

量化专题报告

“量价淘金”选股因子系列研究（十二）

“高频数据+离散化构建方式”在因子研究中的重要性

前言：随着市场的发展与竞争的加剧，尤其是机器学习模型的普及，想要挖掘现有多因子模型以外的增量信息越来越困难。本篇报告提出以下论点：未来量价因子的研究，想要获取更多的、与现有因子低相关性的增量，“逐笔等高频数据”结合“离散化的因子构建方式”至关重要。

连续性的因子构建方式：所谓“连续性”，即基于某一段连续的数据计算因子。以常见的“量价相关性类因子”为例，分别基于日频、分钟、逐笔数据，采用连续性的构建方式，批量生产对应的因子簇。将各自因子簇中的因子等权合成，得到“日频连续”、“分钟连续”、“逐笔连续”因子。考察“分钟连续”、“逐笔连续”因子相对于“日频连续”因子的增量，发现虽然从因子的回测效果上来看，将“分钟连续”、“逐笔连续”叠加到“日频连续”上，因子绩效确有显著提升，但一旦落实到组合层面，“分钟连续”、“逐笔连续”的贡献就变得非常有限。以沪深 300 指数增强组合为例，叠加“分钟连续”、“逐笔连续”因子构建的组合，相对于仅用“日频连续”因子构建的组合，超额年化收益只能提升 0.5%-0.6% 左右。

离散化的因子构建方式：所谓“离散化”，即在过去一段时间内，基于某些特征、抽取几个离散的时间点，只用与这些时间点相关的数据计算因子。分别基于分钟、逐笔数据，采用离散化的方式构建因子簇，同样将因子簇中的因子等权合成，得到“分钟离散”、“逐笔离散”因子。考察离散化因子相对于连续性因子的增量，发现虽然将“分钟离散”、“逐笔离散”叠加到“连续”因子上，因子的回测表现并未得到明显提升，但落实到组合层面，离散化因子却能提供显著的增量。比如基于“连续+逐笔离散”因子构建的沪深 300 指数增强组合，相对基准的超额年化收益为 8.15%，跟踪误差为 3.81%，信息比率为 2.14，月度胜率为 75.68%，最大回撤为 4.27%；相较于只用“连续”因子构建的组合，超额年化收益可以提升 1% 以上。

“逐笔离散”因子的表现：在逐笔数据层面，考察放量时间点之后的量价相关性特征，构建“逐笔离散”因子。回测时间段 2016/01/01-2025/04/30 内，以全体 A 股为研究样本，该因子的月度 IC 均值为 0.073，年化 ICIR 为 2.87，月度 RankIC 均值为 0.099，年化 RankICIR 为 3.37；10 分组多空对冲的年化收益为 38.30%，年化波动为 11.88%，信息比率为 3.22，月度胜率为 81.98%，最大回撤为 11.87%。剔除常用风格与行业的影响后，因子仍然有效，年化 ICIR 为 2.84。

风险提示：以上结论均基于历史数据和统计模型的测算，如果未来市场环境发生明显改变，不排除模型失效的可能性。

作者

分析师 沈芷琦

执业证书编号：S0680521120005

邮箱：shenzhiqi@gszq.com

分析师 刘富兵

执业证书编号：S0680518030007

邮箱：liufubing@gszq.com

研究助理 阮俊烨

执业证书编号：S0680124070019

邮箱：ruanjunye@gszq.com

相关研究

- 1、《“量价淘金”选股因子系列研究（一）：如何将隔夜涨跌变为有效的选股因子？——基于对知情交易者信息优势的刻画》2022-04-26
- 2、《“量价淘金”选股因子系列研究（二）：不同交易者结构下的动量与反转》2022-10-24
- 3、《“量价淘金”选股因子系列研究（三）：如何基于 RSI 技术指标构建有效的选股因子》2023-03-06
- 4、《“量价淘金”选股因子系列研究（四）：高/低位放量：从事件驱动到选股因子》2023-12-18
- 5、《“量价淘金”选股因子系列研究（五）：基于趋势资金日内交易行为的选股因子》2024-05-28
- 6、《“量价淘金”选股因子系列研究（六）：创新高股票中的 Alpha》2024-06-26
- 7、《“量价淘金”选股因子系列研究（七）：盲目追随趋势资金的极端交易行为分析——羊群效应的识别与因子构建》2024-08-06
- 8、《“量价淘金”选股因子系列研究（八）：逐笔买卖差异中的选股信息——条件成交不平衡因子》2024-10-09
- 9、《“量价淘金”选股因子系列研究（九）：Memory Map 在因子生产加速上的应用——以构建羊群效应因子簇为例》2024-11-20
- 10、《“量价淘金”选股因子系列研究（十）：订单簿资金流因子簇的构建与生产加速——基于 Memory Map 的分段读取性能优势》2025-02-11
- 11、《“量价淘金”选股因子系列研究（十一）：基于趋势资金日内交易行为的事件驱动策略》2025-02-26

内容目录

一、前言	3
二、连续性的因子构建方式	5
2.1 日频数据+连续性构建	5
2.2 分钟数据+连续性构建	7
2.3 逐笔数据+连续性构建	8
2.4 “分钟连续”、“逐笔连续”因子的增量	10
三、离散化因子的构建方式	13
3.1 分钟数据+离散化构建	13
3.2 逐笔数据+离散化构建	14
3.3 “分钟离散”、“逐笔离散”因子的增量	16
四、其他重要讨论	19
4.1 “逐笔离散”因子的分年度表现	19
4.2 “逐笔离散”因子剔除常见风格与行业后的表现	19
4.3 “逐笔离散”因子的指数增强组合表现	20
五、总结	23
风险提示	23

图表目录

图表 1: 未来量价因子研究的重点: 高频数据+离散化构建方式	4
图表 2: 日频量价相关性因子的构建方法	5
图表 3: “日频连续”因子 10 分组及多空对冲净值	6
图表 4: “日频连续”因子的沪深 300 指数增强组合净值	7
图表 5: 分钟连续量价相关性因子的构建方法	7
图表 6: “分钟连续”因子 10 分组及多空对冲净值	8
图表 7: 逐笔连续量价相关性因子的构建方法	9
图表 8: “逐笔连续”因子 10 分组及多空对冲净值	9
图表 9: 连续性因子的 IC 及多空绩效	10
图表 10: “日频连续+分钟连续”因子的沪深 300 指数增强组合净值	11
图表 11: “日频连续+逐笔连续”因子的沪深 300 指数增强组合净值	11
图表 12: “日频连续+分钟连续+逐笔连续”因子的沪深 300 指数增强组合净值	12
图表 13: 连续性因子的沪深 300 指数增强组合超额绩效	12
图表 14: 分钟离散量价相关性因子的构建方法	13
图表 15: “分钟离散”因子 10 分组及多空对冲净值	14
图表 16: 逐笔离散量价相关性因子的构建方法	15
图表 17: “逐笔离散”因子 10 分组及多空对冲净值	15
图表 18: “连续”、“离散”、“连续+离散”因子的 IC 及多空绩效	16
图表 19: “连续+分钟离散”因子的沪深 300 指数增强组合净值	17
图表 20: “连续+逐笔离散”因子的沪深 300 指数增强组合净值	17
图表 21: “连续+分钟离散+逐笔离散”因子的沪深 300 指数增强组合净值	18
图表 22: “连续”、“连续+离散”因子的沪深 300 指数增强组合超额绩效	18
图表 23: “逐笔离散”因子分年度表现	19
图表 24: “逐笔离散”因子与 Barra 风格因子相关系数	19
图表 25: “纯净逐笔离散”因子 10 分组及多空对冲净值	20
图表 26: “纯净逐笔离散”因子分年度表现	20
图表 27: “逐笔离散”因子的沪深 300 指数增强组合净值	21
图表 28: “逐笔离散”因子的中证 500 指数增强组合净值	21
图表 29: “逐笔离散”因子的中证 1000 指数增强组合净值	22

一、前言

在多因子选股的研究领域中，关于“挖掘量价类新因子”的研究，涉及的核心因素为：基于何种量价数据、如何构建因子。

对于数据源，根据其频率高低，可分为低频、中频、高频，低频如日度行情数据，中频如分钟行情数据，高频如逐笔成交、逐笔委托数据等。由于中高频数据往往包含更详细的投资者交易行为信息，因此利用中高频数据构建的因子有效性通常更强。

对于因子的构建方式，可以分为连续性构建、离散化构建。连续性的构建方式，即基于某一段连续的数据计算因子，比如计算股票每日第一小时的分钟涨跌幅序列的标准差；而离散化的构建方式，则基于几个离散的数据点计算因子，比如每个股票都根据当天的实际交易情况，找到几个关键的时间点，再利用这些时间点附近的数据构建因子。

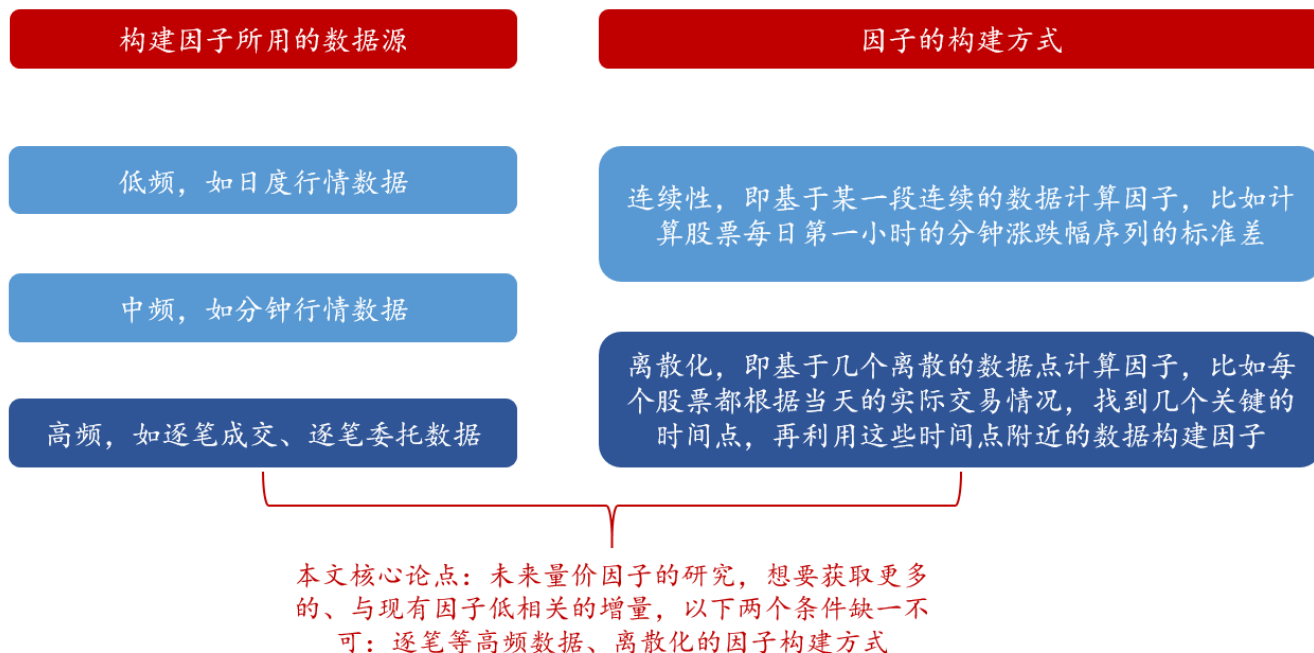
随着市场的发展与竞争的加剧、尤其是机器学习模型的普及，连续性的因子构建方法已经越来越难有增量，即使利用高频数据构建的、单因子回测效果较强的连续性因子，我们也经常发现其与众多低频数据构建的因子存在较高的相关性，一旦考察其相对于低频数据构建因子以外的增量信息，有效性就会显著下降。因此，想要获取与现有多因子模型低相关的增量信息，我们认为采用离散化的方式构建因子至关重要。

近些年，已经有很多因子研究者利用分钟数据构建离散化的因子，但大部分因子的有效性也在快速衰减。因为在分钟层面人为提取的离散化信息，也很容易被机器学习等模型提取到，比如只要在遗传算法中，输入分钟特征、加入离散化的算子，就可以批量、快速挖掘出很多离散化的因子。遗传算法等模型的大量应用，导致分钟数据离散化因子的拥挤度越来越高，有效性自然逐渐下降。

那么未来量价因子研究，获取更多增量信息的方向究竟在哪里？我们认为“逐笔等高频数据+离散化的因子构建方式”。逐笔数据由于其数据量庞大且不规整，目前几乎没有研究者直接将其输入到机器学习等模型中。大家对于逐笔数据的普遍用法，是先将逐笔降频为分钟或者秒级别的数据，再人为或者借助机器学习等模型来挖掘因子，直接利用原始逐笔数据、批量构建离散化因子的做法相对较少。因此我们提出，**未来量价因子的研究，想要获取更多的、与现有因子低相关性的增量，以下两个条件缺一不可：逐笔等高频数据、离散化的因子构建方式。**

本篇报告将从一个具体的例子出发，将不同频率的数据源与不同的构建方式进行搭配，借助本系列报告提出的“因子簇”的生产理念，挖掘不同种类的因子簇，并基于这些因子簇构建常见的指数增强组合，通过对比分析指增组合的表现，展示“高频数据+离散化构建方式”的重要性。

图表1: 未来量价因子研究的重点: 高频数据+离散化构建方式



资料来源: 国盛证券研究所整理

二、连续性的因子构建方式

2.1 日频数据+连续性构建

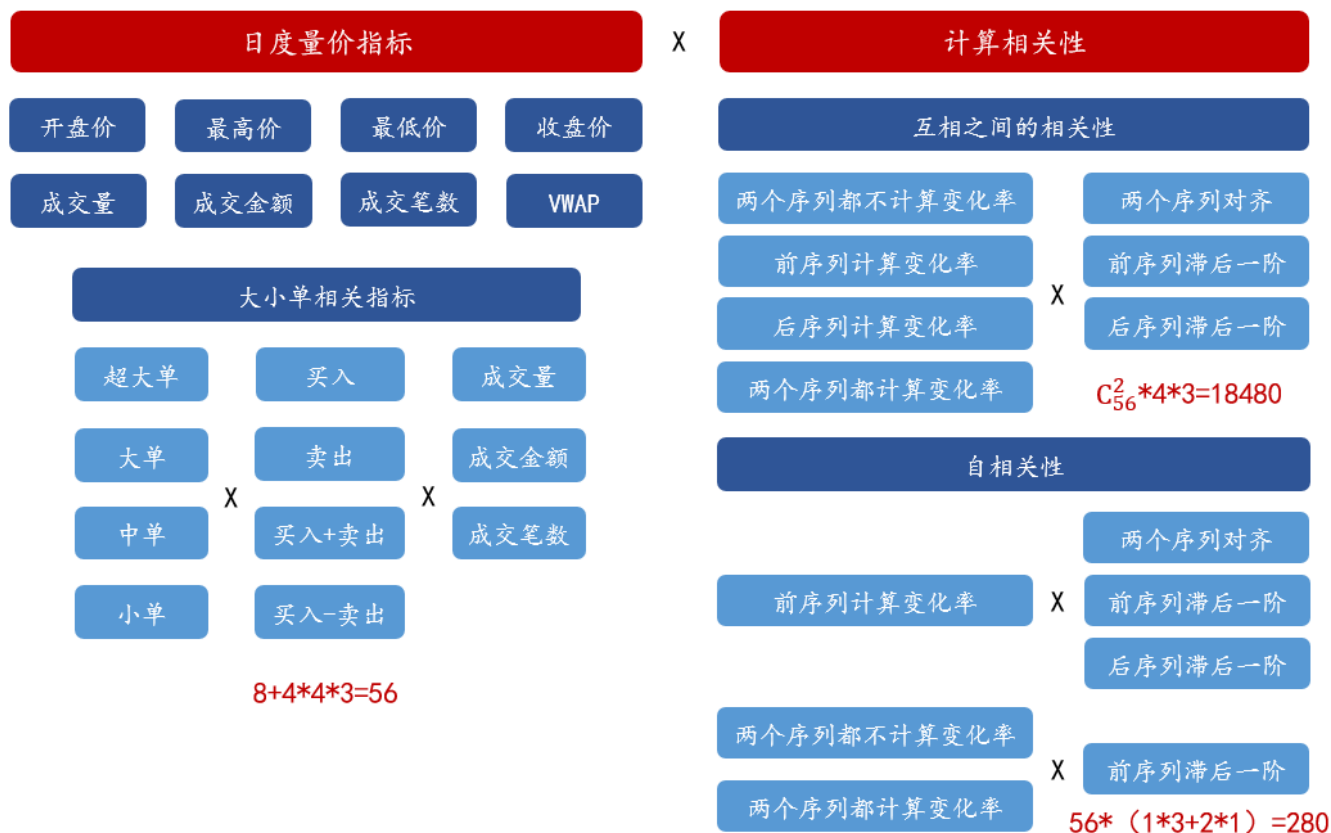
本文以量价因子中常见的“量价相关性类因子”为例，先用日频行情数据，构建一个量价相关性因子簇，作为后续因子的比较基准。特别说明，由于日频行情序列的数据点较少，若再进行离散化提取，则计算两个序列相关系数的误差较大，因此对于日频数据，本文不构建离散化的量价相关性因子，只构建连续性因子。

本小节构建量价相关性类因子的步骤如下：

(1) 设计日度量价指标：日度量价数据，可从 wind 数据库中直接提取，本文用到的数据包括个股每日的开盘价、最高价、最低价、收盘价、成交量、成交金额、成交笔数、VWAP、大小单相关指标。其中，大小单相关指标，可分为 3 个维度，“超大单、大单、中单、小单”，“买入、卖出、买入+卖出、买入-卖出”，“成交量、成交金额、成交笔数”，这 3 个维度互相搭配，可产生 48 种不同的大小单指标。因此，如图表 2 所示，通过这一步骤，我们一共得到了 56 种日度量价指标。

(2) 计算相关性：可分为上述指标两两之间的相关性、或者单一指标的自相关性。对于两两指标互相之间的相关性，根据指标是否计算变化率、计算相关性时是否错位，可分为多种情况，如图表 2 所示，共可计算 18480 个因子；对于单一指标的自相关性，也可通过类似的分类讨论，构建 280 个因子。因此基于日频数据，我们一共得到了 18760 个量价相关性类因子。

图表2：日频量价相关性因子的构建方法



资料来源：国盛证券研究所整理

对于上述因子，我们通过进一步筛选，保留有效且相关性较低的因子，得到“日频量价相关性因子簇”（若无特别说明，本文在批量构建因子后，均通过下述操作进行因子筛选）：

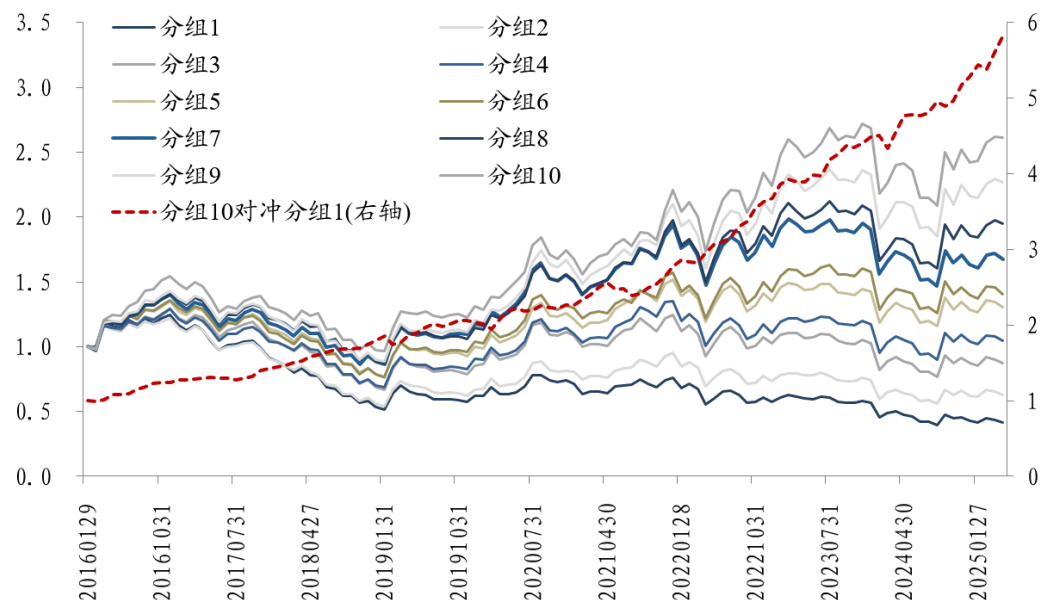
（1）在时间段 2016/01/01-2018/12/31 内，综合考虑因子的 IC、ICIR、10 分组多头超额收益、10 分组多头超额收益的信息比率以及因子互相之间的相关性，挑选效果较优且相关性较低的因子；

（2）对于挑选出的因子，再考虑它们 2019/01/01-2021/12/31 的 IC、ICIR、10 分组多头超额收益、10 分组多头超额收益的信息比率以及因子互相之间的相关性，进一步挑选其中效果较优且相关性较低的因子，称之为基于日频数据构建的量价相关性因子簇（2022 年开始即为样本外）。

随后，我们将因子簇中保留的因子各自横截面标准化、进行等权合成，得到综合因子，为了方便表述，将其简称为“日频连续”因子。

图表 3 展示了“日频连续”因子的 10 分组及多空对冲净值走势。回溯时间段 2016/01/01-2025/04/30 内，以全体 A 股为研究样本，该因子的月度 IC 均值为 0.047，年化 ICIR 为 2.36，月度 RankIC 均值为 0.078，年化 RankICIR 为 3.45；10 分组多空对冲的年化收益为 20.99%，年化波动为 8.11%，信息比率为 2.59，月度胜率为 77.48%，最大回撤为 6.76%。

图表3：“日频连续”因子 10 分组及多空对冲净值

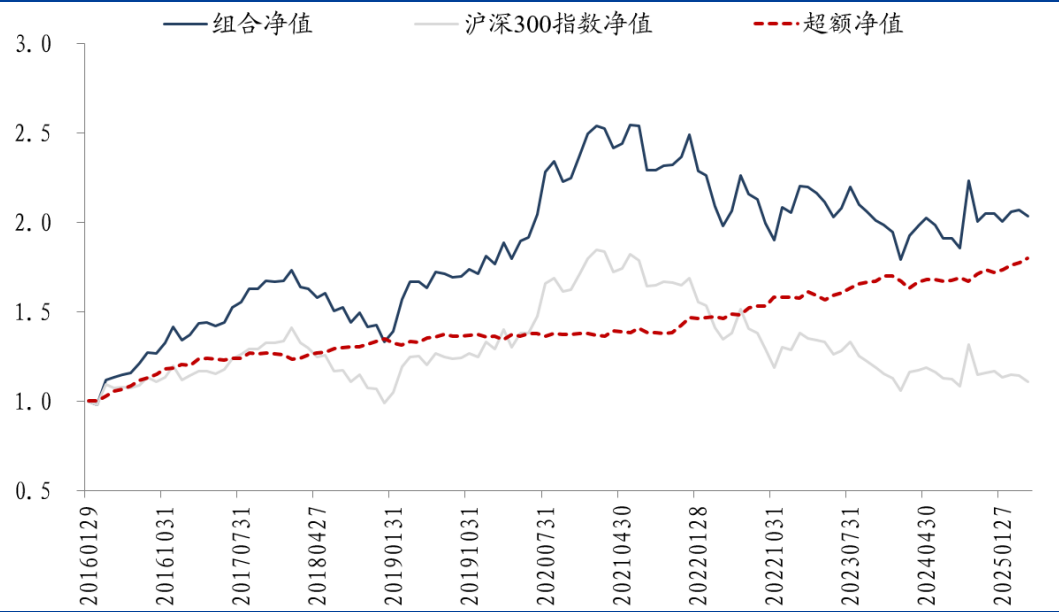


资料来源：Wind，国盛证券研究所

为了后文对比不同因子在构建组合层面的效果，我们先用“日频连续”因子构建一个基础版的股票组合，具体以沪深 300 指数增强组合为例。若无特别说明，本文构建的沪深 300 指数增强组合，调仓频率均为月度，约束条件统一设定为：（1）指数成分股权重不低于 85%；（2）中信一级行业权重偏离不超过 1%；（3）个股权重偏离不超过 0.5%。

图表 4 展示了基于“日频连续”因子构建的沪深 300 指数增强组合的净值走势。2016/01/01-2025/04/30，该组合的年化收益为 7.96%，相对沪深 300 指数的超额年化收益为 6.55%，跟踪误差为 4.17%，信息比率为 1.57，月度胜率为 65.77%，最大回撤为 4.09%。

图表4: “日频连续”因子的沪深 300 指数增强组合净值

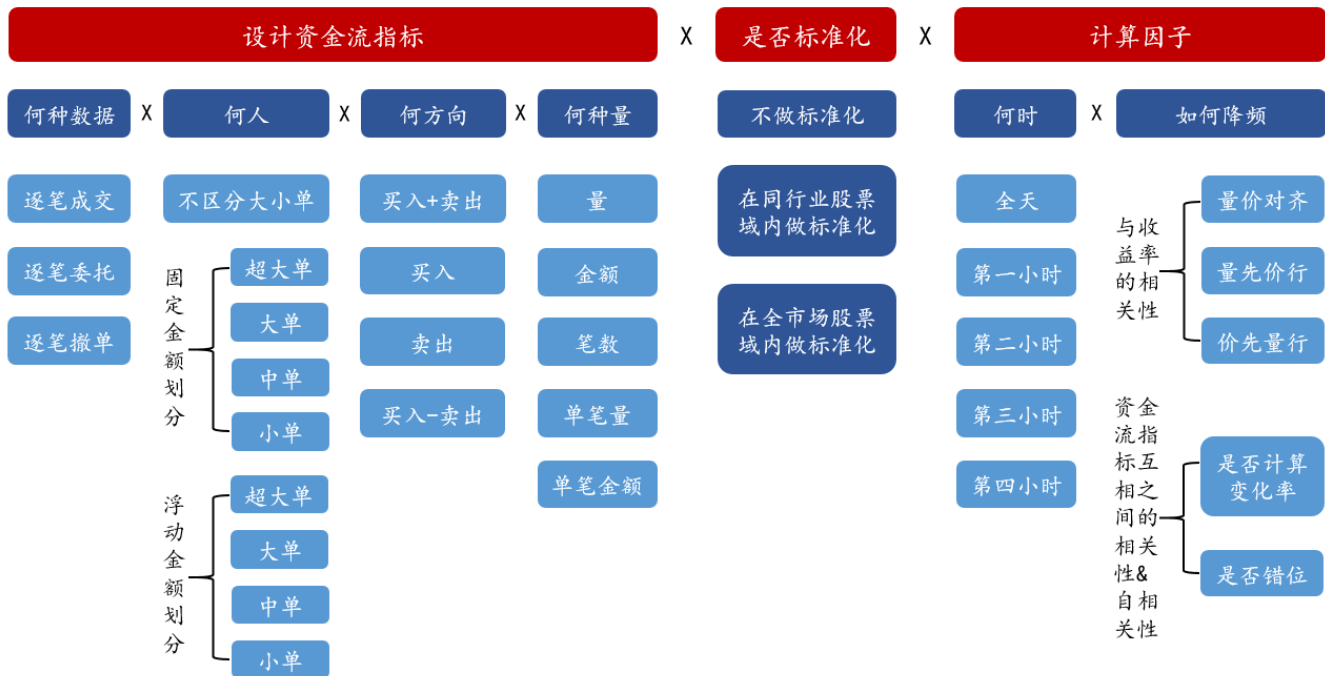


资料来源: Wind, 国盛证券研究所

2.2 分钟数据+连续性构建

本小节内容,我们基于分钟行情数据,用连续性的信息提取方式,构建量价相关性因子簇。因子构建的具体步骤,可参考本系列研究的前述报告《“量价淘金”选股因子系列研究(十):订单簿资金流因子簇的构建与生产加速》,此处只做简要叙述。

图表5: 分钟连续量价相关性因子的构建方法



资料来源: 国盛证券研究所整理

如图表 5 所示,我们把因子的构建流程分为 3 个步骤:

(1) 设计资金流指标: 从数据来源、交易者结构、买卖方向等多个角度出发,提取订单簿数据中不同维度的信息,将逐笔数据降频为 1 分钟级别的资金流指标;

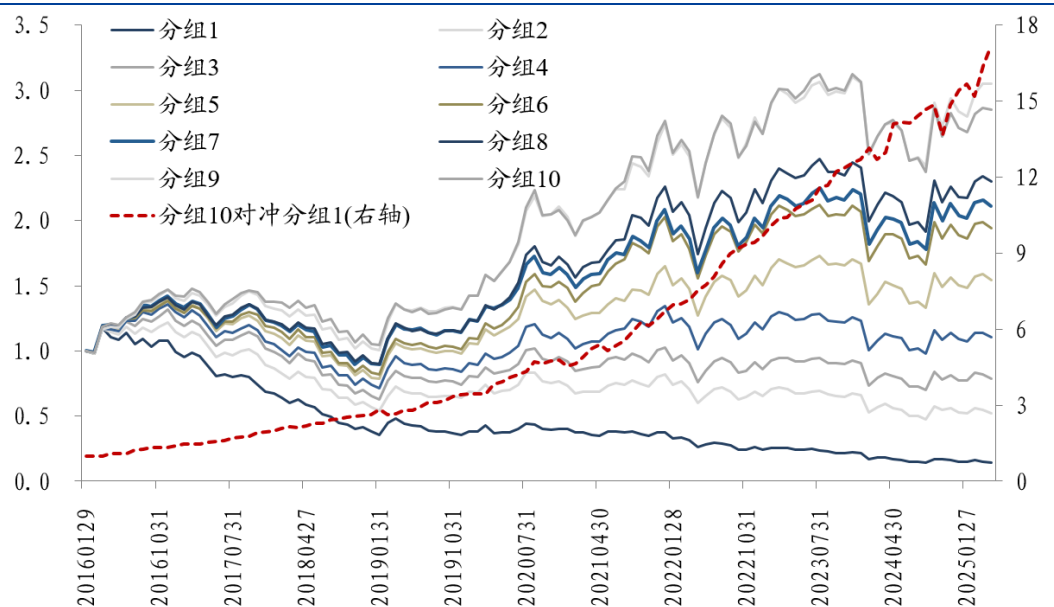
(2) 是否标准化: 可以不对资金流指标做标准化, 也可以将个股的资金流指标在同行业股票域内或者全市场股票域内进行截面标准化处理、剔除行业或者市场因素的影响;

(3) 计算因子: 观测全天、或者聚焦日内不同时间段, 计算资金流与收益率的相关系数、不同资金流指标互相之间的相关系数、单一资金流指标的自相关系数等, 将 1 分钟资金流指标降频为日度因子值。

批量生产因子之后, 与 2.1 节中的因子筛选方法一致, 我们保留其中有效、相关性相对较低的因子, 得到“分钟连续量价相关性因子簇”, 并将因子簇中的因子各自横截面标准化、等权合成, 得到综合因子, 记为“分钟连续”因子。

图表 6 展示了“分钟连续”因子的 10 分组及多空对冲净值走势。回测时间段 2016/01/01-2025/04/30 内, 以全体 A 股为研究样本, 该因子的月度 IC 均值为 0.079, 年化 ICIR 为 3.57, 月度 RankIC 均值为 0.094, 年化 RankICIR 为 4.28; 10 分组多空对冲的年化收益为 35.99%, 年化波动为 10.79%, 信息比率为 3.34, 月度胜率为 88.29%, 最大回撤为 8.10%。

图表6: “分钟连续”因子 10 分组及多空对冲净值



资料来源: Wind, 通联数据, 国盛证券研究所

2.3 逐笔数据+连续性构建

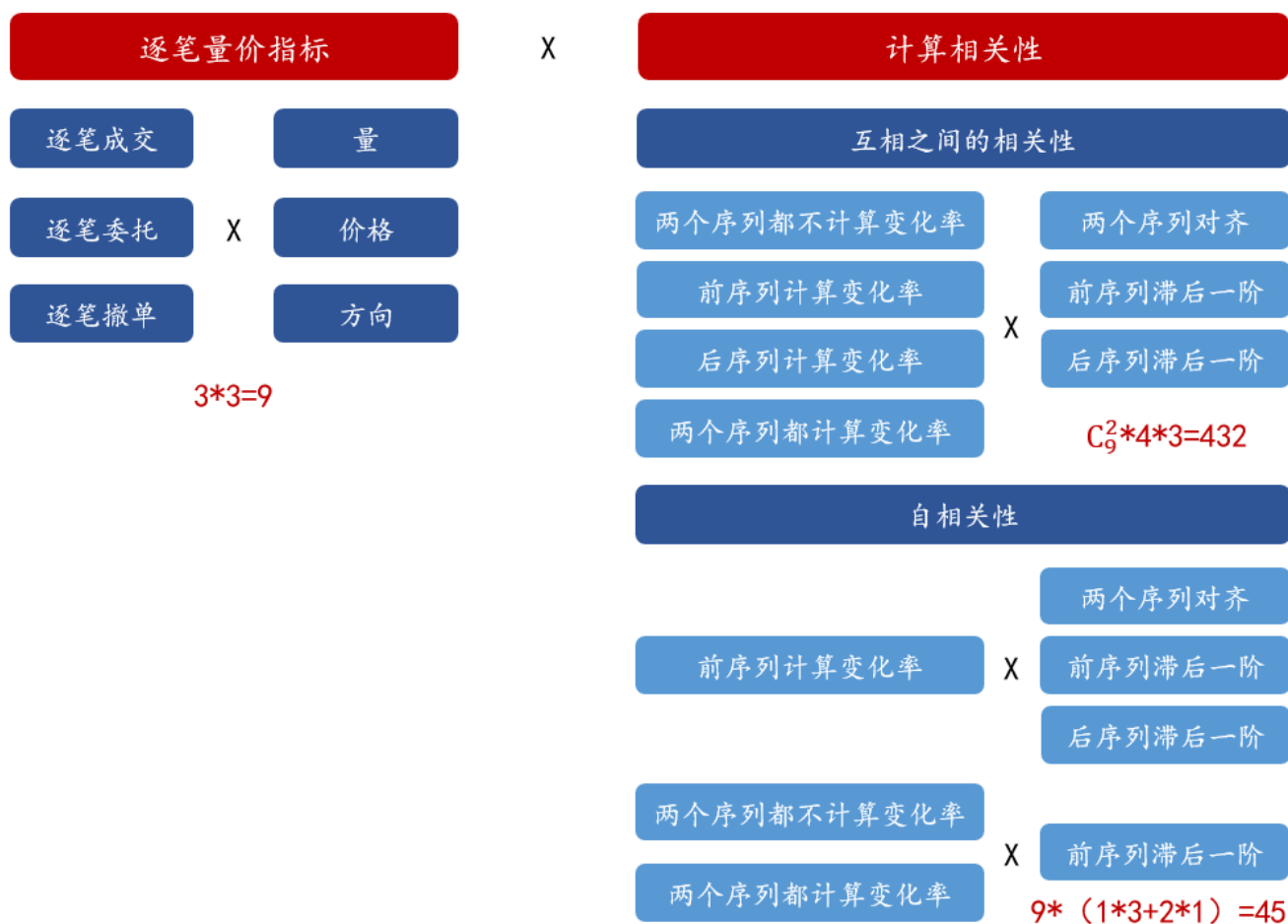
在逐笔数据上、利用连续性的信息提取方式构建因子, 能够生产的因子数量相对较少, 因为在“量价指标设计”环节, 我们无法仿照 2.2 节的内容, 将逐笔数据聚合到分钟级别、构建多种多样的分钟特征。

如图表 7 所示, 本小节构建逐笔连续量价相关性因子的步骤如下:

(1) 设计逐笔量价指标: 直接采用逐笔成交、逐笔委托、逐笔撤单的原始数据, 可以利用的指标为量、价格、方向, 因此共有 9 种量价指标。

(2) 计算相关性: 可分为上述指标两两之间的相关性、或者单一指标的自相关性。对于两两指标互相之间的相关性, 根据指标是否计算变化率、计算相关性时是否错位, 可分为多种情况, 共可计算 432 个因子; 对于单一指标的自相关性, 也可通过类似的分类讨论, 构建 45 个因子。因此基于逐笔数据+连续性的构建方式, 我们一共得到了 477 个量价相关性因子。

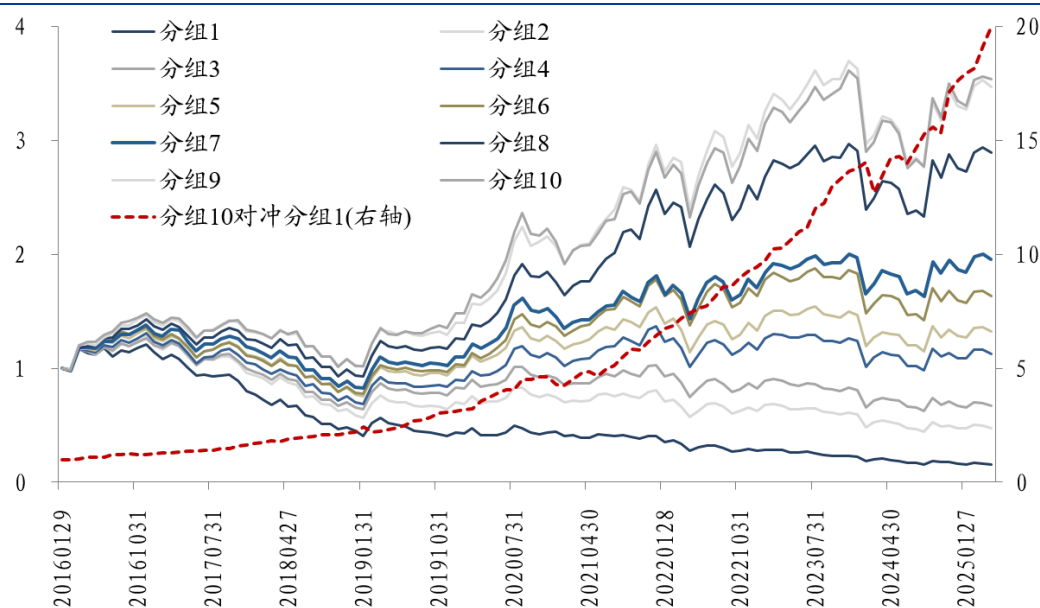
图表7: 逐笔连续量价相关性因子的构建方法



资料来源: 国盛证券研究所整理

批量生产上述因子后, 保留其中有效且相关性相对较低的因子, 得到“逐笔连续量价相关性因子簇”。将因子簇中的因子各自横截面标准化、等权合成, 得到综合因子, 记为“逐笔连续”因子。

图表8: “逐笔连续”因子 10 分组及多空对冲净值



资料来源: Wind, 通联数据, 国盛证券研究所

图表 8 展示了“逐笔连续”因子的 10 分组及多空对冲净值走势。回测时间段 2016/01/01-2025/04/30 内，以全体 A 股为研究样本，该因子的月度 IC 均值为 0.075，年化 ICIR 为 3.20，月度 RankIC 均值为 0.100，年化 RankICIR 为 3.74；10 分组多空对冲的年化收益为 38.22%，年化波动为 11.63%，信息比率为 3.29，月度胜率为 86.49%，最大回撤为 9.29%。

2.4 “分钟连续”、“逐笔连续”因子的增量

前文基于不同频率的数据，构建了“日频连续”、“分钟连续”、“逐笔连续”因子，本小节内容考察“分钟连续”、“逐笔连续”因子相对于“日频连续”因子的增量。

首先，我们将“分钟连续”、“逐笔连续”因子分别或者一起叠加到“日频连续”因子上，考察叠加后的因子回测效果。具体操作如下：

(1) 将“日频连续”、“分钟连续”因子各自横截面标准化、等权合成，得到“日频连续+分钟连续”因子；

(2) 将“日频连续”、“逐笔连续”因子各自横截面标准化、等权合成，得到“日频连续+逐笔连续”因子；

(3) 将“日频连续”、“分钟连续”、“逐笔连续”因子各自横截面标准化、等权合成，得到“日频连续+分钟连续+逐笔连续”因子。

图表 9 汇总展示了上述连续性因子的回测表现。我们可以得到以下结论：

(1) 对于单个因子而言，“分钟连续”因子的稳定性最强，年化 RankICIR 可以达到 4.28，全市场 10 分组多空对冲的信息比率为 3.34、月度胜率接近 90%；

(2) 对于叠加到“日频连续”因子上的增量效果来说，只叠加“分钟连续”或者“逐笔连续”因子的效果相差不大，相对于单个“日频连续”因子均有显著提升；

(3) 若将“分钟连续”、“逐笔连续”同时叠加到“日频连续”之上，则因子的效果可得到进一步提升，合成因子的月度 IC 均值为 0.087，年化 ICIR 为 3.40，月度 RankIC 均值为 0.117，年化 RankICIR 为 4.29；全市场 10 分组多空对冲的年化收益为 43.08%，年化波动为 12.12%，信息比率为 3.56，月度胜率为 88.29%，最大回撤为 9.11%。

图表9: 连续性因子的 IC 及多空绩效

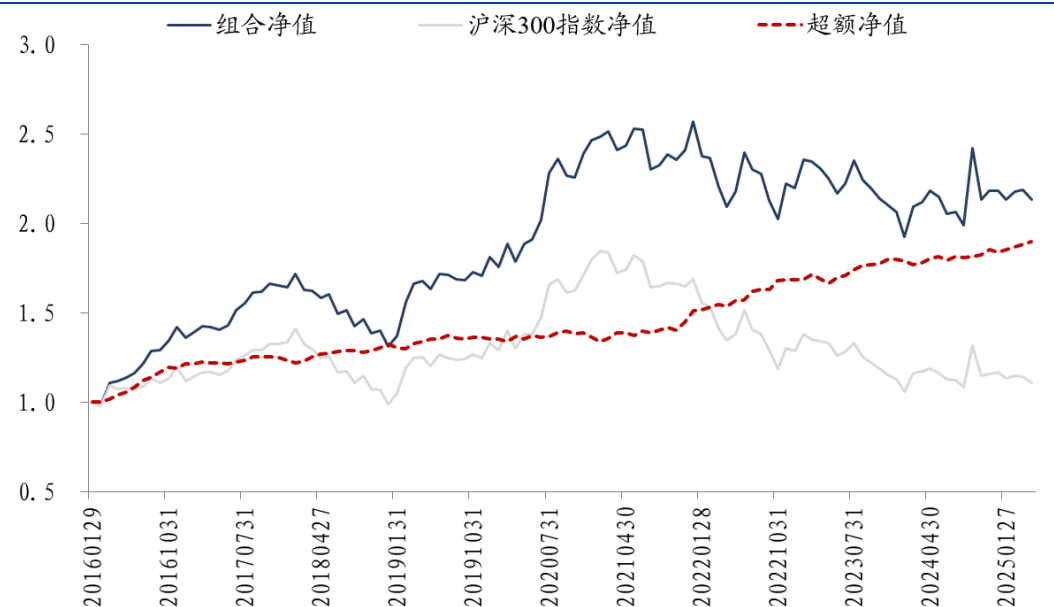
		日频连续	分钟连续	逐笔连续	日频连续+分钟连续	日频连续+逐笔连续	日频连续+分钟连续+逐笔连续
IC 信息	月度 IC 均值	0.047	0.079	0.075	0.078	0.075	0.087
	年化 ICIR	2.36	3.57	3.20	3.27	3.09	3.40
	月度 RankIC 均值	0.078	0.094	0.100	0.107	0.108	0.117
	年化 RankICIR	3.45	4.28	3.74	4.24	3.97	4.29
10 分组多空对冲绩效指标	年化收益率	20.99%	35.99%	38.22%	37.90%	38.38%	43.08%
	年化波动率	8.11%	10.79%	11.63%	11.53%	11.28%	12.12%
	信息比率	2.59	3.34	3.29	3.29	3.40	3.56
	月度胜率	77.48%	88.29%	86.49%	84.68%	86.49%	88.29%
	最大回撤率	6.76%	8.10%	9.29%	8.28%	8.80%	9.11%

资料来源: Wind, 通联数据, 国盛证券研究所

接着，我们分别基于“日频连续+分钟连续”、“日频连续+逐笔连续”、“日频连续+分钟连续+逐笔连续”因子，构建沪深 300 指数增强组合，并与 2.1 节中基于“日频连续”因子构建的组合进行对比。沪深 300 指数增强组合的调仓频率为月度，约束条件与 2.1 节中的内容一致：（1）指数成分股权重不低于 85%；（2）中信一级行业权重偏离不超过 1%；（3）个股权重偏离不超过 0.5%。

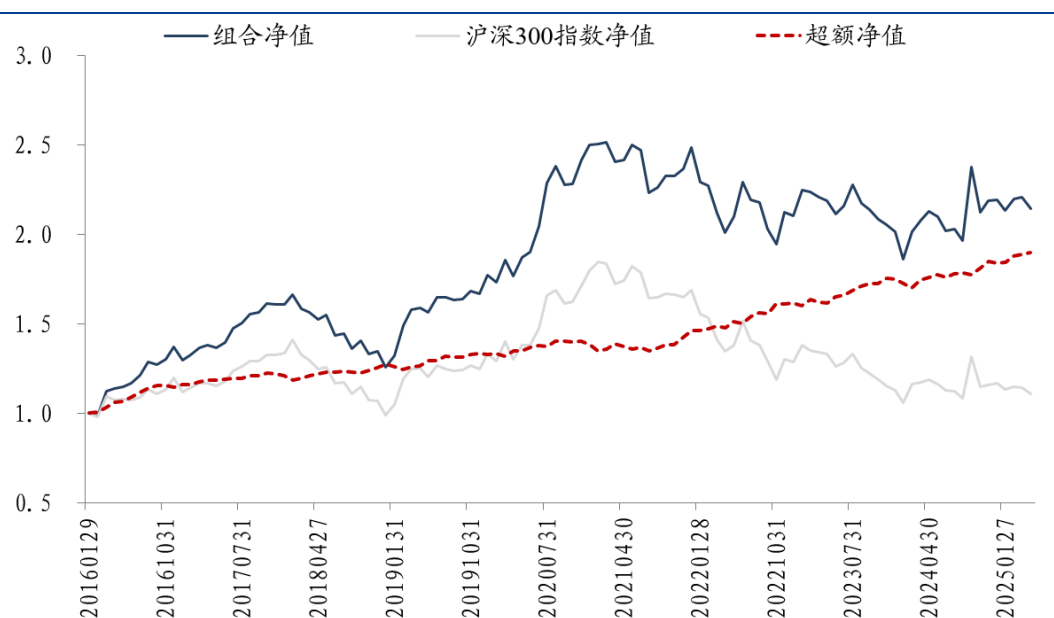
图表 10 至图表 12 分别展示了“日频连续+分钟连续”、“日频连续+逐笔连续”、“日频连续+分钟连续+逐笔连续”因子沪深 300 增强组合的净值，图表 13 则汇总了它们超额收益的各项绩效指标。

图表10: “日频连续+分钟连续”因子的沪深 300 指数增强组合净值



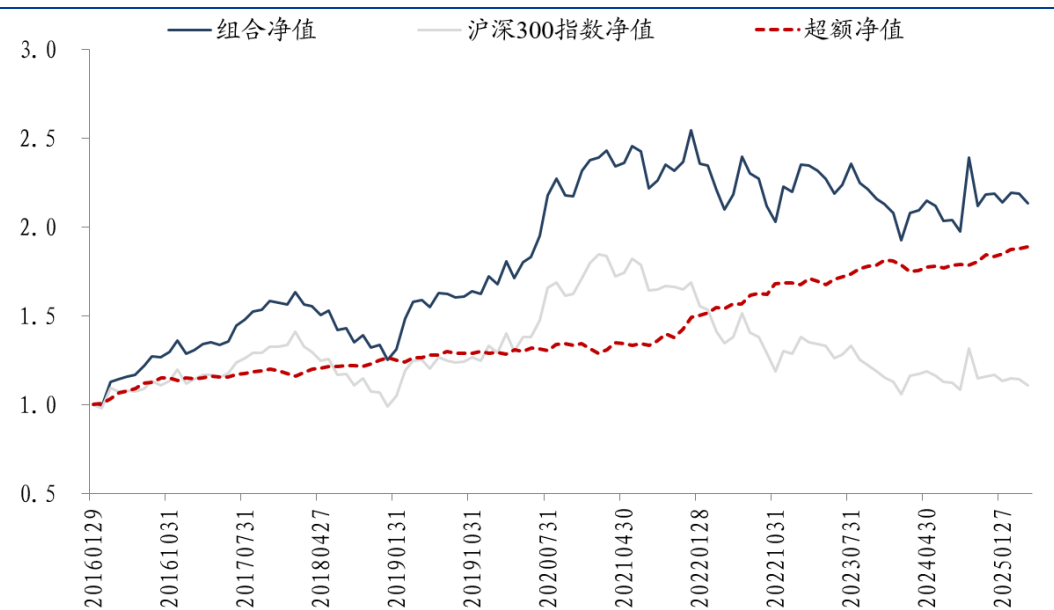
资料来源: Wind, 通联数据, 国盛证券研究所

图表11: “日频连续+逐笔连续”因子的沪深 300 指数增强组合净值



资料来源: Wind, 通联数据, 国盛证券研究所

图表12: “日频连续+分钟连续+逐笔连续”因子的沪深 300 指数增强组合净值



资料来源: Wind, 通联数据, 国盛证券研究所

图表13: 连续性因子的沪深 300 指数增强组合超额绩效

	日频连续	日频连续+分钟连续	日频连续+逐笔连续	日频连续+分钟连续+逐笔连续
年化收益率	6.55%	7.15%	7.16%	7.10%
年化波动率	4.17%	4.17%	4.15%	4.38%
信息比率	1.57	1.72	1.73	1.62
月度胜率	65.77%	67.57%	67.57%	67.57%
最大回撤率	4.09%	4.24%	4.02%	4.33%

资料来源: Wind, 通联数据, 国盛证券研究所

可以发现, 虽然从前文的检验结果来看, 将“分钟连续”、“逐笔连续”叠加到“日频连续”上, 因子的回测表现确有显著提升, 但一旦落实到组合层面, “分钟连续”、“逐笔连续”因子能够提供的增量信息就变得非常有限。图表 13 显示, 基于“日频连续+分钟连续”、“日频连续+逐笔连续”、“日频连续+分钟连续+逐笔连续”因子构建的沪深 300 指数增强组合, 相对于仅用“日频连续”构建的组合, 年化超额收益只能提升 0.5%-0.6% 左右。

三、离散化因子的构建方式

从本节内容开始，我们讨论离散化的因子构建方式。“离散化”，即在过去一段时间内，基于某些特征、抽取几个离散的时间点，只用与这些时间点相关的数据计算因子。本文构建的因子以“量价相关性类因子”为例，如前文 2.1 节所述，由于日频行情序列的数据点较少，若再进行离散化提取，则无法准确计算相关系数。因此，本节内容只对分钟数据、逐笔数据进行离散化处理，构建量价相关性因子。

3.1 分钟数据+离散化构建

如何在分钟数据层面构建离散化的因子，我们在本系列研究的前述报告《“量价淘金”选股因子系列研究（九）：Memory Map 在因子生产加速上的应用——以构建羊群效应因子簇为例》中已经做过讨论。该报告围绕“羊群效应”话题，提炼因子构建流程中的 3 个关键因素，分别进行了详细论述：

（1）事件识别：如何识别趋势资金的行动。通过观测成交量、价格涨跌幅、价格波动率、量价相关性是否发生异常，得到了上千种不同的识别方式。

（2）因子定义：如何刻画羊群效应的强弱。在“趋势资金有所行动”这一事件发生后，通过计算成交量、价格涨跌幅、价格波动率、量价相关性这些量价数据的本身数值、与趋势资金对应时点量价数据的相互关系、与趋势资金事件发生之前量价数据的相互关系，得到了上千种不同的因子构造方式。

（3）数据频率：在何种数据频率上识别事件、定义因子。可用的数据集包括日频数据、分钟数据、逐笔数据、用逐笔数据合成的其他不同频率的数据集等等，我们可以在不同数据维度上识别趋势资金的行动、考察跟随者的交易行为。

如图表 14 所示，本节内容构建的分钟数据离散化因子，大致思路与系列报告（九）类似，也是将因子构建的流程分为事件识别、因子定义，在用逐笔数据降频得到的 1 分钟数据集上，进行因子的批量生产。其中，“事件识别”环节的内容，与系列报告（九）完全一致；“因子定义”环节，我们只构建量价相关性类因子。

图表14: 分钟离散量价相关性因子的构建方法

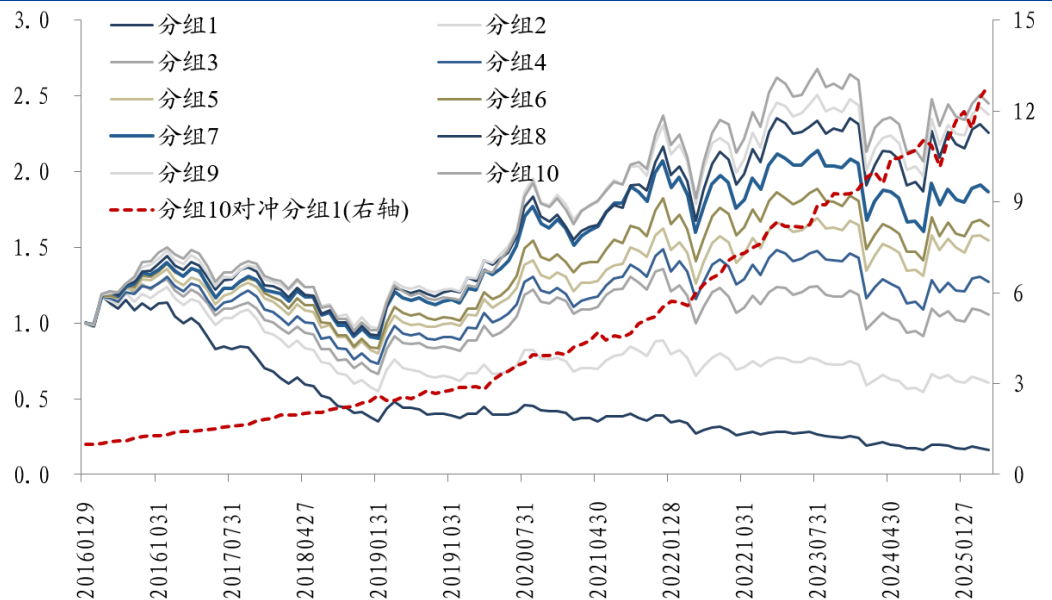


资料来源：国盛证券研究所整理

批量生产上述因子后，我们进行与前文一致的有效性、相关性筛选，留下一簇有效且相关性相对较低的因子，称之为“分钟离散量价相关性因子簇”。将因子簇中的因子各自横截面标准化、等权合成，得到综合因子，记为“分钟离散”因子。

图表 15 展示了“分钟离散”因子的 10 分组及多空对冲净值走势。回测时间段 2016/01/01-2025/04/30 内，以全体 A 股为研究样本，该因子的月度 IC 均值为 0.067，年化 ICIR 为 2.81，月度 RankIC 均值为 0.093，年化 RankICIR 为 3.76；10 分组多空对冲的年化收益为 31.77%，年化波动为 11.16%，信息比率为 2.85，月度胜率为 78.38%，最大回撤为 7.84%。

图表15: “分钟离散”因子 10 分组及多空对冲净值



资料来源: Wind, 通联数据, 国盛证券研究所

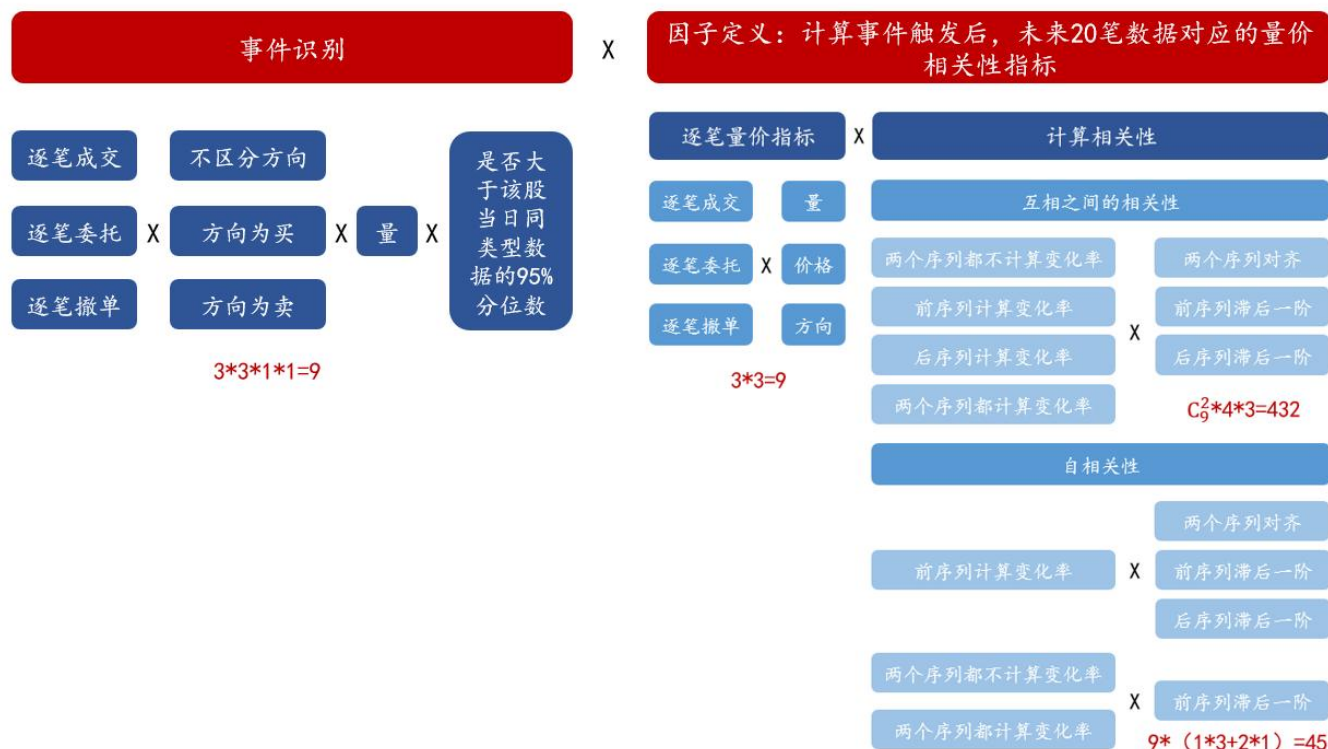
3.2 逐笔数据+离散化构建

在逐笔数据层面进行离散化信息提取、构建因子的逻辑，与分钟层面类似，仍然是先识别量价数据发生异常的事件点，再用事件点附近的数据参与因子计算。但由于逐笔数据的处理耗时相对较长，因此关于事件的定义，我们只选取了“放量”，对于因子的计算，也暂时只考虑事件发生后、量价相关性的本身数值。

具体如图表 16 所示，本小节构建逐笔离散量价相关性因子的步骤如下：

- (1) 事件识别：将“逐笔成交、逐笔委托、逐笔撤单”对应的“整体的量、买的量、卖的量”，超过该股当日同类型数据 95%分位数的时间点，定义为事件发生，因此共有 9 种不同的事件识别方法。
- (2) 因子定义：计算事件触发后，未来 20 笔数据对应的量价相关性指标。此部分内容与 2.3 节“逐笔连续量价相关性因子”的构建方式完全类似，共有 477 种不同的因子计算方法。
- (3) 将上述事件识别与因子定义的方法两两搭配，一共构建了 4293 个“逐笔离散量价相关性因子”。

图表16: 逐笔离散量价相关性因子的构建方法

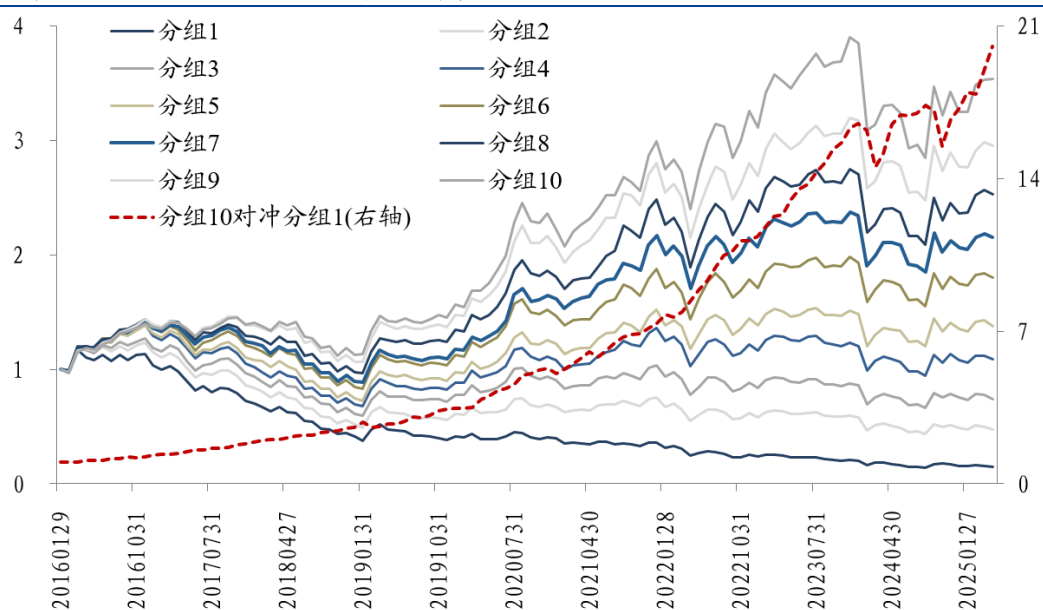


资料来源：国盛证券研究所整理

得到上述因子后，综合考察因子的有效性、相关性，筛选得到“逐笔离散量价相关性因子簇”。将因子簇中的因子各自横截面标准化、等权合成，记为“逐笔离散”因子。

图表 17 展示了“逐笔离散”因子的 10 分组及多空对冲净值走势。回溯时间段 2016/01/01-2025/04/30 内，以全体 A 股为研究样本，该因子的月度 IC 均值为 0.073，年化 ICIR 为 2.87，月度 RankIC 均值为 0.099，年化 RankICIR 为 3.37；10 分组多空对冲的年化收益为 38.30%，年化波动为 11.88%，信息比率为 3.22，月度胜率为 81.98%，最大回撤为 11.87%。

图表17: “逐笔离散”因子 10 分组及多空对冲净值



资料来源：Wind，通联数据，国盛证券研究所

3.3 “分钟离散”、“逐笔离散”因子的增量

前文构建了“分钟离散”、“逐笔离散”因子，本小节内容考察它们相对于第二节内容中连续性因子的增量。为了方便表述，我们将第二节中构建的“日频连续+分钟连续+逐笔连续”因子简称为“连续”因子，其蕴含了前文所有连续性因子的综合信息。

首先，我们将“分钟离散”、“逐笔离散”因子分别或者一起叠加到“连续”因子上，考察叠加后的因子回测效果。具体操作如下：

- (1) 将“连续”、“分钟离散”因子各自横截面标准化、等权合成，得到“连续+分钟离散”因子；
- (2) 将“连续”、“逐笔离散”因子各自横截面标准化、等权合成，得到“连续+逐笔离散”因子；
- (3) 将“连续”、“分钟离散”、“逐笔离散”因子各自横截面标准化、等权合成，得到“连续+分钟离散+逐笔离散”因子。

图表18: “连续”、“离散”、“连续+离散”因子的 IC 及多空绩效

		连续	分钟离散	逐笔离散	连续+分钟离散	连续+逐笔离散	连续+分钟离散+逐笔离散
IC 信息	月度 IC 均值	0.087	0.067	0.073	0.088	0.090	0.093
	年化 ICIR	3.40	2.81	2.87	3.27	3.25	3.28
	月度 RankIC 均值	0.117	0.093	0.099	0.119	0.119	0.124
	年化 RankICIR	4.29	3.76	3.37	4.27	3.93	4.11
10 分组多空对冲绩效指标	年化收益率	43.08%	31.77%	38.30%	44.96%	45.58%	46.56%
	年化波动率	12.12%	11.16%	11.88%	12.65%	12.52%	13.24%
	信息比率	3.56	2.85	3.22	3.56	3.64	3.52
	月度胜率	88.29%	78.38%	81.98%	82.88%	86.49%	83.78%
	最大回撤率	9.11%	7.84%	11.87%	8.92%	9.39%	10.27%

资料来源: Wind, 通联数据, 国盛证券研究所

图表 18 展示了“连续”、“离散”、“连续+离散”因子的回测表现，可以得到以下结论：

(1) 对于单个因子而言，本节内容构建的“分钟离散”、“逐笔离散”因子，回测表现均弱于前文的“连续”因子；

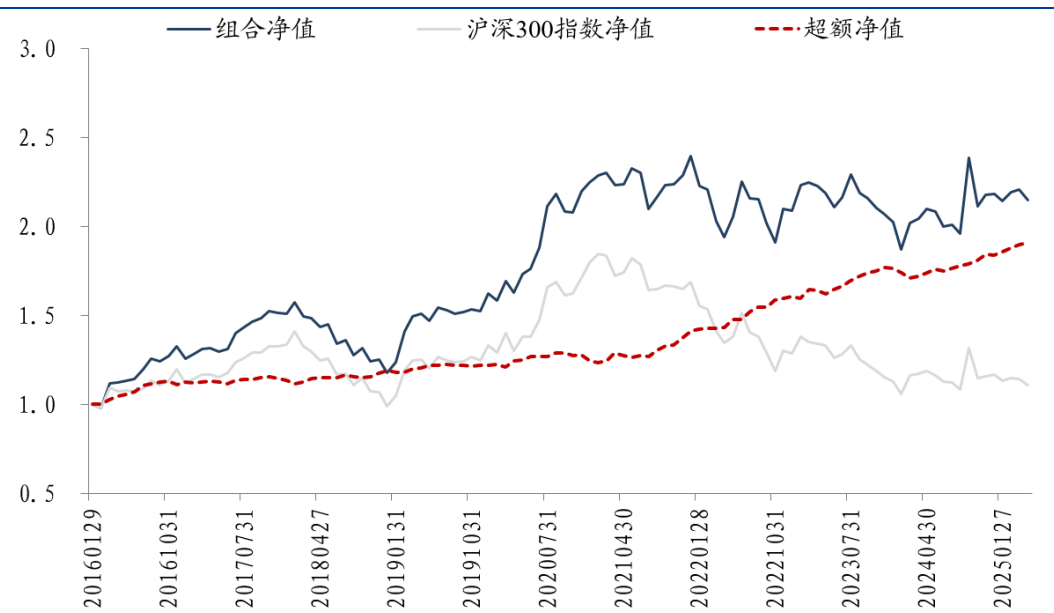
(2) 将“分钟离散”、“逐笔离散”因子叠加到“连续”因子上，因子的 IC、RankIC 略有提升，但 ICIR、RankICIR、10 分组回测的绩效指标整体相差不大。

如此看来，我们构建的离散化因子，能够提供的增量信息似乎非常有限。但在后文的指增组合效果检验中，我们将会看到不一样的景象。

分别基于“连续+分钟离散”、“连续+逐笔离散”、“连续+分钟离散+逐笔离散”因子，构建沪深 300 指数增强组合，并与基于“连续”因子构建的组合进行对比。沪深 300 指数增强组合的调仓频率为月度，约束条件与前文一致：(1) 指数成分股权重不低于 85%；(2) 中信一级行业权重偏离不超过 1%；(3) 个股权重偏离不超过 0.5%。

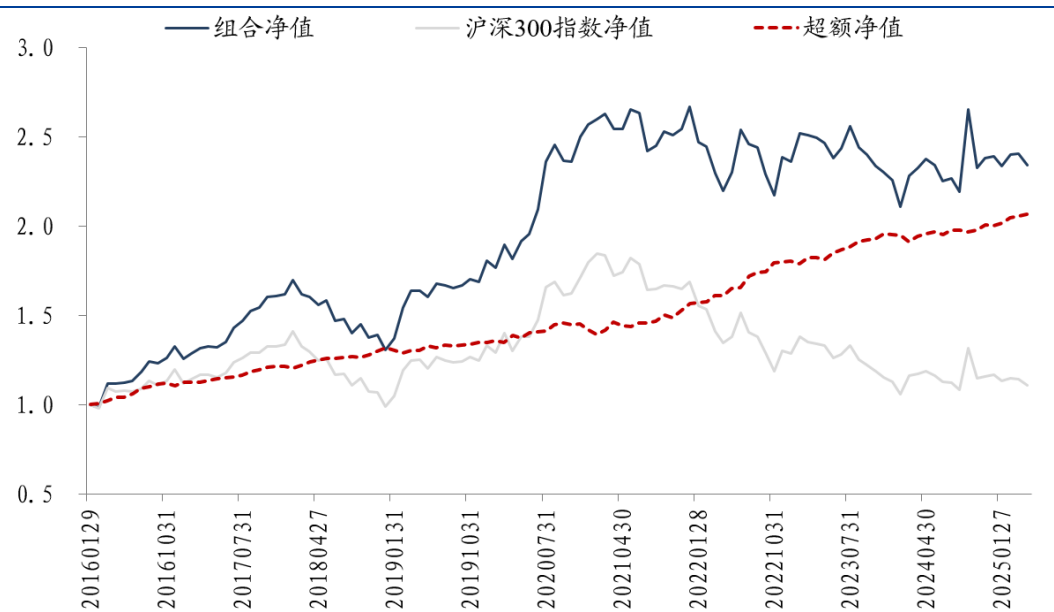
图表 19 至图表 21 分别展示了“连续+分钟离散”、“连续+逐笔离散”、“连续+分钟离散+逐笔离散”因子沪深 300 增强组合的净值，图表 22 则汇总了它们超额收益的各项绩效指标。

图表19: “连续+分钟离散”因子的沪深300指数增强组合净值



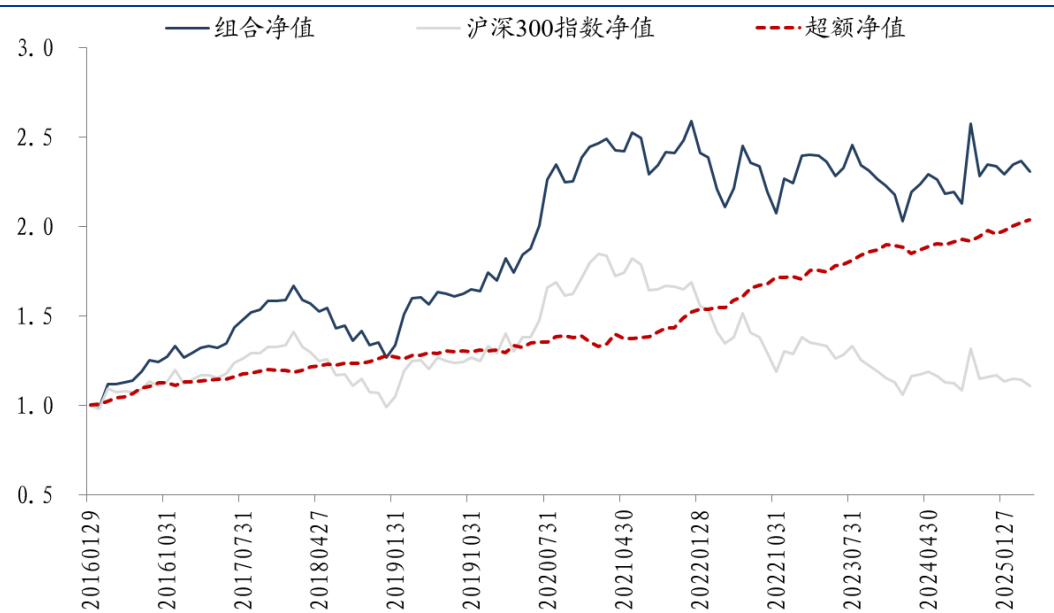
资料来源: Wind, 通联数据, 国盛证券研究所

图表20: “连续+逐笔离散”因子的沪深300指数增强组合净值



资料来源: Wind, 通联数据, 国盛证券研究所

图表21: “连续+分钟离散+逐笔离散”因子的沪深 300 指数增强组合净值



资料来源: Wind, 通联数据, 国盛证券研究所

图表22: “连续”、“连续+离散”因子的沪深 300 指数增强组合超额绩效

	连续	连续+ 分钟离散	连续+ 逐笔离散	连续+分钟离散 +逐笔离散
年化收益率	7.10%	7.22%	8.15%	7.99%
年化波动率	4.38%	4.22%	3.81%	3.89%
信息比率	1.62	1.71	2.14	2.05
月度胜率	67.57%	69.37%	75.68%	70.27%
最大回撤率	4.33%	4.38%	4.27%	4.32%

资料来源: Wind, 通联数据, 国盛证券研究所

可以看到, 离散化因子的检验结果, 与第二节内容中连续性因子的检验结果完全相反。虽然从图表 18 来看, 将“分钟离散”、“逐笔离散”叠加到“连续”因子上, 因子的回测表现并没有得到明显提升, 但落实到组合层面, 离散化因子却能提供增量。根据图表 22, 基于“连续+分钟离散”、“连续+逐笔离散”、“连续+分钟离散+逐笔离散”因子构建的沪深 300 指数增强组合, 相对于仅用“连续”因子构建的组合, 超额年化收益、信息比率都能得到提升。其中, “连续+逐笔离散”因子构建的组合绩效表现最强, 相对基准指数的超额年化收益为 8.15%, 跟踪误差为 3.81%, 信息比率为 2.14, 月度胜率为 75.68%, 最大回撤为 4.27%; 相较于只用“连续”因子构建的组合, 超额年化收益的提升超过 1%。

上述测试结果有力地论证了本文开篇提出的观点, 即用离散化方式构建的选股因子, 尤其是在逐笔数据层面的离散化, 即使从因子回测效果上、未必能显著优于传统的连续性因子, 但落实到组合层面, 往往能够提供连续性因子以外的增量信息, 提升组合的收益与稳定性, 由此可见“高频数据+离散化构建方式”在因子研究中的重要性。

四、其他重要讨论

由于本篇报告重点想要强调的是“逐笔数据+离散化构建方式”，所以本节内容针对前文的“逐笔离散”因子，再做进一步补充讨论。

4.1 “逐笔离散”因子的分年度表现

3.2 节内容只展示了“逐笔离散”因子在全时间段的回测表现，本小节补充展示其分年度的表现情况。

图表23: “逐笔离散”因子分年度表现

年份	年化收益率			分组 10 对冲分组 1 绩效指标			
	分组 1	分组 10	分组 10 对冲分组 1	年化波动率	信息比率	月度胜率	最大回撤率
2016	40.82%	3.82%	33.26%	11.98%	2.78	63.64%	2.09%
2017	2.67%	-32.06%	47.60%	8.28%	5.75	83.33%	1.23%
2018	-19.61%	-41.07%	34.23%	5.19%	6.60	91.67%	0.07%
2019	38.48%	-0.48%	33.67%	13.88%	2.43	75.00%	8.15%
2020	42.25%	-4.05%	46.68%	13.94%	3.35	91.67%	4.22%
2021	34.58%	-8.79%	46.15%	9.32%	4.95	91.67%	4.00%
2022	4.03%	-34.05%	53.90%	8.78%	6.14	83.33%	1.26%
2023	23.50%	-15.53%	45.32%	5.32%	8.52	100.00%	0.00%
2024	-15.53%	-21.21%	4.42%	20.06%	0.22	58.33%	11.87%
2025(至 4 月底)	29.49%	-19.93%	57.73%	10.62%	5.43	75.00%	0.48%

资料来源: Wind, 通联数据, 国盛证券研究所

4.2 “逐笔离散”因子剔除常见风格与行业后的表现

首先考察“逐笔离散”因子与市场常用风格因子的相关性。图表 24 展示了“逐笔离散”因子与 10 个 Barra 风格因子的相关系数，可以看到，该因子与波动率、流动性因子的相关性相对较高，相关系数绝对值在 0.30 左右，与其他因子的相关性较低。

图表24: “逐笔离散”因子与 Barra 风格因子相关系数

“逐笔离散”因子		“逐笔离散”因子	
Beta	-0.104	NonLinearSize	-0.054
BooktoPrice	0.109	Size	-0.084
EarningsYield	0.089	Momentum	-0.017
Growth	0.019	Liquidity	-0.308
Leverage	-0.020	ResidualVolatility	-0.286

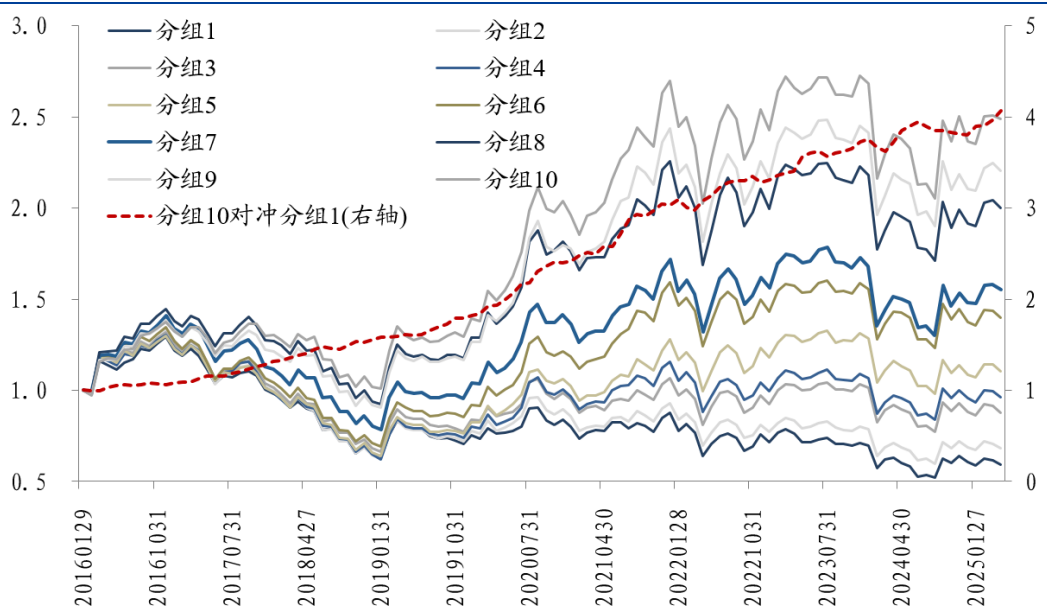
资料来源: Wind, 通联数据, 国盛证券研究所

为了剔除市场常用风格和行业的干扰，我们每月月底将“逐笔离散”因子对 Barra 风格因子和中信一级行业虚拟变量进行回归，取残差作为纯净因子，检验其选股效果。图表 25 展示了纯净因子在全体 A 股中的 10 分组及多空对冲净值走势，图表 26 则汇报了其分年度的表现情况。

剔除常用风格与行业的影响后，因子仍然有效。“纯净逐笔离散”因子的月度 IC 均值为 0.037，年化 ICIR 为 2.84，月度 RankIC 均值为 0.040，年化 RankICIR 为 2.75；10 分

组多空对冲的年化收益为 15.82%，年化波动为 6.15%，信息比率为 2.57，月度胜率为 78.38%，最大回撤为 3.69%。

图表25: “纯净逐笔离散”因子 10 分组及多空对冲净值



资料来源: Wind, 通联数据, 国盛证券研究所

图表26: “纯净逐笔离散”因子分年度表现

年份	年化收益率			分组 10 对冲分组 1 绩效指标			
	分组 1	分组 10	分组 10 对冲分组 1	年化波动率	信息比率	月度胜率	最大回撤率
2016	35.54%	24.19%	8.93%	5.35%	1.67	63.64%	1.62%
2017	-1.31%	-19.43%	21.86%	3.34%	6.55	100.00%	0.00%
2018	-21.83%	-34.19%	17.83%	5.14%	3.47	75.00%	2.00%
2019	36.71%	16.59%	16.98%	4.23%	4.01	83.33%	0.47%
2020	40.94%	6.10%	33.17%	6.81%	4.87	91.67%	0.35%
2021	37.30%	9.30%	25.38%	7.64%	3.32	75.00%	0.64%
2022	-10.06%	-18.20%	9.05%	6.41%	1.41	75.00%	3.69%
2023	10.60%	-2.85%	13.55%	5.07%	2.67	91.67%	1.55%
2024	-12.79%	-15.84%	3.38%	6.78%	0.50	58.33%	3.61%
2025(至 4 月底)	16.71%	-5.75%	23.14%	3.61%	6.42	100.00%	0.00%

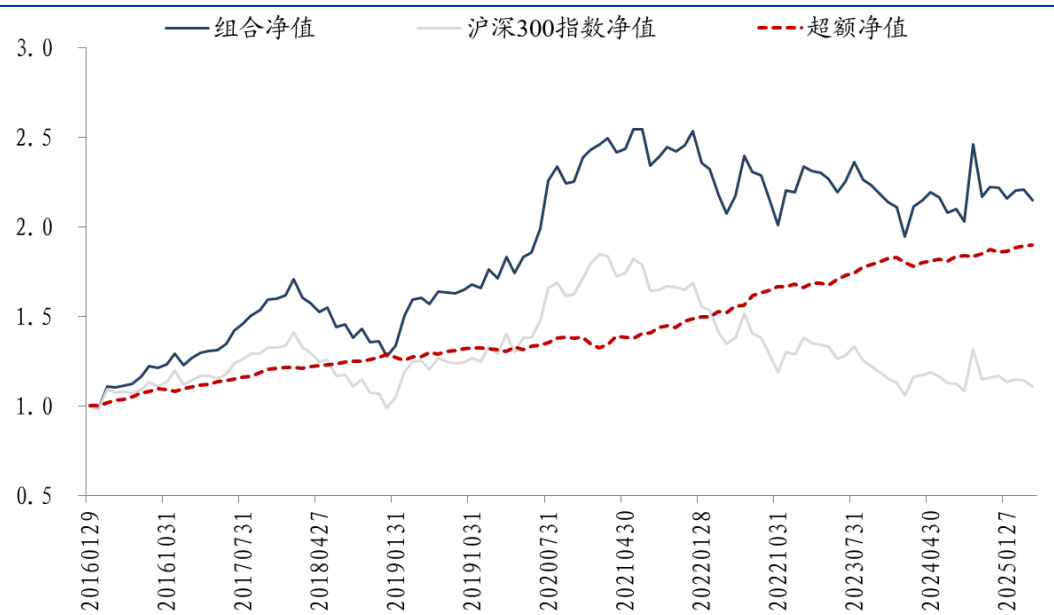
资料来源: Wind, 通联数据, 国盛证券研究所

4.3 “逐笔离散”因子的指数增强组合表现

本小节内容基于“逐笔离散”因子，构建沪深 300、中证 500、中证 1000 指数增强组合，展示其收益表现，组合的调仓频率为月度，约束条件统一设置为：（1）指数成分股权重不低于 85%；（2）中信一级行业权重偏离不超过 1%；（3）个股权重偏离不超过 0.5%。

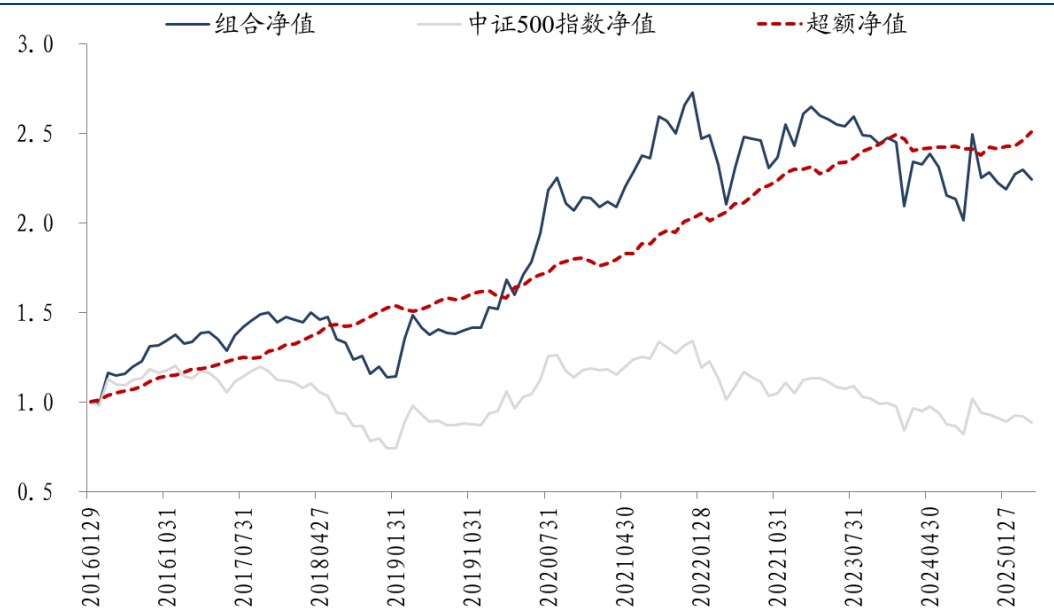
2016/01/01-2025/04/30，沪深 300 指数增强组合的年化收益为 8.61%，相对沪深 300 指数的超额年化收益为 7.17%，跟踪误差为 3.55%，信息比率为 2.02，月度胜率为 74.77%，最大回撤为 4.42%。

图表27: “逐笔离散”因子的沪深300指数增强组合净值



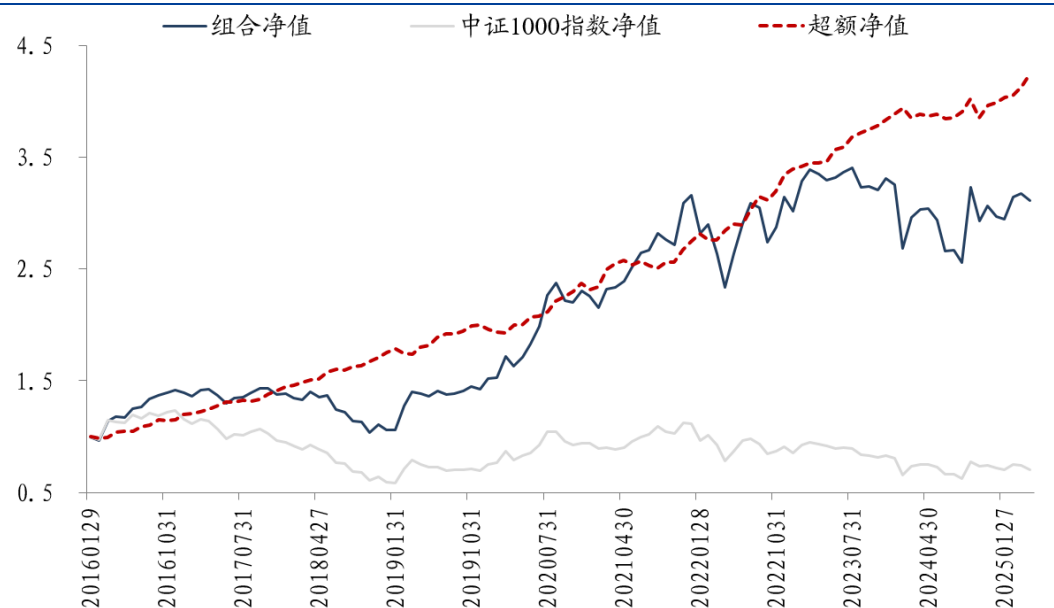
2016/01/01-2025/04/30, 中证 500 指数增强组合的年化收益为 9.12%, 相对中证 500 指数的超额年化收益为 10.43%, 跟踪误差为 3.91%, 信息比率为 2.67, 月度胜率为 81.98%, 最大回撤为 4.65%。

图表28: “逐笔离散”因子的中证500指数增强组合净值



2016/01/01-2025/04/30, 中证 1000 指数增强组合的年化收益为 13.05%, 相对中证 1000 指数的超额年化收益为 16.90%, 跟踪误差为 6.33%, 信息比率为 2.67, 月度胜率为 80.18%, 最大回撤为 4.24%。

图表29: “逐笔离散”因子的中证 1000 指数增强组合净值



资料来源: Wind, 通联数据, 国盛证券研究所

五、总结

随着市场的发展与竞争的加剧、尤其是机器学习模型的普及，想要挖掘现有多因子模型以外的增量信息越来越困难。本文提出一个重要论点：未来量价因子的研究，想要获取更多的、与现有因子低相关性的增量，“逐笔等高频数据”、“离散化的因子构建方式”这两个条件缺一不可。

本文先讨论了“连续性的因子构建方式”，即基于某一段连续的数据计算因子。以常见的“量价相关性类因子”为例，分别基于日频、分钟、逐笔数据，批量生产对应的连续性因子簇。将各自因子簇中的因子等权合成，得到“日频连续”、“分钟连续”、“逐笔连续”因子。考察“分钟连续”、“逐笔连续”因子相对于“日频连续”因子的增量，发现虽然从因子的回测效果上来看，将“分钟连续”、“逐笔连续”叠加到“日频连续”上，因子绩效确有显著提升，但一旦落实到组合层面，“分钟连续”、“逐笔连续”因子能够提供的增量信息就变得非常有限。以沪深 300 指数增强组合为例，叠加“分钟连续”、“逐笔连续”因子构建的组合，相对于仅用“日频连续”因子构建的组合，超额年化收益只能提升 0.5%-0.6% 左右。

随后，本文探索了“离散化的因子构建方式”能否带来更多的增量信息。“离散化”，即在过去一段时间内，基于某些特征、抽取几个离散的时间点，只用与这些时间点相关的数据计算因子。分别基于分钟、逐笔数据，采用离散化的构建方式，批量生产因子簇，同样将各自因子簇中的因子等权合成，得到“分钟离散”、“逐笔离散”因子。考察离散化因子相对于连续性因子的增量，发现虽然将“分钟离散”、“逐笔离散”叠加到“连续”因子上，因子的回测表现并未得到明显提升，但落实到组合层面，离散化因子却能提供显著的增量。比如基于“连续+逐笔离散”因子构建的沪深 300 指数增强组合，相较于只用“连续”因子构建的组合，超额年化收益可以提升 1% 以上。

上述对比测试的结果，有力地论证了本文的观点，即用离散化方式构建的选股因子，尤其是在逐笔数据层面的离散化，即使从因子回测效果上、未必能明显优于传统的连续性因子，但落实到组合层面，往往能够提供连续性因子以外的增量信息，提升组合的收益与稳定性，由此可见“高频数据+离散化构建方式”在因子研究中的重要性。

风险提示

以上结论均基于历史数据和统计模型的测算，如果未来市场环境发生明显改变，不排除模型失效的可能性。

免责声明

国盛证券有限责任公司（以下简称“本公司”）具有中国证监会许可的证券投资咨询业务资格。本报告仅供本公司的客户使用。本公司不会因接收人收到本报告而视其为客户。在任何情况下，本公司不对任何人因使用本报告中的任何内容所引致的任何损失负任何责任。

本报告的信息均来源于本公司认为可信的公开资料，但本公司及其研究人员对该等信息的准确性及完整性不作任何保证。本报告中的资料、意见及预测仅反映本公司于发布本报告当日的判断，可能会随时调整。在不同时期，本公司可发出与本报告所载资料、意见及推测不一致的报告。本公司不保证本报告所含信息及资料保持在最新状态，对本报告所含信息可在不发出通知的情形下做出修改，投资者应当自行关注相应的更新或修改。

本公司力求报告内容客观、公正，但本报告所载的资料、工具、意见、信息及推测只提供给客户作参考之用，不构成任何投资、法律、会计或税务的最终操作建议，本公司不就报告中的内容对最终操作建议做出任何担保。本报告中所指的投资及服务可能不适合个别客户，不构成客户私人咨询建议。投资者应当充分考虑自身特定状况，并完整理解和使用本报告内容，不应视本报告为做出投资决策的唯一因素。

投资者应注意，在法律许可的情况下，本公司及其本公司的关联机构可能会持有本报告中涉及的公司所发行的证券并进行交易，也可能为这些公司正在提供或争取提供投资银行、财务顾问和金融产品等各种金融服务。

本报告版权归“国盛证券有限责任公司”所有。未经事先本公司书面授权，任何机构或个人不得对本报告进行任何形式的发布、复制。任何机构或个人如引用、刊发本报告，需注明出处为“国盛证券研究所”，且不得对本报告进行有悖原意的删节或修改。

分析师声明

本报告署名分析师在此声明：我们具有中国证券业协会授予的证券投资咨询执业资格或相当的专业胜任能力，本报告所表述的任何观点均精准地反映了我们对标的证券和发行人的个人看法，结论不受任何第三方的授意或影响。我们所得报酬的任何部分无论是在过去、现在及将来均不会与本报告中的具体投资建议或观点有直接或间接联系。

投资评级说明

投资建议的评级标准		评级	说明
评级标准为报告发布日后的 6 个月内公司股价（或行业指数）相对同期基准指数的相对市场表现。其中 A 股市场以沪深 300 指数为基准；新三板市场以三板成指（针对协议转让标的）或三板做市指数（针对做市转让标的）为基准；香港市场以摩根士丹利中国指数为基准，美股市场以标普 500 指数或纳斯达克综合指数为基准。	股票评级	买入	相对同期基准指数涨幅在 15%以上
		增持	相对同期基准指数涨幅在 5%~15%之间
		持有	相对同期基准指数涨幅在 -5%~+5%之间
		减持	相对同期基准指数跌幅在 5%以上
	行业评级	增持	相对同期基准指数涨幅在 10%以上
		中性	相对同期基准指数涨幅在 -10%~+10%之间
		减持	相对同期基准指数跌幅在 10%以上

国盛证券研究所

北京

地址：北京市东城区永定门西滨河路 8 号院 7 楼中海地产广场东塔 7 层

邮编：100077

邮箱：gsresearch@gszq.com

南昌

地址：南昌市红谷滩新区凤凰中大道 1115 号北京银行大厦

邮编：330038

传真：0791-86281485

邮箱：gsresearch@gszq.com

上海

地址：上海市浦东新区南洋泾路 555 号陆家嘴金融街区 22 栋

邮编：200120

电话：021-38124100

邮箱：gsresearch@gszq.com

深圳

地址：深圳市福田区福华三路 100 号鼎和大厦 24 楼

邮编：518033

邮箱：gsresearch@gszq.com