单因子测试（上）——因子中性化

原创 量化小白H 量化小白上分记 2018-12-23

收录于话题
#多因子模型
12个 >

之前做了很多因子测试的工作，但一直没有总结，感觉很凌乱，决定花时间把这部分东西写一写，温故知新，也为后续学习打基础。首先写一下单因子测试部分，分三篇，数据预处理一篇， 回归法一篇，分层测试法一篇。本篇首先说明多因子模型是什么，随后着重于单因子测试流程及数据预处理的细节，附代码。

01 因子模型概述

均值方差模型

马科维茨的均值方差组合模型是金融学中最著名的投资组合模型，模型想回答的问题是，当投资者需要进行投资决策时，对于市场上存在的证券，他该如何在这些证券上分配资金，简要叙述如下：

设 $S_i(i=1,2,...,n)$ 为市场上的n个证券，各证券期望收益率分别为 $r=(r_1,r_2,...,r_n)$ ，投资者在各证券上的投资权重为 $\omega=(\omega_1,\omega_2,...,\omega_n)^T$ ，收益率的协方差矩阵为 $\Sigma=(\sigma_{ij})_{n\times n}$ 。则组合的期望收益率为

$$r=\omega_1 r_1+\omega_2 r_2+...+\omega_n r_n=\omega^T r$$

各资产的风险可以用资产收益的标准差衡量，则组合的风险可以表示为

$$\sigma=\omega^T\Sigma\omega$$

从而投资者的最优决策可以为在一定风险水平下，最大化期望收益，或者在一定的期望收益水平下， 最小化风险,即

$$s.t.\begin{cases}\min\ \sigma=\omega^T\sum\omega\\ \omega^T r=r_p\\ \omega_1+\omega_2+...+\omega_n=1\end{cases}$$

理论上来说，这个最优化问题是有解析解的，但实际应用中存在很多问题。

首先，实际情况下，预期收益率是不可知的，因此只能采用历史收益率代替期望收益率，每一期需要进行投资时，利用上一期的各证券收益率估计组合风险，求解优化问题得到最优组合的权重，进行当期投资，这样做的误差很大，因为**用历史数据直接求协方差阵作为未来组合风险的估计量很不准确**。

其次，模型中投资组合的风险计算需要估计组合中每个股票的波动率和两两相关系数，假设股票个数为N，那么都估计的参数个数为 $\frac{N(N-1)}{2}$ ，A股市场有超过3000只股票，如果用这种方法去算，**需要估计的参数非常多，计算速度很慢，最终结果精度很低**。

因此需要新的方法来优化这一模型。

结构化风险因子模型

对于均值方差模型的优化有多种方法，最广为人知的是结构化风险因子模型，简称多因子模型，**多因子模型利用一组共同因子和一个特质因子解释各股票收益率的波动，共同因子对各个股票都有影响，特质因子只对特定股票有影响。**

多因子模型将因子收益率分解为各因子收益率的线性组合：

$$r_j=x_1f_1+x_2f_2+...+x_Kf_K+u_j$$

其中r是股票j的收益率,u是股票j的特质因子收益率， $f_i,i=1,2,...,K$ 是K个共同因子的因子收益率， $x_i,i=1,2,...,K$ 是各共同因子在股票j上的因子暴露（因子值）。此外,因子值为第t期时，收益率值为t+1期，通过因子当期值，预测因子下一期的收益率。将上述因子收益率分解式带入均值方差模型中，可以得到，组合收益率为

$$r = \sum_{i=1}^{n} \omega_i \left(\sum_{j=1}^{K} x_{ij} f_{ij} + u_j \right)$$

假设各特质因子收益率不相关，特质因子收益率与共同因子收益率也不相关，组合协方差可以表示为

$$\sigma = \sqrt{\omega^T \left(X F X^T + \Delta \right) \omega}$$

其中，X为n只股票在K个因子上的因子暴露矩阵（因子载荷阵），F为共同因子收益率的协方差阵，\Delta为特质因子收益率协方差阵，在上述假设下，特质因子收益率协方差阵为对角阵。
综上，就可以把对股票收益率协方差阵的计算转化到了对因子协方差阵的计算上。

为什么要做单因子测试

通过前面的分析，我们建立了从因子协方差阵到组合协方差阵的模型，只要选取的因子数目较小，就可以大大减少计算次数，提高计算速度，但并不是说随便选几个因子，做一个估计就可以效果很好。要提高计算精度，一方面在于对协方差阵的估计方法（这个后面学会了再写），另一方面在于因子的选择，garbage in ,garbage out，必须要选择可以准确刻画股票收益率的因子。因此，我们需要一套方法来评价因子，这就是做单因子测试的原因。

02 单因子测试方法综述

什么样的因子是好因子？

要评价因子好不好，我们要从因子定义和用法上出发。多因子模型中，将股票收益率解释为因子收益率的线性组合，组合的权重就是因子值，认为股票收益率受因子影响，那么一个好的因子就应该能较好解释股票收益率。模型上来说，较好解释股票收益率即就是模型的拟合程度高，因子显著性高。从这一角度出发，就引出了单因子测试的**回归方法**。

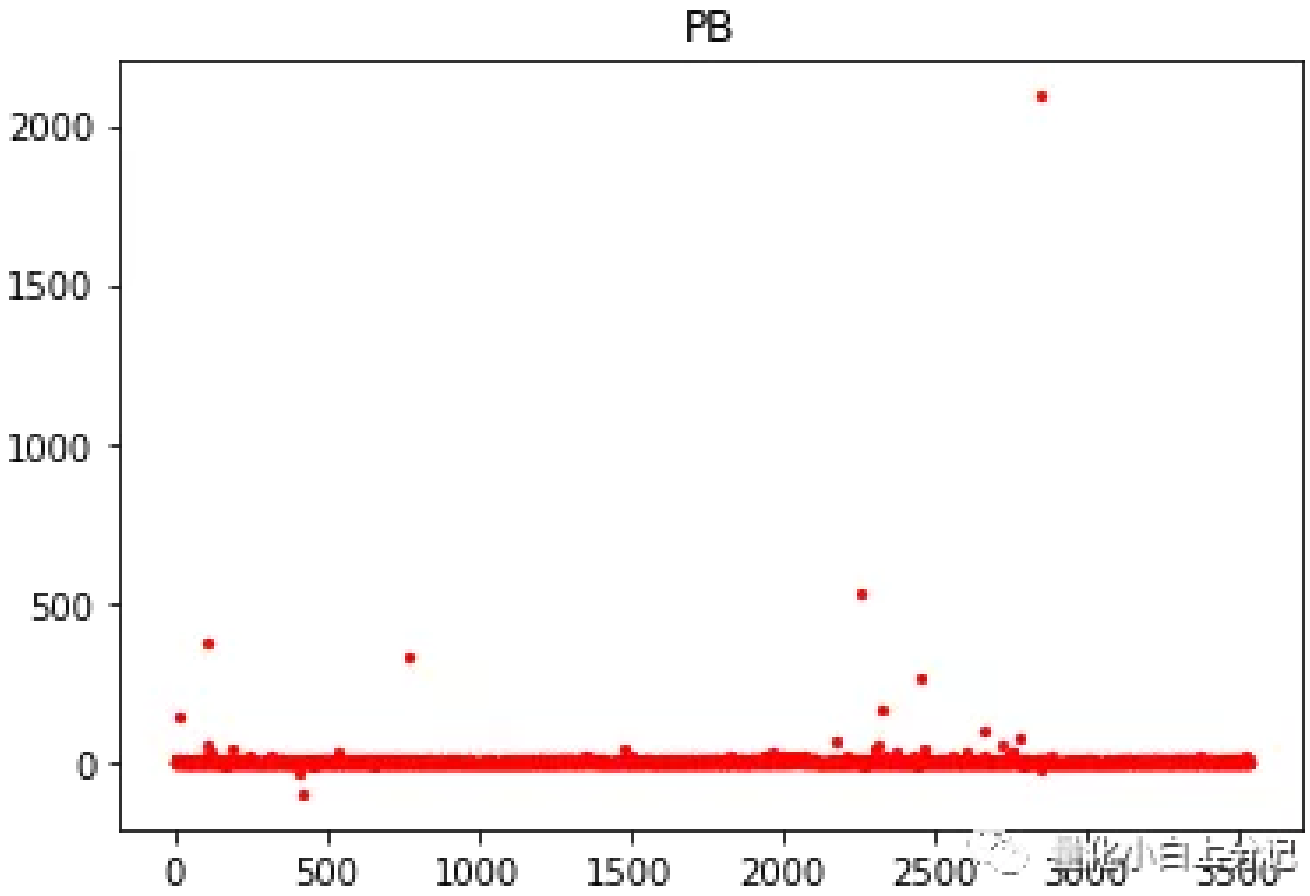
这种方法侧重于对于已经得到的股票收益，分析它和因子值的相关性，用因子解释收益，但这样考虑是不够的，**相关性并不是因果性，而且即使过去有相关性，并不代表未来也有相关性，预测性不强**。如果一个因子是好因子，那么基于因子值筛选出来的股票，也应该表现良好，这样更能说明因子可以解释股票收益，从这一角度出发，就得到了单因子测试的**分层测试法**。

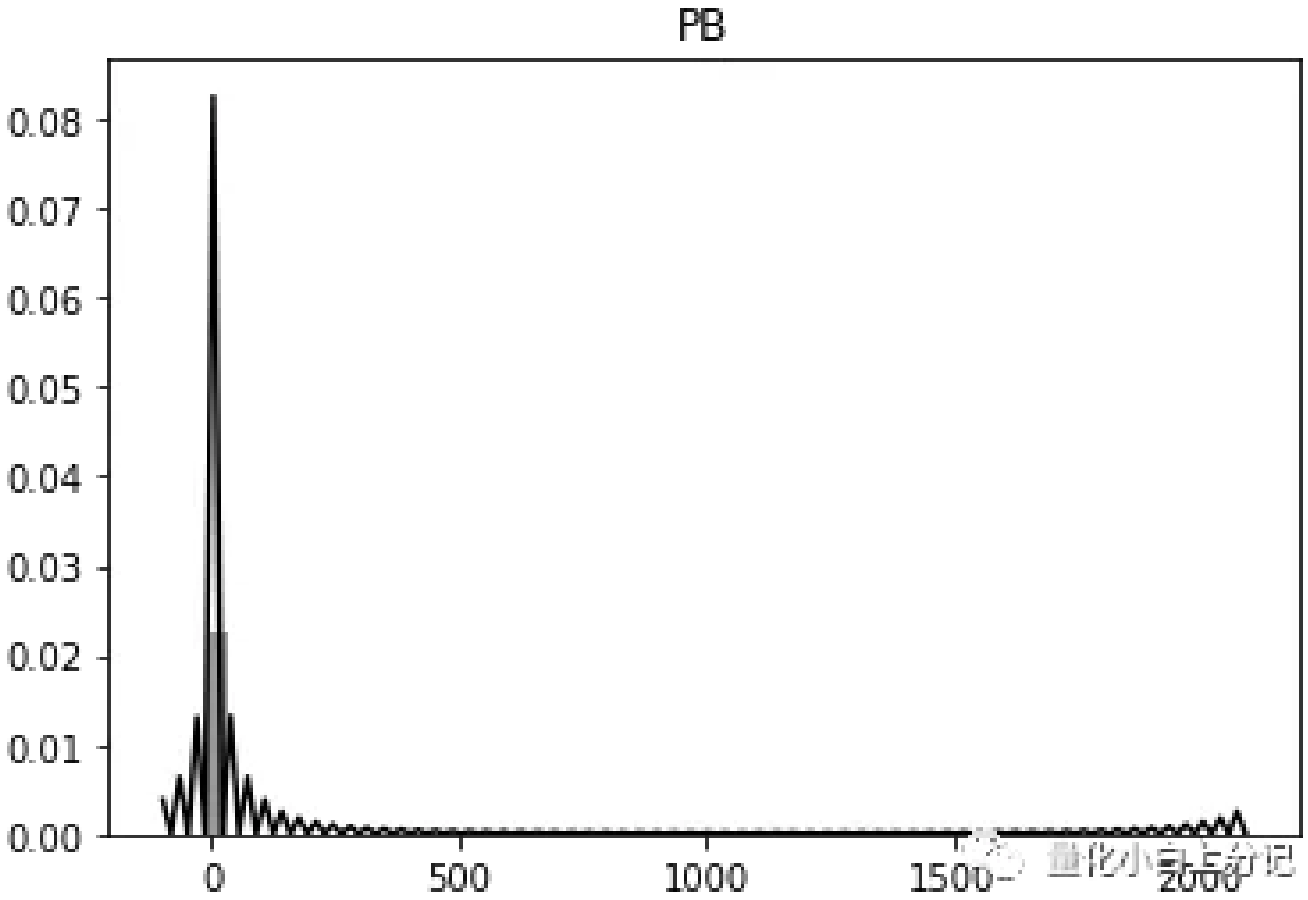
实际应用中，两种方法各有优劣，有不同侧重点，配合使用比较好。具体流程在之后两篇文章中细讲，本文着重于单因子测试的第一步：**因子预处理**。

03 因子预处理

因子预处理的过程即就是清洗数据的过程，一方面，因子中可能存在缺失值、异常值，不进行处理的话，单因子测试的结果是不准确的,需要进行**因子标准化**。

以PB因子为例，取2018年7月10日A股市场所以股票的PB因子作图：


PB
2000
1500
1000
500
0
0
500
1000
1500
2000
2500
3000
3500



第一张图纵轴为PB值，横轴为股票编号，第二张图为PB的直方图和密度曲线。可以看出，绝大部分PB因子的值在0附近，但小部分取值十分异常，甚至上千，导致密度曲线严重拖尾。

另一方面，类似于计量经济学中所说的遗漏变量问题，**当我们分析单个因子对于股票收益率的影响时，得到的结果里可能包含我们没有考虑到的因素造成的影响**，这样得到的结果是有偏的。股票市场中不同市值、不同行业、不同风格的股票，对于因子的响应性不同，因此，在进行因子测试前，我们必须对因子进行处理，剔除掉因子中可能包含的其他因素，处理方法也与计量中的方法类似——加控制变量，这里叫做**因子中性化**，实际操作中，我们一般只考虑市值和行业造成的影响，对这两方面的处理分别称为**市值中性化**和**行业中性化**。

因子标准化

1. 缺失值处理

因子测试时，一般直接删掉缺失值，但在后续建模时，有时需考虑对缺失值进行填补，这里不必考虑。

2. 去极值

去极值一般有三种方法：

均值方差去极值：求出因子的均和标准，把位于$\mu-3\sigma,\mu+3\sigma$以外的值用边界值代替。

中位数去极值：计算因子的中位数med，定义因子的绝对中位数

$$MAD=median(x_i-med)$$

即因子值减去中位数后的中位数。将位于med+3*1.4826MAD和med-3*1.4826MAD以外的数字用边界值代替

分位数去极值法：取因子序列的上下分位数，比如5%分位数和95%分位数，把位于分位数以外的数据用分位数代替。

3. 标准化

标准化一般采用zscore方法

$$x_{zscore}=\frac{x-mean(x)}{std(x)}$$

因子中性化

1. 行业中性化

行业中性化有两种方法，一种是之前所说类似计量中加控制变量的方法，用因子值做因变量，用所属行业（申万行业、中信行业）虚拟变量做自变量进行OLS回归，用回归之后的残差值代替因子值。

另一种方法是对因子分行业进行标准化，即减去行业均值之后再除以行业标准差，**可以证明，两种方法得到的结果是完全一样的**。第一种方法的代码相对简单，并且可以和市值中性化一起进行，因此一般采用第一种方法。

2. 市值中性化

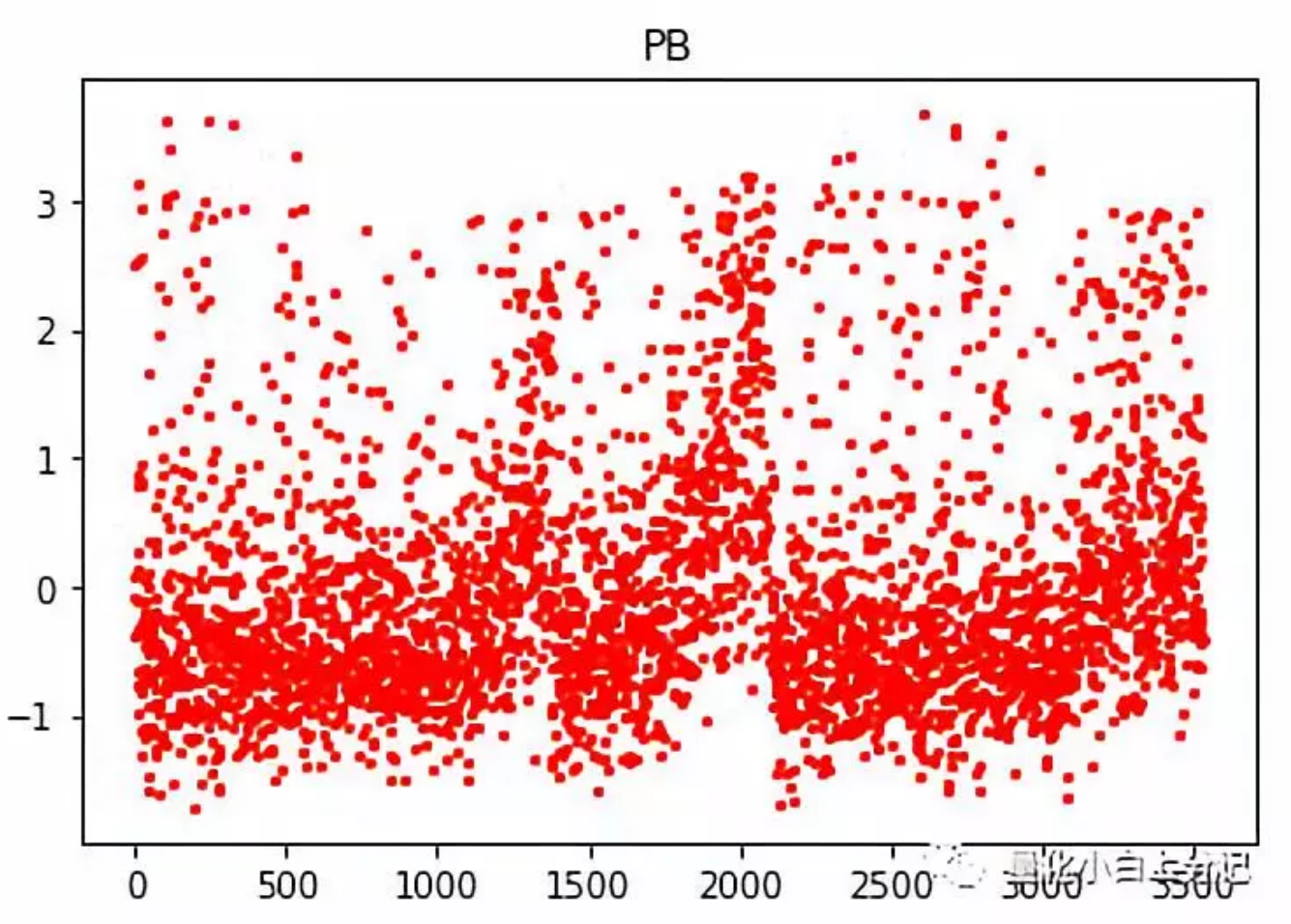
用因子值做因变量，市值做自变量（有时也取市值对数），进行回归，取残差。

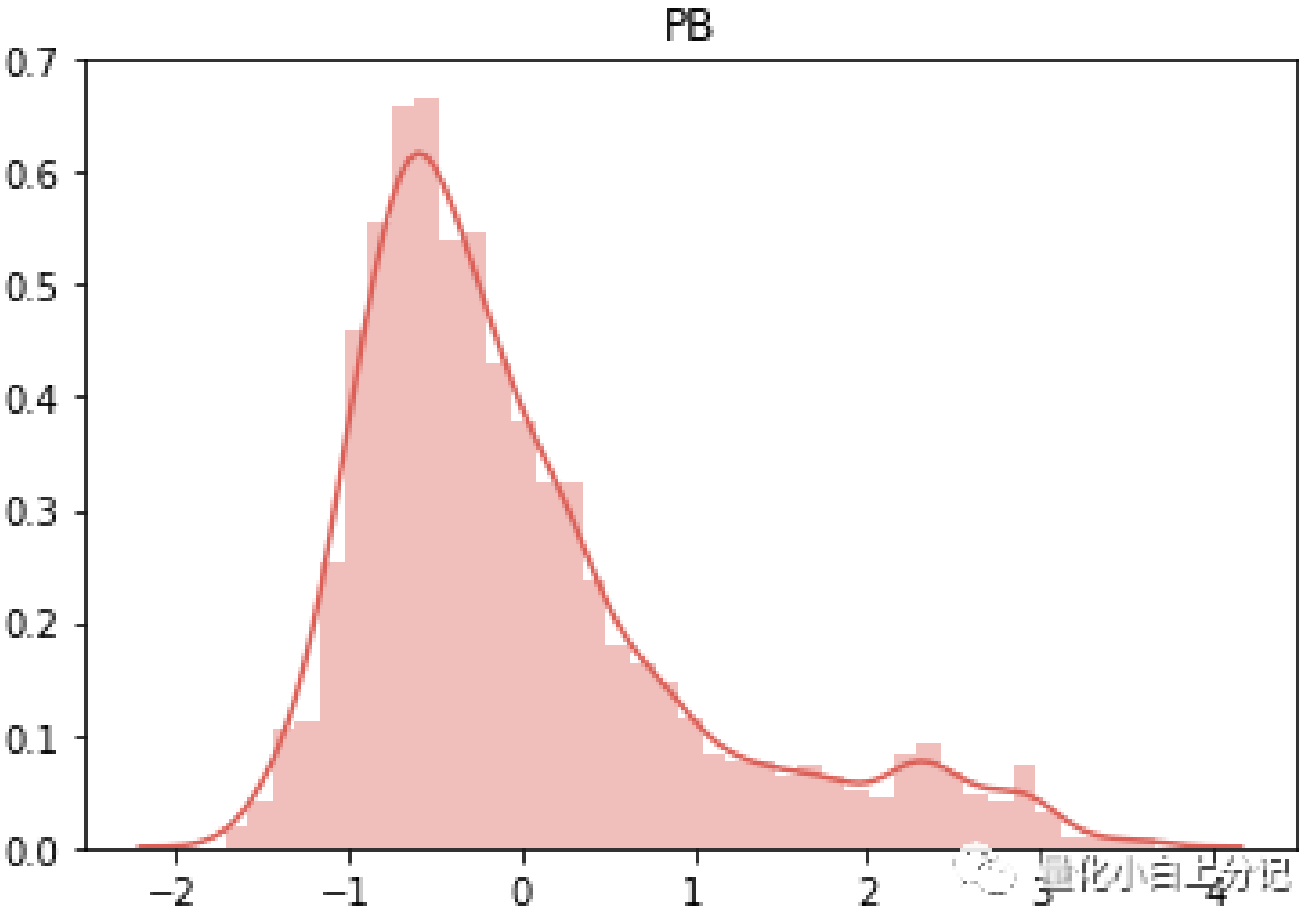
一般将行业虚拟变量和市值同时放在自变量上进行回归，同时进行市值中性化和行业中性化，**理论上可以证明，回归后的残差序列与自变量序列均正交，因此可以认为回归后的残差是因子剔除了行业和市值影响后的纯净的因子**。

这里附上用回归的方法做中性化的python代码，python的**pandas**包里有可以直接生成虚拟变量的函数，回归**statsmodels**包中也有函数，因此整个过程就变得非常简单。

```
# 是否行业中性
if if_neutral_industry:
    indname = industry.unique()    # 获取行业名
    class_var = pd.get_dummies(industry['industry'],columns=['industry'],prefix=['i
                    prefix_sep="_", dummy_na=False, drop_first=False)
    class_var[indname[-1]] = 0

# 是否市值中性
if if_neutral_mktcap:

    # 提取总市值

    class_var['logmktcap'] = np.log(mktcap)

data = pd.concat([data,class_var], axis=1)
if if_neutral_industry | if_neutral_mktcap:
    for j in data.columns:
        x = np.hstack((np.ones((len(data),1)),class_var.values))
        y = data.iloc[:,j].values
        model = sml.OLS(y,x)
        result=model.fit()
        data.iloc[:,j] = y -result.fittedvalues
```

结合前两幅图，对比pb因子标准化、中性化之后的分布图


PB
3
2
1
0
−1
0
500
1000
1500
2000
2500
3000
3500



可以看出，经过标准化和中心化之后，因子序列的分布已经正常了许多。

以上就是因子预处理的整个过程，但是经过上述处理的股票依然不能直接进行单因子测试，还需要对股票的范围进行限定，并非所有的股票都用来测试，需要删除**停牌**的股票，**ST股**、**新股**，这些股票不剔除也会对测试结果造成影响。

做完上面的所有处理之后，就可以对因子进行单因子测试了，因子测试的两种方法之后两篇文章来写，同时也会把进行中性化和不进行中性化的测试结果进行对比。

文章为个人理解，有问题请指出，谢谢！

参考文献

1.国泰君安，数量化专题之五十七：基于组合权重优化的风格中性多因子选股策略

2.申万宏源-申万宏源多因子系列报告之一：因子测试框架及批量测试结果

收录于话题 #多因子模型·12个 >

< 上一篇
规模类因子测试

下一篇 >
单因子测试（中）——分层测试法

喜欢此内容的人还喜欢

查理·芒格震撼演讲：一生抓住少数几个机会，够了！
量化小白上分记

“我的生命走到头了”……90后单亲妈妈住所内藏毒14公斤被抓获
中国禁毒

复利的谎言
财经盘点