



华泰期货  
HUATAI FUTURES

期货研究报告|量化专题报告 2023-08-29

# 原油基本面量化策略

## 研究院 量化组

### 研究员

#### 高天越

☎ 0755-23887993

✉ gaotianyue@htfc.com

从业资格号: F3055799

投资咨询号: Z0016156

### 联系人

#### 李逸资

☎ 0755-23887993

✉ liyizi@htfc.com

从业资格号: F03105861

#### 李光庭

☎ 0755-23887993

✉ liguangting@htfc.com

从业资格号: F03108562

投资咨询业务资格:

证监许可【2011】1289号

## 摘要

原油作为基础能源,是重要的工业原料,经加工能得到柴油、汽油、煤油、沥青等多种石油化工产品,产业链涉及产油、炼油、运输、库存、消费等多个环节。因其涉及面广,原油价格的波动对于整个能化产业链而言称得上是牵一发而动全身的存在。

本文将从原油的产业链逻辑出发,深度挖掘上下游各个环节涉及到的基本面数据,通过精细化的因子衍生及因子降维,并构建出系统科学的因子有效性的打分机制,实现从海量数据中筛选出真正对预测原油价格有效的重点指标;然后再重点因子的基础上,利用机器学习模型对未来的预测值,设计出多因子、多频率的量价交易策略。

目录

摘要 ..... 1

因子挖掘 ..... 4

    ■ 研究背景 ..... 4

    ■ 指标收集 ..... 5

    ■ 初步筛选 ..... 5

    ■ 数据处理 ..... 8

    ■ 组内选优 ..... 8

模型搭建 ..... 9

    ■ 确定因子数量 ..... 9

    ■ 模型互融 ..... 10

    ■ 样本内最优因子 ..... 11

    ■ 预测准度 ..... 12

量化策略 ..... 13

    ■ 基础设定 ..... 13

    ■ 日内策略 ..... 13

    ■ 日间策略 ..... 15

## 图表

图 1:原油产业链   单位: 无	4
图 2:原油的储运、加工与使用   单位: 无	4
图 3: 影响原油价格因子分析框架   单位: 无	5
图 4: RFECV 结果显示最佳因子数量为 11   单位: 无	10
图 5: 网格遍历确定最优权重组合   单位: 无	11
图 6: 日内策略净值对比   单位: 无	14
图 7: 日间策略净值对比   单位: 无	15
图 8: 固定基本面日内策略   单位: 无	16
图 9: 固定基本面日间策略   单位: 无	16
图 10: 滚动基本面日内策略   单位: 无	16
图 11: 滚动基本面日间策略   单位: 无	16
图 12: 固定量价日内策略   单位: 无	16
图 13: 固定量价日间策略   单位: 无	16
表 1: 基本面数据举例 (部分)   单位: 无	6
表 2: 样本内最优回归因子 (前 15 名)   单位: 无	11
表 3: 模型预测准确性对比   单位: 无	12
表 4: 日内策略表现   单位: 无	14
表 5: 日间策略表现   单位: 无	15

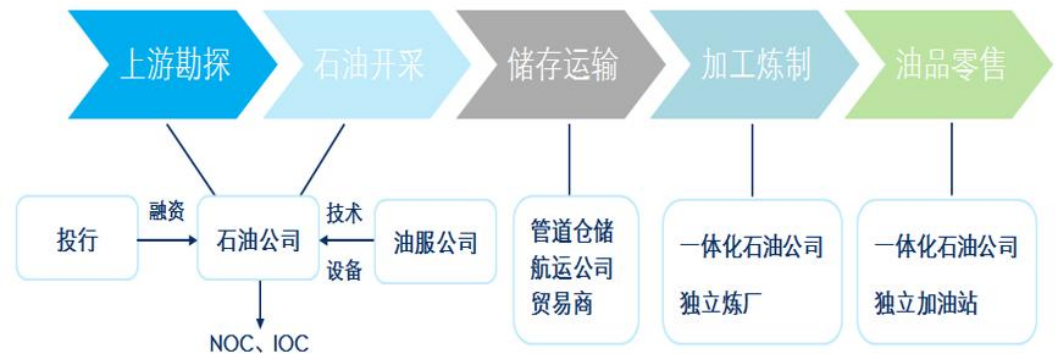
## 因子挖掘

### ■ 研究背景

原油作为基础能源，是重要的工业原料，经加工能得到柴油、汽油、煤油、沥青等多种石油化工产品，产业链涉及产油、炼油、运输、库存、消费等多个环节。因其涉及面广，原油价格的波动对于整个能化产业链而言称得上是牵一发而动全身的存在。

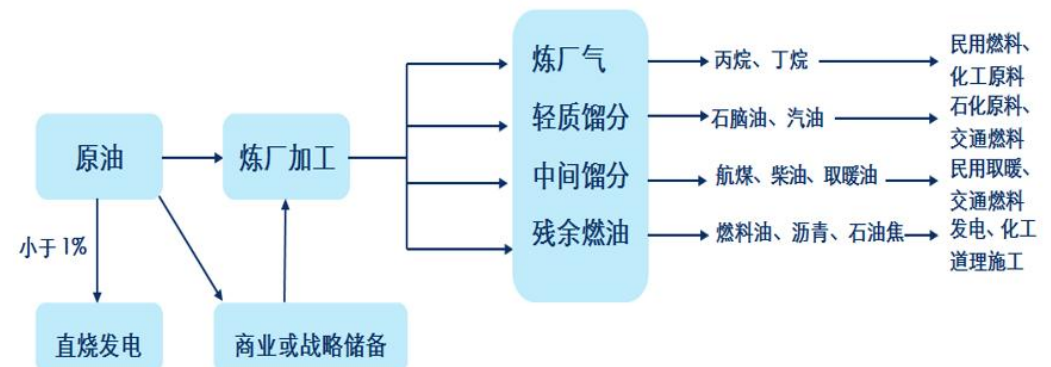
本文将从原油的产业链逻辑出发，深度挖掘上下游各个环节涉及到的基本面数据，通过精细化的因子衍生及因子降维，并构建出系统科学的因子有效性的打分机制，实现从海量数据中筛选出真正对预测原油价格有效的重点指标；然后再重点因子的基础上，利用机器学习模型对未来的预测值，设计出多因子、多频率的量价交易策略。

图 1:原油产业链 | 单位：无



数据来源：华泰期货研究院

图 2:原油的储运、加工与使用 | 单位：无



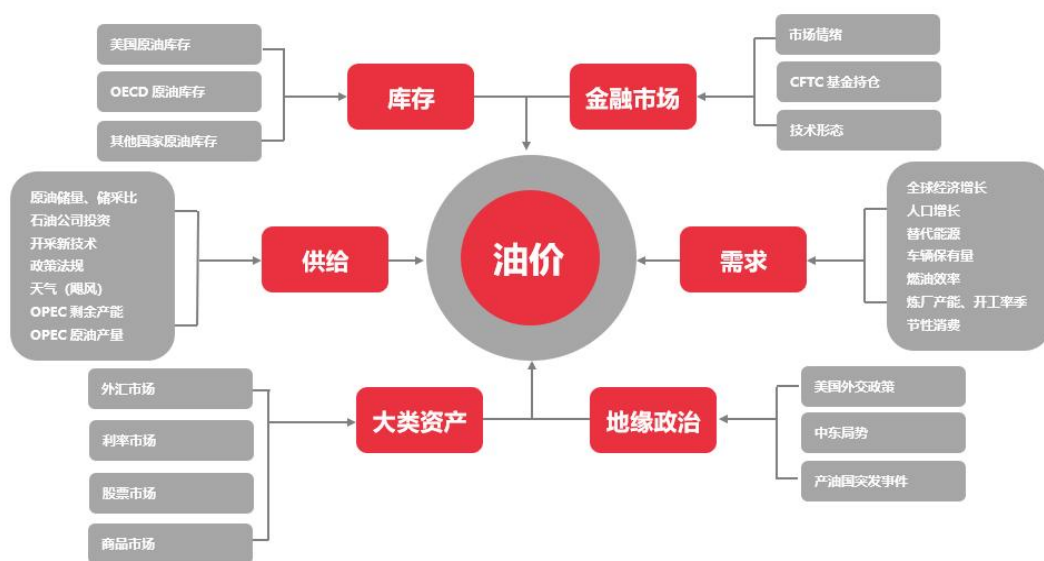
数据来源：华泰期货研究院

## ■ 指标收集

基于对原油产业基本面逻辑和供需结构的研究，并结合有效数据源相关情况，我们着重收集了库存、供给、大类资产、金融市场、需求、地缘政治、进出口、价格等八大方面的与原油价格相关的指标，来源涵盖了 Bloomberg, Wind, 同花顺等，总数据量达到了 2335 个。

除了庞大的基本面数据库外，我们还根据华泰商品因子体系构建出原油相关的量价因子库，共计 11 个量价因子加入到因子挖掘的研究中来。

图 3：影响原油价格因子分析框架 | 单位：无



数据来源：华泰期货研究院

## ■ 初步筛选

基本面数据对比起规整的量价数据而言，比较难处理的问题就在于，基本面数据经常出现缺失值、数据偶尔停更、数据更新频率低等，所以我们在收集到全数据后先对数据进行初步筛选。

本文研究选择标的资产为上海期货交易所的原油期货，其于 2018 年 3 月挂牌上市，于是选取研究时段为 2018 年 3 月至 2023 年 8 月，以下为初步筛选的条件：

- (1) 指标满足研究时段的有效性：具体体现在两个方面，一是指标需要仍在定期更新而不是停更的状态；指标开始发布的时间点需要早于 2019 年 1 月 1 号，以保证数据在时间上的连续性。

- (2) 指标更新频率为**日频/周频**：出于对后续构建模型的考虑，我们希望能根据选用的基本面因子构建出时效性更高、对波动更敏感的预测模型，能够更及时地捕捉到市场上的风吹草动，剔除更新频率太低的指标。
- (3) **剔除连续停更超过 30 天**的指标：有些指标即使满足了发布时间足够早且目前仍在更新的条件，偶尔会出现中间断更的情况，当断更时长超过一个月时，我们予以剔除。

经过初步筛选后，数据量从 2335 个下降至 **368 个**。根据每个指标背后的经济学意义以及其所在的产业链节点，我们人工整理出其所属的大类和二级分类，下表展示了其中一部分：

表 1: 基本面数据举例（部分） | 单位：无

指标大类	二级分类	指标名称	数据来源
库存	美国原油库存	美国原油商业库存、 美国原油战略库存（SPR）	Bloomberg
		阿拉斯加中转库存、 俄克拉荷马州库欣原油库存等	Bloomberg
		PADD I（东海岸）II（中西部）III（墨西哥湾沿岸） IV（洛基山） V（西海岸）原油库存等	Bloomberg
		其他国家原油库 存	Bloomberg
		西北欧原油库存	Bloomberg
		阿姆斯特丹原油库存	Bloomberg
	成品油库存、汽 油库存、航空煤 油库存等	新加坡成品油库存、汽油总库存等	Bloomberg
供给	钻机数	油气活跃钻机数	Bloomberg
		石油活跃钻机数	Bloomberg
		天然气活跃钻机数	Bloomberg
	原油产量	美国本土原油产量	Bloomberg
		阿拉斯加原油产量	Bloomberg
金融市场	市场情绪	恐慌指数	Bloomberg
	CFTC 基金持仓	Nymex WTI 持仓、ICE WTI 持仓、Nymex RBOB 汽油持仓、 Nymex ULSD 柴油持仓、ICE Brent 持仓等	Bloomberg

需求	炼厂总加工量	美国炼厂总加工量	Bloomberg
		PADD I（东海岸）II（中西部）III（墨西哥湾沿岸）	Bloomberg
		IV（洛基山） V（西海岸）炼厂加工量等	
		美国炼厂总产能	Bloomberg
地缘政治	炼厂产能	PADD I（东海岸）II（中西部）III（墨西哥湾沿岸）	Bloomberg
		IV（洛基山） V（西海岸）炼厂产能等	
		指数	
		地缘政治威胁指数	Wind
价格	油种价格	HLS 原油价格	Bloomberg
		LLS 原油价格	Bloomberg
		Bonito 原油价格	Bloomberg
		WTI 原油结算价	Wind
		OPEC：一揽子原油价格	Wind
		Brent 1-3 行价差、Brent 1-6 行价差、	Bloomberg
		Brent 1-12 行价差等	
		Nymex WTI vs ICE Brent、ULSD vs Gasoil	Bloomberg
		Nymex ULSD vs WTI、Nymex RBOB vs Brent、	Bloomberg
		Nymex WTI RBOB ULSD 211 裂解价差等	
进出口	成品油进口	馏分油进口量、	
		航空煤油进口量、	Bloomberg
		汽油与调和组分进口量等	
		美国馏分油出口量、	
大类资产	成品油出口	美国汽油出口量、	Bloomberg
		美国航空煤油出口量等	
	股票市场	标普指数	Bloomberg
		上证综合指数	Bloomberg
	债券市场	美债十年期收益率	Bloomberg
		外汇市场	
		美元指数	Bloomberg

数据来源：Bloomberg, Wind, 华泰期货研究院

可以发现往往同一个二级分类下的指标都高度相似，它们有些是分地区的统计值，有些是分产成品类型的统计值，而这些同质性过高的指标，没有必要重复引入模型，所以接下来的目标是从每个二级分类的小组内挑选出最优的一个指标来代表该小组，这样一方面可以实现数据降维，又能最大程度上保留指标本身的经济意义和可解释性。

## ■ 数据处理

**标的选择：**上海能源交易中心（INE）的原油期货（SC）

**自变量：**基本面因子+量价因子

**因变量：**原油主力期货收盘价

**数据预处理：**

- (1) **因子衍生：**每个二级分类相同的小组，衍生出组内的平均值，以代表组内整体情况
- (2) **平稳性检测：**平稳的指标保留，不平稳的指标做一阶差分直至数据平稳
- (3) **去极值：**高于 99 分位数以及低于 1 分位数的两端数据由上下限数据填充
- (4) **缺失值：**前值填充
- (5) **归一化：**以 12 个月为窗口向后滚动做 z-score 归一化处理

注：(i) 此处对自变量做归一化处理是为了消除自变量之间的量纲差异，提高模型的收敛性；(ii) 滚动往后是为了避免引入未来数据；(iii) 训练集最开始的部分因为滚动窗口较短，归一化处理容易出现极值，于是为了保证归一化处理后的数据的可靠性，我们舍弃了前 2 个月的数据。

- (6) **自变量后移一位：**用今天获得的最新数据预测明天的 SC 期货价格

**样本划分：**样本内数据：2018 年 4 月至 2021 年 12 月，共 33 个月

样本外数据：2022 年 1 月至 2023 年 8 月，共 20 个月

## ■ 组内选优

**预测目标：**原油主力期货下一期收盘价 - 当期收盘价

**选用模型：**3 种不同的回归模型，分别为 Correlation, Linear Regression, Support Vector Machine

**评分模式：**采用单因子评分模式，也就是每个因子轮流作为单个自变量输入模型去拟合预测因变量。

**统计指标：**由 3 种回归模型计算得来的共 5 种指标，分别为：



Correlation: 代表因子与因变量之间的相关系数，越大越好；

Information Coefficient: 代表因子变化率与因变量变化率之间的相关系数，越大越好；

Positive Ratio: 方向准确性，代表因子与因变量同向变大/变小的比例，越大说明变化方向越趋同；

Linear Beta: OLS 线性模型拟合出来的回归系数，代表因子每变动一个单位，因变量变动几个单位，越大说明影响越大；

SVR R square: 支持向量机模型拟合系数，代表单个因子通过 SVM 模型对因变量变动的解释程度，越大越好；

**综合评分：**每个因子在各个统计指标下都有一个得分，赋予权重后加总则可得到该因子的综合得分，最高得分即为选出的该组最优因子。

在对每个小组都进行同样的选优操作后，445 个因子（因子衍生后）成功降维至 **54** 个（外加量价因子）。

## 模型搭建

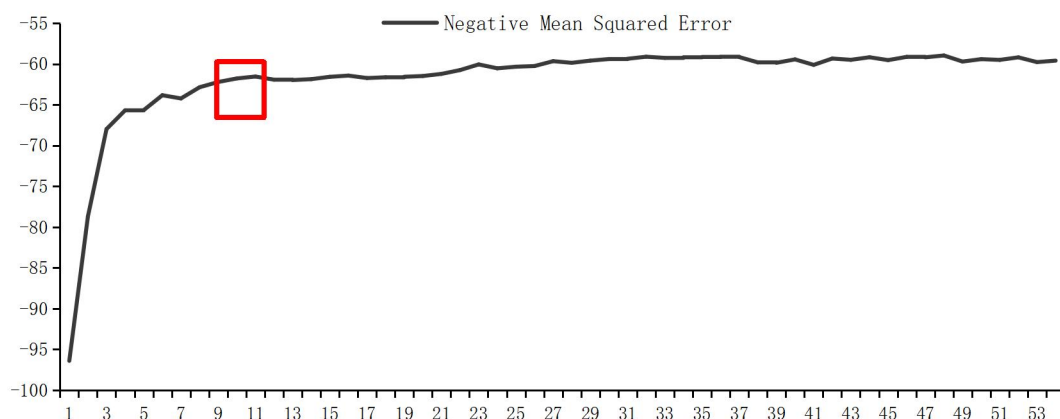
### ■ 确定因子数量

在正式进入搭建模型之前，首先需要确定一个答案，在从两千多个指标中优中选优出来的 54 个指标，我们最终应该引入多少个到模型当中。如果引入的因子数量太少，没办法对因变量的波动有很好的解释力度；如果引入的因子数量过多，模型又会有过拟合的风险。

这里我们使用了交叉验证的递归式特征消除（RFECV），底层生成器采用的是随机森林模型。算法原理是每次获得各个特征的重要程度之后，计算其决策系数之和，然后从当前的特征集合中移除最不重要的特征，并不断重复递归这个步骤，直到最终达到最优的特征数量。

下图为随机抽样重复跑 5 次 RFECV 得到的结果图，横轴为涵盖的因子数量，纵轴为负的均方差，纵坐标越高代表模型预测能力越好。从图中可以看出当因子数量为 11 时，可以兼顾因子数量不太多（降低过拟合风险）且模型预测力高，也就是说在所有的基本面+量价因子集合中，我们最终需要选出最有效的前 11 名因子。

图 4: RFECV 结果显示最佳因子数量为 11 | 单位: 无



数据来源: Bloomberg, Wind, 华泰期货研究院

## ■ 模型互融

在确定完最优的因子数量后, 我们沿用前面介绍的**因子评分模式**来给 54 个因子排序, 不同的是, 前面组内选优的时候只采用了单因子评分模式, 现在还需考虑多因子一起作为自变量输入模型时, 计算得出的特征重要性大小, 这样才更贴近我们的多因子模型效果。

**选用模型:** 在前面的 3 种回归模型基础上, 再引入 Random Forest Regression 和 Xgboost Regression 2 种机器学习模型

**统计指标:** 由 4 种回归模型计算得来的共 7 种指标, 分别为:

Correlation; Information Coefficient; Positive Ratio; Linear Beta

RF Feature Importance: 由随机森林模型运用全因子数据计算得到的特征重要性, 越大越好;

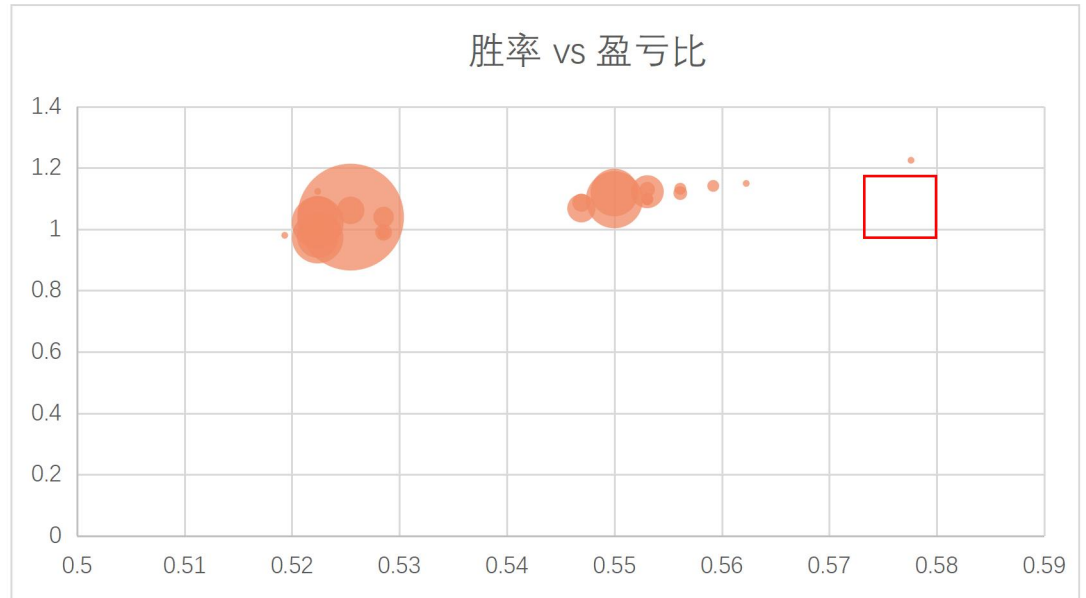
%IncMSE 和 IncNodePurity: 由 Xgboost 模型运用全因子数据计算得到的特征重要性, 越大越好。

**综合评分:** 每个因子在各个统计指标下都有一个得分, 赋予**权重**后加总则可得到该因子的综合得分, 从高到低排列的前 11 名则为模型选出的最优因子。

**最优权重:** 为确定最优的统计指标权重, 使用**网格遍历**的方法, 在共 648 种权重组合中, 使用**样本内数据**选出的不同版本的最优因子组合, 在**样本外数据**中测试模型预测的准确性, 以**胜率**和**盈亏比**为筛选标准, 同时达到胜率和盈亏比最高的则为最优权重。

下图中横轴为胜率，纵轴为盈亏比，散点大小为同样分数的权重组合数量，明显看出右上角有一个权重组合同时取得了最高胜率和最高盈亏比，即为最优权重：

图 5: 网格遍历确定最优权重组合 | 单位：无



数据来源：Bloomberg, Wind, 华泰期货研究院

### ■ 样本内最优因子

通过网格遍历确定完各统计指标的权重后，使用样本内数据计算得来的最优回归因子也就随之得知：

表 2: 样本内最优回归因子（前 15 名） | 单位：无

因子	最终得分	排名
期货结算价(连续):WTI 原油	15.704	1
OPEC:一揽子原油价格	15.387	2
DME Oman 首行原油期货价格	13.531	3
WTI Midland 原油价格	12.900	4
波罗的海运输指数(BDI)	12.819	5
汽油 PADD3 区库存	12.782	6
Bakken UHC 原油价差	12.699	7
美债十年期收益率	12.428	8
Brent1-12 行价差	12.397	9

Nymex ULSD 期货基金空头持仓	12.334	10
航空煤油进口量	11.825	11
西海岸（PADD V）原油库存	11.670	12
恐慌指数 VIX	11.540	13
价值因子	11.392	14
加拿大天然气活跃钻机数	11.095	15

数据来源：Bloomberg, Wind, 华泰期货研究院

### ■ 预测准度

模型选择：Random Forest 和 Xgboost （日频预测）

因子组合设计：

(1) 固定基本面因子组合：根据样本内数据筛选出的 11 个最优因子

【期货结算价(连续):WTI 原油, OPEC:一揽子原油价格, DME Oman 首行原油期货价格, WTI Midland 原油价格, 波罗的海运输指数(BDI), 汽油 PADD3 区库存, Bakken UHC 原油价差, 美债十年期收益率, Brent1-12 行价差, Nymex ULSD 期货基金空头持仓, 航空煤油进口量, 西海岸（PADD V）原油库存, 恐慌指数 VIX, 价值因子, 加拿大天然气活跃钻机数】

(2) 滚动基本面因子组合：月度调整，以每个月起始日往前滚动 12 个月为训练集，重新计算每个因子的权重得分，并选出新的一批最优因子加入模型中

(3) 固定量价因子组合：模型中引入 11 个量价因子，不包含任何基本面因子

【中国 CPIbeta 因子，美元指数 beta 因子，Curve 因子，流动性因子，均价突破因子，动量因子，持仓因子，偏度因子，期限结构因子，价值因子，波动率因子】

预测时间段：2021-01-01 至 2023-08-25

表 3: 模型预测准确性对比 | 单位：无

因子组合	方向预测 准确率	预测正确期间 因变量波动	预测错误期间 因变量波动	理论盈利 空间	F1	多信号 比例	空信号 比例
固定基本面因子	59.57%	2836	1863	973	0.6320	54.64%	45.36%
滚动基本面因子	64.30%	3084	1615	1469	0.6654	51.48%	48.52%
固定量价因子	58.78%	2899	1801	1098	0.6151	51.87%	48.13%

数据来源：Wind 同花顺 SMM 华泰期货研究院

前两组包含基本面因子的组合，在方向预测准确率上都领先于只包含量价因子的第三组，说明对于原油期货的价格预测，基本面因子发挥着不可替代的作用；其中滚动基本面因子组合效果最优，不管是方向准确性还是盈利的维度都遥遥领先，也验证了前面论述的因子挖掘方法是有效的。

## 量化策略

### ■ 基础设定

本金：100 万

单边手续费：3 元

滑点：0.01%

合约乘数：1000 桶/手

保证金比例：10%

多空信号生成规则：

- (1) 若 Random Forest 和 Xgboost 模型同时预测下一期价格变化值  $> 0$ ，则生成多头信号；
- (2) 若 Random Forest 和 Xgboost 模型同时预测下一期价格变化值  $< 0$ ，则生成空头信号；
- (3) 若预测值一正一负，则不产生信号

### ■ 日内策略

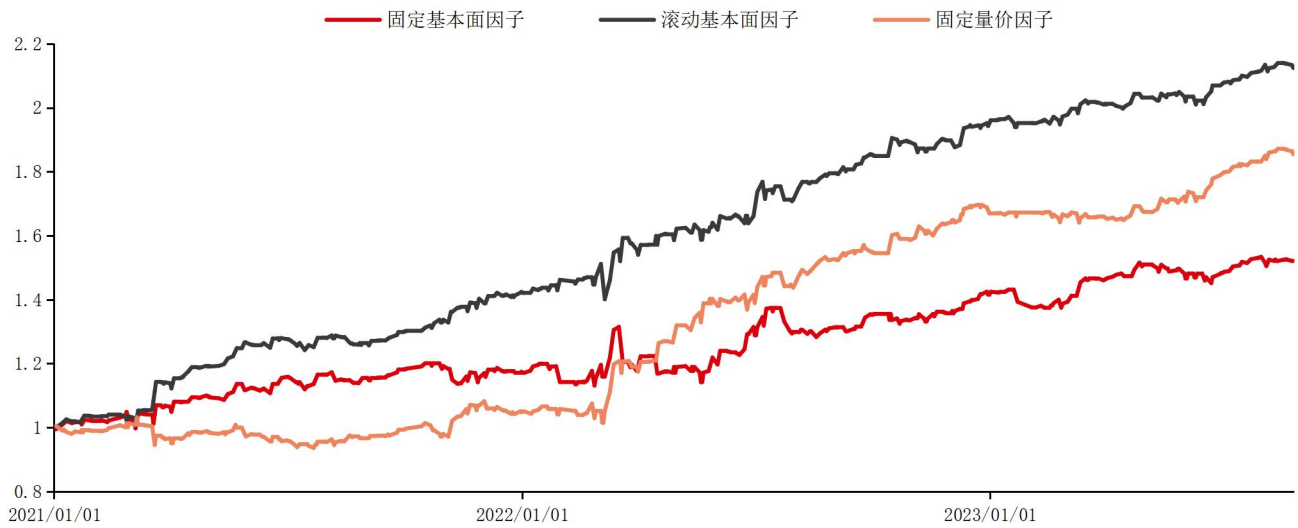
交易逻辑设定：

若前一日信号为多头，当天以开盘价买入，收盘价卖出；

若前一日信号为空头，当天以开盘价卖出，收盘价买入；

若前一日无信号，则空仓不进行任何操作。

图 6: 日内策略净值对比 | 单位: 无



数据来源: Bloomberg, Wind, 同花顺, 华泰期货研究院

表 4: 日内策略表现 | 单位: 无

策略名称	年化收益率 (%)	年化波动率 (%)	最大回撤 (%)	最大回撤天数	夏普比率	卡玛比率
固定基本面因子	22.967	13.91	13.12	32	1.652	1.750
滚动基本面因子	44.983	11.46	7.29	3	3.925	6.167
固定量价因子	35.277	13.74	9.28	73	2.567	3.803

数据来源: Bloomberg, Wind, 同花顺, 华泰期货研究院

通过不同因子组合的日内策略表现对比, 可以发现固定基本面因子组合策略在 22 年之前都略胜于固定量价因子组合策略, 但在 22 年之后被反超, 原因是固定因子组合是基于 2018 至 2021 年的样本内数据计算得到, 对于样本外的 22 年数据预测效果有所下降; 这也从不断更新训练集的滚动基本面因子策略大幅领先的策略表现得到验证。

## ■ 日间策略

交易逻辑设定：

若前一日信号为多头，当天以开盘价买入，一直持有到下一个不同的信号：

若下一个信号为空头，则以第二天的开盘价先平掉之前的多仓再开新的空仓；

若下一个信号为无信号，则以第二天的开盘价平仓。

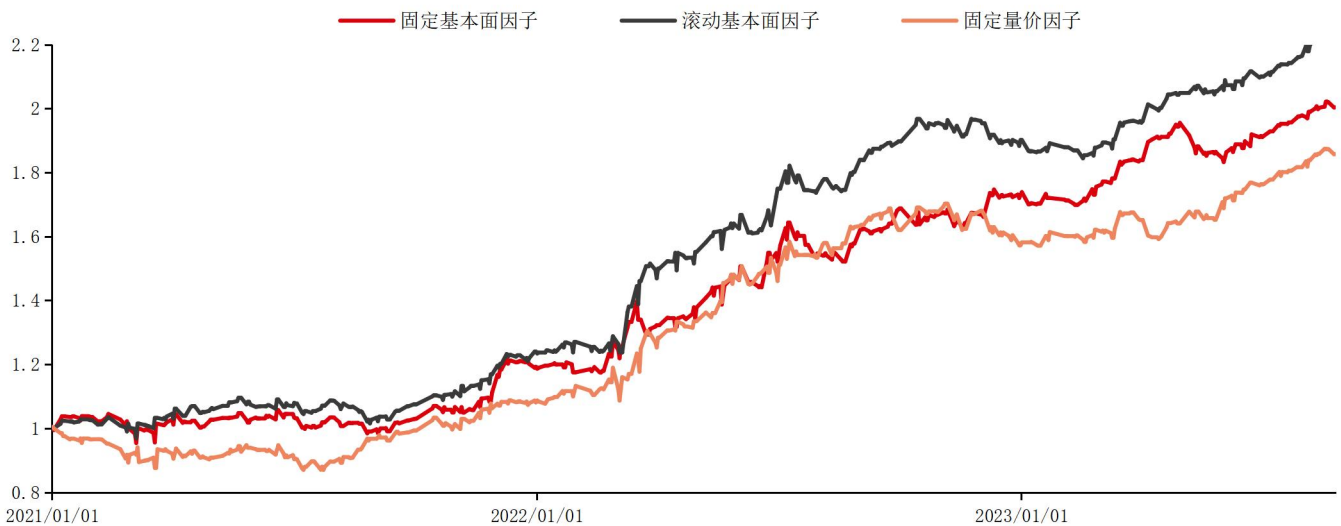
若前一日信号为空头，当天以开盘价卖出，一直持有到下一个不同的信号：

若下一个信号为多头，则以第二天的开盘价先平掉之前的空仓再开新的多仓；

若下一个信号为无信号，则以第二天的开盘价平仓。

若持有期间需要换月的情形，则强制先平掉目前仓位，再根据最新信号开仓。

图 7：日间策略净值对比 | 单位：无



数据来源：Bloomberg, Wind, 同花顺, 华泰期货研究院

表 5：日间策略表现 | 单位：无

策略名称	年化收益率 (%)	年化波动率 (%)	最大回撤 (%)	最大回撤天数	夏普比率	卡玛比率
固定基本面因子	40.86	19.37	8.67	7	2.109	4.711
滚动基本面因子	48.53	18.84	7.32	56	2.576	6.628
固定量价因子	35.37	20.39	13.39	106	1.734	2.641

数据来源：Bloomberg, Wind, 同花顺, 华泰期货研究院

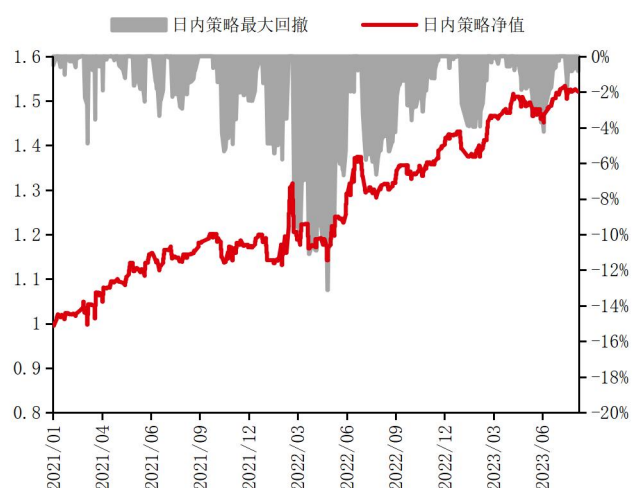
与日内策略不同的是，日间策略不会当天买卖，换手率相对较低。对比不同因子组



合的日间策略表现可知，基本面因子组合不管是固定还是滚动版本，都优于量价因子组合策略，侧面印证了基本面数据在对中长期未来预测有着不错的效果。

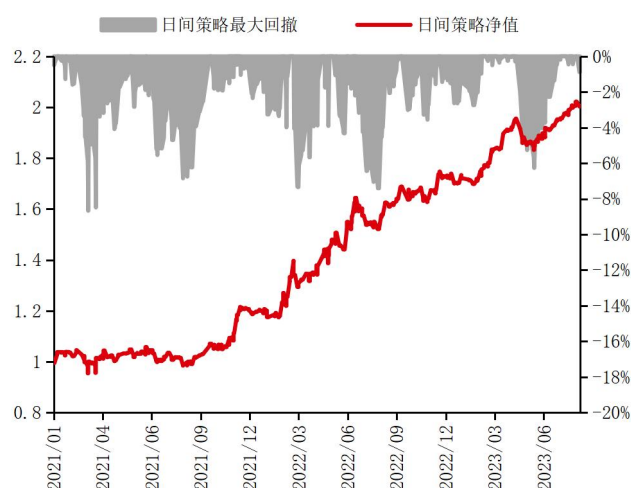
以下为六种策略的具体净值以及最大回撤表现图：

图 8：固定基本面日内策略 | 单位：无



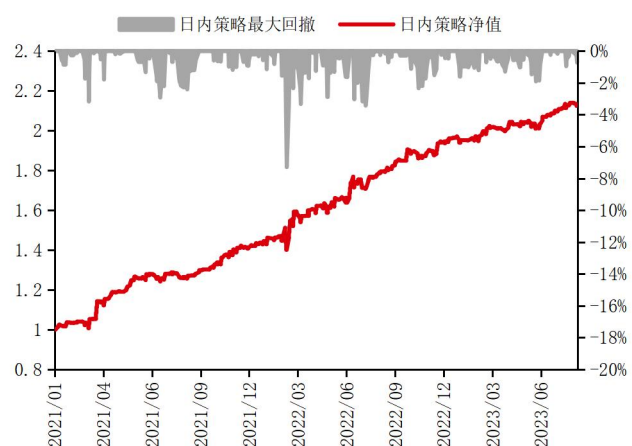
数据来源：Bloomberg, Wind, 同花顺, 华泰期货研究院

图 9：固定基本面日间策略 | 单位：无



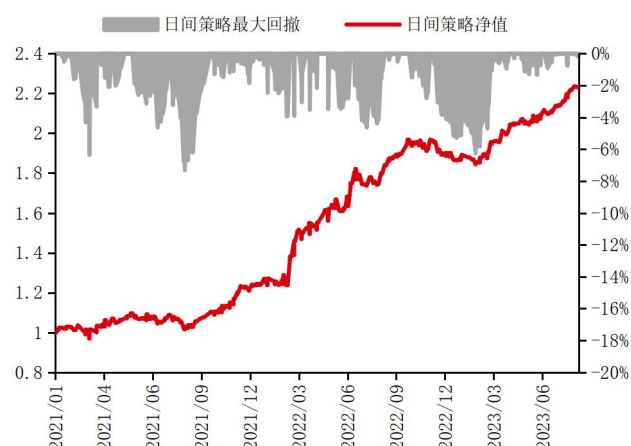
数据来源：Bloomberg, Wind, 同花顺, 华泰期货研究院

图 10：滚动基本面日内策略 | 单位：无



数据来源：Bloomberg, Wind, 同花顺, 华泰期货研究院

图 11：滚动基本面日间策略 | 单位：无

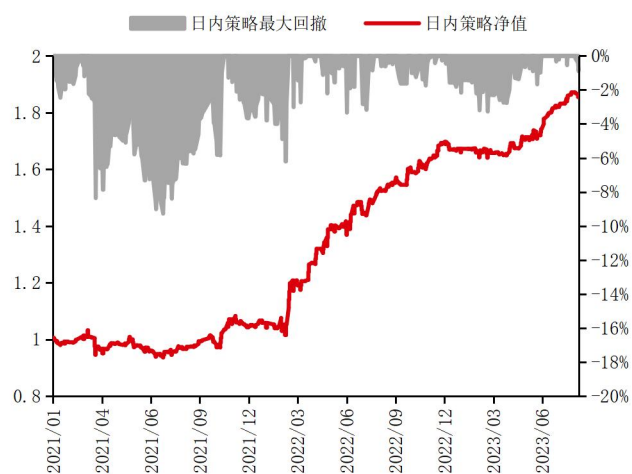


数据来源：Bloomberg, Wind, 同花顺, 华泰期货研究院

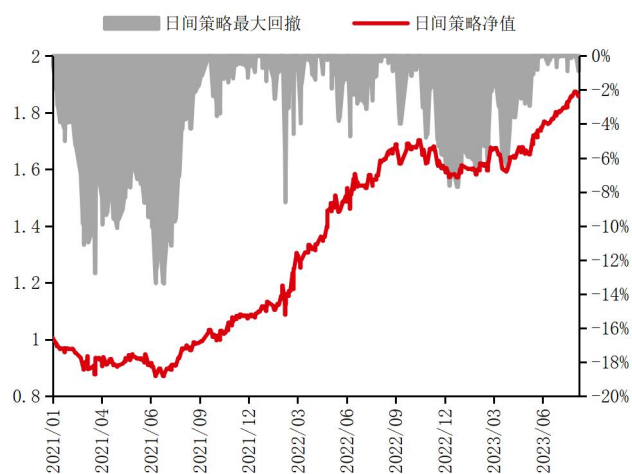
图 12：固定量价日内策略 | 单位：无

图 13：固定量价日间策略 | 单位：无





数据来源: Bloomberg, Wind, 同花顺, 华泰期货研究院



数据来源: Bloomberg, Wind, 同花顺, 华泰期货研究院

## 免责声明

本报告基于本公司认为可靠的、已公开的信息编制，但本公司对该等信息的准确性及完整性不作任何保证。本报告所载的意见、结论及预测仅反映报告发布当日的观点和判断。在不同时期，本公司可能会发出与本报告所载意见、评估及预测不一致的研究报告。本公司不保证本报告所含信息保持在最新状态。本公司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，但本报告所载的观点、结论和建议仅供参考，投资者并不能依靠本报告以取代行使独立判断。对投资者依据或者使用本报告所造成的一切后果，本公司及作者均不承担任何法律责任。

本报告版权仅为本公司所有。未经本公司书面许可，任何机构或个人不得以翻版、复制、发表、引用或再次分发他人等任何形式侵犯本公司版权。如征得本公司同意进行引用、刊发的，需在允许的范围内使用，并注明出处为“华泰期货研究院”，且不得对本报告进行任何有悖原意的引用、删节和修改。本公司保留追究相关责任的权力。所有本报告中使用的商标、服务标记及标记均为本公司的商标、服务标记及标记。

华泰期货有限公司版权所有并保留一切权利。

## 公司总部

广州市天河区临江大道1号之一2101-2106单元 | 邮编：510000

电话：400-6280-888

网址：[www.htfc.com](http://www.htfc.com)