如何基于“集成学习”技术优化多因子复合框架?

中银量化多因子选股系列（四）

本报告基于已发布的中银量化多因子报告中所介绍的基本面财报因子、景气度因子和反转类因子，重点聚焦多因子复合的方式方法，并尝试将集成学习的思想和方法应用于因子复合。实证显示基于改良的 Adaboost 框架得到的复合多因子在多头超额、多空超额等方面显著优于传统等权复合法。

- **传统多因子等权排名复合与等权 z-score 复合效果差异并不显著。**若以多头超额为标准，等权排名（rank）的回测效果略好。
- **集成学习本质是通过构建并结合多个基学习器来完成任务。**对于训练集数据，通过训练若干个个体弱学习器，结合一定的策略就可以最终形成一个强学习器，以达到博采众长的目的。本报告在传统“Adaboost + CART 回归分类器”的模型基础上进行了升级，力求提升因子复合有效性。
- **以“AdaBoost Regressor + 基学习器 Linear Regression 为模型，最优化“因子下一期排名分位数预测值”为目标构建多因子复合框架。**其本质是利用子类因子历史的数据，即最优化预测中证 500 成分股的下期 IC 值，且希望排名越靠前（分位数越接近 1 越优秀）的股票对应的预测结果越准确。我们对样本权重进行了调整以改良多头端超收益，同时对多个参数如因子信息的时间跨度、学习率、弱学习器个数等参数进行了稳健性测试。实证表明，因子的时间序列信息超过 1 年即可实现相对稳健的预测效果、模型对其他参数的敏感性不强。
- **模型表现：**AdaBoost 最终构建的多因子模型多头年化超额 19.2%（样本内）、16.9%（样本外）、多空超额 32.2%（样本内）、23.9%（样本外）、月度中性化后 IC 均值 8.1%，较原始等 Rank 因子复合效果更优。
- **风险提示。**投资者需注意模型失效的风险。

相关研究报告

《中银量化多因子选股系列（一）：基本面财报因子的构建框架初探》20220830
《中银量化多因子选股系列（二）：如何构建盈利景气度因子选股？》20220830
《中银量化多因子选股系列（三）：锚定反转因子构建与增强》20220902

中银国际证券股份有限公司
具备证券投资咨询业务资格

研究领域：金融工程研究

证券分析师：郭策
(8610)66229239
ce.guo@bocichina.com
证券投资咨询业务证书编号：S1300522080002

目录

图表目录

一、单因子策略回顾：财报类因子策略

在中银证券2022年8月30日发布的报告《中银量化多因子选股系列（一）：基本面财报因子的构建框架》中我们以中证500指数增强为例，重点梳理了财报类因子构建的三大核心思想，对传统财报类因子进行批量改进，改进后的因子效果显著增强。

核心思想一：构建“不易操纵”的财报指标

基于企业业务类型，构建更“纯粹”、不易操控的报表指标。企业的业务类型主要包括生产经营、投资、融资以及其他非主营业务。我们可根据业务类型，对三大报表的项目进行重归类，以侧重反映企业各类业务（尤其是生产经营活动）的资产负债、盈利与现金流情况。（具体测算明细请详见报告的附录部分：“基于公司业务的报表项目重新分类”）

核心思想二：以“跨报表验证”思路构建指标

通过观察大量财报评价指标，我们发现这些指标的设计底层原理可总结为基于“跨报表交叉验证法”的方式评估企业财报的基本情况。按不同指标的构建类型，我们可以总结为如下三大类核心要点：

1）**企业资源能力是否与市场盈利脱节？**典型如：ROE、ROA、核心利润/经营资产、存货周转率等

2）**营业收入是否与核心利润脱节？**典型如：主营业务收入/营业收入等；

3）**企业利润是否与经营现金流量脱节？**典型如：经营现金流净额/归母净利润额等。

核心思想三：结合基本面与A股市场定价

传统“市值+行业”中性化没有考虑A股市场对该公司的估值影响。同一行业、市值相似、盈利能力也相似的两个公司，估值较低的公司较估值较高的公司存在低估的可能。借鉴PB-ROE分析框架，可以引入B/P因子对股票进行中性化，具体公式如下。

$$f_{i,t} = \beta_{1,t}\ln(fmv_{i,t}) + \sum_{i=1}^{N}\beta_{j,t}\,Industry_{j,i,t} + \beta_{k,t}\left(\frac{B}{P}\right)_{i,t} + \varepsilon_{i,t}$$

我们将传统财报类因子进行了上述改进后，通过分层聚类优选，共有7个财报单因子入选，主观逻辑上可分为ROA变形、现金与资产匹配、盈利增长以及企业战略扩张四大类，其中4个因子涉及b/p中性化处理（分层聚类优选具体方法请见附录，下文同）。

图表1.财报类经分层聚类优选确定的7个单因子列表及表现

财报因子：聚类精选	bp中性化	类型	IC	多头G1超额样本内	多头G1超额样本外	空头G10超额样本内	空头G10超额样本外	多空G1/G10超额样本内	多空G1/G10超额样本外
ROA_核心利润_张新民_TTM	√	ROA变形	3.8%	9.7%	2.6%	-3.2%	-5.1%	12.9%	7.7%
ROA_核心利润_张新民_TTM		ROA变形	3.1%	9.4%	-1.2%	0.7%	-0.3%	8.8%	-0.9%
ROTC_毛利润比有形资产_TTM		ROA变形	2.3%	7.5%	2.7%	1.3%	-4.3%	6.2%	7.0%
总负债_YoY	√	企业扩张	2.2%	6.5%	7.3%	-2.5%	9.8%	9.0%	-2.5%
净经营现金流比总资产_TTM	√	现金资产匹配	3.2%	8.0%	3.9%	-4.2%	2.6%	12.2%	1.3%
核心利润_张新民_MRQ_YoY		盈利增长	2.4%	7.9%	2.7%	-1.8%	6.8%	9.8%	-4.1%
EBIT_MRQ年变化比营业收入_TTM	√	盈利增长	3.4%	5.5%	9.9%	-3.8%	-3.7%	9.4%	13.6%

注：IC为月度因子Rank IC，业绩比较基准为中证500指数，样本内区间为2010-2020，样本外区间为2021-2022.07

资料来源：万得，中银证券

二、单因子策略回顾：盈利预期类因子策略

中银证券2022年8月30日发布的报告《中银量化多因子选股系列（二）：如何构建盈利景气度因子选股》介绍了股票景气度的构建框架。基于分析师的一致预期，可以分别构建三大类盈利预期因子：

一类因子：净利润预期同比与预期ROE相关指标，代表3-5年的赛道概念；

二类因子：景气度变化因子，为一类因子的季度变化，反映各行业的中短期景气度变化；

三类因子：短期情绪因子，为二类因子的月度变化，反映景气度变化的斜率，即加速上升or减速上升，该类因子常用于日间ETF波动交易，在中长期策略中应用效果相对有限。

因子构建方法

因子构建时我们以一类赛道型因子和二类景气度因子为重点进行因子开发，由于第三类情绪因子需要较高频的换仓周期（通常以日间交易为主），故先不在此报告中进行重点测算。一二类景气度影子构建方法如下。

图表2.盈利预期类因子构建方法

分类	因子	计算公式
一类因子	NETPROFIT_F(k)F(k+1)	NETPROFIT_FY(k+1)/NETPROFIT_FY(k) - 1, k = 0,1,2
	NETPROFIT_F(k)_yoy	NETPROFIT_FY(k)/NETPROFIT_FY(k,去年同期) - 1, k = 1,2,3
	ROE_FY(k)	WindRdf数据库原值
二类因子	NETPROFIT_F(k)F(k+1)_d3m	NETPROFIT_F(k)F(k+1) - NETPROFIT_F(k)F(k+1,上季度), k = 0,1,2
	NETPROFIT_F(k)_qoq	NETPROFIT_FY(k)/NETPROFIT_FY(k,上季度) - 1, k = 1,2,3
	ROE_FY(k)_d3m	ROE_FY(k) - ROE_FY(k,上季度), k = 1,2,3

资料来源：万得，中银证券

等权复合的盈利预期大类因子在样本内多头G1的超额收益表现显著逊色于单因子NetProfit_F3_qoq，因此，我们推荐单因子“NetProfit_F3_qoq”作为盈利预期类因子与其他大类因子进行复合。

图表3.盈利预期类经分层聚类优选确定的1个单因子列表及表现

单因子分组年化超额收益	G1	G2	G3	G4	G5	G6	G7	G8	G9	G10	L/S
NETPROFIT_F3_qoq	13.4%	8.0%	6.3%	1.8%	-0.6%	-1.6%	4.0%	4.4%	1.6%	0.0%	13.4%
ROE_FY3-bp-中性化	9.0%	6.4%	7.6%	2.6%	-4.1%	-0.2%	4.2%	2.9%	2.0%	-2.8%	11.8%
NETPROFIT_F2_yoy_d3m-bp-中性化	8.4%	6.0%	6.8%	3.3%	-6.7%	1.1%	5.3%	4.3%	4.8%	-1.2%	9.6%
NETPROFIT_F1_yoy-bp-中性化	8.4%	7.3%	9.5%	4.1%	6.3%	-0.1%	3.0%	4.4%	0.3%	-2.1%	10.5%
NETPROFIT_F2_yoy-bp-中性化	7.6%	10.3%	7.7%	3.4%	4.3%	4.0%	4.7%	3.1%	0.1%	-2.3%	9.9%

注：单因子分组测试区间为2010.01-2022.07，业绩基准为中证500指数

资料来源：万得，中银证券

三、单因子策略回顾：反转类因子策略

在中银证券 2022 年 9 月 2 日发布的报告《中银量化多因子选股系列（三）：锚定反转因子构建与增强》中我们针对传统反转因子于 2016 年中以后长时间失效的问题，提出了具体改进方案：将 A 股投资者普遍存在的“锚定偏误”、“处置效应”等非理性行为纳入反转因子的构建，开发出新的“锚定反转因子”，并通过一系列增强方法，显著提升了“锚定反转因子”的因子表现。

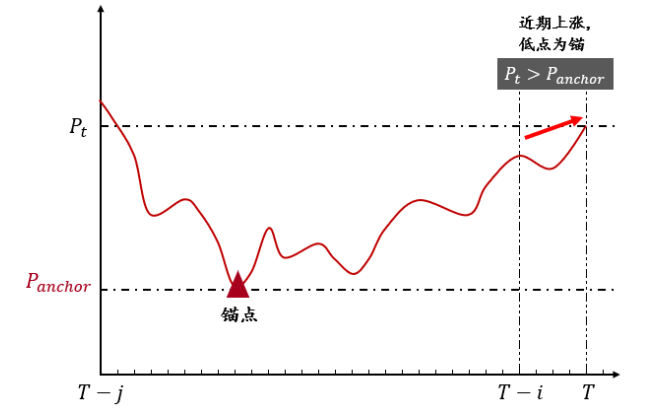
因子构建核心思想与具体构建方法：

- **若股票短期上涨，**反转投资者往往倾向于选择近期涨幅较小的个股。因此**以股票近期低点作为锚点来计算股票的相对涨幅**来衡量个股短期的上涨是否透支了未来的涨幅较为合理。
- **若股票短期下跌，**反转投资者往往倾向于选择近期跌幅较大的个股。因此**以股票近期高点作为锚点来计算股票的相对跌幅**来衡量个股短期的下跌是否能够带来未来更大的上涨空间较为合理。

设定当前时间为T，时间区间$[T-i,\ T]$用于判断股价短期呈现上涨或下跌状态，时间区间$[T-j,\ T]$用于确定股价锚点。若股价短期上涨，即$P_T \geq P_{T-i}$（如下左图），则以时间区间$[T-j,\ T]$的股价低点为锚；若股价短期下跌，即$P_T < P_{T-i}$（如下右图），则以时间区间$[T-j,\ T]$的股价高点为锚，记为P_{ancho} 。因此锚定反转因子的计算公式为：

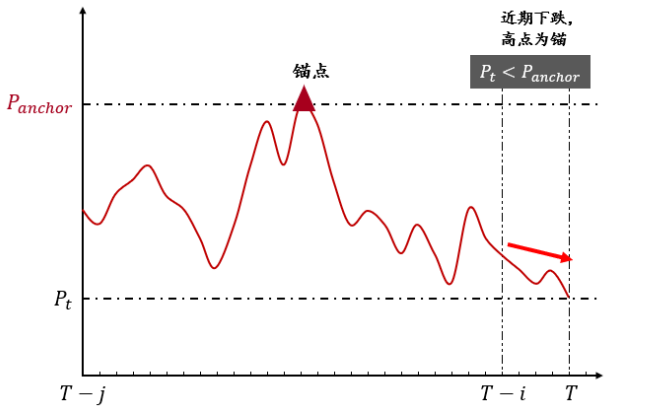
$$\text{锚定反转因子} = \frac{P_t}{P_{anchor}} - 1$$

图表 4. 短期上涨状态个股以近期低点为锚点



资料来源：万得，中银证券

图表 5. 短期下跌状态个股以近期高点为锚点



资料来源：万得，中银证券

增强一：剥离基本面因素（盈利预期与 P/B）。

造成股价超跌的原因可能是投资者情绪过度反应，也可能是股票基本面的恶化。若是投资者短期情绪造成的，股价大概率会在未来一段时间出现均值回复；但若是基本面因素导致的，则股价的下跌可能是趋势性的，而这种趋势性的下跌可能对反转因子的有效性产生不利影响。

具体处理方法:将个股对应的原始因子值对个股对数自由流通市值、中信一级行业以及归母净利润一致预测额（FY1）、B/P 进行 OLS 回归取残差，以剥离基本面（预期）因素对反转因子的干扰，具体公式如下。

$$Reversion(3m)_{vol\ adj_{i,t}} = \beta_{1,t}\ln(fmv_{i,t}) + \sum_{j=1}^{N}\beta_{j,t}\,Industry_{j,i,t} + \beta_{k,t}\,(\frac{B}{P})_{i,t} + \beta_{m,t}\,NetProfit(FY1)_{i,t} + \varepsilon_{i,t}$$

增强二：纳入波动率因素

波动率高的个股在经历下跌后反转效应更剧烈。导致个股波动性不同的原因有很多，除了自身所处行业特征和市值影响外，市场投资者对该股票的关注程度和多空博弈的激烈程度也会对波动性产生影响。我们认为虽然在因子构造时，通常会对行业与自由流通市值进行中性化以剥离行业和市值的影响，但仍未解决因市场关注度不同导致的波动性影响。

对于波动性，我们使用个股在确定锚点时间区间$[T-j,\ T]$的日回报率标准差乘以原始锚定反转因子得到新的因子，公式为：

$$锚定反转因子 = \left(\frac{P_t}{P_{anchor}} - 1\right) * std$$

通过聚类算法精选因子，6 个反转类单因子入选。

图表 6. 反转类经分层聚类优选确定的 6 个单因子列表及表现

反转因子：聚类精选	IC	多头G1超额样本内	多头G1超额样本外	空头G10超额样本内	空头G10超额样本外	多空G1/G10超额样本内	多空G1/G10超额样本外
反转3m-NetProfit_FY1-中性化	3.2%	12.9%	12.6%	-3.7%	1.5%	16.6%	11.0%
反转3m_voladj-NetProfit_FY1-中性化	3.1%	12.6%	14.2%	-5.0%	-0.1%	17.5%	14.3%
锚定反转13_2_w-bp-NetProfit_FY1-中性化	4.5%	12.3%	9.5%	-10.8%	-7.2%	23.1%	16.7%
锚定反转_with_std_13_2_w-bp-NetProfit_FY1-中性化	4.4%	11.6%	6.9%	-11.9%	-7.8%	23.5%	14.7%
反转3m_voladj	3.6%	10.9%	13.0%	-6.6%	2.7%	17.5%	10.3%
反转1m-bp-NetProfit_FY1-中性化	4.4%	10.5%	6.3%	-9.2%	-3.3%	19.8%	9.6%

注：样本内区间为2010-2020，样本外为2021-2022.07

资料来源：万得，中银证券

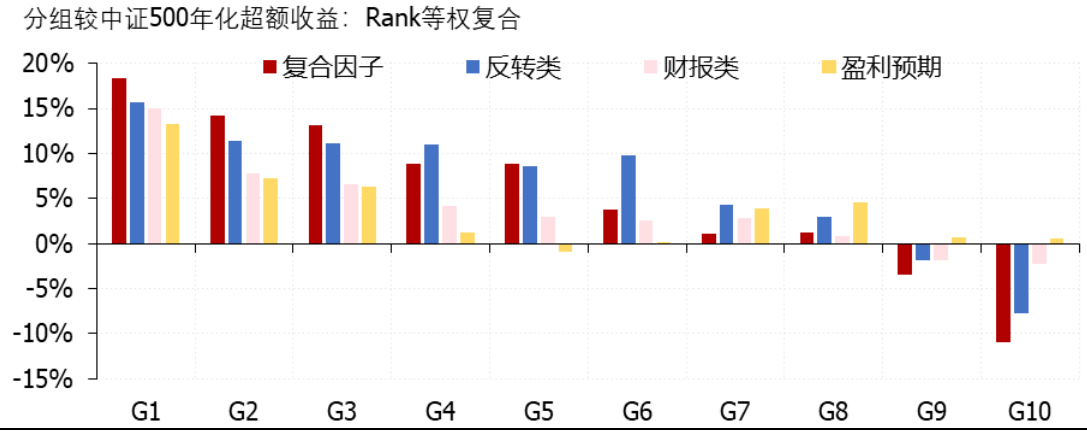
四、 多因子复合框架（传统等权 VS 集成学习）

在传统的多因子复合框架下，通常大家会选择多因子等权复合打分的方法进行多因子复合。我们分别使用 rank 等权复合和 Zscore 等权复合两种方法进行了尝试，结果显示两种方法的差距不大，但复合因子的效果均显著优于单因子效果。

（一）传统大类因子复合：等权复合+周度换仓

复合因子的业绩显著增强。我们将财报基本面将财报基本面因子、盈利预期类因子、技术类因子三个大类因子进行等权复合，得到的大类复合因子表现显著提升。

图表 7.传统 rank 等权复合后，多因子各组超额收益较单因子显著改善



资料来源：万得，中银证券

图表 8.各大类因子分别使用等权 zscore、等权 rank 方式复合效果对比

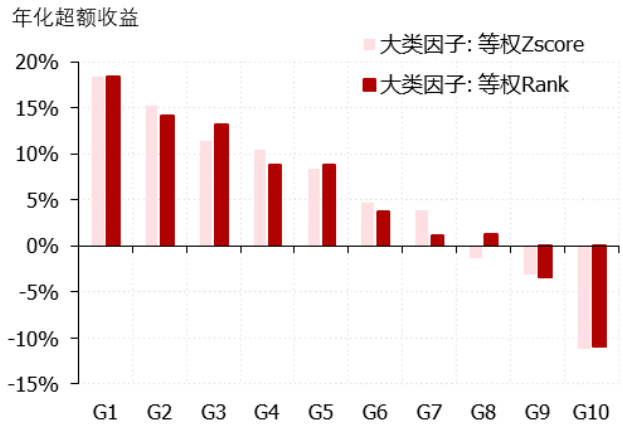
复合因子	IC	G1	G2	G3	G4	G5	G6	G7	G8	G9	G10	L/S
大类因子：等权Zscore	7.5%	18.4%	15.3%	11.4%	10.5%	8.4%	4.7%	3.8%	-1.3%	-3.1%	-11.1%	29.5%
大类因子：等权Rank	7.4%	18.4%	14.2%	13.1%	8.8%	8.9%	3.8%	1.1%	1.3%	-3.4%	-10.9%	29.3%
财报类	**IC**	**G1**	**G2**	**G3**	**G4**	**G5**	**G6**	**G7**	**G8**	**G9**	**G10**	**L/S**
大类因子：等权Zscore	4.5%	9.9%	9.1%	5.3%	4.0%	3.4%	2.8%	2.6%	-0.6%	0.7%	-2.3%	12.3%
大类因子：等权Rank	5.4%	15.1%	7.8%	6.5%	4.2%	3.0%	2.6%	2.9%	0.8%	-1.9%	-2.3%	17.3%
盈利预期类	**IC**	**G1**	**G2**	**G3**	**G4**	**G5**	**G6**	**G7**	**G8**	**G9**	**G10**	**L/S**
Net_Profit_F3_qoq	2.1%	13.2%	7.3%	6.3%	1.2%	-0.9%	0.1%	3.9%	4.5%	0.7%	0.5%	13.4%
反转类	**IC**	**G1**	**G2**	**G3**	**G4**	**G5**	**G6**	**G7**	**G8**	**G9**	**G10**	**L/S**
大类因子：等权Zscore	4.4%	14.9%	10.7%	12.8%	9.1%	13.2%	11.7%	4.2%	2.9%	-0.3%	-9.9%	24.8%
大类因子：等权Rank	4.7%	15.6%	11.4%	11.1%	11.0%	8.6%	9.8%	4.3%	2.9%	-1.9%	-7.8%	23.4%

注：回溯测试的业绩基准为中证500指数，IC为因子月度Rank IC

资料来源：万得，中银证券

从周度测算的回测结果来看，等权 rank 复合与等权 Zscore 复合的差别并不显著。若以 G1 超额收益为标准，等权 rank 的复合方法效果更优。

图表 9. 周度回测等权 zscore 与等权 rank 分组超额对比

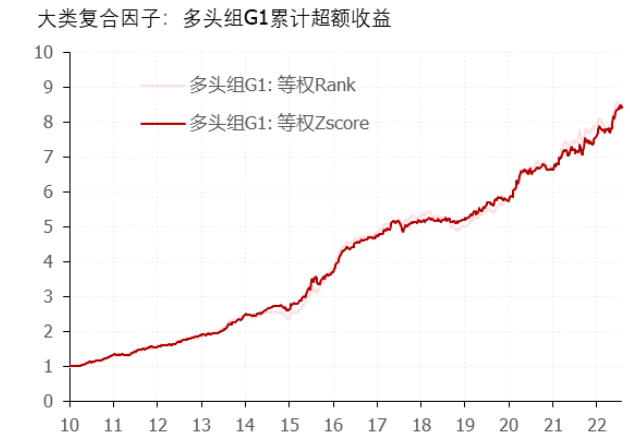


资料来源：万得，中银证券

针对中证 500 成分股，我们的 10 组分组测试显示，大类复合因子的表现显著增强。大类 Rank 等权复合因子月度 IC 为 7.5%, ICIR 为 2.8；周度换仓多头组 G1 年化超额收益超 18.4%，多空年化超额收益接近 30%。

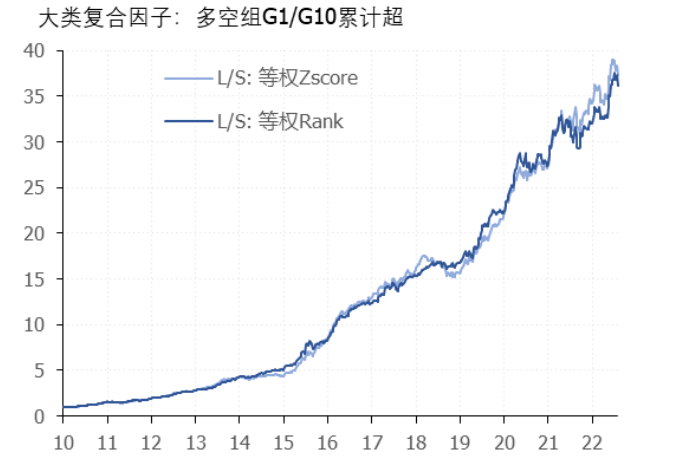
具体分组回测的较中证 500 年化超额收益、多头组 G1 累计超额收益净值以及多空组 G1/G10 累计超额收益净值如下。

图表 10. 两种传统方法复合 G1 组超额累计净值（周度）



资料来源：万得，中银证券

图表 11. 两种传统方法复合多空超额累计净值（周度）



资料来源：万得，中银证券

（二）传统大类因子复合：等权复合 + 月度换仓

上文框架均基于中证 500 成分股以周度换仓频率对因子进行分组回测。本节我们针对月度换仓对因子进行测算。实证可知，对比周度换仓，因子多头组 G1 年化超额收益由 18.4%衰减至 13%，多空组超额收益由 30%衰减至 20%。因子整体表现稳健，符合建模预期。

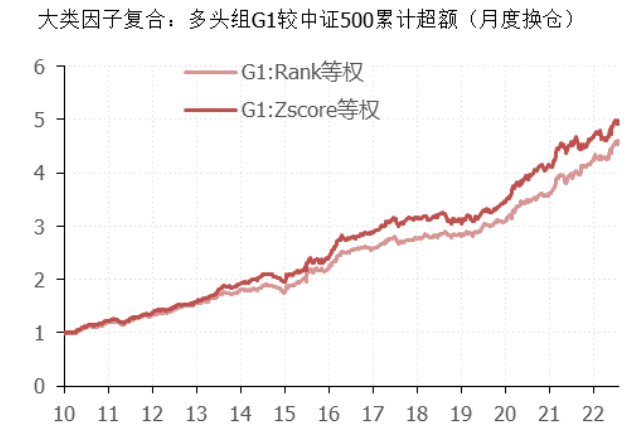
图表 12. 月度回测等权 zscore 与等权 rank 分组超额对比

复合因子	IC	G1	G2	G3	G4	G5	G6	G7	G8	G9	G10	L/S
Zscore等权	7.4%	13.5%	7.3%	7.1%	4.0%	5.0%	5.9%	1.0%	-1.2%	-2.6%	-6.9%	20.4%
Rank等权	7.5%	12.8%	8.6%	9.1%	7.0%	3.8%	1.7%	0.8%	-0.2%	-2.3%	-7.4%	20.1%

注：业绩比较基准为中证500指数，换仓频率为月度（4周），回测区间为2010-2022.07

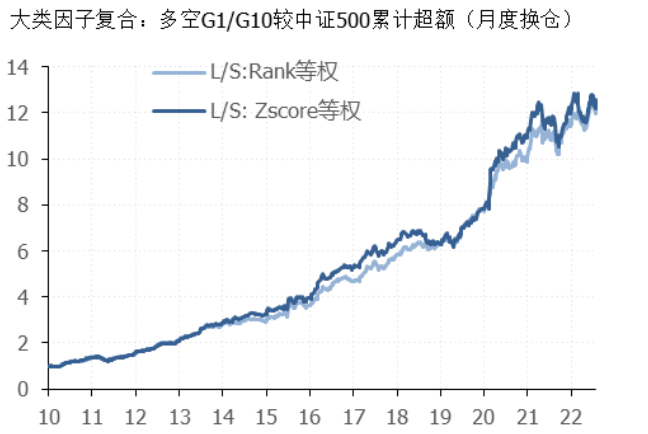
资料来源：万得，中银证券

图表 13. 两种传统方法复合 G1 组超额累计净值（月度）



资料来源：万得，中银证券

图表 14. 两种传统方法复合多空超额累计净值（月度）



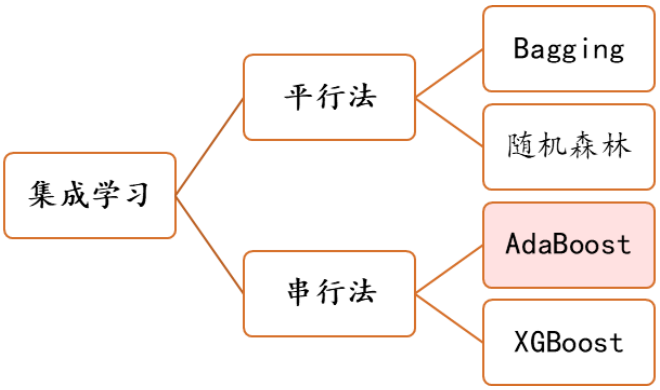
资料来源：万得，中银证券

（三）集成学习概述

集成学习（ensemble learning）：不是一个单独的机器学习算法，而是通过构建并结合多个机器学习器（基学习器）来完成学习任务。对于训练集数据，通过训练若干个个体弱学习器（weak learner），结合一定的策略就可以最终形成一个强学习器（strong learner），以达到博采众长的目的。集成学习构建方法可以分为两类：

- **平行法：**构建多个独立的学习器（通常是同质的弱学习器），个体学习器之间不存在强依赖关系，可以并行生成，最终结果为多学习器的预测平均值；
- **串行法：**多个学习器是依次串行构建的（通常是异质的弱学习器），个体学习器之间存在强依赖关系，最终结果为多学习器预测结果的加权汇总。

图表 15.平行法与串行法集成学习分类



资料来源：万得，中银证券

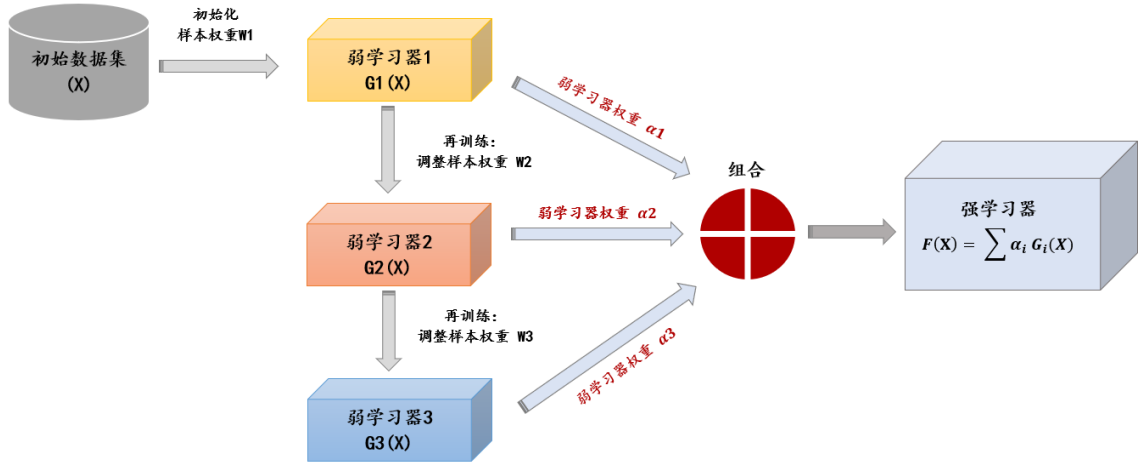
后文以串行法典型代表 AdaBoost（AdaBoost Regressor）为例，构建多因子复合框架。AdaBoost 是英文“Adaptive Boosting”（自适应增强）的缩写，具体算法分为如下三步：

i. **初始化训练数据的权值分布 w1。**假设有 N 个样本，则每个样本开始时均设为相同权值：w1 = 1/N；

ii. **训练弱分类器 G(X)。**具体训练过程：如果某个训练样本点，被弱分类器 Gi 准确地分类，那么再构造下一个训练集中，它对应的样本权值 w 要减小；相反，如果某个训练样本点被错误分类，那么它的样本权值 w 就应该增大。权值的更新过的样本被用于训练下一个弱分类器，整个过程如此迭代下去。

iii. **将各个训练得到的弱分类器 G(X)组合成一个强分类器 F(X)。**各个弱分类器的训练过程结束后，加大分类误差率小的弱分类器的权重α，使其在最终的分类函数中起着较大的决定作用，而降低分类误差率大的弱分类器的权重α，使其在最终的分类函数中起着较小的决定作用。**换而言之，误差率低的弱分类器在最终分类器中占的权重较大，否则较小。**

图表 16.Adaboost 算法示意图



资料来源：万得，中银证券

下图为将基学习器设定为 CART 分类树，进行样本二元分类情景时的 AdaBoost 算法框架，其中弱分类器 G(x)的权重α是基于“最小化分类模型指数误差”推导所得。

图表 17.基于 CART 分类树的二元分类 AdaBoost 算法框架

输入：训练集$D = \{(x_1, y_1), (x_2, y_2), \dots\dots, (x_m, y_m)\}, y \in \{-1, +1\}$
基学习算法
训练论数 T

1：初始化训练数据的权值分布：

$$D_1 = (W_{1,1}, \dots, W_{1,i}, \dots, W_{1,m}), \qquad W_{1i} = \frac{1}{m}, i = 1,2, \dots, m$$

2：*for* $t = 1,2,..T$ ***do***

(a) 使用具有权值分布D_t的训练集学习，得到基分类器$G_t(x)$
(b) 计算$G_t(x)$在训练数据集上的分类错误率：

$$e_t = P(G_t(x_i) \neq y_i) = \sum_{i=1}^{m} w_{ti} I(G_t(x_i) \neq y_i)$$

(c) 计算$G_t(x)$的系数

$$\alpha_t = \frac{1}{2} \ln \frac{1-e_t}{e_t}$$

(d) 更新训练数据的权值分布

$$D_{t+1} = (W_{t+1,1}, \dots, W_{t+1,i}, \dots, W_{t+1,m})$$

$$W_{t+1,i} = \frac{W_{t,i}}{Z_t} \exp(-\alpha_t y_i G_t(x_i)), i = 1,2,3, \dots, N$$

这里Z_t为规范化因子，$Z_t = \sum_{i=1}^{m} w_{t,i} \exp(-\alpha_t y_i G_t(x_i))$

3：构建基本分类器的线性组合

$$f(x) = \sum_{t=1}^{T} \alpha_t G_t(x)$$

4：得到最终分类器

$$G(x) = sign(f(x)) = sign(\sum_{t=1}^{T} \alpha_t G_t(x))$$

资料来源：万得，中银证券

传统情况下，业界常将股票的收益表现进行分类标签，比如在未来一期，能够跑到前 1/3 的股票标注为 1，其他股票标记为 0，然后通过 CART 分类器对股票的分类进行预测，并基于 Adaboost 的框架进行预测准确度的提升，并使用下一期标签为 1 的概率作为复合因子，多股票进行分组排序以及组合优化。本报告我们将尝试一个新的 Adaboost 框架，对多因子复合进行升级，具体见（四）（五）部分。

（四）AdaBoost Regressor：多因子复合框架

预测目标：以最优化“因子下一期排名分位数预测值为目标”构建多因子复合框架。其本质是利用子类因子历史的数据，即最优化预测中证 500 成分股的下期 IC 值，且希望排名越靠前（分位数越接近 1 越优秀）的股票对应的预测结果越准确。

模型选择为“AdaBoost Regressor + 基学习器 Linear Regression”：通过滚动历史因子信息对未来一期的股票排名分位数进行预测，复合因子即为未来一期股票预期排名分位数 quantile(i,t)。

$$quantile_{i,t+1} = \alpha_t + \sum_k w_k f_{k,i,t}$$

模型参数设定：

- **样本权重 sample_weigth：**我们希望对下期排名越靠前的股票预测准度越高，故将样本 y_i 权重设定为 $\frac{e^{yi}}{\sum e^{yi}}$；
- **学习率 Learning Rate ν：**学习策略是一个一个对基学习器进行学习，然后确定每个基学习器的系数，学习率（learning_rate）和弱学习器个数（n_estimators)之间存在权衡关系，合理调节这两个参数可以很好缓解过拟合问题，通常我们将ν设定从 0.1 开始，该参数范围为（0,1]。

$$F_m(x) = F_{m-1}(x) + \nu\alpha G_{m-1}(x)$$

- **弱学习器个数 n_estimators：**参数默认的值为 50。如果参数值太大，容易过拟合；参数值太小，容易欠拟合。由于 Adaboost 算法本身不容易出现过拟合问题，因此在学习率为 0.1 的情况下，可初始化为 500（近似等价于学习率为 1 时，数量为 50）。该参数也可通过样本内交叉验证寻找最优值。
- **Loss 损失函数：**可选的参数有：linear、square、exponential，分别对应于线性损失函数，平方损失函数、指数损失函数，本报告设定为 linear。
- **历史数据与时间区间选择：**滚动 1 个季度、1 年、3 年的因子信息，信息为因子中性化 Zscore 或排名分位数。

（五）AdaBoost Regressor：多因子复合框架实证

单因子优选集合：将 2010-2020 年样本内“财务类”、“盈利预期类”以及“技术类”三大类因子经过凝聚分层聚类优选得到的 14 个因子（财务 7 个 + 盈利预期 1 个 + 反转类 6 个）作为复合因子的单因子优选集合。选择聚类优选因子的原因尽可能规避冗余的因子信息，且显著提升 Adaboost 的计算效率。

回溯时间：通过滚动历史样本来预测未来 1 周的股票排名分位数。实证表明滚动 1 年和滚动 3 年的模型表现差异不大，且显著优于 1 季度结果。为提升模型训练速度，推荐选择滚动 1 年窗口数据。

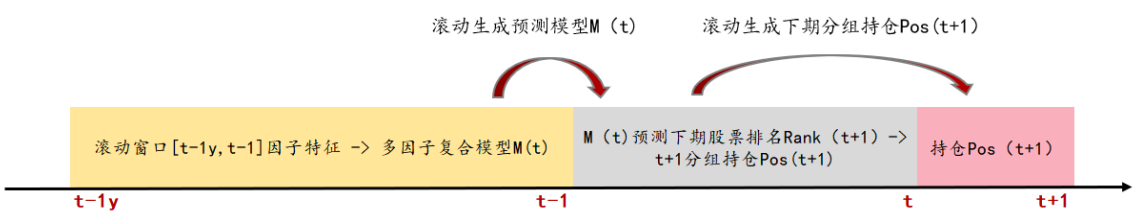
因子特征选择：有两种选择，一是因子的 Zscore，二是因子的分位数（Rank）。实证表明使用因子排名分位数效果略优于 Zscore。

模型选择为“AdaBoost Regressor + 基学习器 Linear Regression”：通过滚动历史因子信息对未来一期的股票排名分位数进行预测，复合因子即为未来一期股票预期排名分位数 quantile(i,t+1)。

$$quantile_{i,t+1} = \alpha_t + \sum_k w_k f_{k,i,t}$$

参数选择：learning rate 设定为(0.01, 0.1,1)，既学习器个数 n_estimators 的设定采用两种模式：统一设定为 500；基于“滚动 n 期样本内交叉验证（cv=3）最优”两种模式。实证表明两种模式模型比较非常相似，因此直接使用（0.1,500）参数组合即可高效完成测试。

图表 18.使用 AdaBoost Regressor 进行多因子复合实证研究示意图



资料来源：万得，中银证券

AdaboostReg 框架下的多因子复合在样内本外表现均较传统方法有进一步提升，尤其在平滑各年度之间超额收益的稳定性层面较传统方法效果进一步提升。

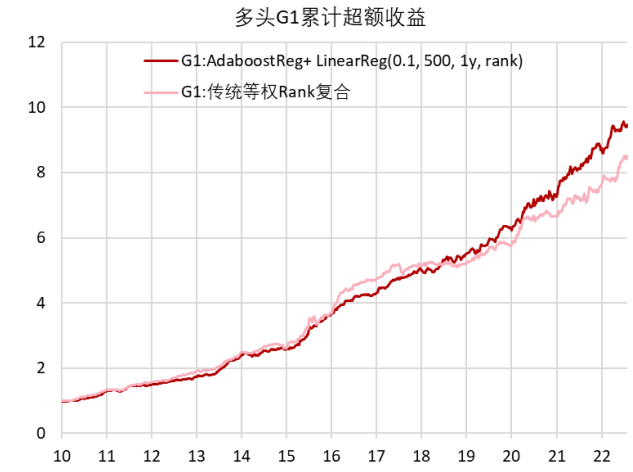
图表 19.各参数组合下 AdaBoost Reg 多因子复合回测结果

learning rate	滚动窗口	学习器个数 n estimator	IC	G1年化超额：样本内	G1年化超额：样本外	G10年化超额：样本内	G10年化超额：样本外	L/S年化超额：样本内	L/S年化超额：样本外
0.1	3y	500	8.0%	19.3%	15.4%	-14.1%	-5.6%	33.4%	21.0%
0.1	1y	500	8.1%	19.2%	16.9%	-13.0%	-7.0%	32.2%	23.9%
0.1	1y	cv = 3	8.0%	19.1%	16.9%	-13.1%	-7.0%	32.2%	23.9%
0.01	1y	1000	8.1%	19.1%	16.0%	-13.2%	-7.8%	32.3%	23.8%
1	1y	500	7.9%	18.5%	16.2%	-12.1%	-3.0%	30.6%	19.2%
0.1	1q	500	7.8%	17.9%	13.5%	-7.9%	4.4%	25.8%	9.1%
传统大类因子复合：Rank等权			7.5%	18.2%	15.6%	-11.8%	-2.5%	30.0%	18.1%
传统大类因子复合：Zscore等权			7.4%	18.2%	16.1%	-11.7%	-4.5%	29.9%	20.6%

注：业绩比较基准为中证500指数，样本内区间为2010-2020.12，样本外区间为2021-2022.07

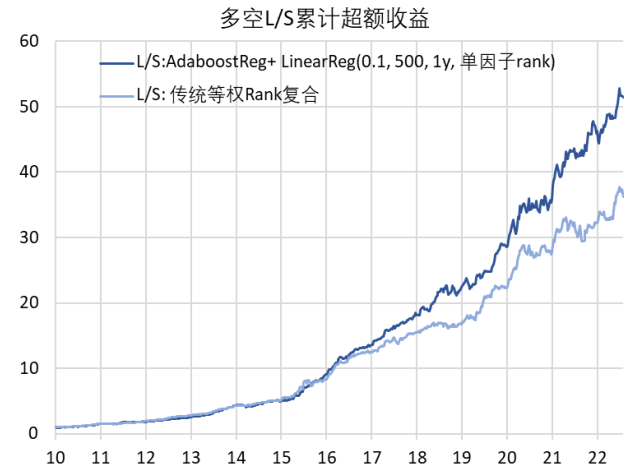
资料来源：万得，中银证券

图表 20. AdaBoost 复合多头组超额优于传统 rank 复合



资料来源：万得，中银证券

图表 21. AdaBoost 复合多空超额优于传统 rank 复合



资料来源：万得，中银证券

五、 附录：相似因子的聚类优选框架

在多因子构建的过程中，我们很容易批量生产大量构建逻辑相似，单因子表现也相似的因子群，如何在因子群中优选单因子，并在一定程度上减少因子的相关性是一个核心难点。

传统因子优选层面，我们可以通过遍历测算多因子等权复合的方法寻找相对靠谱的复合因子池，但这种算法计算量极大，而且非常耗时，且存在一定的过拟合风险。

中银量化团队提出可基于无监督学习的框架对相似因子进行聚类优选，并在每一类中选取代表性的因子来构成精选优质因子池。

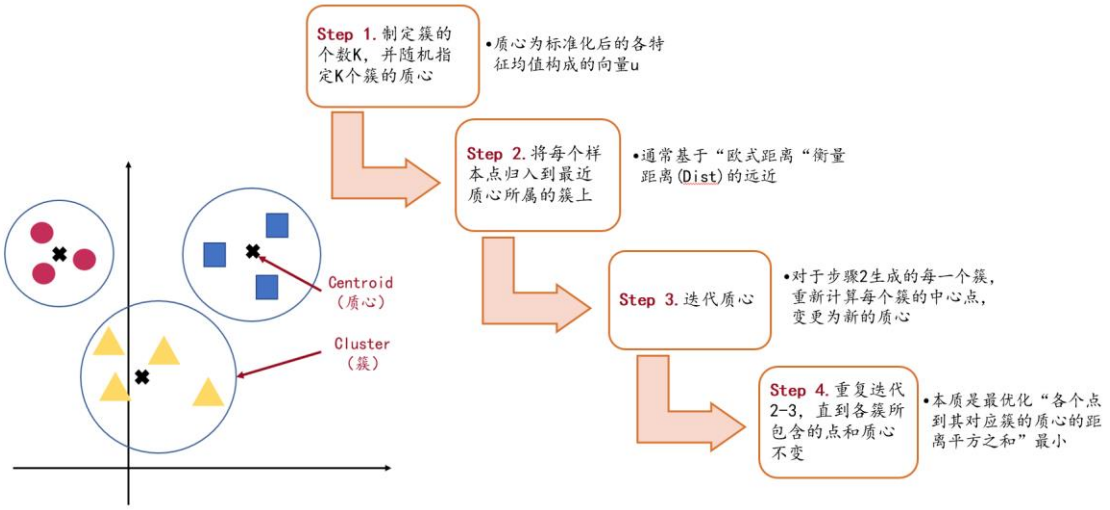
无监督聚类算法

1. K-Means 算法

本报告首先介绍 K-Means 算法对多个相似因子进行聚类。

K-Means 算法：聚类结果非常容易受到“初始化质心”的影响（建议多次测算取平均结果），算法复杂度与样本数量呈线性关系，具体算法流程图如下所示。

附录图表 1. K-Means 算法原理



资料来源：万得，中银证券

2. 层次聚类算法

层次聚类算法分为“自下而上”与“自上而下”两种模式，本报告重点讨论“自下而上”的凝聚层次聚类法，具体思想为：将每个对象作为一个簇，然后合并这些原子簇为越来越大的簇，直到某个终结条件被满足，具体原理如下。

凝聚法的每一步需要合并“距离最小的两个族群”，而不同族群间距离的定义方法决定了不同的聚类结果，关于凝聚法的距离定义主要有两种思想：连接法和 Ward 法。

实证显示“连接法”与 Ward 法（使得聚类导致的类内离差平方和增量最小）聚类效果相似。

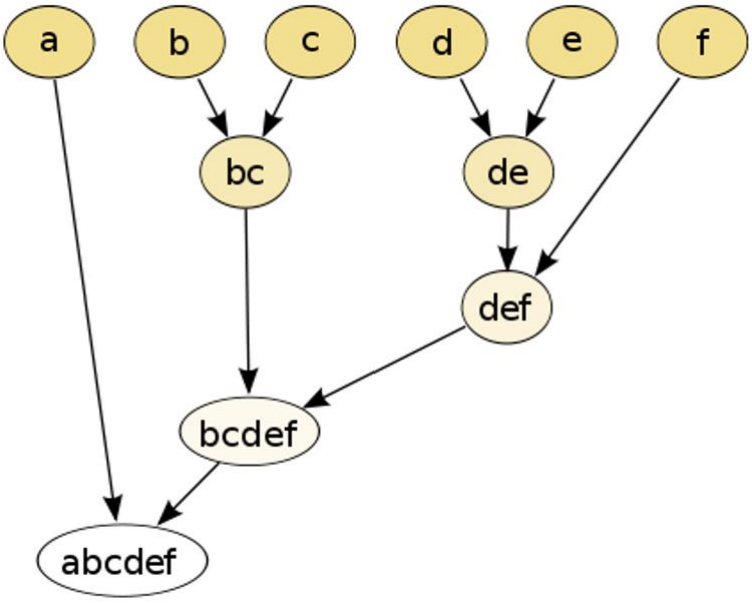
简单连接：定义两族群间相隔最近的两个个体间的距离，为两族群的距离；

完全连接：定义两族群间相隔最远的两个个体间的距离，为两族群的距离；

平均连接：A 群中所有的 N1 个样本与 B 群中所有的 N2 个样本产生的距离（共计 N1xN2 个距离），求平均值作为两个族群的距离；

质心连接：两个群中各自的质心（即样本均值向量），之间的欧式距离，作为两个族群的距离。

附录图表 2.“自下而上”的凝聚层次聚类方法



资料来源：搜狐，中银证券

针对无监督学习的聚类算法，大家都可能面临一个实证过程中的难点：即如何选定合理的簇数 K？即大类因子应该分为几个子分类相对最优呢？这里我们引入一个轮廓系数的概念，来基于样本内区间对因子进行最优聚类。

聚类时如何选择最优 K？

轮廓系数（Peter，1986）一种综合考量聚类内聚度和分类度的聚类效果评价方法。对于某个簇中的每一个样本 i，分别计算样本的轮廓系数 s(i)，然后将所有样本的轮廓系数 s(i)求平均，即可得到该聚类算法的总轮廓系数 S，具体计算流程如下：

1） 簇的内聚度 a(i)：为样本 i 到同一簇内其他点的平均值；

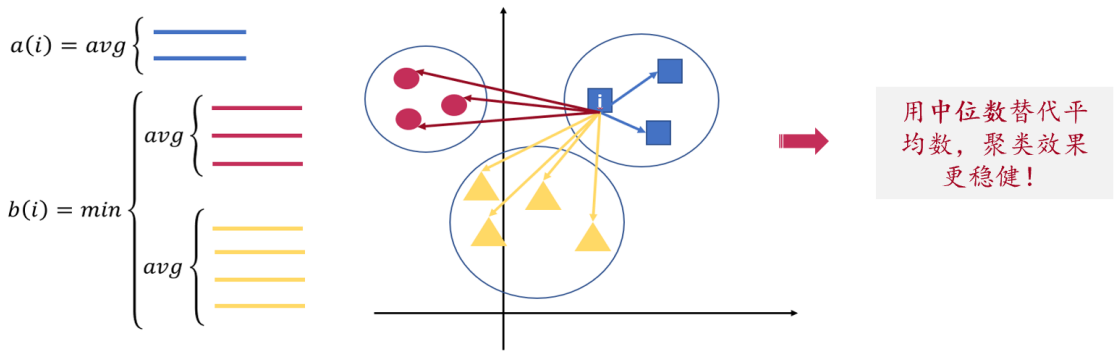
2） 簇间分类度 b(i)：为样本 i 到其他簇的平均距离的最小值；

3） 样本 i 的轮廓系数 s(i)：$s(i) = \frac{b(i)-a(i)}{\max\{a(i),b(i)\}}$

4） 计算该聚类算法的总轮廓系数 S：$S = avg(s(i))$；

5） 轮廓系数 s(i)介于 [-1,1]，越趋近于 1 代表内聚度和分离度都相对较优，聚类效果越好。

附录图表 3. 基于簇内中位数构建的轮廓系数，其聚类效果更为稳健



资料来源：万得，中银证券

多因子聚类筛选的规则与流程如下：

1) **样本区间划分：**将 2010 年 1 月 1 日-2020 年 12 月 31 日定义为样本内区间，2021 年 1 月 1 日-2022 年 7 月 31 日为样本外测试区间；

2) **因子库筛选：**以“样本内区间，单因子 IC>3%，多头 G1 组超额 > 5%”为标准，筛选财报类因子库，共有 65 个财报单因子符合标准；

3) **聚类算法选择：**由于 K-Means 聚类算法结果存在不稳定风险，进行实证测算对比，我们推荐凝聚分层聚类算法；

4) **聚类特征选择：**样本内单因子的 IC 时间序列。因为相较于单因子超额收益时间序列，在进行多因子复合时，我们更推荐关注单因子的 IC 与 ICIR，即因子值映射到股票预期收益率的特征维度；

5) **最优 K 选择：**基于样本内因子特征，我们采用“凝聚分层聚类算法”对单因子的 IC 序列进行聚类，并基于最高的轮廓系数来选择因子的聚类簇数，实证结果显示，K 的最优数量为 7；

6) **最优单因子筛选逻辑：**我们更关注因子多头组获取超额收益的能力，因此在同一簇内，我们选择“多头组 G1 超额收益最高的因子”作为该簇的代表因子。

六、 风险提示

投资者应注意基于历史数据构建的模型失效风险。

披露声明

本报告准确表述了证券分析师的个人观点。该证券分析师声明，本人未在公司内、外部机构兼任有损本人独立性与客观性的其他职务，没有担任本报告评论的上市公司的董事、监事或高级管理人员；也不拥有与该上市公司有关的任何财务权益；本报告评论的上市公司或其它第三方都没有或没有承诺向本人提供与本报告有关的任何补偿或其它利益。

中银国际证券股份有限公司同时声明，将通过公司网站披露本公司授权公众媒体及其他机构刊载或者转发证券研究报告有关情况。如有投资者于未经授权的公众媒体看到或从其他机构获得本研究报告的，请慎重使用所获得的研究报告，以防止被误导，中银国际证券股份有限公司不对其报告理解和使用承担任何责任。

评级体系说明

以报告发布日后公司股价/行业指数涨跌幅相对同期相关市场指数的涨跌幅的表现为基准：

公司投资评级：

买　　入：预计该公司股价在未来 6-12 个月内超越基准指数 20%以上；

增　　持：预计该公司股价在未来 6-12 个月内超越基准指数 10%-20%；

中　　性：预计该公司股价在未来 6-12 个月内相对基准指数变动幅度在-10%-10%之间；

减　　持：预计该公司股价在未来 6-12 个月内相对基准指数跌幅在 10%以上；

未有评级：因无法获取必要的资料或者其他原因，未能给出明确的投资评级。

行业投资评级：

强于大市：预计该行业指数在未来 6-12 个月内表现强于基准指数；

中　　性：预计该行业指数在未来 6-12 个月内表现基本与基准指数持平；

弱于大市：预计该行业指数在未来 6-12 个月内表现弱于基准指数；

未有评级：因无法获取必要的资料或者其他原因，未能给出明确的投资评级。

沪深市场基准指数为沪深 300 指数；新三板市场基准指数为三板成指或三板做市指数；香港市场基准指数为恒生指数或恒生中国企业指数；美股市场基准指数为纳斯达克综合指数或标普 500 指数。

风险提示及免责声明

本报告由中银国际证券股份有限公司证券分析师撰写并向特定客户发布。

本报告发布的特定客户包括：1) 基金、保险、QFII、QDII 等能够充分理解证券研究报告，具备专业信息处理能力的中银国际证券股份有限公司的机构客户；2) 中银国际证券股份有限公司的证券投资顾问服务团队，其可参考使用本报告。中银国际证券股份有限公司的证券投资顾问服务团队可能以本报告为基础，整合形成证券投资顾问服务建议或产品，提供给接受其证券投资顾问服务的客户。

中银国际证券股份有限公司不以任何方式或渠道向除上述特定客户外的公司个人客户提供本报告。中银国际证券股份有限公司的个人客户从任何外部渠道获得本报告的，亦不应直接依据所获得的研究报告作出投资决策；需充分咨询证券投资顾问意见，独立作出投资决策。中银国际证券股份有限公司不承担由此产生的任何责任及损失等。

本报告内含保密信息，仅供收件人使用。阁下作为收件人，不得出于任何目的直接或间接复制、派发或转发此报告全部或部分内容予任何其他人，或将此报告全部或部分内容发表。如发现本研究报告被私自刊载或转发的，中银国际证券股份有限公司将及时采取维权措施，追究有关媒体或者机构的责任。所有本报告内使用的商标、服务标记及标记均为中银国际证券股份有限公司或其附属及关联公司（统称“中银国际集团”）的商标、服务标记、注册商标或注册服务标记。

本报告及其所载的任何信息、材料或内容只提供给阁下作参考之用，并未考虑到任何特别的投资目的、财务状况或特殊需要，不能成为或被视为出售或购买或认购证券或其它金融票据的要约或邀请，亦不构成任何合约或承诺的基础。中银国际证券股份有限公司不能确保本报告中提及的投资产品适合任何特定投资者。本报告的内容不构成对任何人的投资建议，阁下不会因为收到本报告而成为中银国际集团的客户。阁下收到或阅读本报告须在承诺购买任何报告中所指之投资产品之前，就该投资产品的适合性，包括阁下的特殊投资目的、财务状况及其特别需要寻求阁下相关投资顾问的意见。

尽管本报告所载资料的来源及观点都是中银国际证券股份有限公司及其证券分析师从相信可靠的来源取得或达到，但撰写本报告的证券分析师或中银国际集团的任何成员及其董事、高管、员工或其他任何个人（包括其关联方）都不能保证它们的准确性或完整性。除非法律或规则规定必须承担的责任外，中银国际集团任何成员不对使用本报告的材料而引致的损失负任何责任。本报告对其中所包含的或讨论的信息或意见的准确性、完整性或公平性不作任何明示或暗示的声明或保证。阁下不应单纯依靠本报告而取代个人的独立判断。本报告仅反映证券分析师在撰写本报告时的设想、见解及分析方法。中银国际集团成员可发布其它与本报告所载资料不一致及有不同结论的报告，亦有可能采取与本报告观点不同的投资策略。为免生疑问，本报告所载的观点并不代表中银国际集团成员的立场。

本报告可能附载其它网站的地址或超级链接。对于本报告可能涉及到中银国际集团本身网站以外的资料，中银国际集团未有参阅有关网站，也不对它们的内容负责。提供这些地址或超级链接（包括连接到中银国际集团网站的地址及超级链接）的目的，纯粹为了阁下的方便及参考，连结网站的内容不构成本报告的任何部份。阁下须承担浏览这些网站的风险。

本报告所载的资料、意见及推测仅基于现状，不构成任何保证，可随时更改，毋须提前通知。本报告不构成投资、法律、会计或税务建议或保证任何投资或策略适用于阁下个别情况。本报告不能作为阁下私人投资的建议。

过往的表现不能被视作将来表现的指示或保证，也不能代表或对将来表现做出任何明示或暗示的保障。本报告所载的资料、意见及预测只是反映证券分析师在本报告所载日期的判断，可随时更改。本报告中涉及证券或金融工具的价格、价值及收入可能出现上升或下跌。

部分投资可能不会轻易变现，可能在出售或变现投资时存在难度。同样，阁下获得有关投资的价值或风险的可靠信息也存在困难。本报告中包含或涉及的投资及服务可能未必适合阁下。如上所述，阁下须在做出任何投资决策之前，包括买卖本报告涉及的任何证券，寻求阁下相关投资顾问的意见。

中银国际证券股份有限公司
中国上海浦东
银城中路 200 号
中银大厦 39 楼
邮编 200121
电话: (8621) 6860 4866
传真: (8621) 5888 3554

相关关联机构：

中银国际研究有限公司
香港花园道一号
中银大厦二十楼
电话:(852) 3988 6333
致电香港免费电话:
中国网通 10 省市客户请拨打：10800 8521065
中国电信 21 省市客户请拨打：10800 1521065
新加坡客户请拨打：800 852 3392
传真:(852) 2147 9513

中银国际证券有限公司
香港花园道一号
中银大厦二十楼
电话:(852) 3988 6333
传真:(852) 2147 9513

中银国际控股有限公司北京代表处
中国北京市西城区
西单北大街 110 号 8 层
邮编:100032
电话: (8610) 8326 2000
传真: (8610) 8326 2291

中银国际(英国)有限公司
2/F, 1 Lothbury
London EC2R 7DB
United Kingdom
电话: (4420) 3651 8888
传真: (4420) 3651 8877

中银国际(美国)有限公司
美国纽约市美国大道 1045 号
7 Bryant Park 15 楼
NY 10018
电话: (1) 212 259 0888
传真: (1) 212 259 0889

中银国际(新加坡)有限公司
注册编号 199303046Z
新加坡百得利路四号
中国银行大厦四楼(049908)
电话: (65) 6692 6829 / 6534 5587
传真: (65) 6534 3996 / 6532 3371