项目理论介绍

因子选股

因子选股是投资领域中一种基于因子模型的投资策略，它基于对资产收益率背后的系统性因素进行分析与解释。因子选股的基本理论包含以下几个重要概念：

1. **因子是资产收益率的解释变量：**
   因子是一组能够解释资产收益率变动的变量，通常是与资产特定特征相关的经济指标或其他量化的因素。这些因子可以包括公司财务数据、宏观经济指标、市场波动性等。通过识别并利用这些因子，投资者可以更好地理解资产收益率的来源和波动。
2. **因子收益率是对众多资产共同暴露的系统性风险的风险补偿：**
   在因子模型中，资产的收益率被分解为受到系统性风险和非系统性风险的影响。因子收益率代表了众多资产共同暴露的系统性风险，这种风险通常是无法通过分散化投资来消除的。因此，投资者对这种系统性风险的承受被认为是一种风险补偿，即因子收益。
3. **通过选择适当的因子来进行投资组合的创建，从而获取超额收益：**
   投资者可以通过选择和配置不同的因子，构建一个具有较低系统性风险的投资组合。通过将因子暴露纳入投资决策过程，投资者可以更有效地管理风险并寻求获取超额收益。这就是所谓的因子投资策略，其中投资组合的构建基于对因子收益率的预测和管理。

因子选股的基本理论背后的关键假设是，资产的收益率不仅仅取决于特定资产的特征，还受到一系列系统性因素的影响。通过识别和利用这些因子，投资者可以提高他们的投资组合的效率，更好地理解市场的变化，并在不同市场环境中取得更为稳定的投资表现。

经典风格因子

经典风格因子是投资组合中常用的一组因子，它们有助于解释资产的收益率变动。包括规模因子、盈利因子、价值因子、动量因子、换手率因子：

1. **规模因子（Size Factor）：**
   规模因子关注的是公司的市值大小。通常，投资者认为小市值公司相对于大市值公司更具有成长潜力和风险。规模因子捕捉到了这种不同市值公司之间的收益率差异。在投资策略中，规模因子通常表现为投资者倾向于投资小型公司以获取潜在的高收益。
2. **盈利因子（Profitability Factor）：**
   盈利因子关注公司的盈利能力。公司的盈利水平被认为是一个重要的投资考量，因此盈利因子捕捉到了高盈利公司和低盈利公司之间的差异。投资者可能更倾向于投资盈利能力较好的公司，因为它们具有更强的财务健康和稳定性。
3. **价值因子（Value Factor）：**
   价值因子考察了公司股票的估值。通常，价值因子将公司的市值与其财务指标（如盈利、现金流等）进行比较，以确定股票是否被低估或高估。价值因子的投资策略通常包括寻找低估值的股票，认为它们具有较大的上涨潜力。
4. **动量因子（Momentum Factor）：**
   动量因子关注资产价格的趋势。该因子认为在一段时间内表现良好的资产将在未来继续表现良好，而在一段时间内表现差的资产可能会继续表现差。动量因子的投资策略通常包括追逐走势较好的资产，以期望能够利用其继续上涨的趋势。

5. **换手率因子（Turnover Factor）：**
   换手率因子考察了资产的交易频率。高换手率通常反映了更频繁的买卖活动，而低换手率则反映了相对较稳定的投资组合。这个因子关注投资者对于资产的持有期限，以及他们对于主动管理投资组合的态度。

排序法

排序法是因子选股中最经典的策略之一，它通过对股票池中的股票进行排序、分组、做多和做空操作，构建一个多空对冲的投资组合。排序法的基本步骤如下：

1. **确定股票池：**
   首先，投资者需要确定一个特定的股票池，这可以是整个市场或特定行业、板块。股票池中包含了投资者希望分析和选择的股票。
2. **排序变量的取值：**
   投资者选择一个排序变量，通常是某个因子，如价值、动量、盈利等。对于该因子，股票池中的全部股票在截面上按照排序变量的取值高低进行排序。如果是反向因子，排序则从小到大。
3. **分组：**
   排序后，将全部股票分为若干组，通常分为L组（例如，L=10）。这样，每组中的股票都具有相似的排序变量取值。分组后，投资者就能够根据这一排序变量的强弱来进行投资组合的构建。
4. **做多和做空：**
   在排序法中，投资者选择做多排名最高的第一组内的股票，同时做空排名最低的最后一组内的股票。这意味着投资者构建了一个多头头寸和一个空头头寸，形成了一个多空对冲的投资组合。
5. **加权方式：**
   多空两个投资组合通常需要按照一定的方式加权，以确保它们在构建价差组合时具有相同的权重。市值加权和等权是两种常见的加权方式，投资者根据自身的偏好和策略选择适当的加权方式。
6. **定期调仓：**
   排序法是一种定期调仓的策略，每个投资周期结束后，重新构建新的因子模拟投资组合。在调仓时，投资者重新排序、分组、选择做多和做空的股票，以适应市场的变化。通过定期调仓，投资者有机会捕捉到因子收益率的变动，从而优化投资组合的表现。
7. **评价指标：**

   排序法的有效性通过多个指标检验，包括因子序列的统计性检验、投资组合收益率单调性、信息系数IC、信息比率IR以及夏普率。统计性检验验证因子在历史数据上的表现，而IC、IR和夏普率则衡量了因子对股票收益的解释能力、稳定性和风险调整后的表现，进一步评估了排序法在实际投资中的可行性。

项目功能

本项目实现了以下功能：

1. 实现排序法选股
2. 同时计算多个因子在不同股票池中回测结果
3. 分析投资组合的行业分布结果
4. 计算相关回测指标和可视化呈现

项目代码结构

backtest_framework

multi_factors_backtest

封装所有功能函数

generate_single_factor_pool_object()
使用传入实例的a个因子和b个股票池，生成a×b个单因子实例并完成所有实例的数据清洗

Parameters

factor_data : pd.DataFrame - MultiIndex (date, asset)
　　因子数据，每一个因子为一个字段，字段名是因子名。
price_data : pd.DataFrame - MultiIndex (date, asset)
　　价格数据，只有唯一一个price字段。
benchmark_data : pd.DataFrame - MultiIndex (date, asset)
　　基准数据，只有唯一一个benchmark_price字段。
pool_data : Dict[str, pd.DataFrame - MultiIndex (date, asset)]
　　选股池数据，每一个选股池是一个DataFrame且无字段。
start_date : str, optional
　　开始日期，格式如：'20130101'，默认为None。
end_date : str, optional
　　结束日期，格式如：'20221231'。默认为None。
is_daily_factor : bool or Dict[str, bool], optional
　　该因子是否为日频因子，默认为True。若为bool，则所有因子共享该标记；若为字典，则各因子可分别设置。
group_data : pd.DataFrame - MultiIndex (date, asset), optional
　　DataFrame的字段是所属行业。
direction : int or Dict[str, int], optional
　　因子方向，默认为1。若为int，则所有因子共享该方向；若为字典，则各因子可分别设置。
quantiles : int, optional
　　回测分组数，默认为5

get_backtest
获取指定因子和股票池的backtest对象

single_factor_backtest

封装所有功能函数

Parameters

factor_data : pd.DataFrame - MultiIndex (date, asset)
　　因子数据。
price_data : pd.DataFrame - MultiIndex (date, asset)
　　价格数据。
benchmark_data : pd.DataFrame - Index (date)
　　基准数据。
pool_data : pd.DataFrame - MultiIndex (date, asset)
　　池子数据。
factor_name : str
　　因子名称。
pool_name : str
　　池子名称。
start_date : str or None, optional
　　开始日期，默认为None。
end_date : str or None, optional
　　结束日期，默认为None。
is_daily_factor : bool, optional
　　是否是日度因子，默认为True。
group_data : pd.DataFrame, optional
　　分组数据，默认为None。
direction : int, optional
　　因子方向，默认为1。

赋值Backtest对象的以下属性：
　　因子频率清洗后数据factor_freq_clean_data
　　日频清洗后数据daily_freq_clean_data
　　带有所属行业的因子频率清洗后数据grouped_factor_freq_clean_data

generate_clean_data()
内部调用
- get_factor_freq_clean_data()
- get_daily_freq_clean_data()
- get_grouped_factor_freq_clean_data()

get_factor_coverage()
plot_factor_coverage()
获取、展示因子覆盖率

analyse_factor_descriptive_statistics()
按quantile对因子进行描述性统计

plot_factor_distribution()
绘制因子分布直方图和密度图

功能函数

IC相关
- get_ic() / analyse_ic() / plot_ic()：计算、分析、展示IC
- get_quantile_ic() / analyse_quantile_ic() / plot_quantile_ic()：计算、分析、展示RankIC
- get_grouped_ic() / analyse_grouped_ic() / plot_grouped_ic()：计算、分析、展示行业IC
- analyse_ic_decay() / plot_ic_dacay()：分析、展示IC衰退

get_factor_autocorrelation()
analyse_factor_autocorrelation()
计算、分析因子自相关

get_factor_turnover()
analyse_factor_turnover()
计算、分析换手率

analyse_factor_group_distribution()
analyse_factor_group_distribution_topN_per_year()
plot_factor_group_distribution()
分析、展示行业分布

get_quantile_return_data()
get_benchmark_return_array()
get_quantile_return_array()
get_net_value_array()
get_single_net_value_array()
计算净值

analyse_return_array()
analyse_return_briefly()
分析收益

plot_annual_return_heatmap()
plot_quantile_annualized_return()
plot_quantile_accumulated_net_value()
plot_long_short_accumulated_net_value()
展示结果

结果展示

数据读取

1. 价格数据：每只股票当日收盘价。
2. 股票池：上市超过一年的沪深A股，剔除ST股、停牌股。
3. 基准收益数据：中证500。
4. 行业数据：每只股票分属行业。

```
df_factor = pd.read_parquet('../Data_test/factor.parquet')
df_price = pd.read_parquet('../Data_test/price.parquet')
df_pool = pd.read_pickle('../Data_test/pool.pkl')
df_benchmark = pd.read_pickle('../Data_test/benchmark.pkl')
df_group = pd.read_pickle('../Data_test/group.pkl')
```

本项目包含两个因子，分别是动量因子和换手率因子。因子计算公式如下图所示。

$$\text{动量因子} = \frac{t-1\text{月末的收盘价}}{t-12\text{月末的收盘价}} - 1$$

$$\text{换手率因子} = \frac{\text{过去20个交易日的平均换手率}}{\text{过去240个交易日的平均换手率}}$$

基本设定

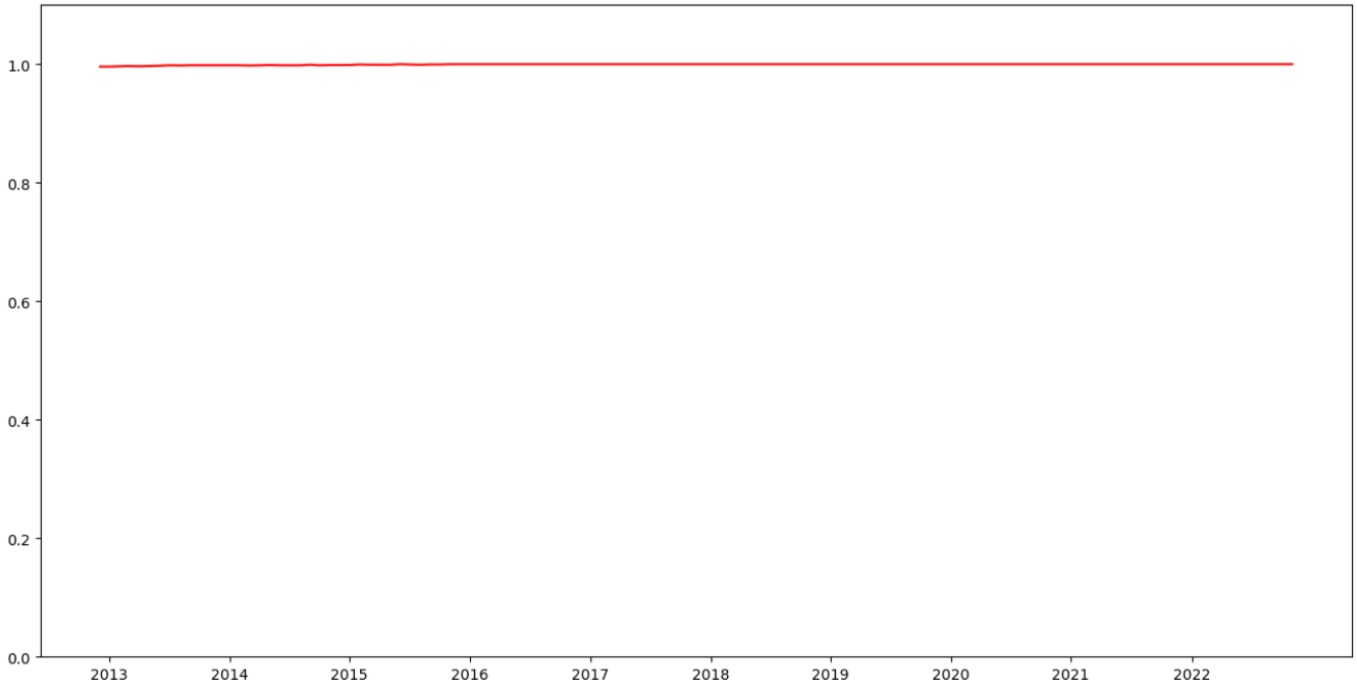
调仓频率为月度调仓，设定每个月第一个交易日开盘为调仓期。
动量因子为正向因子，换手率为负向因子。
根据因子的数值进行排序将股票池内股票分为十组。

```
Backtest = MultiFactor_MultiPool_BackTest(
    factor_data=df_factor,
    price_data=df_price,
    benchmark_data=df_benchmark,
    pool_data=df_pool,
    start_date='20130101',
    end_date='20221231',
    is_daily_factor= False,
    group_data=df_group,
    direction= {'momentum240_20':1,'turnover240_20':-1},
    quantiles=10)
Backtest.generate_single_factor_pool_object()
```

因子覆盖率

展示换手率因子覆盖率，因子覆盖率为有因子的股票数占当期股票池股票数的比例。

```
Backtest.plot_factor_coverage(factor_name= 'turnover240_20',pool_name = '000905')
```



因子描述性统计

按quantile对换手率因子进行描述性统计。

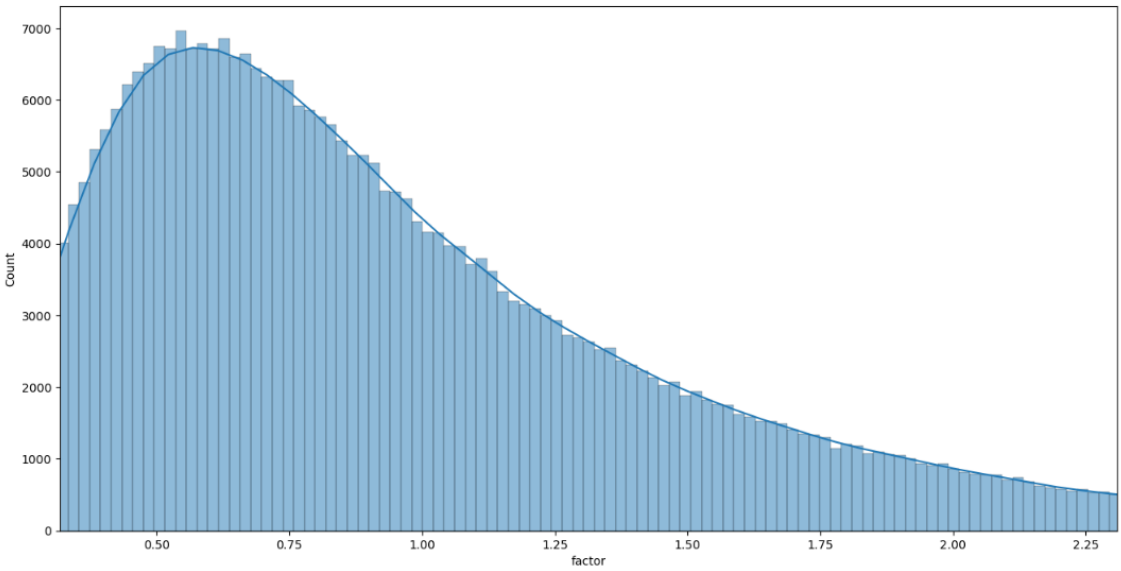
```
Backtest.analyse_factor_descriptive_statistics(factor_name= 'turnover240_20',pool_name = '000905')
```

	样本量	均值	标准差	偏度	峰度	最小值	p10	p25	p50	p75	p90	最大值	中位数绝对偏差
factor_quantile													
1	34946	0.366056	0.165054	1.615242	3.587144	0.010248	0.206290	0.262519	0.330032	0.425142	0.572989	1.145442	0.077324
2	34883	0.526297	0.202862	1.754248	3.309621	0.259396	0.343042	0.390575	0.466486	0.595134	0.789398	1.395326	0.090737
3	34870	0.627698	0.230345	1.679410	2.981284	0.315765	0.416848	0.471477	0.557939	0.711349	0.928416	1.592097	0.106430
4	34890	0.720563	0.252940	1.605365	2.700542	0.367832	0.483878	0.547902	0.645807	0.814907	1.061082	1.797334	0.119082
5	34885	0.815712	0.273187	1.541691	2.545519	0.415034	0.554430	0.630111	0.737478	0.923721	1.186413	1.993435	0.130717
6	34861	0.921712	0.292403	1.463766	2.372619	0.473150	0.636165	0.724047	0.840773	1.046422	1.332269	2.194239	0.144085
7	34875	1.048602	0.313012	1.359597	2.103728	0.532401	0.732998	0.834402	0.966898	1.189257	1.501631	2.422918	0.160850
8	34885	1.221057	0.338347	1.204811	1.753212	0.610409	0.862178	0.983131	1.140819	1.386017	1.720316	2.706485	0.186924
9	34868	1.503555	0.385231	0.943944	1.108199	0.745804	1.070763	1.226068	1.431517	1.712825	2.048091	3.202208	0.234463
10	34932	2.416905	0.895418	1.623359	4.023436	0.985410	1.523740	1.799409	2.217549	2.788360	3.554927	9.268274	0.472960
总体	348895	1.016888	0.685883	2.229547	8.508015	0.010248	0.392644	0.558412	0.836354	1.270780	1.851908	9.268274	0.326266

因子分布直方图和密度图

绘制换手率因子的分布直方图和密度图。

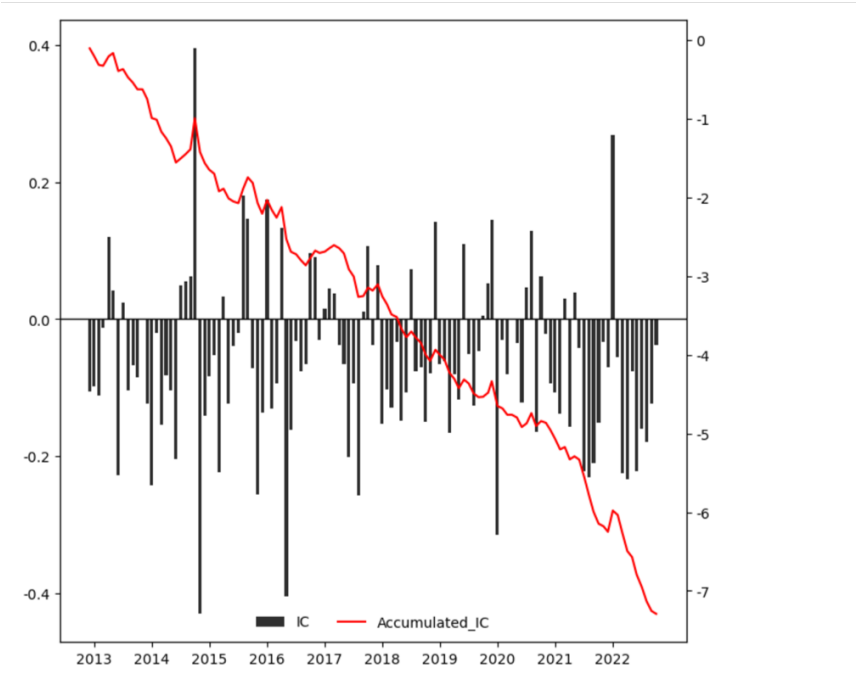
```
Backtest.plot_factor_distribution(factor_name='turnover240_20',pool_name = '000905')
```



IC和累计IC

展示换手率因子各个月的IC（每个周期下,因子值与下周期收益率的相关系数），可以看到IC值大多为负，证明换手率因子和收益率有较为显著的负相关性。

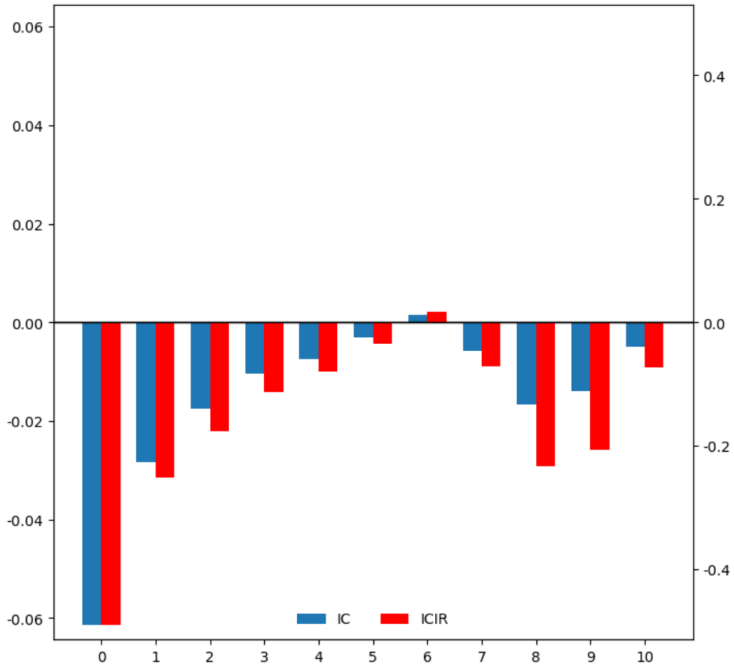
```
Backtest.plot_ic(factor_name='turnover240_20',pool_name='000905',bar_figure=True)
```



IC衰减柱形图

根据换手率因子的IC及ICIR（IC均值除以IC标准差再年化）绘制因子IC衰减柱形图。

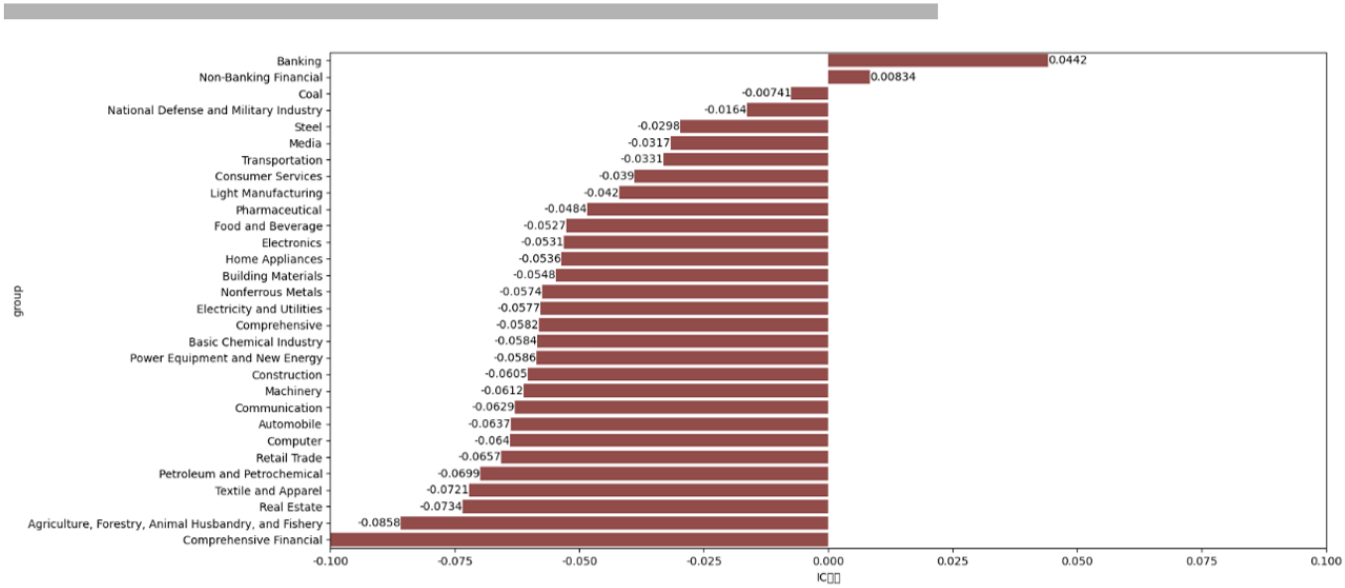
```
Backtest.plot_ic_dacay(factor_name='turnover240_20',pool_name = '000905')
```



分行业IC

分行业展示IC，可以看到换手率因子在综合金融和农林牧渔两个行业最为显著，而在银行业和非银行金融业的效果较差。

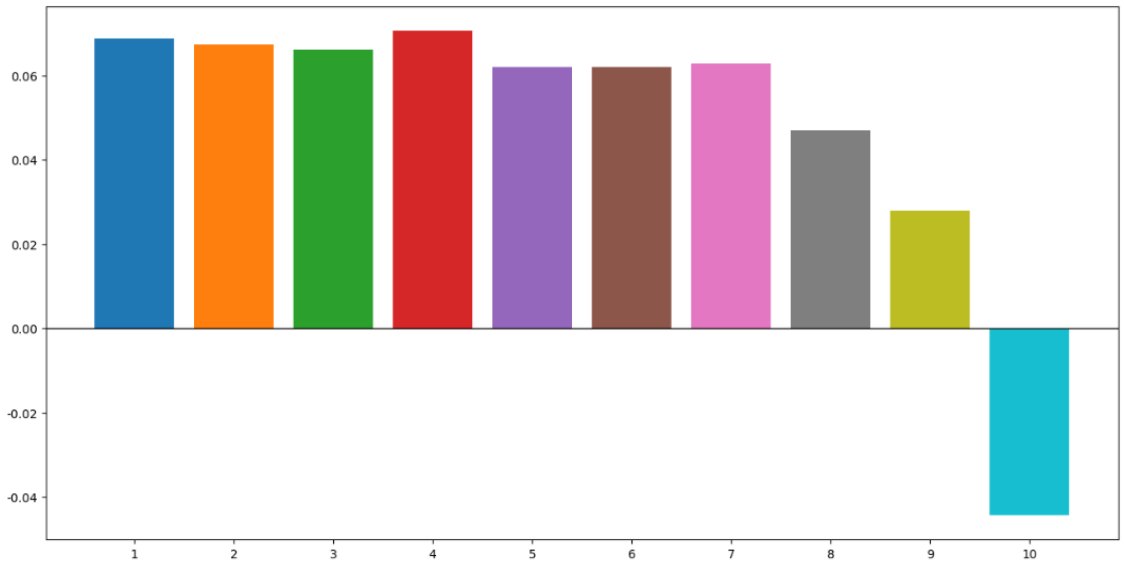
```
Backtest.plot_ic(factor_name='turnover240_20',pool_name='000905',ic_type='grouped_ic',
bar_figure=True)
```



不同分组的年化收益柱形图

展示所分十个组的超额年化收益率，可以看到换手率因子最小的一组也就是负相关最明显的一组超额年化收益率最高。

```
Backtest.plot_quantile_annualized_return(factor_name='turnover240_20',pool_name='000905')
```



因子分组收益情况的详细分析

展示因子各分组、基准、多空收益评估。

```
Backtest.analyse_return_array(factor_name='turnover240_20',pool_name='000905')
```

	年化收益	年化波动率	夏普比率	最大回撤	卡玛比率	超额年化收益	超额年化波动率	信息比率	超额最大回撤	超额收益卡玛比率	相对基准胜率	相对整体胜率	盈亏比
factor_quantile													
1	0.109201	0.283250	0.332574	-0.631044	0.173049	0.068707	0.162225	0.423531	-0.208598	0.329376	0.621849	0.220339	1.125728
2	0.106670	0.287056	0.319344	-0.604559	0.176442	0.067324	0.160078	0.420573	-0.212415	0.316948	0.605042	0.127119	1.224322
3	0.104852	0.289872	0.309969	-0.573998	0.182669	0.066106	0.160436	0.412038	-0.211740	0.312202	0.563025	0.059322	1.485806
4	0.108793	0.293312	0.319772	-0.539899	0.201506	0.070510	0.161544	0.436472	-0.183981	0.383245	0.605042	0.076271	1.374553
5	0.099964	0.292660	0.290316	-0.540660	0.184893	0.061928	0.160756	0.385228	-0.173013	0.357936	0.596639	0.042373	1.304829
6	0.099150	0.296084	0.284208	-0.548605	0.180730	0.061853	0.160784	0.384693	-0.159130	0.388693	0.588235	0.050847	1.365925
7	0.099231	0.303269	0.277742	-0.561567	0.176703	0.062806	0.166766	0.376612	-0.177076	0.354683	0.638655	0.084746	1.157785
8	0.081984	0.308447	0.217167	-0.580538	0.141222	0.046952	0.169229	0.277446	-0.176368	0.266216	0.546218	0.135593	1.392158
9	0.061153	0.318534	0.144892	-0.650143	0.094061	0.027993	0.177504	0.157703	-0.180209	0.155336	0.546218	0.076271	1.058789
10	-0.016007	0.344610	-0.089976	-0.792381	-0.020201	-0.044358	0.203340	-0.218145	-0.597163	-0.074280	0.411765	0.127119	0.922060
benchmark	0.035140	0.250321	0.080455	-0.651957	0.053899	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
long_short	0.104560	0.148117	0.604654	-0.299726	0.348852	0.104560	0.148117	0.705925	-0.299726	0.348852	0.697479	NaN	1.023619

净值曲线

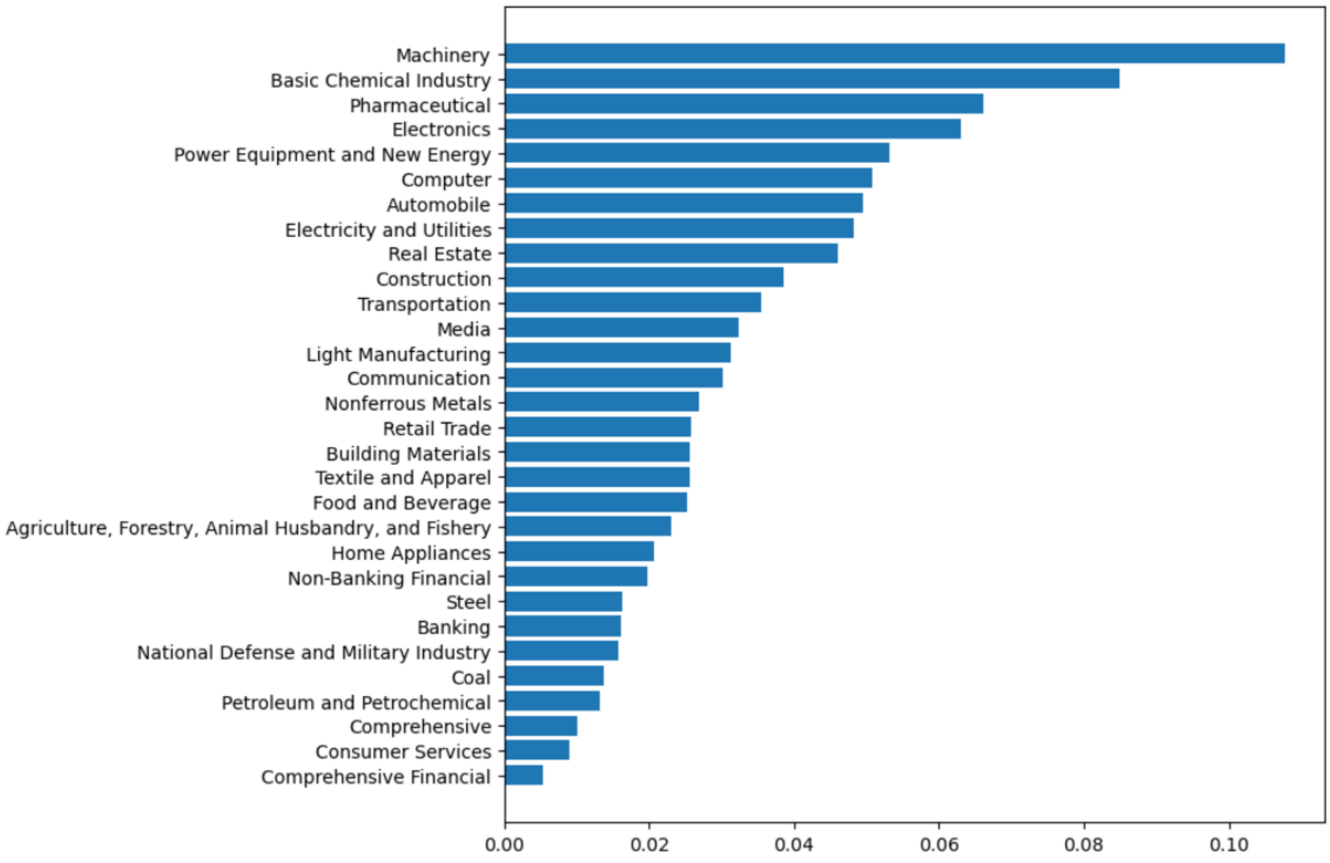
展示多头组和空头组的净值曲线，蓝色线为空头组，黄色线为空头组，绿色线为同时做多多头并做空空头。可以看到多头组合的收益要显著高于基准收益。

```
Backtest.plot_accumulated_net_value(factor_name='turnover240_20',pool_name='000905',plot_type='long_short')
```

多头组合的行业分布图

展示多头组合，也就是第一组中各个行业的占比，可以看到机械行业和化工行业所占比重最大。

```
Backtest.plot_factor_group_distribution(factor_name='turnover240_20',pool_name='000905')
```


Machinery
Basic Chemical Industry
Pharmaceutical
Electronics
Power Equipment and New Energy
Computer
Automobile
Electricity and Utilities
Real Estate
Construction
Transportation
Media
Light Manufacturing
Communication
Nonferrous Metals
Retail Trade
Building Materials
Textile and Apparel
Food and Beverage
Agriculture, Forestry, Animal Husbandry, and Fishery
Home Appliances
Non-Banking Financial
Steel
Banking
National Defense and Military Industry
Coal
Petroleum and Petrochemical
Comprehensive
Consumer Services
Comprehensive Financial
0.00
0.02
0.04
0.06
0.08
0.10