

因子设计与回测说明

一、数据描述

本项目主要使用了如下数据：

- 高频分钟数据

路径： `/content/drive/MyDrive/factor/data/data_5m/`

内容：包含每只股票每5分钟的成交数据，包括 `datetime`、`code`（股票代码）、`close`（收盘价）、`volume`（成交量）、`money`（成交额）等字段。

用途：用于计算每只股票每日的收盘价与成交量相关系数（`close_volume_corr`），反映日内量价关系。

- 日频数据

路径： `/content/drive/MyDrive/factor/data/data_daily/`

内容：包含每只股票每日的基础行情数据，如开盘价、收盘价、最高价、最低价、成交量等。

用途：与分钟数据生成的因子合并，作为后续因子计算的基础。

- 收益数据

路径： `/content/drive/MyDrive/factor/data/data_daily_pkl/ret.pkl`

内容：每只股票每日收益率。

用途：用于因子的去收益化处理（截面回归残差）。

- 回测辅助数据

- 交易日历： `/content/drive/MyDrive/factor/data/date.pkl`

- 未来1日收益： `/content/drive/MyDrive/factor/data/data_ret1`

- 涨跌停信息： `/content/drive/MyDrive/factor/data/data_ud`

- 因子输出路径

- 量价相关均值因

- 子： `/content/drive/MyDrive/factor/feature/factor_1_8_pv_corr_avg`

- 量价相关波动因

- 子: `/content/drive/MyDrive/factor/feature/factor_1_8_pv_corr_std`
- 去收益化因
子: `/content/drive/MyDrive/factor/feature/factor_1_8_pv_corr_deRet20`
- 趋势因
子: `/content/drive/MyDrive/factor/feature/factor_1_8_pv_corr_trend`
- 复合因子: `/content/drive/MyDrive/factor/feature/factor_1_8_cpv`

二、设计思路

本课设围绕“收盘价-成交量相关性”构建量价相关因子，旨在挖掘高频数据中的价格与成交量关系对股票收益的预测能力。具体思路如下：

1. 分钟级数据处理

- 以每只股票为单位，计算其在每个交易日内收盘价与成交量的相关系数（`close_volume_corr`），反映日内量价关系的强弱。

2. 日频因子生成

- 对每只股票的 `close_volume_corr` 进行20日滚动均值（`PV_corr_avg`）和标准差（`PV_corr_std`）处理，分别反映量价相关性的趋势和波动性。

3. 去收益化处理

- 为消除收益本身对因子的影响，对 `PV_corr_avg` 和 `PV_corr_std` 进行20日均值收益的截面回归，提取残差部分（`deRet20`），并做标准化处理。

4. 趋势因子

- 对 `PV_corr_avg` 进行20日滚动时序回归，提取时间趋势项（`pv_corr_trend`），衡量量价相关性的变化趋势。

5. 复合因子

- 将去收益化后的均值、标准差因子与趋势因子进行加权组合，得到最终复合因子（`cpv`）。

三、改进方式

• 高频数据利用

传统因子多基于日频数据，本项目充分利用5分钟级别高频数据，捕捉更细粒度的市场微观结构信息。

• 多维度量价关系刻画

不仅关注量价相关性的均值，还引入波动性（标准差）和趋势（回归斜率）等多

通过截面回归剔除收益影响，提升因子的纯度和泛化能力，并对因子做横截面标准化，消除极端值影响。

• 去收益化与标准化

通过截面回归剔除收益影响，提升因子的纯度和泛化能力，并对因子做横截面标准化，消除极端值影响。

• 并行与分布式计算

利用 multiprocessing 和 ray 等工具加速大规模数据处理，显著提升因子计算效率。

四、回测表现

• 单因子表现

- PV_corr_avg 和 PV_corr_std 因子在2021年全年的回测中，分组收益呈现一定的单调性，部分分组收益显著高于基准，表明量价相关性对未来收益具有一定的预测能力。

• 去收益化因子

- 去收益化后的因子 (deRet20) 表现更为稳健，分组间收益差异更加明显，极端分组的超额收益提升。

• 趋势因子

- 趋势因子 (pv_corr_trend) 能够捕捉到量价关系变化带来的收益机会，部分阶段表现优异。

• 复合因子 (cpv)

- 复合因子 cpv 综合了多维度信息，回测结果显示分组收益曲线更为平滑，极端分组的年化收益和夏普比率均有提升，具备较好的实用价值。

五、结论

本项目通过高频量价相关性因子的系统设计与改进，验证了量价关系在A股市场中的有效性。因子在回测中表现出较好的分组收益和风险调整后收益，具备一定的选股和风险管理价值。未来可进一步结合行业、风格等信息，提升因子的稳定性和泛化能力。

注：

详细回测结果及分组收益曲线可见于课设代码输出及相关图表。