指数增强策略研究报告

基于沪深 300 指数

张墨轩

目录

项目概览：

本项目通过构建多因子模型实现了沪深 300 指数增强策略，流程包括原始数据获取、因子生成、因子预处理、单因子检验、收益预测模型及风险模型等，具体流程如图：

代码框架：

1.raw_data_fetch.py/原始数据获取.py：原始数据获取功能文件，用于从 wind/tushare 数据源获取原始数据，生成代理指标，数据以指标名作为文件名保存到磁盘文件中。

2.factor_generate.py/因子生成.py：因子生成功能文件，用于通过对各种指标的计算生成相应各种因子，因子最终以截面形式保存到文件。

3.factor_preprocess.py/因子预处理.py：因子预处理功能文件，用于对计算后的原始因子数据进行预处理，包括缺失值处理、中位数法去极值、标准化及相对行业和市值中性化处理。

4.single_factor_test.py/单因子检验.py：单因子检验功能文件，同时实现回测功能模块，用于对预处理后的因子逐个进行回归，IC 检验、T 检验、分层回测，对有效因子进行初步筛选。

5.index_enhance.py/指数增强模型.py：指数增强模型功能文件，用于对有效因子进行因子合成与正交，进行收益预测，并通过风险模型进行权重优化，然后利用回测引擎进行回测，生成回测报告，最后对结果进行行业绩归因。

本模型针对沪深 300 成分股，最终选取估值因子、动量因子、流动性因子、成长因子、作为阿尔法因子，选取波动率因子、市值因子、 Beta 因子和行业因子作为风险因子，采用月

频方式调仓，根据历史月频横截面回归计算因子收益率，并在每个月末根据最新的因子暴露预测下个月全部指数成分股收益，并通过风险模型以最大化组合的预测收益为目标，同时控制组合相对基准指数的行业及市值暴露，实现对组合中个股的权重优化，以沪深 300 增强为例，每个月通过模型从 300 只成分股中选出约 60-100 只个股，以最优权重构建组合。回测周期为 2014 年 1 月 - 2019 年 12 月。模型采用 24 个月指数加权方式预测 T+1 期因子收益率，并进行市值中性化和行业中性化处理。经过回测，模型年化超额收益 8.3%，并在 2014-2019 年 期间每一年都跑赢指数，平均跟踪误差 6.2%，平均双边换手率 12 倍。

多因子模型构建步骤：

一.因子生成

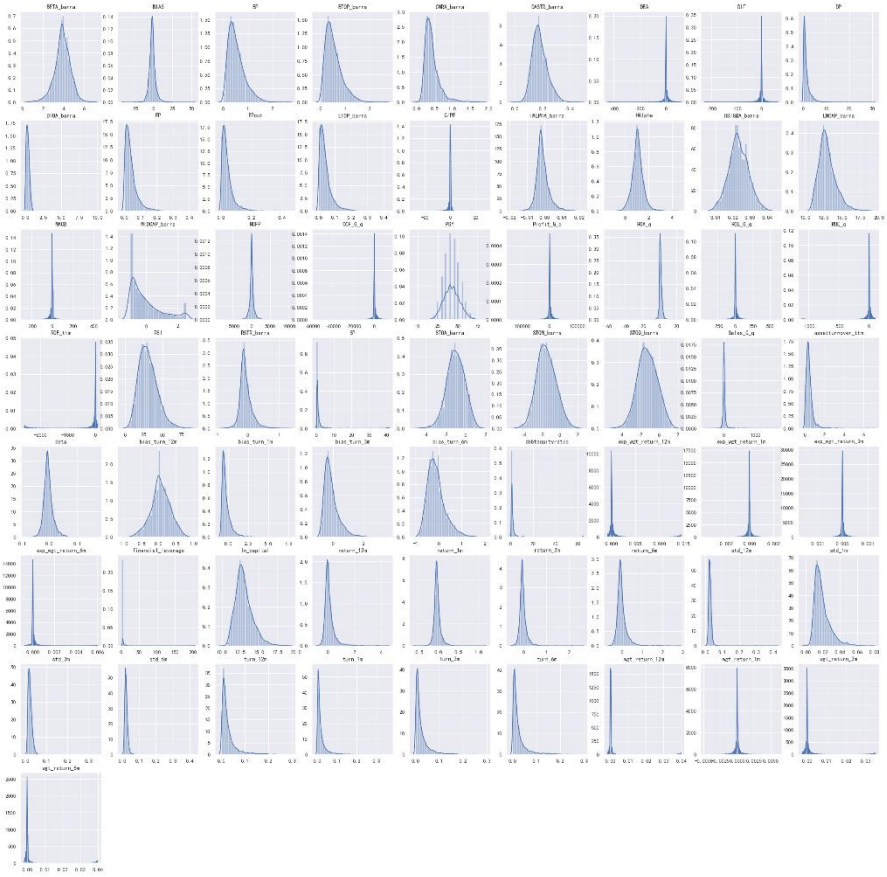
通过原始数据及指标生成各种因子，其中主要的关注点是财务指标的生成，一定要防止使用未来数据，财报数据需要进行数据对齐，因为各上市公司财报的发布日期之间不一定都是同一天，为避免前视偏差，在提取数据的时候需要对日期进行修正，保证因子数据为当时那个时刻所能获取的最新财报数据。

二.数据预处理

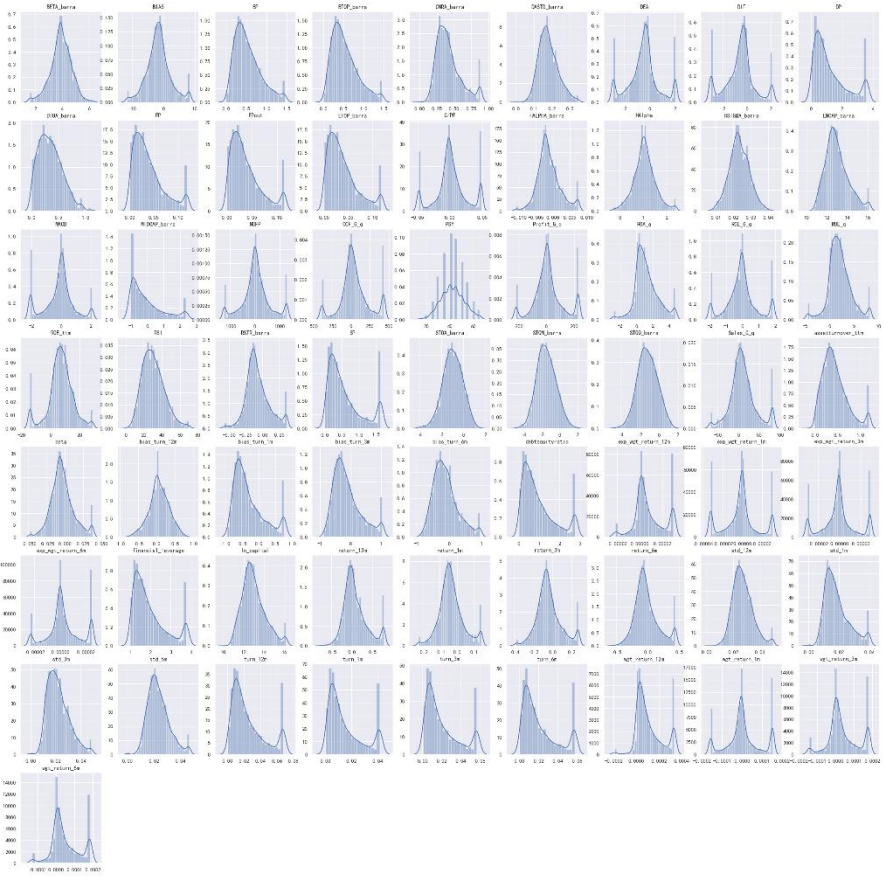
原始因子数据需经过缺失值处理、去极值、中性化、标准化等步骤后，才能进行下一步的单因子检验。

1．缺失值处理：原始因子会因为各种原因出现缺失值，当缺失值少于比如 10%的情况下，可以使用行业中位数代替，但是如果缺失值过多，那么最好更换数据源或使用其他因子。填充后某期因子如下图:

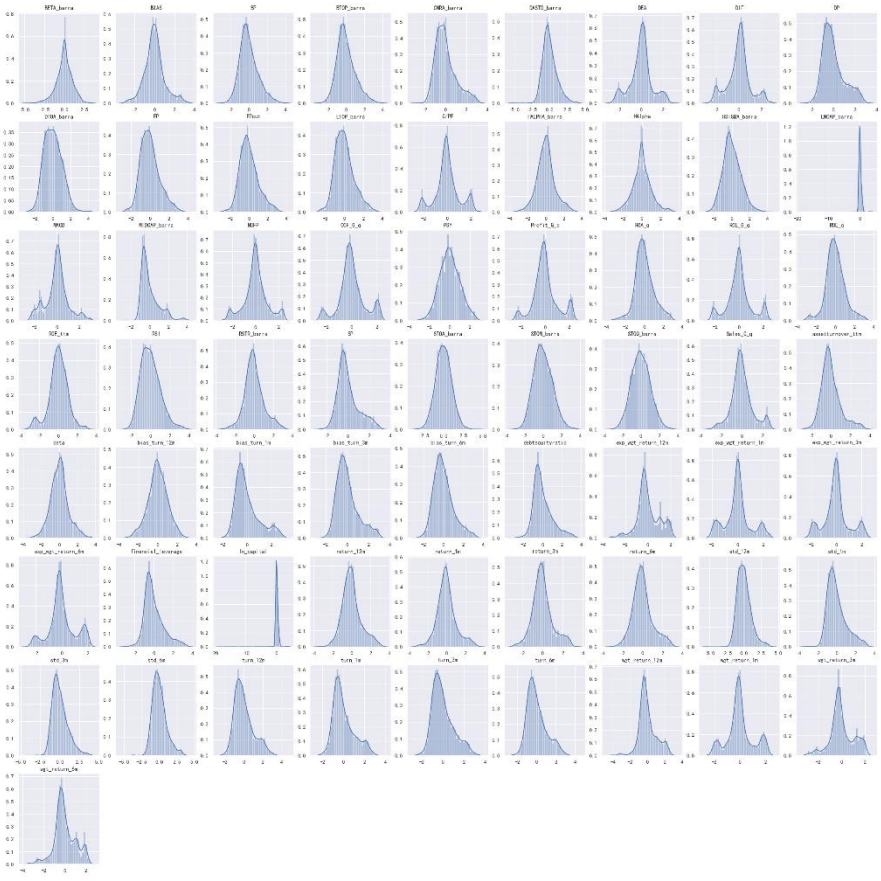
缺失值填充后-2019-11-29



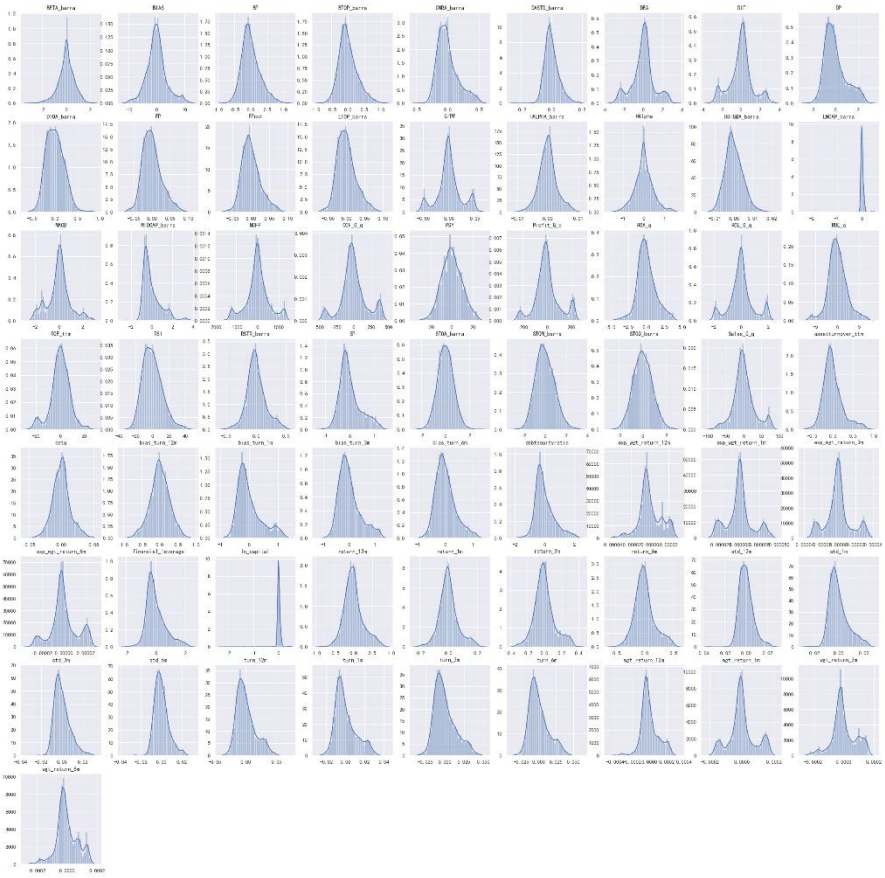
2．去极值：使用“中位数去极值法”，将超过上下限的极端值用上下限值代替，这样可以尽量防止极端值对回归产生过多影响。去极值后某期因子如下图：



3．标准化：不同因子量纲不同，为了使其具有可比性，需要对其进行 ZScore 标准化处理，使因子序列近似成为一个符合 N(0,1)正态分布的序列。标准化后某期因子如下图：



4．行业市值中性化：为了剔除数据中多余的风险暴露，防止投资组合过于集中，将因子序列对流动市值与行业哑变量做线性回归，取残差作为新的因子值。中性化后某期因子如下图：

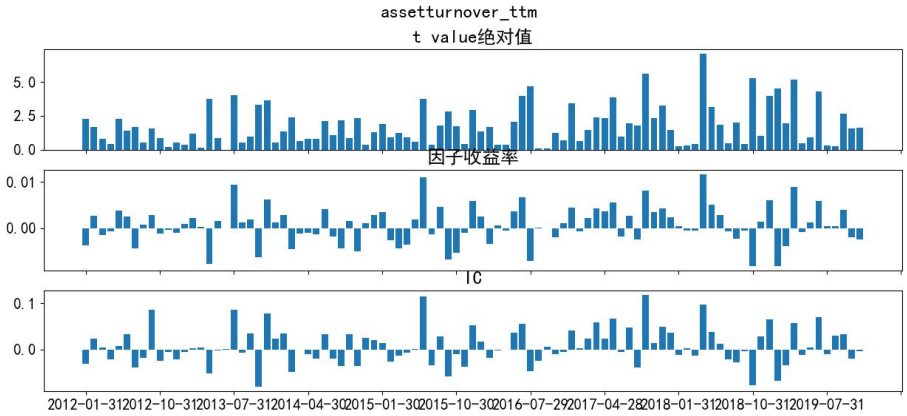


三.单因子检验

单因子检测一般分为两种方式，一种是偏数学的统计检验法，另外一种是偏实际的分层回测法，两种方法可以结合使用互相印证。

1．统计检验法

依据 Barra 手册中关于因子显著性测试的内容，对因子进行横截面回归，同时需要考虑行业与市值的影响。而 IC（ Information Coefficient）是衡量因子收益预测能力的重要参数，它的计算方法是将每一期的因子值作为因变量，与行业哑变量和市值变量进行回归，取其残差，作为剔除行业与市值影响后的因子值。再计算新因子值与下一期股票收益序列间的 Spearman 相关系数。某因子的 IC 等检验图如下:



最终每个因子统计结果如下图所示:

	IC>0概率	ICIR	IC平均值	IC标准差	t值绝对值>2概率	t值绝对值平均值	因子收益>0概率	因子收益t值	因子收益平均值	因子收益标准差
2012	0.333333333	-0.39211143	-0.04205473	0.10725199	0.75	5.999797654	0.333333333	-0.846904961	-0.002847861	0.011152712
2013	0.416666667	-0.26874805	-0.01843986	0.068613931	0.5	3.284224946	0.25	-1.607972768	-0.003380018	0.006971667
2014	0.416666667	0.039177729	0.002362388	0.060299251	0.5	3.321551905	0.333333333	0.177343934	0.000486957	0.009106896
2015	0.416666667	-0.00791813	-0.00048779	0.061603991	0.583333333	3.447239301	0.416666667	-0.221356873	-0.000840437	0.012592395
2016	0.583333333	-0.0196119	-0.00091811	0.046814051	0.916666667	3.033637131	0.5	-0.185905007	-0.000286515	0.005111554
2017	0.416666667	-0.05305311	-0.00259353	0.048885624	0.666666667	2.808106623	0.416666667	-0.226935493	-0.000338338	0.004944755
2018	0.75	0.508041152	0.023482843	0.046222325	0.583333333	3.156214017	0.583333333	0.824608726	0.001409726	0.005670001
2019	0.545454545	0.235880473	0.01445994	0.061301978	0.545454545	3.60601023	0.545454545	0.997307068	0.002063291	0.006542316

评估指标说明:

IC>0 概率：衡量模型预测收益方向性是否一致的指标;

IRIC： IC 平均值/IC 标准差;

IC 平均值：衡量模型预测能力的指标;

IC 标准差：衡量模型预测能力是否稳定的指标;

t 值绝对值>2 概率：衡量因子显著性是否稳定;

t 值绝对值平均值：衡量因子整体显著性的指标;

因子收益>0 概率：衡量因子收益率方向性是否一致的指标;

因子收益 t 值：衡量因子收益率统计上是否显著不为 0 的指标;

因子收益平均值：衡量因子收益能力大小的指标;

因子收益标准差：衡量因子收益能力波动率的指标;

最终可以依据以上每个因子的评估指标选出合适的因子, 比如 t 值绝对值平均值大于 2, ICIR 稳定的因子等。

2．分层回测法

在每个截面期的最后一个交易日，提取样本内股票因子值，并剔除因子值缺失的股票。按照因子将样本内股票排序，并按照序号从大到小平均分为 5 组，当然也可以分成 10 组。在下一个截面期的首个交易日，以当天的收盘价换仓并剔除当天因停牌等因素不能交易的股票。 对 5 组股票的历史收益率进行回测，并计算其年化收益率、波动率、夏普比率等值。然后查看每一组的股票收益率是否有较好的分层单调性，判断因子是否在截面上对股票有很好的区分能力。某因子分层图如下:

某因子分年收益图如下:

某因子第一组减去第五组收益图如下:

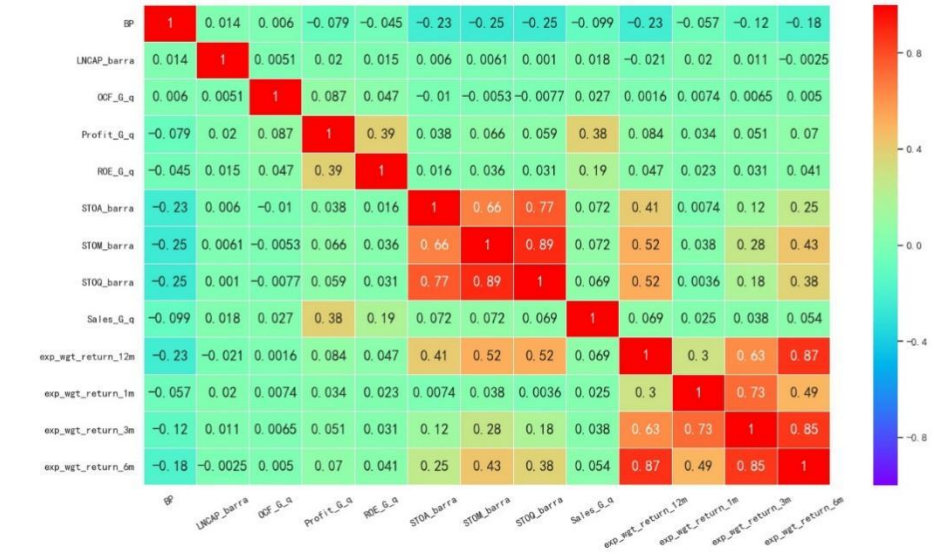
通过单因子检验最终选择以下这些因子作为阿尔法因子:

因子名称	所属大类	单位	因子定义
BP	估值	无（比值）	净资产/总市值
OCF_G_q	成长	百分比(×100%)	经营性现金流同比增长率
ROE_G_q	成长	百分比(×100%)	ROE同比增长率
STOA_Barra	流动性	无（比值）	平均换手率（12个月）
STOM_Barra	流动性	无（比值）	平均换手率（1个月）
STOQ_Barra	流动性	无（比值）	平均换手率（3个月）
Sales_G_q	成长	百分比(×100%)	营业收入同比增长率
exp_wgt_return_12m	动量反转	百分比(×100%)	
exp_wgt_return_1m	动量反转	百分比(×100%)	个股最近N个月内用每日换手率乘以函数exp(-x_i/N/4)再乘以每日收益率求算术平均值，x_i为该日距离截面日的交易日的个数，N=1，3，6，12
exp_wgt_return_3m	动量反转	百分比(×100%)	
exp_wgt_return_6m	动量反转	百分比(×100%)	
Profit_G_q	成长	百分比(×100%)	净利润同比增长率

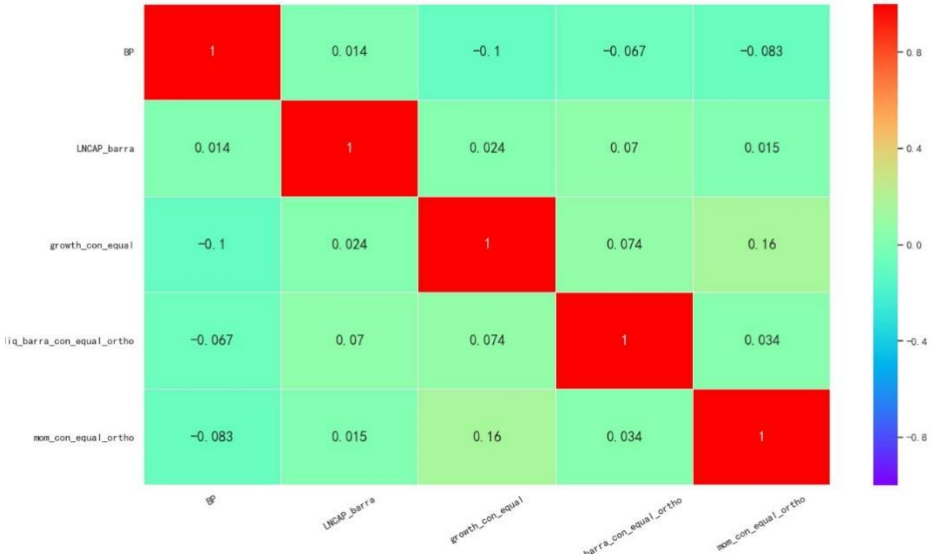
四.指数增强策略构建

虽然上一步我们已经选出了在主观上与收益率有显著关系的因子池，但因子彼此之间（尤其是同一大类下所属的因子）有可能存在很强的相关性。这就是多重共线性问题，它会对投资组合有很不利的影响。因此有必要对因子进行大类因子合成以及因子正交，如果不做处理，投资组合会在同种因子类型上暴露过多风险，这种多重共线性问题会导致多元线性回归的结果偏差。一般在具体处理上可以有多种方法，比如对非常相似的因子可选取单因子检验效果最好的因子进入模型，剔除其他因子，也可以利用等权法、历史收益率加权法、 ICIR 加权法等处理手段，对因子进行合并，生成新的因子。因子正交也有施密特正交、规范正交和对称正交等。

因子处理前相关系数图如下:



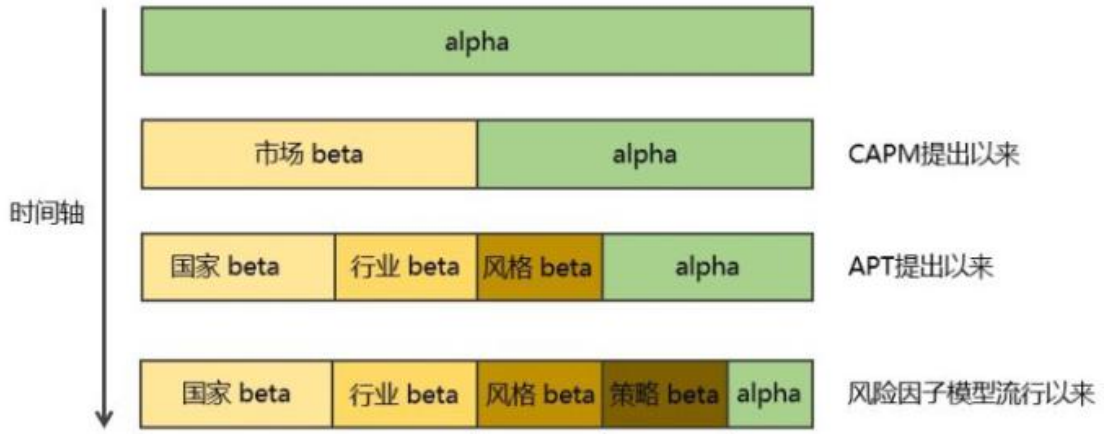
因子处理后相关系数图如下:



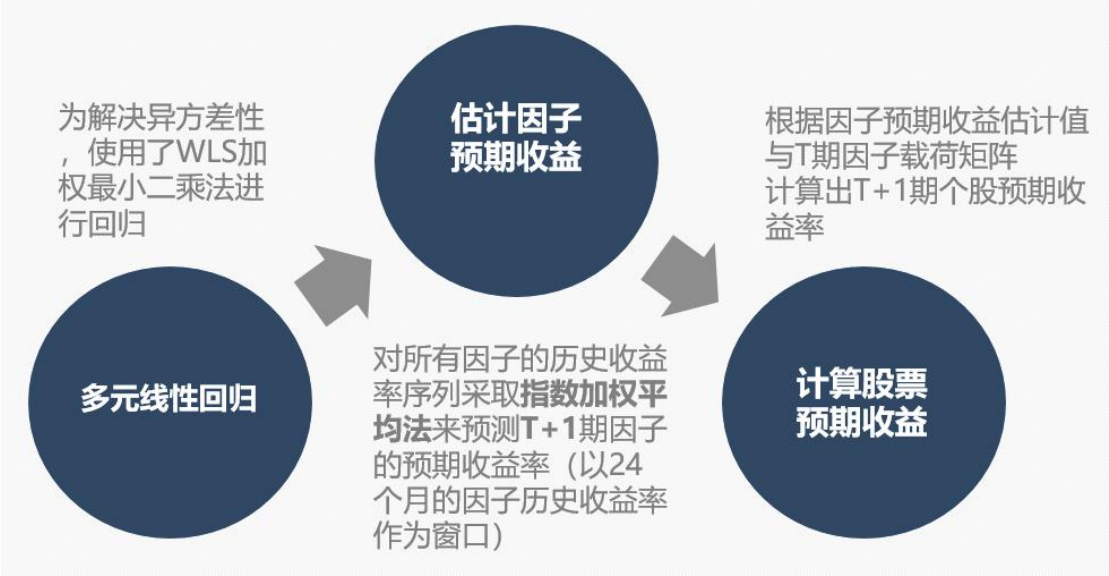
从以上处理前后的因子相关系数图对比可以看到，大类因子合成以及因子正交还是很有必要的。因子数目也大大减少，处理后合成了新的动量反转因子，流动性因子以及成长因子等少数因子。

接下来开始计算每一期的因子收益率，因子收益率的估计可以通过横截面回归得到（可以参考 Barra 模型），使用下一期个股收益率对上一期个股因子暴露进行回归，回归系数即为因子收益率。为解决异方差性，使用了 WLS 加权最小二乘法进行回归。

接下来开始计算个股收益率，股票收益率可以表示为市场收益率、行业收益率、风格收益率、阿尔法因子收益率以及特质收益率的线性组合。类似下图:



得到所有因子的历史收益率序列后，就可以去估计 T+1 期因子预期收益率了，当然估计的方法有很多种，比如历史数据平均法，指数加权移动平均法，ARMA，ARCH，GARCH，滤波，神经网络等等，这里我采样窗口期为 24 个月的指数加权移动平均法来估计。然后通过因子预期收益率就可以计算出个股预期收益率。流程如图:



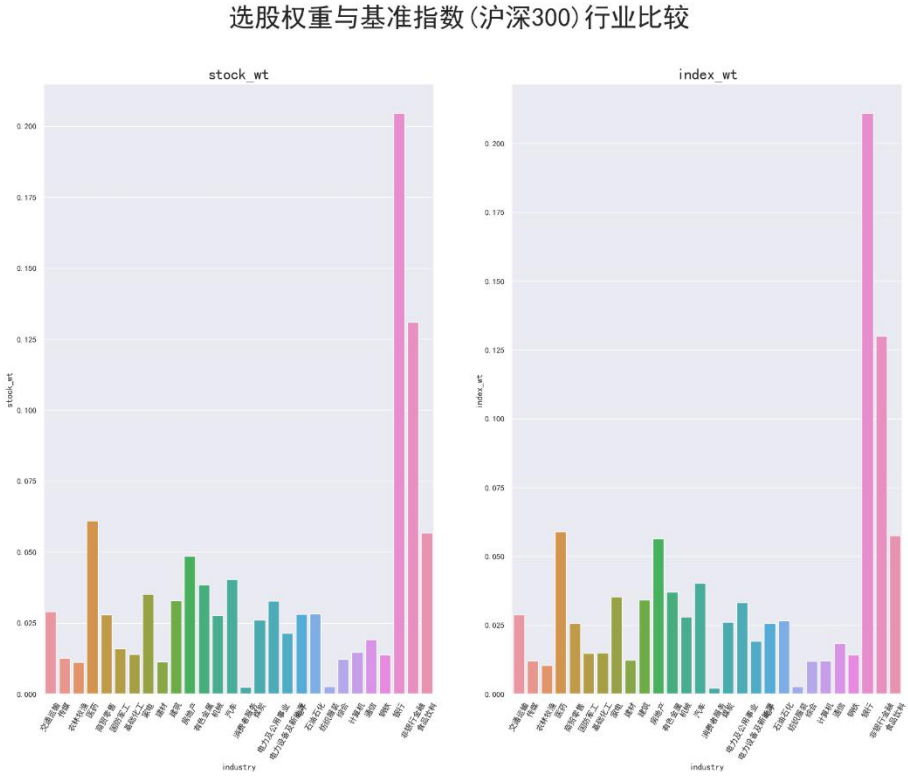
最后以波动率因子、市值因子、Beta 因子、行业因子等作为风险因子，加上行业中性约束，市值中性约束等通过动态规划的方法计算出每期的股票最优权重。

某期输出优化权重结果如下图:

1		stock_wt	industry	sec_name	mkt_cap_fl	turn	amt	close
2	000027.SZ	0.005333	电力及公月	深圳能源	1830448	0.054735	98532.62	19.1
3	000039.SZ	0.010133	机械	中集集团	3263942	0.032121	104777.2	26.29
4	000157.SZ	0.0128	机械	中联重科	5591069	0.107022	588699.1	8.93
5	000402.SZ	0.014613	房地产	金融街	3777269	0.028279	106673.9	12.65
6	000413.SZ	0.002347	电子	东旭光电	2589952	0.109402	286896.2	12.83
7	000425.SZ	0.0096	机械	徐工机械	3454632	0.041012	142273.1	14.66
8	000568.SZ	0.004053	食品饮料	泸州老窖	3530565	0.022965	81149.13	25.21
9	000623.SZ	0.013333	医药	吉林敖东	3076470	0.040357	125045.8	38.99
10	000651.SZ	0.010347	家电	格力电器	18578825	0.034465	610169.7	62.21
11	000709.SZ	0.010667	钢铁	河钢股份	5361530	0.036423	195172.8	5.05
12	000725.SZ	0.021333	电子	京东方A	11770068	0.144917	1720317	5.04
13	000729.SZ	0.006933	食品饮料	燕京啤酒	2815533	0.025206	69603.03	11.26
14	000776.SZ	0.038933	非银行金融	广发证券	15289530	0.018351	283053.5	25.83
15	000858.SZ	0.022933	食品饮料	五粮液	10183651	0.019572	200370.6	26.83
16	000876.SZ	0.005973	农林牧渔	新希望	3804006	0.032063	122569.3	22.01
17	000917.SZ	0.000533	传媒	电广传媒	4554542		0	42.89
18	002024.SZ	0.012693	商贸零售	苏宁易购	9155143	0.071588	656443.5	18.53
19	002142.SZ	0.009067	银行	宁波银行	5426184	0.031021	169311.1	18.89
20	002202.SZ	0.0144	电力设备及	金风科技	4940301	0.054423	270756.5	23
21	002399.SZ	0.004267	医药	海普瑞	3427257	0.014468	48811.39	42.83
22	002422.SZ	0.0064	医药	科伦药业	1579615	0.04118	64196.8	44.8
23	002465.SZ	0.0112	国防军工	海格通信	3748015	0.050585	186693.3	41.5
24	600011.SH	0.0192	电力及公月	华能国际	11361000	0.009503	108004.4	10.82
25	600019.SH	0.008213	钢铁	宝钢股份	13648576	0.011708	160687.2	8.31
26	600023.SH	0.003733	电力及公月	浙能电力	3049784	0.27691	804110.4	11.94
27	600027.SH	0.008533	电力及公月	华电国际	4998048	0.022932	114951	8.5
28	600028.SH	0.01984	石油石化	中国石化	68897153	0.006952	479753.3	7.21
29	600030.SH	0.077333	非银行金融	中信证券	29807128	0.040356	1213819	30.37
30	600036.SH	0.140267	银行	招商银行	37276503	0.012679	468943.5	18.07
31	600038.SH	0.004267	国防军工	中直股份	3174183	0.042185	131724.2	80.84
32	600050.SH	0.020693	通信	中国联通	18186680	0.037951	698659.3	8.58
33	600060.SH	0.010133	家电	海信视像	4191065	0.02695	111660	32.03
34	600079.SH	0.00352	医药	人福医药	2308515	0.033524	76256.77	45.63
35	600089.SH	0.0096	电力设备及	特变电工	5369388	0.050591	270561.5	16.96
36	600100.SH	0.017067	计算机	同方股份	5958040	0.093717	557273.5	28.78
37	600104.SH	0.020907	汽车	上汽集团	26439309	0.008822	234324.9	23.98
38	600143.SH	0.006933	基础化工	金发科技	3783680	0.061106	222420.6	14.78
39	600153.SH	0.012267	交通运输	建发股份	5092020	0.036565	184425.4	17.96
40	600166.SH	0.0064	汽车	福田汽车	2380738	0.060639	140745.5	9.38
41	600177.SH	0.004053	纺织服装	雅戈尔	5032142	0.03445	172693.4	22.6
42	600352.SH	0.004693	基础化工	浙江龙盛	5243193	0.050943	263260.5	34.27
43	600362.SH	0.009067	有色金属	江西铜业	5159065	0.059828	306845.6	24.86
44	600369.SH	0.0128	非银行金融	西南证券	5464971	0.028797	157901.7	23.53
45	600373.SH	0.005333	传媒	中文传媒	3910378	0.028385	112193.4	32.98

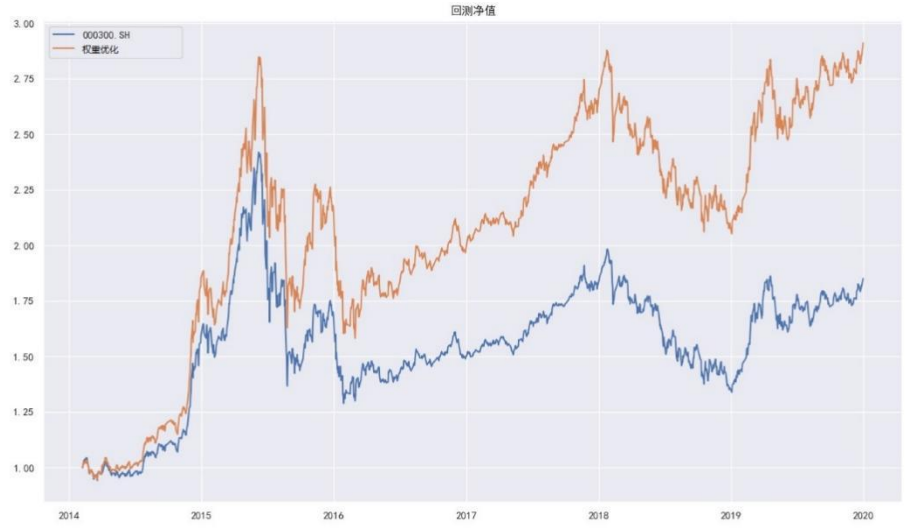
46	600383.SH	0.027733	房地产	金地集团	5964663	0.014904	89336.37	13.28
47	600547.SH	0.010667	有色金属	山东黄金	4062872	0.028822	117823.6	28.55
48	600585.SH	0.00768	建材	海螺水泥	10227239	0.041291	421019.5	25.57
49	600642.SH	0.010347	电力及公用	申能股份	5225740	0.045208	230903.5	11.48
50	600648.SH	0.0048	商贸零售	外高桥	3566230	0.019972	71338.4	38.15
51	600664.SH	0.0048	医药	哈药股份	2341247	0.039874	89776.01	12.21
52	600718.SH	0.00544	计算机	东软集团	3778535	0.042517	159310.8	30.78
53	600839.SH	0.013867	家电	四川长虹	3904668	0.056384	219212.4	8.47
54	600873.SH	0.008	食品饮料	梅花生物	3478106	0.043551	151859.6	11.19
55	601006.SH	0.00928	交通运输	大秦铁路	18107752	0.014427	263474.5	12.18
56	601009.SH	0.0144	银行	南京银行	5635035	0.05105	286115.5	18.98
57	601088.SH	0.00608	煤炭	中国神华	33905574	0.006706	228497.2	20.56
58	601098.SH	0.008	传媒	中南传媒	4928224	0.009693	48444.41	27.44
59	601168.SH	0.010667	有色金属	西部矿业	2666577	0.043563	116571.5	11.19
60	601169.SH	0.007036	银行	北京银行	13443124	0.03315	445340.8	12.73
61	601333.SH	0.0112	交通运输	广深铁路	4092220	0.066622	268377.5	7.24
62	601398.SH	0.000324	银行	工商银行	1.37E+08	0.001998	273079.5	5.07
63	601601.SH	0.0512	非银行金融	中国太保	20262034	0.0156	320068	32.23
64	601607.SH	0.011733	医药	上海医药	5186156	0.030291	157238.9	26.97
65	601668.SH	0.061547	建筑	中国建筑	27793348	0.032878	912764.9	9.31
66	601699.SH	0.005867	煤炭	潞安环能	3002915	0.029472	88714.41	13.05
67	601899.SH	0.001387	有色金属	紫金矿业	11015251	0.063576	698814.5	6.97
68	601928.SH	0.0064	传媒	凤凰传媒	4988004	0.027272	135921.3	19.6
69	601929.SH	0.0064	传媒	吉视传媒	2445510	0.039893	97164.77	16.66
70	601989.SH	0.010667	国防军工	中国重工	31731531	0.06157	1907764	17.67
71	601998.SH	0.011733	银行	中信银行	23577916	0.005504	129956.8	7.39

某期持仓权重对比基准指数在各行业对比图:



上图中左边是模型输出的优化权重，右边是基准指数的权重，可见模型输出的优化权重与基准指数成份股权重在行业上的分布是相当的接近，达到了行业中性化的效果。

将咱们模型每期输出的优化权重通过回测引擎进行回测，最终得到的策略净值图如下:



模型收益情况统计如下:

	年度收益	年度波动	夏普比率	最大回撤	年度超额收益	跟踪误差	信息比率	日胜率	换手率
2014	0.85454773	0.199748231	4.0778721	0.0851711	0.163208815	0.0599012	2.0568666	0.56696429	11.63065
2015	0.147493196	0.419393179	0.2563065	0.4275384	0.130411972	0.088106	1.0415108	0.50819672	12.999826
2016	-0.009113062	0.251517412	-0.195267	0.2244625	0.044800745	0.0519093	0.0930659	0.55737705	12.174738
2017	0.299520002	0.119920608	2.1640984	0.0659643	0.078565089	0.0487553	0.8123877	0.5204918	11.119795
2018	-0.223381569	0.220387131	-1.195086	0.2836252	0.055213639	0.0465335	0.3453412	0.53909465	10.474423
2019	0.417115147	0.206840398	1.823218	0.1274871	0.027689008	0.0592479	-0.209741	0.51229508	12.076521
总计	0.198642646	0.259442437	0.6114753	0.4444217	0.083394677	0.0620901	0.698899	0.53361053	12.706497

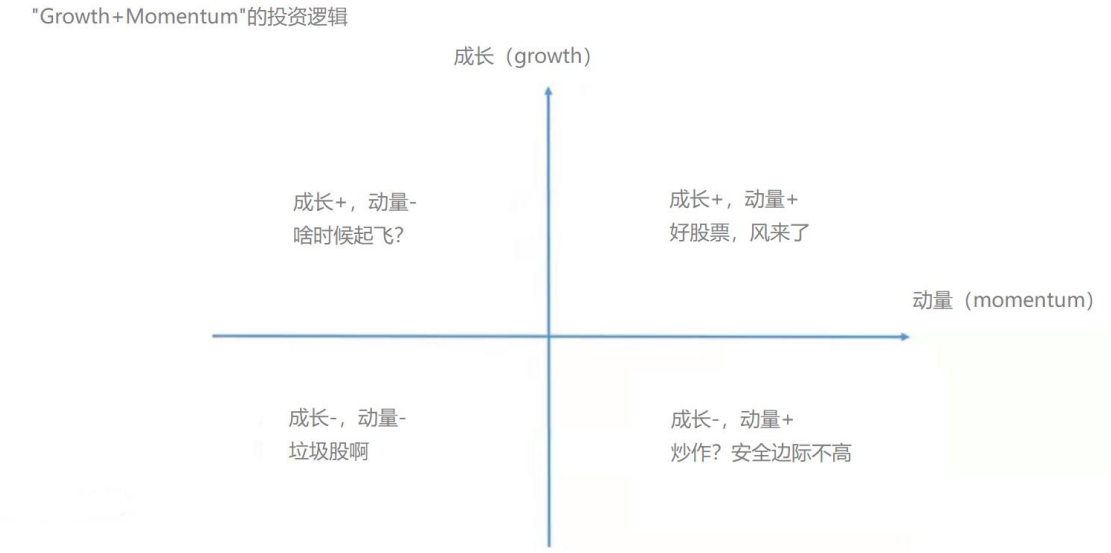
最后为了弄明白组合收益究竟来自哪些因子，而哪些因素又会对组合收益产生不利影响，我们需要对组合进行业绩归因。当然业绩归因的方法也有很多种，比如多因子模型归因，比如 Brinson 模型归因等等。最终业绩归因结果图如下：

	BP	growth_con_equal	liq_barra_con_equal_ortho	mom_con_equal_ortho
2014/12/31	0.574682534	-0.014165853	-0.038776225	0.028517975
2015/12/31	0.1239438	0.09308144	-0.124594326	-0.170318577
2016/12/31	0.232937333	0.165962562	-0.01765473	-3.63E-05
2017/12/31	-0.12933514	0.17482111	-0.141548291	0.26522974
2018/12/31	0.020735239	0.104092423	-0.034039444	0.130207726
2019/12/31	-0.50239649	0.221631929	0.063851554	0.190806277

通过业绩归因，让我们能够更加清楚组合的收益与风险来源，也能够对投资过程进行更详细的监控。

总结：

本模型以沪深 300 作为基准，最终选取动量因子、流动性因子、成长因子、估值因子等作为阿尔法因子，以波动率因子、Beta 因子、市值因子、行业因子作为风险因子构建指数增强策略，策略的基本投资逻辑如下图：



策略信息如下：

回测时间： 2014 年 1 月-2019 年 12 月；

基准指数： 沪深 300；

选股空间： 沪深 300 指数成份股（剔除 ST 股票、剔除调仓时停牌股票、剔除每月停牌超过 10 天的股票）；

交易成本：暂不考虑交易成本；

调仓频率：月频，每月第一个交易日；

调仓价格：收盘价；

行业及市值约束：行业、市值相对于基准的暴露为 0；

策略回测从 2014 年 1 月起，每个月初从沪深 300 成分股中选出约 60-100 只个股，构建投资组合，其年化超额收益为 8.3%，每一年都能跑赢指数，平均跟踪误差为 6.2%。当然这还仅仅只是基础的研究，所谓研究无止境，勇攀更高峰。这些年量化领域的竞争日趋激烈，多因子模型应该是其中最为成熟的方法论，它即是科学也是艺术，我们通过前人埋下的基石谨慎的前行，才能走的更远，走的更好。致广大而尽精微，极高明而道中庸。

未来计划：

1.增加基准指数，尝试中证 500，中证 800 等，在基准不变的情况下扩大选股空间，比如在全 A 进行选股。增加更多的调仓频率，比如增加周频调仓的功能。

2.秉承工匠精神，在每个步骤尝试更多的方法，尽力将每个细节做到最好，选择最能提高收益的方法，具体如下：

a) 因子去极值尝试对比更多方法，如 3σ法，百分位法，MAD 法，Beat G. Briner 方法、箱线图方法等。
b) 因子标准化包括 z-score 标准化与排名标准化。其中 z-score 标准化是将数据转换为标准正态分布，一定程度的保留了因子截面信息，但是容易受到极端值的影响。排名标准化将数据标准化成均匀分布，完全忽略了因子截面的距离信息，但是可以避免极端值的影响。
c) 单因子检验需要计算 IC，通常有两种：Pearson 相关系数（IC）与 Spearman 相关系数（Rank IC），其中 Pearson 相关系数考察的是两个变量之间的线性相关关系。Rank IC 是先排名再计算，考察的是两个变量之间的单调性关系。
d) 因子合成方式也有多种，如等权法、历史收益率(半衰)加权法、历史 IC(半衰)加权法、最大化 ICIR 加权法、最大化 IC 加权法、主成分分析法等。
e) 尝试更多的选股方法，比如随机森林，它相比于传统线性回归模型具有更直观，参数少，抗干扰，不易出现过拟合等优点。当然还可以尝试更多方法，如 SVM，AdaBoost，XGBoost 等。也可以改变预测目标，比如绝对收益率，相对收益率，排名，分类标签(上涨或下跌)等。
f) 尝试不同的优化组合方法：
   Risk parity：控制了风险，鲁棒性好，不考虑预期收益率
   Minimum variance：控制了风险，理论上更佳，不考虑预期收益率

Alpha Risk Parity: 盈利能力更强，但是不考虑相关性，风险配置更高
Black-Litterman: 贝叶斯思想，主观结合客观

3.挖掘更多有效因子，使用更多种类数据，如另类数据，如利用高频数据低频化等。还有比如资金流因子(大小单，基金持仓，机构投资者持仓，沪深通资金流)，市场情绪因子(分析师预期评级，散户行为，机构行为，社交媒体大数据分析)，事件类因子(财报超过或不及预期，财报预期调整公告，限售股解禁，高管增减持，股权质押，指数成份股变更)等等。