XGBoost 算法在多因子模型中的应用

纽约大学 张嘉昕
纽约大学 袁子琪

不同的因子对应着不同的信息，如何将这些信息组合起来形成一个策略是大家一直在探究的问题。近几年，随着人工智能与机器学习的发展与普及，越来越多的研究人员开始将机器学习运用于量化金融领域。在因子挖掘领域，相比传统的多因子模型，机器学习模型在海量数据下有更好的性能，并且大多数机器学习模型都具有非线性拟合能力，可以有效利用非线性因子。机器学习模型有机会帮助我们在多因子选股方向得到新的突破。
本研究的主要目的是探究机器学习模型 XGBoost 在多因子组合方面的表现。XGBoost 是一种梯度提升算法的高效实现。由于其非线性拟合能力强、速度快等特点，成为一种常用的机器学习模型。
研究主要分为三部分：

1) 单因子的挖掘以及构建因子挖掘代码模块。
2) 利用第一部分挖掘的因子，使用 XGBoost 搭建模型，合成新的因子。
3) 测试因子表现并且对比。

第一部分 因子挖掘

1. 因子模块构建：

因子模块由于财务数据的缺失，暂时仅支持量价因子的更新和测试。未来在财务数据可得的情况下，会将财务数据也本地化，并封好 lag periods、TTM 的计算函数。

1.1 取数据模块

本篇报告的数据源是复旦大学策略大赛的比赛平台。平台财务数据质量待考，所以我们这里只运用了从 2012.12.01 至 2020.2.28 的量价数据。其中包括 close, preclose, open, volume, adj close, high, low, turnover 等。我们将数据本地化，并且构建了一个 API，使得后面的计算中取数据变得非常迅速且便捷。

1.2 写因子模块

先构建了一个因子创建的模版，将取数据的接口封入，在因子构建模块(factorZoo)中可以较为简便地构建量价因子。

1.3 因子更新模块

根据需求将因子进行月频、周频或日频更新，并存入 csv 中，便于以后的测试和策略的运用。

1.4 因子测试模块

因子测试基于 alphalens 包。以月频因子为例，每个月最后一个交易日的因子值对应下月第一个交易日的开盘价，将 DataFrame (dateTime, securityId, factorValue) 输入 alphalens 包进行因子测试。

2. 因子选择标准

2.1 IC

因子值与下一周期收益率的相关性的平均值。IC 的绝对值在 0.03 以上的因子被视为较好的因子。除此之外，information ratio 也是个很好的选股指标。Information ratio 为下一周期的收益率对该周期因子值截面回归的 beta 的平均值，当 information ratio 大于 1 时，该因子被视为有用的因子。我们这里选择 IC 作为因子的测试标准，当因子的 IC 的绝对值超过 0.03，并且因子的单调性较好时就视为有效因子。

2.2 单调性

在每一期，因子值从小到大将股票池分为 10 组，并计算下一期的收益率。将多期的分组收益率做平均，得到 10 个平均收益率，分别对应因子值从小到大的 10 组。将这 10 个平均收益率可视化后进行评估。我们希望因子单调性，尤其是多空两方的单调性较好，这样这个因子区分股票的能力越强。若因子值最大的那组对应的平均收益率较高，则可以认为该因子是个正向因子，反之，认为该因子是个负向因子。正向因子对应 IC 为正，负向因子对应 IC 为负。

2.3 其他标准

因子的筛选还有许多其他的标准，如覆盖率，IC 的 t 值，市值行业中性化后的 IC IR，相关性等。我们所选择的因子覆盖率都在 90%以上，量价因子的相关性普遍较低。

3. 因子选择及结果

因子来源：

由于暂时仅有量价数据，我们从 worldquant alpha101 着手，alpha101 中的因子都是由量价数据通过一些固定的公式来构建的。我们做了一个 alpha101 的通用的代码版本，将所有的公式和取数据的函数封装，使得实现 alpha101 中的因子变得非常容易。我们测试了 30 个左右的因子（前 20 和后 10），选择了三个表现突出的因子：alpha003 和 alpha015 以及 alpha044 输入我们的算法。其他的因子来源于 wind 上的一些研报。下面将对每个因子的构建方法进行简单的介绍。

因子定义：

因子名称	因子定义
alpha003	(-1 * correlation(rank(open), rank(volume), 10))
alpha015	(-1 * sum(rank(correlation(rank(high), rank(volume), 3)), 3))
alpha044	(-1 * correlation(high, rank(volume), 5))

ADTM_20	用开盘价的向上波动幅度和向下波动幅度的距离差值来描述人气高低的指标。（详见附录）
Ret_HL_60d	低于历史 60 日复合收益率的日收益波动率除以高于历史 60 日复合收益率的日收益率波动的对数
Turn_CoV_1M	过去一个月换手率的标准差/过去一个月换手率的均值
MA_5D	MA 简单移动平均（5 日）（该因子来自复旦大学量化比赛平台）

因子测试结果：
为了测试因子的有效性，我们将因子测试的时间段定为：2013.1.1-2019.12.31，因子频率定为月频。下表展示了这七个因子的 IC 值，更具体的因子测试 alphalens 测试结果(包括单调性图)见附录。

因子名称	IC
alpha003	0.048
alpha015	0.031
alpha044	0.046
ADTM_20	-0.050
Ret_HL_60d	0.044
Turn_CoV_1M	-0.045
MA_5D	-0.053

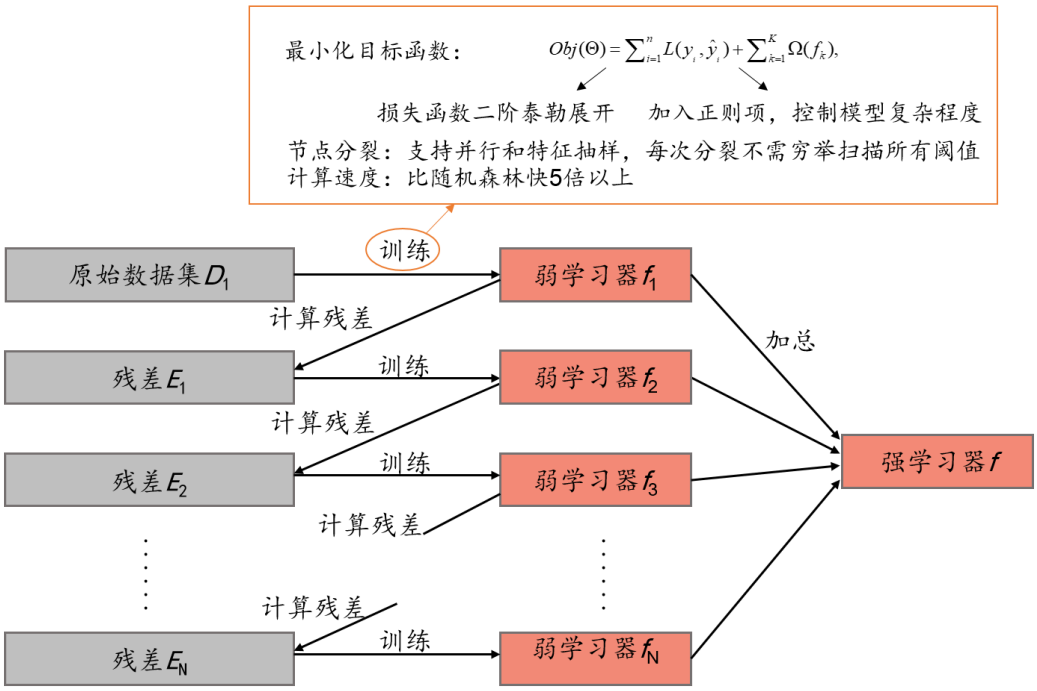
第二部分 XGBoost 算法

1. XGBoost 算法简介

XGBoost 是陈天奇等人开发的一个开源机器学习项目，它高效地实现了 GBDT(Gradient Boosting Decision Tree)算法，并且在损失函数、正则化、切分点查找和并行化设计这些方面进行了改进。相比传统的线性模型，XGBoost 由于使用决策树为基学习器，具有非线性拟合能力。
XGBoost 模型可以简单的归纳为以下几个步骤：

1) 通过不断地进行特征分裂来添加新的树。每添加一个树，就是学习一个新函数，去拟合上次预测的残差。
2) 当训练完成得到 k 棵树，模型会根据这个样本的特征，在每棵树中找到对应的叶子节点，每个叶子节点就对应一个分数。
3) 将每棵树对应的分数加起来就是该样本的预测值。

XGBoost 算法流程示意：



资料来源:XGBoost，华泰证券研究所

XGBoost 算法的基本步骤与 GBDT 类似，不同的是在损失函数的设计上，XGBoost 加入了正则项，用以控制模型复杂度，有利于防止过度拟合，并且对损失函数做了二阶泰勒展开来近似。传统的 GBDT 在每轮迭代时使用全部的数据，XGBoost 则采用了与随机森林相似的策略，支持对数据进行采样。除此之外，XGBoost 还能够自动学习出缺失值的处理策略。

2. 多因子模型构建：

上图为运用机器学习模型组合多因子的流程。提取数据与计算单因子已经在前文详细介绍,在此不再赘述。

选择特征单因子：

选取上文所述的表现较好的 4 个因子以及 alpha101 中经过因子测试 IC 值在 3%或以上的 alpha003，alpha015 以及 alpha044 三个因子。再加上基础因子 volume,将这总共 8 个单因子作为样本的原始特征。

数据预处理：

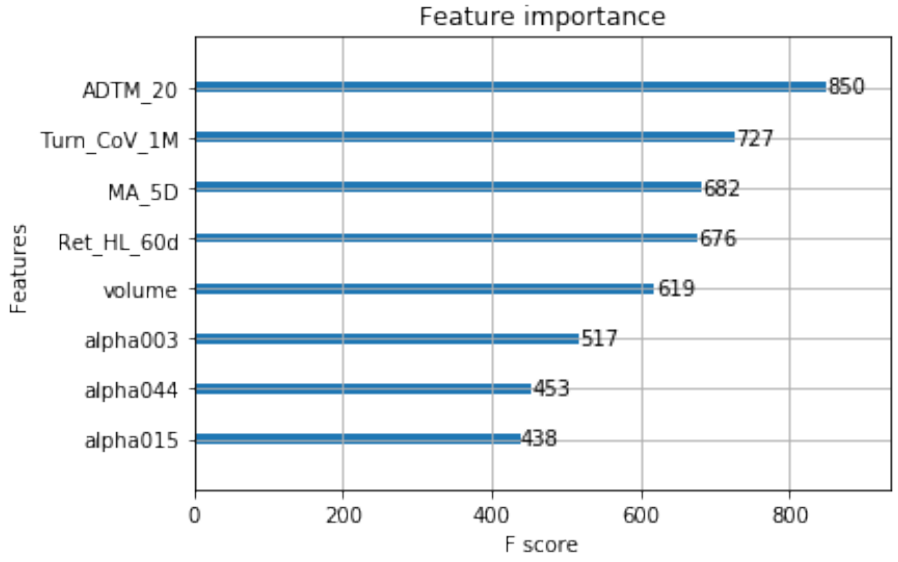
1) 首先将所有数据用日期以及股票代码为基准，合并到一起。
2) 用 close price 计算出每期下个月的个股收益作为样本的标签。
3) 去除缺失值。
4) 将数据平均分为样本内和样本外数据。样本内：2013-01-31 到 2016-12-30， 样本外：2016-12-30 到 2019-10-31。

因为我们使用 XGBoost 的树状模型，数值缩放并不会影响分裂点位置，对树模型的结构不造成影响，所以不需要将数据归一化和标准化。

样本内交叉验证：

用 Python sklearn 中的 XGBoost API 进行模型的构建。模型输入为上述 8 个因子，训练目标为下个月的个股收益。用样本内数据进行 5-fold 交叉验证。测量标准为预测值与实际标签的均方根误差(RMSE)。每一轮交叉验证取随机重复 10 次，取平均 RMSE。调整 learning_rate、max_depth 等参数，使得参数组合达到较小的 RMSE。

下图为模型中每个单因子的特征重要性：



样本外合成因子：

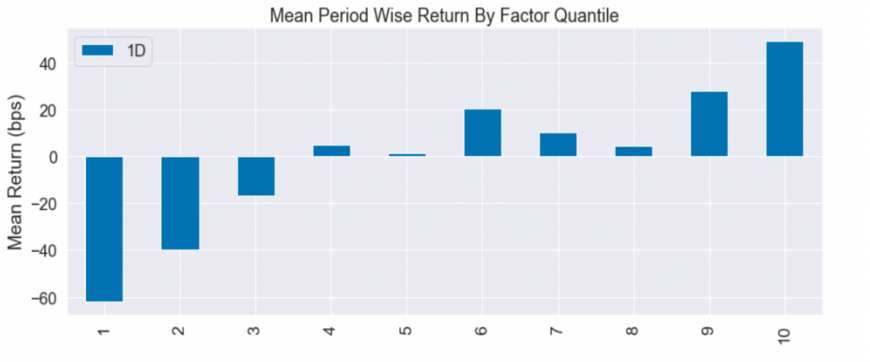
确定模型最优参数后，用样本内数据训练模型。向该模型输入样本外数据（无标签），得到样本外预测值，将预测值视作合成后的因子。

第三部分 结果及结论

我们将因子等权平均作为基准因子，来对比 XGBoost 合成因子的表现。其具体做法为，将七个因子都调成正向因子，并进行归一化，最后所得平均作为一个新的因子。由于我们需要将 XGBoost 组合出来的因子与基准进行对比，所以基准因子和 XGBoost 因子测试时间段要一致，都为 2016.12.1-2019.11.30。

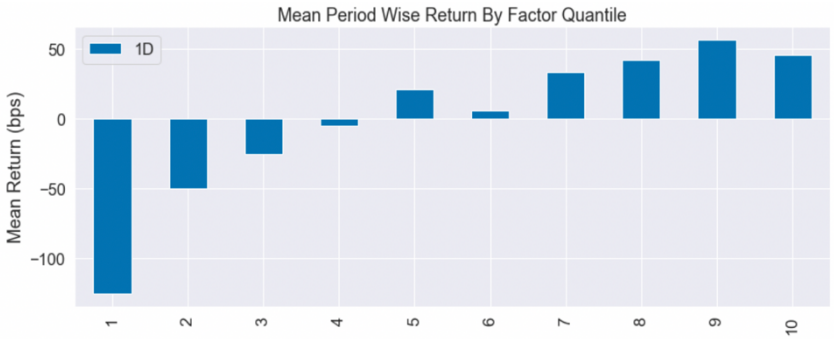
基准因子的测试结果：

IC Mean **0.047**
IC Std. 0.043
Risk-Adjusted IC 1.106
t-stat(IC) 6.544
p-value(IC) 0.000
IC Skew -0.115
IC Kurtosis 0.007



XGBoost 因子测试结果：

IC Mean **0.059**
IC Std. 0.081
Risk-Adjusted IC 0.727
t-stat(IC) 4.241
p-value(IC) 0.000
IC Skew -0.268
IC Kurtosis -0.422



我们复制了 worldquant alpha101 中的大多数因子，选取了几个 IC 较高且单调性较好的因子作为算法的 input。我们的数据源是复旦大学策略大赛的比赛平台。本篇报告验证了 XGBoost 算法组合因子的可行性。由于可获得的数据较少，且更新周频的因子较为耗时，所以在本篇报告中，仅测试月频因子。为了验证算法的可行性，我们将 2013.1.1-2019.12.31 中 2013.1.1-2016.11.31 的数据作为训练集，预测 2016.12.1-2019.11.30 期间的月频收益率作为一个新“因子”，并测试这个因子的选股能力。研究发现使用 XGBoost 模型合成的因子选股能力较强，IC 值相较于单因子和基准混合因子都有显著提高，且单调性非常好。

改进方向以及研究计划

1、因为我们现在条件有限，没有办法获取更全面的数据，所以用的数据频率较低，不能支撑我们做更加详细的计算。今后可以用更加高频的数据进行计算，并且加入行业数据，对特征数据做行业中性化处理。
2、因子模块由于财务数据的缺失，暂时仅支持量价因子的更新和测试。未来在财务数据可得的情况下，会将财务数据也本地化，并封好 lag periods、 TTM 的计算函数。
3、今后可以搭建更完善的回测系统，测试运用多因子模型策略的收益情况、最大回撤、换手率影响等。
4、金融市场是瞬息万变的，我们可以通过用滚动训练的方式使机器学习模型更加好的适应市场特征变化。

附录

ADTM_20 的定义：（来源于广发证券研报）

ADTM 是用开盘价的向上波动幅度和向下波动幅度的距离差值来描述人气高低的 指标。
DTM=IF(O<=REF(O,1),0,MAX((H-O),(O-REF(O,1))))
DBM=IF(O>=REF(O,1),0,MAX((O-L),(O-REF(O,1))))
STM=SUM(DTM,P)
SBM=SUM(DBM,P)
ADTM=IF(STM>SBM,(STM-SBM)/STM,IF(STM=SBM,0,(STM-SBM)/SBM))
如果开盘价≤昨日开盘价，DTM=0 如果开盘价>昨日开盘价，DTM=(最高价-开 盘价)和(开盘价-昨日开盘价)的较大值。

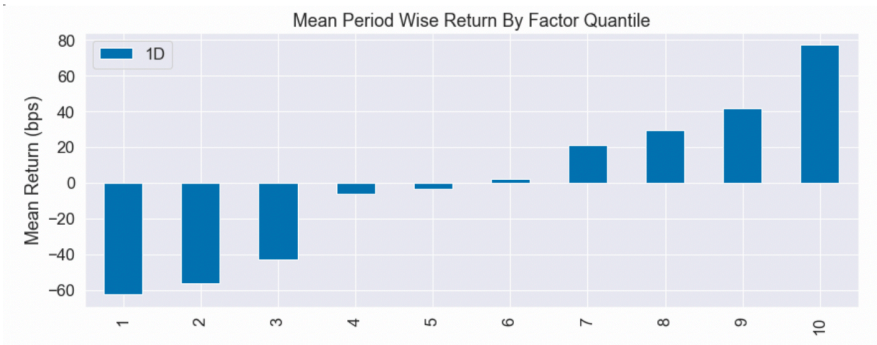
如果开盘价≥昨日开盘价，DBM=0 如果开盘价<昨日开盘价，DBM=(开盘价-最 低价)和(昨日开盘价-开盘价)的较大值。

STM=DTM 在 N 日内的和
SBM=DBM 在 N 日内的和

如果 STM>SBM,则 ADTM=(STM-SBM)/STM. 如果 STM<SBM, 则 ADTM=(STM-SBM)/SBM. 如果 STM=SBM,则 ADTM=0。

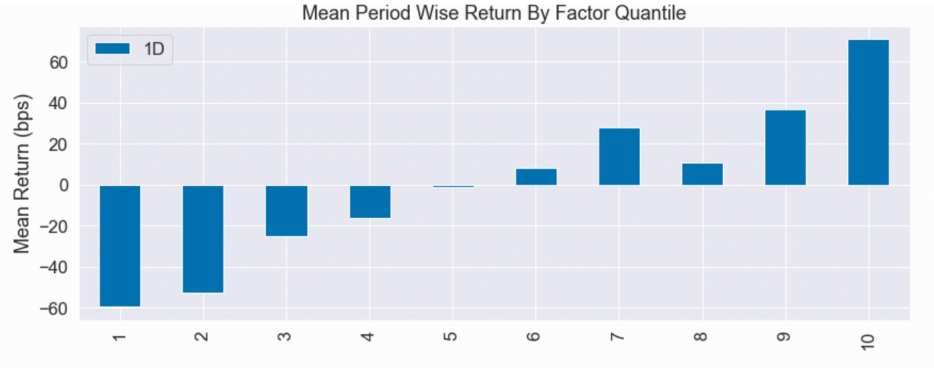
因子的详细测试结果

1. alpha003
   IC Mean 0.048
   IC Std. 0.059
   Risk-Adjusted IC 0.803
   t-stat(IC) 7.270
   p-value(IC) 0.000
   IC Skew 0.323
   IC Kurtosis -0.199
   单调性

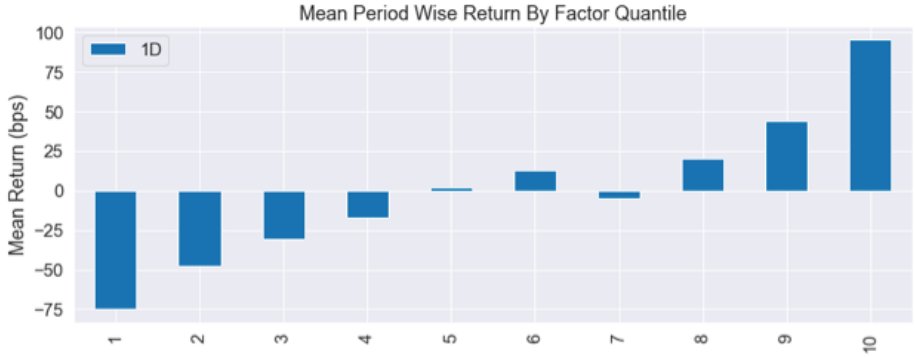


2. alpha015

IC Mean 0.031
IC Std. 0.064
Risk-Adjusted IC 0.481
t-stat(IC) 4.354
p-value(IC) 0.000
IC Skew -0.892
IC Kurtosis 2.061
单调性

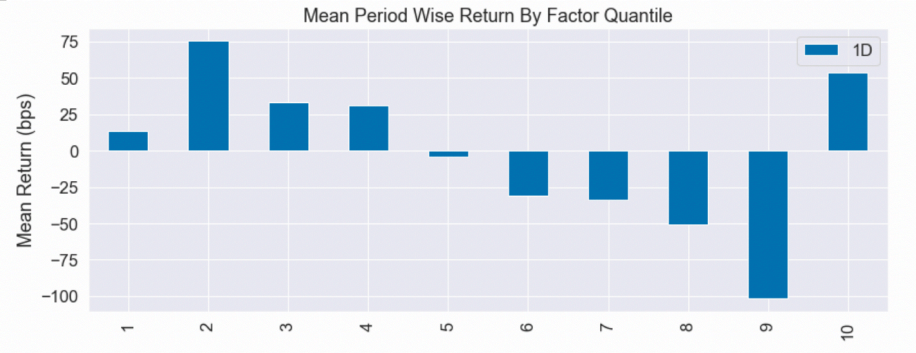


3. alpha044
IC Mean 0.046
IC Std. 0.057
Risk-Adjusted IC 0.818
t-stat(IC) 7.408
p-value(IC) 0.000
IC Skew -0.291
IC Kurtosis 0.422
单调性

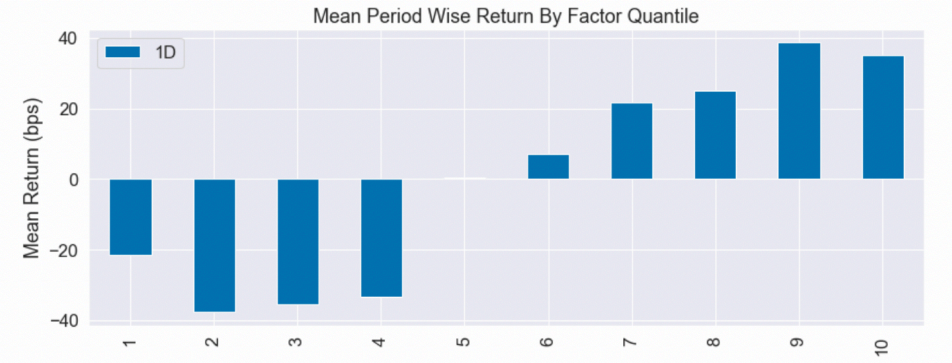


4. ADTM_20
IC Mean -0.050
IC Std. 0.120
Risk-Adjusted IC -0.413
t-stat(IC) -3.738
p-value(IC) 0.000
IC Skew -0.123

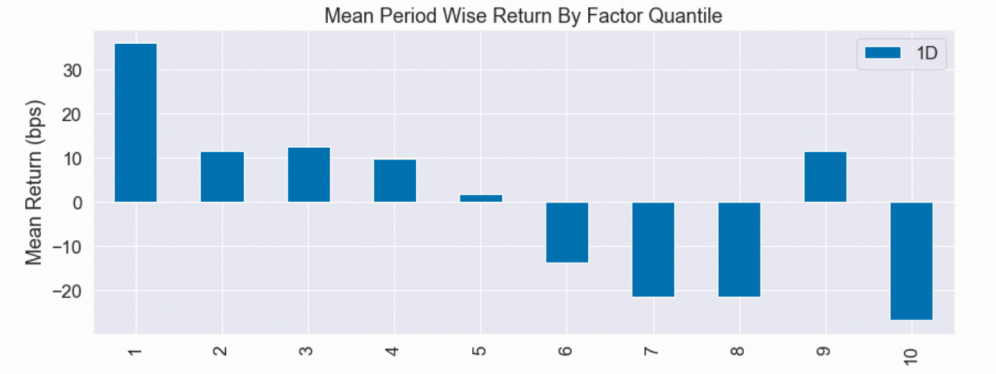
IC Kurtosis 0.501
单调性



5. Ret_HL_60d
   IC Mean 0.044
   IC Std. 0.083
   Risk-Adjusted IC 0.536
   t-stat(IC) 4.857
   p-value(IC) 0.000
   IC Skew -0.333
   IC Kurtosis 0.245
   单调性



6. Turn_CoV_1M
   IC Mean -0.045
   IC Std. 0.075
   Risk-Adjusted IC -0.605
   t-stat(IC) -5.477
   p-value(IC) 0.000
   IC Skew -0.570
   IC Kurtosis 0.253
   单调性



7. MA_5D
   IC Mean -0.053
   IC Std. 0.165
   Risk-Adjusted IC -0.319
   t-stat(IC) -2.892
   p-value(IC) 0.005
   IC Skew -0.391
   IC Kurtosis -0.653
   单调性