2021 年第二届“大湾区杯”粤港澳

金融数学建模竞赛

题 目　　卖方券商的研报对公司股票走势的影响和投资策略

摘　　要：

本文基于卖方券商研报进行分析，将研报中非结构化数据结构化，并通过衰减函数转换为时间序列数据，构造分析师情绪因子。同时将研报中对公司的财务分析提取出来构建中期财务质量维度因子且加上 70 个短期价量因子，将各因子进行因子检验来构造因子库。然后构建一种将目标公司的股票根据当天的收益率变化进行标签化，运用 Random Forest 算法提取因子库中较强的因子，最后通过 LightGBM 来分类训练，判断买入点和卖出点。此外，由于舆情事件的突发性，对目标公司股票价格的短期影响较大，为此本文构建一种影响效应统计模型。经安慰剂检验，模型有效，能很好地反映舆情事件的对股票价格的冲击。最后将测算舆情事件冲击大小的模型加入交易策略模型中，利用模型修正目标公司的理论价格，为解决面对外部影响时，投资策略不会失效。

关键词：研报，分析师情绪因子，LightGBM，舆情事件

0.引言

正如赛题中所述，卖方券商提供的典型公司研报可能会包括：公司的相关数据、经营情况、重大事件、重要信息，以及关于公司的盈利预测与投资建议，公司的财务预测数据、估值结果与风险提示等。还可能会给出相关报告评级、历史推荐等级和目标价等。

由于券商提供的公司研报包含丰富和及时的信息，显然，券商研报是我们投 资决策重要的参考资料，以研报为基础得到的投资策略也是最有参考价值。基于研报信息的投资方法，一种可行的思路是从研报中提取特征指标，根据历史数据 完成深度学习，利用学习结果确定投资策略。这些特征指标可以用于构造市场流行的股票因子分析法的因子。

根据信息论，卖方券商的研报是每个专业分析师对目标公司的基本情况分析的成果，相对于研报中其它财务预测等分析，分析师对目标公司的评级和目标价格的观点已经完全包含研报中其他信息，因此本文根据信息包含，构建带分析师的“态度”和目标价格的情绪因子，然后对外部舆情事件对目标公司股价的影响进行建模和试测，最后综合选取的研报因子、中期财务因子和短期价量因子构建量化策略模型。

本文安排如下：第一部分，运用股票基本面数据对提供的 30 支股票进行选股；第二部分，如何运用 NLP 算法和表格识算法提取研报中有效部分，构造特征因子并检验特征因子。同时检验因子的有效性，分析筛选出的因子对股票价格的影响，并构建多因子投资策略；第三部分，建模探讨舆情事件突发对目标公司股票价格的影响；第四部分，建立带舆情事件的多因子投资模型和策略；第五部分，后续研究。

1.基本面选股

股票的基本面指对现行宏观经济环境，其所处的行业情况和公司的基本情况进行分析，包括公司经营增长性、公司财务报表等，选出优质股。而本文采用莫伦卡选股规则，通过过去 3 年平均净资产收益率高于 10%、市盈率低于 40 并大于 0、经营现金流为正和新期的净利润大于前 3 年的净利润这四个指标要求进行选股。选股流程图如下图 1 所示。

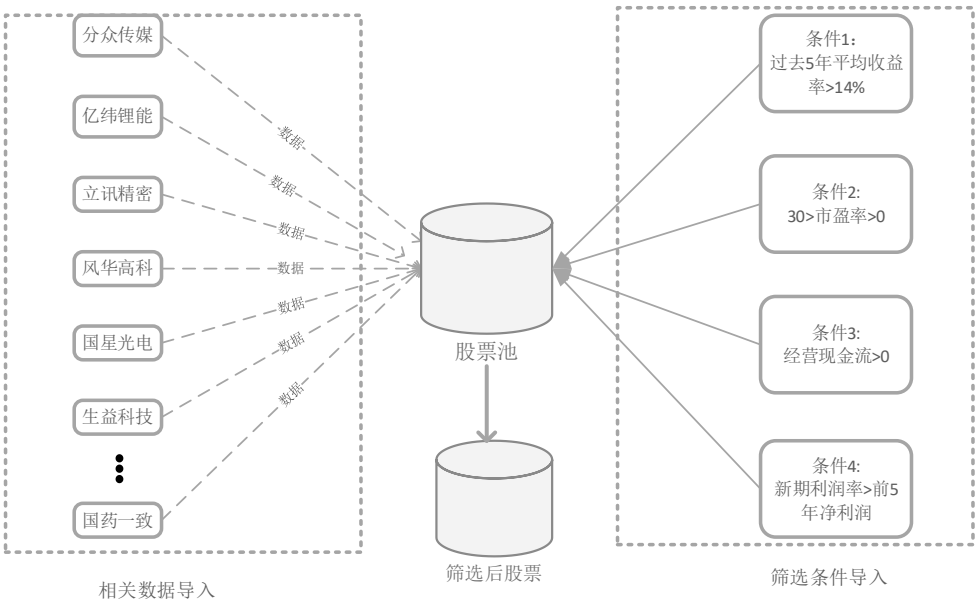


图 1 选股流程图

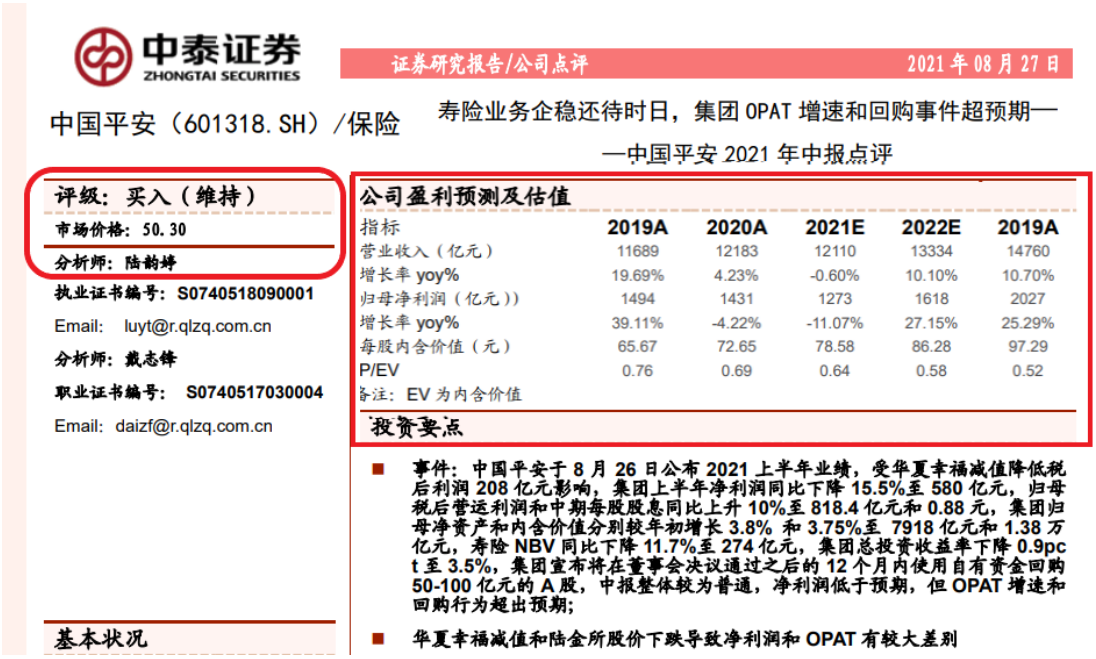
本文根据上述条件，在聚宽平台进行条件定义，从 30 只股票中筛选到保利地产、白云山、亿纬锂能、风华高科、中国平安、中顺洁柔、瀚蓝环境这 7 只股票，为选到 10 只，将条件宽松，再次筛选到华帝股份、招商积余和分众传媒这三只股票。

2.特征因子构造与检验

卖方券商研报作为投资者对目标公司股票的买卖参考，为此本文针对研报的分析师基于股票基本面分析给予的投资推荐等级（买入、增持、维持、减持和卖出）与未来盈利预测结果这非结构化内容进行结构化特征提取，用来表征分析师的情绪因子；再与目标股票的中期财务数据与其股票价格与交易量之间的关系构建短期价量因子等数据相结合，通过因子测试检验，确定符合条件的特征因子，构造因子库，进一步提出投资交易策略或股票风险预测。

2.1 特征因子提取方法

针对研报中盈利预测与估值部分涉及到的内容进行结构化提取，其中图 2 展示了中泰证券针对中国平安公司发布的研究报告中对中国平安的评级部分和盈利预测及估值情况。



中泰证券 ZHONGTAI SECURITIES

证券研究报告/公司点评 2021 年 08 月 27 日

中国平安（601318.SH）/保险

寿险业务企稳还待时日，集团 OPAT 增速和回购事件超预期——

——中国平安 2021 年中报点评

评级：买入（维持）

市场价格：50.30

分析师：陆韵婷

执业证书编号：S0740518090001

Email: luyt@r.qlzq.com.cn

分析师：戴志锋

职业证书编号：S0740517030004

Email: daizf@r.qlzq.com.cn

基本状况

公司盈利预测及估值

指标	2019A	2020A	2021E	2022E	2019A
营业收入（亿元）	11689	12183	12110	13334	14760
增长率 yoy%	19.69%	4.23%	-0.60%	10.10%	10.70%
归母净利润（亿元））	1494	1431	1273	1618	2027
增长率 yoy%	39.11%	-4.22%	-11.07%	27.15%	25.29%
每股内含价值（元）	65.67	72.65	78.58	86.28	97.29
P/EV	0.76	0.69	0.64	0.58	0.52

备注：EV 为内含价值

投资要点

- 事件：中国平安于 8 月 26 日公布 2021 上半年业绩，受华夏幸福减值降低税后利润 208 亿元影响，集团上半年净利润同比下降 15.5%至 580 亿元，归母税后营运利润和中期每股股息同比上升 10%至 818.4 亿元和 0.88 元，集团归母净资产和内含价值分别较年初增长 3.8% 和 3.75%至 7918 亿元和 1.38 万亿元，寿险 NBV 同比下降 11.7%至 274 亿元，集团总投资收益率下降 0.9pct 至 3.5%，集团宣布将在董事会决议通过之后的 12 个月内使用自有资金回购 50-100 亿元的 A 股，中报整体较为普通，净利润低于预期，但 OPAT 增速和回购行为超出预期；
- 华夏幸福减值和陆金所股价下跌导致净利润和 OPAT 有较大差别

图 2 券商研究报告-评级部分和盈利预测及估值部分

由于通过慧博、Wind 和东方财富网下载的卖方券商研报的格式均为 pdf 格式，且研报中要提取的部分多由文字和表格构成。因此对于研报中特征因子的提取总体上划分为两个步骤：提取和序列转化。如下图 3 所示提取流程。

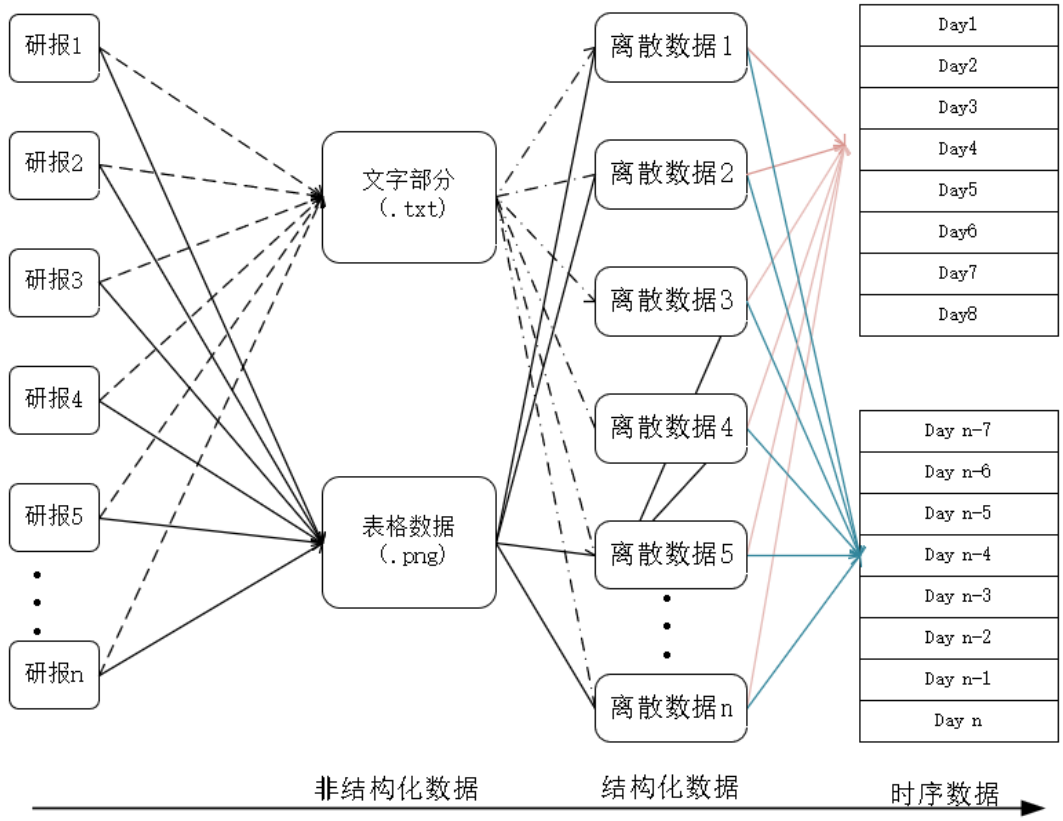


图 3 研报特征因子提取流程图

纵观典型研报，文字部分多为分析师根据相关数据提出的自身主观意见和分析，如东方证券研究所分析师赵旭翔的一份题为“业绩负增长，毛利率持续承压”的对万科 A000002.SZ 的研报分析，其中投资评级为“买入”。投资评级是卖方研报中分析师针对目标公司的基本面分析给出的投资推荐级别，但是不同券商机构对投资评级的等级不同，根据爬取的研报统计，本文将投资评级主要划分为买入、增持、维持、减持和卖出这 5 个级别。在研报中买入级别占多数，而卖出等级的较少，有统计发现，卖出的研报更具有参考价值，为此将这些顶级对应 1-5 各个量化数值。对于这些词语的提取，步骤如下：将带.pdf 格式的研报转化为带.txt 的文本文件，然后采用构造的词云去匹配.txt 文件中这些词语出现的次数，构成分析师的推荐特征因子。但由于研报的发布存在时间区间限制，不是连续性的数据，为此本文采用研报效应递减将离散数据进行时序数据转换，构造分析师情绪因子。

而研报中的表格多为公司的客观情况，如主要财务数据变动分析表、核心财务报表、财务报表预测和估值数据汇总等财务数据。因此，可以从研报中获取目标公司的一些典型传统量化因子。为提取研报中的表格，首先将.pdf 后缀的研报转化为.png 格式的图片，然后采用图片表格识别算法 PaddleOCR 进行提取，导入 EXCEL 文件，整理分析，构造中期质量因子。

2.1 特征因子说明

2.1.1 分析师情绪因子

由于卖方券商发布研究报告的时间不固定，提取的非结构化数据经过分值转换为结构化数据，并通过衰减效应将结构化数据转化为时间序列数据。因此，假设 L_t 为目标公司股票在时间 t 的分析师情绪等级，券商研究报告在时间 t 的数量篇数集合 $D_t=\{D^1,D^2,...,D^i\}$，每份研报 D^i 的分析师情绪等级为 L_t^i，因此在时间 t 的分析师情绪因子计算公式如下：

$$L_t=\frac{\sum_{i=1}^{n}D^iL_t^i}{Length(D_t)}$$

假设研报在 $t-j$ 时段内，没有卖方券商发布研究报告，为将离散数据转换为时间序列数据，故定义一种指数函数的衰减系数 φ，则在时间 $t-j$ 的分析师情绪因子为下列计算公式：

$$L_{t-j}=L_t\varphi^{-j},\varphi>1$$

2.1.2 中期质量维度

本文参考学者 Jason Hsu 和 Vitali Kalesnik 的论文思想来构建传统的衡量中期质量维度指标体系，如盈利能力、盈利稳定性、资本结构、盈利成长性、会计质量、股票发型摊薄与投资能力。同一质量维度下，指标具有同质性，各位单因子。

表 1 中期质量维度指标表

中期质量维度	质量分析指标
盈利能力	总资产收益率(ROA) 净资产收益率(ROE) 投资资本收益率(ROIC) 资本收益率(ROC)
盈利稳定性	每股盈利增长波动率(EPS Growth Variability) 每股盈利波动率(EPS Variability) 每股分红波动率(DPS Variability)
投资	总资产增长率(Growth in Total Assers)
会计质量	应计比率(Accrual Ratio)
盈利增长性	每股盈利增长(EPS Growth) 每股分红增长(DPS Growth) 资本周转率增长(Sales/Assets Growth)
资本结构	债务权益比(D/E) 债务现金流比(Debt/Cash Flow)
派息/摊薄	派息（股息+回购） 净派息（股息+回购+股票发行）

2.1.3 短周期价量维度

在股票交易中，投资者的交易行为在短期内对股票价格起着决定性的影响，尤其是近几年来，市场风格不断发生轮动变化，各传统策略的稳定性受到了一定冲击，实务界通过实战发现市场风格的变化对策略的影响远大于 Alpha 因子本身产生的波动，导致大量 Alpha 多因子策略失效，因此尝试找寻短期价量这种交易型套利空间是必要的，故本文根据国泰君安证券一篇《基于短周期价量特征的多因子选股体系》研究报告的思想，增加了短周期价量维度指标中多种短期价量 Alpha 因子，试着构建了 70 个基于价格和成交量数据的短周期 Alpha 因子，数据维度为日历史交易维度。因子构建如附件代码所示。如下面几个显著的价量特征例子：

（1）价量背离：短周期内成交量逐步提升，价格不断下降；或成交量逐步下降，价格不断提升。根据定义，计算公式为：

$$Alpha_t^i = -1 * corr(Stock\_p_{t-d:t}^i, Stock\_v_{t-d:t}^i)$$

其中 $Stock\_p$ 为股票的价格， $Stock\_v$ 为股票的成交量。

（2）开盘缺口：当日股票跳空高开或低开。计算公式为：

$$Alpha_t^i = Open\_price_t^i / Close\_price_{t-1}^i$$

（3）异常成交量：当日成交量较短周期均值异常放大、减少。计算公式为：

$$Alpha_t^i = -1 * Volume_t^i / mean(Volume_{t-d:t}^i)$$

2.2 因子的有效性检验

对单因子测试方法：本文采用一种检验单因子显著性及有效周期检验，步骤如下：

（1）因子中性化，去除风格和行业影响，取目标因子残差截面 ε_k^t，即

$$X_k^t = \beta_{industry} X_{industry} + \beta_{style} X_{style} + \varepsilon_k^t$$

（2）针对给定预测周期 d，通过回归方程计算单因子收益率 f_k：

$$R_{t+d} = f_{industry} X_{industry} + f_{style} X_{style} + f_k \varepsilon_k + \varepsilon_{t+d}$$

（3）计算因子收益率序列的年化收益 $E(f_k)$、信息比率 $IR(f_k)$ 及信息系数 IC_A：

$$E(f_k) = 252 * (f_k / d)$$

$$IR(f_k) = \sqrt{252} * (f_k / d) / \delta(f_k / d)$$

$$IC_A = correlation(f_{kt}, f_{k+1,t+1})$$

（4）对于不同的收益预测周期 d'，重复第 2、3 步。

因此，本文基于所选的 10 只股票构建股票池，在聚宽平台和同花顺量化金融实验室进行每个单因子的有效性和显著性分析。

2.2.1 分析师情绪因子

本节检验将在同花顺量化金融实验室平台上进行检验，将构建的分析师情绪因子数据上传构建分析师情绪因子，然后采用因子有效性检验。下图为 2021 年期间在东方财富网站爬取的研报分析，针对平安银行的研报投资推荐等级量化结果。

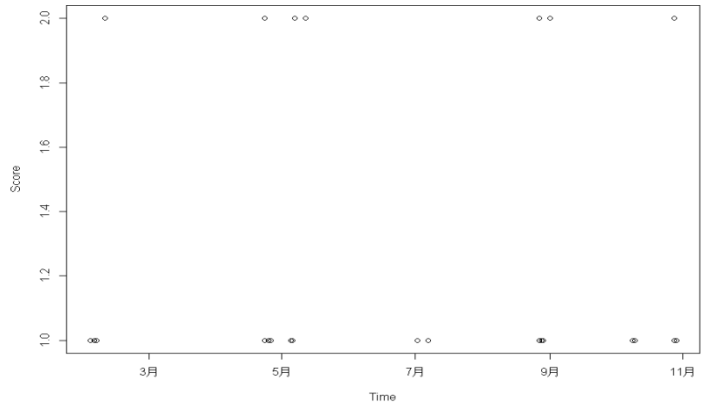


图 4 研报评级推荐等级量化结果

然后本文选取的衰减系数φ为 1.05，将推荐等级结构化数据转换为时间序列数据。然后通过与收益率之间进行单因子回归，测得中国平安研报的情绪因子对中国平安的股价影响效果较显著，系数为 0.012456，在 0.1 的置信区间下 P 值通过检验。

2.2.2 短周期价量维度

本节采用聚宽平台，分别对 70 个短期价量 Alpha 因子做单因子分析，由于因子众多和论文篇幅的限制，只展示对两个单因子分析的结果。两个单因子分别为 Alpha045 和 Alpha065。

设置起止时间为 2020 年 10 月 31 日到 2021 年 10 月 31 日、调仓周期为 3，6，10 天、分层数为股票池中股票个数，且只考虑做多机制，不考虑做空。

（1）Alpha045 单因子分析

Alpha045 因子的计算公式如下所示：

Alpha045=RANK(DELTA((((CLOSE*0.6)+(OPEN*0.4))),1))*
RANK(CORR(VWAP,MEAN(VOLUME,150),15))

通常来讲，对因子效果检验时，主要观察的两个指标是 IC 值和 IR 值。从图 1 分析结果可以得出，在持仓 3 天、6 天、10 天的分组中，持仓 6 天和 10 天的收益 IC 均值相同，IR 信息比率相同，但是 IC 均值和 IR 值的绝对值在整个过程中都小于 0.03，说明该因子效果有待提升。

	period_3	period_6	period_10
IC Mean	-0.005	-0.011	-0.011
IC Std.	0.329	0.329	0.324
IR	-0.016	-0.034	-0.034
t-stat(IC)	-0.241	-0.501	-0.496
p-value(IC)	0.809	0.617	0.621
IC Skew	0.074	0.059	-0.003
IC Kurtosis	-0.199	-0.596	-0.459

图 5 各时期 IC 结果（聚宽平台运行结果截图）

另外，收益率分析的原理是假设每一次调仓时都以因子值为权重购买相应的股票组合，然后计算出每一期可以获得的收益率。从收益率分析，也表明 Alpha045 单因子的效果并不理想。

	period_3	period_6	period_10
Ann. alpha	-0.054	-0.065	-0.052
beta	-0.028	0.042	0.058
Mean Period Wise Return Top Quantile (bps)	-11.076	-11.072	-8.060
Mean Period Wise Return Bottom Quantile (bps)	9.356	-7.057	-11.904
Mean Period Wise Spread (bps)	-19.495	-3.242	3.764

图 6 收益分析结果（聚宽平台运行结果截图）

再从各分位数平均收益来看，大部分分位收益为负数。

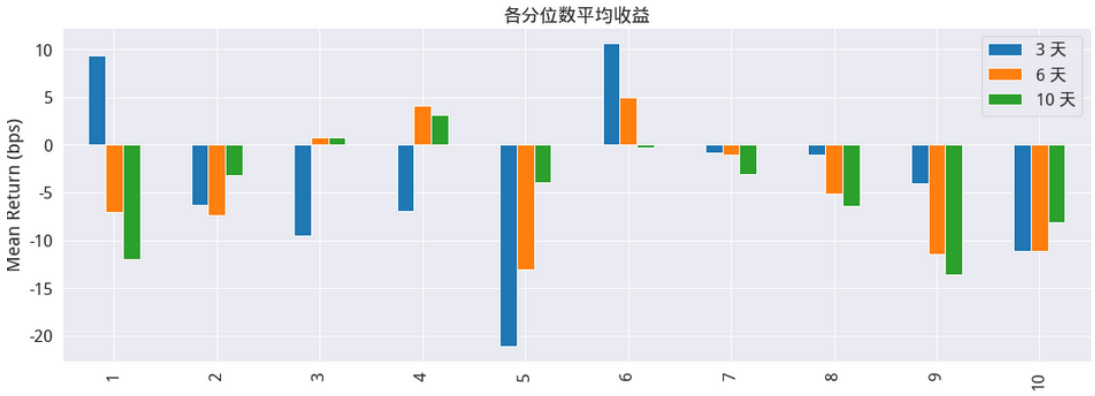


图 7 各分位数平均收益（聚宽平台截图）

从换手率分析，即计算相邻两期选股组合中股票的平均换手率，如果一个因子挑选出来的每一期股票的变动分为不大即换手率不高，那么交易成本就比较低，所有换手率尽量低的因子最好。

	period_10	period_3	period_6
Quantile 1 Mean Turnover	0.884	0.757	0.806
Quantile 2 Mean Turnover	0.865	0.883	0.863
Quantile 3 Mean Turnover	0.908	0.860	0.896
Quantile 4 Mean Turnover	0.937	0.874	0.877
Quantile 5 Mean Turnover	0.903	0.879	0.938
Quantile 6 Mean Turnover	0.903	0.897	0.929
Quantile 7 Mean Turnover	0.855	0.897	0.882
Quantile 8 Mean Turnover	0.913	0.860	0.872
Quantile 9 Mean Turnover	0.870	0.850	0.900
Quantile 10 Mean Turnover	0.889	0.855	0.919

图 8 换手率分析结果（聚宽平台截图）

综上所述，舍弃 Alpha045 单因子，在量化策略中不考虑此单因子。

（2）Alpha003 单因子分析

Alpha003 因子的计算公式如下：

$$Alpha003 = SUM\left(\left(\begin{array}{l} CLOSE = DELAY\left(CLOSE,1\right)?0:CLOSE - \\ \left(CLOSE > DELAY\left(CLOSE,1\right)?MIN\left(LOW,DELAY\left(CLOSE,1\right)\right):\right. \\ \left.MAX\left(HIGH,D\ ELAY\left(CLOSE,1\right)\right)\right. \end{array}\right),6\right)$$

从 IC 分析来看，Alpha003 的 IC 值和 IR 值相对来说都比较高，且在实务界中，IC 的绝对值大于 0.03，即表明因子有效。所以从 IC 值和 IR 值方面，Alpha003 因子可以被保留。

	period_3	period_6	period_10
IC Mean	0.060	0.091	0.060
IC Std.	0.323	0.335	0.320
IR	0.185	0.272	0.187
t-stat(IC)	2.699	3.964	2.729
p-value(IC)	0.008	0.000	0.007
IC Skew	0.104	0.104	-0.076
IC Kurtosis	-0.134	-0.350	-0.363

图 9 IC 分析结果(聚宽平台结果截图)

从收益分析方面来看，下图为 Alpha003 因子的收益分析图，整个收益情况良好，尤其是调仓期限为 3 天时，在测试期间，累积收益率达 25%。

	period_3	period_6	period_10
Ann. alpha	0.315	0.276	0.167
beta	-0.036	0.013	0.112
Mean Period Wise Return Top Quantile (bps)	-4.597	-2.680	-2.252
Mean Period Wise Return Bottom Quantile (bps)	-17.836	-13.323	-2.445
Mean Period Wise Spread (bps)	13.720	11.438	0.803

图 10 收益分析（聚宽平台结果截图）

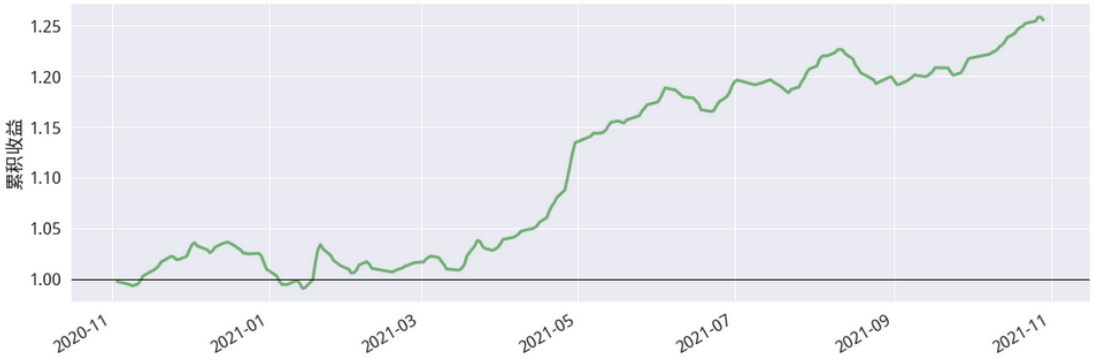


图 11 因子值加权加多组合累积收益（3 天平均）

从换手率方面来看，相对来说，在不同时期换手率都较高。

	period_10	period_3	period_6
Quantile 1 Mean Turnover	0.817	0.751	0.845
Quantile 2 Mean Turnover	0.861	0.823	0.883
Quantile 3 Mean Turnover	0.871	0.842	0.908
Quantile 4 Mean Turnover	0.901	0.852	0.893
Quantile 5 Mean Turnover	0.871	0.837	0.893
Quantile 6 Mean Turnover	0.847	0.876	0.854
Quantile 7 Mean Turnover	0.921	0.880	0.908
Quantile 8 Mean Turnover	0.876	0.885	0.879
Quantile 9 Mean Turnover	0.886	0.828	0.869
Quantile 10 Mean Turnover	0.866	0.694	0.854

图 12 换手率分析（聚宽平台结果截图）

综上分析，保留 Alpha003 单因子，将其纳入因子体系中。其余因子都采用相同的办法进行检验，将无效的短期价量因子进行舍弃，有效的因子保留，最终从 70 个价量因子中筛选出 34 个短期价量因子，分别是 Alpha003，Alpha009，Alpha010，Alpha011，Alpha012，Alpha014，Alpha018，Alpha021，Alpha023，Alpha024，Alpha025，Alpha026，Alpha029，Alpha031，Alpha033，Alpha034，Alpha035，Alpha040，Alpha041，Alpha043，Alpha047，

Alpha049，Alpha050，Alpha051，Alpha052，Alpha053，Alpha057，Alpha058，Alpha061，Alpha065，Alpha066，Alpha067，Alpha068 和 Alpha070。（因子计算公式如附录 1 所示）

2.2.3 中期财务质量维度

由于中期财务质量维度是西方学者根据国外股票数据测得有效，但是由于国内股票存在差异，因此本节将检验各因子的有效性，构建财务质量多因子策略，直接回测因子的有效性。下图为策略收益情况：

图 13 财务质量因子策略收益情况

经测得，中期财务质量因子中盈利稳定性、盈利能力和盈利增长性维度的指标相对于其他维度的指标更具有效果，因此，本文筛选掉其他维度因子，将保留的因子纳入因子池。

2.3 投资策略构建

本文主要以卖方券商的研究报告为重要特征来源，结合其他有关目标公司的财务有效指标和短期价量指标，在通过因子有效性检验之后，构建出下列用于构建量化投资策略和风险预测的特征因子库。特征因子库主要指标包括：分析师情绪因子、中期财务因子维度中盈利能力、盈利稳定性和盈利增长性指标以及 30 只短期价量因子。本节采用 LightGBM 模型去训练因子库中数据，进行买入点和卖出点的计算。

相对来说，因子库中数据较高维，且存在数据稀疏情况，而 LightGBM 算法更能处理相关这种问题，处理过程是利用 EFB 算法利用特征空间的稀疏性来有效减少特征数量，通过将互斥的特征捆绑在一起，来减少特征数目。互斥特征意味着它们几乎很少同时出现非零值，并且 LightGBM 也表明：找到最优互斥特征捆绑是 NP(Non-deterministic Polynomial)难问题，但是贪婪算法能够获得非常好的近似概率。所以本文选择 LightGBM 模型去判定买入和卖出点。具体步骤图如下所示：

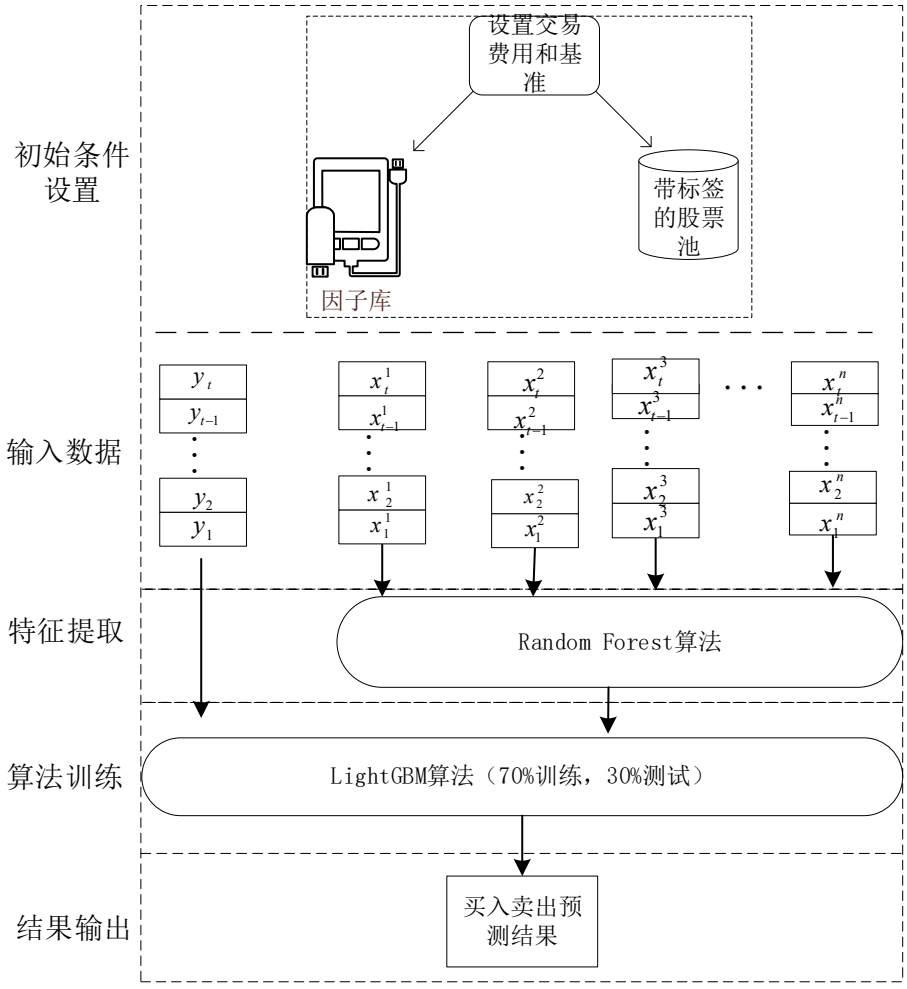


图 14 基于 LightGBM 算法的买入和卖出交易模型

其中，带标签的股票是指假设当天收盘价大于开盘价，则标注为 1；反之，标注为 0。Random Froest 算法用来提取因子库中强有效因子，然后通过 LightGBM 算法去分类训练，最后通过分类结果去买卖当天的股票。

3. 探讨舆情事件的影响

为讨论舆情事件对目标公司的影响，本文采用 Hsiao 教授的原理方法构建一种“反事实”的突发事件影响效应评估模型，考察突发事件的闪现对目标公司股价波动的影响。模型构建过程如下所示：

假设目标公司的股票价格与同行业内的股票价格存在关系，对目标公司的股票价格有参考作用，因此本文引入目标同行公司处于同行业且业务相似的股票价格，共构成 N 个相似

公司股票价格指标，每个指标由K个公共因子，研究区间为$t=1,...,T$时刻，在T_1时刻，标的指标i（目标公司股价）受与目标公司相关的舆情事件的影响，其他非标的指标不受该舆情事件的影响。

首先，构建股票价格的回归模型：

$$y_{it}^{0}=b_i^{'}F_t+\alpha_i+\varepsilon_{it},\ \ i=1,\ldots,N,t=1,\ldots,T$$

公式（1）中：

y_{it}^{0} 舆情事件前未影响标的股票i的情况下，i在t时刻的股票价格；

$b_i^{'}=(b_{i1},...,b_{ik})$ $b_i^{'}$为$1\times K$的矩阵，表示K个公因子的系数；

$F_t=(f_{i1},...,f_{ik})'$，表示F_t为$K\times 1$的矩阵，表示y_{ik}^{0}受K个公共因子影响；

α_i，表示个体的固定截距；

ε_{it}，表示i个体残差部分，其中$E(\varepsilon_{it})=0$.

在式（1）中，F_t表示时间序列的公共因子驱动面板数据中的所有截面数据，α_i表示每个个体的异质性影响因素，ε_{it}表示个体的残差项。假设每个个体的异质性部分不相关，各个截面单位间的相关性由公共因子F_t驱动。同时，各个截面单位受F_t驱动的影响可能不同，即$b_i\neq b_j$。

再假设N个指标的y_{it}^{0}表示为y_t^{0}，则

$$y_t^{0}=\alpha+BF_t+\varepsilon_t$$

公式（2）中：

$y_t^{0}=(y_{1t},...,y_{Nt})'$：$t$时刻$N$个指标股票价格的列矩阵；

$\alpha=(\alpha_1,...,\alpha_N)'$：$N$个指数的固定截距与残差的列矩阵；

N个指数中每个$B=(b_1,...,b_N)_K'$：指数公因子的系数的$N\times K$阶矩阵；

$\varepsilon_t=(\varepsilon_{1t},...,\varepsilon_{Nt})'$：$N$个指标残差的列矩阵；

$F_t=(f_1,...,f_K)'$：影响N个指标的K个公因子。

假定其中：$rank(B)=K$,每个指标至少会受一个公共因子的影响；

现在，在构建处置效应模型公式前，假设从T_1开始对第一个个体（目标公司）开始受舆情事件的影响。在T_1之前，我们将观测到的未加干预的y_{1t}表示为：

$y_{1t}=y_{1t}^0$, t=1,...,T_1.

T_1后，y_{1t}在舆情事件影响下的观测值记为：

$y_{1t}=y_{1t}^1$, t=$T_1+1,...,T$.

对其他指标没有受到这种舆情事件的影响，因此我们有：

$y_{it}=y_{it}^0$, i= 2,...,N, t=1,...,T.

因为不能够同时观测标的指标受舆情事件的影响与不受舆情事件的影响两种不同状态，因此我们定义一个虚拟变量为：

$d_{it}=\begin{cases}1\\0\end{cases}$，其中$d_{it}=1$时，表示在时刻$t$，$i$受到舆情事件的影响；反之未受影响。

则y_{it}可表示为公式：

$y_{it}=d_{it}y_{it}^1+(1-d_{it})y_{it}^0$

再假设 $E(\varepsilon_{is}\mid d_{1t})=0$,其中 i=2,...,N 并且$s\geq t$，其他指标不受舆情事件的影响。

在无舆情事件影响的情况下，用实际股票价格与预测股票价格之差来衡量y_{1t}的处置效应，如下：

$\Delta_{1t}=y_{1t}^1-y_{1t}^0, t=T_1+1,...,T$

主要问题是很难测量Δ_{1t}，我们不能观察到T_1时期后y_{1t}^0的值。有学者通过指定y_{1t}的条件股票价格模型来预测Δ_{1t}，但是仍然存在缺陷，为解决问题，本人应用 Hsiao 的方法，使用来自$\tilde{y}_t^0=(y_{2t}^0,...,y_{Nt}^0)'$来预测$y_{1t}^0$。

那等式（1）可以表示成：

$y_{1t}^0=\tilde{\alpha}+\tilde{\alpha}'\tilde{y}_t^0+\varepsilon_{1t}-\tilde{\alpha}'\tilde{\varepsilon}_t$

其中$\tilde{\alpha}=(\alpha_2,...,\alpha_N)',\tilde{\varepsilon}=(\varepsilon_{2t},...,\varepsilon_{Nt})'$，定义$\tilde{\varepsilon}_{1t}=\varepsilon_{1t}-\tilde{\alpha}'\tilde{\varepsilon}_t$，然后可得：

$y_{1t}^{0}=\tilde{\alpha}+\tilde{\alpha}'\tilde{y}_{t}^{0}+\tilde{\varepsilon}_{1t}$

我选择$(\bar{\alpha},\tilde{\alpha})$去最小化，得：

$$\frac{1}{T_1}\sum_{t=1}^{T_1}(y_1^0-\bar{\alpha}-\tilde{\alpha}\tilde{y}_t^0)'(y_{1t}^0-\bar{\alpha}-\tilde{\alpha}'\tilde{y}_t^0).$$

y_{1t}^{0}的估计量可以定义为：

$\hat{y}_{1t}^{0}=\hat{\bar{\alpha}}+\hat{\tilde{\alpha}}\tilde{y}_{t}^{0}$，

并且Δ_{1t}可以通过下面等式估计：

$$\hat{\Delta}_{1t}=y_{1t}-\hat{y}_{1t}^{0} \quad t=T_1+1,...,T$$

根据 Hsiao 所提出的模型，可得：

$$E(\hat{\Delta}_{1t}\mid\tilde{y}_t)=\Delta_{1t} \quad t=T_1+1,...,T$$

需要注意的是，在估计效应时也许存在一系列相关性。因此，本文使用 Box-Jenkins 的方法构造一个 ARMA 模型：

$$\tilde{\alpha}(L)\hat{\Delta}_{1t}=\tilde{\mu}+\tilde{\theta}(L)v_t$$

其中$\tilde{\mu}$测量事件的长期效应，而 t 统计能够用于测试$\tilde{\mu}$是否明显不同于 0.如果$\tilde{\Delta}_{1t}$是一个平稳过程，则长期效果可以通过简单的平均数除于 t，如下所示：

$$\underset{(T-T_1)\to\infty}{p\lim}\frac{1}{T-T_1}\sum_{t=T_1+1}^{T}\hat{\Delta}_{1t}=\Delta_1$$

3.1 数据说明

为验证舆情事件对目标公司股票价格的影响，本节选择的目标公司和舆情事件为华帝股份 2018 年 7 月 1 日的“法国夺冠，华帝退全款”，以华帝股份股票的日交易数据为实验组，控制池选择与华帝股份同行业的 46 家上市公司。由于同行业的上市公司数量较大，为进一步减少控制池中样本个体来提高整个模型的预测效果。我们采用 CAPM 模型中的 Beta(β)系数来选取与目标公司最接近的样本，采用下列等式计算指标：

$$D_{ij}=\frac{1}{T}\sum_{t=1}^{T}(\beta_{it}-\beta_{jt})^2$$

其中，β_{it} 为目标公司 i 在时间 t 的 Beta 系数，β_{jt} 为控制池中第 j 只股票在时间 t 的 Beta 系数。

将计算结果由小到大排序，找出控制池中与目标公司最相近的公司，所以选择 TCL 科技、奥佳华、飞科电器、格力电器、海尔智家、海信家电、海信视像、九阳股份、莱克电气、老板电器、美的集团、日出东方、三花智控、四川九洲、苏泊尔、新宝股份、兆驰股份和浙江美大这 18 家上市公司。为消除各公司股票的价格大小的影响，采用 20 日收益标准差的简单移动平均波动率进行探讨。此外，要探讨舆情事件对目标公司的影响，需要选择舆情事件发生前和发生后的数据，因此我们最终选取的时间窗口区间为 2016 年 1 月 4 日到 2018 年 10 月 15 日，其中 2016 年 1 月 4 日—2018 年 7 月 1 日为舆情事件前窗口期，而 2018 年 7 月 2 日-2018 年 10 月 15 日为舆情事件后窗口期。20 日平均波动率的数据来源于锐思数据库，总数据为 13540 条。

运行软件是 R 语言，所用的工具包有 pampe 和 leaps。下表为各股票 20 日平均波动率数据的描述性统计。

表 2 描述性统计结果

Variable	Obs	Mean	Std. Dev.	Min	Max
TCL科技	677	0.01444	0.01059	0	0.04364
奥佳华	677	0.02504	0.01008	0	0.06069
飞科电器	677	0.01979	0.01110	0	0.06794
格力电器	677	0.01730	0.01151	0	0.04612
海尔智家	677	0.01864	0.00742	0	0.04273
海信家电	677	0.02666	0.00890	0.01101	0.06057
海信视像	677	0.01996	0.00861	0	0.05439
华帝股份	677	0.02665	0.00857	0	0.04647
九阳股份	677	0.18507	0.00831	0	0.04745
莱克电气	677	0.02435	0.00781	0.00951	0.05009
老板电器	677	0.02073	0.00688	0.00606	0.04027
美的集团	677	0.02005	0.00774	0	0.04390
日出东方	677	0.02355	0.00881	0	0.04430
三花智控	677	0.02490	0.00973	0	0.05124
四川九洲	677	0.02560	0.01116	0.00619	0.05786
苏泊尔	677	0.02199	0.00679	0	0.04147
新宝股份	677	0.02151	0.00684	0.01115	0.04400
兆驰股份	677	0.21111	0.00967	0	0.05039
浙江美大	677	0.25215	0.01121	0.00886	0.06930

3.2 最优选择准则结果

虽然通过 Beta 系数减少了控制池中的控制变量，但是依然存在欠缺。本文还根据 Akaike

信息准则 AIC 指标和 Bayesian 信息准则(BIC)选择控制组中具体应包含的 k 个公司股票。

$$AIC = \ln(R^2) + \frac{2*(k+1)}{n}$$

$$BIC = \ln(R^2) + \frac{(k+1)*\ln(n)}{n}$$

其中：$R^2 = e'e/(n-k-1)$ 是模型的拟合优度，$2*(k+1)/n$ 和 $\frac{(k+1)*\ln(n)}{n}$ 为测度模型的复杂程度。

根据本文拟定的选择标准进行控制组对象和方程的选择后，AIC 和 BIC 准则都指向同一控制组，包含海尔智家、海信家电、海信视像、莱克电气、老板电器、三花智控、四川九洲、苏泊尔、新宝股份和浙江美大这 10 个公司股票，拟合情况如表 2 所示。

表 3 时间窗口中各控制组公司的权重

	Estimate	Std. Error	t value	Pr(>|t|)
Intercept	-0.00145920	0.00099465	-1.4670	0.142890
海尔智家	0.20045257	0.02643796	7.5820	1.311e-13 ***
海信家电	0.11807099	0.02698825	4.3749	1.434e-05 ***
海信视像	-0.17149648	0.04363899	-3.9299	9.495e-05 ***
莱克电气	0.12074845	0.03904286	3.0927	0.002076 **
老板电器	0.19410148	0.04683161	4.1447	3.896e-05 ***
三花智控	0.19085911	0.02667806	7.1542	2.480e-12 ***
四川九洲	-0.10308423	0.02317415	-4.4482	1.033e-05 ***
苏泊尔	0.18351661	0.04410830	4.1606	3.642e-05 ***
新宝股份	0.18858484	0.04047007	4.6599	3.906e-06 ***
浙江美大	0.27354282	0.02935574	9.3182	< 2.2e-16 ***

$R^2 = 0.804, Adjusted\ R^2 = 0.800$，其中***为 0，**为 0.01 条件

在上述最优模型下，下图 2 是 2016 年 1 月 4 日—2018 年 7 月 1 日期间对华帝股份的平均波动率拟合情况以及 2018 年 7 月 2 日-2018 年 10 月 15 日对华帝股份的平均波动率的预测。（红线左边是舆情事件发生之前，右边是发生之后）由图可见，控制池中公司股票的平均波动率能够很好地拟合华帝股份的平均波动率，不同波动趋势也能很好地拟合，因此可以构建模型预测未发生舆情事件时，华帝股份应该的平均波动率。

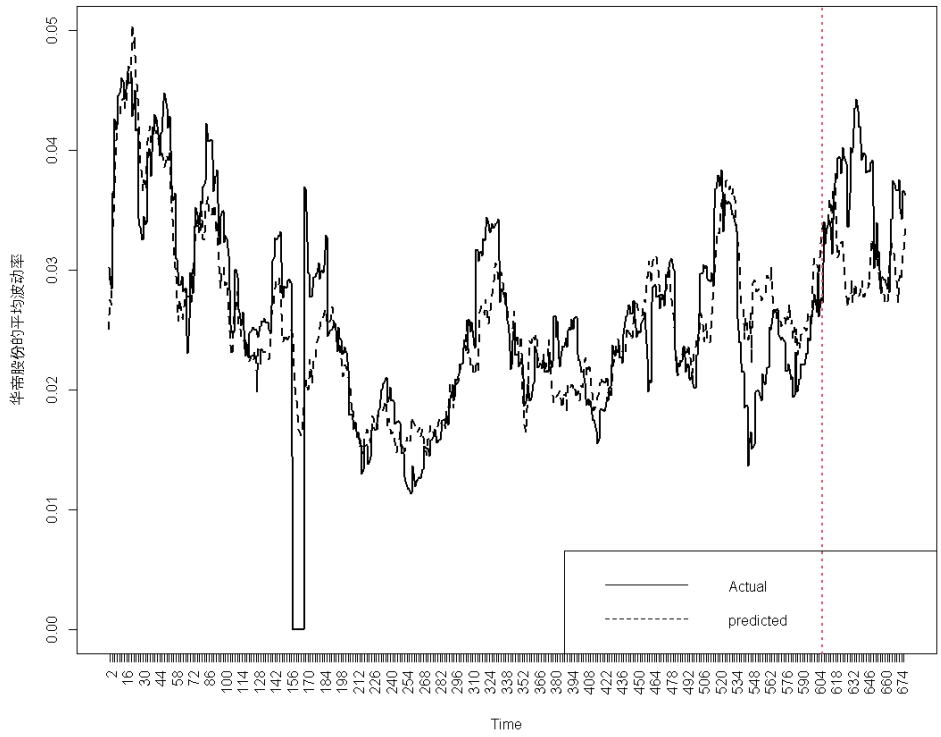


图 15 华帝股份的平均波动率的实际值和预测值

因此，根据上述模型，计算舆情事件的处置效应，如下图 16 所示。图 16 很好地反应实际值与预测值之间的差距，在“法国夺冠，华帝退全款”事件爆出后，2018 年 7 月 2 日当天华帝股份的日波动跌幅达 10.63%，日后都出现大幅度跳水。通过模型拟合结果计算，处置效应达到 11.484%。

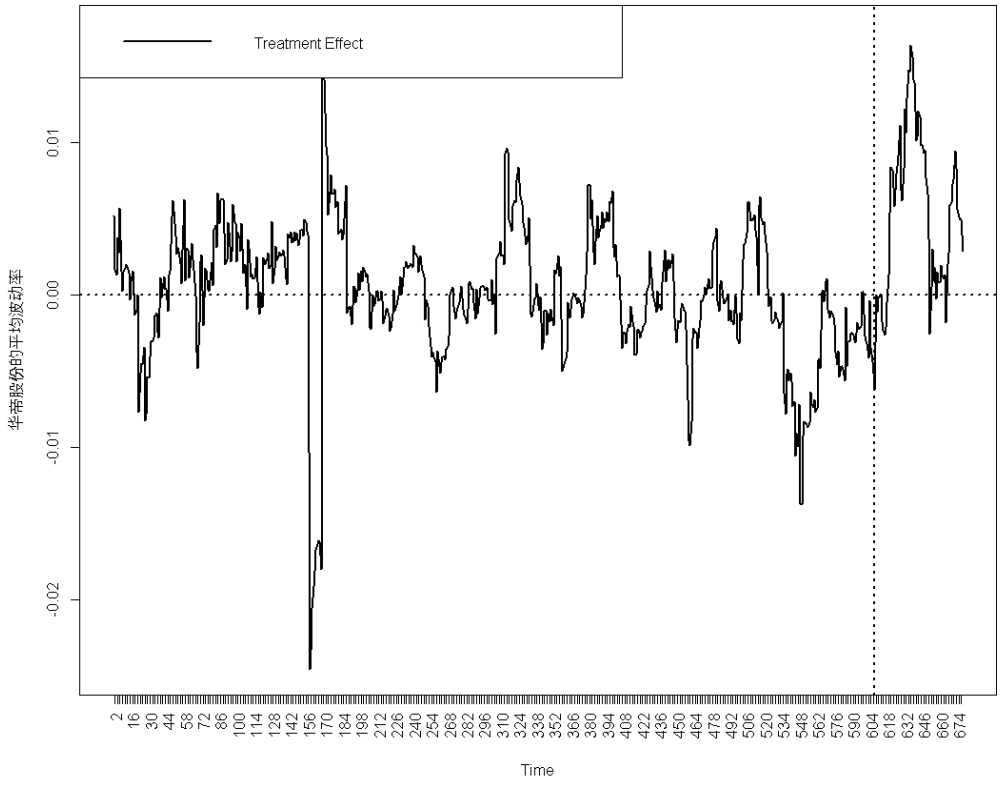


图 16 舆情事件的处置效应

3.3 安慰剂检验

为了检验模型的稳定性和预测的有效性的检验，本节对构建的模型进行安慰剂检验和稳健性检验。本文的安慰剂检验思想是通过将处理重新分配给其他未处理的单元，来迭代面板数据方法的应用；或在事件尚未发生时，将处置效应新分配到其它干预前阶段。

下图 17 为控制分配的安慰剂检验图，将华帝股份 2016 年 8 月份“非公开发行 A 股”事件纳入考虑范围，检验模型对事件冲击效应的评估。图 17 很好地检验了模型的有效性，模型稳定。

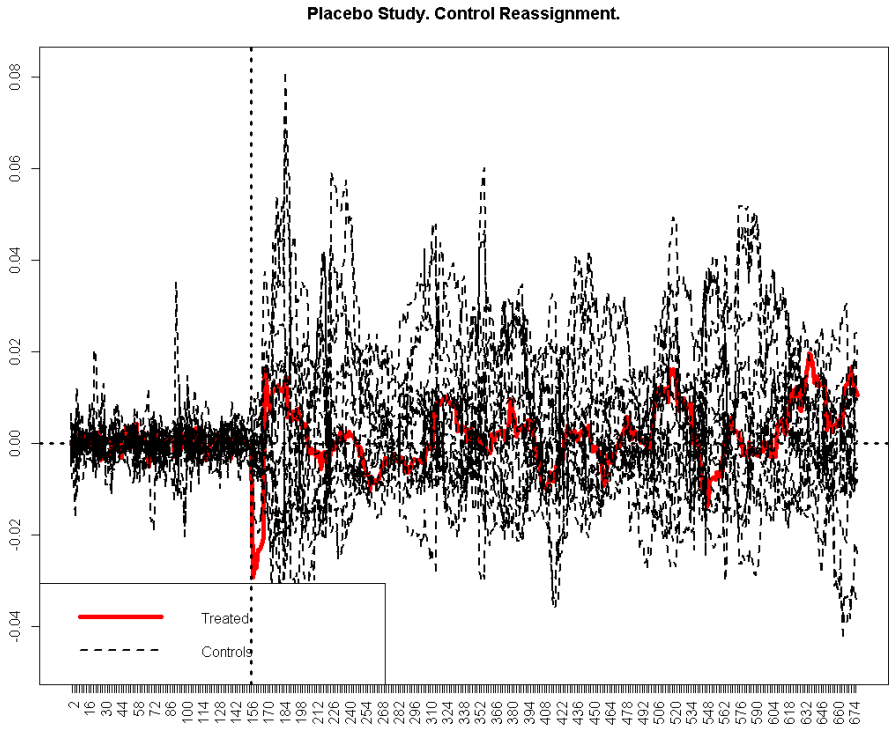


图 17 控制分配的安慰剂检验图

同时也进行 Time Reassignment 检验，检验结果如图 5 所示，以非公开发行 A 股事件为例，结果显示模型在处理事件上具有稳定性。

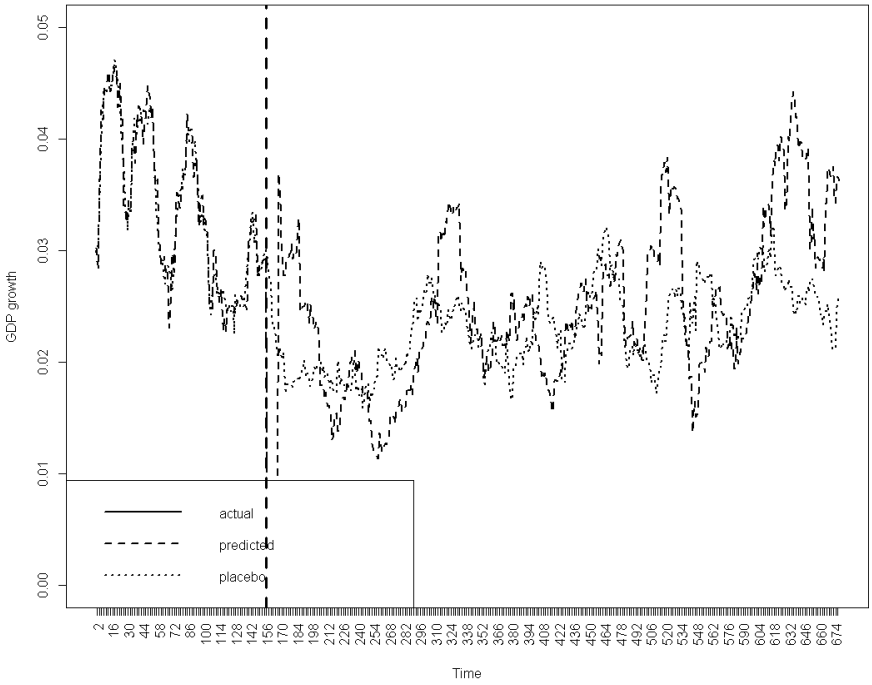


图 18 Time Reassignment 检验结果图

4. 量化策略再研究

由于外部舆情事件具有突发性，且通过第三节的建模，这种事件在短时间内对目标公司的股票价格影响较大。当这种事件不足以让目标公司倒闭，那么事件的影响会随着时间而衰弱，最终使得公司股票价格回归到理性价格。因此，可以通过第 3 节的模型方法不断测试同行业内的股票价格与目标公司股票之间的关系，对目标公司股票的理性价格进行预测。投资策略设计如下：

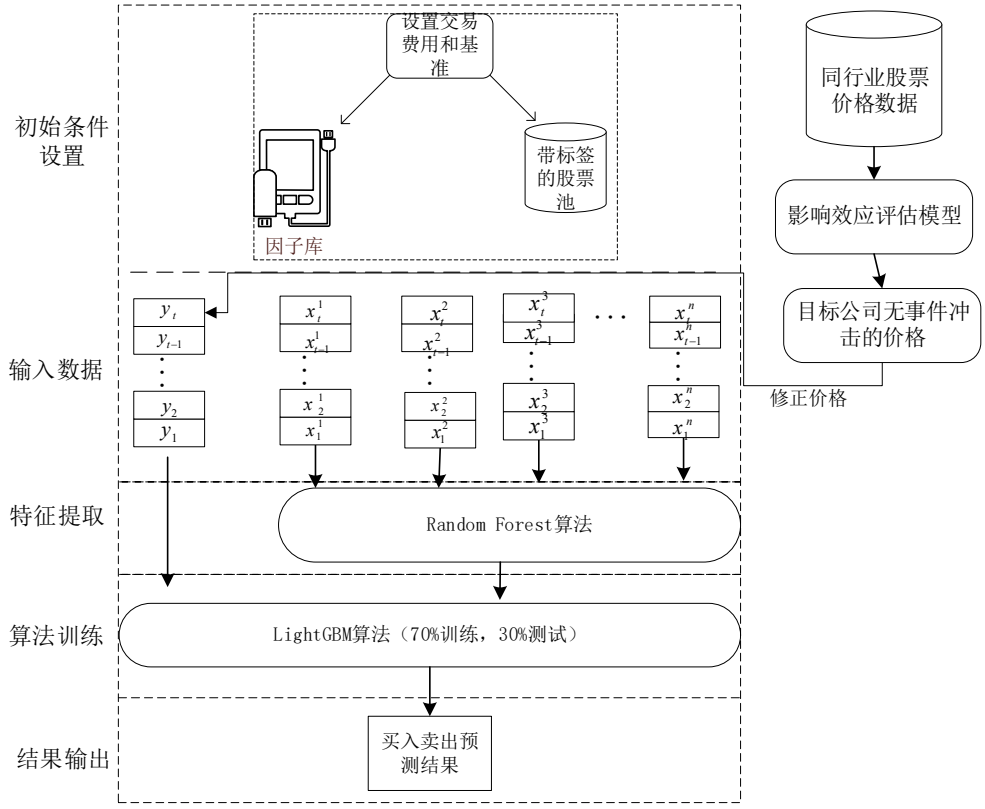


图 19 带舆情事件的量化投资策略

利用第 3 节中的模型修正受舆情事件影响的股票价格，然后利用机器学习的方法判断买卖点。

5.后续研究

由于本论文只探讨了研报中分析师对目标公司股票的评级和财务状况，提取非机构性因子，但是研报的作用远远大于这些，包括预测净利润增长率、预测主营业务增长率、盈利预测调整等预测数据。对于后续研究有两点：

（1）丰富研报因子库。调整研报因子提取方法，将多种非结构数据转换为结构性数据，如研报中带情绪的词语出现的字数。因为仅仅对分析师的评级或推荐等级不能完全反映分析师的情感，而且根据研报的推荐统计，发现绝大部分研报的推荐等级为买入，很少出现卖出，这与实际不符合。

（2）增加投资者或市场情感因子。中国股票市场投资者结构中，散户投资者占比较大，且散户投资者易受外部环境的影响，投资具有羊群效应。获取信息的渠道狭窄，对外部信息的获取和处理能力较弱，主要来源于新闻，因此在后续研究中可以尝试从散户投资者获取新闻的渠道提取散户的投资情绪，丰富因子库。

参考文献

[1] Meesad, P., & Li, J. (2014). Stock trend prediction relying on text mining and sentiment analysis with tweets. In the 4th IEEE World Congress on Information and Communication Technologies (WICT) (pp. 257-262). https://doi.org/10.1109/WICT.2014.7077275.

[2] Jason Hsu, Vitali Kalesnik & Engin Kose (2019) What Is Quality?, Financial Analysts Journal, 75:2, 44-61, DOI: 10.1080/0015198X.2019.1567194.

[3] 李展和刘富兵，国泰君安证券，2017.06.15《基于短周期价量特征的多因子选股体系》.

[4] Ainhoa Vega-Bayo.(2015). An R Package for the Panel Approach Method for Program Evaluation: pamper.

[5] Dechow, P. M., & Ge, W. (2006). The persistence of earnings and cash flows and the role of special items: Implications for the accrual anomaly. Review of Accounting Studies, 11(2), 253-296.

[6] David Hirshleifer, Kewei Hou, Siew Hong Teoh, Yinglei Zhang.(2004). Do investors overvalue firms with bloated balance sheets?. Journal of Accounting and Economics 38 (2004) 297–331.

[7] Kamaladdin Fataliyev, Aneesh Chivukula, Mukesh Prasad, and Wei Liu. 2021. Text-based Stock Market Analysis: A Review. 1, 1 (July 2021), 30 pages.

[8] Belo, F., X. Lin, and M. Vitorino. 2014. Brand capital and firm value. Review of Economic Dynamics 17: 150–69.

[9] Bai,Hang, Kewei Hou, Howard Kung, Erica X.N.Li, and Lu Zhang, 2019, The CAPM strikes back? An equilibrium model with disasters, Journal of Financial Economics 131, 269-298.

[10]韩雨薇.基于深度学习的多因子股票风险预测方法研究[D].导师：周波.浙江大学,2020.

[11]Dain C. Donelson,Robert J. Resutek.The predictive qualities of earnings volatility and earnings uncertainty[J].Review of Accounting Studies,2015,20(1):.

[12]Samir M. El-Gazzar,Afaf A. Shalaby.Accounting practices, earnings volatility, and earnings forecast: The case of the oil and gas producing companies[J].International Advances in Economic Research,1995,1(1):.

1.附录一

因子顺序	因子构建方式
Alpha3	*SUM((CLOSE=DELAY(CLOSE,1)?0:CLOSE-(CLOSE>DELAY(CLOSE,1)?MIN(LOW,DELAY(CLOSE,1)):MAX(HIGH,DELAY(CLOSE,1)))),6)*
Alpha9	*SMA(((HIGH+LOW)/2-(DELAY(HIGH,1)+DELAY(LOW,1))/2)*(HIGH-LOW)/VOLUME,7,2)*
Alpha10	*(RANK(MAX(((RET < 0) ? STD(RET, 20) : CLOSE)^2),5))*
Alpha11	*SUM(((CLOSE-LOW)-(HIGH-CLOSE))./(HIGH-LOW).*VOLUME,6)*
Alpha12	*(RANK((OPEN - (SUM(VWAP, 10) / 10)))) * (-1 * (RANK(ABS((CLOSE - VWAP)))))*
Alpha14	*CLOSE-DELAY(CLOSE,5) A*
Alpha18	*CLOSE/DELAY(CLOSE,5)*
Alpha21	*REGBETA(MEAN(CLOSE,6),SEQUENCE(6))*
Alpha23	*SMA((CLOSE>DELAY(CLOSE,1)?STD(CLOSE:20),0),20,1)/(SMA((CLOSE>DELAY(CLOSE,1)?STD(CLOSE,20):0),20,1)+SMA((CLOSE<=DELAY(CLOSE,1)?STD(CLOSE,20):0),20,1))*100*
Alpha24	*SMA(CLOSE-DELAY(CLOSE,5),5,1)*
Alpha25	*((-1 * RANK((DELTA(CLOSE, 7) * (1 - RANK(DECAYLINEAR((VOLUME / MEAN(VOLUME,20)), 9)))))) * (1 + RANK(SUM(RET, 250))))*
Alpha26	*((((SUM(CLOSE, 7) / 7) - CLOSE)) + ((CORR(VWAP, DELAY(CLOSE, 5), 230))))*
Alpha29	*(CLOSE-DELAY(CLOSE,6))/DELAY(CLOSE,6)*VOLUME*
Alpha31	*(CLOSE-MEAN(CLOSE,12))/MEAN(CLOSE,12)*100*
Alpha33	*((((-1 * TSMIN(LOW, 5)) + DELAY(TSMIN(LOW, 5), 5)) * RANK(((SUM(RET, 240) - SUM(RET, 20)) / 220))) * TSRANK(VOLUME, 5))*
Alpha34	*MEAN(CLOSE,12)/CLOSE*
Alpha35	*(MIN(RANK(DECAYLINEAR(DELTA(OPEN, 1), 15)), RANK(DECAYLINEAR(CORR((VOLUME), ((OPEN * 0.65) + (OPEN *0.35)), 17),7))) * -1)*
Alpha40	*SUM((CLOSE>DELAY(CLOSE,1)?VOLUME:0),26)/SUM((CLOSE<=DELAY(CLOSE,1)?VOLUME:0),26)*100*
Alpha41	*(RANK(MAX(DELTA((VWAP), 3), 5))* -1)*
Alpha43	*SUM((CLOSE>DELAY(CLOSE,1)?VOLUME:(CLOSE<DELAY(CLOSE,1)?-VOLUME:0)),6)*
Alpha47	*SMA((TSMAX(HIGH,6)-CLOSE)/(TSMAX(HIGH,6)-TSMIN(LOW,6))*100,9,1)*
Alpha49	*SUM(((HIGH+LOW)>=(DELAY(HIGH,1)+DELAY(LOW,1))?0:MAX(ABS(HIGH-DELAY(HIGH,1)),ABS(LOW-DELAY(LOW,1)))),12)/(SUM(((HIGH+LOW)>=(DELAY(HIGH,1)+DELAY(LOW,1))?0:MAX(ABS(HIGH-DELAY(HIGH,1)),ABS(LOW-DELAY(LOW,1)))),12)+SUM(((HIGH+LOW)<=(DELAY(HIGH,1)+DELAY(LOW,1))?0:MAX(ABS(HIGH-DELAY(HIGH,1)),ABS(LOW-DELAY(LOW,1)))),12))*
Alpha50	*SUM(((HIGH+LOW)<=(DELAY(HIGH,1)+DELAY(LOW,1))?0:MAX(ABS(HIGH-DELAY(HIGH,1)),ABS(LOW-DELAY(LOW,1)))),12)/(SUM(((HIGH+LOW)<=(DELAY(HIGH,1)+DELAY(LOW,1))?0:MAX(ABS(HIGH-DELAY(HIGH,1)),ABS(LOW-DELAY(LOW,1)))),12)+SUM(((HIGH+LOW)>=(DELAY(HIGH,1)+DELAY(LOW,1))?0:MAX(ABS(HIGH-DELAY(HIGH,1)),ABS(LOW-DELAY(LOW,1)))),12))-SUM(((HIGH+LOW)>=(DELAY(HIGH,1)+DELAY(LOW,1))?0:MAX(ABS(HIGH-DELAY(HIGH,1)),ABS(LOW-DELAY(LOW,1)))),12)/(SUM(((HIGH+LOW)>=(DELAY(HIGH,1)+DELAY(LOW,1))?0:MAX(ABS(HIGH-DELAY(HIGH,1)),ABS(LOW-DELAY(LOW,1)))),12)+SUM(((HIGH+LOW)<=(DELAY(HIGH,1)+DELAY(LOW,1))?0:MAX(ABS(HIGH-DELAY(HIGH,1)),ABS(LOW-DELAY(LOW,1)))),12))*
Alpha51	*SUM(((HIGH+LOW)<=(DELAY(HIGH,1)+DELAY(LOW,1))?0:MAX(ABS(HIGH-DELAY(HIGH,1)),ABS(LOW-DELAY(LOW,1)))),12)/(SUM(((HIGH+LOW)<=(DELAY(HIGH,1)+DELAY(LOW,1))?0:MAX(ABS(HIGH-DELAY(HIGH,1)),ABS(LOW-DELAY(LOW,1)))),12)+SUM(((HIGH+LOW)>=(DELAY(HIGH,1)+DELAY(LOW,1))?0:MAX(ABS(HIGH-DELAY(HIGH,1)),ABS(LOW-DELAY(LOW,1)))),12))*
Alpha52	*SUM(MAX(0,HIGH-DELAY((HIGH+LOW+CLOSE)/3,1)),26)/SUM(MAX(0,DELAY((HIGH+LOW+CLOSE)/3,1)-L),26)*100*

Alpha53	COUNT(CLOSE>DELAY(CLOSE,1),12)/12*100
Alpha57	SMA((CLOSE-TSMIN(LOW,9))/(TSMAX(HIGH,9)-TSMIN(LOW,9))*100,3,1)
Alpha58	COUNT(CLOSE>DELAY(CLOSE,1),20)/20*10
Alpha61	(MAX(RANK(DECAYLINEAR(DELTA(VWAP, 1), 12)), RANK(DECAYLINEAR(RANK(CORR((LOW),MEAN(VOLUME,80), 8)), 17))) * -1)
Alpha65	MEAN(CLOSE,6)/CLOSE
Alpha66	(CLOSE-MEAN(CLOSE,6))/MEAN(CLOSE,6)*100
Alpha67	SMA(MAX(CLOSE-DELAY(CLOSE,1),0),24,1)/SMA(ABS(CLOSE-DELAY(CLOSE,1)),24,1)*100
Alpha68	SMA(((HIGH+LOW)/2-(DELAY(HIGH,1)+DELAY(LOW,1))/2)*(HIGH-LOW)/VOLUME,15,2
Alpha70	STD(AMOUNT,6)

2.重要代码部分：

```
#--------舆情事件部分-------------------------
library(readxl)
library(tidyr)
library(tidyverse)
library(stargazer)
data <- read_excel("D:/data2.xls")
View(data)

data <- as_tibble(data)
data$最新股票名称_Lstknm <- as.factor(data$最新股票名称_Lstknm)
data <- data %>%
  group_by(最新股票名称_Lstknm) %>%
  spread(最新股票名称_Lstknm, '日收益标准差_20 日移动平均_Dstd20')
head(data)

library(VIM)
aggr(data)
data <- data[,c(-15)]
data$荣泰健康 <- ifelse(is.na(data$荣泰健康) == TRUE, 0, data$荣泰健康)
data$三花智控 <- ifelse(is.na(data$三花智控) == TRUE, 0, data$三花智控)
data$格力电器 <- ifelse(is.na(data$格力电器) == TRUE, 0, data$格力电器)
data$美的集团 <- ifelse(is.na(data$美的集团) == TRUE, 0, data$美的集团)
data$TCL 科技 <- ifelse(is.na(data$TCL 科技) == TRUE, 0, data$TCL 科技)
data$奥佳华 <- ifelse(is.na(data$奥佳华) == TRUE, 0, data$奥佳华)
data$飞科电器 <- ifelse(is.na(data$飞科电器) == TRUE, 0, data$飞科电器)
```

```
data$苏泊尔 <- ifelse(is.na(data$苏泊尔) == TRUE, 0, data$苏泊尔)
data$海尔智家 <- ifelse(is.na(data$海尔智家) == TRUE, 0, data$海尔智家)
data$奥佳华 <- ifelse(is.na(data$奥佳华) == TRUE, 0, data$奥佳华)
data$海信视像 <- ifelse(is.na(data$海信视像) == TRUE, 0, data$海信视像)
data$华帝股份 <- ifelse(is.na(data$华帝股份) == TRUE, 0, data$华帝股份)
data$九阳股份 <- ifelse(is.na(data$九阳股份) == TRUE, 0, data$九阳股份)
data$兆驰股份 <- ifelse(is.na(data$兆驰股份) == TRUE, 0, data$兆驰股份)
data$万和电气 <- ifelse(is.na(data$万和电气) == TRUE, 0, data$万和电气)
data$日出东方 <- ifelse(is.na(data$日出东方) == TRUE, 0, data$日出东方)
aggr(data)
data <- as.data.frame(data)
stargazer(data, title = "Table 1. descriptive statistic",type ="html",
          out='D:/descriptive statistic.txt')

#--------------------模型检验-----------------------------
#install.packages("pampe")
#install.packages("leaps")
library(pampe)
library(leaps)
??pample

treated <- "华帝股份"
time.pretr <- 1:607
time.tr <- 608:677
possible.ctrls <- c('奥佳华','TCL 科技','飞科电器','格力电器','海尔智家','海信家电',
                    '莱克电气','老板电器','美的集团','海信视像','九阳股份','四
川九洲',
                    '苏泊尔','新宝股份','三花智控','浙江美大','荣泰健康','日出
东方','兆驰股份')

pol.integ1 <- pampe(time.pretr = time.pretr, time.tr = time.tr, treated = treated,
                    controls = possible.ctrls, data = data,select = "AIC")
pol.integ2 <- pampe(time.pretr = time.pretr, time.tr = time.tr, treated = treated,
                    controls = possible.ctrls, data = data,select = "BIC")
```

```
summary(pol.integ1)
summary(pol.integ2)
#stargazer(pol.integ2, title = "Results", align = TRUE)
plot(pol.integ)
plot(pol.integ2)

#    A plot of the actual Hong Kong together with the predicted path
matplot(c(time.pretr, time.tr), pol.integ2$counterfactual, type = "l", xlab = "",
              ylab = "华帝股份的平均波动率", ylim = c(0, 0.05), col = 1, lwd = 2, xaxt
= "n")

axis(1, at = c(time.pretr, time.tr)[c(seq(2, length(c(time.pretr, time.tr)),
                        by  =  2))],  labels  =  c(rownames(data$ 日 期 _Date)[c(time.pretr,
time.tr)
                        [c(seq(2, length(c(time.pretr, time.tr)), by = 2))]]), las = 3)
title(xlab = "Time", mgp = c(3.6, 0.5, 0))
legend("bottomright", c("Actual", "predicted"),
           col = 1, lty = c(1, 2), lwd = 1)
abline(v = time.pretr[length(time.pretr)], lty = 3, lwd = 2,col = 2)

# A plot of the estimated treatment effect
tr.effect <- pol.integ2$counterfactual[, 1] - pol.integ2$counterfactual[, 2]
plot(c(time.pretr, time.tr), tr.effect, type = "l", ylab = "华帝股份的平均波动率",
       xlab = "", col = 1, lwd = 2, xaxt = "n", ylim = c(-0.05, 0.05))
axis(1, at = c(time.pretr, time.tr)[c(seq(2, length(c(time.pretr, time.tr)),
                        by = 2))], labels = c(rownames(data)[c(time.pretr, time.tr)
                        [c(seq(2, length(c(time.pretr, time.tr)), by = 2))]]), las = 3)
title(xlab = "Time", mgp = c(3.6, 0.5, 0))
legend("topleft", "Treatment Effect", col = 1, lty = 1, lwd = 2)
abline(v = time.pretr[length(time.pretr)], lty = 3, lwd = 2)
abline(h = 0, lty = 3, lwd = 2)

# 安慰剂检验
pol.integ.placebos <- pampe(time.pretr = time.pretr, time.tr = time.tr,
```

```
                                    treated = treated, controls = possible.ctrls,
                                    data = data, placebos = "Both", select = "BIC")
mspe <- pol.integ.placebos$placebo.ctrl$mspe
linewidth <- matrix(2, 1, ncol(mspe) - 1)
linewidth <- append(linewidth, 5, after = 0)
matplot(c(time.pretr, time.tr), pol.integ.placebos$placebo.ctrl$tr.effect,
        type = "l", xlab = "", ylab = "华帝股份的平均波动率",
        col = c("red", matrix(1, 1, ncol(mspe) - 1)),
        lty = c(1, matrix(2, 1, ncol(mspe) - 1)), lwd = linewidth,
        ylim = c(-0.05, 0.05), xaxt = "n")

axis(1, at = c(time.pretr, time.tr)[c(seq(2, length(c(time.pretr, time.tr)),
                    by = 2))], labels = c(rownames(data)[c(time.pretr, time.tr)
                    [c(seq(2, length(c(time.pretr, time.tr)), by = 2))]]), las = 3)
title(xlab = "Time", mgp = c(3.6, 0.5, 0))
legend("bottomleft", c("Treated", "Controls"), col = c("red", 1),
       lty = c(1, 2), lwd = c(5, 2))
abline(h = 0, lty = 3, lwd = 2)
abline(v = time.pretr[length(time.pretr)], lty = 3, lwd = 2)

#-------------------------------------------------------------------------------
treated <- "华帝股份"
time.pretr <- 1:156
time.tr <- 157:677
possible.ctrls <- c('奥佳华','TCL 科技','飞科电器','格力电器','海尔智家','海信家电',
                    '莱克电气','老板电器','美的集团','海信视像','九阳股份','四
川九洲',
                    '苏泊尔','新宝股份','三花智控','浙江美大','荣泰健康','日出
东方','兆驰股份')
pol.integ.placebos <- pampe(time.pretr = time.pretr, time.tr = time.tr, treated = treated,
                    controls = possible.ctrls, data = data,select = "BIC",placebos
= "controls")
summary(pol.integ.placebos)
#stargazer(pol.integ2, title = "Results", align = TRUE)
```

```
plot(pol.integ.placebos)

#-------------------------------------------------------------------------------
placebo.in.time1 <- pol.integ.placebos$placebo.time$tr.effect[, 2] +
   data[c(time.pretr, time.tr), 1]
matplot(c(time.pretr, time.tr), cbind(data[c(time.pretr, time.tr), 1],
                pol.integ.placebos$counterfactual, placebo.in.time1), type = "l",
          ylab = "GDP growth", xlab = "", ylim = c(0, 0.05), col = 1,
          lwd = 2, xaxt = "n")
axis(1, at = c(time.pretr, time.tr)[c(seq(2, length(c(time.pretr, time.tr)),
                          by = 2))], labels = c(rownames(data)[c(time.pretr, time.tr)
                        [c(seq(2, length(c(time.pretr, time.tr)), by = 2))]]), las = 3)

title(xlab = "Time", mgp = c(3.6, 0.5, 0))
legend("bottomleft", c("actual", "predicted", paste("placebo",
          colnames(pol.integ.placebos$placebo.time$tr.effect)[2], sep = " ")),
          col = 1, lty = c(1, 2, 3), lwd = 2)
abline(v = time.pretr[length(time.pretr)], lty = 2, lwd = 3)
abline(v = which(colnames(pol.integ.placebos$placebo.time$tr.effect)[2]
                              == rownames(data)), lty = 3, lwd = 3)

#-------单因子检验------------------------------
import jqdata
from jqlib.alpha191 import *
from jqfactor import *

def initialize(context):
      set_benchmark('000300.XSHG')
      set_option('use_real_price', True)
      log.info('初始函数开始运行且全局只运行一次')
      g.i = 0

      set_order_cost(OrderCost(close_tax=0.001,              open_commission=0.0003,
close_commission=0.0003, min_commission=5), type='stock')
```

```
    run_daily(before_market_open, time='before_open')
    run_daily(market_open, time='open')

def before_market_open(context):
    if g.i%8 == 0:
        log.info('函数运行时间(before_market_open)：'+str(context.current_dt.time()))
        current_security = context.portfolio.positions.keys()
        current_date = context.previous_date

        codes = ['002027.XSHE','000636.XSHE','600323.XSHG','002035.XSHE','300014.XSHE','001914.XSHE','601318.XSHG','002511.XSHE','600048.XSHG','600332.XSHG']
        alpha_stocks = alpha_003(codes,current_date).dropna().order(ascending=True)
        alpha_head = alpha_stocks.head(5).index
        log.info('\n',alpha_stocks.head(5))
        g.stocks_to_buy = list(set(alpha_head)-set(current_security))
        g.stocks_to_sell = list(set(current_security)-set(alpha_head))
        g.i += 1
    else:
        g.stocks_to_buy = []
        g.stocks_to_sell = []
        g.i += 1

def market_open(context):
    for stock in g.stocks_to_sell:
        order_target(stock,0)
        log.info("卖出 %s" % (stock))
    try:
        g.cash = context.portfolio.available_cash/len(g.stocks_to_buy)
    except:
        g.cash = 0
    for stock in g.stocks_to_buy:
```

```
            order_value(stock, g.cash)
            log.info("买入  %s" % (stock))

#-----单因子短期价量检验-----------
import numpy as np
import pandas as pd
import jqfactor
from jqlib.alpha191 import *
import datetime

#获取日期列表
def get_tradeday_list(start,end,frequency=None,count=None):
    if count != None:
        df = get_price('000001.XSHG',end_date=end,count=count)
    else:
        df = get_price('000001.XSHG',start_date=start,end_date=end)
    if frequency == None or frequency =='day':
        return df.index
    else:
        df['year-month'] = [str(i)[0:7] for i in df.index]
        if frequency == 'month':
            return df.drop_duplicates('year-month').index
        elif frequency == 'quarter':
            df['month'] = [str(i)[5:7] for i in df.index]
            df =  df[(df['month']=='01') | (df['month']=='04') | (df['month']=='07') |
(df['month']=='10') ]
            return df.drop_duplicates('year-month').index
        elif frequency =='halfyear':
            df['month'] = [str(i)[5:7] for i in df.index]
            df = df[(df['month']=='01') | (df['month']=='06')]
            return df.drop_duplicates('year-month').index

# 设置起止时间
start='2020-10-31'
```

```
end='2021-10-31'
# 设置调仓周期
periods=(3,6,10)
securities =    [
        '002027.XSHE',
        '000636.XSHE',
        '600323.XSHG',
        '002035.XSHE',
        '300014.XSHE',
        '001914.XSHE',
        '601318.XSHG',
        '002511.XSHE',
        '600048.XSHG',
        '600332.XSHG'

]

# 设置分层数量
quantiles=len(securities)
#获取日期列表
date_list = get_tradeday_list(start=start,end=end, count = None)
date_list

#定义要计算的动量因子
#def factor_cal(pool,date):
#       df = get_price(pool,end_date=date,count=21,fields=['close'])['close']
#       far = df.iloc[-1,:]/df.iloc[0,:] - 1
#       return far

#定义一个空的 dataframe 记录因子值
factor_df = pd.DataFrame()
#循环计算给定日期范围的因子值
mark = 1
```

```
for d in date_list:
    pool = securities

    far1=alpha_068(pool, end_date=d)
    far2=alpha_069(pool, end_date=d)
    if mark == 1:
        factor_df1 = far1
        factor_df2 = far2
        mark = 0
    else:
        #逐日合并 factor_df
        factor_df1 = pd.concat([far1,factor_df1],axis=1,sort=True)
        factor_df2 = pd.concat([far2,factor_df2],axis=1,sort=True)

#将 columns 更改为可以日期标签
factor_df1.columns = date_list
factor_df2.columns = date_list

from jqfactor import *
#数据清洗、包括去极值、标准化、中性化等,并加入 y 值
for date in date_list:
    #对数据进行处理、标准化、去极值、中性化
    factor_df1 = winsorize_med(factor_df1, scale=3, inclusive=True, inf2nan=True,
axis=0) #中位数去极值处理
    se1 = standardlize(factor_df1[date], inf2nan=True) #对每列做标准化处理
    se1 = neutralize(se1, how=['liquidity'], date=date)#剔除原始因子值与流动性相
关的部分
    factor_df1[date] = se1

    factor_df2 = winsorize_med(factor_df2, scale=3, inclusive=True, inf2nan=True,
axis=0) #中位数去极值处理
    se2 = standardlize(factor_df2[date], inf2nan=True) #对每列做标准化处理
    se2 = neutralize(se2, how=['liquidity'], date=date)#剔除原始因子值与流动性相
关的部分
```

```
    factor_df2[date] = se2

#进行转置，调整为分析可用的格式
factor_df1 = factor_df1.T
factor_df2 = factor_df2.T

far1 = analyze_factor(factor=factor_df1, start_date=date_list[0], end_date=date_list[-1], weight_method='avg', industry='jq_l1', quantiles=quantiles, periods=periods,max_loss=0.3)
far2 = analyze_factor(factor=factor_df2, start_date=date_list[0], end_date=date_list[-1], weight_method='avg', industry='jq_l1', quantiles=quantiles, periods=periods,max_loss=0.3)
#调用因子分析方法，进行因子信息全览:分位数统计
#far.create_full_tear_sheet(demeaned=False, group_adjust=False, by_group=False, turnover_periods=None, avgretplot=(5, 15), std_bar=False)
# 打印信息比率（IC）相关表
far1.plot_information_table(group_adjust=False, method='rank')
far2.plot_information_table(group_adjust=False, method='rank')
```