中信期货
CITIC Futures

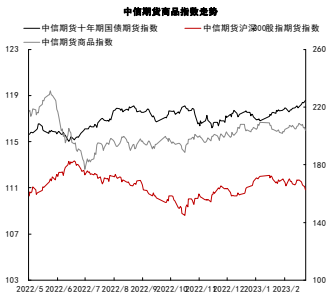
中信期货研究|金融工程专题报告

2023-03-13

利用遗传规划挖掘商品期货截面因子

——期货多因子系列七

投资咨询业务资格：
证监许可【2012】669 号



金融工程研究团队

研究员：
周通
010-80401733
从业资格号 F3078183
投资咨询号 Z0018055

报告要点

本文详细介绍了遗传规划算法的原理以及如何使用它进行因子挖掘。具体到实操流程上，我们使用到了 Python 中专门实现遗传规划的 gplearn 包进行代码实现。在对 gplearn 包进行修改优化后，我们以 2016 至 2022 年作为输入数据在商品期货上挖掘了 5 个具有较好选期能力的截面 alpha 因子，并回测了其在全样本上的表现。回测结果表明，遗传规划挖掘出的因子在样本内外均具备一定的有效性。

摘要：

本文利用遗传规划算法在商品期货上挖掘了5个具有较好选期能力的截面 alpha 因子，并回测了其在全样本上的表现。回测结果如下：

- **Alpha1：** 年化收益:9.32%，年化波动:5.42%，夏普:1.72，最大回撤:4.15%，卡玛比率:2.25；
- **Alpha2：** 年化收益:11.07%，年化波动:7.00%，夏普:1.58，最大回撤:6.05%，卡玛比率:1.83；
- **Alpha3：** 年化收益:10.24%，年化波动:7.74%，夏普:1.32，最大回撤:9.86%，卡玛比率:1.05；
- **Alpha4：** 年化收益:9.35%，年化波动:7.41%，夏普:1.26，最大回撤:8.16%，卡玛比率:1.15；
- **Alpha5：** 年化收益:8.40%，年化波动:7.11%，夏普:1.18，最大回撤:9.00%，卡玛比率:0.93；

回测下来，由遗传规划算法挖掘出来的因子在样本内外均具备一定的有效性，表明利用遗传规划可以帮助我们归纳并总结出具有一定 alpha 能力的因子。

风险提示：本报告中所涉及的资产配比和模型应用仅为回溯举例，并不构成推荐建议。

目 录

图表目录

一、因子挖掘方法论

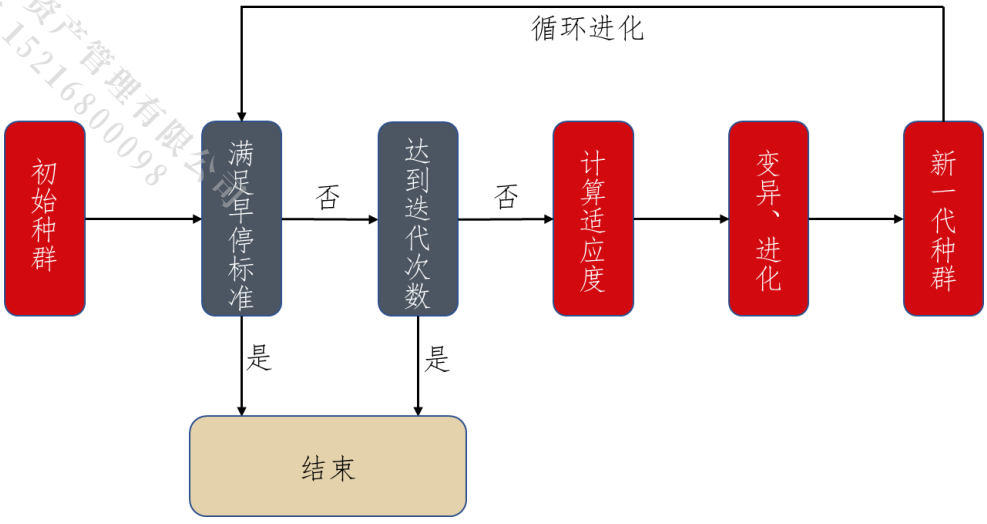
在我们之前的期货多因子系列报告中，各类因子往往是基于一定的经济学逻辑或是历史经验人工构造出来的。如经典的截面动量因子，就是我们认为资产之间的相对强弱关系会延续（动量效应），那么依据“强者恒强，弱智恒弱”的逻辑我们就可以构造出动量因子。这样的构建因子的方式属于“先有逻辑，后有公式”的“演绎法”。在因子构建过程中，还有一种方式，即我们可以先通过机器学习的方法将大量的历史数据与收益率序列相拟合，生成海量因子，再从中归纳总结出背后的逻辑。这种方式就属于“先有公式，后有逻辑”的“归纳法”。

在“归纳法”法中，遗传规划作为一种仿生启发式算法，在我们对目标因子知之甚少的情况下可以帮助我们挖掘因子。它的优势在于通过启发式的搜索，可以挖掘出人工难以构造、复杂的因子。本节我们将重点介绍遗传规划算法以及如何使用 Python 的 gplearn 包进行因子挖掘。

（一）遗传算法与遗传规划简介

遗传算法（Genetic Algorithm）最初由美国密歇根大学的 J. Holland 提出，是一种通过模拟自然界生物进化过程（“物竞天择、适者生存”）搜索最优解的算法，其本质上也是一种监督学习算法。对于一个最优化的问题，它借鉴了生物学中的现象（遗传、突变、杂交等）使一定数量的初始解按照适应度的方向进化为更优的解。进化过程从完全随机生成的个体种群开始，一代代进化。每一代中，会基于个体的适应度筛选出较优个体并在个体中发生变异、进化进而生成新的种群，新种群则继续进行迭代直至生成最优（适应度最高）种群。

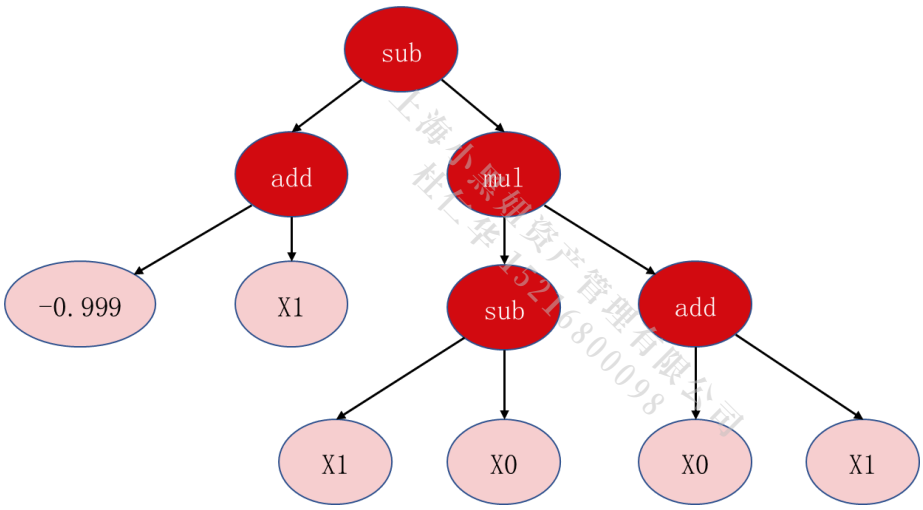
图表1：遗传算法图解



资料来源：中信期货研究所

遗传规划/遗传编程（Genetic Progamming）是遗传算法中的一个分支，相较于遗传算法，它最大的不同在于将进化对象编码成树而不是简单的线性串。得益于此，遗传规划可以表达更为复杂的语义。下图简单展示了在遗传规划算法中可以作为进化个体的公式树（$X_0^2 - X_1^2 + X_1 - 0.999$）。

图表2：公式树图解



资料来源：gplearn、中信期货研究所

我们沿着公式树从下往上即可还原整个公式，其中红色的节点表征函数，其他的粉红色叶子则表征变量和常数。在因子挖掘过程中，公式树也可以视作我们的因子表达式，所以公式树的进化过程实际上就是挖掘因子。

（二）适应度函数

适应度函数用来计算每一代种群中个体的适应度，适应度则衡量了个体与最终的目标个体的相符程度。每一代进化过程中，只保留适应度较高的个体作为后续进化的种群，适应度较低的个体则被淘汰。从因子挖掘的角度来说，我们可以使用用来评判因子有效程度的指标做为个体适应度，如使用因子的 RankIC、RankICIR 可以挖掘与未来收益率相关程度较高的线性因子；互信息（Mutual information）可以用来挖掘与识别非线性的因子；或是直接使用因子回测后的夏普率（sharpe ratio）作为适应度。

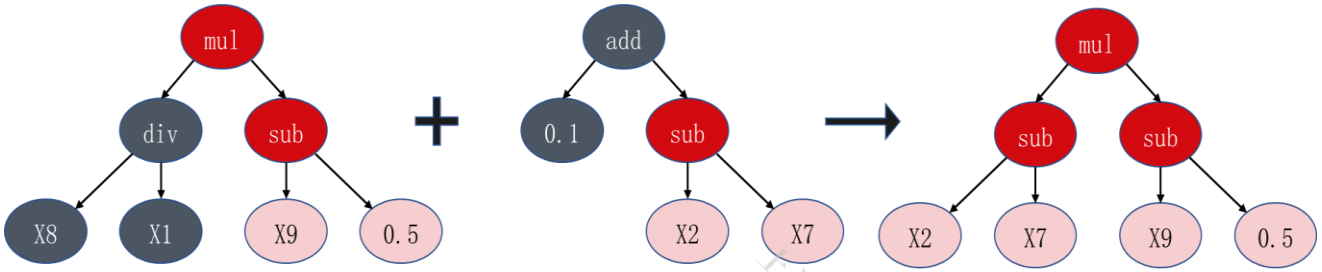
（三）公式树的进化方式

1. 杂交变异

杂交变异是遗传规化中的主要变异方式，是将两个个体中的遗传物质进行混合交换。在公式树进化过程中，杂交变异需要选择两个种群，其中一个种群中的最高适应度个体作为“父代”，从“父代”中随机挑选一个子树以待替换；并从另一个种群中随机挑选一个个体作为“捐赠者”，也从“捐赠者”中随机挑选一

个子树用以替换“父代”中的子树。最终生成的公式树用以后续进化。

图表3：杂交变异示意图

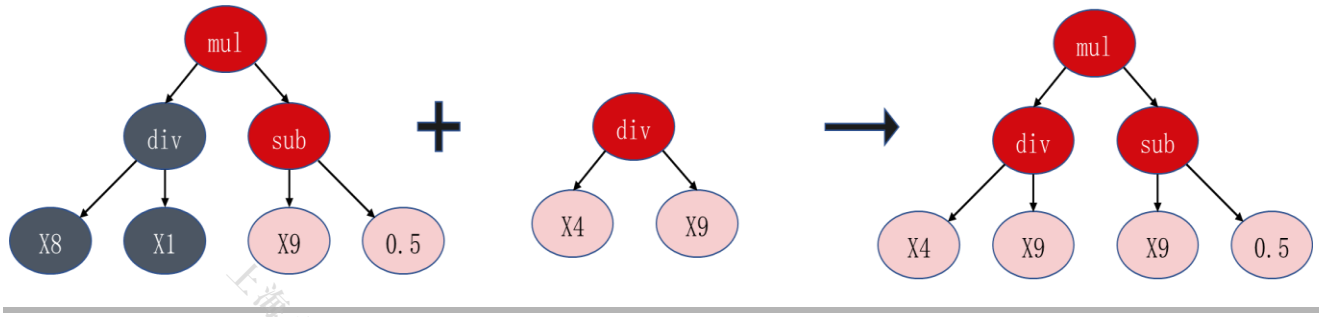


资料来源：gplearn、中信期货研究所

2. 子树变异

子树变异是一种更为激进的变异方式，称它更激进是因为子树变异如同对公式树进行“移植手术”般，直接将公式树中的子树替换成最初生成的随机子树。这样做可以将已经淘汰的子树重新引入进化过程中以增强遗传物质的多样性。

图表4：子树变异示意图

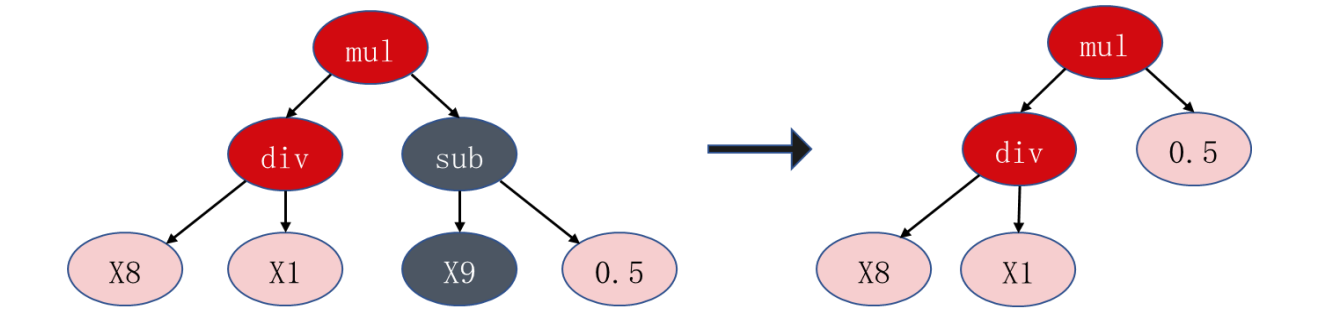


资料来源：gplearn、中信期货研究所

3. Hoist 变异

Hoist（抬升）变异也是一种较为激进的变异方式。主要是为了防止公式树过于“臃肿庞大”而对公式树中的随机子树进行“截断”，并保留子树中的一小部分用以替换原子树。

图表5：Hoist 变异示意图

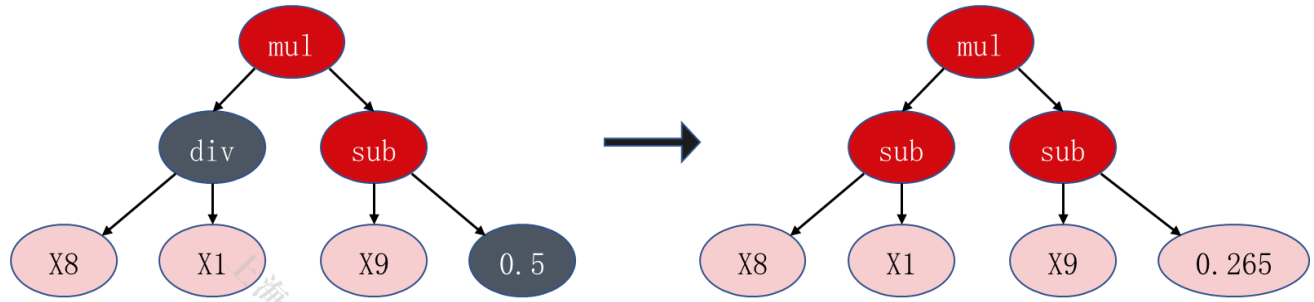


资料来源：gplearn、中信期货研究所

4. 点变异

点变异是遗传规划中最常见的变异方式，与子树变异类似都是为了增添遗传物质的多样性而将过去淘汰的节点重新引入进化过程中。不同于子树变异激进的变异方式，点变异较为温和，仅仅只替换公式数中的随机某个节点而不是整个子树。

图表6：点变异示意图



资料来源：gplearn、中信期货研究所

以上即为公式树的主要变异方式，它们四者的概率之和应小于 1。从因子的可解释性角度来说，我们应该尽量减少因子挖掘过程中子树变异、Hoist 变异与点变异的概率，而提升杂交变异概率。这么做的目的也是为了防止因子表达式过于难以解释。

（四）gplearn 的使用与改进

1. 参数说明与设置

本文主要通过 Python 中的专门实现遗传规划的 gplearn 包进行商品期货的截面因子挖掘，有关 gplearn 的详细介绍，感兴趣的读者可以查阅官方文档（https://gplearn.readthedocs.io/en/stable/）。

gplearn 中的特征转化器 SymbolicTransformer，可以将初代种群的原始特征通过公式树转换为适应度最优的新特征，因此可以用于因子的挖掘。其中 SymbolicTransformer 模型的主要参数说明与设置如下：

图表7：SymbolicTransformer 参数说明与设置

参数	定义	取值	说明
feature_name	定义输入特征的名称，默认为 X0，X1…	open, close, high, low, volume, amount, avgprice, openinterest, return_close warehouse, warehouse_chg	我们输入开盘价、收盘价、最高价、最低价、成交量、成交额、成交均价、持仓量、当日收盘价收益率、仓单数、仓单变化数作为初代种群且未经其他特征工程。
generation	公式进化的世代数量。世代数量越多，消耗算力越多，公式的进化次数越多，默认为 20。	2	通过测试公式进化 2 代或者 3 代，种群适应度已较高。继续进化不仅消耗大量算力且因子可解释性较低。出于节省算力的考虑，本文选择进化 2 代。
population_size	每一代公式群体中的公式数量，默认为 1000。	1000	公式数量越大，消耗算力越多，公式之间组合的空间越大，这里使用默认值。
function_set	公式树中用于遗传和进化时所使用的函数集合，可自定义更多函数。	原始函数集+自定义函数集	**详细自定义函数集请见附录**
init_depth	公式树的初始化深度。init_depth 是一个二元组(min_depth, max_depth)，默认为(2, 6)。	(1, 4)	树的初始深度在 min_depth 与 max_depth 之间。最大深度越深，可能得出越复杂的因子，但是因子的意义更难解释。
torunament_size	在每一代的所有公式中，随机选择 torunament_size 个公式，其中适应度最高的公式能进行变异或繁殖生成下一代公式默认为 20	50	数值越小，随机选择范围越小，选择的结果越不确定。
metric	适应度指标，如 pearson 皮尔逊相关系数，spearman 斯皮尔曼秩相关系数。可自定义更多指标，默认为 pearson。	自定义适应度指标	**详细自定义适应度指标请见下小节**
p_crossover	父代进行交叉变异进化的概率，默认为 0.9。	0.5	交叉变异是最有效的进化方式，可以设置为较大概率。
p_subtree_mutation	父代进行子树变异进化的概率，默认为 0.01。	0.01	子树变异的结果不太稳定，概率不宜过大。这里使用默认值。
p_hoist_mutation	父代进行 Hoist 变异进化的概率，默认为 0.01。	0.01	Hoist 变异的结果不太稳定，概率不宜过大。这里使用默认值。
p_point_mutation	父代进行点变异进化的概率，默认为 0.01。	0.01	点变异的结果不太稳定，概率不宜过大。这里使用默认值。
p_point_replace	即点变异中父代每个节点进行变异进化的概率。默认为 0.05。	0.4	父代每个节点的点变异的概率已经很小了，可设置为较大概率保证点变异的执行。

参数	定义	取值	说明
n_components	选定最后的 n_components 个公式，默认为 10	10	这里使用默认值。
parsimony_coefficient	节俭系数，默认为 0.001。	0.001	惩罚过于复杂的公式，使用默认值。
max_samples	最大采样比例，默认为 1。	1	比率越大使用到的数据越多，考虑到输入变量带有时序性质，随机采样不能很好的区分样本内与样本外，因此默认使用全样本。
verbose	显示运行日志，默认为 1。	1	显示运行过程中的进化代数、适应度、花费时间等。
n_jobs	并行计算使用的核心数量，默认为 1。	-1	等于 -1 时,表示使用 cpu 里的所有的核进行工作，以提高运行速度。
random_state	随机状态，默认为 None。	666	随机种子数。
const_range	公式树中常数节点的取值范围，默认为（-1,1）。	None	为了提高因子的可解释性，我们不希望因子公式中出现常数，因此设置为 None。
stopping_criteria	早停指标，默认为 1。	10	适应度指标的阈值，满足相应条件即停止训练。本文为了适应自定义适应度指标，更改为 10。

资料来源：gplearn、中信期货研究所

2. gplearn 的改进

直接使用 gplearn 挖掘截面选期因子，是无法实现或者效果往往是不尽人意的，其原因有三点：

- gplearn 原始模型输入变量只能是 2 维数据，即 NumPy 中的 2 维数组。对应到因子挖掘过程中，输入变量只能是包含时间序列以及各项特征的单一资产，无法扩充到截面上多个资产。
- gplearn 所提供的原函数集（加、减、乘、除、开方、取对数、取绝对值等）较为简单且无法考虑时间序列上的运算（滚动窗口运算）。
- gplearn 自带适应度函数（IC、RankIC）较为简单，且出来的挖掘因子往往样本内的效果较好，样本外效果较差，即过拟合程度较高。

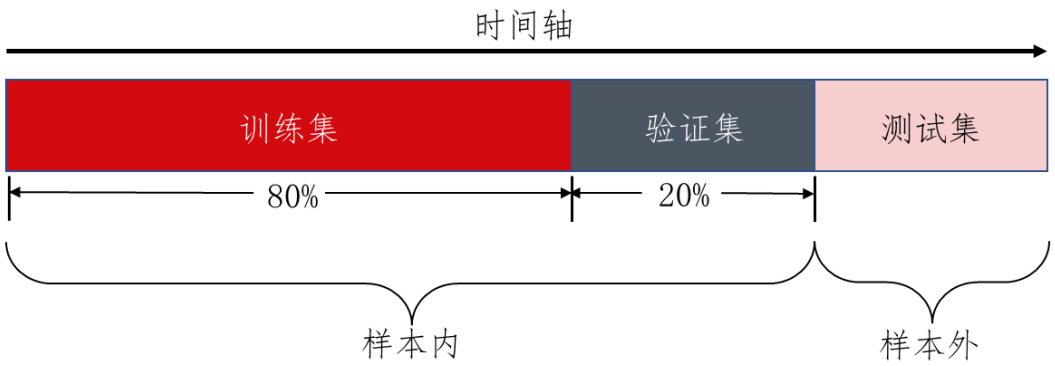
针对第一点，我们将输入的训练变量更改为 3 维数组（3D array），增添截面这一维度。修改过后的 3 维数组的 shape 为（x,y,z），其中 x 表征 OHLC、成交量、持仓量、仓单数等变量，y 表征时间序列，z 表征截面上的品种个数。我们通过修改原包中的大量函数，用以适配 3 维数组的运算。

针对第二点，我们对 gplearn 源代码进行修改后引入了一系列可以进行时间序列运算的函数以及 Ta-lib 包中的函数并将之汇总扩充到原函数集中，用以提高因子挖掘的能力，**具体算子函数表请见附录**。

最后我们针对第三点，自定义了适应度函数，用以提升因子挖掘效果。在初步尝试中，我们直接使用单因子回测后策略的夏普比率作为个体额适应度指标，但多次测试下来，因子也存在样本外失效较快的情况。对于此，我们进行简单的交叉验证，即将样本内的训练集进一步划分为 80%训练集+20%验证集，并分别计算训练集与验证集的夏普率。得出两者值后，再根据以下条件筛选因子：

- 训练集夏普率绝对值大于 1（取绝对值为了考虑因子方向）；
- 验证集夏普率/训练集夏普率取值较高者，继续进化；

图表8：简单交叉验证示意图



资料来源：中信期货研究所

以上便是我们对于 gplearn 的改造与优化，下一节我们将展示具体的因子挖掘流程与回测结果。

二、截面选期因子挖掘流程与结果

（一）截面选期因子挖掘流程

本节我们将遵循以下步骤进行因子挖掘：

1. 样本选择与回测细节

- 商品期货池：我们从国内期货市场筛选出历史流动性较好的 41 个品种具体如下表所示。

图表9：商品品种选择

类别	具体品种
黑色类	螺纹钢、热轧卷板、焦炭、焦煤、铁矿石、玻璃、纯碱
有色类	沪铜、沪铝、沪锌、沪镍、沪锡、不锈钢
能源类	原油、石油沥青、低硫燃料油、LPG、燃料油
化工类	PTA、乙二醇、短纤、甲醇、聚乙烯，聚丙烯，PVC，苯乙烯、尿素
软商品类	棉花、白糖、纸浆、橡胶
农产品类	豆粕、菜粕、棕榈油、豆油、菜油、玉米、生猪、鸡蛋、豆一、玉米淀粉

资料来源：中信期货研究所

- 交易价格：我们使用各个商品期货品种中主力合约的复权收盘价计算收益。
- 回测区间：我们使用 2016/1/1-2022/1/1 作为样本内训练集，2022/1/2-2023/3/10 作为样本外测试集，总回测区间为 2016/1/1-2023/3/10。
- 交易成本：暂不考虑任何交易中产生的费用。
- 杠杆倍数：一倍杠杆。
- 输入变量（初代种群）：如上文图表 7 所提，各个品种开盘价、收盘价、最高价、最低价、成交量、成交额、成交均价、持仓量、当日收盘价收益率、仓单数、仓单变化数 11 个原始因子。
- 目标变量：各个品种 2 个交易日后收益率，我们回测时在 T 日收盘计算因子、T+1 日做入、T+2 日产生收益。
- 单因子回测方法：五组分层回测法，默认使用因子排序前 20%品种与排序后 20%品种构建多空组合并计算收益（等权）。
- 调仓周期：每日调仓。

2. 使用 gplearn 进行因子挖掘

- 将我们的训练集与测试集输入模型，使用 SymbolicTransformer.fit()方法即可开始训练模型。在训练过程中，首先通过我们图表 7 中的算子函数集随机生成大量公式（因子表达式）。
- 通过自定义适应度函数计算每个个体的适应度值，并按照图表 1 中的遗传规划流程循环进化。在上一节自定义适应度函数中已经提到，我们会将样本内数据进一步拆分并进行简单交叉验证以对抗过拟合。
- 对挖掘出来的因子进行单因子分层回测，考察因子在样本内与样本外的表现。
- 尝试对回测效果较好的因子内在含义进行归因解释。

（二）截面选期因子挖掘结果及整体回测结果

经过多次挖掘，我们总结出了以下 5 个在全样本上都表现较好的截面选期 Alpha 因子。

图表10：Alpha 因子挖掘结果

因子名称	因子表达式	因子方向
Alpha1	ts_midpoint(ts_pct_change(ts_inverse_cv(ts_AROONOSC(volume,avgprice,10),10),5),486)	正向
Alpha2	ts_sum(ts_maxmin(ts_maxmin(warehouse,126),126),63)	负向
Alpha3	ts_ema(ts_sum(ts_rsi(ts_kama(openinterest, 486), 63), 243), 63)	负向
Alpha4	ts_ema(ts_inverse_cv (ts_corr(volume,avgprice,21),42) ,105)	负向
Alpha5	ts_dema(ts_median(ts_cov(open, volume,21),21),15)	负向

资料来源：中信期货研究所

下面是 Alpha 因子在样本内、样本外、以及全样本上的单因子回测结果。

图表11：Alpha 因子回测结果（样本内）

样本内	年化收益率	年化波动率	夏普比率	最大回撤	Calmar 比率
Alpha1	9.13%	5.22%	1.75	4.15%	2.20
Alpha2	12.52%	7.09%	1.77	6.05%	2.07
Alpha3	11.03%	8.33%	1.32	9.86%	1.12
Alpha4	8.44%	7.32%	1.15	8.16%	1.03
Alpha5	9.37%	7.09%	1.32	5.84%	1.61

资料来源：中信期货研究所

图表12：Alpha 因子回测结果（样本外）

样本外	年化收益率	年化波动率	夏普比率	最大回撤	Calmar 比率
Alpha1	9.97%	6.09%	1.64	3.29%	3.03
Alpha2	4.31%	6.40%	0.67	4.59%	0.94
Alpha3	8.34%	6.33%	1.32	3.37%	2.47
Alpha4	14.19%	7.82%	1.81	7.10%	2.00
Alpha5	4.22%	7.21%	0.59	9.00%	0.47

资料来源：中信期货研究所

图表13：Alpha 因子回测结果（全样本）

全样本	年化收益率	年化波动率	夏普比率	最大回撤	Calmar 比率
Alpha1	9.32%	5.42%	1.72	4.15%	2.25
Alpha2	11.07%	7.00%	1.58	6.05%	1.83
Alpha3	10.24%	7.74%	1.32	9.86%	1.05
Alpha4	9.35%	7.41%	1.26	8.16%	1.15
Alpha5	8.40%	7.11%	1.18	9.00%	0.93

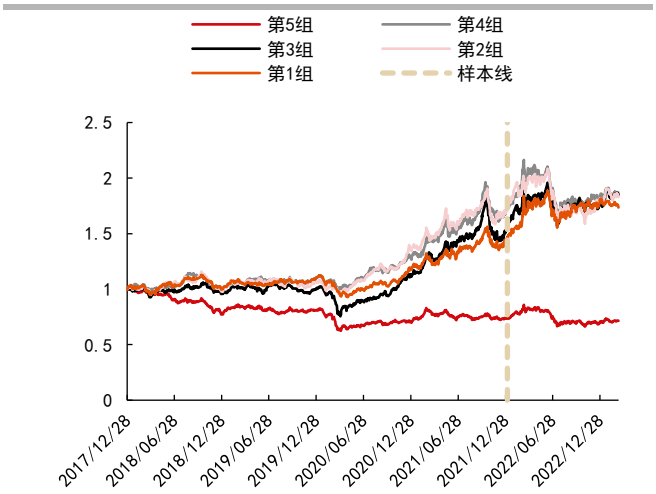
资料来源：中信期货研究所

（三）截面选期因子回测表现及简单归因

1.Alpha1 因子

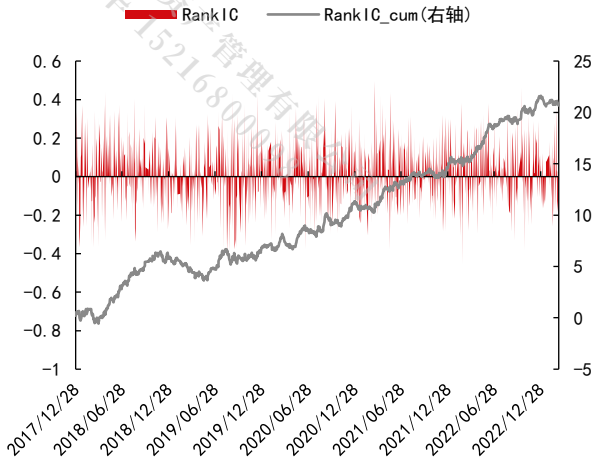
Alpha1 的因子表达式为 ts_midpoint(ts_pct_change(ts_inverse_cv(ts_AROONOSC(volume,avgprice,10),10),5),486)，该因子属于量价相关性类因子。它描述了短期量价相关性趋势的波动的变化率的长期走势。该因子为正向因子，则表明量价相关性趋势的波动的变化程度越高越好。下面是该因子在全样本上的具体回测结果，其中样本线以左表明样本内，样本线以右表明样本外。

图表14：Alpha1 分层回测净值



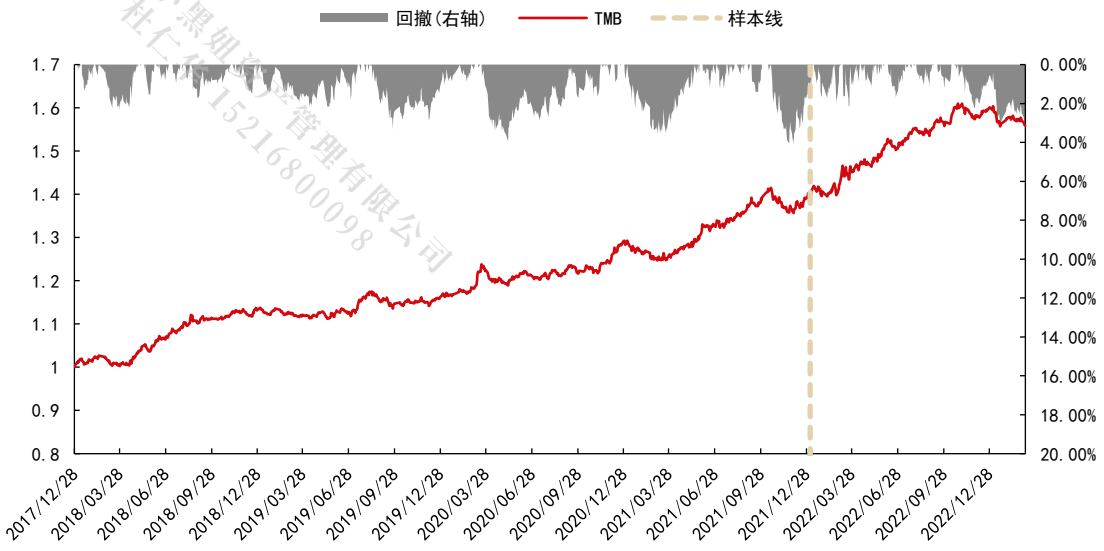
资料来源：中信期货研究所

图表15：Alpha1 RankIC 表现



资料来源：中信期货研究所

图表16：Alpha1 多空组合净值



资料来源：中信期货研究所

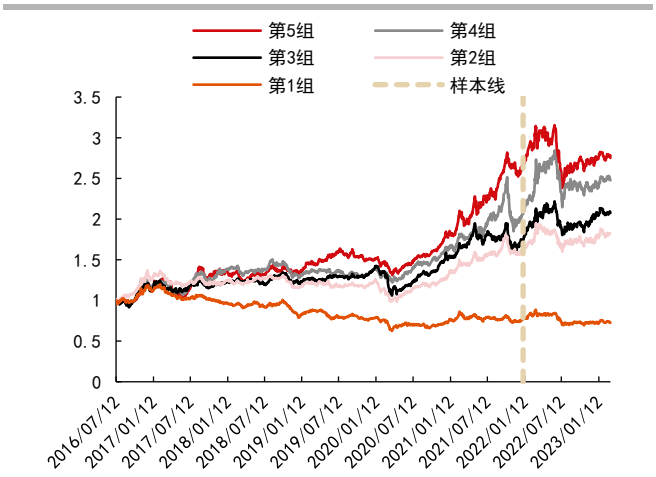
回测下来，该因子分层的单调性虽不明显，但是在时序上整体的 RankIC 累

计值与净值的走势都比较流畅，稳定性强，回测基本上小于 4%。

2. Alpha2 因子

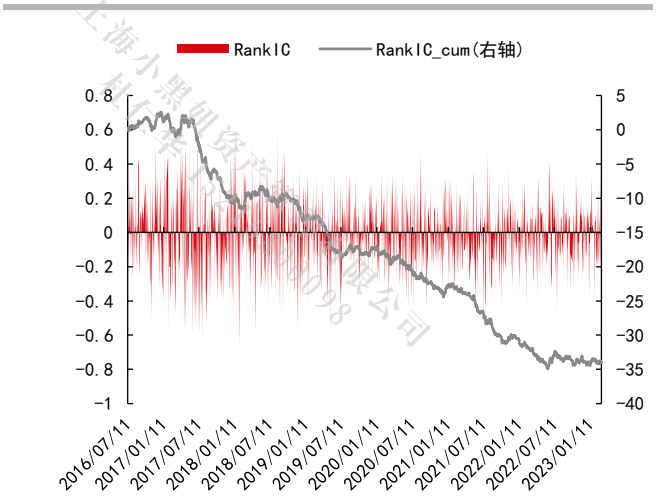
Alpha2 的因子表达式为 ts_sum(ts_maxmin(ts_maxmin(warehouse,126),126),63)，该因子属于基本面仓单类因子的衍生物。

图表17：Alpha2 分层回测净值



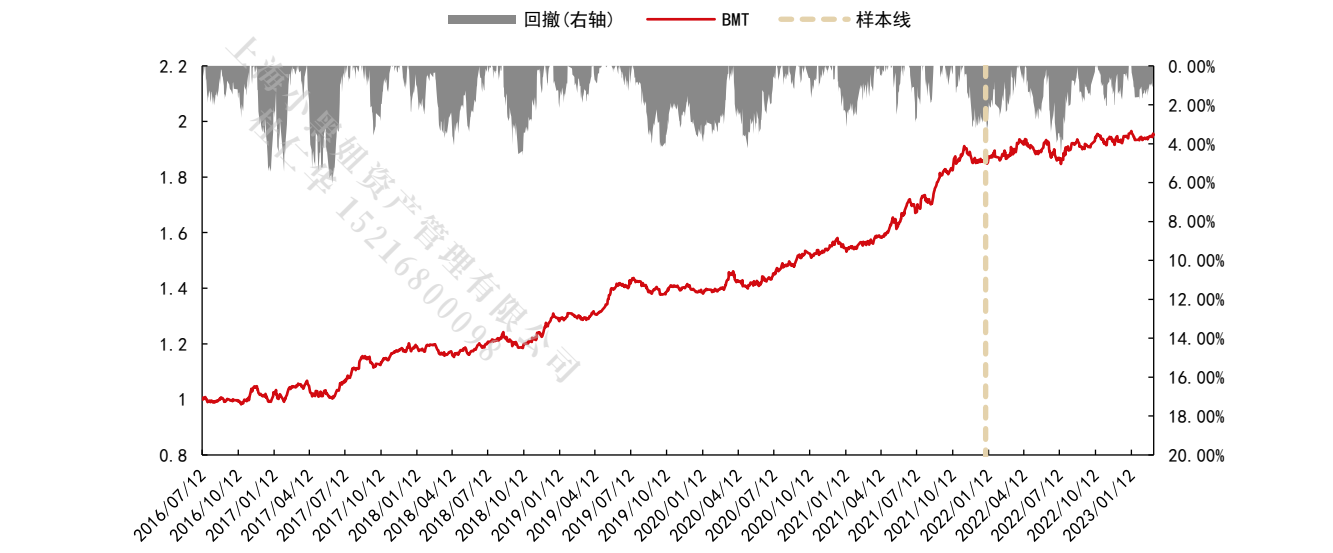
资料来源：中信期货研究所

图表18：Alpha2 RankIC 表现



资料来源：中信期货研究所

图表19：Alpha2 多空组合净值



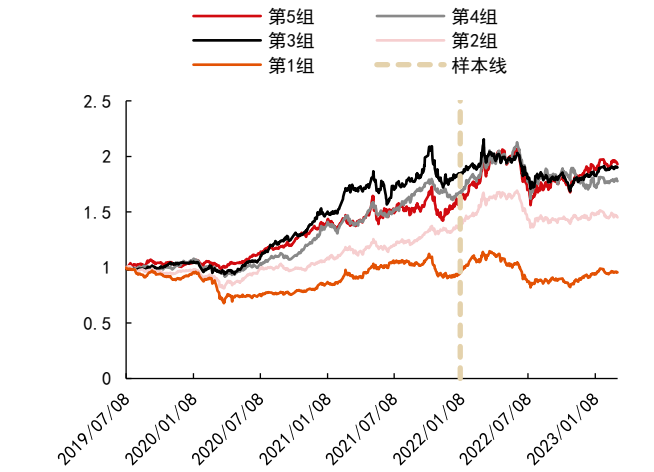
资料来源：中信期货研究所

回测结果表明，该因子分层的单调性强，RankIC 累计值与净值走势也呈现出强稳定性。样本外选期能力虽有所衰减，但仍然具备一定的有效性。

3. Alpha3 因子

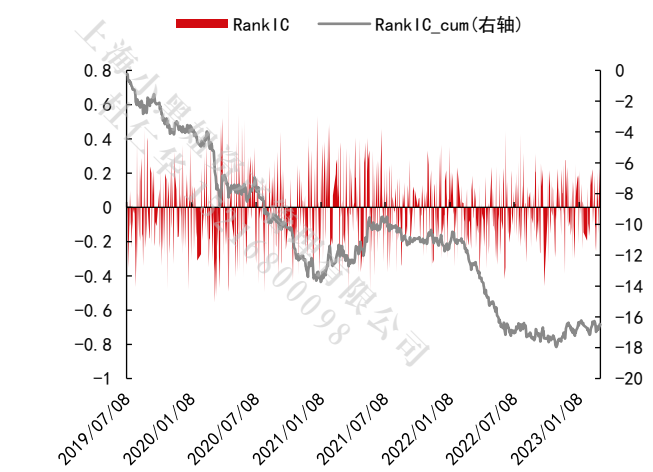
Alpha3 的因子表达式为 ts_ema(ts_sum(ts_rsi(ts_kama(openinterest, 486), 63), 243), 63)，它描述了持仓量的库夫曼移动均线在过去一段时间内的相对强弱程度之和的平均水平。该因子为负向因子，表明品种长期持仓量的均线持续走高时，倾向于做空该品种。

图表20：Alpha3 分层回测净值



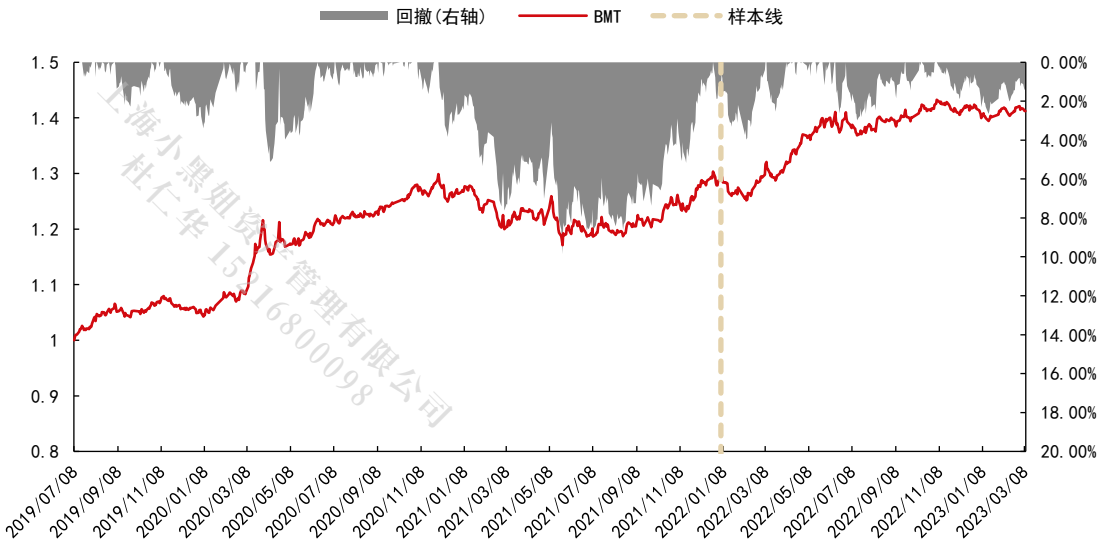
资料来源：中信期货研究所

图表21：Alpha3 RankIC 表现



资料来源：中信期货研究所

图表22：Alpha3 多空组合净值



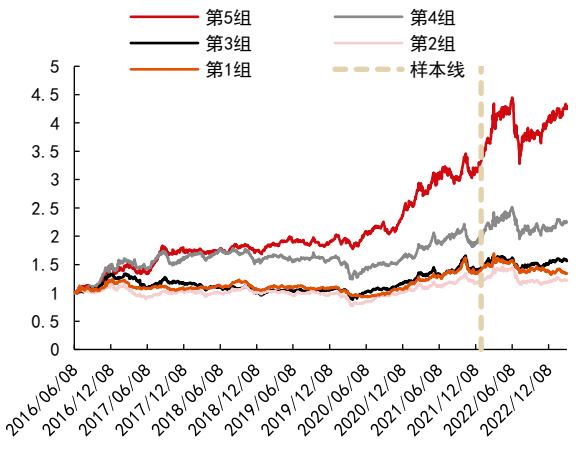
资料来源：中信期货研究所

Alpha3 因子在整个时序上分层的单调性较为一般，从 RankIC 累计值走势来看，样本内该因子预测方向也发生过较长时间的反转。但其在样本外的表现较为稳定，回测幅度较小。

4. Alpha4 因子

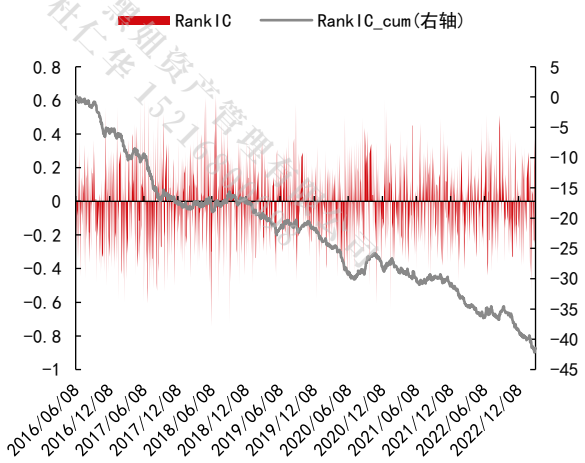
Alpha4 的因子表达式为 ts_ema(ts_inverse_cv(ts_corr(volume,avgprice,21),42),105)。它反映了回看期 21 日（一个月）的成交均价与成交量的相关性变化程度的指数移动平均。根据因子回测结果，该因子为负向因子，表明品种过去一个月的量价相关性在过去一段时间的变化程度越低越好。

图表23：Alpha4 分层回测净值



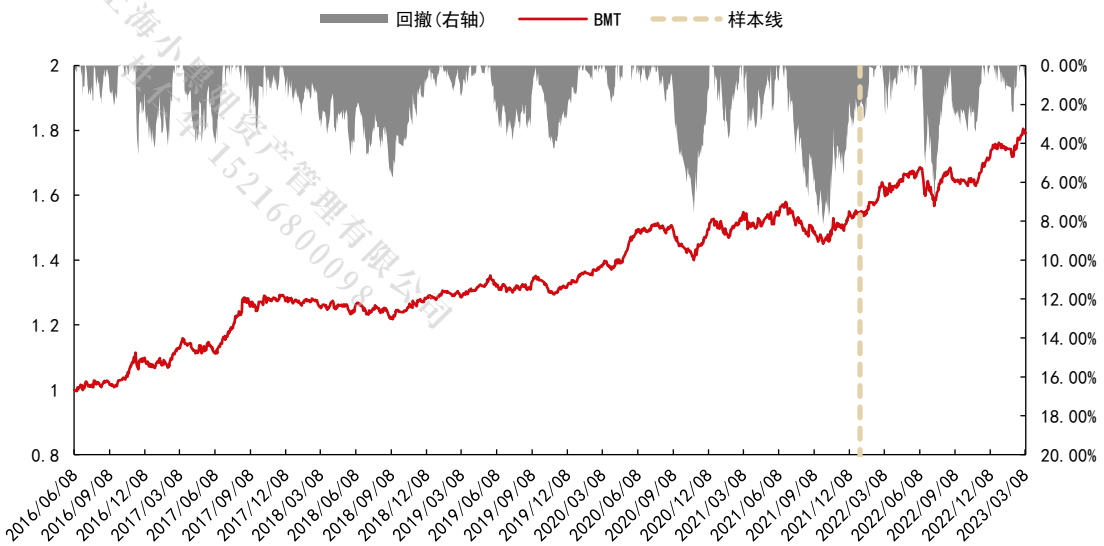
资料来源：中信期货研究所

图表24：Alpha4 RankIC 表现



资料来源：中信期货研究所

图表25：Alpha4 多空组合净值



资料来源：中信期货研究所

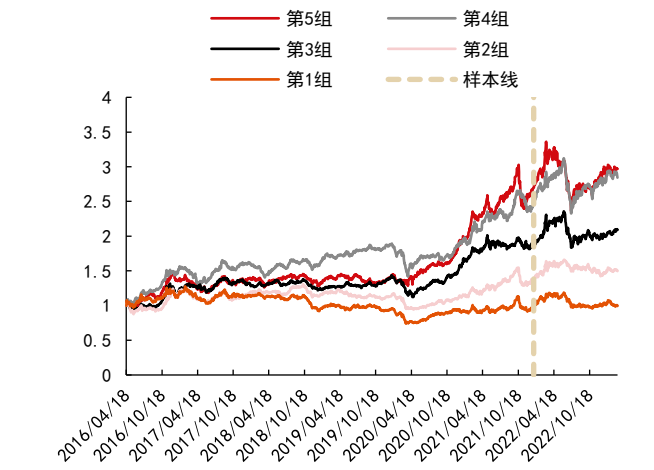
从回测结果来看，Alpha1 因子不仅在分层回测中表现出来较好的单调性（分

层明显）且整个时序上保持了较好的负向选期能力（RankIC 累计值稳定下行）。在样本外，Alpha1 因子也保持了较强的 alpha 能力且整体的回撤水平较低。

5. Alpha5 因子

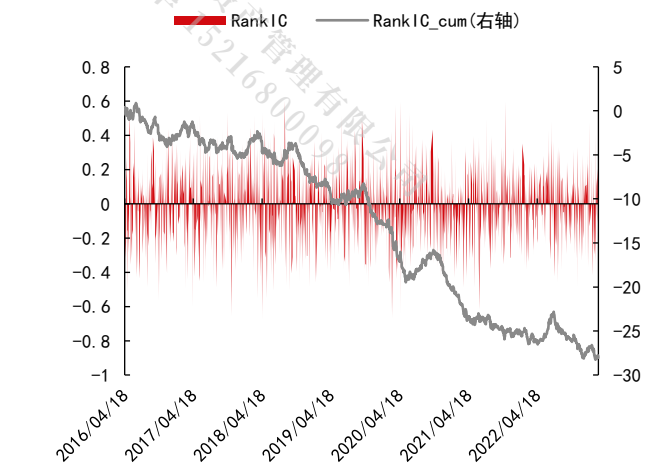
Alpha5 的因子表达式为 ts_dema(ts_median(ts_cov(open, volume, 21), 21), 15)，它同样可以归属为量价相关性类因子。它描述了开盘价与成交量的相关性中枢的移动平均水平。

图表26：Alpha5 分层回测净值



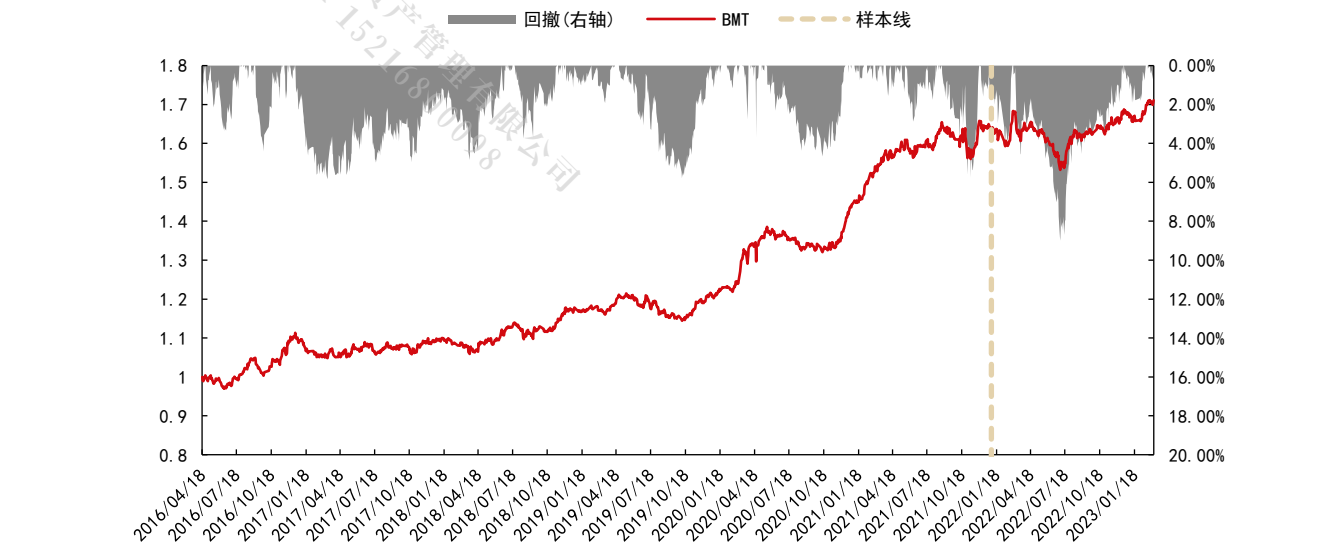
资料来源：中信期货研究所

图表27：Alpha5 RankIC 表现



资料来源：中信期货研究所

图表28：Alpha5 多空组合净值



资料来源：中信期货研究所

回测结果表明，该因子可以很好的区分截面上品种的强弱关系，且 RankIC 累计值在全样本上能维持稳定的下行走势。Alpha5 虽在样本外发生了最大回撤，但回撤时间不长且近期仍维持了较好的走势。

三、总结与思考

本文详细介绍了遗传规划算法的原理以及如何使用它进行因子挖掘。具体到实操流程上，我们使用到了 Python 中专门实现遗传规划的 gplearn 包进行代码实现。

在因子挖掘的过程中，我们对 gplearn 包中的源代码进行了改进与优化：

- 扩充了原始用于进化的算子函数集，使因子在进化过程中可以进行时间序列上运算。
- 修改了输入变量的维度使其适配于截面因子的挖掘。
- 自定义了适应度函数，以对抗过拟合程度并提高因子的可解释性。

在完成了以上修改后，我们利用了遗传规划算法，以 2016 至 2022 年作为输入数据在商品期货上挖掘了 5 个具有较好选期能力的截面 alpha 因子，并回测了 5 个因子在全样本上的表现：

- **Alpha1**：年化收益:9.32%,年化波动:5.42%,夏普:1.72,最大回撤:4.15%,卡玛比率:2.25;
- **Alpha2**：年化收益:11.07%,年化波动:7.00%,夏普:1.58,最大回撤:6.05%,卡玛比率:1.83;
- **Alpha3**：年化收益:10.24%,年化波动:7.74%,夏普:1.32,最大回撤:9.86%,卡玛比率: 1.05;
- **Alpha4**：年化收益:9.35%,年化波动:7.41%,夏普:1.26,最大回撤:8.16%,卡玛比率: 1.15;
- **Alpha5**：年化收益:8.40%,年化波动:7.11%,夏普:1.18,最大回撤:9.00%,卡玛比率:0.93;

回测下来，由算法挖掘出来的因子在样本内外均具备一定的有效性，表明利用遗传规划可以帮助我们归纳并总结出具有一定 alpha 能力的因子。但这种方法仍然存在着许多缺点。第一，因子的挖掘过程中具有很强的随机性，不同的输入特征变量、模型的超参数、算子函数、适应度函数都会影响最终的因子挖掘结果。第二，很容易产生过拟合的问题，即挖掘出来的因子往往在样本内表现较好而样

本外失效较快，即使本文使用了简单的交叉验证以对抗过拟合，但也无法保证样本外因子表现持续较好。第三，遗传规划挖掘的因子复杂程度普遍较高，绝大部分并不具有明确的经济意义。

因此在后续的研究中，可以从两方面进行改进以提升我们因子挖掘的效果，首先对输入特征变量进行特征工程使其转变成有一定的经济含义的因子，再进行挖掘，缩小挖掘的随机性并提高因子的可解释性。其次是可以在挖掘过程中加入时序交叉验证的步骤以更好地对抗因子过拟合的问题。

四、附录

我们在遗传规划挖掘因子的过程中使用到了以下算子函数，其中 X1，X2 为表征不同特征变量（open、close、…）的 2 维数组，每行对应每一个交易日，每列对应每一个品种，d 为时序中的回看天数。

图表29：算子函数集

类型	函数名	定义
原函数集	add(X1,X2)	返回 X1、X2 相加后的 2 维数组
原函数集	sub(X1,X2)	返回 X1、X2 相减后的 2 维数组
原函数集	mul(X1,X2)	返回 X1、X2 对应元素相乘后的 2 维数组
原函数集	div(X1,X2)	返回 X1、X2 对应元素相除后的 2 维数组
原函数集	abs（X1）	返回取绝对值后的 X1 数组
原函数集	sqrt（X1）	返回取开方后的 X1 数组
原函数集	log（X1）	返回取对数后的 X1 数组
原函数集	inv（X1）	返回取倒数后的 X1 数组
时间序列函数	ts_delay（X1，d）	滞后 d 日的 X1 数组
时间序列函数	ts_delta（X1，d）	X1 数组减滞后 d 日的 X1 数组
时间序列函数	ts_mean（X1，d）	X1 数组过去 d 日移动平均
时间序列函数	ts_pct_change（X1，d）	X1 数组过去 d 日的变化率
时间序列函数	ts_mean_return（X1，d）	X1 数组过去 1 日的变化率的 d 日移动平均
时间序列函数	ts_max（X1，d）	X1 数组过去 d 日最大值
时间序列函数	ts_min（X1，d）	X1 数组过去 d 日最小值
时间序列函数	ts_sum（X1，d）	X1 数组过去 d 日之和
时间序列函数	ts_product（X1，d）	X1 数组过去 d 日乘积
时间序列函数	ts_std（X1，d）	X1 数组过去 d 日标准差
时间序列函数	ts_median（X1，d）	X1 数组过去 d 日中位数
时间序列函数	ts_midpoint（X1，d）	X1 数组过去 d 日最大值与最小值的均值
时间序列函数	ts_skew（X1，d）	X1 数组过去 d 日偏度
时间序列函数	ts_kurt（X1，d）	X1 数组过去 d 日峰度
时间序列函数	ts_inverse_cv（X1，d）	X1 数组过去 d 日变异系数的倒数
时间序列函数	ts_cov（X1，X2，d）	X1 数组与 X2 数组过去 d 日的协方差
时间序列函数	ts_corr（X1，X2，d）	X1 数组与 X2 数组过去 d 日的相关系数
时间序列函数	ts_maxmin（X1，d）	（X1-ts_min（X1，d））/（ts_max（X1，d）- ts_min（X1，d））
时间序列函数	ts_zscore（X1，d）	X1 数组过去 d 日的 z-score 值
时间序列函数	ts_regression_beta（X1，X2，d）	X1 数组与 X2 数组过去 d 日的回归系数
时间序列函数	ts_linear_slope（X1，d）	X1 数组与时序（t=1,2,…,d）的回归系数
时间序列函数	ts_linear_intercept（X1，d）	X1 数组与时序（t=1,2,…,d）的回归截距
时间序列函数	ts_argmax（X1，d）	X1 数组过去 d 日的最大值索引值
时间序列函数	ts_argmin（X1，d）	X1 数组过去 d 日的最小值索引值
时间序列函数	ts_argmaxmin（X1，d）	ts_argmax（X1，d）- ts_argmin（X1，d）
时间序列函数	ts_rank（X1，d）	X1 数组过去 d 日的从小到大排序值
Ta-lib 函数	ts_ema（X1，d）	X1 数组过去 d 日的指数移动平均

类型	函数名	定义
Ta-lib 函数	ts_dema（X1，d）	X1 数组过去 d 日的双指数移动平均
Ta-lib 函数	ts_kama（X1，d）	X1 数组过去 d 日的考夫曼自适应移动平均
Ta-lib 函数	ts_wma（X1，d）	X1 数组过去 d 日的加权移动平均
Ta-lib 函数	ts_mom（X1，d）	X1 数组过去 d 日的动量指标
Ta-lib 函数	ts_cmo（X1，d）	X1 数组过去 d 日的钱德动量摆动指标
Ta-lib 函数	ts_roc（X1，d）	X1 数组过去 d 日的变动率指标
Ta-lib 函数	ts_AROONOSC（X1，X2，d）	X1 数组与 X2 数组过去 d 日的阿隆震荡指标

资料来源：Ta-lib、中信期货研究所

免责声明

除非另有说明，中信期货有限公司拥有本报告的版权和/或其他相关知识产权。未经中信期货有限公司事先书面许可，任何单位或个人不得以任何方式复制、转载、引用、刊登、发表、发行、修改、翻译此报告的全部或部分材料、内容。除非另有说明， 本报告中使用的所有商标、服务标记及标记均为中信期货有限公司所有或经合法授权被许可使用的商标、服务标记及标记。未经中信期货有限公司或商标所有权人的书面许可，任何单位或个人不得使用该商标、服务标记及标记。

如果在任何国家或地区管辖范围内，本报告内容或其适用与任何政府机构、监管机构、自律组织或者清算机构的法律、规则或规定内容相抵触，或者中信期货有限公司未被授权在当地提供这种信息或服务，那么本报告的内容并不意图提供给这些地区的个人或组织，任何个人或组织也不得在当地查看或使用本报告。本报告所载的内容并非适用于所有国家或地区或者适用于所有人。

此报告所载的全部内容仅作参考之用。此报告的内容不构成对任何人的投资建议，且中信期货有限公司不会因接收人收到此报告而视其为客户。

尽管本报告中所包含的信息是我们于发布之时从我们认为可靠的渠道获得，但中信期货有限公司对于本报告所载的信息、观点以及数据的准确性、可靠性、时效性以及完整性不作任何明确或隐含的保证。因此任何人不得对本报告所载的信息、观点以及数据的准确性、可靠性、时效性及完整性产生任何依赖，且中信期货有限公司不对因使用此报告及所载材料而造成的损失承担任 何责任。本报告不应取代个人的独立判断。本报告仅反映编写人的不同设想、见解及分析方法。本报告所载的观点并不代表中信期货有限公司或任何其附属或联营公司的立场。

此报告中所指的投资及服务可能不适合阁下。我们建议阁下如有任何疑问应咨询独立投资顾问。此报告不构成任何投资、法律、会计或税务建议，且不担保任何投资及策略适合阁下。此报告并不构成中信期货有限公司给予阁下的任何私人咨询建议。

深圳总部

地址：深圳市福田区中心三路 8 号卓越时代广场（二期）北座 13 层 1301-1305、14 层

邮编：518048

电话：400-990-8826

传真：(0755)83241191

网址：http://www.citicsf.com