专题报告

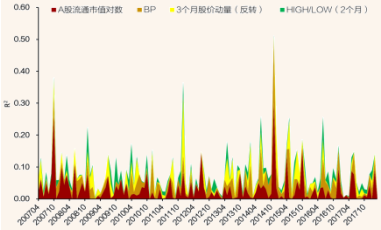
因子筛选与投资组合构建

2018 年 10 月 23 日

因子模型系列之十一

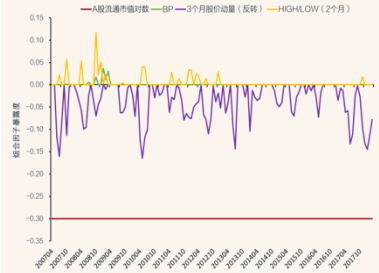
本文对单因子测试下的多个指标进行了主观权衡与分析，先初步筛选出 16 个比较有影响力的因子。然后以最大化模型整体解释力为原则，进一步精选因子，在已有的数据和全 A 股的样本空间下，暂定选定 4 个因子进入模型。通过逐层增量信息解释方式，统计了各因子的增量信息对于超额收益的解释能力。最后提了三种适合本框架的投资组合构建方法，并且着重介绍了其中的纯因子法和二次规划方法，并以中证 500 成分股作为股票池进行算例演示。

模型整体解释能力面积叠加图



资料来源：招商证券、Wind 资讯

二次规划法各因子暴露走势



资料来源：招商证券、Wind 资讯

- 在实际数据的测算中发现波动量能、最高累计收益和最高趋势收益、夏普值等各项收益指标关系密切，相互影响，这些指标有“此消彼长”的关系。
- 因子不同特性之间要做出权衡与妥协，本文试图综合考察因子的众多指标，来挑选适合策略的因子。
- 对各指标进行综合考察判断，初选 16 个因子。依据包括：首先，相关系数高的因子尽量不同时入选；然后，着重考虑波动量能大的因子；最后综合考察累积收益指标、Sharpe 值与因子背后的经济学逻辑。
- 在筛选因子的量化精选阶段，着眼于因子模型的整体解释能力，用少量因子尽可能提高模型的整体解释能力。即尽可能挑选那些能带来较大解释能力的因子。最后选出的因子有：A 股流通市值对数、BP、3 个月股价动量（反转）、HIGH/LOW（2 个月）。
- 我们用逐层增量解释的方法在对因子收益贡献能力进行剥离，使得因子对于超额收益的贡献划分更为明晰，后续进行因子调整的时候也更有针对性。
- 最后提了三种适用于本模型的构建投资组合的方法，着重介绍了纯因子法与二次规划方法，并且以算例进行解释。

相关报告

《各大类单因子有效性汇总比较分析》2018-4

《基于增量信息逐层解释的因子模型框架搭建》2017-11

叶涛
021-68407343
yetao@cmschina.com.cn
S1090514040002

研究助理
崔浩瀚
cuihaohan@cmschina.com.cn

正文目录

图表目录

因子模型提要

在前期的系列报告中，我们已经对 8 大类 78 个因子都做了细致的单因子收益比较。着重研究了与收益序列相关的各项指标（并以周报的形式持续跟踪），包括波动量能、最高累计收益和最高趋势收益、夏普值等；研究因子的超额暴露对于超额收益的解释力，展示了因子收益和截距项之间的关系；也对因子超额暴露原值和因子收益的相关性、自相关性进行了分析。

上述各项收益指标关系密切，相互影响，在实际数据的测算中发现这些指标有“此消彼长”的关系。比如，因子的波动量能描述的是若每期都对因子的方向做出正确判断的情况下，理论上在该因子上暴露所能获得的最高收益。累积收益展示的是固定因子暴露方向所能获得的最高收益。因子波动量能和该因子累积收益之间的差异空间反映的是对该因子择时的努力所能获得的回报。然而实际测算的结果是：往往波动量能和累积收益之间差异越大的因子，下一期个股暴露原值与之前几期的因子暴露原值相关性也越低，预测难度较大。

正是因为因子不同特性之间要做出权衡与妥协，本文试图综合考察因子的众多指标，来挑选适合策略的因子。

因子主观初选

因子模型是解释性模型，其最大功能是对解释超额收益所需的影响因素进行降维，从考察 n 只股票的多种影响因素降维成 m 个因子（m≪n）。我们前期考察的因子有 78 个，仍然是比较高维的数据，而且部分因子呈现出较为显著的相关，因而有必要对它们进行进一步降维，来缩小候选因子的数量。

由于我们前期已经对各个因子的收益数据和相关系数等进行了详细测算（详见前期报告），综合考察众多指标之后，我们重点关注以下 16 个因子，并将以下 16 个因子作为多因子模型的候选因子。

表 1：初步筛选得出候选因子

因子名称	因子代码	所属大类	构建说明
每股收益（EPS）增长率	Grow_eps_dilu	成长类	每股收益 12 个月的同比增长率
净利润增长率	Grow_net_prof	成长类	上市公司净利润同比增长率
速动比率	Liqu_quick	流通性类	Wind 数据库提取的速动比
总资产周转率（同比）	Liqu_yoy_asturn	流通性类	上市公司总资产周转率 12 个月的同比增长
销售净利率	Prof_netpmar	盈利类	净利润占销售收入的占比
ROE	Prof_roe_mrq	盈利类	净利润 TTM 与最近季度的股东权益的比值
ROE（同比）	Prof_yoy_roe_mrq	盈利类	ROE 的 12 个月同比增长
A 股流通市值对数	Size_ln_fltcap_a	规模类	A 股流通市值对数
3 个月股价动量（反转）	Tech_mtm_3m	技术指标类	个股最近 3 个月股价涨跌幅
24 日相对强弱指数	Tech_rsi_24d	技术指标类	个股最近 24 日相对强弱指数
BP	Valu_bp	估值类	市净率的倒数
EBITDA / EV	Valu_ebi2ev	估值类	企业价值倍数倒数，用 EBIDTA 除以企业价值
EP	Valu_ep	估值类	市盈率倒数

因子名称	因子代码	所属大类	构建说明
HIGH/LOW（2 个月）	Vola_high2low_2m	波动类	个股 2 个月最高价最低价比值的对数
1 个月价格波动	Vola_pricvola_1m	波动类	个股 1 个月最高价减去最低价的差和期初价的比
24 个月收益率标准差	Vola_std_dev_24m	波动类	个股 24 个月收益率标准差

资料来源：招商证券

选择以上 16 个因子，具体参考的依据为：

1. 相关系数高的因子尽量不同时入选。由于高相关系数（包括因子超额暴露原值的相关性和因子收益的相关性）的因子在解释力上有较大的重叠，出现在一个多因子模型中，边际解释能力并不显著，甚至还可能引起多重共线性的问题，从而会对模型的解释结果产生不良影响，因而候选的这 16 个因子两两之间并不存在很强的相关性。在实际数据测算中发现同一大类因子的相关性较强，在选择这 16 个因子的时候，同一大类因子不超过 3 个。

2. 着重考虑波动量能大的因子。因子的波动量能指的是在观测窗口内单因子单期收益的累积平方和。正如前面所说，因子的波动量能背后所包含的实际意义是：若每期都对因子的方向做出正确判断的情况下，理论上在该因子上暴露所能获得的最高收益。波动量能划定了因子理想收益的上限，实际测算数据中发现一部分因子的波动量能比较有限，对这些因子的预测的努力带来的回报也相应不高，可操作空间不大，因而我们认为没有必要将过多的精力放在波动量能不高的因子上，这类因子不应该包括在这 16 个因子中。

3. 综合考察累积收益指标、Sharpe 值等指标，以及因子背后的经济学逻辑。累积收益是指策略在确定因子方向后不再进行调整，因子在观测窗口内所能获得的最高收益，也是投资者关注最多的指标之一。结合因子收益的波动情况，综合考虑，确定候选池中的 16 个因子。

在确定这 16 个因子的过程中，我们依据前期有效性考察得到的重要数据指标，在综合考察的阶段，由于需要参照的判别依据多，且各指标有此消彼长的现象，难以完全量化，所以不可避免地，此步骤含有主观判断的成分。

模型因子量化精选——最大化整体解释度

因子模型随着因子数量的增加，边际解释能力的衰减是十分迅速的，模型中因子数量过多，并不一定能带来更好的整体解释度，我们认为入选多因子模型的因子个数无需太多，3 至 5 个足矣。

在前述步骤中，我们已经对波动量能、最高累计收益和最高趋势收益、夏普值等各项指标进行了考虑，在筛选因子的最后阶段，着眼于因子模型的整体解释能力，用少量因子尽可能提高模型的整体解释能力。即尽可能挑选那些能带来较大解释能力的因子。

我们先对两个因子进行组合，构成两因子模型，因子的组合一共有$C_{16}^{2}=120$种。我们对这 120 个模型进行遍历，做加权最小二乘回归，分别求解各个模型在每个回归截面上的R^2，对这 120 个模型的R^2均值由大到小进行排序，并统计每种组合在所有截面上出现最大R^2的频率。罗列R^2均值排名前 20 的两因子组合如下：

表 2：各两因子模型R^2指标

因子 1	因子 2	R^2均值	出现最大R^2频率
A 股流通市值对数	**BP**	**0.0549**	**5.83%**
A 股流通市值对数	3 个月股价动量（反转）	0.0517	5.83%
A 股流通市值对数	24 日相对强弱指数	0.0510	3.33%
A 股流通市值对数	HIGH/LOW（2 个月）	0.0498	4.17%
ROE	A 股流通市值对数	0.0464	1.67%
A 股流通市值对数	1 个月价格波动	0.0455	0.83%
速动比率	A 股流通市值对数	0.0445	5.00%
A 股流通市值对数	EP	0.0444	2.50%
销售净利率	A 股流通市值对数	0.0443	0.83%
A 股流通市值对数	24 个月收益率标准差	0.0439	3.33%
A 股流通市值对数	EBITDA / EV	0.0438	0.83%
3 个月股价动量（反转）	BP	0.0395	1.67%
BP	HIGH/LOW（2 个月）	0.0380	1.67%
24 日相对强弱指数	BP	0.0380	1.67%
每股收益（EPS）增长率	A 股流通市值对数	0.0378	1.67%
ROE	BP	0.0371	0.83%
速动比率	BP	0.0369	4.17%
净利润增长率	A 股流通市值对数	0.0368	0.00%
ROE（同比）	A 股流通市值对数	0.0367	1.67%
BP	EP	0.0359	0.00%

资料来源：招商证券、Wind 资讯

由于在我们的多因子模型中，在构建被解释变量的时候，已经剥离了市场组合对于超额收益的解释，即已经是经系统风险调整后的被解释变量。因而两因子情况下R^2均值最高到达 0.0549 是正常的，不算过低。

“A 股流通市值对数+BP”的组合R^2均值最高（0.0549），两因子组合下，能解释剩下的超额收益的部分也最多；紧随其后的是“A 股流通市值对数+3 个月股价动量（反转）”组合。以上者两种组合在全部截面上出现最大R^2的频率也最高。

在上述 20 个组合中，A 股流通市值对数这个因子出现在绝大部分的组合中，规模类的因子在 A 股市场上确实对超额收益的解释力非常强，这也是国内许多投资者非常关注该类因子的原因。

从我们的分析结果来看，模型中最先应该被确定的因子应当是 A 股流通市值对数和 BP。确定前两个因子之后，再确定第三个因子。用“A 股流通市值对数+BP+第 3 个因子”的模型进行加权最小二乘回归，遍历剩下的 14 个因子，同样罗列出各组合模型的R^2均值，并统计每种组合在所有截面上出现最大R^2的频率。

表 3：各三因子模型R^2指标

因子 1	因子 2	因子 3	R^2均值	出现最大R^2频率
A 股流通市值对数	**BP**	**3 个月股价动量**	**0.0682**	**10.83%**
		HIGH/LOW（2	0.0677	12.50%
		24 日相对强弱	0.0675	12.50%

因子 1	因子 2	因子 3	R^2均值	出现最大R^2频率
		ROE	0.0647	5.00%
		EP	0.0638	5.83%
		销售净利率	0.0635	3.33%
		24 个月收益率	0.0621	11.67%
		速动比率	0.0619	11.67%
		1 个月股价波动	0.0601	2.50%
		每股收益(EPS)	0.0596	8.33%
		ROE（同比）	0.0587	5.00%
		净利润增长率	0.0584	3.33%
		总资产周转率	0.0577	4.17%
		EBITDA / EV	0.0575	3.33%

资料来源：招商证券、Wind 资讯

在 A 股流通市值对数和 BP 两个因子的基础上加入第三个因子，“A 股流通市值对数+BP+3 个月股价动量（反转）”组合的R^2均值最大，为 0.0682；而“A 股流通市值对数+BP+HIGH/LOW（2 个月）”组合与“A 股流通市值对数+BP+24 日相对强弱指数”组合在观测窗口期出现最大R^2的频率最高。

这里我们暂定前三个因子为 A 股流通市值对数、BP 和 3 个月股价动量（反转）。

表 4：各四因子模型R^2指标

因子 1	因子 2	因子 3	因子 4	R^2均值	出现最大R^2频率
			HIGH/LOW	**0.0772**	**11.67%**
			ROE	0.0772	5.83%
			EP	0.0763	7.50%
			销售净利率	0.0760	4.17%
			24 日相对强	0.0759	11.67%
			速动比率	0.0738	15.00%
A 股流通市值对数	**BP**	**3 个月股价动量（反转）**	24 个月收益	0.0733	12.50%
			每股收益	0.0727	11.67%
			1 个月股价波	0.0726	2.50%
			ROE（同比）	0.0711	5.83%
			净利润增长	0.0710	2.50%
			EBITDA / EV	0.0706	3.33%
			总资产周转	0.0703	5.83%

资料来源：招商证券、Wind 资讯

用一样的方式，整理了四因子模型的整体解释度。“A 股流通市值对数+BP+3 个月股价动量（反转）+HIGH/LOW（2 个月）”组合的整体解释度最高。同时，随着因子数量的增加，R^2的边际增加值会迅速减少，继续追加因子并不会对整体解释度带来明显改善，反而会对后续构建投资组合带来更多的约束，因而我们不再追加因子，将因子模型的因子数量暂定为 4 个。分别是规模类因子中的 A 股流通市值对数，估值类因子中的 BP，技术指标类因子中的 3 个月股价动量（反转）以及波动类因子中 HIGH/LOW（2 个月）。

以上 4 个因子是在过去 10 年数据、以及全 A 股为样本时选定的因子，当时间窗口发生

改变或者样本空间发生改变时，不一定会选出上述 4 个因子，都可以依据实际情况进行调整。

逐层增量解释

前期在我们进行因子初选的时候，已经考虑过因子之间的相关性，相关性过高的因子没有同时入选到 16 个因子当中，然而因子与因子之间仍然会存在一定的相关性，此处，我们用逐层增量解释的方法在对因子收益贡献能力进行剥离，使得因子对于超额收益的贡献划分更为明晰，后续进行因子调整的时候也更有针对性。

因子模型的基本形式如下：

$$\left(\Delta r^{B,M}_{i,[t_0,t_1]}\right)_{n\times 1}=\left(c_{[t_0,t_1]}\right)_{n\times 1}+\left(\Delta\beta^{(j)}_{i,t_0}\right)_{n\times k}\cdot\left(r^{(j)}_{F,[t_0,t_1]}\right)_{k\times 1}+\left(\varepsilon_{i,[t_0,t_1]}\right)_{n\times 1}\cdots\cdots\text{式（1）}$$

式（1）中，$\Delta r^{B,M}_{i,[t_0,t_1]}$为被解释变量，表示经系统性风险调整后的超额收益；$\left(\Delta\beta^{(j)}_{i,t_0}\right)$为解释变量，表示因子的超额暴露；$r^{(j)}_{F,[t_0,t_1]}$为待估参数，表示的是因子收益（具体解释详见《基于增量信息逐层解释的因子模型框架搭建》）。逐层增量信息解释的具体步骤如下：

基于单因子有效性测试、因子暴露截面排序的稳定性等，确定优先排序(1)因子。第(1)层横截面回归与单因子有效性测试完全相同，被解释变量$y^{(1)}=\Delta r^{B,M}_{i,[t_0,t_1]}$，即因子模型的被解释变量；解释变量$x^{(1)}=\Delta\beta^{(1)}_{i,t_0}$，即因子(1)超额暴露的标准化赋值，待估计参数分别为$a^{(1)}$，$b^{(1)}$拟合优度$\gamma^{(1)}$代表因子(1)对因子模型整体解释度的贡献，残差项$Res^{(1)}=y^{(1)}-\tilde{a}^{(1)}-\tilde{b}^{(1)}\cdot x^{(1)}$代表未能被因子(1)解释的横截面股价表现差异。

从剩余有效单因子中挑选优先排序(2)因子，构造超额因子暴露原值$\Delta\tau^{(2)}_{i,t_0}$与$\Delta\tau^{(1)}_{i,t_0}$的辅助横截面回归，提取新入因子(2)所提供的增量解释信息，待估计参数分别为$c^{(1)}$，$d^{(1)}$，若辅助横截面回归的拟合优度R^2：

1. 过高，则说明当前测试因子不应作为优先排序(2)因子；
2. 过低，则可直接取用$\Delta\tau^{(2)}_{i,t_0}$生成对应的标准化赋值$\Delta\beta^{(2)}_{i,t_0}$；
3. 残差项$Net^{(2)}=\Delta\tau^{(2)}_{i,t_0}-\tilde{c}^{(1)}-\tilde{d}^{(1)}\cdot\Delta\tau^{(1)}_{i,t_0}$并生成对应的标准化赋值$\Delta\beta^{(2)}_{i,t_0}$。

输出结果$\Delta\beta^{(2)}_{i,t_0}$代表因子(2)相对于前序优先因子(1)所贡献的新增量解释信息。

第(2)层横截面回归中的解释变量$y^{(2)}=Res^{(1)}$，解释变量$x^{(2)}=\Delta\beta^{(2)}_{i,t_0}$，待估计参数分别为$a^{(2)}$，$b^{(2)}$，拟合优度$\gamma^{(2)}$，残差项$Res^{(2)}=Res^{(1)}-\tilde{a}^{(2)}-\tilde{b}^{(2)}\cdot x^{(2)}$，代表未能被因子(1)、因子(2)解释的横截面股价表现差异，依据回归显著性确定优先排序(2)因子，因子(2)对因子模型整体解释度的贡献为$\left(1-\gamma^{(1)}\right)\gamma^{(2)}$，因子(1)、因子(2)协同形成的因子模型整体解释度为$\gamma=1-\left(1-\gamma^{(1)}\right)\left(1-\gamma^{(2)}\right)$。

依次类推，第(k)层横截面回归用于挑选优先排序(k)因子。被解释变量$y^{(k)}=Res^{(k-1)}=\Delta r^{B,M}_{i,[t_0,t_1]}-\sum_{j=1}^{k-1}\tilde{a}^{(j)}-\sum_{j=1}^{k-1}\tilde{b}^{(j)}\cdot\Delta\beta^{(j)}_{i,t_0}$，代表所有前序优先因子$(1,2,\cdots,k-1)$未能解释的横截面股价表现差异。解释变量$x^{(k)}=\Delta\beta^{(k)}_{i,t_0}$为第$(k-1)$层辅助横截面回归残差项

$Net^{(k)}$对应的标准化赋值变量，代表因子(k)相对于所有前序优选因子所贡献的新增解释信息。$a^{(k)}$，$b^{(k)}$为第(k)层横截面回归的待估计参数。

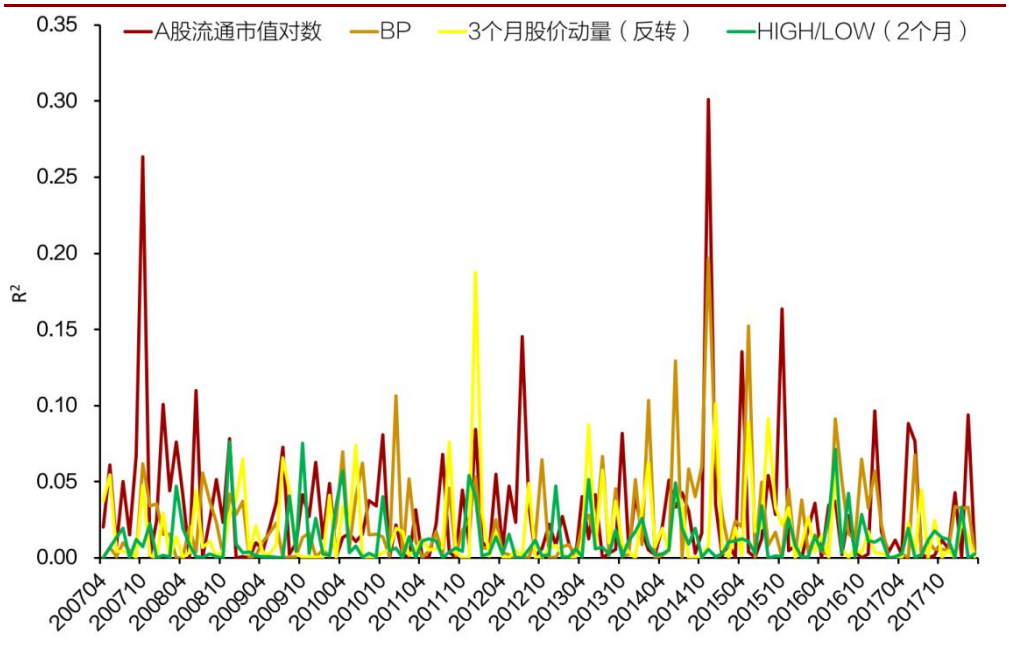
$\forall j = 1,2,\cdots,k-1$，$\tilde{a}^{(j)}$，$\tilde{b}^{(j)}$为在第$(j)$层横截面回归中已得到估计的参数，初值：$y^{(1)} = \Delta r_{i,[t_0,t_1]}^{B,M}$，$\tilde{a}^{(0)} = \tilde{b}^{(0)} = 0$。残差项$Res^{(k)}$代表未能被因子$(1,2,\cdots,k)$解释的横截面股价表现差异，作为第$(k+1)$层横截面回归的被解释变量，$y^{(k+1)} = Res^{(k)}$。

拟合优度$\gamma^{(k)}$，因子(k)对因子模型整体解释度的贡献为$\gamma = 1 - \prod_{j=1}^{k-1}(1-\gamma^{(j)})\gamma^{(k)}$，因子$(1,2,\cdots,k)$协同形成的因子模型整体解释度为$\gamma = 1 - \prod_{j=1}^{k}(1-\gamma^{(j)})$。依据回归显著性确定优先排序$(k)$因子，依据因子模型整体解释度$\gamma$确定纳入模型的因子个数$k$。各层横截面回归的待估计参数即为对应因子的收益估计，即$r_{F,[t_0,t_1]}^{(j)} = \tilde{b}^{(j)}$。

综上所述，因子逐层增量解释是以新纳入因子所提供的、前序优先因子未含有的新增解释信息对前序优先因子未能解释的股价表现差异进行解释，逐层堆积因子模型整体解释度，直至实现因子模型对截面股价表现差异的有效降维归纳。第$(k-1)$层辅助横截面回归用于提取因子(k)相对于前序优选因子$(1,2,\cdots,k-1)$所提供的新增解释信息，被解释变量为因子(k)的超额暴露原值$\Delta\tau_{i,t_0}^{(k)}$，解释变量为$Net^{(1)}, Net^{(2)}, \cdots, Net^{(k-1)}$，初值：$Net^{(0)} = (0)_{n\times 1}$，$Net^{(1)} = \Delta\tau_{i,t_0}^{(1)}$，输出结果为残差项$Net^{(k)}$，作为第$(k)$层辅助横截面回归新增的解释变量。

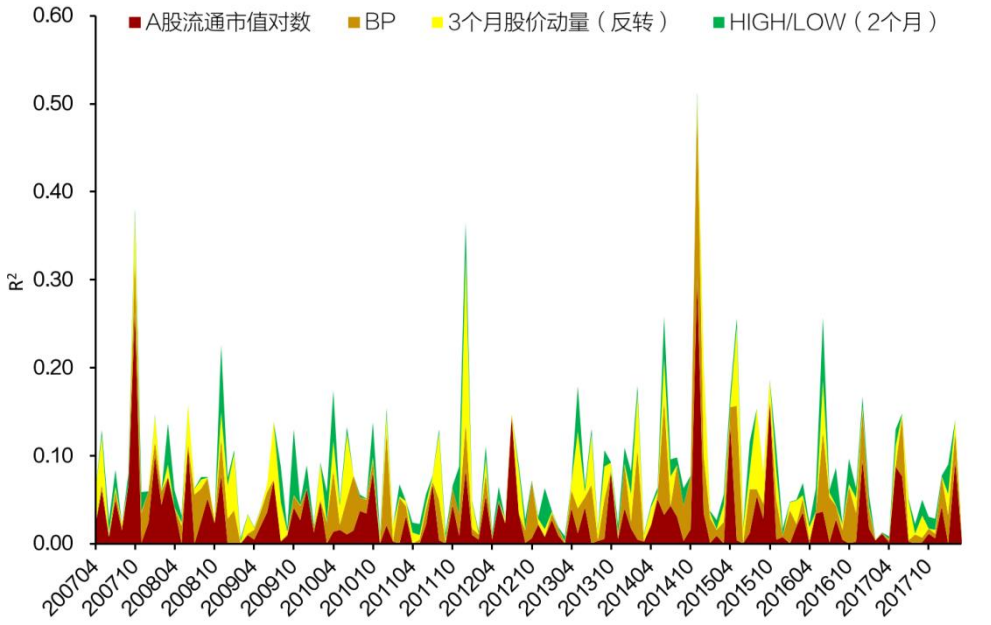
根据我们对于各个因子的测试，落实到实际模型上，依据解释度强弱，将 A 股流通市值对数作为第一个因子，BP 为放入模型的第二个因子，3 个月股价动量（反转）为第三个因子，HIGH/LOW（2 个月）为第四个，以上述方式进行各增量信息的求解，得到增量信息解释能力走势如下图所示：

图 1 各因子逐层增量解释能力走势图



资料来源：招商证券、Wind 资讯

图 2 各因子逐层增量解释能力面积叠加图



资料来源：招商证券、Wind 资讯

实际测算数据也说明，随着因子的增加，后面加入的因子增量信息对于超额收益的解释能力变得越来越有限。进一步增加因子个数，一来无法对模型整体解释能力进行显著提升，二来随着受控条件的增多，也会限制投资组合构建的灵活性。权衡利弊之下，我们模型暂定为 4 个因子。

因子投资组合构建

在本系列报告的开篇，我们已经强调过，因子模型最主要的作用是实现对个股股价表现差异的影响因素进行有效降维。根据对下期因子方向的判断来配置组合的因子目标暴露，而非直接用于选股。个股只是携带因子暴露配置信息的基础可交易载体，充当实现组合因子目标配置的填充材料。

因子模型本身只是一个解释性的模型，若只研究因子模型的内生变量，无法起到预测的作用，要对下期因子收益方向的预测以实现对下期因子暴露的配置需要研究模型以外的外生变量。通过对外生变量的研究，来对因子未来期方向和收益率进行判断，而后根据对收益的预期来确定组合在下一期的因子暴露值。对于比较有把握的因子，可以主动寻求在该因子上的暴露；若对某因子下期收益的方向没有把握，则将下期组合在该因子上的暴露调整为 0，以规避风险。

根据特定的因子目标暴露，来确定组合中各个股的权重的方法有很多。目前业界采用较多是类似 Fama E. F. & French K, R.在上世纪 90 年代初的做法，即根据个股按照在因子上的暴露从大到小进行排序（打分），做多排名靠前个股，做空排名靠后的个股，来建立多空组合。这种做法的优点是操作简便，易于理解；缺点是无法保证该投资组合对其他因子的因子暴露为 0。而且在正常情况下，在谋求组合在某个因子上暴露的同时不对其他因子加以控制的话，其他因子上的暴露必然也不为 0。

第二种方法是 Barra 在 2010 年提出的纯因子投资组合的方法。纯因子投资组合最初是为了正确量化因子的收益和风险而从纯数学的角度构建的。建立时没有考虑可投资性的要求，需要进一步用线性规划来确定个股权重。能使得因子业绩归因更具体化，但灵活度低，且在实际操作中，因子暴露和可投资性之间需要作出妥协。

第三种方法则是利用二次规划的方法直接解出个股在组合中权重。即根据目标暴露与实际投资条件建立个股权重的可行域，而后根据其他最优化目标（如换手率最小）来求出最优解。这方式相对于纯因子模型灵活性大幅提升，对个股暴露的控制十分清晰，但是较为复杂，因子暴露和可投资性之间也需要作出妥协。

以下，对第二种和第三种方法进行阐释，并以中证 500 成分股为例，进行算例演示。

纯因子组合方法构建投资组合

纯因子组合指的是在某因子上暴露度为 1 而在其他受控因子上暴露度为 0 的组合。求出纯因子组合之后再对纯因子组合进行权重配置，已达到组合因子暴露目标配置。

用纯因子方式进行投资组合构建实际上是分两步走。先解纯因子组合，再对纯因子组合进行权重配置，得到最后个股在组合中的权重。这种方法相对于第二种方法的简便之处在于：求因子收益的加权最小二乘求解过程中，已经给出了个股权重的解析解。且大幅减少了需要解线性规划确定的权重数量。

由于我们采用的横截面模型会存在异方差性，在前面的系列报告中我们采用加权最小二乘法（WLS）来求解待估参数，原模型等式两遍同乘以权重的平方根$\boldsymbol{D}$：

$$\boldsymbol{D^T r = D^T \beta f + D^T \mu}$$

其中r为个股经系统性风险调整后的超额收益，$\boldsymbol{\beta}$是个股在各因子上的暴露（可观测），$\boldsymbol{f}$为因子收益（待估参数），$\boldsymbol{D^T}$为权重的平方根，即$\boldsymbol{D^T D = \Omega}$。对于权重的选择有很多，Barra 采用的是市值的平方根，我们这里权重$\boldsymbol{\Omega}$（对角矩阵）设置如下：

$$\boldsymbol{\Omega} = \begin{bmatrix} \left(\tilde{\sigma}_1^2 T_1(t_0,t_1)\right)^{-1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \left(\tilde{\sigma}_n^2 T_n(t_0,t_1)\right)^{-1} \end{bmatrix}_{\boldsymbol{n\times n}}$$

$\tilde{\sigma}_n^2$为个股n在t_0截面上估计超额系统性风险暴露$\Delta\beta^M$时得到的随机误差，T(·) 为实际可交易天数（具体请参见因子系列报告一《基于增量信息逐层解释的因子模型框架搭建》）。

求解上述加权回归：

$$\boldsymbol{f = (\beta^T \Omega \beta)^{-1} \beta^T \Omega r}$$

可以解出无偏有效的待估参数，同时，上式也给出了纯因子组合的个股权重：

$$\boldsymbol{(\beta^T \Omega \beta)^{-1} \beta^T \Omega = W = \begin{bmatrix} w_1^T \\ \vdots \\ w_k^T \end{bmatrix}_{k\times 1}}$$

w_k^T即为因子k的纯因子组合中的个股权重。计算构建纯因子组合之后，再结合其他要求

对纯因子组合进行线性组合，就能达到我们需要的各因子的目标配置。

我们用中证 500 成分股为例，来对纯因子构建进行演示。

- 数据窗口从 2007 年 4 月至 2018 年 4 月。根据每日可计算条件，剔除中证 500 成分股存在股价异动和停牌的个股（具体请参见因子系列报告二《市值类因子有效性剖析》）。
- 提取个股在 A 股流通市值对数、BP、3 个月股价动量（反转）和 HIGH/LOW（2 个月）这四个因子上的因子暴露，并对异常值进行删失处理，对缺失值用序列均值进行赋值。
- 用分级靠档的方法对因子暴露值进行标准化赋值，将个股在各因子的暴露统一至 $[-0.5, 0.5]$。

而后用上述方法来构建纯因子组合。

各纯因子组合内部权重见附录；以下罗列各纯因子组合在各因子上的暴露：

表 5：纯因子组合在各因子上的暴露（2018 年 3 月截面）

纯因子组合	A 股流通市值对数	BP	3 个月股价动量	HIGH/LOW
A 股流通市值对数组合	1.0	5.2692E-17	3.9682E-17	2.4720E-17
BP 组合	4.7705E-17	1.0	4.2501E-17	-4.6838E-17
3 个月股价动量组合	-1.0235E-16	3.6971E-17	1.0	0.0000E+00
HIGH/LOW 组合	-1.1579E-16	7.0365E-17	-2.4340E-17	1.0

资料来源：招商证券、Wind 资讯

各纯因子组合在对应的因子上的暴露度为 1，在其他因子上的暴露度接近于 0。对上述纯因子组合在进行线性组合就能得到我们所需要的各因子的目标配置。

纯因子线性组合的情况下，全额投资条件是自动满足的。在实际操做中，需要添加一些必要的约束条件，比如在不允许做空的市场中，还应该满足：

$$\boldsymbol{w_c^T} + v_{flt}\boldsymbol{w_{flt}^T} + v_{bp}\boldsymbol{w_{bp}^T} + v_{m3m}\boldsymbol{w_{m3m}^T} + v_{h2l2}\boldsymbol{w_{h2l2}^T} \geqslant \boldsymbol{0}$$

$\boldsymbol{w_c}$表示常数项纯因子组合个股权重，$\boldsymbol{w_{flt}}$表示 A 股流通市值对数纯因子组合个股权重，$\boldsymbol{w_{bp}}$表示 BP 纯因子组合个股权重，$\boldsymbol{w_{m3m}}$表示三个月股价动量（反转）纯因子组合个股权重，$\boldsymbol{w_{h2l2}}$表示 HIGH/LOW（2 个月）纯因子组合个股权重。$v_{flt}, v_{bp}, v_{m3m}, v_{h2l2}$分别表示投资组合在对应因子上的暴露（可以根据暴露要求设置）。

算例中假设要使得 A 股流通市值因子在组合中的暴露负向最大化，并将其他因子作为受控因子，使其暴露度为 0。则目标函数和约束条件为：

$$\boldsymbol{minimize} \qquad v_{flt} \qquad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\text{式（2）}$$

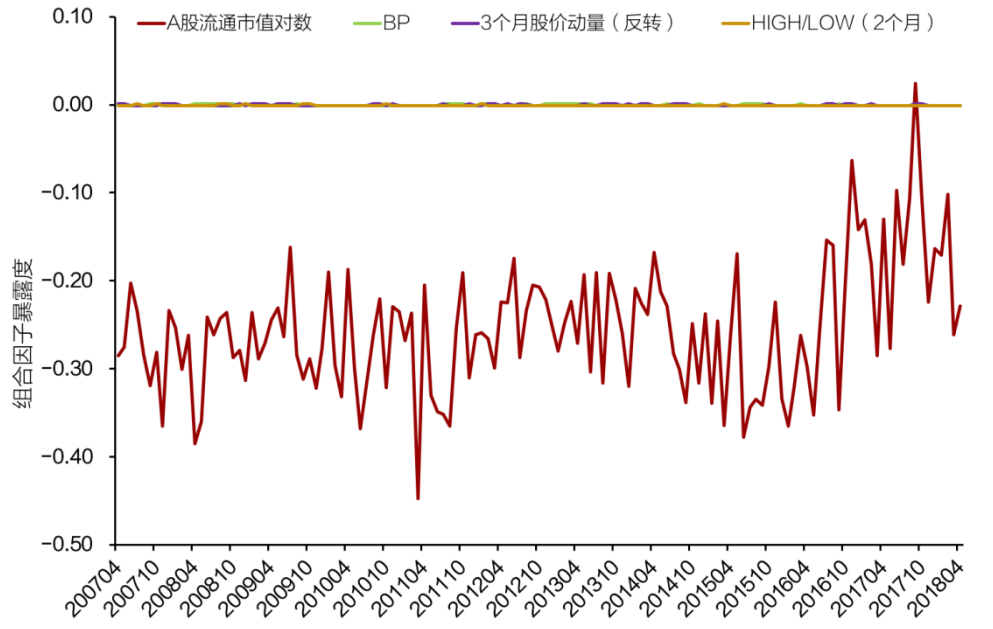
$$\boldsymbol{subject\ to} \quad \boldsymbol{w_c^T} + v_{flt}\boldsymbol{w_{flt}^T} + v_{bp}\boldsymbol{w_{bp}^T} + v_{m3m}\boldsymbol{w_{m3m}^T} + v_{h2l2}\boldsymbol{w_{h2l2}^T} \geqslant -10^{-2}$$

$$-10^{-3} \leqslant v_{bp}, v_{m3m}, v_{h2l2} \leqslant 10^{-3}$$

Other constraints

由于纯因子组合是从纯数学的角度去考虑问题的，所以实际操作中，需要对约束条件做一些妥协，以保证能得到数值解。比如不允许做空的条件修改成$\geqslant -10^{-2}$，受控因子的暴露度不能严格设置为 0，而是要设置一定的变动空间$[-10^{-3}, 10^{-3}]$。即便如此，最终组合的因子暴露仍然不能达到理论上的最大化：

图 3 投资组合在各因子上的暴露走势（纯因子组合法）



资料来源：招商证券、Wind 资讯

BP、3 个月股价动量（反转）和 HIGH/LOW（2 个月）这三个受控因子的暴露基本为 0，符合我们们的要求。A 股流通市值对数这个因子在 2017 年之前可以给出-0.3 左右的暴露度，但是在 2017 年后，因子之间相关度上升，越来越难以配出负向的暴露度，甚至在有些月份只能配置正向暴露。

二次规划方法构建投资组合

相对而言，二次规划方法是比纯因子法更灵活的组合构建方式。其灵活性主要体现在：

1. 二次规划方法不像纯因子法局部限制了各股票之间相对权重，因此可行域更大，在

配置目标暴露的时候更能接近理想的暴露值。这使得组合的可投资性显著提升。

2. 二次规划可以设置除目标因子暴露以外的优化目标，根据投资的实际需求来设置优化对象和约束条件。

这里我们假设需要在配齐投资组合在各因子上的暴露的前提下，使得换手率最低，以减少交易成本。以矩阵形式表述：

$$\boldsymbol{minimize} \quad \|\boldsymbol{w_p} - \boldsymbol{w_0}\|^2 \cdots\cdots\cdots\cdots \text{式（3）}$$

$$\boldsymbol{subject\ to} \quad \boldsymbol{f^T} \cdot \boldsymbol{w_p} = \boldsymbol{X}$$

$$\boldsymbol{w_p} \geqslant \boldsymbol{0}$$

$$\boldsymbol{D^T} \cdot \boldsymbol{w_p} = 1$$

其中，$\boldsymbol{w_p}$是需要求解的本期个股权重向量；$\boldsymbol{w_0}$是上期组合中个股的权重向量（已知），求解第一期的$\boldsymbol{w_p}$时，$\boldsymbol{w_0}$等权配置；$\boldsymbol{f^T}$为个股在各因子上的暴露；$\boldsymbol{X}$为组合的在各因子上的暴露目标配置。$\boldsymbol{D^T}$为元素均为 1 的行向量，$\boldsymbol{D^T} \cdot \boldsymbol{w_1} = 1$代表的是全额投资，若非全额投资，可以修改等式右边参数。

同样，我们用中证 500 成分股为例，来对二次规划组合构建进行演示。同样先对个股数据进行每日可计算条件筛选，提取个股在 A 股流通市值对数（用X^{flt}表示）、BP（用X^{bp}表示）、3 个月股价动量（反转）（用X^{m3m}表示）和 HIGH/LOW（2 个月）（用X^{h2l2}表示）这四个因子上的因子暴露，进行删失和标准化处理。用 Python 求解二次规划的模块 cvxopt 计算个股权重。

式（3）是理想情况下的求解公式，在实际操作中，理想的限制条件下无法解出最优解，需要放宽限制、扩大可行域。将组合的目标暴露配置从常数扩展为$[\boldsymbol{X_D}, \boldsymbol{X_U}]$可行区间。即：

$$\boldsymbol{minimize} \quad \|\boldsymbol{w_p} - \boldsymbol{w_0}\|^2 \cdots\cdots\cdots\cdots \text{式（4）}$$

$$\boldsymbol{subject\ to} \quad \boldsymbol{f^T} \cdot \boldsymbol{w_p} \leqslant \boldsymbol{X_U}$$

$$\boldsymbol{f^T} \cdot \boldsymbol{w_p} \geqslant \boldsymbol{X_D}$$

$$\boldsymbol{w_p} \geqslant \boldsymbol{0}$$

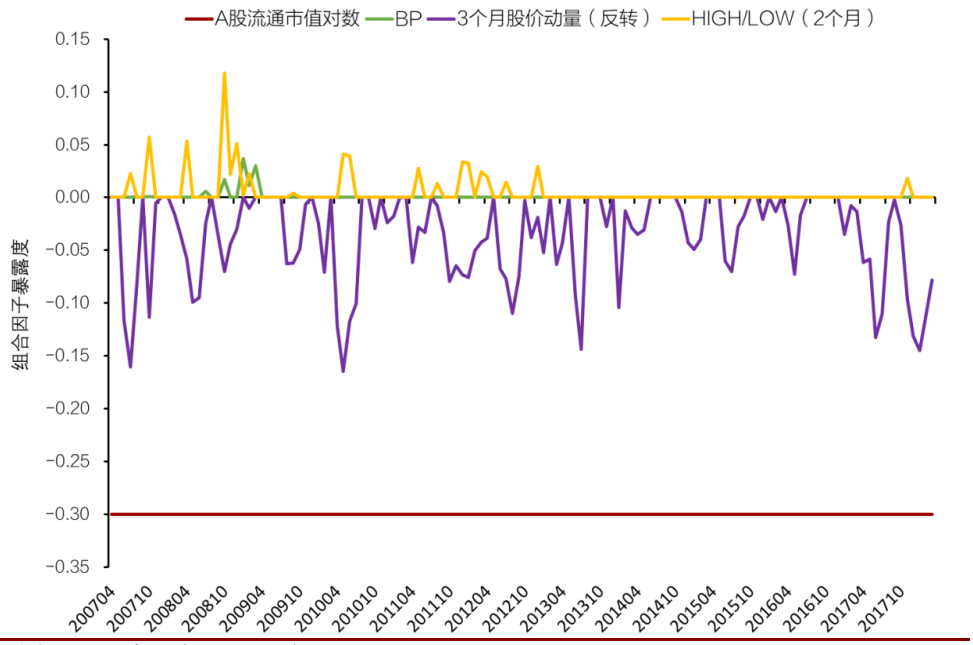
$$\boldsymbol{D^T} \cdot \boldsymbol{w_p} = 1$$

这里放宽的主要是组合在因子上的暴露度，从定值改成有上下限的可行变动范围。因子

模型本身是解释性的模型，本身不具备因子方向的预测功能。此处本文只做算例上的演示，来说明二次规划方法的操作流程，因此根据经验数据设置组合因子目标暴露。

设置$X_D^{flt} \in [-0.5,-0.3]$, $X_D^{bp} \in [0,0.5]$, $X_D^{m3m} \in [-0.5,0]$, $X_D^{h2l2} \in [0,0.5]$，代入式(4)。计算得到组合每期在各因子上的暴露如下图所示：

图 4 投资组合在各因子上的暴露走势（二次规划法）



资料来源：招商证券、Wind 资讯

为了保证能在每个截面上计算出数值解，我们在算例中设定的取值范围是比较宽的。在实际操作中，可以根据各时点上的具体要求来针对性地设置因子的目标暴露可选范围。

根据算例要求，范围设置得较精细的为 A 股流通市值对数因子，在实际测算中，基本能给出接近-0.3 的暴露度，基本满足要求。另外三个因子我们只做了方向上的约束，其中 3 个月股价动量（反转）因子在整个测试期内大部分时间能给出一个负向的暴露，另两个因子能在少部分时间给出正向暴露。

从可投资性和因子暴露的角度来说，二次规划方法比纯因子法已经有了较大的改善，至少不会出现与预期配置暴露方向相反的情况。

总结

本文基本完结了我们对于因子模型框架的构建。

在前面的报告中，我们提出了因子独具特色的因子模型的搭建方式，在该因子模型框架下，进行了单因子多个指标的测算，并进行了汇总比较。

本文更是对测算过的多个指标进行主观权衡，初步选择了 16 个因子；而后用最大化整体解释度原则对因子再进行量化精选，在现有情况与数据下，确定模型因子。最后提了三种适用于本模型的构建投资组合的方法，着重介绍了纯因子法与二次规划方法，以算

例进行解释。

后续因子系列的研究，将不断完善我们的模型架构，并逐步将重心转向对于因子择时的研究。

附录

表 6：中证 500 纯因子组合 2018 年 3 月份截面个股权重（股票代码排名前 10 只）

股票代码	股票简称	常数项	A 股流通市值	BP 权重	3 个月股价	HIGH/LOW
000006.SZ	深振业 A	0.0020	0.0048	-0.0023	-0.0307	0.0123
000009.SZ	中国宝安	0.0014	0.0105	-0.0098	-0.0234	0.0033
000012.SZ	南玻 A	0.0025	0.0071	-0.0080	-0.0164	-0.0037
000021.SZ	深科技	0.0006	0.0024	-0.0014	-0.0033	-0.0001
000025.SZ	特力 A	0.0014	-0.0034	-0.0026	-0.0015	0.0113
000027.SZ	深圳能源	0.0061	0.0410	0.0331	0.0131	-0.0236
000028.SZ	国药一致	0.0008	0.0136	-0.0080	0.0016	-0.0099
000039.SZ	中集集团	0.0029	0.0245	0.0014	-0.0425	0.0223
000049.SZ	德赛电池	0.0007	-0.0059	-0.0148	-0.0151	-0.0018
000050.SZ	深天马 A	0.0011	0.0225	-0.0068	-0.0250	0.0028

资料来源：招商证券、Wind 资讯

表 7：二次规划组合 2018 年 3 月截面个股权重（权重排名前 10 只）

股票代码	股票简称	个股权重
000816.SZ	*ST 慧业	0.0355
600939.SH	重庆建工	0.0354
601811.SH	新华文轩	0.0350
603528.SH	多伦科技	0.0341
600996.SH	贵广网络	0.0324
603355.SH	莱克电气	0.0317
000572.SZ	海马汽车	0.0315
002392.SZ	北京利尔	0.0314
000727.SZ	华东科技	0.0278
002818.SZ	富森美	0.0276

资料来源：招商证券、Wind 资讯

表 7 中有一只 ST 股入选，原因是在 2018 年 3 月份时还不是特别处理股票，也是中证 500 成分股。到今年 6 月份的时候才从成分股中剔除。若想避免这种情况发生，可以先对股票池进行筛选，剔除财务状况不好与市值过小的个股，就不会出现选出 ST 股的情况了。

分析师承诺

负责本研究报告的每一位证券分析师，在此申明，本报告清晰、准确地反映了分析师本人的研究观点。本人薪酬的任何部分过去不曾与、现在不与，未来也将不会与本报告中的具体推荐或观点直接或间接相关。

叶涛：首席分析师。上海交通大学管理学硕士，2005 年起从事金融工程研究，曾先后任职于易方达基金机构投资部、上投摩根基金研究部、申万菱信基金投资管理总部、长江证券研究部、广发证券发展研究中心，2014 年 3 月加盟招商证券研究发展中心。

崔浩瀚：研究助理。浙江大学经济学硕士，2017 年 7 月加盟招商证券研究发展中心金融工程组。

投资评级定义

公司短期评级

以报告日起 6 个月内，公司股价相对同期市场基准（沪深 300 指数）的表现为标准：

强烈推荐：公司股价涨幅超基准指数 20%以上
审慎推荐：公司股价涨幅超基准指数 5-20%之间
中性：公司股价变动幅度相对基准指数介于±5%之间
回避：公司股价表现弱于基准指数 5%以上

公司长期评级

A：公司长期竞争力高于行业平均水平
B：公司长期竞争力与行业平均水平一致
C：公司长期竞争力低于行业平均水平

行业投资评级

以报告日起 6 个月内，行业指数相对于同期市场基准（沪深 300 指数）的表现为标准：

推荐：行业基本面向好，行业指数将跑赢基准指数
中性：行业基本面稳定，行业指数跟随基准指数
回避：行业基本面向淡，行业指数将跑输基准指数

重要声明

本报告由招商证券股份有限公司（以下简称“本公司”）编制。本公司具有中国证监会许可的证券投资咨询业务资格。本报告基于合法取得的信息，但本公司对这些信息的准确性和完整性不作任何保证。本报告所包含的分析基于各种假设，不同假设可能导致分析结果出现重大不同。报告中的内容和意见仅供参考，并不构成对所述证券买卖的出价，在任何情况下，本报告中的信息或所表述的意见并不构成对任何人的投资建议。除法律或规则规定必须承担的责任外，本公司及其雇员不对使用本报告及其内容所引发的任何直接或间接损失负任何责任。本公司或关联机构可能会持有报告中所提到的公司所发行的证券头寸并进行交易，还可能为这些公司提供或争取提供投资银行业务服务。客户应当考虑到本公司可能存在可能影响本报告客观性的利益冲突。