的决策树模型在我国科技创新板块内筛选股票，然而并未取得良好结果，焦健等人认为这是由于两国股票市场间的较大差异造成的，因此在原模型的基础上进行改进，选取市现率和资产增长率等六个指标，回测结果表明新的选股指标可以在A股市场上取得较好收益。

随机森林算法于2001年被Breiman正式提出。随机森林算法将多个决策树进行整合，每一颗决策树单独决策得到各自的预测结果，最后在全部的决策树中使用投票的方法得到最终预测结果，这样就由很多弱分类器通过集成获得了一个强分类器。和单颗决策树相比，随机森林的优点在于对缺失值不敏感，且不容易过拟合。Genuer(2010)利用随机森林算法将重要特征筛选出来以解决二分类问题，并取得较好的结果，其中用到的算法也是随机森林分类模型进行分类的基础。Ladyzynski(2013)等人利用随机森林算法预测股价走势。Manojlovi(2015)采用随机森林算法对股票进行分类，分类结果表明随机森林算法在股票分类问题上具备很好地泛化性能。Abraham Rebecca (2022)利用随机森林根据当日收盘价预测次日开盘价的变化情况，分别构建了基于遗传算法和随机森林的分类器，并证明随机森林在股指预测方面的有效性。BehrMerle (2022)从投资者情绪的角度出发，对影响投资者情绪的因子按重要性排序，并使用主成分分析法衡量证券收益率与情绪变动的关系，优化后的预测准确度和耗时有了很大提升。

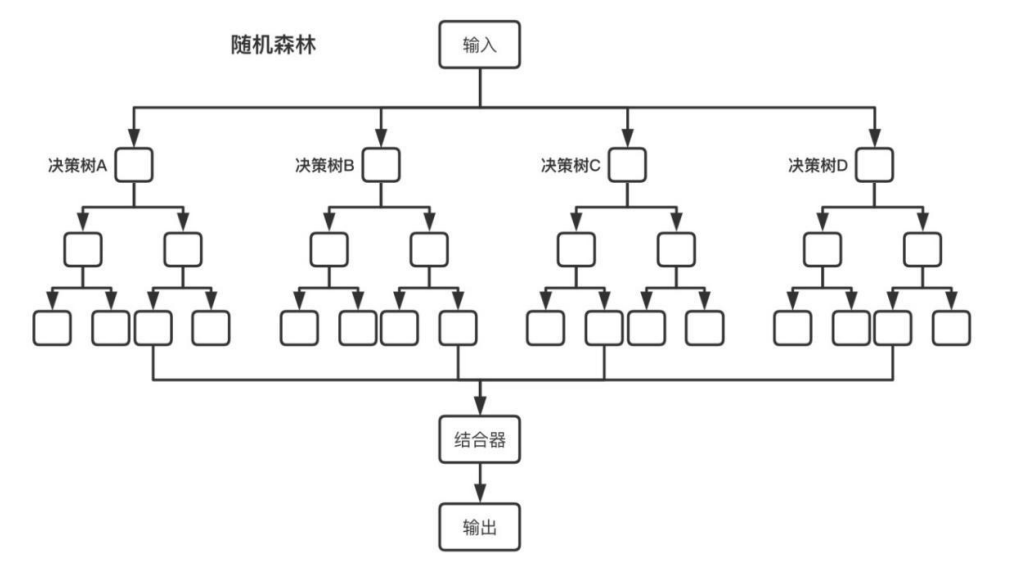


图1. 随机森林算法模型，本质上是多颗决策树的组合

### 行业轮动效应与拥挤交易

行业轮动现象是指证券市场行情演变的过程中，不同行业的景气度和市场表现在一段时间内出现交替变化的现象。不同行业的发展受到多个因素的影响，包括经济环境、技术进步、政策变化等。由于这些因素的不断变化，某个行业的表现可能会逐渐衰退，而其他行业则可能逐渐崛起。轮动现象通常是周期性的，投资者可以通过研究市场趋势、行业前景和其他相关因素来预测和应对轮动现象，以获得更好的投资收益。

国外对行业轮动的研究有很多，不过和股票收益分析相比，对行业轮动的研究的出现要晚得多。Bradley(2002)对美国股市上不同行业的股价展开研究，发现美国股市中存在着显著的行业轮动效应。George和Hwang (2004)的研究表明价值型股票和成长型股票之间存在风格轮动现象，投资者在价值型风格组合与成长型风格组合之间轮换投资来获取更高的收益。2004年，美林证券提出了产业时钟理论。将经济周期分为过热、滞涨、衰退和复苏四个阶段，不同行业在经济周期的同一阶段会有不同的表现，建议在不同的经济阶段进行动态行业配置。

大量实证研究表明，A股存在非常明显的行业轮动效应，并且比欧美市场轮动速度更快。

宗淑杰(2017)从货币周期和经济周期两方面解释了行业轮动的原因，利用多因子模型将行业分为周期性行业和非周期性行业，并搭建识别行业轮动因子的模型，构建了基于行业轮动的投资组合。武文超(2014)通过动量反转的思想，借助沪深300的行业指数对A股行业轮动进行了实证检验，结果表明在在日和月的区间上行业动量特别明显，而在周的时间区间内则表现为行业轮动现象。黄前前(2016)基于熊市和牛市下A股历史数据，经实证分析得出：中国股市存在明显的短期动量效应。欧美市场的短期和中期的动量盈利持有期长达1年，而我国股市的动量效应只有3到4周，证明我国市场行业轮动周期较短。

对于出现行业轮动效应的原因，学者们有着诸多解释，有人从宏观经济周期的理论展开分析，也有人从投资者行为的角度进行解释，总的来说无外乎实体经济和行为金融两个角度。但是国内外轮动周期相差如此巨大的原因，更多的还是归结于市场机制和投资者群体。拥挤交易理论在我国特殊的国情下可以解释行业轮动快的原因。

最开始研究拥挤交易的是Stein(2009)，他把投资者分为观望者和套利者两种。由于套利者无法判断有多少人和他采取同样的交易策略，从而会发生大量同质化资金购买同一类资产的现象，Stein将这种情况定义为拥挤交易，并通过构建模型得出套利者过多会引发拥挤交易，降低市场有效性。Menkveld(2016)认为机构投资者集中大量资金于同一交易会造成拥挤交易，并认为机构的拥挤交易是诱发金融危机的重要原因。Kinlaw(2019)认为拥挤交易会使投资者心理发生巨大变化，导致投资者踩踏引发资产价格的大幅波动。同时作者也认为，拥挤交易往往会导致泡沫和泡沫的急速破裂。Kinlaw(2019)提出了一种泡沫识别方法，Kinlaw基于美国历史上的市场泡沫，使用“资金集中度”作为拥挤交易的代理指标对出现泡沫的行业加以识别，并结合行业的相对估值指标区分泡沫的形成期和破裂期，取得了比较好的识别效果。

国内对拥挤交易的研究尚处于初级阶段，目前有关拥挤交易的研究较少。周丽云(2017)构建了投资者情绪指标和日度个股投资者拥挤交易行为指标，发现拥挤交易行为和投资者情绪指标都对股价有系统性的重要影响，且对小盘股影响更显著。同时，不同情绪股票组合均受到各组合投资者拥挤交易行为的显著影响，股票超额收益随着投资者情绪的上涨而增加。王婷婷(2021)构建行业轮动策略组合，使用了Kinlaw(2019)中提出的“资金集中度”作为拥挤交易的代理指标，与“相对估值”因子结合可以在泡沫破裂前发出卖出信号，实现及时逃顶，在A股市场上取得了不错的收益和风险控制能力。

本文将采取王婷婷(2021)中提出的行业周期轮动信号作为选取轮动行业的标准，结合随机森林算法在该行业内选择上涨概率较大的股票构建投资组合，以获取更高的超额收益。

### 国内外研究评述

1. 国内对拥挤交易研究较少

由于拥挤交易是一个新出现的概念，在国内的研究中尚未完全挖掘。国内学界和业界对行业轮动的研究比较深入，不过，国内机构投资往往从经济周期的角度研究行业轮动策略：即识别当前在经济周期中的阶段（如美林时钟理论中的过热、滞涨、衰退和复苏四个阶段），由于每个阶段内都有对应的优势行业，可以配置这些行业以获得超额收益。者常采用的行业轮动策略主要有两种：一种是从经济周期角度出发，认为在每个周期都有相应的优势行业与之对应，可以通过配置相应周期下的优势行业来谋取超额收益。然而，过去三十年间中国经济快速发展，且受政策因素影响大，在多个经济周期的各个阶段中所涌现出的优势行业也有变化，且关系不稳定。比如2008年金融危机后的滞涨衰退期，按理来说家用电器行业会随着房地产的阶段下行而进入阶段的衰退期，然而伴随着家电下乡计划，过剩库存得到了去除，家电行业也借此迎来了新生，而这一次偶然在后续的经济周期中并未复现。因此，这一方式并不能稳定的识别出优势行业，仍需仰赖投资者的专业经验，对个人投资者的参考价值不大。

本文将从拥挤交易的角度出发，研究股市板块轮动现象，并基于拥挤交易的不同阶段，制定适时的买入卖出策略，以及时逃顶，避免泡沫破裂带来的损失。

1. 使用随机森林选股时，股票池中的股票往往来自于不同行业的股票

大部分使用随机森林模型进行股价预测的论文中，所选取的股票来自各种行业，这一方面提高模型的多样性和鲁棒性，使得模型能够更好地适应不同行业板块的变化，但另一方面由于不同行业板块之间可能存在市场表现、盈利能力、成长性等方面的差异，会增加模型的复杂程度，并使模型更难以捕捉到股价变化的共性规律，从而降低了预测准确率。如果针对不同的行业分别构建随机森林模型，使每个行业都有对应的模型，可以在一定程度上提高在该行业内部进行预测的准确率和稳定性，这也是本文重点考虑的内容。

## 研究内容和论文结构

### 研究内容

第一步，在同花顺数据中心获得28个申万一级行业从2014年1月到2024年1月的行业指数、行业市值、市净率数据，以及这些一级行业下接近2000支股票在2019年1月到2024年1月的日度开盘价、收盘价、最高价、最低价、成交量、涨跌幅等数据；

第二步，根据28个申万一级行业的行业指数和行业市值构建“资金集中度”因子，用以衡量行业板块拥挤交易的程度。根据市净率数据构建“相对估值”指标，结合“资金集中度”因子用以衡量拥挤交易的不同阶段；

第三步，对第二步中构建的指标进行有效性检验，包括IC分析、分层测试法；

第四步，使用随机森林分别对每个行业构建多因子选股模型，并测试不同因子集、训练窗口下的模型表现，尝试获取表现良好的随机森林多因子选股模型；

第五步，将行业轮动与随机森林结合起来，首先选择处于泡沫形成阶段（资金集中度较高且相对估值处在低位）的行业，然后使用随机森林在这些行业中选择上涨概率最大的股票。对以上策略进行回测。

本文拟解决的问题如下：

其一，针对中国股市牛短熊长，长期震荡，板块轮动效应明显的特点，结合拥挤交易的理论，设计出适用于A股市场的板块轮动预测指标；

其二，探讨随机森林算法在高度集中的行业内的预测准确性，使用随机森林算法在活跃的板块中找出所谓“股性好”的股票，以期在市场震荡的状态下获得较为稳健的收益。

### 论文结构

本研究共分为5章，基本框架如下：

第1章为绪论，主要介绍本文的研究背景及研究意义，结合国内外论文对研究内容进行评述。

第2章主要介绍本篇文章的理论背景，包括多因子模型，由浅入深地介绍随机森林算法的背景、应用和评价指标，以及对Kinlaw(2019)中提到的集中度因子及其中所用到的主成分分析法进行了详细的介绍。

第3章为基于拥挤交易的行业轮动策略介绍。首先根据Kinlaw(2019)的论文构建集中度和相对估值因子，然后通过IC因子分析和分层打分法验证因子的有效性，最后结合行业指数在一段时期内进行回测展示行业轮动投资策略的可行性

第4章在第3章的基础上，构建随机森林选股模型。首先明确了评判标准，经过筛选构建了最终因子集，进行随机森林模型的调参，选择最合适的参数。最后将行业轮动策略和随机森林算法结合起来进行回测，验证了该策略是一个有效的管理风险并获得收益的投资模型。

第5章为总结与展望，分析了策略中仍有提升空间的部分，对未来A股行情做了一些展望。

# 理论概述

## 多因子选股模型

多因子选股模型是量化投资早期的一种简单但广泛应用的模型，其理论研究相对较为全面。该模型主要用于股票选取，其核心在于构建原始因子池。随着越来越多的学者和量化研究员的挖掘，可供研究的原始因子越来越多，但股票交易和公开的财务信息所反映出的信息终归有限，进行因子挖掘的边际效益也在逐步递减。为了更好的提高模型的准确性，使用较为全面的因子是必须的，但是使用太多的因子反而会使真正有效的信息掩盖在庞杂的因子库中，因此需要按照各个因子与股票收益率的相关性作为评估因子有效性的标准，进行降序排列以筛选出较为有效的因子，用以建立有效的因子库。

股票的收益受到众多因素的综合影响，这些因素包括公司的财务状况、行业的发展趋势、市场情绪、宏观经济环境等等。而这些因素在不同国家的证券市场上表现出的有效性也有所不同。在市场条件较弱的情况下，即使是构建多因子选股模型，也需要假设这些因子对股票收益的影响是持续性的。在此基础上，通过深入研究和分析，寻找在特定市场条件下表现较为稳定和有效的因子。这些因子可能包括但不限于公司的盈利能力、成长潜力、财务健康状况等。在确定了这些有效因子后，应重点关注那些相对被低估的股票，同时，也应综合考虑各方面的有效信息，包括技术分析、基本面分析、市场情绪等，以构建一个全面的选股策略。这样的策略不仅需要在历史数据上进行验证，更需要在真实市场环境中产生持续的收益。

当前的量化模型主要集中在两个方面进行改进：首先是因子的选择需要具有更好的有效性，以提高模型对数据的拟合程度，从而增强预测的准确性。其次是在确保因子有效性的同时，构建择股模型，以尽可能提高选择股票的准确度。现代模型将动量、价值、技术、成长等多种指标结合起来，以涵盖多个维度的因素，并使得量化模型具有更加合理的理论解释。这样的模型不仅能够更全面地捕捉市场特征，还能够更准确地指导投资决策，为投资者提供更可靠的投资策略。

## 决策树算法和随机森林

### 决策树算法

决策树算法是一种常用的机器学习算法，其原理基于对数据集进行递归地划分，以生成一个树形结构的决策规则。其主要原理如下：算法首先从数据集中选择一个特征，作为当前节点的划分标准。通常采用的特征选择方法包括信息增益、信息增益率、基尼系数等；根据选定的特征，将数据集划分为多个子集。每个子集对应于特征的一个取值或取值范围。这样的划分过程将在树的每个节点上进行，直到满足停止条件；对于每个子集，重复上述过程，选择新的特征，并将子集进一步划分。这是一个递归的过程，直到满足某种停止条件，如达到最大深度、节点中的样本数小于某个阈值等；当达到停止条件时，生成叶节点，并确定该叶节点的类别或值。对于分类问题，叶节点通常是一个类别；而对于回归问题，叶节点通常是一个数值；在建立完整的决策树后，可以对其进行剪枝操作，以减少过拟合的风险。剪枝的目标是删除一些冗余的节点，使得决策树更加简洁且具有更好的泛化能力。

决策树的关键在于如何选择最优的属性进行分裂。常见的属性选择方法包括信息增益、信息增益比、基尼指数等，这些方法都旨在使得每次分裂后的子集更加纯净。

按照选择最优属性进行分裂的方式，决策树可大致分为以下3种：

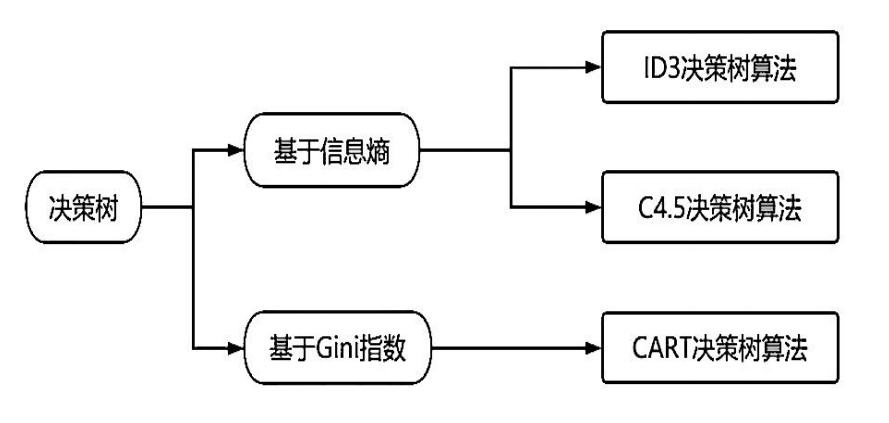


图2. 决策树类别

1. ID3决策树算法

ID3决策树算法的核心是信息熵：

标题: E n t r o y 左小括号 D 右小括号 空格 等于 空格 加总 从 1 到 c 对 空格 减 空格 p 下标 i log 下标 2 p 下标 i - 说明 {"mathml":"<math style=\"font-family:stix;font-size:16px;\" xmlns=\"http://www.w3.org/1998/Math/MathML\"><mstyle mathsize=\"16px\"><mi>E</mi><mi>n</mi><mi>t</mi><mi>r</mi><mi>o</mi><mi>y</mi><mfenced><mi>D</mi></mfenced><mo>&#xA0;</mo><mo>=</mo><mo>&#xA0;</mo><mstyle displaystyle=\"false\"><munderover><mo>&#x2211;</mo><mn>1</mn><mi>c</mi></munderover></mstyle><mo>&#xA0;</mo><mo>-</mo><mo>&#xA0;</mo><msub><mi>p</mi><mi>i</mi></msub><msub><mi>log</mi><mn>2</mn></msub><msub><mi>p</mi><mi>i</mi></msub></mstyle></math>","origin":"MathType for Microsoft Add-in"} (2-1)

其中，D为未划分的数据集，第个子类别出现的概率为，子类别的个数为c。划分数据集后，ID3决策树算法可以提取出信息熵增益最大的特征，进而构建各级子节点：

标题: G a i n 左小括号 t 右小括号 空格 等于 空格 E n t r o p 左小括号 D 右小括号 空格 减 空格 加总 从 j 等于 1 到 m 对 分数 左 绝对值 D 右 绝对值 分之 左 绝对值 D 下标 j 右 绝对值 结束分数 空格 E n t r o p 左小括号 D 下标 j 右小括号 - 说明 {"mathml":"<math style=\"font-family:stix;font-size:16px;\" xmlns=\"http://www.w3.org/1998/Math/MathML\"><mstyle mathsize=\"16px\"><mi>G</mi><mi>a</mi><mi>i</mi><mi>n</mi><mfenced><mi>t</mi></mfenced><mo>&#xA0;</mo><mo>=</mo><mo>&#xA0;</mo><mi>E</mi><mi>n</mi><mi>t</mi><mi>r</mi><mi>o</mi><mi>p</mi><mfenced><mi>D</mi></mfenced><mo>&#xA0;</mo><mo>-</mo><mo>&#xA0;</mo><mstyle displaystyle=\"false\"><munderover><mo>&#x2211;</mo><mrow><mi>j</mi><mo>=</mo><mn>1</mn></mrow><mi>m</mi></munderover></mstyle><mfrac><mfenced open=\"|\" close=\"|\"><msub><mi>D</mi><mi>j</mi></msub></mfenced><mfenced open=\"|\" close=\"|\"><mi>D</mi></mfenced></mfrac><mo>&#xA0;</mo><mi>E</mi><mi>n</mi><mi>t</mi><mi>r</mi><mi>o</mi><mi>p</mi><mfenced><msub><mi>D</mi><mi>j</mi></msub></mfenced></mstyle></math>","origin":"MathType for Microsoft Add-in"} (2-2)

其中，标题: 左 绝对值 D 右 绝对值 - 说明 {"mathml":"<math xmlns=\"http://www.w3.org/1998/Math/MathML\" style=\"font-family:stix;font-size:16px;\"><mfenced open=\"|\" close=\"|\"><mi>D</mi></mfenced></math>","origin":"MathType for Microsoft Add-in"}为数据集样本总数量，标题: 左 绝对值 D 下标 j 右 绝对值 - 说明 {"mathml":"<math style=\"font-family:stix;font-size:16px;\" xmlns=\"http://www.w3.org/1998/Math/MathML\"><mstyle mathsize=\"16px\"><mfenced open=\"|\" close=\"|\"><msub><mi>D</mi><mi>j</mi></msub></mfenced></mstyle></math>","origin":"MathType for Microsoft Add-in"}为第j类数据集样本的数量，总样本熵是Entrop(D)。划分后，所有子样本数据集的加权熵综合是：

标题: 空格 加总 从 j 等于 1 到 m 对 分数 左 绝对值 D 右 绝对值 分之 左 绝对值 D 下标 j 右 绝对值 结束分数 空格 E n t r o p 左小括号 D 下标 j 右小括号 - 说明 {"mathml":"<math style=\"font-family:stix;font-size:16px;\" xmlns=\"http://www.w3.org/1998/Math/MathML\"><mstyle mathsize=\"16px\"><mo>&#xA0;</mo><mstyle displaystyle=\"false\"><munderover><mo>&#x2211;</mo><mrow><mi>j</mi><mo>=</mo><mn>1</mn></mrow><mi>m</mi></munderover></mstyle><mfrac><mfenced open=\"|\" close=\"|\"><msub><mi>D</mi><mi>j</mi></msub></mfenced><mfenced open=\"|\" close=\"|\"><mi>D</mi></mfenced></mfrac><mo>&#xA0;</mo><mi>E</mi><mi>n</mi><mi>t</mi><mi>r</mi><mi>o</mi><mi>p</mi><mfenced><msub><mi>D</mi><mi>j</mi></msub></mfenced></mstyle></math>","origin":"MathType for Microsoft Add-in"} (2-3)

1. C4.5决策树算法

C4.5相较于ID3最大的变化在于：C4.5选择信息增益比选择特征属性，而不是信息增益，从而避免了ID3算法再选择具有大量属性的数据集时可能出现的偏好选择问题。C4.5 决策树算法的计算方法与ID3类似，不再赘述。

1. CART决策树算法

CART即分类回归树(Classification Regression Tree),假设划分和的子样本内有k个样本，则衡量最优特征属性的标准Gini系数如下：

标题: G i n i 左小括号 t 右小括号 空格 等于 空格 1 空格 减 空格 加总 从 i 等于 1 到 k 对 p 下标 i 上标 2 - 说明 {"mathml":"<math style=\"font-family:stix;font-size:16px;\" xmlns=\"http://www.w3.org/1998/Math/MathML\"><mstyle mathsize=\"16px\"><mi>G</mi><mi>i</mi><mi>n</mi><mi>i</mi><mfenced><mi>t</mi></mfenced><mo>&#xA0;</mo><mo>=</mo><mo>&#xA0;</mo><mn>1</mn><mo>&#xA0;</mo><mo>-</mo><mo>&#xA0;</mo><mstyle displaystyle=\"false\"><munderover><mo>&#x2211;</mo><mrow><mi>i</mi><mo>=</mo><mn>1</mn></mrow><mi>k</mi></munderover></mstyle><msubsup><mi>p</mi><mi>i</mi><mn>2</mn></msubsup></mstyle></math>","origin":"MathType for Microsoft Add-in"} (2-4)

标题: p 下标 i - 说明 {"mathml":"<math style=\"font-family:stix;font-size:16px;\" xmlns=\"http://www.w3.org/1998/Math/MathML\"><mstyle mathsize=\"16px\"><msub><mi>p</mi><mi>i</mi></msub></mstyle></math>","origin":"MathType for Microsoft Add-in"}表示第i个样本类别出现的概率，根据特征属性t将D划分为m个子集，Gini系数可表示为：

标题: G i n i 左小括号 t 右小括号 空格 等于 空格 加总 从 i 等于 1 到 m 对 分数 n 分之 n 下标 i G i n i 左小括号 i 右小括号 - 说明 {"mathml":"<math style=\"font-family:stix;font-size:16px;\" xmlns=\"http://www.w3.org/1998/Math/MathML\"><mstyle mathsize=\"16px\"><mi>G</mi><mi>i</mi><mi>n</mi><mi>i</mi><mfenced><mi>t</mi></mfenced><mo>&#xA0;</mo><mo>=</mo><mo>&#xA0;</mo><mstyle displaystyle=\"false\"><munderover><mo>&#x2211;</mo><mrow><mi>i</mi><mo>=</mo><mn>1</mn></mrow><mi>m</mi></munderover></mstyle><mfrac><msub><mi>n</mi><mi>i</mi></msub><mi>n</mi></mfrac><mi>G</mi><mi>i</mi><mi>n</mi><mi>i</mi><mfenced><mi>i</mi></mfenced></mstyle></math>","origin":"MathType for Microsoft Add-in"}(2-5)

其中Gini(i)代表了第i个子数据集的Gini系数，n表示总数据量，

然而，决策树容易过拟合训练数据，特别是在处理复杂问题或数据噪声较多时。为了缓解这个问题，通常会使用剪枝等方法，但相对于其他分类模型还是比较严重

### 随机森林算法

随机森林算法是一种基于集成学习的机器学习算法，用于解决分类和回归问题。它通过构建多个决策树并将它们组合成一个强大的集成模型，以提高整体的预测性能和泛化能力。

随机森林最早于2001年由Breiman提出，他开创性地将决策树组合成随机森林，这是一种利用决策树对数据进行改造的方法：在构建每棵决策树时，随机森林算法会从原始数据集中进行有放回或无放回的随机抽样。这意味着每棵决策树都是基于不同的数据子集进行训练的；在每棵决策树的构建过程中，算法也会随机选择一部分特征进行训练。这样可以确保每棵树之间的差异性，增加模型的多样性；随机森林中的每棵决策树都是通过对随机抽样的数据子集进行递归地划分而生成的，这与传统的决策树算法相似。但在每个节点上，随机森林会使用随机选择的特征进行节点划分，而不是对所有特征进行尝试。当需要对新样本进行预测时，随机森林中的每棵决策树都会对其进行分类或回归，并给出一个预测结果。最后，随机森林会将每棵决策树的预测结果进行投票或平均，以确定最终的预测结果。

随机森林通过组合多棵决策树的预测结果，可以有效地减少过拟合的风险，提高模型的泛化能力。由于其具有并行化计算的特性，随机森林也适用于处理大规模数据集。同时，随机森林还能够提供特征重要性评估，帮助用户理解数据和模型。

在二分类问题中，我们通常用真阳（True Positive，TP）、假阳（False Positive，FP）、假阴（False Negative，FN）和真阴（True Negative，TN）这四种情况来衡量模型的分类结果。

真阳（TP）表示真实类别为正，模型预测结果也为正，即模型成功地将正类样本分类为正类。

假阳（FP）表示真实类别为正，但模型预测结果为负，即模型错误地将负类样本分类为正类。

假阴（FN）表示真实类别为负，但模型预测结果为正，即模型错误地将正类样本分类为负类。

真阴（TN）表示真实类别为负，模型预测结果也为负，即模型成功地将负类样本分类为负类。

这四种情况涵盖了所有可能的分类结果，通过对它们的计数和比较，可以评估模型的分类性能和准确度。

真阳性率（TPR）表示在所有实际正例样本中，模型正确预测为正例的样本数量与实际正例样本总数的比例。计算公式为TP/(FN+TP)。假阳性率（FPR）表示在所有实际负例样本中，模型错误地预测为正例的样本数量与实际负例样本总数的比例。计算公式为FP/(TN+FP)。ROC曲线用于评估随机森林算法的分类性能，其中横轴表示FPR，纵轴表示TPR。当TPR增加时，意味着模型正确地识别了更多的实际正例样本；而FPR的增加则意味着模型错误地将更多的实际负例样本预测为正例。

当随机森林算法进行分类任务时，每个样本都会得到一个概率值，代表其属于正类的可能性。这个概率值是通过随机森林中所有决策树的投票或平均得到的。通常情况下，我们会将概率值与一个阈值进行比较，如果概率值大于阈值，则将样本判定为正类，否则判定为负类。这个阈值的选择会影响最终的分类结果。为了评估模型在不同阈值下的表现，我们可以通过计算真阳性率（True Positive Rate，TPR）和假阳性率（False Positive Rate，FPR）来绘制ROC曲线。

首先，我们需要对数据集进行排序，将样本按照模型输出的概率值从高到低排列。然后，我们逐步调整阈值，依次计算每个阈值下的TPR和FPR。根据这些TPR和FPR的值，我们就可以绘制出ROC曲线。ROC曲线的横轴是FPR，表示假阳性率；纵轴是TPR，表示真阳性率。

通过观察ROC曲线，我们可以直观地了解模型在不同阈值下的性能表现。一般来说，ROC曲线越靠近左上角（0,1）点，说明模型的性能越好，因为此时TPR较高且FPR较低。ROC曲线下的面积AUC（Area Under the Curve）也可以作为评价模型性能的指标，AUC越大，表示模型的性能越好。而AUC值越小，即ROC曲线围成的面积越小，则意味着分类器还有改进的空间。但如果AUC非常小，意味着该分类器给出的结果成为了一个较好的反向预测指标。

在实际应用中，样本类别经常存在不平衡的情况。当测试集的正负样本分布发生变化时，ROC曲线能够保持不变。这意味着，当根据样本平衡性进行调整时，ROC曲线和AUC值大体上不会发生变化。因此，利用AUC来评估分类效果的有效性具有一定的广泛性。

随机森林算法集成了多个学习分类器的结果，相较于单一决策树，它主要有下面三个方面的优点：

1. 更强的泛化能力：由于随机森林是通过多个决策树的投票或取平均来进行分类或回归，因此在面对新数据时，往往具有更好的泛化能力，预测误差较小，有助于提高模型的稳定性和可靠性。
2. 适用于高维数据：随机森林能够有效地处理高维数据，而且还可以输出各个特征的重要性程度，这有助于了解数据集的特征和结构，帮助分析人员在特征选择和数据理解方面做出更准确的决策。
3. 鲁棒性强：随机森林对于缺失值和异常值的处理能力较强，能够在一定程度上容忍数据质量问题，因此在实际应用中更加稳健。此外，随机森林对于不平衡数据集也具有良好的容错性，能够有效应对类别不平衡的情况，避免模型在少数类别上的过度拟合。

## 拥挤交易

### PCA主成分分解算法

主成分分析（PCA）的目的是通过线性变换将高维数据集投影到一个低维子空间中，从而找到数据中的主要特征或模式。具体来说，PCA的目标是找到一个新的坐标系统，使得数据在这个新的坐标系统中的方差最大化。通过保留最重要的特征，可以减少数据的维度，同时保留数据中最重要的信息，从而简化后续的数据分析和建模过程，降低计算复杂度，去除数据中的噪声和冗余信息，以及可视化高维数据。

首先，我们来重新审视一下矩阵乘法的意义

标题: 左小括号 表格 列 开始保存格 p 下标 1 结束保存格 结束列 列 开始保存格 p 下标 2 结束保存格 结束列 列 垂直省略号 结束列 列 开始保存格 p 下标 R 结束保存格 结束列 结束表格 右小括号 空格 左小括号 表格 列 开始保存格 a 下标 1 结束保存格 开始保存格 a 下标 2 结束保存格 结束列 结束表格 中线省略号 空格 a 下标 M 空格 右小括号 空格 等于 空格 空格 左小括号 表格 列 开始保存格 p 下标 1 a 下标 1 结束保存格 开始保存格 p 下标 1 a 下标 2 结束保存格 中线省略号 开始保存格 p 下标 1 a 下标 M 结束保存格 结束列 列 开始保存格 p 下标 2 a 下标 1 结束保存格 开始保存格 p 下标 2 a 下标 2 结束保存格 中线省略号 开始保存格 p 下标 2 a 下标 M 结束保存格 结束列 列 垂直省略号 垂直省略号 下斜省略号 垂直省略号 结束列 列 开始保存格 p 下标 R a 下标 1 结束保存格 开始保存格 p 下标 R a 下标 2 结束保存格 中线省略号 开始保存格 p 下标 R a 下标 M 结束保存格 结束列 结束表格 右小括号 - 说明 {"mathml":"<math style=\"font-family:stix;font-size:16px;\" xmlns=\"http://www.w3.org/1998/Math/MathML\"><mstyle mathsize=\"16px\"><mfenced><mtable><mtr><mtd><msub><mi>p</mi><mn>1</mn></msub></mtd></mtr><mtr><mtd><msub><mi>p</mi><mn>2</mn></msub></mtd></mtr><mtr><mtd><mo>&#x22EE;</mo></mtd></mtr><mtr><mtd><msub><mi>p</mi><mi>R</mi></msub></mtd></mtr></mtable></mfenced><mo>&#xA0;</mo><mfenced><mrow><mtable><mtr><mtd><msub><mi>a</mi><mn>1</mn></msub></mtd><mtd><msub><mi>a</mi><mn>2</mn></msub></mtd></mtr></mtable><mo>&#x22EF;</mo><mo>&#xA0;</mo><msub><mi>a</mi><mi>M</mi></msub><mo>&#xA0;</mo></mrow></mfenced><mo>&#xA0;</mo><mo>=</mo><mo>&#xA0;</mo><mo>&#xA0;</mo><mfenced><mtable><mtr><mtd><msub><mi>p</mi><mn>1</mn></msub><msub><mi>a</mi><mn>1</mn></msub></mtd><mtd><msub><mi>p</mi><mn>1</mn></msub><msub><mi>a</mi><mn>2</mn></msub></mtd><mtd><mo>&#x22EF;</mo></mtd><mtd><msub><mi>p</mi><mn>1</mn></msub><msub><mi>a</mi><mi>M</mi></msub></mtd></mtr><mtr><mtd><msub><mi>p</mi><mn>2</mn></msub><msub><mi>a</mi><mn>1</mn></msub></mtd><mtd><msub><mi>p</mi><mn>2</mn></msub><msub><mi>a</mi><mn>2</mn></msub></mtd><mtd><mo>&#x22EF;</mo></mtd><mtd><msub><mi>p</mi><mn>2</mn></msub><msub><mi>a</mi><mi>M</mi></msub></mtd></mtr><mtr><mtd><mo>&#x22EE;</mo></mtd><mtd><mo>&#x22EE;</mo></mtd><mtd><mo>&#x22F1;</mo></mtd><mtd><mo>&#x22EE;</mo></mtd></mtr><mtr><mtd><msub><mi>p</mi><mi>R</mi></msub><msub><mi>a</mi><mn>1</mn></msub></mtd><mtd><msub><mi>p</mi><mi>R</mi></msub><msub><mi>a</mi><mn>2</mn></msub></mtd><mtd><mo>&#x22EF;</mo></mtd><mtd><msub><mi>p</mi><mi>R</mi></msub><msub><mi>a</mi><mi>M</mi></msub></mtd></mtr></mtable></mfenced></mstyle></math>","origin":"MathType for Microsoft Add-in"} (2-6)

为了准确描述向量，我们首先要确定一组基，然后给出向量再基所在的方向上的投影。如果基发生了变化，要求同样的向量在新基下的坐标需要一些奇妙的线性变换。

在上面的公式中，我们首先假设行向量pi表示第i个新基在旧基座系下的表示，列向量aj表示第j个列向量在旧基座标系的表示，而两者相乘的结果则是原列向量组a1…aM在新基下的坐标。

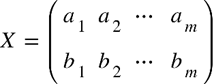
以上分析从坐标变换的视角给矩阵相乘一个直观的解释，即：两个矩阵相乘的意义是将右边矩阵中的每一列向量aj变换到左边矩阵中以每一行行向量为基所表示的空间中去。也就是说一个矩阵可以表示一种线性变换。

在一维空间中，我们可以使用方差来衡量数据的分散程度。然而，在高维空间中，我们使用协方差来描述变量之间的关系。协方差反映了两个变量的相关性。为了使得变量能够尽可能地表达原始数据的信息，我们希望它们之间的线性相关性尽可能地弱。协方差公式如下：

标题: C o v 左小括号 a 逗号 b 右小括号 空格 等于 空格 分数 m 空格 减 空格 1 分之 1 结束分数 加总 从 i 等于 1 到 m 对 左小括号 a 下标 i 空格 减 空格 mu （ 小写 ） 下标 a 右小括号 左小括号 b 下标 i 空格 减 空格 mu （ 小写 ） 下标 b 右小括号 - 说明 {"mathml":"<math style=\"font-family:stix;font-size:16px;\" xmlns=\"http://www.w3.org/1998/Math/MathML\"><mstyle mathsize=\"16px\"><mi>C</mi><mi>o</mi><mi>v</mi><mfenced><mrow><mi>a</mi><mo>,</mo><mi>b</mi></mrow></mfenced><mo>&#xA0;</mo><mo>=</mo><mo>&#xA0;</mo><mfrac><mn>1</mn><mrow><mi>m</mi><mo>&#xA0;</mo><mo>-</mo><mo>&#xA0;</mo><mn>1</mn></mrow></mfrac><munderover><mo>&#x2211;</mo><mrow><mi>i</mi><mo>=</mo><mn>1</mn></mrow><mi>m</mi></munderover><mfenced><mrow><msub><mi>a</mi><mi>i</mi></msub><mo>&#xA0;</mo><mo>-</mo><mo>&#xA0;</mo><msub><mi>&#x3BC;</mi><mi>a</mi></msub></mrow></mfenced><mfenced><mrow><msub><mi>b</mi><mi>i</mi></msub><mo>&#xA0;</mo><mo>-</mo><mo>&#xA0;</mo><msub><mi>&#x3BC;</mi><mi>b</mi></msub></mrow></mfenced></mstyle></math>","origin":"MathType for Microsoft Add-in"} (2-7)

如果能确保变量a和b的均值为0，而且在数据量较大时m和m-1的差距可以忽略不记，那么协方差公式可以表示为

标题: C o v 左小括号 a 逗号 b 右小括号 空格 等于 空格 分数 m 分之 1 加总 从 i 等于 1 到 m 对 a 下标 i b 下标 i - 说明 {"mathml":"<math style=\"font-family:stix;font-size:16px;\" xmlns=\"http://www.w3.org/1998/Math/MathML\"><mstyle mathsize=\"16px\"><mi>C</mi><mi>o</mi><mi>v</mi><mfenced><mrow><mi>a</mi><mo>,</mo><mi>b</mi></mrow></mfenced><mo>&#xA0;</mo><mo>=</mo><mo>&#xA0;</mo><mfrac><mn>1</mn><mi>m</mi></mfrac><munderover><mo>&#x2211;</mo><mrow><mi>i</mi><mo>=</mo><mn>1</mn></mrow><mi>m</mi></munderover><msub><mi>a</mi><mi>i</mi></msub><msub><mi>b</mi><mi>i</mi></msub></mstyle></math>","origin":"MathType for Microsoft Add-in"} (2-8)

假设将a和b两个变量按行组成矩阵X：

那么a与b构成的协方差矩阵可以表示为：

标题: 分数 m 分之 1 空格 X X 的 T 次方 空格 等于 空格 左小括号 表格 列 开始保存格 分数 m 分之 1 加总 从 i 等于 1 到 m 对 a 下标 i 上标 2 结束保存格 开始保存格 分数 m 分之 1 加总 从 i 等于 1 到 m 对 a 下标 i 上标 空白 b 下标 i 结束保存格 结束列 列 开始保存格 分数 m 分之 1 加总 从 i 等于 1 到 m 对 a 下标 i 上标 空白 b 下标 i 结束保存格 开始保存格 分数 m 分之 1 加总 从 i 等于 1 到 m 对 b 下标 i 上标 2 结束保存格 结束列 结束表格 右小括号 空格 等于 空格 左小括号 表格 列 开始保存格 C o v 左小括号 a 逗号 a 右小括号 结束保存格 开始保存格 C o v 左小括号 a 逗号 b 右小括号 结束保存格 结束列 列 开始保存格 C o v 左小括号 b 逗号 a 右小括号 结束保存格 开始保存格 C o v 左小括号 b 逗号 b 右小括号 结束保存格 结束列 结束表格 右小括号 - 说明 {"mathml":"<math xmlns=\"http://www.w3.org/1998/Math/MathML\" style=\"font-family:stix;font-size:16px;\"><mfrac><mn>1</mn><mi>m</mi></mfrac><mo>&#xA0;</mo><mi>X</mi><msup><mi>X</mi><mi>T</mi></msup><mo>&#xA0;</mo><mo>=</mo><mo>&#xA0;</mo><mfenced><mtable><mtr><mtd><mfrac><mn>1</mn><mi>m</mi></mfrac><mstyle displaystyle=\"false\"><munderover><mo>&#x2211;</mo><mrow><mi>i</mi><mo>=</mo><mn>1</mn></mrow><mi>m</mi></munderover></mstyle><msubsup><mi>a</mi><mi>i</mi><mn>2</mn></msubsup></mtd><mtd><mfrac><mn>1</mn><mi>m</mi></mfrac><mstyle displaystyle=\"false\"><munderover><mo>&#x2211;</mo><mrow><mi>i</mi><mo>=</mo><mn>1</mn></mrow><mi>m</mi></munderover></mstyle><msubsup><mi>a</mi><mi>i</mi><mrow/></msubsup><msub><mi>b</mi><mi>i</mi></msub></mtd></mtr><mtr><mtd><mfrac><mn>1</mn><mi>m</mi></mfrac><mstyle displaystyle=\"false\"><munderover><mo>&#x2211;</mo><mrow><mi>i</mi><mo>=</mo><mn>1</mn></mrow><mi>m</mi></munderover></mstyle><msubsup><mi>a</mi><mi>i</mi><mrow/></msubsup><msub><mi>b</mi><mi>i</mi></msub></mtd><mtd><mfrac><mn>1</mn><mi>m</mi></mfrac><mstyle displaystyle=\"false\"><munderover><mo>&#x2211;</mo><mrow><mi>i</mi><mo>=</mo><mn>1</mn></mrow><mi>m</mi></munderover></mstyle><msubsup><mi>b</mi><mi>i</mi><mn>2</mn></msubsup></mtd></mtr></mtable></mfenced><mo>&#xA0;</mo><mo>=</mo><mo>&#xA0;</mo><mfenced><mtable><mtr><mtd><mi>C</mi><mi>o</mi><mi>v</mi><mfenced><mrow><mi>a</mi><mo>,</mo><mi>a</mi></mrow></mfenced></mtd><mtd><mi>C</mi><mi>o</mi><mi>v</mi><mfenced><mrow><mi>a</mi><mo>,</mo><mi>b</mi></mrow></mfenced></mtd></mtr><mtr><mtd><mi>C</mi><mi>o</mi><mi>v</mi><mfenced><mrow><mi>b</mi><mo>,</mo><mi>a</mi></mrow></mfenced></mtd><mtd><mi>C</mi><mi>o</mi><mi>v</mi><mfenced><mrow><mi>b</mi><mo>,</mo><mi>b</mi></mrow></mfenced></mtd></mtr></mtable></mfenced></math>","origin":"MathType for Microsoft Add-in"} (2-9)

推广到一般情况，假设有m个n维记录，将其排列为矩阵，优化的目的是把这个n维记录降低至K维，同时尽最大可能保证降至K维后的数据方差最大化。假设标题: C 空格 等于 空格 分数 m 分之 1 X X 的 T 次方 - 说明 {"mathml":"<math style=\"font-family:stix;font-size:16px;\" xmlns=\"http://www.w3.org/1998/Math/MathML\"><mstyle mathsize=\"16px\"><mi>C</mi><mo>&#xA0;</mo><mo>=</mo><mo>&#xA0;</mo><mfrac><mn>1</mn><mi>m</mi></mfrac><mi>X</mi><msup><mi>X</mi><mi>T</mi></msup></mstyle></math>","origin":"MathType for Microsoft Add-in"}，则C是一个对称矩阵，对角线元素维各个变量的方差，第i行第j列的元素表示i和j两个变量的协方差。

按照PCA分解的思路，我们要实现的目标是寻找一个线性变换P，经过线性变换Y=PX后Y的协方差矩阵除对角线的元素均为0，并且在对角线上将元素按大小从上到下排列（变量方差尽可能大）。

假设原始数据矩阵 X 对应的协方差矩阵为 C，而 P 是一组基按行组成的矩阵，设 Y=PX，则 Y 为 X 对 P 做基变换后的数据。假设设 Y 的协方差矩阵为 D，要求D是只有主对角线元素非0的矩阵，我们推导一下 D 与 C 的关系：

标题: D 空格 等于 空格 分数 m 分之 1 Y Y 的 T 次方 等于 空格 分数 m 分之 1 左小括号 P X 右小括号 左小括号 P X 右小括号 的 T 次方 空格 等于 空格 分数 m 分之 1 P X X 的 T 次方 P 的 T 次方 空格 等于 空格 P 左小括号 分数 m 分之 1 X X 的 T 次方 右小括号 P 的 T 次方 空格 等于 空格 P C P 的 T 次方 - 说明 {"mathml":"<math style=\"font-family:stix;font-size:16px;\" xmlns=\"http://www.w3.org/1998/Math/MathML\"><mstyle mathsize=\"16px\"><mi>D</mi><mo>&#xA0;</mo><mo>=</mo><mo>&#xA0;</mo><mfrac><mn>1</mn><mi>m</mi></mfrac><mi>Y</mi><msup><mi>Y</mi><mi>T</mi></msup><mo>=</mo><mo>&#xA0;</mo><mfrac><mn>1</mn><mi>m</mi></mfrac><mfenced><mrow><mi>P</mi><mi>X</mi></mrow></mfenced><msup><mfenced><mrow><mi>P</mi><mi>X</mi></mrow></mfenced><mi>T</mi></msup><mo>&#xA0;</mo><mo>=</mo><mo>&#xA0;</mo><mfrac><mn>1</mn><mi>m</mi></mfrac><mi>P</mi><mi>X</mi><msup><mi>X</mi><mi>T</mi></msup><msup><mi>P</mi><mi>T</mi></msup><mo>&#xA0;</mo><mo>=</mo><mo>&#xA0;</mo><mi>P</mi><mfenced><mrow><mfrac><mn>1</mn><mi>m</mi></mfrac><mi>X</mi><msup><mi>X</mi><mi>T</mi></msup></mrow></mfenced><msup><mi>P</mi><mi>T</mi></msup><mo>&#xA0;</mo><mo>=</mo><mo>&#xA0;</mo><mi>P</mi><mi>C</mi><msup><mi>P</mi><mi>T</mi></msup></mstyle></math>","origin":"MathType for Microsoft Add-in"} (2-10)

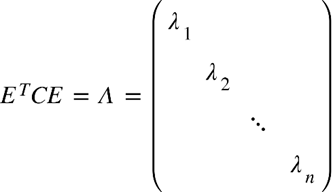
由此可见，新的基矩阵P是能使原始协方差矩阵对角化的矩阵，换句话说，PCA分解的目标是寻找一个矩阵P，使得标题: P C P 的 T 次方 - 说明 {"mathml":"<math style=\"font-family:stix;font-size:16px;\" xmlns=\"http://www.w3.org/1998/Math/MathML\"><mstyle mathsize=\"16px\"><mi>P</mi><mi>C</mi><msup><mi>P</mi><mi>T</mi></msup></mstyle></math>","origin":"MathType for Microsoft Add-in"}是一个对角矩阵，且对角线上的元素从大到小顺序排列，那么P的前K行即是需要的基，用P的前K行组成的矩阵乘以X就使得X从N维降到了K维。

将实对称矩阵C对角化需要用到特征向量。实对称矩阵有一些非常好的性质，能在对角化时带来许多便捷之处：

1. 实对称矩阵不同特征值对应的特征向量正交；
2. 假设某个特征值的重数为n，则该特征值对应着n个线性无关的特征向量，可将这n个特征向量单位正交化。

所以，对于实对称矩阵Cn,n必然存在n个单位正交的特征向量标题: e 下标 1 逗号 e 下标 2 逗号 中线省略号 逗号 e 下标 n - 说明 {"mathml":"<math style=\"font-family:stix;font-size:16px;\" xmlns=\"http://www.w3.org/1998/Math/MathML\"><mstyle mathsize=\"16px\"><msub><mi>e</mi><mn>1</mn></msub><mo>,</mo><msub><mi>e</mi><mn>2</mn></msub><mo>,</mo><mo>&#x22EF;</mo><mo>,</mo><msub><mi>e</mi><mi>n</mi></msub></mstyle></math>","origin":"MathType for Microsoft Add-in"}，将其按列组成矩阵：标题: E 空格 等于 空格 左小括号 e 下标 1 逗号 e 下标 2 逗号 中线省略号 逗号 e 下标 n 右小括号 - 说明 {"mathml":"<math xmlns=\"http://www.w3.org/1998/Math/MathML\" style=\"font-family:stix;font-size:16px;\"><mi>E</mi><mo>&#xA0;</mo><mo>=</mo><mo>&#xA0;</mo><mfenced><mrow><msub><mi>e</mi><mn>1</mn></msub><mo>,</mo><msub><mi>e</mi><mn>2</mn></msub><mo>,</mo><mo>&#x22EF;</mo><mo>,</mo><msub><mi>e</mi><mi>n</mi></msub></mrow></mfenced></math>","origin":"MathType for Microsoft Add-in"}

则有：

 (2-11)

其中Ʌ是对角矩阵，对角线上的元素为各个特征向量对应的特征值，对比两式可以发现：标题: P 空格 等于 空格 E 的 T 次方 - 说明 {"mathml":"<math style=\"font-family:stix;font-size:16px;\" xmlns=\"http://www.w3.org/1998/Math/MathML\"><mstyle mathsize=\"16px\"><mi>P</mi><mo>&#xA0;</mo><mo>=</mo><mo>&#xA0;</mo><msup><mi>E</mi><mi>T</mi></msup></mstyle></math>","origin":"MathType for Microsoft Add-in"}

综上，总结一下PCA降维算法的步骤：假设有m条n维数据，想将数据降至K维

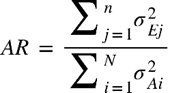
1. 将原始数据按列组成n行m列矩阵X；
2. 将X的每一行进行零均值化
3. 按照标题: C 空格 等于 空格 分数 m 分之 1 X X 的 T 次方 - 说明 {"mathml":"<math style=\"font-family:stix;font-size:16px;\" xmlns=\"http://www.w3.org/1998/Math/MathML\"><mstyle mathsize=\"16px\"><mi>C</mi><mo>&#xA0;</mo><mo>=</mo><mo>&#xA0;</mo><mfrac><mn>1</mn><mi>m</mi></mfrac><mi>X</mi><msup><mi>X</mi><mi>T</mi></msup></mstyle></math>","origin":"MathType for Microsoft Add-in"}求出协方差矩阵
4. 求出协方差矩阵的特征值和特征向量
5. 将特征向量按对应的特征值的大小自上而下按列排行成矩阵，取前k行组成矩阵P
6. 标题: Y 空格 等于 空格 P X - 说明 {"mathml":"<math style=\"font-family:stix;font-size:16px;\" xmlns=\"http://www.w3.org/1998/Math/MathML\"><mstyle mathsize=\"16px\"><mi>Y</mi><mo>&#xA0;</mo><mo>=</mo><mo>&#xA0;</mo><mi>P</mi><mi>X</mi></mstyle></math>","origin":"MathType for Microsoft Add-in"}，获得降至K维后的结果

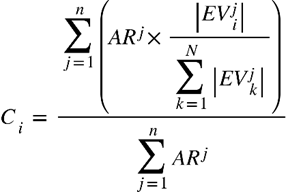
### 资金集中度

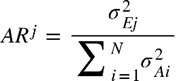
在解释“资金集中度”这一指标前，首先要了解拥挤交易现象。拥挤交易是指一群交易者动用大量金购买或出售一项资产或一组具有相似特征的资产，从而导致资产价格发生重大变化。如果是因为基本面变化导致的资产价格波动，那么资产价格会偏离原先的均衡水平很长一段时间，但是如果在基本面没有变化的前提下由于拥挤交易产生资产价格的大幅波动，那么价格将在一段时间内回到原来的水平。

如何识别出正在发生拥挤交易的行业板块，并借此机会盈利，是广大股民孜孜不倦追求的目标。如果仅仅根据机构披露出的持仓明细来获知拥挤交易是否发生是远远不足的，因为机构往往延迟一个季度披露持仓明细，按图索骥会带来很大的误导性。本文借鉴了Kinlaw(2019)提出的基于资金集中度的拥挤交易代理指标，先对这一指标做详细阐释。

Kinlaw(2019)直接对行业指数和市值的变化进行分析间接得到资金在某些行业的集中程度：他构造了“资金集中度”的指标，用以分析资产在其所在的板块中影响收益变化的程度，从而衡量整个板块的拥挤交易现象，具体公式如下：

 (2-12)

 (2-13)

 (2-14)

在式（2-12）中，N代表行业板块的数量，n等于PCA降维后的目标维数，即吸收比率中特征向量的个数，标题: sigma （ 小写 ） 下标 A i 结束下标 上标 2 - 说明 {"mathml":"<math style=\"font-family:stix;font-size:16px;\" xmlns=\"http://www.w3.org/1998/Math/MathML\"><mstyle mathsize=\"16px\"><msubsup><mi>&#x3C3;</mi><mrow><mi>A</mi><mi>i</mi></mrow><mn>2</mn></msubsup></mstyle></math>","origin":"MathType for Microsoft Add-in"}代表第i个行业板块的方差，标题: sigma （ 小写 ） 下标 E j 结束下标 上标 2 - 说明 {"mathml":"<math style=\"font-family:stix;font-size:16px;\" xmlns=\"http://www.w3.org/1998/Math/MathML\"><mstyle mathsize=\"16px\"><msubsup><mi>&#x3C3;</mi><mrow><mi>E</mi><mi>j</mi></mrow><mn>2</mn></msubsup></mstyle></math>","origin":"MathType for Microsoft Add-in"}代表第j个特征向量的方差，亦即第i个特征值，AR即整个市场的吸收比率的总和。在式（2-13）中，标题: E V 下标 i 上标 j - 说明 {"mathml":"<math style=\"font-family:stix;font-size:16px;\" xmlns=\"http://www.w3.org/1998/Math/MathML\"><mstyle mathsize=\"16px\"><mi>E</mi><msubsup><mi>V</mi><mi>i</mi><mi>j</mi></msubsup></mstyle></math>","origin":"MathType for Microsoft Add-in"}表示第i个行业在第j个特征向量的载荷，该比值反映了该行业在第j个特征向量的权重。

吸收比率的建立利用了主成分分析法（PCA）的降维思想。它通过对原始变量进行重新组合和降维，计算能够最大限度解释原始变量信息的互不相关的主成分数量。同时，通过计算各行业收益总方差能够被一定数量的特征向量解释的比率，来分析主成分随时间变化对原始变量信息的解释能力，进而推测各变量之间关联性的紧密程度，以此来衡量内部风险分散程度。总的来说，吸收比率展示了特征向量对样本总方差的解释力度，吸收比率越高，表示资产内部的系统性风险越集中，各种风险类型较为统一，分散风险功能较弱，更容易受到负面消息的冲击；而吸收率越低，对应的系统性风险也较低，内部风险更为分散，分散风险功能更强，临时抵御能力更强，不太容易因负面消息而发生崩塌。

“资金集中度”指标本质上就是每个特征向量的吸收比率乘以各行业在该特征向量中的权重，再除以所有构成该特征向量的子集的吸收比率之和，衡量了行业对样本总方差的影响力。该指标显示了某个行业能在多大程度上推动其他关联板块收益波动。如果一个行业的集中度较高，意味着其能够带动其他行业板块收益波动的程度就更大，即该行业板块相对于其他行业板块而言，波动性和关联性更强，这可能代表该行业出现了拥挤交易现象。

造成这一现象的主要原因包括以下两点：首先是关于波动性。当某一行业出现拥挤交易时，大量资金涌向该行业，导致大量不平衡订单（或买多于卖，或卖多于买），进而引发价格大幅调整，该行业的波动率随着成交量的增加而增大。此外，由于许多投资者将该行业视为一个整体进行配置，而不是分配交易到行业内的个股，导致行业内个股之间的关联性增强，使得行业的风险来源更加统一，其风险分散能力降低，因此该行业的波动性也会增大。其次是关于关联性。当某一行业出现拥挤交易时，随着大量资金涌入，该行业将成为市场的焦点，这种市场信息将推动其他行业的变化，使得行业之间的关联性增强。

# 基于拥挤交易的行业轮动策略

## 样本选取与数据处理

在探究行业轮动规律时，行业的分类至关重要。Kinlaw(2019)选择按照标准普尔的行业分类选择行业，包括工业(Industry)、医疗保健(Medical & Health)、基础材料(Materials)、消费者非必需品(Unnecessary Consume)、消费者常用品(Necessary Consume)、能源(Energy)、电信服务(Telecom)、金融(Finance)、信息技术(Information technology)、公用事业(Utilities)和地产业(Real Estate)这11个门类。考虑到中美之间的国情差异，这11个门类的划分实在过于粗糙。A股市场的一个非常明显的特点是炒概念现象非常严重，一旦一支股票蹭上了某个热门的概念，在多方力量的共同作用下，其股价马上就会快速拉升，最后在山顶留下大量散户投资者。目前在五千余家上市公司中已经涌现出接近400个概念板块，也有一部分股票横跨多个概念板块。由于本人精力实在有限，无法对如此多的概念板块进行筛选分类，最终，我选择了申万一级行业分类作为本文的行业分类标准。目前最新的申万一级行业分类中包含28项，分别是：采掘、通信、非银金融、银行、化工、有色金属、钢铁、交通运输、汽车、综合、房地产、商业贸易、电子、家用电器、公用事业、轻工制造、医药生物、电气设备、农林牧渔、计算机、机械设备、休闲服务、建筑装饰、国防军工、传媒、建筑材料、食品饮料和纺织服装，其中有部分行业直至2024年2月才开始启用。因此本文选取2014年2月20日至2024年1月19日申万一级行业指数的相关信息作为分析对象。

本文从同花顺iFind数据库中提取了28个申万一级行业在上述时间区间内的指数值、市净率及行业市值等数据，由于这些指数数据比较偏宏观，数据缺失的情况较少，经过简单的筛选后得到可用数据。

## “资金集中度”因子的构建

按照KinLaw(2019)中给出的步骤，结合中国28个申万一级行业指数在2014年2月20日至2024年1月19日间的相关数据，构建“资金集中度”的步骤如下：

1. 计算这28个申万一级行业的日度收益率序列，利用StandardScaler()函数对日度收益率进行标准化处理，即减去均值再除以方差；
2. 根据每日行业市值数据，计算各个行业占总市值权重，再用标准化后的日度收益率乘以权重的平方根，用以获得加权后的各个行业日度收益率曲线。这里对权重进行开方处理的原因是作者认为：市值大的行业和其他行业关联性更强，但当市值达到一定程度时关联性会达到饱和水平
3. 以1年（252个交易日）为窗口期滚动计算集中度。考虑到远期数据对当前时点的作用效果会逐步减弱，作者采取了指数衰减的方式对数据进行加权处理，以提高近期数据的权重。这样就不能直接调用PCA()函数获得主成分分析的结果，而是要回到PCA算法的本源：首先用半衰期半年（126个交易日）对原始数据进行处理，然后生成协方差矩阵，使用numpy.linalg下的eig函数获得协方差矩阵的特征值和特征向量，对特征值进行排序，筛选出前2大(n=2)的特征值和对应的特征向量；
4. 通过式（）计算每个行业的集中度因子。

计算结果如下图所示：

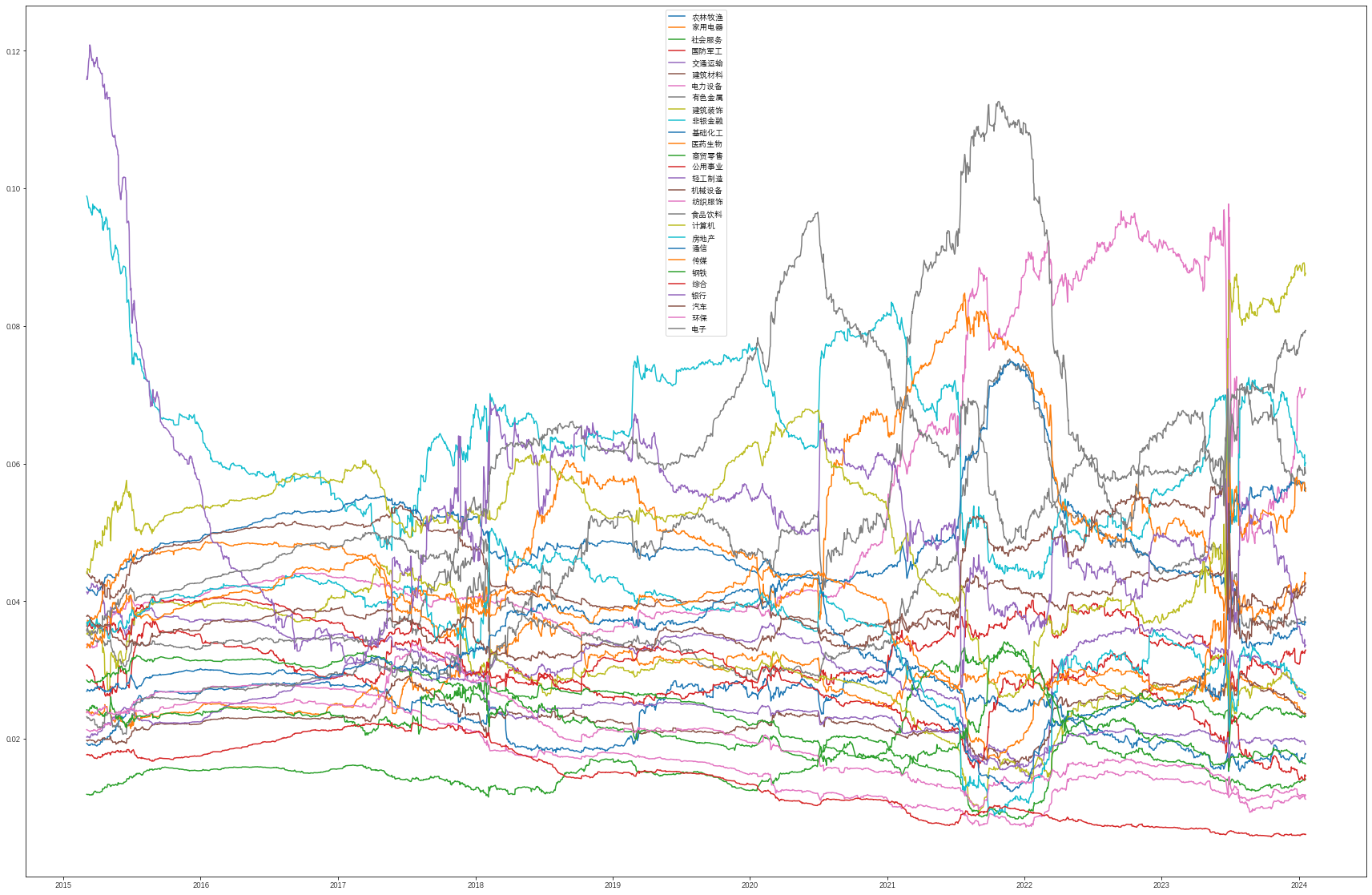


图3. 28个申万一级行业集中度因子

## 相对估值指标的构建

集中度因子只能在一定程度上反应在某个行业内资金的活跃程度，但是拥挤交易也会在投资者争相抛售资产的时候发生，如果这个时候买入，势必会造成巨大的亏损。因此，单单使用集中度一个因子是不够的，还需要另一些指标判断拥挤交易是买入还是卖出导致的。

一般而言，当投资者认为当前资产估值水平过高时，会考虑卖出，当投资者形成一致预期时，集中抛售就会导致资产价格大幅下跌。因此，可以选择估值指标作为卖出信号。目前最常用的两个估值指标是市盈率(PR)和市净率(PE)，其中，市盈率等于市值除以净利润，市净率等于市值除以净资产。对于周期性行业，其利润不稳定，为了保证指标的相对稳定性，选取市净率作为估值指标。但是由于行业间的性质不一样，市净率的绝对水平会有较大差异，因此构建相对估值指标，即行业估值在历史上的百分位水平，再将各个行业的相对估值在横截面上进行比较，具体计算步骤如下：

1. 计算申万一级行业在每个交易日的市净率；
2. 将当前市净率除以过去2年的平均值，得到每个行业的标准化市净率；
3. 将每个行业标准化后的市净率除以其他所有27个行业标准化市净率的平均，得到每个行业的相对标准化市净率，该值作为横截面上的相对价值参考指标

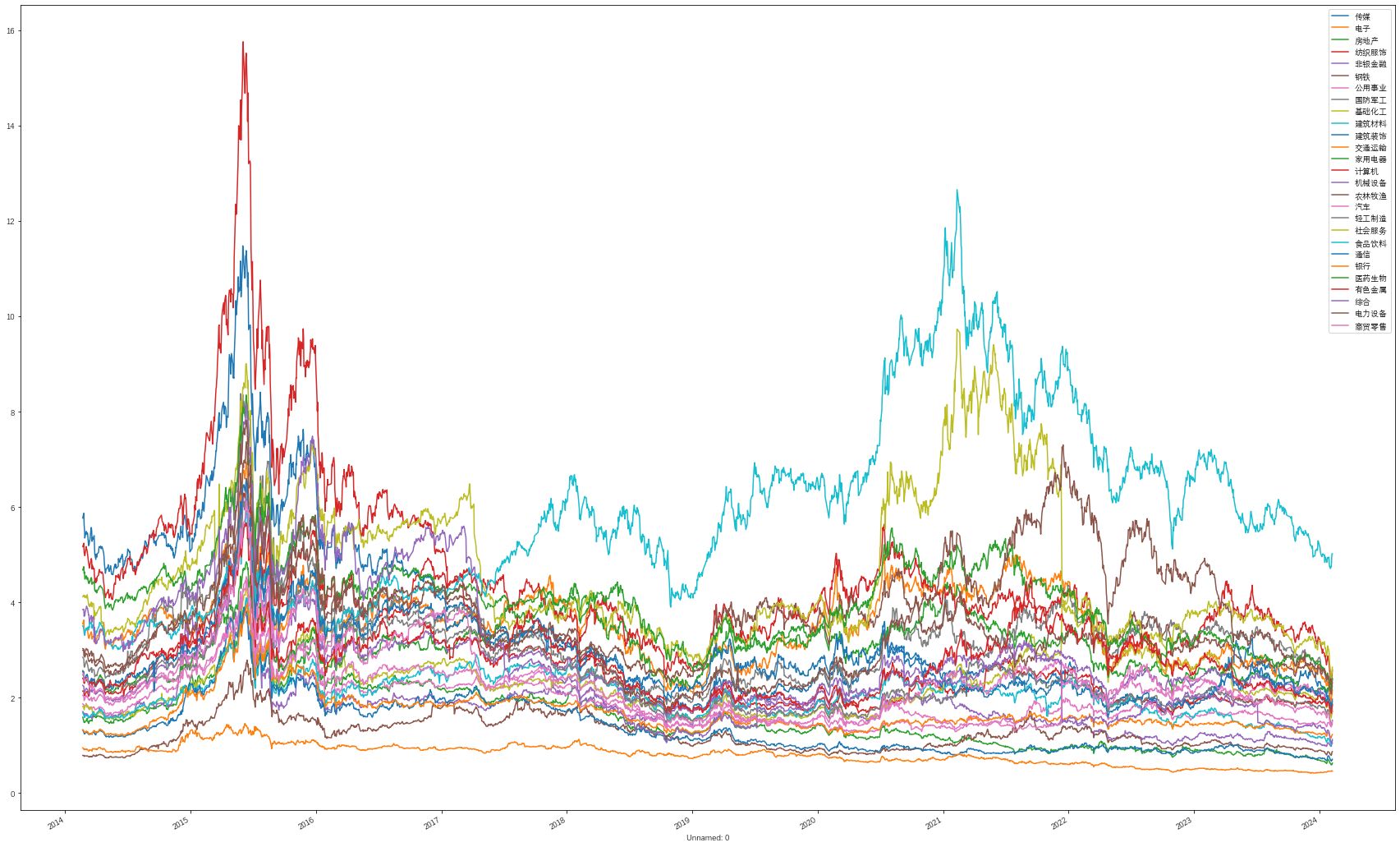


图4. 28个申万一级行业市净率

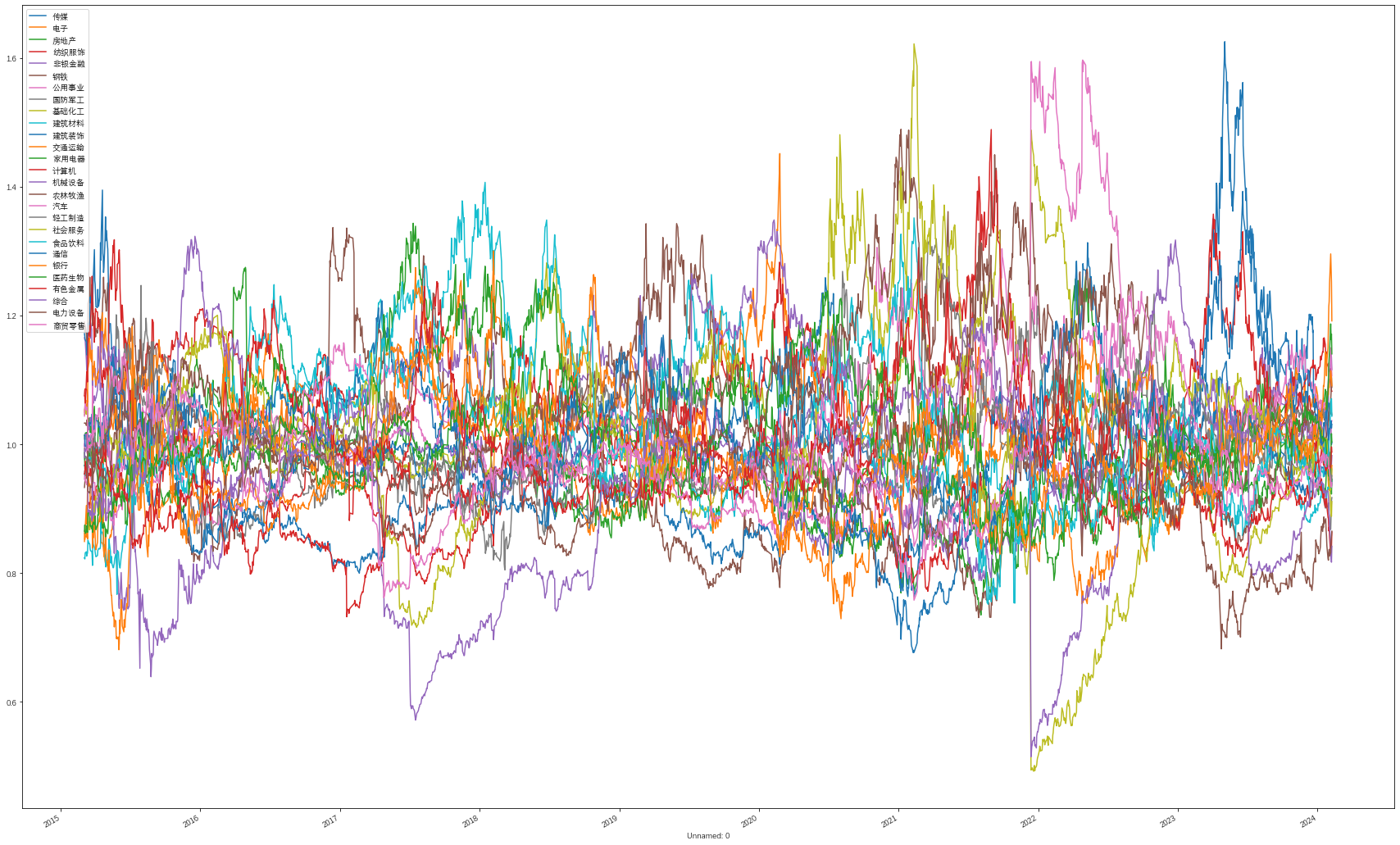


图5. 28个申万一级行业相对估值

## 因子有效性的检验

根据上一小节计算出的“资金集中度”和“相对估值”因子，构建如下行业轮动策略：当资金集中度较高且相对估值在中等偏下的水平时，意味着该行业处于拥挤交易泡沫形成期；当该行业相对估值较高的时候如果集中度升高，则意味着泡沫可能破裂，此时应卖出处于该行业的资产，获得超额收益。

为了评判该策略的有效性，首先使用排序打分法和分层测试法检验集中度因子在行业轮动效应中的有效性，而后结合相对估值指标，计算行业轮动策略的回测表现，筛选出最合适的买入、卖出时点。

### 排序打分法检验

排序打分法是一种基于信息系数（IC）值和信息比率（IR）值的统计检验方法。IC值是指所选资产当期的因子值与下一期收益率之间的Pearson相关系数，数学描述如下：

标题: I C 的 T 次方 空格 等于 空格 c o r 左小括号 F 的 T 次方 逗号 R 的 T 加 1 次方 右小括号 - 说明 {"mathml":"<math style=\"font-family:stix;font-size:16px;\" xmlns=\"http://www.w3.org/1998/Math/MathML\"><mstyle mathsize=\"16px\"><mi>I</mi><msup><mi>C</mi><mi>T</mi></msup><mo>&#xA0;</mo><mo>=</mo><mo>&#xA0;</mo><mi>c</mi><mi>o</mi><mi>r</mi><mfenced><mrow><msup><mi>F</mi><mi>T</mi></msup><mo>,</mo><msup><mi>R</mi><mrow><mi>T</mi><mo>+</mo><mn>1</mn></mrow></msup></mrow></mfenced></mstyle></math>","origin":"MathType for Microsoft Add-in"}

其中标题: R 的 T 加 1 次方 - 说明 {"mathml":"<math style=\"font-family:stix;font-size:16px;\" xmlns=\"http://www.w3.org/1998/Math/MathML\"><mstyle mathsize=\"16px\"><msup><mi>R</mi><mrow><mi>T</mi><mo>+</mo><mn>1</mn></mrow></msup></mstyle></math>","origin":"MathType for Microsoft Add-in"}代表第T+1期的收益率序列，标题: F 的 T 次方 - 说明 {"mathml":"<math xmlns=\"http://www.w3.org/1998/Math/MathML\" style=\"font-family:stix;font-size:16px;\"><msup><mi>F</mi><mi>T</mi></msup></math>","origin":"MathType for Microsoft Add-in"}代表第T期的因子值序列，标题: I C 的 T 次方 - 说明 {"mathml":"<math xmlns=\"http://www.w3.org/1998/Math/MathML\" style=\"font-family:stix;font-size:16px;\"><mi>I</mi><msup><mi>C</mi><mi>T</mi></msup></math>","origin":"MathType for Microsoft Add-in"}为第T期因子的IC值。

IC值可以在统计意义上评估当期因子对下期收益率的预测能力，若IC值均值为正表示因子值越大越好，反之表示因子值越小越好。一般认为，当IC值均值的绝对值大于0.03时，表明该因子的选股能力较强。信息比率（IR）等于信息系数的均值除以信息系数的方差，用以衡量IC值的稳定性。一般而言，当IR值大于0.5时，意味着因子具有稳定获得超额收益的能力。

IC值的具体计算方法如下：首先获得申万一级行业知识每月最后一个交易日的集中度因子值以及下月收益率，并将集中度因子针对历史水平进行标准化；而后在每月最后一个交易日的截面上求取Pearson截面相关系数；最后求取IC值均值和标准差并得到IR值

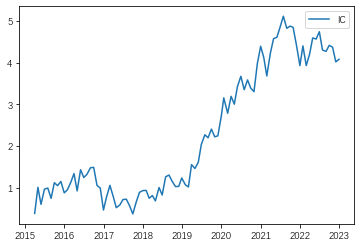


图6. IC值累计曲线

表1. 排序打分法有效性检验

|  |  |  |
| --- | --- | --- |
| IC值均值 | IC值标准差 | IR值 |
| 0.043389 | 0.120773 | 0.359259 |

结果显示，集中度的IC值均值的绝对值为0.043389，大于0.03，表明集中度因子与下一期收益之间存在相关性。同时，IC累计曲线向右上方倾斜，表明集中度因子能有效预测下一期的强势行业。

### 分层测试法检验

相较于排序打分法，分层测试法能够展示更多的细节。其本质是利用单个因子构建投资组合进行回测，并对投资组合的绩效进行全面评估，以此来评价因子的有效性。具体操作是，在每月的最后一个交易日将各个行业的集中度因子由小到大进行排序，每一期构建5个投资组合，集中度因子最大的记为第5分位，同时加入第5分位与第1分位的差作为多空组合，计算每个不同投资方案的累计净值。具体结果如下：

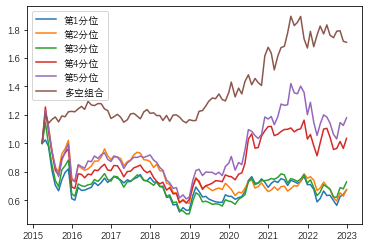


图7. 分层测试法累计净值结果

可以看出：集中度在第4和第5分位（即行业集中度排名在前40%）的行业投资策略的累计净值大幅领先其余的投资组合，产生了明显的分层；且做多第5分位，做空第1分为行业的多空组合取得了更明显的收益，随着时间推移多空组合的净值稳步增加，说明集中度因子较高的行业未来是收益确实会表现得更好。

## 基于集中度因子和相对估值指标的行业轮动策略

### 评价标准

交易策略的评价标准主要从业绩和风险两个方面进行评价，主要有累计收益率、年化收益率、夏普比率、最大回撤。

### 回测结果

下面，我将通过回测模拟的方式，验证该策略能否紧跟市场行情，选择出交易活跃的行业并获得超额收益。选取2019年9月30日至2024年1月19日的区间进行检验，之所以选择这段时间，是因为在这接近五年的时间里同时涵盖了熊市、牛市、震荡市及千股跌停等各种市场行情，能够有效的检验策略的稳定性。

策略具体规则可以概括为以下两点：在每个月的最后一个交易日，计算每个行业的资金集中度因子。当一个行业的资金集中度因子排名在前10且相对估值的排名在后55%至85%时，认为该行业相对被低估，目前处于拥挤交易的泡沫形成阶段，未来上涨的概率较大，因此买入；当该行业相对估值排名在前35%或集中度因子排在10名之后时，认为该行业估值过高或者目前已不属于市场中的热点行业了，横久必跌，未来下跌的概率较大，因此卖出。

以行业指数代表资产价格的收盘价，该策略的累计净值以及相较沪深300和中证1000的表现如下：

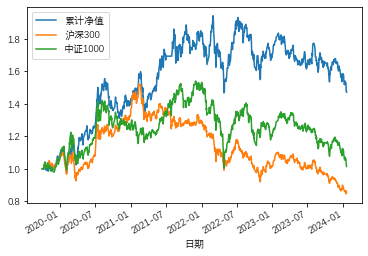


图8. 行业轮动策略的累计净值

可以直观看出，策略在大部分时间内都跑赢了沪深300和中证1000，量化的评价标准如下表所示：

表2. 策略与沪深300和中证1000的业绩指标和风险指标对比

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 累计收益率 | 年化波动率 | 年化收益率 | 夏普比率 | 最大回撤 |
| 行业轮动策略 | 47.10% | 22.40% | 9.75% | 0.345 | 24.42% |
| 沪深300 | -14.28 | 18.74% | -3.65% | -0.301 | 44.40% |
| 中证1000 | 1.42% | 21.98% | 0.34% | -0.83 | 34.62% |

行业轮动策略的最大回撤低于沪深300和中证1000，在2022年4月、2022年10月的恐慌期和2023年7月以后的熊市中下跌幅度较少。由上表可以看出，该行业轮动策略的所有风险指标均优于其他两种指数，总体表现非常不错。

表3. 策略在每个月最后一个交易日选出的行业

|  |  |
| --- | --- |
| 日期 | 行业 |
| 2019/9/30 | 汽车 |
| 2019/10/31 | 汽车 |
| 2019/11/29 | 通信, 非银金融, 交通运输 |
| 2019/12/31 | 交通运输, 通信, 国防军工, 非银金融 |
| 2020/1/23 | 房地产, 交通运输, 通信, 非银金融 |
| 2020/2/28 | 农林牧渔, 交通运输 |
| 2020/3/31 | 国防军工 |
| 2020/4/30 | 国防军工 |
| 2020/5/29 | 农林牧渔, 汽车 |
| 2020/6/30 | 农林牧渔, 汽车 |
| 2020/7/31 | 农林牧渔, 非银金融 |
| 2020/8/31 | 农林牧渔, 非银金融 |
| 2020/9/30 | 农林牧渔, 非银金融 |
| 2020/10/30 | 农林牧渔, 非银金融, 银行 |
| 2020/11/30 | 农林牧渔, 非银金融, 银行, 医药生物 |
| 2020/12/31 | 非银金融, 银行, 医药生物 |
| 2021/1/29 | 农林牧渔, 医药生物 |
| 2021/2/26 | 医药生物 |
| 2021/3/31 | 国防军工, 医药生物 |
| 2021/4/30 | 国防军工 |
| 2021/5/31 | 国防军工 |
| 2021/7/30 | 社会服务, 食品饮料 |
| 2021/8/31 | 社会服务, 食品饮料, 医药生物 |
| 2021/9/30 | 社会服务, 食品饮料, 医药生物 |
| 2021/10/29 | 社会服务, 食品饮料, 医药生物 |
| 2021/11/30 | 建筑材料, 食品饮料, 医药生物 |
| 2021/12/31 | 食品饮料 |
| 2022/1/28 | 有色金属 |
| 2022/2/28 | 汽车 |
| 2022/3/31 | 纺织服饰, 汽车 |
| 2022/4/29 | 纺织服饰, 电力设备 |
| 2022/5/31 | 纺织服饰, 电力设备 |
| 2022/6/30 | 纺织服饰, 电力设备 |
| 2022/7/29 | 纺织服饰, 建筑材料 |
| 2022/8/31 | 纺织服饰, 电力设备, 建筑材料 |
| 2022/9/30 | 纺织服饰, 电力设备, 建筑材料 |
| 2022/10/31 | 纺织服饰, 电力设备, 建筑材料 |
| 2022/11/30 | 纺织服饰, 电力设备, 建筑材料, 银行 |
| 2022/12/30 | 建筑装饰, 电力设备, 建筑材料, 银行, 汽车 |
| 2023/1/31 | 电力设备, 建筑材料, 交通运输, 银行 |
| 2023/2/28 | 银行, 房地产, 建筑材料 |
| 2023/3/31 | 建筑材料, 房地产, 非银金融, 银行 |
| 2023/4/28 | 电子, 建筑材料 |
| 2023/5/31 | 建筑材料, 商贸零售, 电子 |
| 2023/6/30 | 电子, 非银金融, 建筑材料 |
| 2023/7/31 | 建筑材料, 电子 |
| 2023/8/31 | 建筑装饰, 商贸零售, 电子 |
| 2023/9/28 | 建筑装饰, 电子, 商贸零售 |
| 2023/10/31 | 建筑装饰, 机械设备, 商贸零售 |
| 2023/11/30 | 建筑装饰, 机械设备, 房地产, 商贸零售 |
| 2023/12/29 | 建筑装饰, 国防军工, 商贸零售 |

# 基于随机森林的多因子模型构建

在上一章中，我们构建了一个可用的基于拥挤交易的行业轮动投资策略，对比沪深300和中证1000都取得了较明显的超额收益。如果能在涨势好的行业中，选出所谓“股性好”的龙头，收益率应该还会有进一步的提升。因此，本文拟借助于随机森林算法，根据股票的技术指标，预测股票未来上涨的概率。

## 数据源及因子提取

选取了28个申万一级行业的成分股在2019年1月至2024年1月五年间1215个交易日的开盘价、收盘价、最高价、最低价、涨跌幅、成交量的数据，剔除2019年1月前未上市的公司，共计3228支股票3650868行数据。股票数据按行业分类，在每个行业内，一只股票未来1个月内（21个交易日）的收益若排在前20%，则作为正例，标签设为“1”；收益若排在后20%，则作为负例，标签设为“-1”；收益在中间60%的股票标签为“0”。

TA-Lib 库是量化分析中的有力武器，涵盖了重叠指标，动量指标，交易量指标，周期指标，价格变换，波动率指标，模式识别，统计函数等多个范围的200多个因子。本文将使用TA-Lib库，根据股票的开盘价、收盘价、最高价、最低价、涨跌幅、成交量等基础数据，获得其中50个常用的特征因子。原始的特征因子列表如下：

表4. 本文所用到的技术因子及中文名

|  |  |  |
| --- | --- | --- |
| **类别** | **因子** | **中文名** |
| 重叠指标 | EMA | 指数平均数 |
| DEMA | 双移动平均线 |
| WMA | 加权移动平均法 |
| KAMA | 考夫曼自适应移动平均线 |
| MA | 移动平均线 |
| SAR | 抛物线指标 |
| SMA | 简单移动平均线 |
| T3 | 三重指数移动平均线 |
| 动量指标 | ADX | 平均趋向指数 |
| CMO | 钱德动量摆动指标 |
| DX | 动向指标或趋向指标 |
| MFI | 资金流量指标 |
| MOM | 动量 |
| PPO | 价格震荡百分比指数 |
| RSI | 相对强弱指数 |
| ULTOSC | 终极波动指标 |
| WILLR | 威廉指标 |
| CCI | 顺势指标 |
| ROC | 变更率指标 |
| BOP | 均势指标 |
| 成交量指标 | AD | 累积/派发线 |
| OBV | 能量潮 |
| ADOSC | 震荡指标 |
| 波动率指标 | ATR | 真实波动幅度均值 |
| NATR\_14 | 归一化波动幅度均值 |
| TRANGE | 真正的范围 |
| 价格指标 | AVGPRICE | 平均价格 |
| MEDPRICE | 中位数价格 |
| TYPPRICE | 表明性价格 |
| WCLPRICE | 加权收盘价 |
| 周期指标 | HT\_DCPHASE | 希尔伯特变换-主导循环阶段 |
| HT\_DCPERIOD | 希尔伯特变换-主导周期 |
| HT\_TRENDMODE | 希尔伯特变换-趋势与周期模式 |
| 统计类指标 | LINEARREG\_SLOPE | 线性回归斜率 |
| BETA | 贝塔系数 |
| LINEARREG | 线性回归 |
| CORREL | 皮尔逊相关系数 |
| LINEARREG\_ANGLE | 线性回归角度 |
| LINEARREG\_INTERCEPT | 线性回归截距 |
| STDDEV | 标准偏差 |
| TSF | 时间序列预测 |
| VAR | 方差 |
| 数学方法 | ADD | 向量加法运算 |
| DIV | 向量除法运算 |
| SUM | 周期内求和 |
| 数学变换 | COS | 反余弦函数 |
| SIN | 反正弦函数 |
| TAN | 反正切值函数 |

## 因子筛选

### 相关系数筛选

之前选取的因子中存在较多同质因子，举例来说，EMA指数平均数、DEMA双移动平均线、WMA移动加权平均、KAMA考夫曼的自适应移动平均线、MA移动平均线等移动平均线指标，观测周期重复且计算方法类似，可能会造成严重的自相关问题。因此需要进行相关性检验剔除同质性因子。

下面以计算机行业为例，介绍因子的筛选过程：首先因子间的 pearson相关系数，输出相关系数矩阵，绘制热力图以直观展示相关关系，如下图所示。

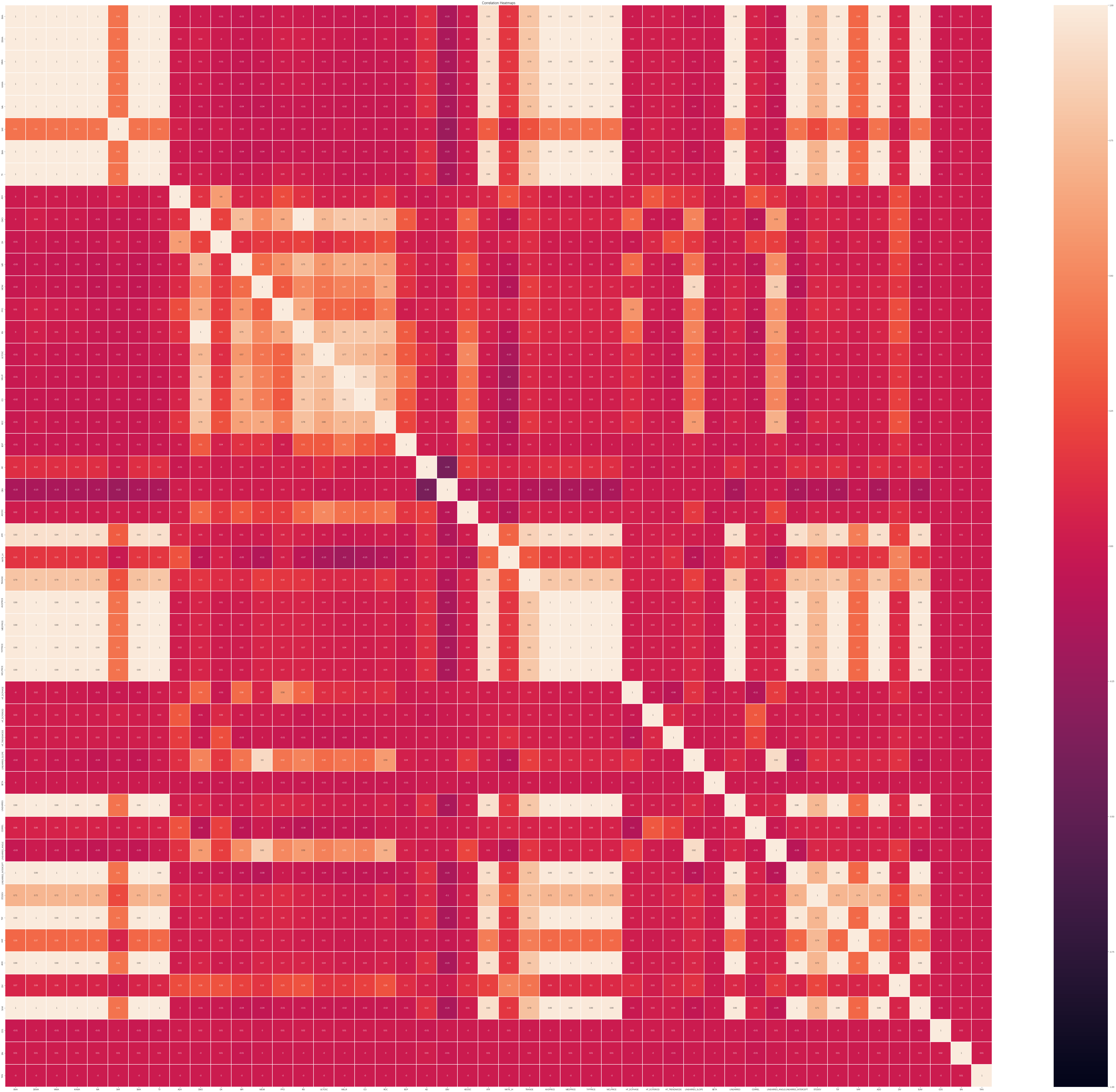


图9. 原始因子集的热力图

可以看出，几大移动平均线类的特征因子相关性接近1，且与其他的因子相关性几乎没有区别，因此这些因子中保留一个即可。经计算，EMA与其他因子的总相关性最低，也就是说在移动平均数这一系列因子中EMA与其他因子有最大的区分度，因此在这些特征因子中保留EMA。

对于其他特征因子都进行类似的筛选，经过处理后，还剩35个因子，这些因子之间相关性较低，

筛选后的相关系数热力图如下：

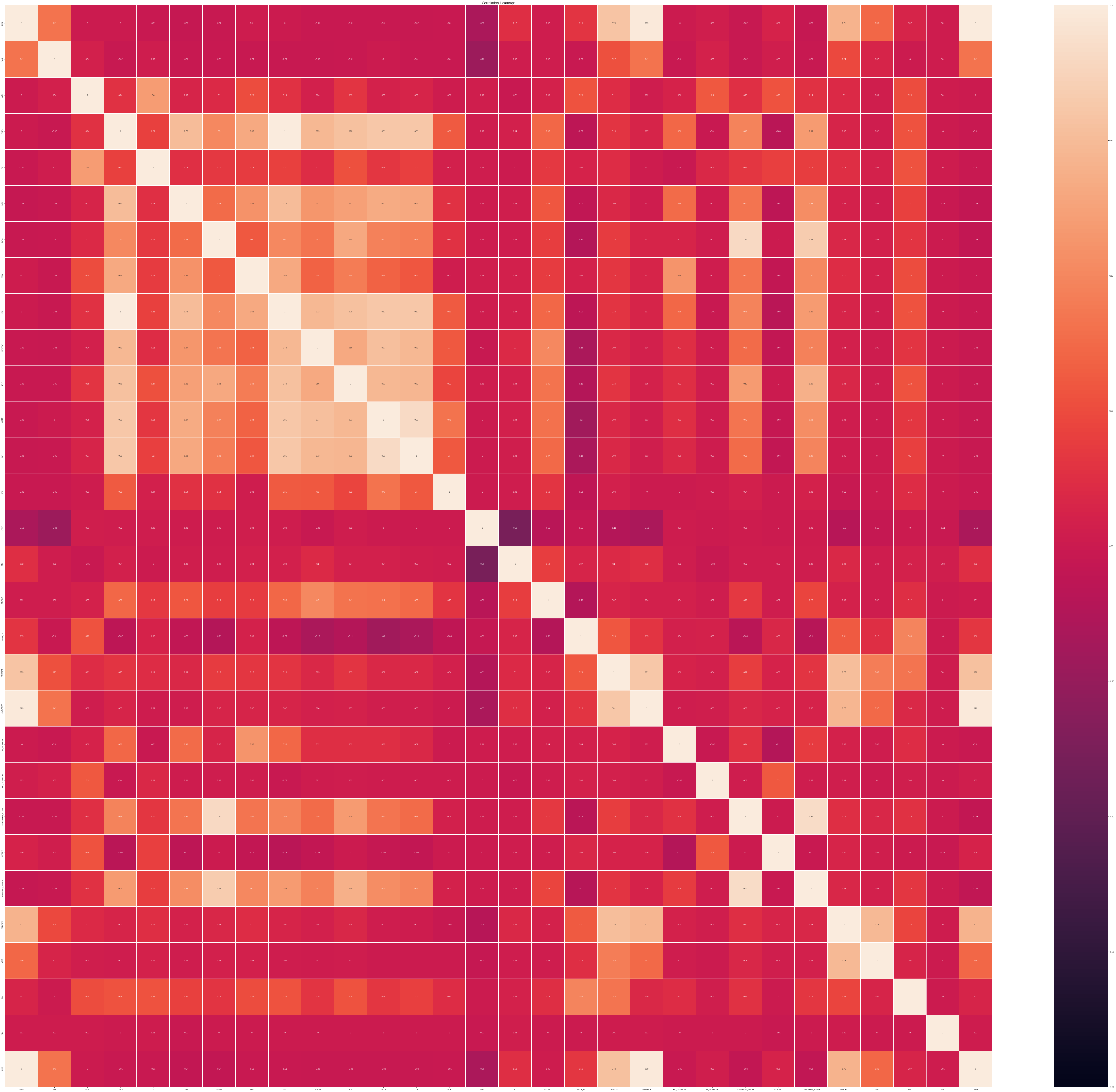


图10. 筛选后的因子热力图

在简单地使用相关系数矩阵剔除同质度较高的因子后，下面使用LightGBM进行特征因子重要性评分。

### LightGBM特征因子评分

LightGBM（Light Gradient Boosting Machine）是一个基于梯度提升决策树框架的高效、快速的机器学习算法，主要针对大规模数据集和高维特征的训练进行了优化，适用于处理大规模数据集和高维特征的分类和回归问题。使用LightGBM算法可以以较快的速度训练模型。可以首先使用LightGBM进行与训练，通过feature\_importances方法可以获得特征重要性，筛选出比较重要的特征，以较少的时间消耗反映出更多有关结果的信息，使预测更加迅速。

设置LightGBM 模型的n\_estimators为200，learing\_rate为0.1，random\_state为42，max\_depth为None。训练后调用feature\_importances\_方法获得特征因子的重要性，对重要性降序排序，筛选出重要性排前30的因子，可视化结果如下：

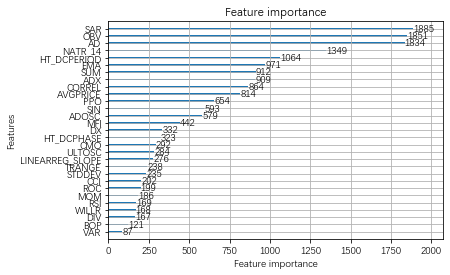


图11. LightGBM模型给出的因子重要性评分

可以看出，SAR、OBV、AD这些反应成交量和动量的指标对收益率有着显著影响，这也符合A股的投资经验。将这些筛选出的特征因子作为后续随机森林模型的输入。

## 训练随机森林模型

对随机森林模型的训练，着重点在调整随机森林模型的参数，主要体现在两个参数上，一是n\_estimators，它决定了随机森林中决策树的个数，如果太小会导致过拟合，而太大则会导致训练所需资源增加，耗时较长，且提升精度的边际效用非常少；二是max\_depth，它限制了决策树的最大深度，该值太小会导致决策链太浅，精度较低，太大同样会导致训练时间过长。

选取“综合”行业2019年9月至2022年9月的因子和收益率数据作为调参所用的数据源，阐述随机森林的调参过程。采用 sklearn 包下的 GridSearchCV 方法进行网格搜索法交叉验证调参，候选参数集n\_estimators设置为[100,150,200,250,300], max\_depth设置为[15,20,25,30,None],调优目标设置为accuracy，进行5折交叉验证。最优参数及调优目标如下

表5. 随机森林较优的一组参数

|  |  |  |
| --- | --- | --- |
| 最佳参数 | | 测试集预测结果 |
| n\_estimators | max\_depth | accuracy |
| 200 | None | 0.720 |

使用这组参数构建随机森林模型，使用训练集的数据进行训练，然后在预测集上检验模型的效果，对比模型的预测值和真实值，结果如下：



图12. 测试集混淆矩阵

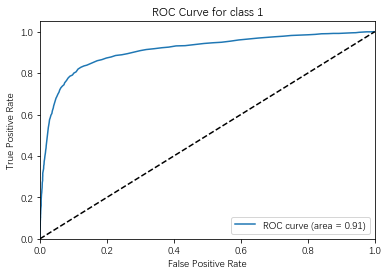


图13. 测试集ROC曲线

表6. 随机森林模型的分类结果

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Precision** | **Recall** | **F1-Score** | **Support** |
| -1 | 0.76 | 0.57 | 0.65 | 1644 |
| 0 | 0.73 | 0.92 | 0.81 | 4114 |
| 1 | 0.87 | 0.51 | 0.64 | 1707 |
| Accuracy |  |  | 0.75 | 7465 |
| Macro Avg | 0.79 | 0.67 | 0.70 | 7465 |
| Weighted Avg | 0.77 | 0.75 | 0.74 | 7465 |

可以看出，模型的识别率较高，能很好地预测次月涨幅在前20%的股票（即分类1）。假阴率较低，表明模型给出的预测正确率较高；假阳率较高，表明还有很多能在次月涨幅排在前20%的股票没有被预测出来，容易踏空。分类为1的ROC\_AUC为0.91，总体的预测能力比较强。

# 行业轮动策略和随机森林选股多因子模型的整体回测

## 策略详细描述

第3章和第4章分别介绍了基于集中度因子的行业轮动策略和随机森林模型，现将这两者结合到一起，在行业轮动策略的基础上，使用随机森林在热门行业中选择上涨概率大的股票进行投资。下面对策略的步骤进行详细的描述：

（1） 在每个月的最后一个交易日，基于集中度和相对估值因子选择行业，具体策略是：当某个行业的资金集中度因子排名在前10且相对估值的排名在后55%至85%时，选择买入该行业的股票；当该行业相对估值排名在前35%或集中度因子排在10名之后时，选择卖出该行业的股票。

（2） 在选择出最后要买入或仍将持有的行业后，使用随机森林模型在这些行业中选股。每个被选中的行业在调仓日这一天，滚动使用过去3个月的数据作为训练集，训练随机森林模型，而后根据每月最后一个交易日的数据对下月股票收益率能否排在前20%做出预测。随机森林模型给出的预测是一个概率值，选择概率在前3%的个股进入投资组合（若某行业股票数量太少，导致3%的个股数量不足2，则选择2支股票），下表展示了滚动学习调仓的示意。注意，每个调仓日，资金在选取的行业间均匀分配，在每个行业中的股票均匀分配。

图14. 滚动学习示意图（假设训练集窗口为3个月）

## 模型回测与评价

同样选取2019年9月30日至2024年1月10日作为回测区间，在行业轮动策略基础上的随机森林选股策略表现如下：



图15. 策略累计净值曲线

从图中可以直观看出，行业轮动＋随机森林的强化选股模型收益率大幅超越沪深300、中证1000和单纯的行业轮动模型。更详细的指标如下：

表7. 策略间业绩指标和风险指标对比

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 累计收益率 | 年化波动率 | 年化收益率 | 夏普比率 | 最大回撤 |
| 行业轮动+随机森林 | 161.11% | 26.44% | 26.21% | 0.924 | 25.44% |
| 行业轮动策略 | 47.10% | 22.40% | 9.75% | 0.345 | 24.42% |
| 沪深300 | -14.28 | 18.74% | -3.65% | -0.301 | 44.40% |
| 中证1000 | 1.42% | 21.98% | 0.34% | -0.83 | 34.62% |

从回测表现看，行业轮动＋随机森林的强化选股模型取得了最佳的策略表现：获得了161.11%的累计收益率的同时，年化波动率26.44%，使得夏普比率为0.924，大幅高于其他策略，在控制风险相同的情况下获得了最高的策略收益。最大回撤略高于行业轮动策略，但考虑到卓越的累计收益率，最大回撤稍高也可以接受。

相较于以行业指数为投资标的的行业轮动策略，经过随机森林模型进行选股优化的净值曲线有两大特点：

其一，由于随机森林的优化指标是尽可能寻找在行业内收益率排名靠前的个股，因此所选股票的收益大概率高于行业总体的收益率。在大盘反弹的时段，随机森林优化模型反弹幅度更大，在市场恐慌时，有价值的股票被相对低估，一旦市场从恐慌情绪中回复，行业内的优质股票就会以强劲的势头率先上涨，带来较大的超额收益；

其二，策略在熊市行情下相对抗跌。2023年经济复苏的预期转变后，基本面走弱的趋势彻底暴露了出来，A股市场进入了漫长的下跌区间，市场行情逐渐变得混乱。沪深300、中证1000、中证2000均出现了不同程度的下跌，在7月份之后这一趋势更加严重。但即使是在这样的行情下，该策略仍能保持较好的稳健性，甚至能在下跌中继中逆势上涨，展现了其良好的对抗风险的能力。

表8. 随机森林模型在每月最后一个交易日选出的股票代码

|  |  |
| --- | --- |
| 日期 | 股票代码 |
| 2019/9/30 | 002265.SZ, 603767.SH, 603776.SH, 601965.SH, 300745.SZ |
| 2019/10/31 | 300745.SZ, 603023.SH, 603178.SH, 002906.SZ, 603730.SH |
| 2019/11/29 | 600095.SH, 600621.SH, 300571.SZ, 300308.SZ, 603885.SH, 601919.SH, 300350.SZ |
| 2019/12/31 | 600834.SH, 000828.SZ, 600179.SH, 300308.SZ, 300571.SZ, 300699.SZ, 002446.SZ, 600643.SH, 000987.SZ |
| 2020/1/23 | 000429.SZ, 300350.SZ, 600179.SH, 600383.SH, 600743.SH, 000809.SZ, 300571.SZ, 300308.SZ, 600599.SH, 601162.SH |
| 2020/2/28 | 000582.SZ, 601872.SH, 600153.SH, 002852.SZ, 002772.SZ |
| 2020/3/31 | 688287.SH, 300177.SZ |
| 2020/4/30 | 002214.SZ, 300600.SZ |
| 2020/5/29 | 300652.SZ, 600698.SH, 600297.SH, 000572.SZ, 600653.SH, 000713.SZ, 002311.SZ |
| 2020/6/30 | 603596.SH, 300695.SZ, 603158.SH, 603006.SH, 603701.SH, 603566.SH, 300119.SZ |
| 2020/7/31 | 002673.SZ, 600053.SH, 002481.SZ, 300498.SZ |
| 2020/8/31 | 601066.SH, 601628.SH, 002311.SZ, 603477.SH |
| 2020/9/30 | 002647.SZ, 600816.SH, 600975.SH, 002548.SZ |
| 2020/10/30 | 600036.SH, 000001.SZ, 601628.SH, 601336.SH, 300087.SZ, 601952.SH |
| 2020/11/30 | 000001.SZ, 600926.SH, 601066.SH, 601601.SH, 000663.SZ, 600354.SH, 603222.SH, 000597.SZ, 002622.SZ, 603127.SH, 300194.SZ, 002821.SZ, 300725.SZ, 603233.SH |
| 2020/12/31 | 601318.SH, 601099.SH, 601166.SH, 600926.SH, 603368.SH, 300318.SZ, 000813.SZ, 300583.SZ, 603520.SH, 600211.SH, 603939.SH, 002102.SZ |
| 2021/1/29 | 603976.SH, 603259.SH, 000538.SZ, 002102.SZ, 300558.SZ, 603229.SH, 600511.SH, 603222.SH, 600965.SH, 300094.SZ |
| 2021/2/26 | 600763.SH, 600511.SH, 002030.SZ, 300759.SZ, 600161.SH, 603882.SH, 603301.SH, 002821.SZ |
| 2021/3/31 | 000733.SZ, 000638.SZ, 600161.SH, 603883.SH, 300206.SZ, 000710.SZ, 300558.SZ, 600276.SH, 300676.SZ, 002932.SZ |
| 2021/4/30 | 300101.SZ, 002231.SZ |
| 2021/5/31 | 002231.SZ, 300696.SZ |
| 2021/7/30 | 600300.SH, 603779.SH, 002858.SZ, 002638.SZ |
| 2021/8/31 | 600300.SH, 002329.SZ, 002589.SZ, 600200.SH, 000766.SZ, 600535.SH, 002880.SZ, 002172.SZ, 300171.SZ, 600568.SH, 002306.SZ, 002638.SZ |
| 2021/9/30 | 603779.SH, 600300.SH, 000403.SZ, 600645.SH, 000999.SZ, 300677.SZ, 000788.SZ, 002294.SZ, 000590.SZ, 300111.SZ, 603136.SH, 002858.SZ |
| 2021/10/29 | 300106.SZ, 603777.SH, 300725.SZ, 002437.SZ, 300642.SZ, 002399.SZ, 300171.SZ, 300534.SZ, 300111.SZ, 300181.SZ, 002858.SZ, 000978.SZ |
| 2021/11/30 | 600298.SH, 002557.SZ, 000877.SZ, 603737.SH, 300683.SZ, 002287.SZ, 603567.SH, 300702.SZ, 300630.SZ, 603896.SH, 000999.SZ, 300573.SZ |
| 2021/12/31 | 000799.SZ, 000596.SZ |
| 2022/1/28 | 000970.SZ, 600807.SH, 002192.SZ |
| 2022/2/28 | 603348.SH, 300681.SZ, 603596.SH, 301192.SZ, 600653.SH |
| 2022/3/31 | 000901.SZ, 600469.SH, 002553.SZ, 603596.SH, 301192.SZ, 002574.SZ, 002875.SZ |
| 2022/4/29 | 002546.SZ, 603861.SH, 300173.SZ, 000982.SZ, 002630.SZ, 300116.SZ, 002656.SZ, 002486.SZ |
| 2022/5/31 | 600525.SH, 603333.SH, 300173.SZ, 002129.SZ, 002692.SZ, 600468.SH, 002087.SZ, 002721.SZ |
| 2022/6/30 | 002518.SZ, 300450.SZ, 603507.SH, 000821.SZ, 300438.SZ, 603031.SH, 601718.SH, 002656.SZ |
| 2022/7/29 | 002088.SZ, 300093.SZ, 002731.SZ, 002656.SZ |
| 2022/8/31 | 603737.SH, 002791.SZ, 002879.SZ, 603628.SH, 002090.SZ, 603050.SH, 300360.SZ, 603819.SH, 603908.SH, 002485.SZ |
| 2022/9/30 | 600819.SH, 002791.SZ, 601567.SH, 300427.SZ, 002471.SZ, 002706.SZ, 603606.SH, 300360.SZ, 002042.SZ, 603908.SH |
| 2022/10/31 | 603737.SH, 002791.SZ, 002692.SZ, 603829.SH, 600537.SH, 002358.SZ, 300427.SZ, 002709.SZ, 002740.SZ, 603958.SH |
| 2022/11/30 | 002247.SZ, 002088.SZ, 601939.SH, 601988.SH, 002309.SZ, 002471.SZ, 300153.SZ, 300670.SZ, 002865.SZ, 002364.SZ, 002193.SZ, 601339.SH |
| 2022/12/30 | 601818.SH, 601988.SH, 603737.SH, 002694.SZ, 002662.SZ, 002594.SZ, 000980.SZ, 601058.SH, 603319.SH, 600477.SH, 603959.SH, 603909.SH, 300750.SZ, 000049.SZ, 002112.SZ, 300670.SZ, 300432.SZ, 300407.SZ |
| 2023/1/31 | 002271.SZ, 002372.SZ, 000001.SZ, 600036.SH, 601021.SH, 002183.SZ, 603069.SH, 600241.SH, 002169.SZ, 600732.SH, 002692.SZ, 300035.SZ, 601727.SH |
| 2023/2/28 | 002142.SZ, 601128.SH, 601636.SH, 002088.SZ, 600340.SH, 000056.SZ, 000002.SZ |
| 2023/3/31 | 600801.SH, 002088.SZ, 002807.SZ, 002839.SZ, 600503.SH, 600340.SH, 600208.SH, 000627.SZ, 601211.SH |
| 2023/4/28 | 601636.SH, 002613.SZ, 603626.SH, 002655.SZ, 600707.SH, 603297.SH, 002199.SZ, 002371.SZ |
| 2023/5/31 | 000401.SZ, 002233.SZ, 000020.SZ, 002916.SZ, 002841.SZ, 002288.SZ, 300671.SZ, 300566.SZ, 002818.SZ, 002803.SZ |
| 2023/6/30 | 600958.SH, 000532.SZ, 002302.SZ, 000401.SZ, 300398.SZ, 300446.SZ, 000020.SZ, 002388.SZ, 002079.SZ, 300303.SZ |
| 2023/7/31 | 603616.SH, 002694.SZ, 002681.SZ, 603629.SH, 603933.SH, 300650.SZ, 300701.SZ, 300657.SZ |
| 2023/8/31 | 002482.SZ, 603909.SH, 300492.SZ, 603685.SH, 000727.SZ, 603989.SH, 002141.SZ, 002130.SZ, 300301.SZ, 600365.SH, 600828.SH |
| 2023/9/28 | 002251.SZ, 000007.SZ, 002325.SZ, 600610.SH, 300649.SZ, 000413.SZ, 600751.SH, 002141.SZ, 002681.SZ, 600666.SH, 600601.SH |
| 2023/10/31 | 600327.SH, 000882.SZ, 002325.SZ, 300495.SZ, 002663.SZ, 600545.SH, 000425.SZ, 000008.SZ, 300606.SZ, 002021.SZ, 000157.SZ, 300195.SZ, 002786.SZ |
| 2023/11/30 | 600415.SH, 002251.SZ, 600603.SH, 600641.SH, 600007.SH, 300668.SZ, 000010.SZ, 002586.SZ, 300276.SZ, 300606.SZ, 002685.SZ, 601717.SH, 603488.SH, 002009.SZ, 300554.SZ, 600499.SH |
| 2023/12/29 | 603123.SH, 601028.SH, 002200.SZ, 300668.SZ, 600853.SH, 601989.SH, 300397.SZ |

# 研究结论与展望

## 研究总结

本文使用Kinlaw(2019)中提出的“资金集中度”和相对估值指标，对中国的拥挤交易现象进行了量化分析，试图构建一个基于行业轮动规律的行业选择模型。在此基础上，使用行业内股票的技术因子，利用随机森林模型选择行业内表现相对强劲的股票，以期获得超额收益。

全文以28个申万一级行业作为研究对象，以这些行业的指数、市值、市净率等指标，参考Kinlaw(2019)，构建了适合中国市场行情的“资金集中度”因子。通过IC值分析和分层测试法，证明了资金集中度因子对未来收益的有效性。并结合“相对估值”指标，构建行业轮动策略，在2019年9月至2024年1月的回测周期内证明了该策略相较于沪深300和中证1000的优越性。用回测检验证明了在中国市场进行行业轮动投资的可行性。

在此基础上，使用随机森林算法在行业中选取收益排名靠前的个股。首先对原始数据进行重标签，将未来21个交易日内收益率排名前20%的标记为1，后20%的标记为-1，中间60%标记为0。因子集选取TA-Lib量化因子库中的50个因子，经过相关系数和LightGBM筛选后选择出重要性最强的因子，用以避免多重共线性问题，缩减因子规模减少训练时间。使用选取的因子在指定训练集上训练，借助GridSearchCV 方法进行网格搜索法交叉验证调参，选择出随机森林模型的最佳参数。最后，同样在2019年9月至2024年1月的周期内进行回测验证，获得了年化收益率26.21%，夏普比率0.924的理想结果，同时策略在2023年至今的熊市内保持了较小的回撤，表明策略具有较好的稳健性。

本文的研究结果在证明了A股市场上存在着拥挤交易现象，投资者可以借此在拥挤交易的泡沫产生的初期获利。同时这也给监管部门一个衡量风险的指标，尽量减少争相退出交易时造成的踩踏。

## 未来展望

在A股市场中，对概念题材的炒作经久不衰，由于A股长期缺乏个股的做空机制，一些基本面不良的股票因种种原因获得炒作后会获得很高的估值。参考东方财富的统计数据，目前在仅仅5000余家上市公司中就已经出现了400多个概念题材，未来随着热点事件的继续发生概念题材还有继续增加的趋势。受个人的精力和硬件设备的限制，本人无法在完成本篇论文的时间内对如此多个概念板块进行分门别类的分析验证。但不得不承认的是：在概念题材中还有很多超额收益的挖掘空间，是未来值得研究的对象。

2024年初，A股市场经历了股灾的洗礼，若没有国家队的救市行动，A股的最低点或将进一步突破大家的预期。在股灾这种特殊的市场状况下，投资者最好的办法就是及时割肉止损，不要妄言抄底，以所谓“网格交易”摊薄成本，而应该耐心等待强烈的底部信号出现。这给策略的优化提供了一个方向：即加入极端行情的识别信号，当股灾出现时全部清仓等待反转。至于极端行情的识别，则需要全面参考市场的各种数据，包括交易数据和情绪数据等，需要根据投资者的经验进行判断。

最后，虽然该投资策略在过去5年，尤其是在2023年取得了相对不错的收益，但我仍不建议投资者在A股市场中火中取栗。截至本人完成本篇论文时，在“神秘资金”的拉动和政策激励下，上证指数从2635的低点重新站上3000点，实现了八连阳。俗话说“一阳改三观”，经过本次拉升，许多投资者又恢复了乐观的心态，幻想着牛市的到来。但是需要思考的是，如果牛市真的会来，那么这轮牛市的基础是什么？难道是跌到2600就能把筹码洗干净吗？在房地产行业持续下行的今天，还有什么行业能够代替房地产“撑起A股的脊梁”？目前来看，房地产下行的趋势并没有得到改变，市场依旧缺乏信心，外加舆论保持着很大的压力，房地产行业复苏的概率较低。我们或许可以期待结构性的行情，但是全面的牛市很难到来。目前市场追逐出海和高股息的股票，也在一定程度上反映了市场对未来经济形势的看法。另外，为了应对本轮股灾，监管层采取限制做空等方式强行扭转多空双方的力量，这种做法固然能在一定程度上起到刺激效果，但在长期内确实对正常市场机制的一种破坏，未来或许会导致成交量的萎缩。

投资有风险，入市须谨慎，建议投资者根据自身的风险承受能力妥善安排资金。

## 参考文献

[1] 黄兴旺,胡四修,郭军.中国股票市场的二因素模型[J].当代经济科学,2002(05):50-57+95.

[2] 王涛. Fama-French三因子模型及其添加市盈率因子模型在中国股市的适用性研究[D].西南财经大学,2012.

[3] 吴荻. 中国股票市场选股模型实证分析[D].复旦大学,2011.

[4] 李倩倩. H企业供应商绩效管理优化研究[D].河南大学,2019.

[5] 何路. 多因子量化选股及投资者情绪择时策略的实证检验[D]. 南京大学, 2020.

[6] 刘微,罗林开,王华珍.基于随机森林的基金重仓股预测[J].福州大学学报(自然科学版),2008(S1):134-139.

[7] 雍凯. 随机森林的特征选择和模型优化算法研究[D]. 哈尔滨工业大学, 2008

[8] 方匡南,朱建平,谢邦昌.基于随机森林方法的基金收益率方向预测与交易策略研究[J].经济经纬,2010(02):61-65.DOI:10.15931/j.cnki.1006-1096.2010.02.032.

[9] 曹正凤,纪宏,谢邦昌.使用随机森林算法实现优质股票的选择[J].首都经济贸易大学学报,2014,16(02):21-27.DOI:10.13504/j.cnki.issn1008-2700.2014.02.003.

[10] 罗超.面向高维数据的随机森林算法优化探讨[J].商,2016(04):207+199.

[11] 刘凯. 随机森林自适应特征选择和参数优化算法研究[D]. 长春工业大学2018.

[12] 李东,贾郭军. 基于遗传算法的随机森林算法优化研究[J]. 首都师范大学学报(自然科学版), 2018, 39(4): 15-21.

[13] 李杰. 基于随机森林算法的多因子选股模型研究[D].哈尔滨工业大学,2019.

[14] 延昭. 基于随机森林方法进行量化投资交易的实证研究[D].西南财经大学,2019.

[15] 周亮. 基于随机森林模型的股票多因子投资研究[J]. 金融理论与实践, 2021(7): 97-103

[16] 宗淑杰. 基于行业轮动的证券投资组合选择[D].天津财经大学,2017.

[17] 武文超.中国A股市场的行业轮动现象分析——基于动量和反转交易策略的检验[J].金融理论与实践,2014(09):111-114.

[18] 黄前前.A股不同市场状态下动量策略盈利性研究[J].金融理论探索,2016(05):48-53.DOI:10.16620/j.cnki.jrjy.2016.05.007.

[19] 周丽云. 投资者情绪、投资者拥挤交易行为与资产定价研究[D].华南理工大学,2017.

[20] 王婷婷. 基于拥挤交易的行业轮动投资交易策略[D].上海师范大学.

[21] Markowitz, H. (1952), PORTFOLIO SELECTION\*. The Journal of Finance, 7: 77-91.

[22] Toward a Theory of Market Value of Risky Assets. In Treynor on Institutional Investing, J.L. Treynor (Ed.)

[23] Sharpe, W.F. (1964) Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk. Journal of Finance, 19, 425-442.

[24] Paul A. Samuelson The Review of Economics and Statistics, Vol. 36, No. 4. (Nov., 1954),

[25] The Pricing of Options and Corporate Liabilities. Fischer Black and Myron Scholes. Journal of Political Economy, 1973

[26] JOURNAL OF ECONOMIC THEORY 13, 341-360 (1976) The Arbitrage Theory of Capital Asset Pricing STEPHEN A. ROSS\*

[27] Fama and French (1993) Three-Factor Model: Evidence from Conventional and Shariah-Compliant Portfolios in Bursa Malaysia

[28] Carhart, M.M. (1997), On Persistence in Mutual Fund Performance. The Journal of Finance

[29] Liew, J., & Vassalou, M. (2000). Can Book-to-Market, Size and Momentum Be Risk Factors That Predict Economic Growth The Journal of Finance

[30] Sabehi G, Loy A, Jung KH, Partha R, Spudich JL, et al. (2005) New insights into metabolic properties of marine bacteria encoding proteorhodopsins.

[31] Joseph, J. (2000), A Realist Theory of Hegemony. Journal for the Theory of Social Behaviour

[32] Fama E F, French K R. A five-factor asset pricing model[J]. Journal of FinancialEconomics, 2015, 116(1): 1-22.

[33] Sorensen, J.B. And Stuart, T.E. (2000) Aging, Obsolescence and Organizational Innovation. Administrative Science Quarterly, 45, 81-112

[34] Breiman, L. (2001) Random Forests. Machine Learning, 45, 5-32.

[35] Kumar, S. and Phrommathed, P. (2005) Research Methodology. In Kumar, S. and Phrommathed, P., Eds., New Product Development, Springer, Berlin, 43-50

[36] Genuer R., Poggi J.-M. and Tuleau C. Variable selection using random forests. To appear in Pattern Recognition Letters (2010)

[37] Piotr Ladyzynski, Przemyslaw Grzegorzewski:Particle swarm intelligence tunning of fuzzy geometric protoforms for price patterns recognition and stock trading. Expert Syst. Appl. 40(7): 2391-2397 (2013)

[38] Manojlović, Teo and Ivan Štajduhar. “Predicting stock market trends using random forests: A sample of the Zagreb stock exchange.” 2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO) (2015): 1189-1193.

[39] Abraham Rebecca,Samad Mahmoud El,Bakhach Amer M., et al. Forecasting a Stock Trend Using Genetic Algorithm and Random Forest[J]. Journal of Risk and Financial Management, 2022,15(5):188.

[40] Behr Merle,Wang Yu,Li Xiao, et al. Provable Boolean interaction recovery from tree ensemble obtained via random forests.[J]. Proceedings of the National Academy of Sciences of the United States of America, 2022, 119(22):119

[41] GEORGE, T.J. and HWANG, C.-Y. (2004), The 52-Week High and Momentum Investing. The Journal of Finance, 59: 2145-2176.

[42] STEIN, J.C. (2009), Presidential Address: Sophisticated Investors and Market Efficiency. The Journal of Finance, 64: 1517-1548.

[43] Menkveld, Albert J. et al. “Shades of Darkness: A Pecking Order of Trading Venues.” Capital Markets: Market Microstructure eJournal (2016): n. pag.

[44] Kinlaw, W., Kritzman, M., & Turkington, D. (2019). Crowded Trades: Implications for Sector Rotation and Factor Timing. The Journal of Portfolio Management, 45(5)