

投资咨询业务资格：
证监许可【2012】669号

风险提示：本报告中所涉及的交易信号
和标的信息仅为回溯举例，并不构成推
荐建议

利用遗传规划挖掘商品期货截面因子

研究所 金融工程组
2023年3月



中信期货有限公司
CITIC Futures Company Limited

【重要提示：本报告难以设置访问权限，若给您造成不便，敬请谅解。我司不会因为关注、收到或阅读本报告内容而视相关人员为客户；市场有风险，投资需谨慎。】

周通
021-80401733
从业资格号F3078183
投资咨询号Z0018055



01

因子挖掘方法论

遗传规划算法简介+gplearn使用与改进

02

因子挖掘流程与结果

因子挖掘结果+因子回测结果

03

总结与思考

回顾+思考

01

因子挖掘方法论

遗传规划算法简介+gplearn使用与改进

1.1 因子挖掘方法论

- 因子挖掘方法：

- 演绎法VS归纳法

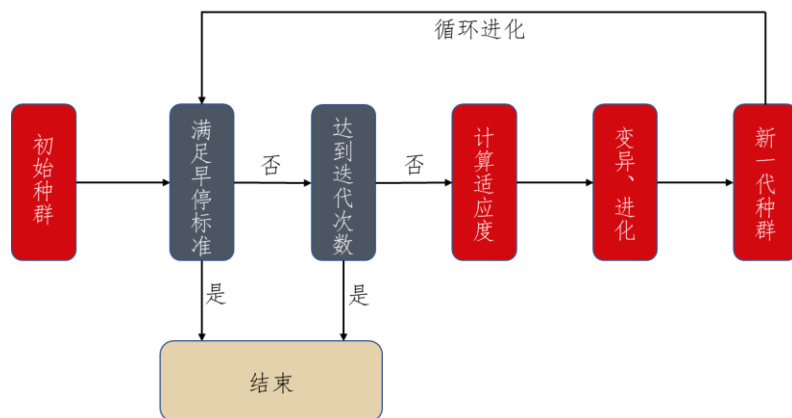
- 演绎法：“先有逻辑，后有公式”，如传统截面动量因子。

- 归纳法：“先有公式，后有逻辑”，遗传规划算法。

1.2 遗传规划算法简介

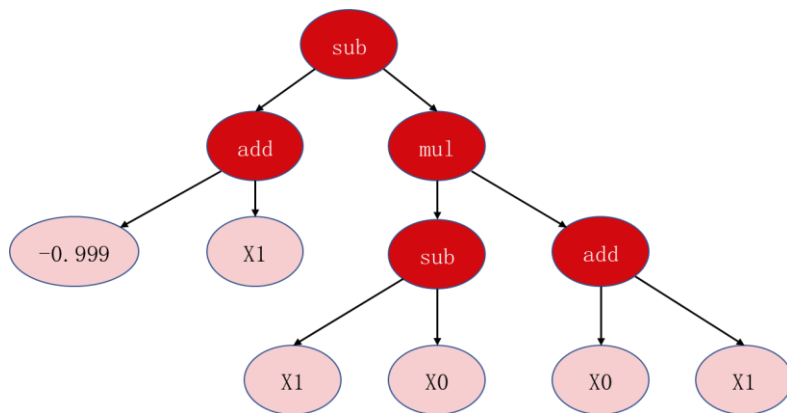
- 遗传算法是一种通过模拟自然界生物进化过程（“物竞天择、适者生存”）搜索最优解的算法。

遗传算法图解



- 遗传规划是遗传算法中的一个分支，主要用于符号回归：

公式树

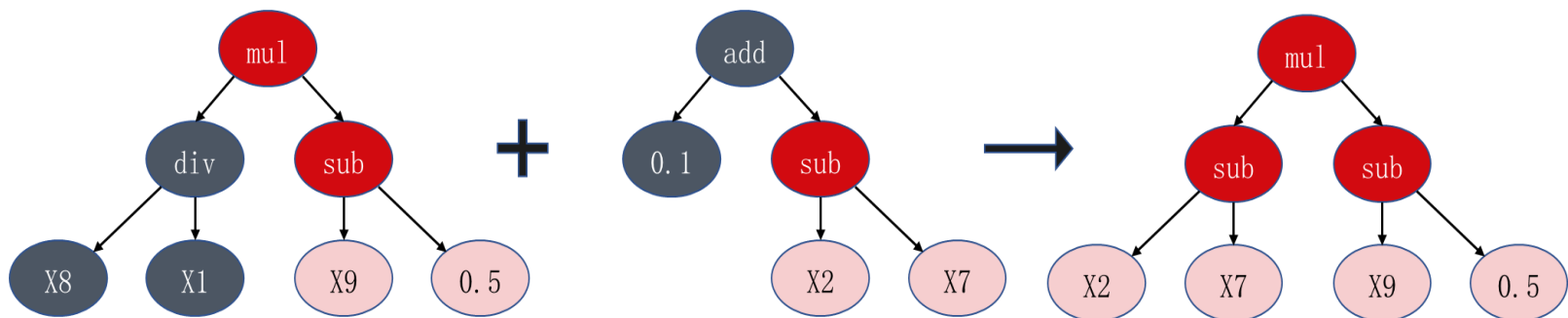


1.3 适应度函数

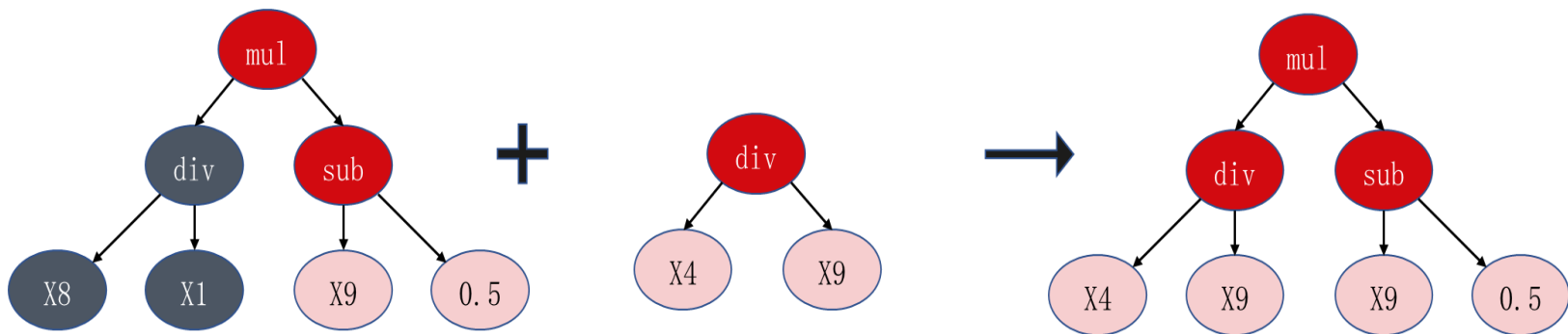
- 适应度函数用来计算每一代种群中个体的适应度，适应度则衡量了个体与最终的目标个体的相符程度。
- 每一代进化过程中，只保留适应度较高的个体作为后续进化的种群，适应度较低的个体则被淘汰。
- 从因子挖掘的角度来说，我们可以使用用来评判因子有效程度的指标做为个体适应度，比如：
 - 因子的RankIC、RankICIR可以挖掘与未来收益率相关程度较高的线性因子；
 - 互信息（Mutual information）可以用来挖掘与识别非线性的因子；
 - 或是直接使用因子回测后的夏普率（sharpe ratio）作为适应度；

1.4 公式树的进化方式

- 杂交变异：最主要的变异方式，两个个体遗传物质的分配融合。

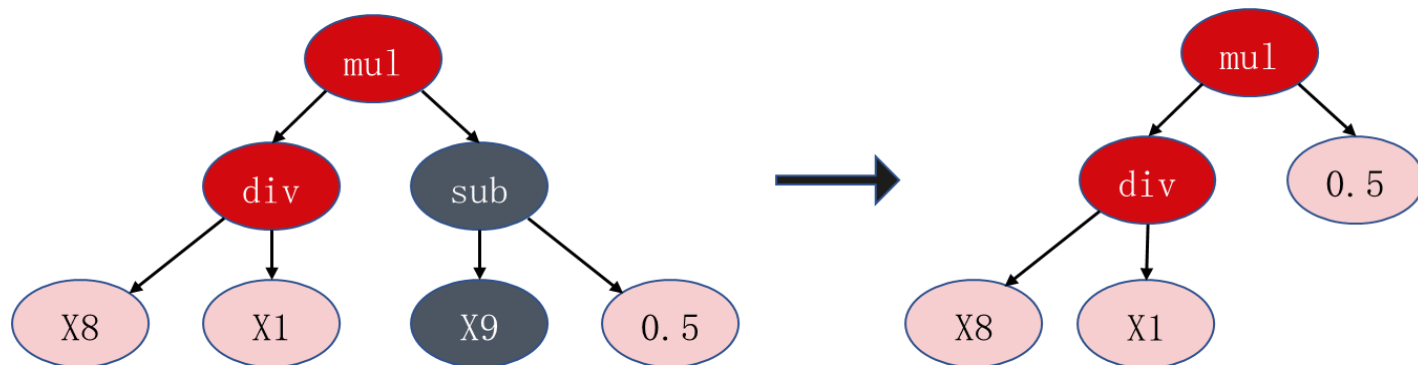


- 子树变异：十分激进的变异方式，整个子树变异。

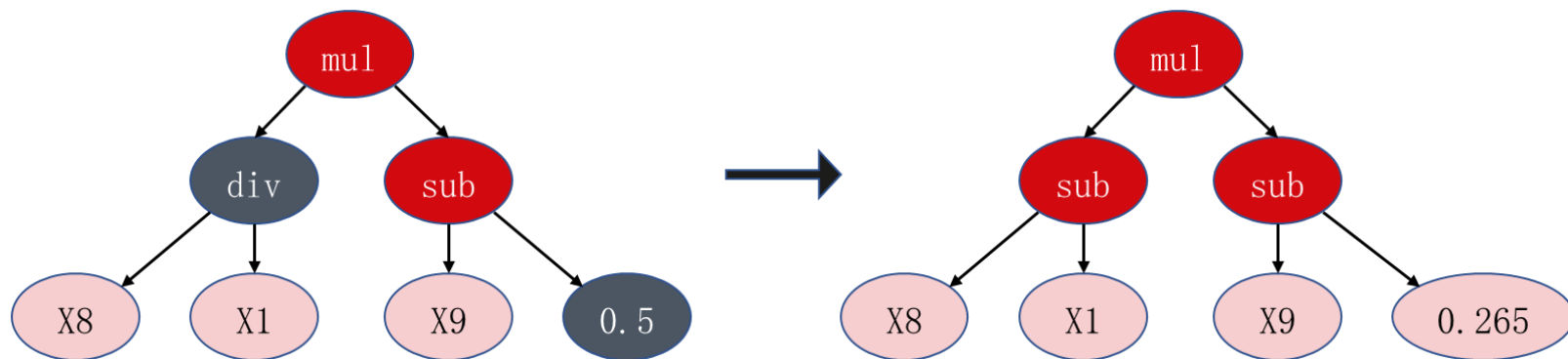


1.4 公式树的进化方式

- Hoist（抬升）变异：十分激进的变异方式，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”。



- 点变异：常见的变异方式，随机某个节点变异。



1.5 gplearn的使用与改进

- 我们通过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，其中gplearn中的特征转化器SymbolicTransformer可以用来挖掘因子。
- SymbolicTransformer主要参数设置与说明如下：

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, avgp rice, openinterest, retur n_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。
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
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
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。

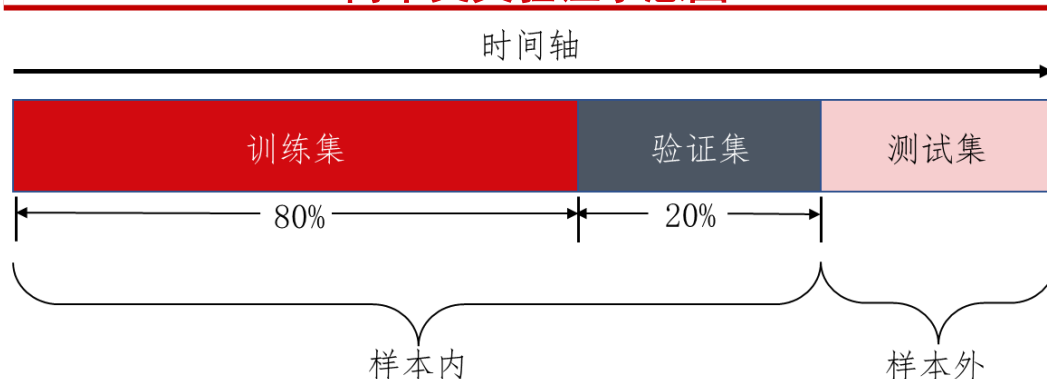
1.5 gplearn的使用与改进

- gplearn无法直接用于截面因子挖掘，其原因是：
 - gplearn原始模型输入变量只能是2维数据，无法考虑截面多个资产。
 - 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。
- 针对以上三点我们做出了以下改进：
 - 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。
 - 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。附录

1.5 gplearn的使用与改进

- 自定义适应度函数为单因子回测后的夏普率，并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；

简单交叉验证示意图



02

因子挖掘流程与结果

因子挖掘结果+因子回测结果

2.1 样本选择与回测细节

■ 商品期货池：

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

■ 交易价格：主力合约复权收盘价

■ 回测区间：总回测区间为2016/1/1-2023/3/10， 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。

■ 交易成本：暂不考虑。

■ 杠杆倍数：一倍杠杆。

■ 输入变量：上文所提开盘价、收盘价、成交量、仓单数等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，因子值前20%，后20%构建多空组合。

■ 调仓周期：日度

2.1 因子挖掘结果及回测结果

- 经过多次挖掘，我们总结出了以下5个Alpha因子：

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

- 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。

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

2.1 因子挖掘结果及回测结果

■ 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

■ 全样本：

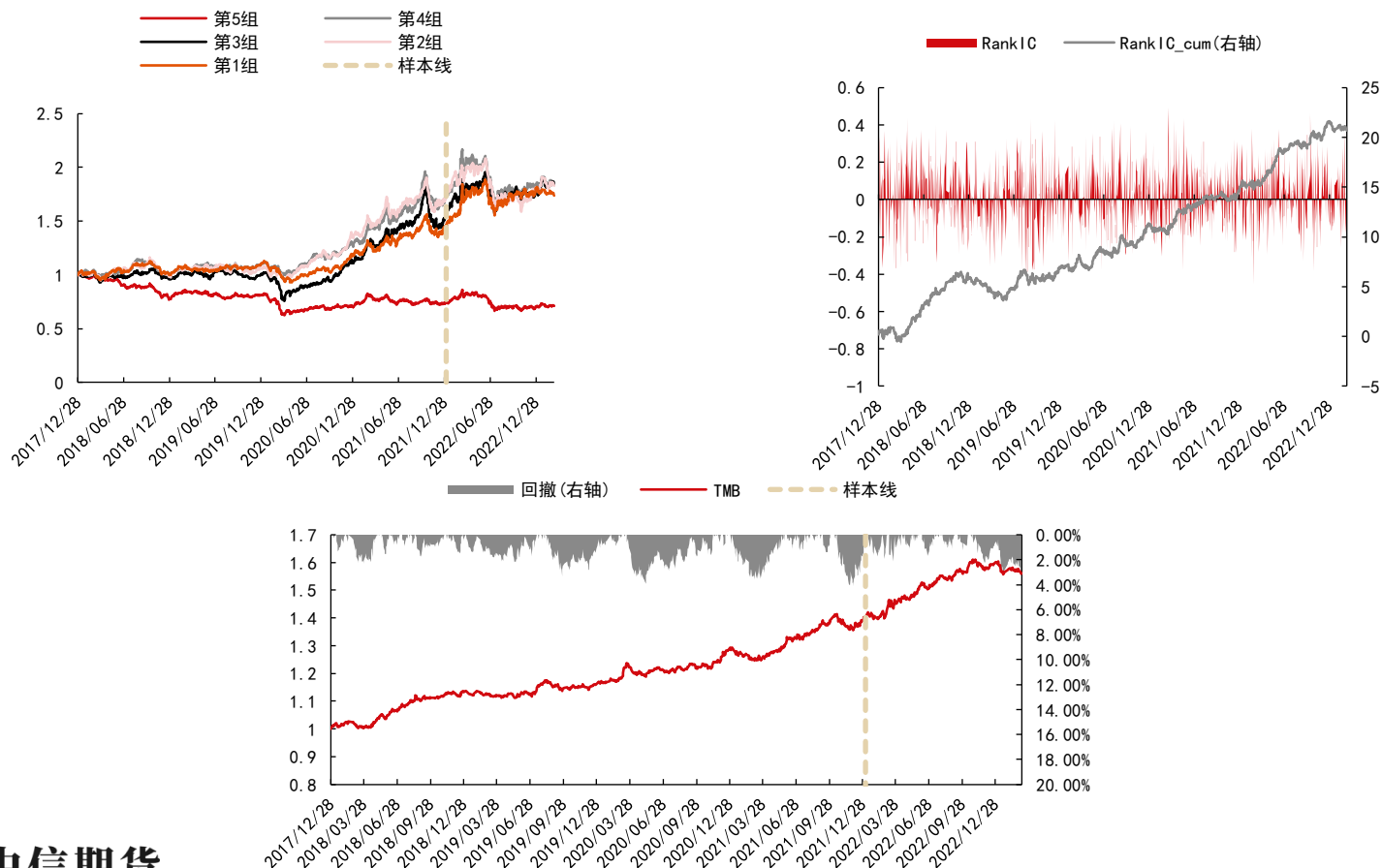
全样本	年化收益率	年化波动率	夏普比率	最大回撤	Calmar比率
Alpha1	9.32%	5.42%	1.72	4.15%	2.25
Alpha2	11.07%	7.00%	1.58	6.05%	1.83
Alpha3	10.24%	7.74%	1.32	9.86%	1.05
Alpha4	9.35%	7.41%	1.26	8.16%	1.15
Alpha5	8.40%	7.11%	1.18	9.00%	0.93

2.2 Alpha1

Alpha1的因子表达为

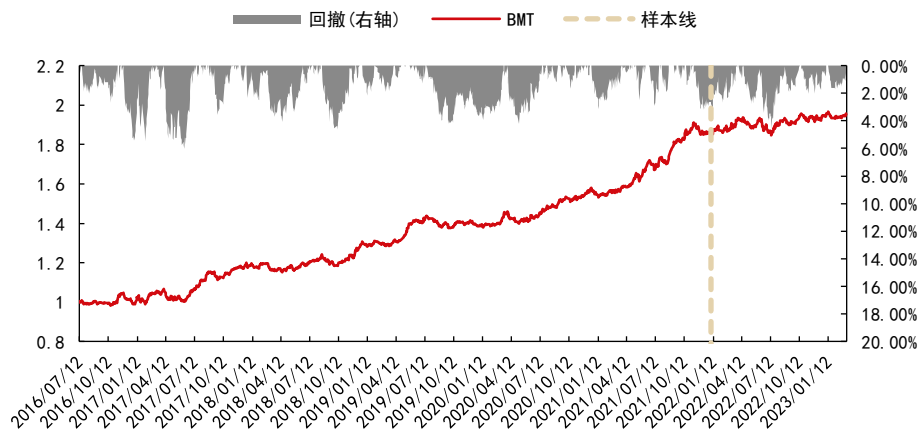
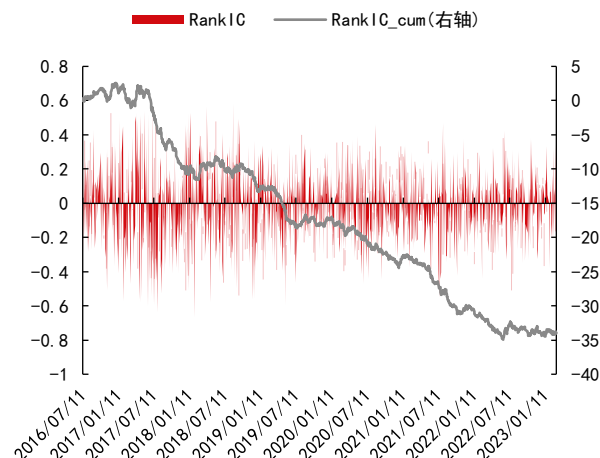
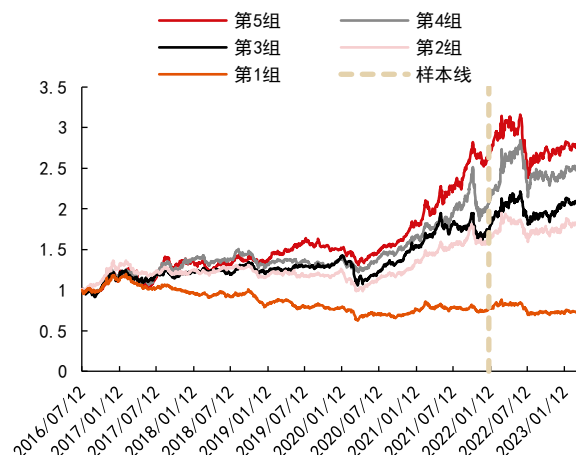
$ts_midpoint(ts_pct_change(ts_inverse_cv(ts_AROONOSC(volume, avgprice, 10), 10), 5), 486)$

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。



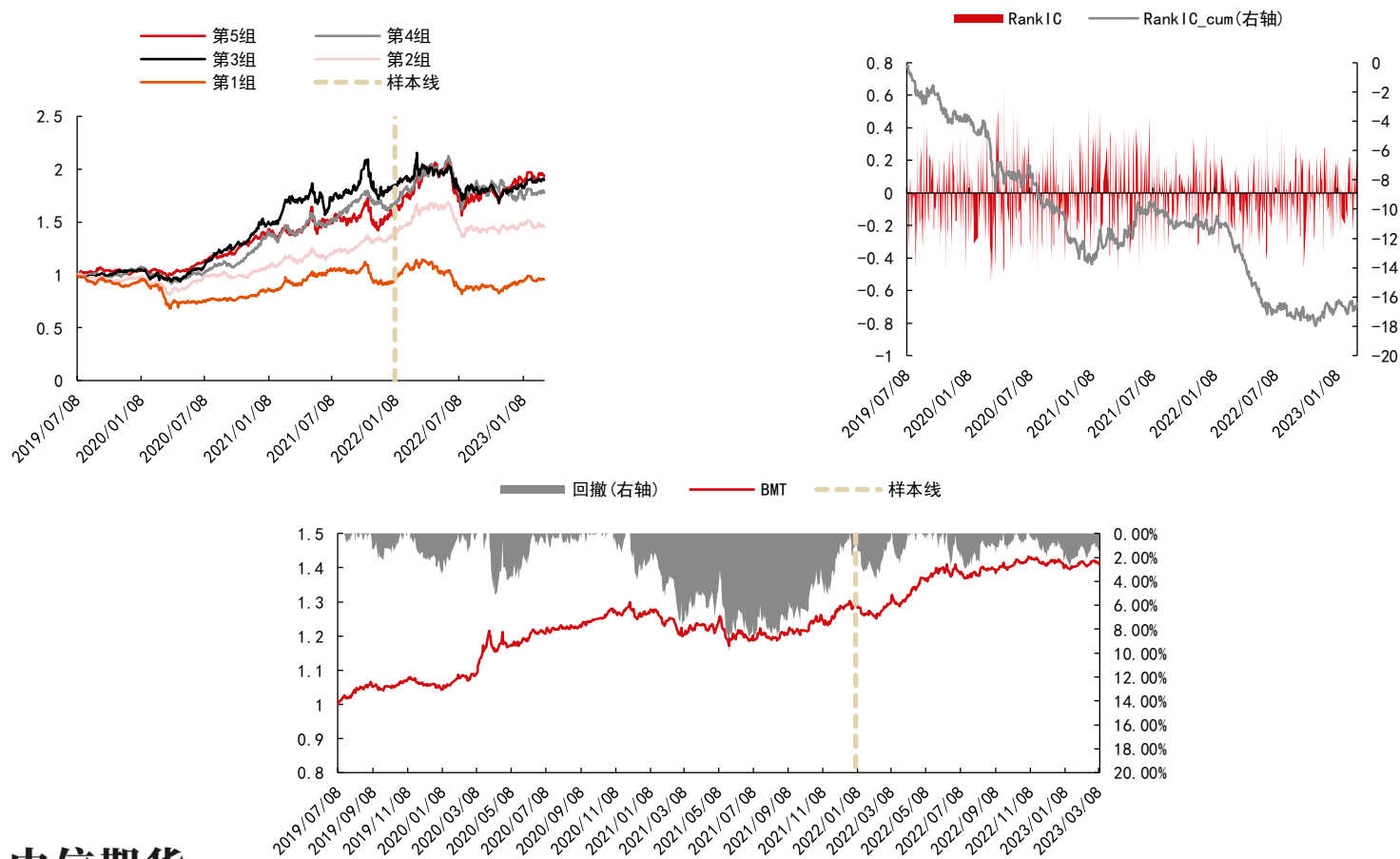
2.2 Alpha2

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



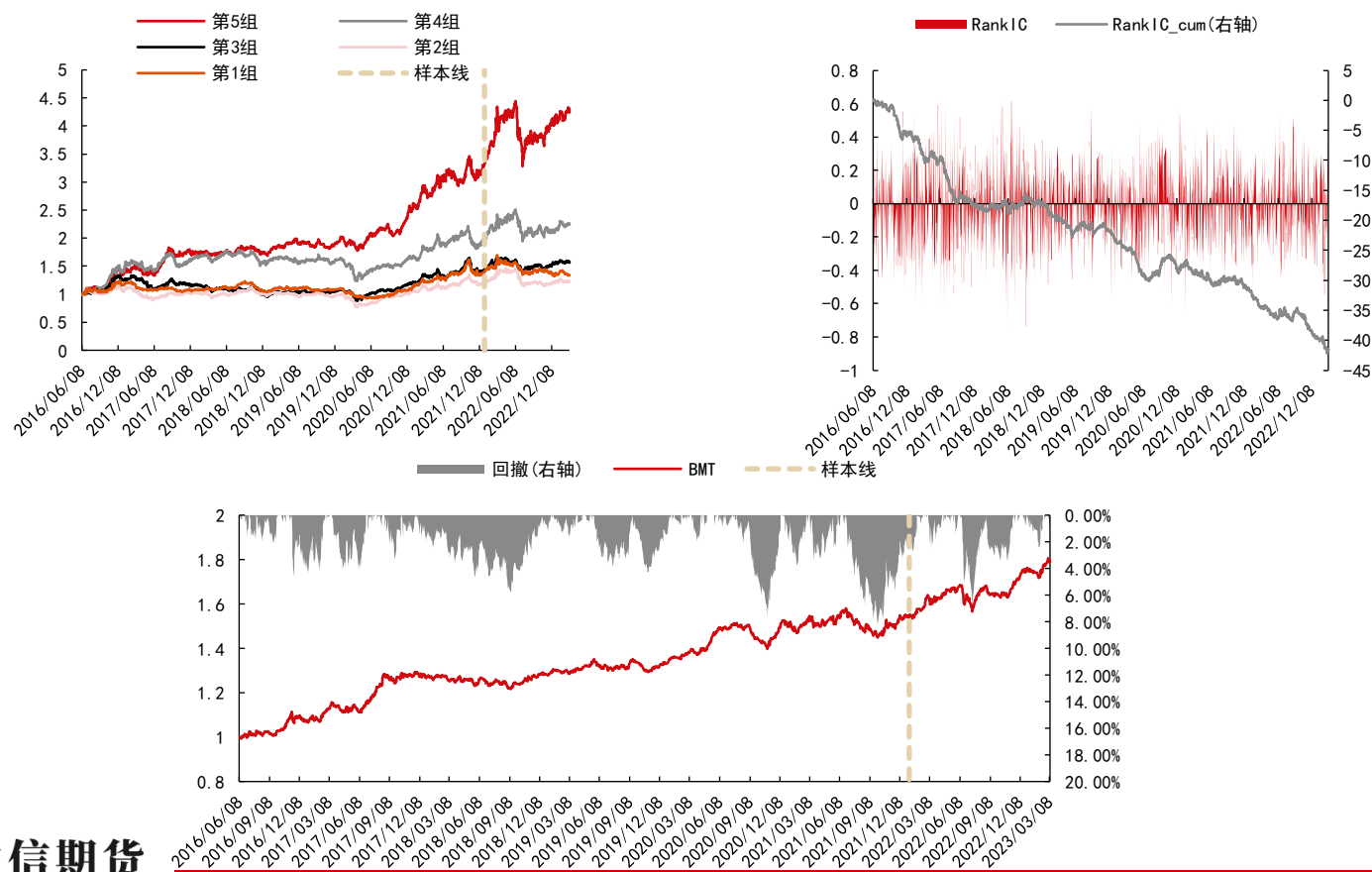
2.2 Alpha3

- Alpha3的因子表达为 $ts_ema(ts_sum(ts_rsi(ts_kama(openinterest, 486), 63), 243), 63)$ 。
- 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。



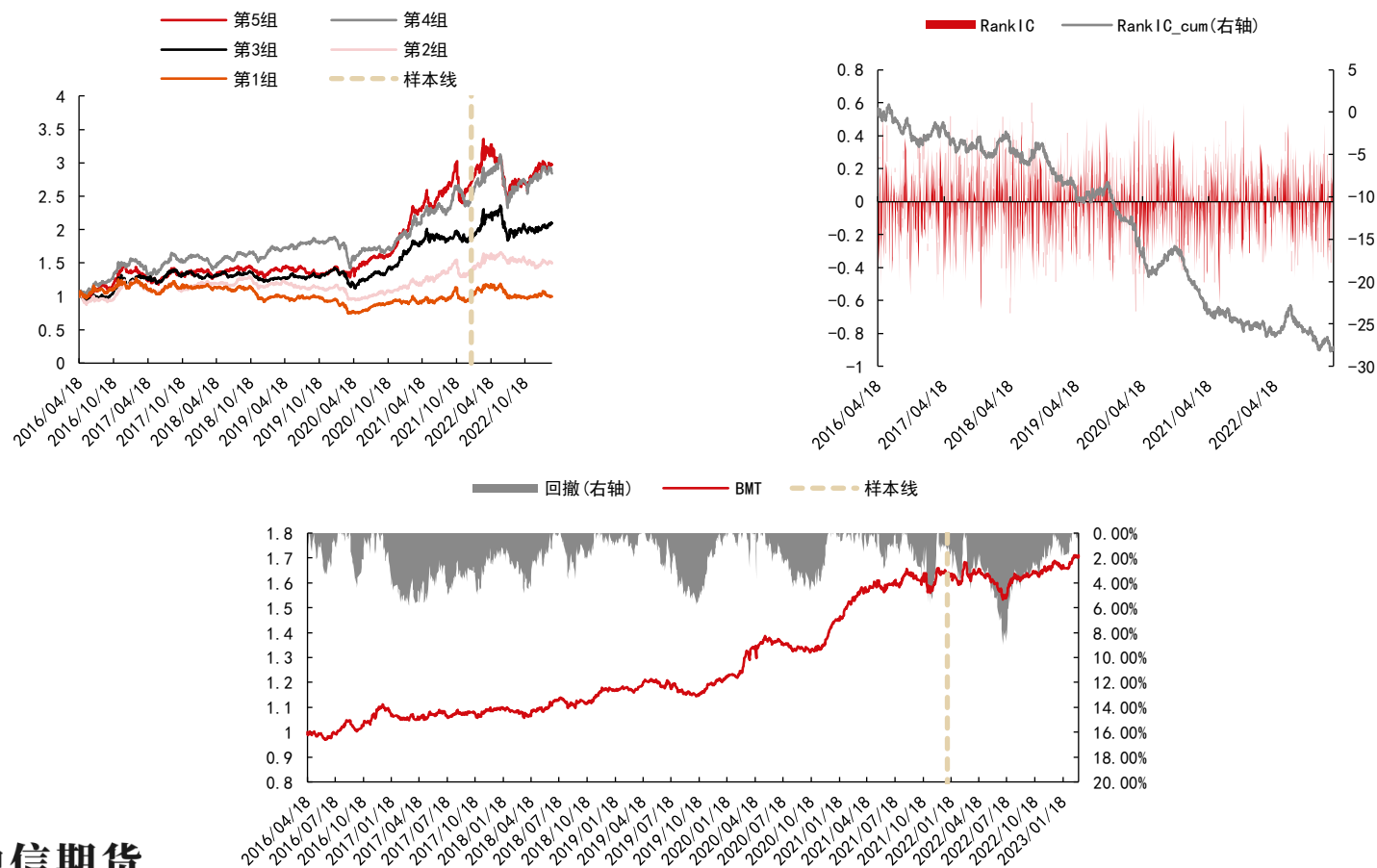
2.2 Alpha4

- Alpha4的因子表达为 $\text{ts_ema}(\text{ts_inverse_cv}(\text{ts_corr}(\text{volume}, \text{avgprice}, 21), 42), 105)$
- 该因子属于量价相关性类因子，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



2.2 Alpha5

- Alpha5的因子表达为 $ts_dema(ts_median(ts_cov(open, volume, 21), 21), 15)$
- 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。



03

总结与思考

回顾+思考

3 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。
- 回测结果：
 - 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；

3 总结与思考

- 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。
- 缺点：
 - 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。
 - 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。
- 改进方向：
 - 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。

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。

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

资料来源：中信期货研究所

[返回](#)

类型	函数名	定义
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
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：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，且不承担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。

中信期货有限公司

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



资料来源：中信期货研究所

【重要提示：本报告难以设置访问权限，若给您造成不便，敬请谅解。我司不会因为关注、收到或阅读本报告内容而视相关人员为客户；市场有风险，投资需谨慎。】



中信期货
CITIC Futures

中信期货有限公司

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室、14层

上海地址：

上海市浦东新区杨高南路799号陆家嘴世纪金融
广场
3号楼23层

致謝