金融工程

证券研究报告
2020 年 02 月 25 日

市场微观结构探析系列之三：分时 K 线中的 alpha

作者

吴先兴 分析师
SAC 执业证书编号：S1110516120001
wuxianxing@tfzq.com
18616029821

缪铃凯 联系人
miaolingkai@tfzq.com

相关报告

1 《金融工程：市场微观结构探析系列之二：订单簿上的 alpha》2019-09-05

高频因子

高频选股因子与低频因子具有较大的差异。以月、季为持仓周期的低频选股因子主要来自于财务指标，其从盈利、成长、估值等维度综合评估上市公司；而以日为持仓周期的高频因子主要从股票量价信息中衍生而得。

高频因子挖掘与低频因子相比更显复杂和神秘。相比于低频基本面因子的挖掘由主动管理的投资逻辑所驱动，高频因子的挖掘更倾向于由数据所驱动。而股票交易产生的量价数据频率远高于财务信息，通过遍历量价衍生指标以筛选因子并非易事。

遗传编程

遗传编程通过模拟“物竞天择，适者生存”的进化思想，基于个体对于环境适应度，通过“自然选择”和“基因变异”方式从父代中迭代生成新的子代种群。我们构建了因子表达式到个体基因之间的映射，以因子绩效为进化目标让种群迭代繁衍以搜索有效的选股指标。

因子表达方面，本文基于树结构构建了表达式到个体基因之间的映射，并以树结构为基础实现了基因间的交叉互换以及变异；数据输入方面，本文以股票 30 分钟 K 线数据作为模型输入，其信息量是日频 K 线的 8 倍，有效地降低了降频所带来的交易信息损失；适应度方面，我们从信息系数 IC、多头超额、分组收益单调评估因子基因的适应度。

挖掘分时 K 线中的 alpha

基于遗传编程算法，本文以 2017 至 2018 年数据作为输入挖掘了 100 个有效的高频因子。利用 2019 年数据作为样本外测试集合，我们以选股指标 covariance(12,amount,high) 、 sub(close,ts_Mean(8,low)) 和 指 标 stddev(6,delta(11,log(volume)))为例展示了因子绩效，各因子分组收益单调，ICIR 分别达到了-10.98、-7.77 和-13.16，在样本外测试中仍然保持稳健选股能力。

100 个高频因子间保持较高独立性，两两之间相关系数绝对值均低于 0.70，平均值为 0.28。从样本内至样本外，因子 alpha 随时间出现衰减，ICIR 均值从 8.67 下降到 7.30，多空 IR 均值从 7.14 下降到 5.56，但因子整体在样本外仍然保持了显著的选股能力。

风险提示：因子失效风险，模型失效风险，市场风格变动风险

内容目录

图表目录

1. 高频 alpha

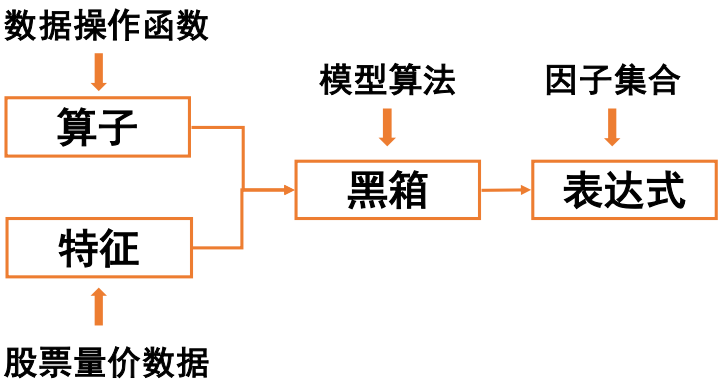
如果将构建多因子模型类比成烹饪菜肴，选股因子作为原料将对出锅“股票”的好坏起着决定作用。根据交易频率差异，多因子模型所使用选股因子存在明显差别：以月度或季度为持仓周期的低频模型，因子主要来自于基本面财务指标，从盈利、成长、估值、分析师预期等多维度综合评价上市公司；而以隔日为持仓周期的高频 alpha 模型主要关注股票量价特征，利用趋势跟踪、均值回归、量价相关性等捕捉短期统计套利机会。

多因子模型的运行原理简洁明了且框架也已非常成熟，因此差异化的模型取决于对细节的把控，即捕获所谓“工匠 alpha”。对于线性模型而言，我们希望寻找与股票收益呈线性相关的选股指标。良好的选股因子应该是由逻辑驱动而挖掘的，这样才能在实盘确保拥有更强的稳健性。因而低频模型所用选股因子基本以上市公司财务指标为主，模型包含因子数量也较少。从主动管理投资逻辑出发，在财务报表数据中遍历筛选出有效的选股特征集合在计算方面并不复杂。

高频 alpha 的主要数据来源是股票的量价特征，相较于基本面财务特征具有明显差异。首先，股价日频 K 线的数据频率是季频财报的 20 倍，如果涉及到高频分时 K 线、Tick 数据，复杂度将更为庞大；其次，在模型盈利逻辑层面，低频基本面模型更类似于因子投资或者“smart beta”，高频模型则将盈利希望寄托于大数定律所反应的统计规律，其中的规律更加难以归纳总结；最后，相比于低频基本面模型中一般数十个因子的使用量级，日间高频模型所使用的因子可能以成百上千计。

因此，对于初建高频因子库的投资者而言，手工搭建庞大的因子库需要耗费极大的精力；对于已有成熟因子库的投资者，寻找全新逻辑以挖掘增量 alpha 也将充满挑战。

图 1：数据驱动因子挖掘流程



资料来源：天风证券研究所

遗传编程（遗传规划，Genetic Programming）是解决高频因子挖掘的取巧方式之一。遗传编程借助于生物进化的思想，利用遗传算法在量价数据衍生特征集合中进行局部搜索，让种群快速朝着预期的进化方向收敛，其避免了对特征全集进行全局遍历；同时，相比于神经网络等模型，遗传编程可以清晰展示选股因子的逻辑表达式，这让我们对因子投资逻辑的后续验证成为了可能。

以往券商研报中对于遗传规划算法已有介绍，但一般局限于以股票日频、月频 K 线数据作为输入特征，日内数据中蕴含 alpha 则被忽略。**本文将尝试以股票分时 K 线作为输入特征，利用遗传规划算法挖掘有效的日频 alpha 因子。**

2. 遗传编程

由于难以对量价信息所衍生特征集合进行全局遍历，而手工构建高频因子库面临庞大工程量，因此在本章我们将介绍利用遗传编程挖掘有效的选股指标。

遗传编程（遗传规划，Genetic Programming）通过模拟生物“物竞天择，适者生存”的进化思想，通过种群筛选以及基因变异等方式不断在父代中迭代生成下一代种群。

遗传编程主要包含种群初始化、适应度评估、自然选择、基因交叉变异等步骤：

a) 初始化种群：初始化生成包含一定数量个体的种群；

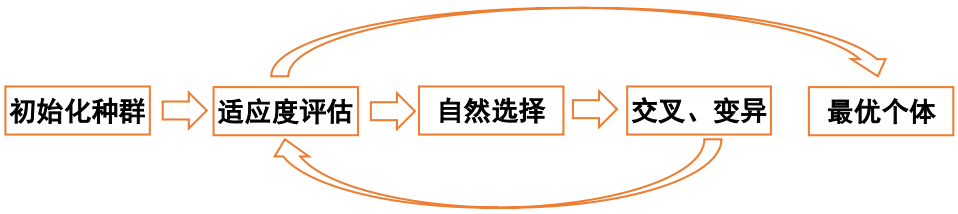
b) 适应度评估：根据评估标准度量个体的适应度；

c) 自然选择：按照优胜劣汰思想在当前种群中筛选一定数量个体进入下一代；

d) 基因交叉变异：由父代生成子代过程中模仿生物进化过程进行基因交叉互换和基因突变；

不断迭代适应度评估、自然选择以及基因交叉变异的过程生成子代种群，当算法触发设定的停止标准后我们可在种群中挑选出适应度最优的个体集合。

图 2：遗传编程流程

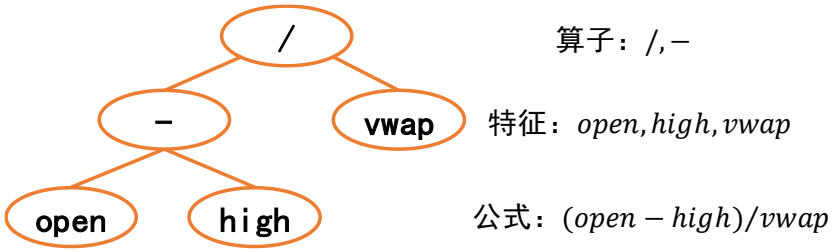


资料来源：天风证券研究所

2.1. 基因表达

对于遗传编程而言，个体基因的表示是其核心之一。我们将选股因子看作遗传编程算法中的种群个体，那么因子的具体表达式即为个体独有的基因。例如，结合股票日行情数据以及四则表达式我们可构造指标“(open-high)/vwap”，以树形结构可将其简洁地表示如下。

图 3：公式树示例

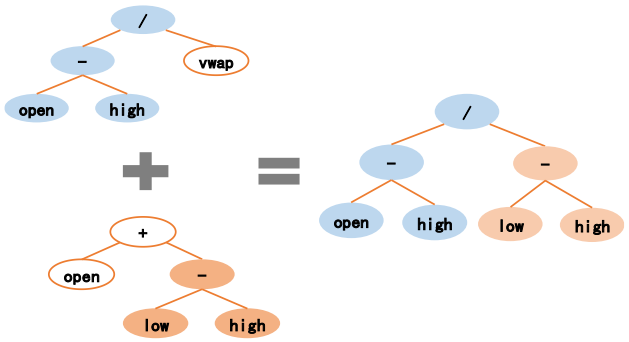


资料来源：天风证券研究所

通过树结构，因子基因的交叉与变异也能清晰地展现。在进化中种群中往往存在不同个体间基因的交叉互换以生成新的个体，其原理如下左图所示，树 1 将子树“vwap”与树 2 的子树“low-high”交叉互换得到新的树结构“(open-high)/(low-high)”。

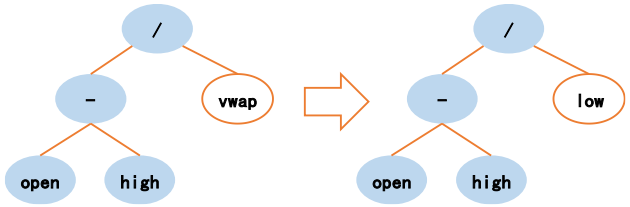
而个体自身在进化过程中还存在基因变异，本文所使用变异类型包含点变异、子树变异、Hoist 变异。在下右图中展示了点变异，因子树“(open-high)/vwap”的节点“vwap”通过点变异转变为“low”，进而得到新的树结构“(open-high)/low”。

图 4：基因交叉互换



资料来源：天风证券研究所

图 5：基因点变异



资料来源：天风证券研究所

2.1. 特征与算子

如前文因子树图中所展示，公式树由特征和算子两部分组成。在本文中，特征是我们输入模型的基础数据，即每只股票所拥有的量价信息。本文用以构建因子的量价数据包含股票在 K 线图上的开盘价、最高价、最低价、收盘价、均价、成交额、成交量共 7 类数据类型。

表 1：模型输入行情数据

特征	含义
open	开盘价
high	最高价
low	最低价
close	收盘价
vwap	均价
volume	成交量
amount	成交额

资料来源：天风证券研究所

算子即对特征进行运算以产生全新表达的函数，算子通过组合基础特征可产生全新的特征，进而得到更复杂的基因表达。根据类型的差异，不同算子接受参数个数也存在区别。以算子“correlation(n,x,y)”为例，其接受 x 和 y 两个参数，其计算了变量 x 和变量 y 在过去 n 天的相关系数；其中正整数 n 为算子调用过程中自动生成的随机数，x 和 y 为基础特征或经算子操作得到的基因表达式。本文所使用的算子集合如下表所示。

表 2：算子列表

算子符号	算子含义
add(x,y)	向量加法
sub(x,y)	向量减法
mul(x,y)	向量乘法
div(x,y)	向量除法
sqrt(x)	开平方根
log(x)	取自然对数
abs(x)	返回绝对值

neg(x)	返回相反数
actan(x)	反正切函数
max(x,y)	二者中最大值
min(x,y)	二者中最小值
rank(x)	截面排序值
correlation(n,x,y)	x 和 y 在过去 n 天相关系数
delta(n,x)	x 减去 n 天前 x 的取值
delay(n,x)	x 在 n 天前的取值
stddev(n,x)	过去 n 天 x 的波动率
ts_Max(n,x)	过去 n 天 x 的最大值
ts_Min(n,x)	过去 n 天 x 的最小值
ts_Rank(n,x)	x 在过去 n 天取值的排序
covariance(n,x,y)	x 和 y 在过去 n 天的协方差
ts_Mean(n,x)	过去 n 天 x 的平均值
REGbeta_ts(n,x,y)	过去 n 天 x 对 y 回归系数
REGresid_ts(n,x,y)	过去 n 天 x 对于回归残差
pctange_ts(n,x)	过去 n 天 x 的变化率

资料来源：天风证券研究所

2.2. 适应度评估

我们将因子表达式作为个体基因，通过因子表达式可计算具体因子值。因子挖掘的目标是寻找具有显著 alpha 的选股因子，因此我们通过因子的选股绩效度量个体适应度，进而让因子种群朝着“alpha”的方向进化。

图 6：基因到适应度传导流程

资料来源：天风证券研究所

高频因子绩效刻画首先要处理的是高频模型对于撮合价格的敏感性。相较于以往低频选股因子我们简化地以收盘价或者开盘价计算因子收益，但是日间 alpha 对成交价格非常敏感，实际交易时难以在收盘或开盘的一个时间点完全全部成交，因此本文以股票每日开盘前 30 分钟均价作为股票的撮合价格。

信息系数 IC 一般被作为选股因子 alpha 强弱的度量指标，IC 在整体上刻画了因子值与股票未来收益的线性相关性。由于做空机制问题，实际操作中我们更关注因子在多头端的表现，因此我们将因子多头收益也作为适应度评估的一个维度。此外，由于我们以线性模型复合因子，因子分组单调性也将对模型表现产生影响。

综上，我们将个体基因翻译成因子取值，进而综合因子 IC、多头收益、分组单调性三个绩效维度评估个体适应度。

图 7：个体适应度评估维度

资料来源：天风证券研究所

2.3. 自然选择

遗传算法新颖之处即在于其模仿了生物进化的过程，根据“适者生存”原则让种群中的胜者向下一代遗传其基因。因此我们根据个体的适应度以确定其被选入下一代的概率。按照轮盘赌算法，假设个体 i 的适应度$fitness_i$，我们从种群中重抽样时，个体 i 被抽取中的概率ρ_i满足：

$$\rho_i = fitness_i / \sum_i fitness_i$$

为丰富基因多样性，个体进入下一代时将按照前文所述基因交叉互换以及变异。假设α_1、α_2将区间[0,1]切割成 3 部分：

$$0 < \alpha_1 < \alpha_2 < 1$$

我们定义基因交叉的概率为α_1，基因变异概率为$\alpha_2 - \alpha_1$，直接复制个体进入下一代概率为$1 - \alpha_2$。通过随机数 $\beta \in [0,1]$ 落入具体区间的情形，我们可以确定个体进化过程中的基因遗传与变异的方式

3. 挖掘分时 K 线中的 alpha

本章我们将以股票的分时 K 线信息为基础，利用遗传编程算法挖掘分时量价信息中所蕴涵的高频 alpha 信息。

首先，我们将遗传编程算法看作一个黑箱，模型输出完全依赖于其被投喂的数据，训练数据对结果的影响至关重要。高频因子一般具有较短的迭代周期，如果输入数据的时间跨度过长，所生成因子表达式可能已经不适应最新的市场情境。因此，在每个时间点，我们以最近 2 年数据作为输入，至于生成因子的迭代周期则需要在算力和需求间综合权衡。

此外，以往券商研报中所涉及的遗传编程一般是基于股票的日频 K 线进行，日频 K 线来自于个股逐笔交易信息的日级别降频，其信息损耗非常明显。模型的输入信息量应该在机器算力以及信息损耗之间取最优权衡，本文的输入为股票的 30 分钟 K 线数据，其信息量是日频 K 线的 8 倍。

最后，为避免模型过度拟合，两年的输入数据被均分成训练集和测试集，我们根据训练集生成因子表达式的同时将在测试集上检验表达式的有效性。最终生成的选股因子在本文中均默认进行行业和市值中性化处理。

以 2019-01-01 时间点为例，本文展现了基于遗传编程算法的因子挖掘流程以及模型生

成的选股因子介绍和列表。

以 2017 年和 2018 年的股票 30 分钟 K 线作为输入，其中 2017 年数据用于训练生成因子表达式，2018 年数据用于测试因子绩效。根据适应度的三个评估维度信息系数 IC、多头超额和分组单调性，我们设定筛选阈值为 0.03、0.15、0.8 后，模型按照如下流程进行迭代[1]：

1. 根据随机数种子生成符合种群数目的随机数表达式；
2. 从 IC、多头收益、分组单调性三维度评估个体适应度；记录各维度都达到阈值个体至因子池；
3. 根据适应度按照轮盘赌法筛选下一代个体，并对个体进行基因交叉互换以及变异等操作，而后转到步骤 2；
4. 循环步骤 2、3 至给定深度后，在因子池中按两两相关性系数阈值剔除高度相关指标得到选股因子集合；

3.1. 因子介绍

按照上述流程本文挖掘了 100 个基于 30 分钟 K 线的选股因子，按照相关系数阈值 0.7 的上限我们在挖掘过程中保证了因子间两两相关系数均在设定的阈值以下。下面展示了部分因子 2017 年至 2019 年间的绩效表现，其中前两年为样本内数据，2019 年为样本外，因子按次日开盘前 30 分钟均价买卖股票以计算收益。

3.1.1. covariance(12,amount,high)

covariance(12,amount,high)利用最近 12 根 30 分钟 k 线度量了量、价的协方差。该指标是个反向指标，即量价负相关且波动越大股票在次日拥有更高的相对收益。

因子从 2017 年以来 IC 均值为-4.9%，ICIR 为 10.98，多头超额收益 15.6%，多空 IR 为 6.74，信号表现出稳健的选股能力。

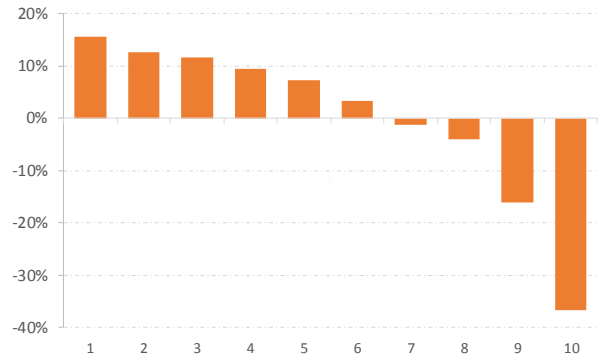
表 3：因子绩效-covariance(12,amount,high)

IC	ICIR	多头超额	多空收益	多空 IR
-4.9%	-10.98	15.6%	52.2%	6.74

资料来源：Wind，天风证券研究所

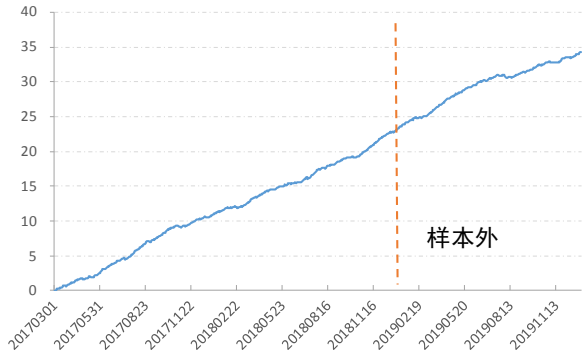
covariance(12,amount,high)指标的分组收益以及 IC 累计值如下图所示，因子单调性良好，IC 在样本外仍然保持稳健增长。

图 8：分组收益-covariance(12,amount,high)



资料来源：Wind，天风证券研究所

图 9：IC 累计值（方向调整后）-covariance(12,amount,high)



资料来源：Wind，天风证券研究所

[1] 超额收益为扣费前收益，年收益按日收益*242 计算。

3.1.2. sub(close,ts_Mean(8,low))

sub(close,ts_Mean(8,low))是个反转指标，其刻画了日收盘价与最近 8 根 30 分钟 K 线最低价均值间的价差。因子从 2017 年以来 IC 均值为-4.9%，ICIR 为 10.98，多头超额收益 15.6%，多空 IR 为 6.74，信号表现出稳健的选股能力。

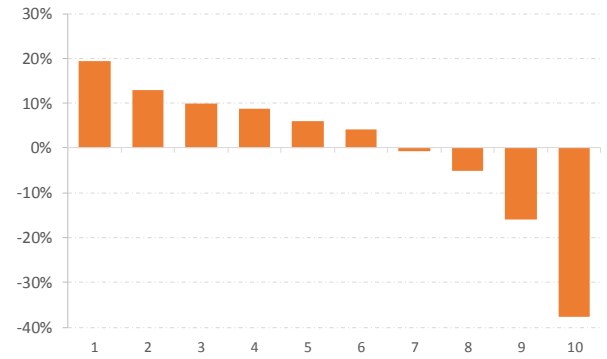
表 4：因子绩效-sub(close,ts_Mean(8,low))

IC	ICIR	多头超额	多空收益	多空 IR
-4.1%	-7.77	19.4%	57.1%	6.77

资料来源：Wind，天风证券研究所

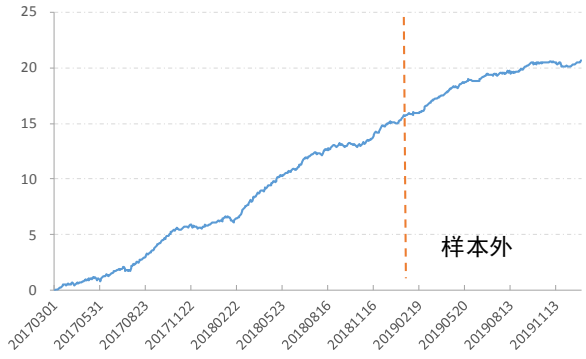
sub(close,ts_Mean(8,low))指标的分组收益以及 IC 累计值如下图所示，因子单调性良好，IC 在样本外仍然保持稳健增长。

图 10：分组收益-sub(close,ts_Mean(8,low))



资料来源：Wind，天风证券研究所

图 11：IC 累计值（方向调整后）-sub(close,ts_Mean(8,low))



资料来源：Wind，天风证券研究所

3.1.3. stddev(6,delta(11,log(volume)))

stddev(6,delta(11,log(volume)))指标刻画了成交量变化的波动率，因子从 2017 年以来 IC 均值为-3.6%，ICIR 为 13.16，多头超额收益 15.0%，多空 IR 为 8.40。

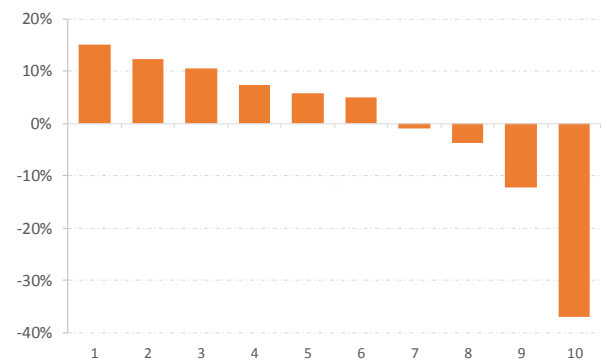
表 5：因子绩效-stddev(6,delta(11,log(volume)))

IC	ICIR	多头超额	多空收益	多空 IR
-3.6%	-13.16	15.0%	52.0%	8.40

资料来源：Wind，天风证券研究所

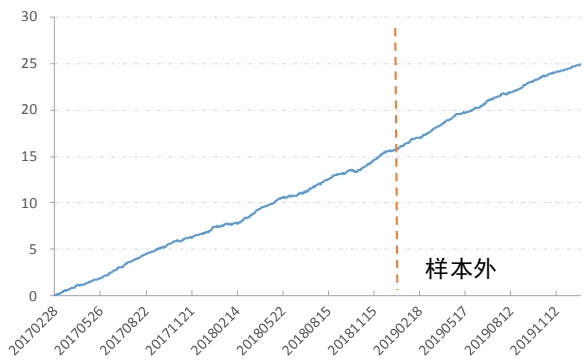
指标的分组收益以及 IC 累计值如下图所示，因子单调性良好，IC 增长稳健。

图 12：分组收益-stddev(6,delta(11,log(volume)))



资料来源：Wind，天风证券研究所

图 13：分组收益-stddev(6,delta(11,log(volume)))



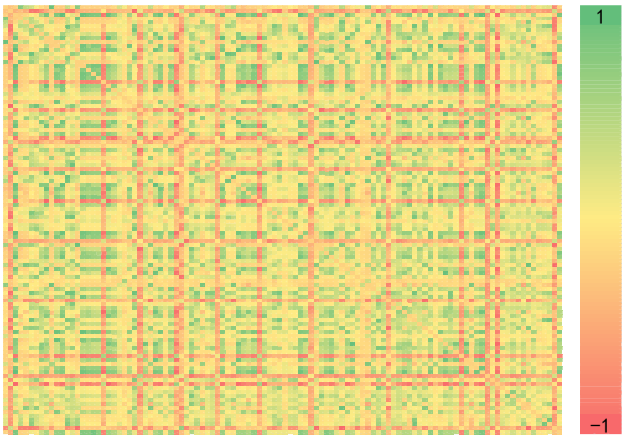
资料来源：Wind，天风证券研究所

3.2. 因子分析

根据遗传算法我们挖掘了 100 个有效的选股指标，样本内区间为 2017 年和 2018 年，样本外检验区间为 2019。

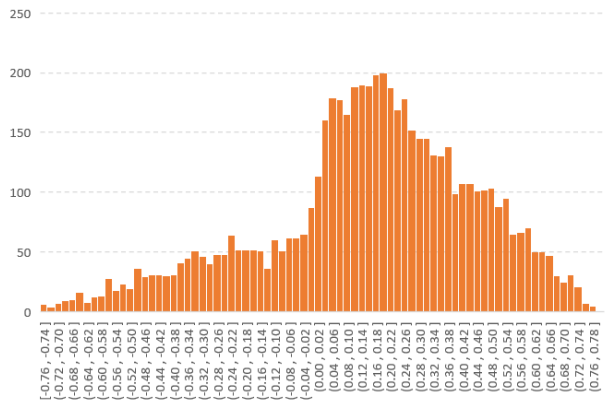
我们所挖掘的 100 个因子间整体保持较高独立性。相关系数热力图以及分布图如下所示，因子间两两相关性较低，相关系数绝对均值为 0.28。

图 14：高频因子相关系数热力图



资料来源：Wind，天风证券研究所

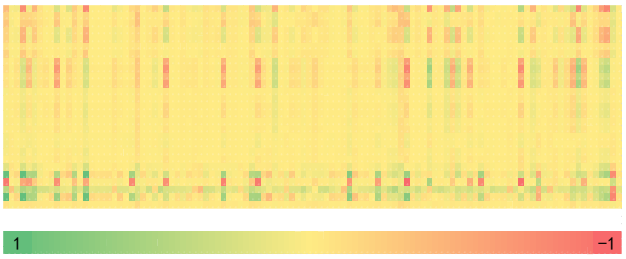
图 15：高频因子相关系数分布图



资料来源：Wind，天风证券研究所

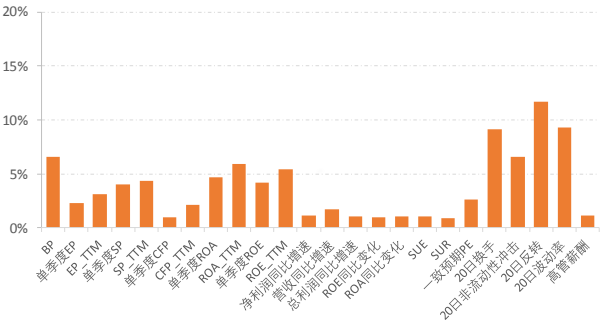
此外，我们对比本中的高频因子与常见低频选股因子间的相关系数，如下图所示。左图中横轴是高频因子，纵轴是低频因子，可以看出二者间相关性基本在 0 附近波动。高、低频因子之间相关系数绝对值均值如右图所示，各低频因子与高频因子的平均相关性均在 20%以下。

图 16：高频因子与低频指标相关性热力图



资料来源：Wind，天风证券研究所

图 17：高频因子与低频因子平均相关性



资料来源：Wind，天风证券研究所

由于 2017 至 2018 年为样本内数据，因此我们以 2019 年为样本外对比因子在样本内外的绩效表现。因子 ICIR 前后变化如下所示，在样本外 ICIR 出现一定程度下滑，ICIR 均值从 8.67 下降到 7.30，中位数从 8.34 下降到 6.59。但从 ICIR 维度评估，因子在样本外仍然保持着显著的选股能力。

表 6：因子绩效在样本内外差异

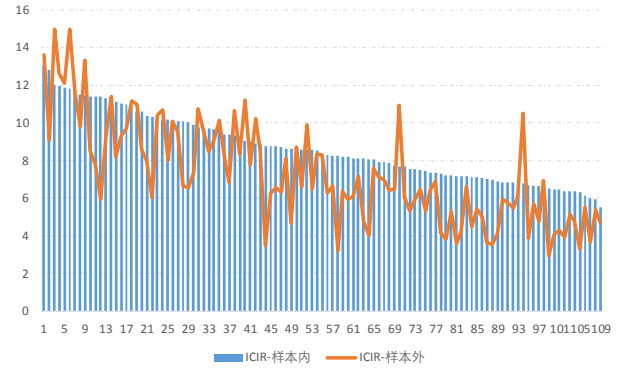
数据类型	ICIR 均值	ICIR 中位数	多空均值	多空中位数
样本内	8.67	8.34	7.14	7.02
样本外	7.29	6.59	5.56	5.65

资料来源：Wind，天风证券研究所

类似的，因子在样本内外多空收益 IR 对比图如下所示。因子多空 IR 整体在样本外出现一定程度下滑，平均值从 7.14 下降到 5.56，中位数从 7.02 下降到 5.65。但从多空 IR 维度

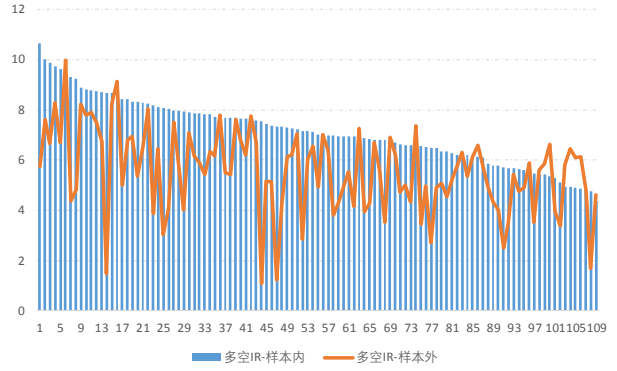
评估，因子在样本外仍然保持着显著的选股能力。

图 18：分样本 ICIR 对比



资料来源：Wind，天风证券研究所

图 19：分样本多空 IR 对比



资料来源：Wind，天风证券研究所

本文所挖掘的100个基于30分时K线的高频因子表达式如下表所示，其中open、high、low、close、和vwap分别为股票K线的开高低收和均价，volume、amount为成交数量和金额，算子定义如前文。

表 7：因子列表-分时 K 线

序号	表达式
1	ts_Max(11,covariance(8,ts_Max(5,stddev(10,mul(volume,vwap))),ts_Mean(12,amount)))
2	sub(correlation(11,sig(close),correlation(10,delta(2,open),delta(2,amount))),correlation(12,add(high,volume),high))
3	sub(delta(5,vwap),sqrt(ts_Rank(8,neg(close))))
4	ts_Max(11,covariance(8,ts_Max(5,low),ts_Mean(12,amount)))
5	div(covariance(9,volume,amount),ts_Mean(8,REGbeta_ts(13,amount,high)))
6	covariance(12,amount,high)
7	ts_Max(6,correlation(12,volume,high))
8	sqrt(min(log(mul(arctan(volume),correlation(12,volume,high))),delta(6,open)))
9	min(max(sig(vwap),delta(3,volume)),pctchange_ts(4,high))
10	stddev(15,add(max(stddev(11,close),ts_Max(14,volume)),stddev(11,close)))
11	delta(2,log(close))
12	min(delta(3,low),delta(6,close))
13	rank(covariance(9,volume,high))
14	correlation(9,log(volume),max(close,close))
15	stddev(14,max(ts_Max(9,max(delta(15,vwap),rank(amount))),covariance(9,vwap,amount)))
16	delta(3,close)
17	REGresid_ts(7,delay(8,mul(volume,volume)),close)
18	REGresid_ts(7,delay(8,open),close)
19	REGresid_ts(7,REGbeta_ts(12,open,low),close)
20	add(mul(neg(low),min(ts_Max(11,open),delta(5,close))),div(ts_Rank(9,vwap),min(low,low)))
21	div(min(vwap,stddev(8,volume)),ts_Max(8,low))
22	arctan(REGresid_ts(8,sig(REGbeta_ts(7,abs(volume),sqrt(close))),low))
23	add(delta(4,delta(15,close)),ts_Max(13,rank(amount)))
24	add(delta(6,rank(arctan(volume))),stddev(9,log(volume)))
25	abs(sub(REGbeta_ts(6,amount,low),div(close,open)))
26	correlation(15,volume,vwap)
27	REGbeta_ts(6,log(mul(add(volume,high),arctan(amount))),sub(stddev(14,div(close,open)),low))
28	ts_Mean(11,div(close,open))

29	ts_Max(13,REGbeta_ts(9,vwap,amount))
30	div(delta(6,min(low,vwap)),div(sqrt(open),abs(open)))
31	pctchange_ts(5,sub(add(close,high),sub(abs(high),REGresid_ts(12,open,close))))
32	REGbeta_ts(14,close,stddev(6,abs(close)))
33	covariance(4,vwap,ts_Rank(14,volume))
34	neg(REGresid_ts(8,amount,low))
35	neg(ts_Max(12,correlation(14,amount,close)))
36	correlation(12,vwap,ts_Mean(8,volume))
37	rank(covariance(6,amount,close))
38	REGbeta_ts(8,log(volume),vwap)
39	add(volume,ts_Mean(11,covariance(12,amount,vwap)))
40	correlation(9,sig(rank(vwap)),volume)
41	correlation(9,sig(arctan(close)),max(pctchange_ts(11,open),delta(9,volume)))
42	correlation(5,delta(4,high),neg(volume))
43	mul(max(sub(open,volume),max(volume,vwap)),delta(4,close))
44	div(div(neg(close),sub(high,vwap)),rank(covariance(8,max(high,vwap),volume)))
45	div(div(neg(close),ts_Mean(13,amount)),rank(covariance(8,max(high,vwap),volume)))
46	delta(6,high)
47	min(pctchange_ts(1,close),pctchange_ts(4,high))
48	min(sub(sig(min(low,amount)),div(vwap,close)),pctchange_ts(5,close))
49	REGresid_ts(7,mul(arctan(ts_Max(8,amount)),ts_Rank(6,ts_Rank(8,REGbeta_ts(9,open,amount)))),close)
50	div(ts_Max(5,vwap),high)
51	min(delay(12,REGbeta_ts(15,mul(open,volume),covariance(10,arctan(high),add(max(high,volume),vwap)))),pctchange_ts(5,close))
52	min(correlation(13,vwap,sqrt(ts_Max(6,amount))),delta(2,open))
53	min(correlation(13,vwap,volume),covariance(14,min(vwap,amount),stddev(8,low)))
54	arctan(add(correlation(9,volume,high),pctchange_ts(1,ts_Max(14,volume))))
55	pctchange_ts(13,min(log(delay(2,amount)),REGresid_ts(7,delay(5,vwap),add(rank(volume),close))))
56	correlation(14,delta(11,volume),add(correlation(9,volume,high),close))
57	correlation(14,add(low,max(volume,low)),pctchange_ts(5,log(vwap)))
58	min(log(delay(2,amount)),REGresid_ts(7,rank(amount),add(high,close)))
59	min(log(delay(2,amount)),REGresid_ts(7,delay(5,vwap),add(add(vwap,close),close)))
60	sub(neg(correlation(11,min(sqrt(volume),open),ts_Max(10,volume))),delta(5,min(low,volume)))
61	mul(ts_Max(13,correlation(9,volume,close)),delay(7,max(sqrt(amount),correlation(9,volume,close))))
62	REGbeta_ts(6,sub(log(volume),low),log(add(low,close)))
63	REGresid_ts(10,delay(10,high),mul(open,close))
64	REGresid_ts(7,min(ts_Rank(4,open),log(vwap)),vwap)
65	correlation(7,add(volume,close),sqrt(pctchange_ts(5,close)))
66	min(ts_Max(10,amount),ts_Max(7,REGbeta_ts(12,vwap,amount)))
67	REGresid_ts(6,ts_Rank(7,ts_Rank(14,sub(vwap,close))),arctan(sqrt(vwap)))
68	mul(low,min(pctchange_ts(7,low),ts_Min(11,volume)))
69	sub(ts_Mean(15,correlation(11,vwap,volume)),sub(open,high))
70	stddev(6,delta(11,log(volume)))
71	stddev(6,delta(11,neg(volume)))
72	rank(correlation(4,amount,close))
73	arctan(mul(stddev(14,volume),REGbeta_ts(10,amount,vwap)))
74	max(delta(3,close),correlation(9,low,volume))
75	mul(rank(min(covariance(11,amount,vwap),ts_Mean(4,REGbeta_ts(15,amount,low)))),ts_Min(9,div(delta(2,amount),stddev(8,close))))

76	correlation(9,abs(volume),pctchange_ts(14,close))
77	add(delta(5,low),arctan(ts_Min(4,covariance(14,volume,close))))
78	add(delta(5,low),arctan(high))
79	min(div(vwap,sub(ts_Min(11,neg(vwap)),abs(amount))),delta(5,vwap))
80	mul(ts_Mean(6,vwap),correlation(10,high,amount))
81	add(rank(REGbeta_ts(6,volume,close)),neg(REGbeta_ts(14,close,low)))
82	covariance(6,ts_Rank(5,max(low,volume)),max(div(rank(volume),mul(high,volume)),add(vwap,close)))
83	max(REGbeta_ts(14,high,amount),REGbeta_ts(6,volume,open))
84	REGresid_ts(6,REGresid_ts(4,high,amount),sqrt(close))
85	mul(vwap,div(sig(ts_Rank(4,volume)),abs(REGbeta_ts(4,amount,low))))
86	ts_Mean(2,sub(vwap,open))
87	pctchange_ts(7,vwap)
88	sub(delta(7,close),neg(min(log(volume),correlation(10,low,volume))))
89	mul(log(abs(volume)),delta(7,rank(close)))
90	REGresid_ts(10,arctan(delay(4,pctchange_ts(6,min(delay(13,close),pctchange_ts(11,arctan(rank(vwap)))))),neg(close))
91	mul(delay(7,amount),delta(7,close))
92	div(sub(low,sub(stddev(6,correlation(7,min(low,open),stddev(5,low))),div(sub(amount,volume),arctan(volume)))),stddev(5,low))
93	REGresid_ts(10,div(close,REGbeta_ts(14,high,low)),close)
94	sub(max(close,low),ts_Mean(8,low))
95	sub(ts_Max(5,REGbeta_ts(15,close,amount)),log(amount))
96	correlation(11,mul(covariance(7,vwap,amount),ts_Mean(12,low)),stddev(6,abs(amount)))
97	sub(sig(sub(high,volume)),pctchange_ts(6,log(close)))
98	min(REGbeta_ts(15,add(amount,low),rank(open)),REGbeta_ts(9,volume,vwap))
99	min(covariance(6,sqrt(volume),max(ts_Rank(8,close),vwap)),REGbeta_ts(14,covariance(10,high,vwap),min(open,close)))
100	add(covariance(6,sqrt(volume),max(ts_Rank(8,close),vwap)),rank(close))

资料来源：Wind，天风证券研究所

4. 总结

高频选股因子与低频因子具有较大的差异。以月、季为持仓周期的低频选股因子主要来自于财务指标，其从盈利、成长、估值等维度综合评估上市公司；而以日为持仓周期的高频因子主要从股票量价信息中衍生而得。

高频因子挖掘与低频因子相比更显复杂和神秘。相比于低频基本面因子的挖掘由主动管理的投资逻辑所驱动，高频因子的挖掘更倾向于由数据所驱动。而股票交易产生的量价数据频率远高于财务信息，通过遍历量价衍生指标以筛选因子并非易事。本文中，我们介绍了如何利用遗传编程算法高效地挖掘高频选股因子。

遗传编程通过模拟“物竞天择，适者生存”的进化思想，基于个体对于环境适应度，通过“自然选择”和“基因变异”方式从父代中迭代生成新的子代种群。我们构建因子表达到个体基因之间的映射，以因子绩效为进化目标让种群迭代繁衍以搜索有效的选股指标。

因子表达方面，本文基于树结构构建了因子表达到个体基因之间的映射，并以树结构为基准实现了基因间的交叉互换以及变异；数据输入方面，本文以股票 30 分钟 K 线数据作为模型输入，其信息量是日频 K 线的 8 倍，有效地降低了降频所带来的交易信息损失；适应度方面，我们从信息系数 IC、多头超额、分组收益单调评估因子基因的适应度。

基于遗传编程算法，本文以 2017 至 2018 年数据作为输入挖掘了 100 个有效的高频因子。利用 2019 年数据作为样本外测试集合，我们以指标 covariance(12,amount,high)、

sub(close,ts_Mean(8,low))和 stddev(6,delta(11,log(volume)))为例展示了因子绩效，各因子分组收益单调，ICIR 分别达到了-10.98、-7.77 和-13.16，在样本外测试中仍然保持稳健选股能力。

100 个因子间保持较高独立性，两两相关系数绝对值均低于 0.70，平均值为 0.28。对比因子在样本内至样本外绩效，因子集合 ICIR 均值从 8.67 下降到 7.30，多空 IR 均值从 7.14 下降到 5.56，因子 alpha 随时间出现衰减，但仍然保持了显著的选股能力。

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，本报告所表述的所有观点均准确地反映了我们对标的证券和发行人的个人看法。我们所得报酬的任何部分不曾与，不与，也将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。

一般声明

特别声明

投资评级声明

类别	说明	评级	体系
股票投资评级	自报告日后的 6 个月内，相对同期沪深 300 指数的涨跌幅	买入	预期股价相对收益 20%以上
		增持	预期股价相对收益 10%-20%
		持有	预期股价相对收益-10%-10%
		卖出	预期股价相对收益-10%以下
行业投资评级	自报告日后的 6 个月内，相对同期沪深 300 指数的涨跌幅	强于大市	预期行业指数涨幅 5%以上
		中性	预期行业指数涨幅-5%-5%
		弱于大市	预期行业指数涨幅-5%以下

天风证券研究

北京	武汉	上海	深圳
北京市西城区佟麟阁路 36 号	湖北武汉市武昌区中南路 99 号保利广场 A 座 37 楼	上海市浦东新区兰花路 333 号 333 世纪大厦 20 楼	深圳市福田区益田路 5033 号平安金融中心 71 楼
邮编：100031	邮编：430071	邮编：201204	邮编：518000
邮箱：research@tfzq.com	电话：(8627)-87618889	电话：(8621)-68815388	电话：(86755)-23915663
	传真：(8627)-87618863	传真：(8621)-68812910	传真：(86755)-82571995
	邮箱：research@tfzq.com	邮箱：research@tfzq.com	邮箱：research@tfzq.com