沪深300单因子选股测试

目录

[沪深300单因子选股测试 1](#_Toc43065834)

[背景介绍 2](#_Toc43065835)

[研究对象 2](#_Toc43065836)

[Alpha因子 2](#_Toc43065837)

[因子评价 3](#_Toc43065838)

[因子选择 4](#_Toc43065839)

[基本面指标 4](#_Toc43065840)

[技术面指标 9](#_Toc43065841)

[总结 14](#_Toc43065842)

[结论 14](#_Toc43065843)

[不足与改进 15](#_Toc43065844)

叶旭枫

2019211904

# 背景介绍

## 研究对象

本课题使用2020年6月份获得的沪深300成分股作为研究对象。包含上海证券交易所与深圳证券交易所中选取的300支实值大、流通性好的A股股票。

由于股票数据较多，并且沪深300的股票池中上市日期不一，因此在本研究中，对于基本面指标的因子选用的历史数据时间跨度较大，并且选择了全体股票。对于技术面指标因子由于计算较为复杂，选取了较短的时间跨度和300支股票的前100个进行回测。

并且由于沪深300的股票行业分布共有76个不同细分类行业，因此本课题不对分行业的资产配比进行分析，这样做可能会使得风险的分布过于集中与某类行业。

本课题选用的历史数据来源于Tushare，技术指标包括了每日的股票行情数据，诸如股票代码、交易日期、开盘价、最高价、最低价、收盘价、昨日收盘价、涨跌幅、涨跌额、成交量、成交额。另外对于个股的每日基本面指标包括了换手率、量比、市盈率、市净率、滚动市盈率、滚动市销率、市销率、总股本、流通股本、总市值等信息。

## Alpha因子

Alpha因子与Alpha策略是指根据CAPM(资本资产定价模型)，投资者在市场中交易面临系统性风险和非系统性风险，

传统阿尔法策略是在基金经理建立了β部位的头寸后，通过衍生品对冲β部位的风险，从而获得正的阿尔法收益。本课题回测时没有对的β风险进行对冲，仅仅探讨对应于alpha的超额收益。

常见的Alpha因子的来源包括：基本面分析、动量策略、波动捕获策略、风格轮动、行业轮动等。

## 因子评价

因子非常重要的一个性质是有效性。测试有效性一般常用的方法有

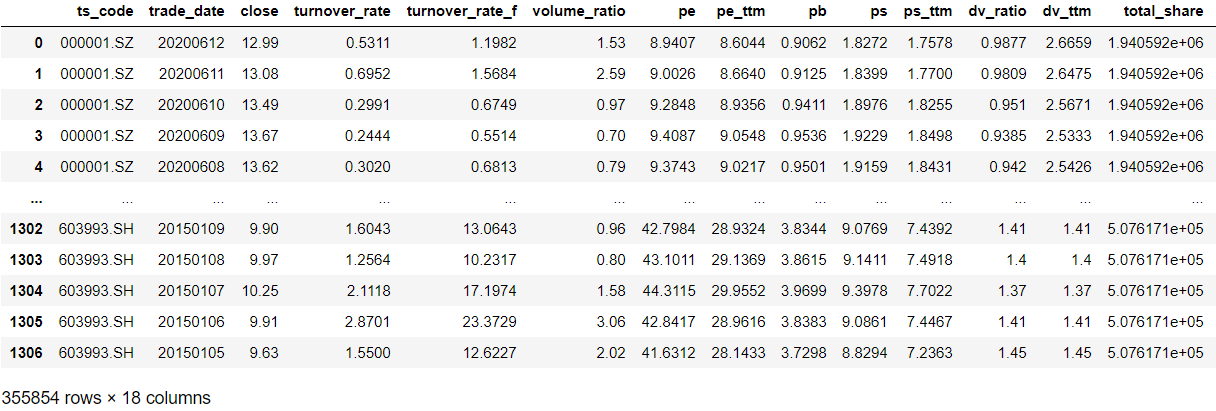
1. 回归法：回归法是一种常用的用于测试因子有效性的方法，即将T期的因子暴露度与T+1期的股票收益向量进行回归，得到的回归系数即因子收益率。同时还能得到该因子的显著性水平t值。一般认为t值的绝对值大于2即认为本期的因子对于下一期的持股收益率具有明显的解释作用。另外可以计算在完整的历史周期中，显著性大于2的占比来判断因子的稳定性。
2. IC分析法：信息系数(Information Coefficient,简称 IC)，代表因子预测股票收益的能力。IC的计算方法是：计算全部股票在调仓周期期初排名和调仓周期期末收益排名的线性相关度(Correlation)。IC越大的因子，选股能力就越强。实际使用中往往使用Spearman秩相关系数。另外还有利用IC值计算得到的IR值进行判断，信息比率 (Information Ratio, 简称 IR)= IC 的多周期均值 / IC 的标准方差，代表因子获取稳定 Alpha 的能力。
3. 分层回测：根据每个股票的因子值，从大到小排列以后，等分位多个分层组合。分层回测可以发掘因子对收益预测的非线性规律，可以有效避免因为不同类型股票对于因子的暴露程度不同使得总体的因子暴露趋于平庸。

本课题选用国外较为致命的因子回测的包Alphalens进行这一部分操作。并且由于A股市场无法进行做空操作，因此在进行回测时放弃做空操作，即不对因子进行中心化操作，在计算权重时不减去因子的平均值，使得所有的因子保留原来的正负性。

# 因子选择

## 基本面指标

先利用Tushare接口获取沪深300股票列表并且获取从2015年1月1日开始的历史每日行情数据，数据结构如下图：

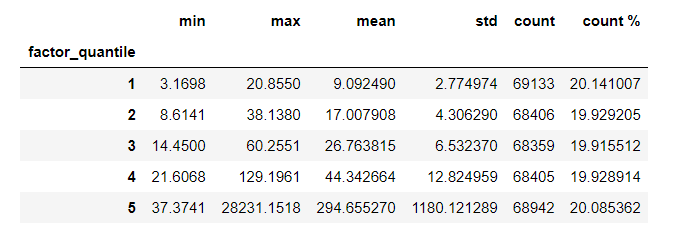


300个股票共计35万条数据。约有12个按日的基本面数据。之后对数据进行合适的格式调整与数据清洗如重复数据删除。通过观察发现在2015年1月之初，沪深300只有约235支股票在交易，在临近2020年6月份时为300支交易股票。这是因为诸多股票实在2015年至2020年之间上市的。

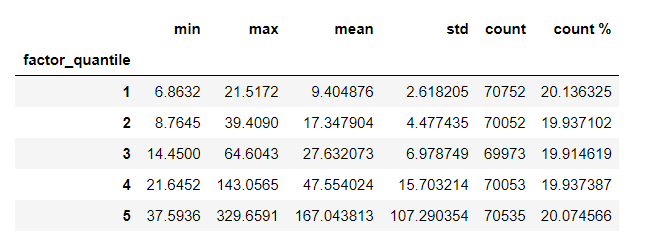
本课题对常见的市盈率、市销率、总市值等因子进行了回测，利用前文的三种因子评价标准进行描述。将300支股票按照因子从大到小排列的分位数分为了5组，并且按照5日、10日、15日(分别对应于一至三周)的持股周期进行回测，之所以没有按照最小的一天的进行回测是因为考虑到本课题没有设置交易滑点与手续费，因此过于高频的交易将使这部分忽略影响结果的准确性，因此选用一个相对合理的持股周期。

#### 市盈率：

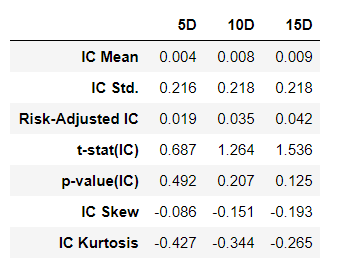
如图为市盈率作为引子时的因子分布:

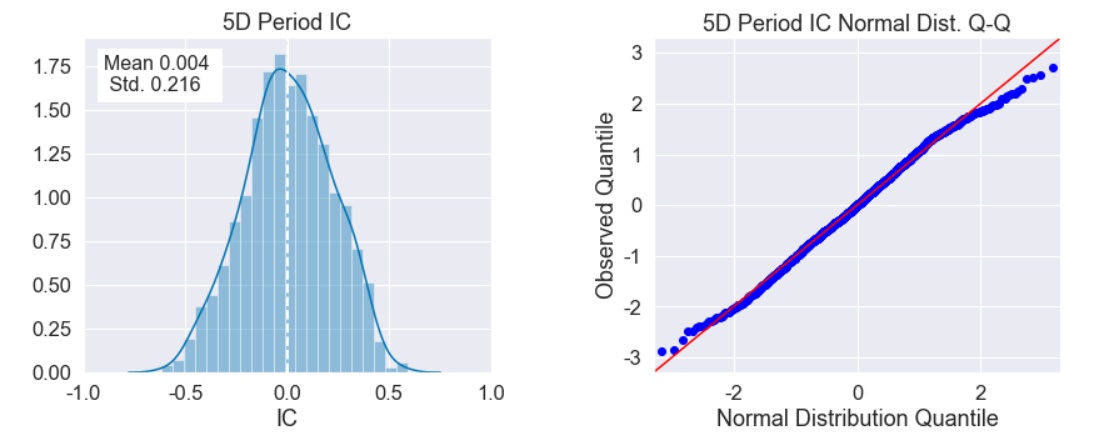


可以见到市盈率的最大值突破天际，虽然这不会影响Spearman Rank IC值的计算，因为Alphalens中的IC值取决于排行，但是这会影响计算持股权重时的各个股票的权重值，会使得在因子大小不同的股票分层中个别股票占据过大的权重，因此温莎平均是很好的处理极大值的方法，既保留了较大值的优势又不会过于影响平衡。本课题对市盈率进行了分位数5%的温莎平均，重新运行因子评价。此时因子分布值为：



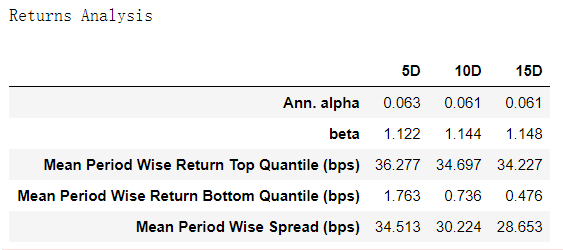
IC值(rank IC)为：



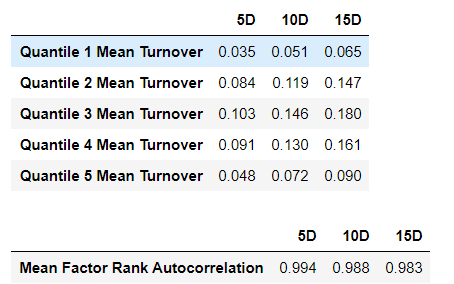


可以看到IC的分布接近于正态分布，并且均值较小，观察IC值的t值也没有大于2，因此认为市盈率作为因子实际上不是很有效。

通过收益率分析可以看到：



获取的大部分收益均来自于beta而不是Alpha。

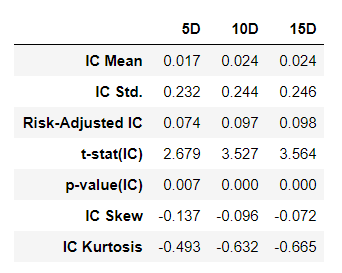


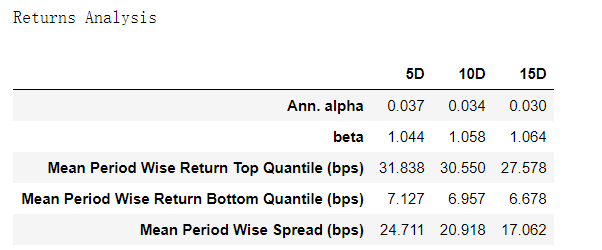
而因子的换手率与自相关性分析可以看到市盈率作为因子其自相关性很高，换手率也很低，因此我们认为市盈率作为因子并不适合作为Alpha因子，不适合进行高频交易，更适合长期持股。

#### 市销率、市净率

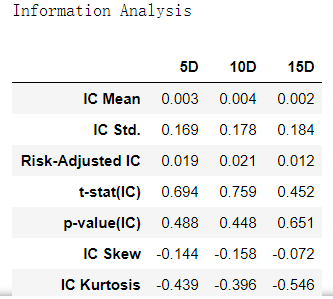
通过分位数分析，发现这两个因子同样需要进行预处理，同样选用温莎平均处理以后进行因子测评。

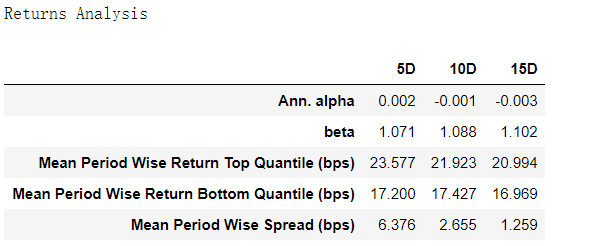
市净率：





市销率：

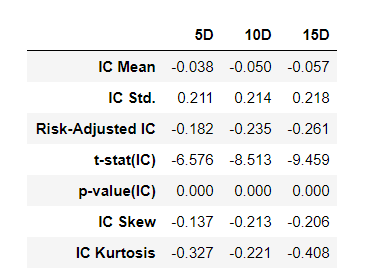


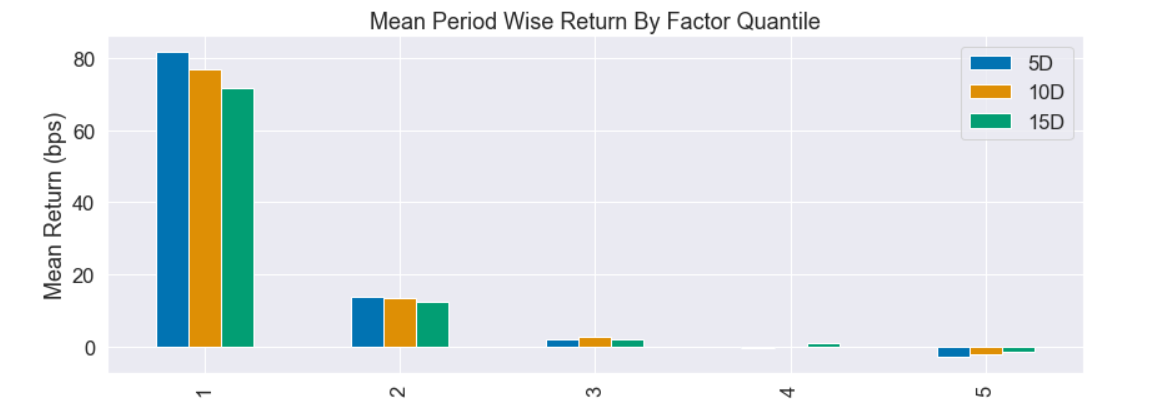


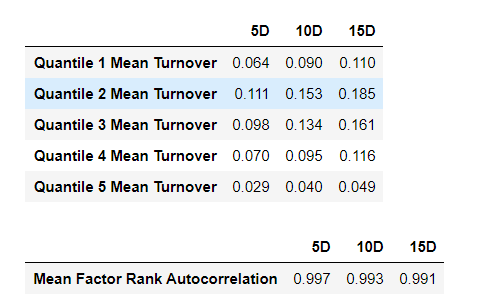
简而言之，这两个指标和市盈率因子半斤八两，IC值均比较小，并且t统计量绝对值较小。无法很明显地体现出费系统性风险带来的超额收益。

#### 市值

由于不同股票之间市值差距非常大，因此对市值指标进行放缩到0,1之间的变量，再进行测试。结果如下：







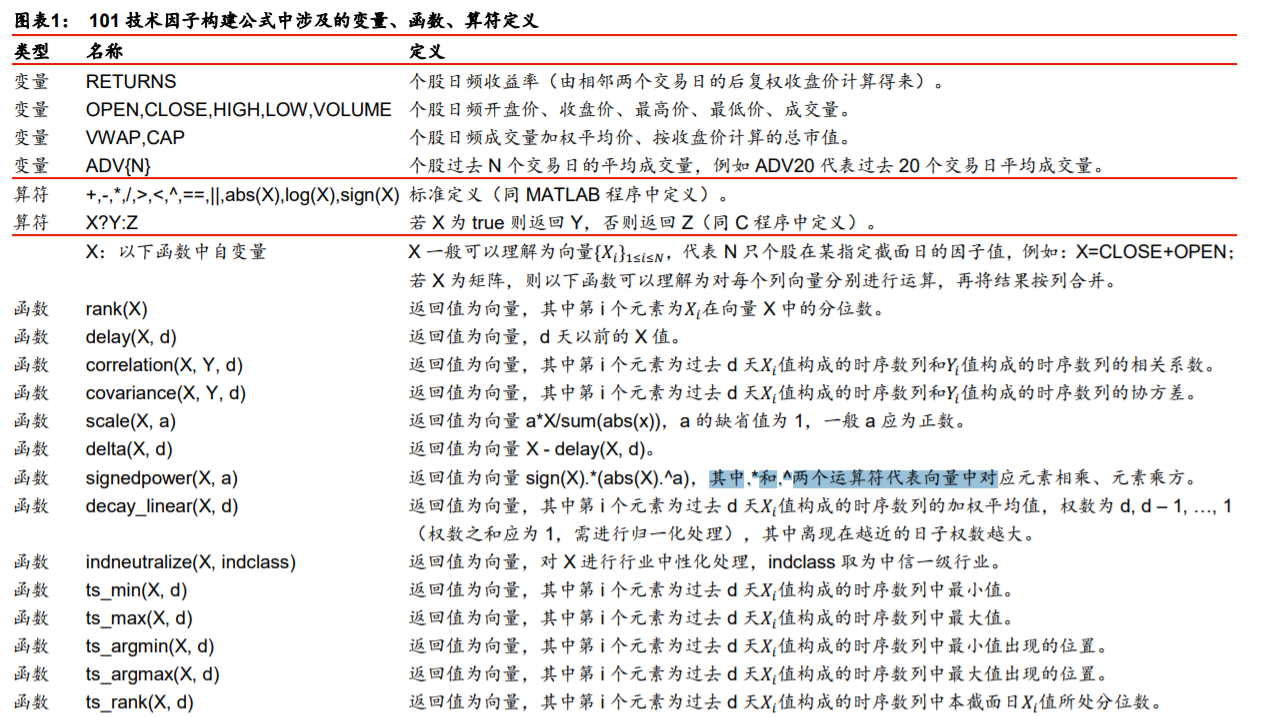
可以看到市值作为一个反向的alpha因子具有不错的IC值，t值的绝对值也比较大，因此认为市值是一个比较优秀的用于沪深300的因子，但是使用时，需要购买那些市值比较小的股票。同时看到换手率依然比较低，这表明市值因子的变动不是很大。这可以通过A股最近几年比较偏爱相对更小型的初创企业或者精品企业有关。

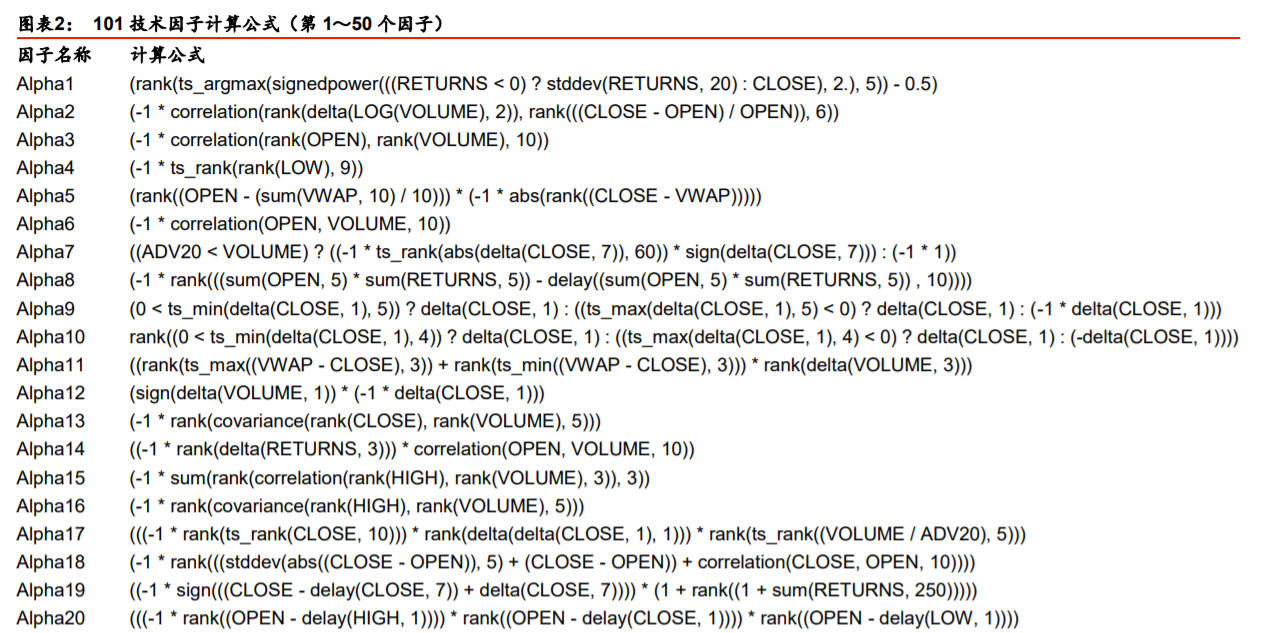
## 技术面指标

对于技术面指标，我们认为这个指标相对变化较快，因此我们选取的回测窗口期相对比较短，为1天、5天、10天。而计算技术面指标用到的窗口数据根据之前5-10天的计算得到。

这里我使用现成的来自于WorldQuant的101个因子作为因子池进行评价。这101个因子都是通过股票池的价量数据进行计算，以5天、10天等作为窗口期滚动计算得到。由于性能与时间限制，本课题对于因子的计算依赖于每获得一支股票的每日价量数据便进行一次因子计算，对于一些复杂的因子计算相对复杂，因此本课题仅仅对于其中前20个以及个别因子并且从2018年1月1日至今共两年半进行测试。

因子的计算式子如下：



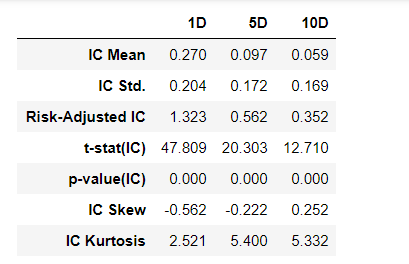


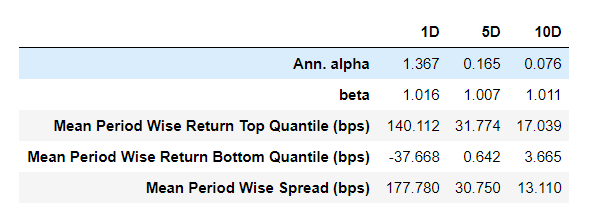
完整的内容可以参见华泰证券出品的《华泰单因子测试之海量技术因子》或者WorldQuant出品的《101 Formulaic Alphas》。

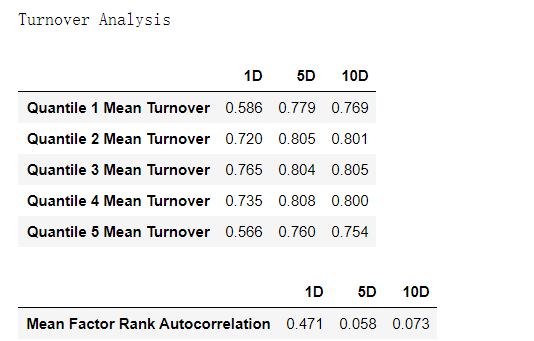
这些技术因子的计算方式虽然复杂但是有一些共同点，许多是通过选择每日价量数据中的若干个变量，然后回溯5-10日的数据计算数量之间或者秩之间的相关系数或者协方差，再构建相关关系的函数来判断是否出现了价量的背离。

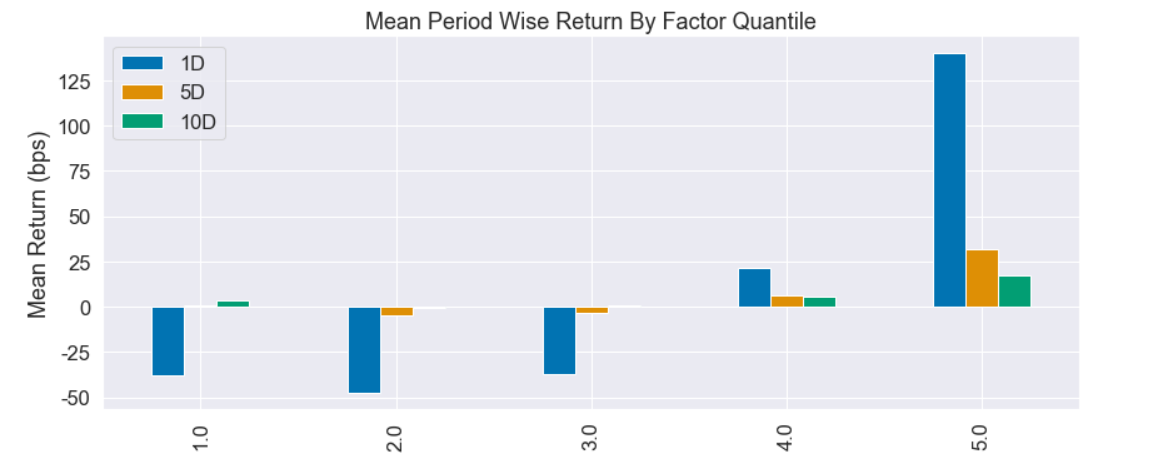
#### Alpha1

Alpha1是筛选出每只股票所有时间点上下跌的交易日，然后计算这些天回溯20天窗口期的涨跌幅的标准差，将负的下跌幅用标准差代替以后计算涨跌幅的平方在过去5天的最大值出现日期的百分位数。计算方式描述相当不直观，但是如果股票在上涨，这个分位数应当会保持不变或者维持在较大的数值。因子评估如下：





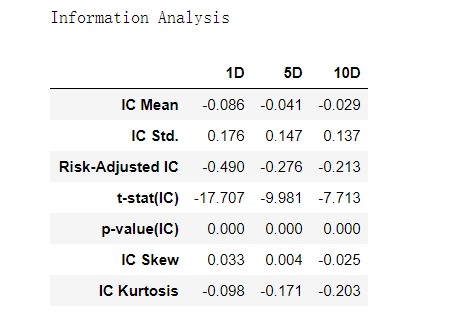


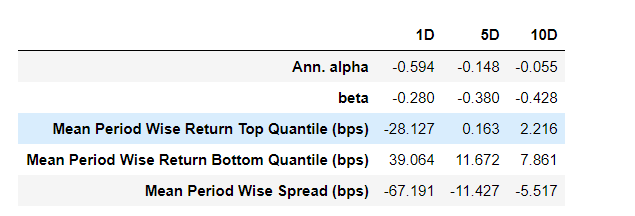


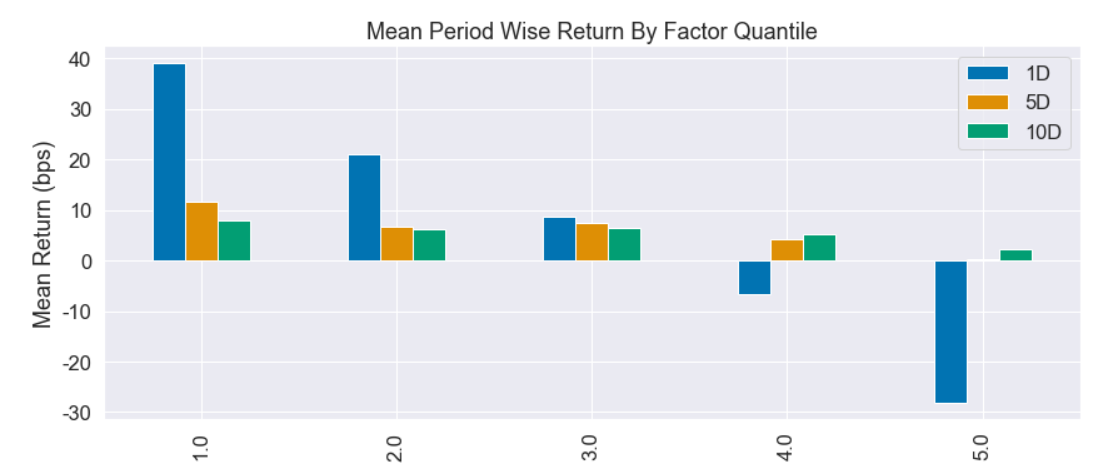
可以看到首先该因子的IC值很大，尤其是单日的持股周期时，并且根据收益图也可以看到单日周期的收益率很高。其次1D、5D、10D的t统计量也显著大于2。观察收益分析也可以看到大部分的收益属于alpha收益即非系统性风险部分。并且因子的换手率也比较高，说明这是一个相对高频变化的量。因此这个因子表现很不错。

#### Alpha2

Alpha2因子是通过6天的回溯期中收盘价相对开盘价增加的程度与成交量对数值差值的百分位数的相关系数的负数。简单来说如果这个因子越小(相当于相关系数越大)表示股价正在上涨并且成交量和每日涨幅在同步上升。因此这个因子是相关系数的负数所以，期待的结果是具有负的IC值。结果如下







#### Alpha3-20

由于每个因子都有较多图表，因此本课题对这些因子仅仅记录各个因子的1D,5D,10D的IC值。图表如下：

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | alpha001 | alpha002 | alpha003 | alpha005 | alpha006 | alpha008 | alpha009 | alpha010 | alpha011 |
| 1D | 0.270495 | -0.08602 | -0.02891 | -0.2937 | -0.03418 | -0.20278 | -0.64653 | -0.56923 | -0.02333 |
| 5D | 0.0969 | -0.04053 | 0.014608 | -0.06768 | 0.011074 | -0.07102 | -0.26868 | -0.23711 | 0.01431 |
| 10D | 0.059324 | -0.02923 | 0.035457 | -0.00656 | 0.032037 | -0.04687 | -0.18353 | -0.16203 | 0.016419 |
|  | alpha012 | alpha013 | alpha014 | alpha015 | alpha016 | alpha017 | alpha018 | alpha019 | alpha020 |
| 1D | -0.14598 | -0.13735 | -0.11365 | -0.06822 | -0.12979 | -0.56748 | -0.59062 | -0.15423 | 0.05387 |
| 5D | -0.07062 | -0.06235 | -0.02891 | -0.00891 | -0.05389 | -0.2518 | -0.23925 | -0.05535 | 0.015415 |
| 10D | -0.04712 | -0.03908 | 0.001474 | -0.0012 | -0.03613 | -0.17971 | -0.15807 | -0.03959 | 0.009935 |

如图可以看到其中IC值绝对值比较大的几个因子包括Alpha001、Alpha009、Alpha010、Alpha005、Alpha017、Alpha018。另外还有一些因子的表现不错，具有一定的实用价值。由于性能有限，暂时仅仅列出了部分因子的部分股票历史回测的部分指标，如果条件允许还可以进行更深一步的探索。

#### 因子相关性分析

通过观察诸多因子的构建逻辑可以发现，他们都是在刻画个股的价量关系的背离性质，因此可以预见的是这些因子之间会存在较强的相关性，通过相关系数的检验也证实了这一点。因此通过这一性质可以考虑进行多因子策略的构建，可以利用PCA等方法对alpha因子进行组合降维构建出合适的多因子策略。

# 总结

## 结论

可以在测试的因子中，市值因子具有一定的alpha意义，通过购买市值较小的股票可以获取一定的超额收益，并且基本面因子具有一个共性即他们的的因子换手率一般都较低。

对于技术面指标，我们利用WorldQuant发布的多个因子进行了测试，发现其因子表现远远超过了基本面指标。并且诸多因子的不过其IC值绝对值比较大，其换手率也比较高，换手率比较高意味着其获得超额收益率占主导。

此外，本课题也了解到了存在更多的Alpha因子，但是本课题构建的回测体系还是可以有效地反应这些因子在沪深300中的表现。

通过这次课题，也让我对于因子的作用、因子的评价与技术选股等方面有了初步的了解。并且自己通过对于Alphalens、因子构建程序代码的阅读也有了很大的收获。所以总体上我认为本课题还算是比较充实的。

## 不足与改进

由于本课题使用的Python用于回测评价因子的包Alphalens的限制，暂时无法将手续费率、交易费用、滑点等因素考虑其中，这一点我们仅仅通过换手率来作一个综合的主观的考虑和比较。

另外由于性能限制，为了追求程序运行速度，对于历史数据和股票池选取了一个较小的样本，对于因子的筛选也选取了一个窗口，但是本课题的程序的瓶颈在于程序运行速度，如果可以通过对现成包的源代码进行分析精简等手段可以有效地优化计算速度，也可以筛选更多的因子。

最后本课题虽然列举了个别比较有代表性的因子，但是我实际上收集了更多因子，由于性能和时间限制没有一一进行回测。但我认为这个评价因子的框架对于开发新的因子还是很有意义的。对于新的因子也可以很方便的选取不同的历史数据进行回测。

在对因子选股的过程中由于股票池较浅，且获得的公司基本面数据中行业分布较为分散，无法进行有效地行业分析。可能会使得对于部分因子存在风险过于集中。另外对于不同行业与不同股票的择时也没有实现，这在实际应用中也是需要考虑的部分，但由于这部分需要考量的较为复杂，暂时缺乏有效的理论支撑和技术支撑，留待以后了解。

本课题在历史数据窗口选择上较为狭窄，而A股市场存在明显的风格切换，此时很多因子会失效，我们的回测过程没有体现这一点。