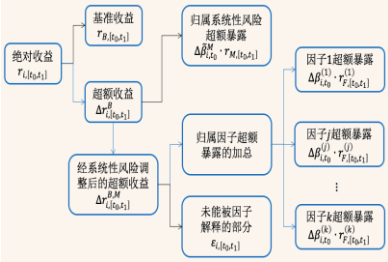
专题报告

基于增量信息逐层解释的因子模型框架搭建

2017 年 11 月 22 日

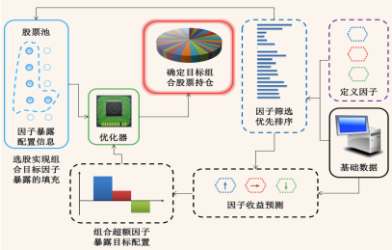
因子模型系列之一

超额收益的分解



资料来源：招商证券

因子模型搭建与优化过程示意



资料来源：招商证券

相关报告

1、《基于同质性分析的市场及风格描述》2016-10

2、《基于同质性分析的行业特征研究》2016-08

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

研究助理
崔浩瀚
cuihaohan@cmschina.com.cn

本篇报告是招商金工因子模型系列的第一篇报告。报告系统地阐释了招商多因子模型的理论基础并介绍了模型的框架设计。同时比较了横截面模型和时间序列模型之间的异同和各自的优缺点，强调选因子而非选股的构建理念。随后细致描述了后续研究预计会使用到的基础数据收集和整理方法、单因子检验步骤、逐层增量解释的多因子检验、因子暴露目标配置、股票组合目标持仓和绩效归因与模型调整的方法。提纲挈领地为后续的因子系列研究做了铺垫。

- **本报告在经典学术理论的基础上，试图以新的视角审视因子模型，并提出独具特色的因子模型构建方法。**重新回顾了经典均衡模型的原理，较为仔细地区分了均衡模型和统计模型的差异，比较了因子模中时间序列模型和横截面模型的异同。
- **阐释了因子模型选因子而非选股票的理念，将个股仅仅看作是因子视角下携带因子暴露配置信息的基础可交易载体。**通过单因子测试、多因子排列来选出最优的因子配置方案，而能配置出最优因子方案的个股组合并非只有一种。我们的因子模型始终以个股因子暴露对基准的偏离来解释超额收益。
- **较为详实地阐述了因子模型系列研究所要做的工作。**具体包括基础数据收集和整理方法、单因子检验步骤、逐层增量解释的多因子检验、因子暴露目标配置、股票组合目标持仓和绩效归因与模型调整的方法。
- **报告在多因子组合排序阶段提出了逐层增量解释的方法，该方法为多因子组合提供了新的思路。**采用逐层增量解释的意义在于明确超额收益的来源，通过因子对组合超额收益（一阶）、股价差异整体解释度（二阶）形成对模型可纳入因子明确归因的可加形式。有利于后续的有目的性的改善因子组合。
- **提纲挈领的为后续的研究做了奠基工作。**本篇报告是招商金工多因子系列的第一篇报告。后续的系列报告将会按照本报告中提到的构建方法构建因子模型，并根据实际的数据测算的情况，可能会在某些方面做出必要的调整，以期形成一套较为完整的因子策略。

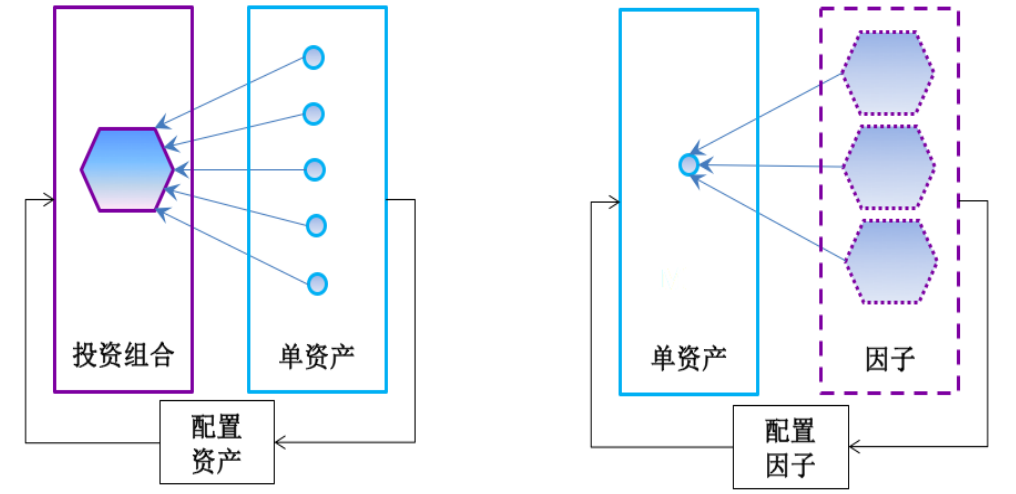
正文目录

图表目录

模型理论基础

从因子模型的双视角来看，投资组合可以看作是单资产的线性组合，将单资产作为基础可交易资产，对单资产的配置构成投资组合；同时，又可将单资产视为一系列因子的线性组合，因子是独立的定价要素，对因子进行配置合成单资产。

图 1 传统投资组合视角与因子视角



资料来源：招商证券

经典均衡模型

资本资产定价模型

马科维茨(Harry Markowitz，1952)的分散投资与效率组合投资理论第一次以严谨的数理工具为手段向人们展示了一个风险厌恶的投资者在众多风险资产中如何构建最优资产组合的方法。但是由于根据马科维茨的模型来构建最优投资组合所需要计算的协方差数量将是一个天文数字，受限于当时的计算机计算能力，马科维茨理论难以运用到实践之中。

在此背景下，以夏普(William F. Sharpe，1964)为代表的一些经济学家开始从实践的角度出发，提出了资本资产定价模型（Capital Asset Pricing Model，CAPM）。资本资产定价模型的一般形式为：

$$E[R_S] - r_f = \beta_S^M \left(E[R_M] - r_f\right) \qquad \text{式（1）}$$

其中，$\mathrm{E}[R_S]$表示投资组合S的期望收益，$E[R_M]$代表市场组合M的期望收益。令$\omega_S^M = \beta_S^M, \omega_S^f = 1 - \beta_S^M$，式（1）可以写成：

$$E[R_S] = \omega_S^M \cdot E[R_M] + \omega_S^f r_f \qquad \text{式（2）}$$

式（2）所代表的现实含义是，在均衡条件下，投资者所持有的资产S中，市场组合的配置权重为ω_S^M，无风险资产的配置权重为ω_S^f。

套利定价理论

另一个经典理论是套利定价理论(Arbitrage Pricing Theory，APT)，由套利定价理论得出的模型和资本资产定价模型一样，都是均衡模型。套利定价理论的思想是，资产的价格会受到多个因素的影响，这些因素会影响资产的风险，从而改变资产的收益。套利定价理论可以由以下数学形式给出：

$$E[R_S] - r_f = \sum_{j=1}^{k} \beta_S^j \left(E\left[R_F^j\right] - r_f\right) \qquad \text{式（3）}$$

其中，$E[R_S]$表示资产S的期望收益，$E\left[R_F^j\right]$代表纯因素组合j的期望收益，令$\omega_S^j = \beta_S^j$，$\omega_S^j = 1 - \sum_{j=1}^{k} \beta_S^j$，式（3）可以写成：

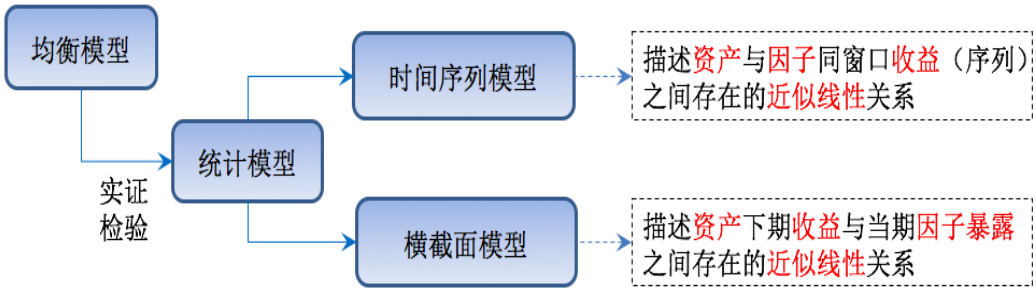
$$E[R_S] = \sum_{j=1}^{k} \omega_S^j \cdot E\left[R_F^j\right] + \omega_S^f r_f \qquad \text{式（4）}$$

式（4）的现实含义是当市场达到均衡时，投资者将持有一组纯因素组合与无风险资产，其中每项纯因素组合配置权重为ω_S^j，$j = 1,2,\cdots,k$，共配置$\sum_{j=1}^{k}\left(\omega_S^j \cdot \text{纯因素组合}j\right)$，并持有权重为$\omega_S^f$的无风险资产。

统计模型

资本资产定价模型和套利定价理论得出的模型都是均衡模型，描述了资产预期风险与预期收益率之间存在的线性关系。均衡模型构建了一种理想的状态，在这种理想状态下，每个投资者的行为都可以被预测。然而这两个均衡模型的假设条件在现实的资本市场中基本不可能实现，因此利用均衡模型在直接面向“未来、预期”的实证检验中存在困难。在实证中，需要转换为统计模型，来描述资产价格的决定因素，基于历史数据对均衡模型进行实证检验。

图 2 均衡模型与统计模型



资料来源：招商证券

统计模型包括两类基础模型形式：时间序列模型和横截面模型。

时间序列模型以资产收益序列作为被解释变量，描述资产与因子同窗口收益序列之间存在的近似线性关系，其一般形式可以写成：

$$r_{S,t} = \alpha_S + \sum_k \beta_S^k \cdot r_{k,t} + \varepsilon_{S,t} \qquad \text{式（5）}$$

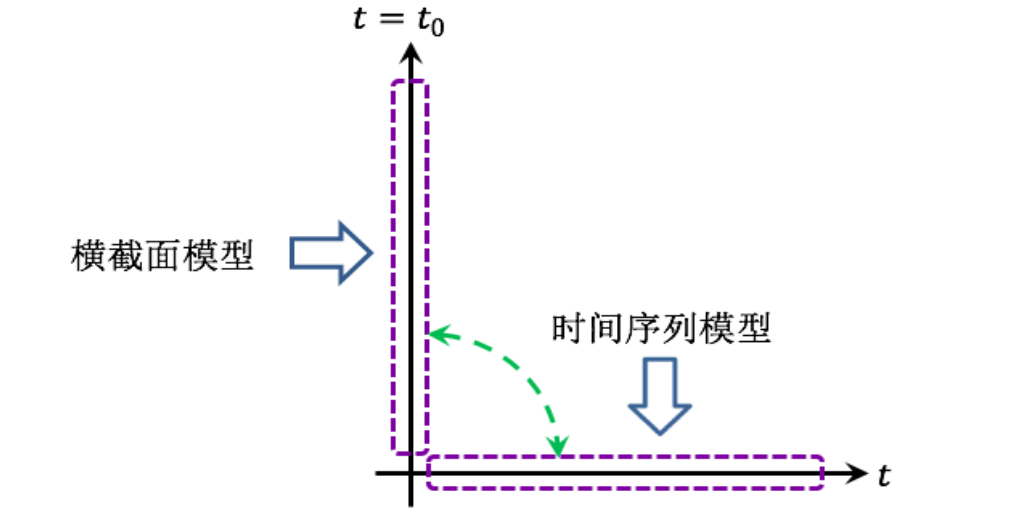
在时间序列模型中，因子k的收益序列$r_{k,t}$可观测，作为解释变量，而资产S对因子k在序列观测窗口内平均的暴露β_S^k需要估计。

而横截面模型以同期多资产收益作为被解释变量，描述资产下期收益与当期因子暴露之间存在的近似线性关系。其一般形式可以表述为：

$$\left(r_{i,[t_0,t_1]}\right)_{n\times 1}=\left(c_{[t_0,t_1]}\right)_{n\times 1}+\left(\beta_{i,t_0}^{j}\right)_{n\times k}\cdot\left(r_{F,[t_0,t_1]}^{j}\right)_{k\times 1}+\left(\varepsilon_{i,[t_0,t_1]}\right)_{n\times 1}\quad\text{式（6）}$$

在横截面模型中，资产i对因子j的当期暴露β_{i,t_0}^{j}可观测，作为解释变量；因子j的下期收益$r_{F,[t_0,t_1]}^{j}$需要估计。

图 3 两类基础统计模型在建模方向上的垂直关系



资料来源：招商证券

时间序列模型与横截面模型存在以下两点共性。第一，这两类模型都给出了资产收益、波动(收益率方差)分解的线性、可加形式；第二，这两类模型都要求解释变量互不相关，即正交化。

同时，这两个模型的区别也很明显。时间序列模型描述资产与因子收益、波动率的联动，因子的设定需有已存在的载体(用于计算因子收益序列)。由于模型采用滚动窗口估计方法，因此时间序列模型对资产价格波动的解释相对迟缓。在搭建模型时，时间序列的解释变量的数据处理相对容易，但是存在回归残差的自相关性。反观横截面模型，它描述不同资产收益差异和因子暴露差异之间的关联，因子设定可以比较灵活，由于横截面模型用多资产方式来捕捉解释变量的变异度，采用单片截面滚动估计，对不同资产价格波动差异的解释更为及时。相对于时间序列模型来说，横截面模型的解释变量数据处理较为复杂，同时在参数估计时，由于不同个股波动率结构有差异，而存在异方差性，在后续的估计中，需要对异方差性进行处理。

模型框架

横截面因子模型一般形式

基于上述分析的模型特性，我们选择横截面模型作为我们因子模型系列的估计模型，具体形式可以表述为：

$$\left(\Delta r_{i,[t_0,t_1]}^{B,M}\right)_{n\times 1}=\left(c_{[t_0,t_1]}\right)_{n\times 1}+\left(\Delta\beta_{i,t_0}^{(j)}\right)_{n\times k}\cdot\left(r_{F,[t_0,t_1]}^{(j)}\right)_{k\times 1}+\left(\varepsilon_{i,[t_0,t_1]}\right)_{n\times 1}\quad\text{式（7）}$$

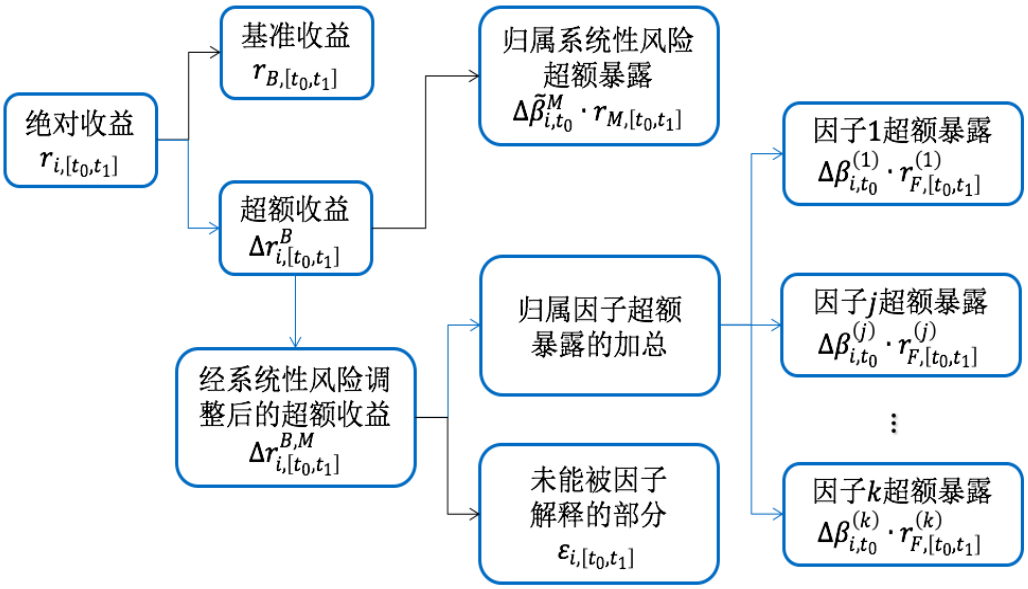
式（7）左侧，被解释变量为个股$i\,(i=1,2,\cdots n)$在时间窗口$[t_0,t_1]$内经系统性风险暴露调整后的超额收益，这一超额收益序列数据可由式（8）获得：

$$\Delta r_{i,[t_0,t_1]}^{B,M}=\Delta r_{i,[t_0,t_1]}^{B}-\Delta\tilde{\beta}_{i,t\le t_0}^{M}\cdot r_{M,[t_0,t_1]}\quad\text{式（8）}$$

式（8）右侧$\Delta r^{B}_{i,[t_0,t_1]}$代表个股$i$在$[t_0,t_1]$内的超额收益，$B$表示投资基准，即$\Delta r^{B}_{i,[t_0,t_1]}=r_{i,[t_0,t_1]}-r_{B,[t_0,t_1]}$。$\Delta\tilde{\beta}^{M}_{i,t\le t_0}$代表个股 i 在t_0截面的超额系统性风险暴露，具体地，$\Delta\tilde{\beta}^{M}_{i,t\le t_0}=\tilde{\beta}^{M}_{i,t\le t_0}-\tilde{\beta}^{M}_{B,t\le t_0}$，$M$表示市场组合，$\tilde{\beta}^{M}_{i,t\le t_0}$，$\tilde{\beta}^{M}_{B,t\le t_0}$为以$t_0$截面为终点的区间序列估计值。同时，由于是横截面模型，被解释变量会存在截面异方差，即$\varepsilon_{i,[t_0,t_1]}\sim N\left(0,\sigma^2\left[\varepsilon_{i,[t_0,t_1]}\right]\right)$。由于异方差的存在，需要在后续的估计过程中进行调整修正。

式（7）右侧，j为因子被模型纳入的优先排序值，$j=1,2,3,\cdots,k$，因子j在$[t_0,t_1]$中的收益$r^{(j)}_{F,[t_0,t_1]}$为模型待估参数，解释变量为个股i在t_0截面对因子(j)的超额暴露$\Delta\beta^{(j)}_{i,t_0}$。这里的因子暴露需要与基准进行对标，使得解释变量与被解释变量均具有相同的基点，即$\Delta\beta^{(j)}_{B,t_0}=0\Leftrightarrow\Delta r^{B,M}_{B,[t_0,t_1]}=\Delta\bar{r}^{B,M}_{i\in B,[t_0,t_1]}=0$，同时，因子的超额暴露需要满足标准化赋值要求，即$\forall j\le k$，均有$max_i\left\{\Delta\beta^{(j)}_{i,t_0}\right\}-min_i\left\{\Delta\beta^{(j)}_{i,t_0}\right\}=1$。此外，我们的因子模型采用“逐层增量解释”的方法来添加新因子，因而$\forall 2\le j\le k$，$\Delta\beta^{(j)}_{i,t_0}$均代表前序优先变量$\Delta\beta^{(1)}_{i,t_0},\cdots,\Delta\beta^{(j-1)}_{i,t_0}$未包含的新增解释信息，即$\forall l\ne m\le k$均有$\left(\Delta\beta^{(l)}_{i,t_0}\right)_{n\times 1}\perp\left(\Delta\beta^{(m)}_{i,t_0}\right)_{n\times 1}$。我们的因子模型对超额收益的分解层次如图 4 所示。

图 4 超额收益的分解



资料来源：招商证券

因子模型将截面股价表现差异归因于个股因子暴露配置信息的差异，从截面股价信息中提取独立共性成分（定价要素）的运动方向与速度（因子收益），以实现对股价表现差异影响因素的有效降维归纳。因子模型分解和归因了超额收益，以超额因子暴露与因子收益乘积的线性叠加形成对期望收益的估计。

在我们的因子模型中，因子是能够基于个股当期截面信息对下期股价表现差异进行有效区分、排序、解释的数值指标、分组规则或者属性标签，充当独立定价要素。因子暴露是因子截面取值(原值)的单调映射，计量个股对独立定价要素的配置，有效的因子暴露与截面个股收益也应形成近似的单调关联。而个股是因子视角下携带因子暴露配置信息的基础可交易载体。因子收益既是因子模型输出的结果，也是因子模型对市场动态变化自适应的结果，因此因子本身并不存在定性上固有的正向或者负向属性。

模型有效性与整体解释度

作为横截面解释型模型，因子模型的有效性应体现为模型纳入的因子组合及优先排序方式对截面股价表现差异的整体解释度，影响模型有效性的因素，包括但不仅限于：

1、模型纳入因子的数量是否足够；

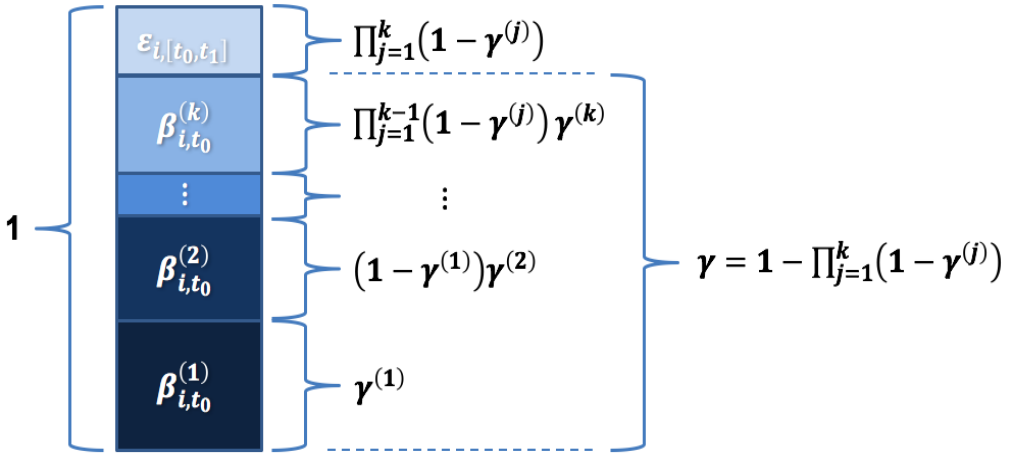
2、单因子对截面股价表现差异的解释度是否足够；

3、后续因子相对于前序优先因子是否提供了足够多的新增解释信息,即因子被模型纳入的优先排序设置是否合理。

由于我们的模型采用逐层增量方式来选取因子，因此需要关注引入的因子对于模型整体解释度的贡献。若因子(1)的解释度$\gamma^{(1)}$，指仅使用因子(1)作为解释变量时模型被解释掉的差异比例；若因子(k)的解释度$\gamma^{(k)}$，指经前序优先因子$(1),\cdots,$因子$(k-1)$对被解释变量依次解释后，被解释变量的剩余未能被解释的部分中，能够被因子(k)解释的占比，那么，因子(k)对模型解释度的贡献为$\prod_{j=1}^{k-1}(1-\gamma^{(j)})\cdot\gamma^{(k)}$，则模型的整体解释度：

$$\gamma=\gamma^{(1)}+\sum_{j=2}^{k}\prod_{i=1}^{j-1}(1-\gamma^{(i)})\gamma^{(j)}=1-\prod_{j=1}^{k}(1-\gamma^{(j)}) \qquad \text{式（9）}$$

图 5 因子模型整体解释度



资料来源：招商证券

我们认为，因子模型仅是横截面解释模型，而并非预测性模型，因此模型的有效性体现为对截面股价表现差异的解释程度是否足够。模型整体解释度，作为模型有效性的检验指标，用于判别是否找到了能够对截面股价表现差异进行有效降维归纳的因子体系，即被模型纳入的因子种类、数量以及因子优先排序是否足够优秀。

期望超额收益

在我们的因子模型中：

$$\text{下期超额收益的期望}=\sum(\text{本期末超额因子暴露}\times\text{下期因子收益预测})$$

因子收益预测体现了策略构建中的主动管理成分，被模型纳入的因子依据对超额暴露的选择可分为α因子与风险因子两类：

下期收益的可预测性较强，且依据管理目标主动寻求超额暴露的因子定义为α因子，体现了“主动 + 进攻”，获取期望超额收益；下期收益的预测难度较大，或依据管理目标严格规避超额暴露的因子定义为风险因子，体现了“被动 + 防守”，控制跟踪误差。α因子与风险因子的划分并非因子的固有属性，会随因子的可预测性与主动管理目标的变化发生切换。

因子模型：选股还是选因子？

通过对截面股价表现差异的降维归纳以及对因子收益的预测，目标是获取能由因子解释的下期期望超额收益，但未能被因子解释的波动部分确实会带来“非预期、计划外”的不确定性。由于个股资产在时间序列方向上的个体风险占比普遍较高，超额收益的波动量能被因子解释的占比相当有限，在量级上难以压制“未能被因子解释的波动”带来的（非预期的）不确定性，因此因子模型并不适用于直接选股。

以超额因子暴露$\left(\Delta\beta_{P,t_0}^{(j),target}\right)_{k\times1}$为配置目标，构建等权组合$P$，持仓股票个数为$N_{P,t_0}$，即满足$\left(\overline{\Delta\beta_{i\in P,t_0}^{(j)}}\right)_{k\times1}=\left(\Delta\beta_{P,t_0}^{(j),target}\right)_{k\times1}$，同时关闭等权组合$P$的超额系统性风险暴露，即$\Delta\tilde{\beta}_{P,t_0}^{M}=\overline{\Delta\tilde{\beta}_{i\in P,t_0}^{M}}=0$。因子$(j)$收益的波动估计为$\sigma\left[r_{F,[t_0,t_1]}^{(j)}\right]$，估计值源于因子历史收益序列，而并非是横截面因子模型中待估计参数$r_{F,[t_0,t_1]}^{(j)}$的估计误差。个股i的超额收益在时间序列方向上未能被因子解释的波动估计$\sigma\left[\mu_{i,[t_0,t_1]}\right]$并非是横截面因子模型中随机项的波动估计$\sigma\left[\varepsilon_{i,[t_0,t_1]}\right]$。

定义等权组合P的超额收益中能被因子解释的波动与未能被因子解释的波动的比值为$\theta_{P,[t_0,t_1]}$，即被因子牵动的波动对非因子波动的压制倍数。为确保因子对超额收益波动解释的显著性，设定波动率比值的下限为θ^{inf}，那么等权组合P持仓股票个数的下限N_{P,t_0}^{inf}：

$$N_{P,t_0}\geq N_{P,t_0}^{inf}=\frac{\theta^{inf}\cdot\bar{\sigma}^2\left[\mu_{i\in P,[t_0,t_1]}\right]}{\sum_{j=1}^{k}\left(\Delta\beta_{P,t_0}^{(j),target}\right)^2\cdot\theta^2\left[r_{F,[t_0,t_1]}^{(j)}\right]} \qquad \text{式（10）}$$

组合因子暴露目标配置、波动比值下限与组合持仓股票个数下限形成相互制约，若控制波动率比值不变：组合超额因子暴露越大，则股票持仓个数越少，持仓越集中，组合的主动性、进攻性就越强；组合超额因子暴露越小，则股票持仓个数越多，持仓越分散，组合的被动、防守特征就越明显。

因而，因子模型作用是为组合选择因子暴露目标配置，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。能够满足同一因子暴露目标配置的股票组合未必是唯一解，这些股票组合的表现在统计意义上应当比较接近。

因子模型搭建步骤

在上述构建理论和因子模型框架原理的基础上搭建我们的因子模型，具体实施步骤如下：

1、获取基础数据。包括分类并筛选股票池、构建市场组合与投资基准组合、收益率数据、因子模型的被解释变量与因子暴露原值。

2、单因子检验。包括对因子暴露原值处理，单调性检验，标准化赋值、显著性检验、单因子有效性测试。

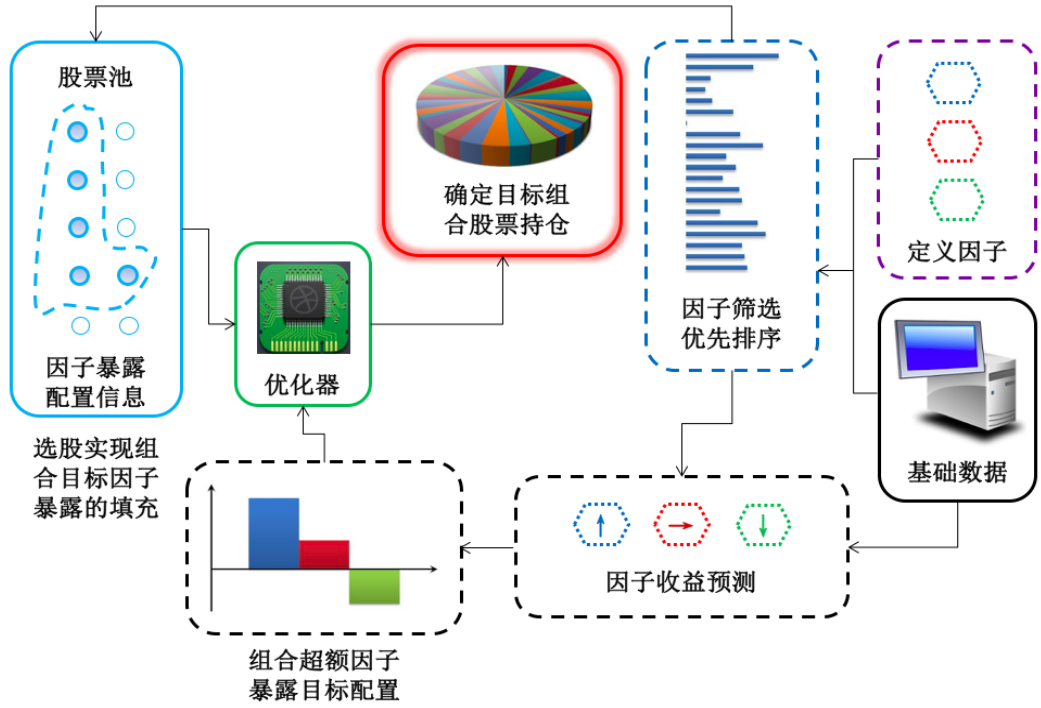
3、多因子检验。我们将采用逐层增量解释的方法来对因子进行组合确认，对各因子的优先级进行排序。

4、因子暴露目标配置。包括因子收益预测、划分α因子与风险因子、设定投资组合超额因子暴露目标配置。

5、股票组合目标持仓。由因子暴露目标配置、持仓股票个数、组合波动率、调仓换手率等，构建优化问题求解股票组合目标持仓。

6、绩效归因与模型调整。基于因子模型的整体解释度、因子收益预测效果，对组合收益与波动做归因分析，并调整模型。

图 6 因子模型搭建与优化过程示意图



资料来源：招商证券

以下对几个关键步骤进行必要说明。

基础数据处理

股票池分类与筛选

在进行因子模型的搭建之前，需要搭建一个可选股票池和投资基准股票池。由于个股价格并不是在所有时段都是“正常”波动的，在一些时候（比如新股发行、证券更名等）个股会遇到引起股价发生异常波动的事件。这些偶然发生的事件往往会导致股价出现较大的异动，而因子模型所要捕捉的并不是这些一次性、不可重复的事件带来的波动。因而必须要将这些事件引起的股价异动数据从样本中剔除，否则会出现 GIGO 的情况。我们将按交易日截面对股票池进行筛选调整，考虑个股投资基准成份股变更、上市时间、长期停牌、更名、摘帽、重大事项、行业变更、涨跌停（一字板）等情况，对数据进行

必要的剔除或者修正，按日截面构建可选股票池和投资基准股票池。

市场组合与投资基准组合构建

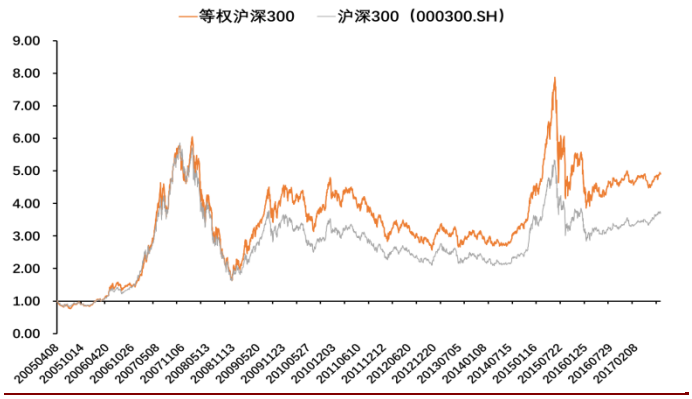
在可选股票池和投资基准股票池的基础上建立等权的市场组合与等权的投资基准组合。之所以自建等权组合而不直接利用已有的基准指数，是因为不管是中证 500、沪深 300 还是上证 50，这些常用的基准指数（标准指数）都是基于成分股（自由流通）市值来编制的，在市值维度上并不是中性的，而是更偏向于市值较大的公司个股，而等权重的编制方式能克服这一弊端。

另外，我们用满足可计算条件的成分股编制的等权基准指数还存在以下特点和优点：首先，编制的等权指数更能反映市场实际可成交成分股的整体走势。第二，由于捕捉到了市场的反转效应带来的收益，编制的等权指数在走势上要好于标准指数（如图 7 和图 8 所示），因此若策略组合走势跑赢等权指数，那么也将在很大概率上跑赢标准指数。最后，依据等权方式来编制市场组合和投资基准组合，能对后续的因子处理带来便利。

图 7 等权中证 500 指数与中证 500 指数

资料来源：招商证券、Wind 资讯

图 8 等权沪深 300 指数与沪深 300 指数



资料来源：招商证券、Wind 资讯

因子模型的被解释变量

我们的因子模型始终以个股因子暴露对基准的偏离来解释超额收益。因子模型等式（7）左侧被解释变量为经系统性风险暴露调整后的超额收益：

$$\Delta r^{B,M}_{i,[t_0,t_1]} = \Delta r^{B}_{i,[t_0,t_1]} - \Delta\tilde{\beta}^{M}_{i,t_0} \cdot r_{M,[t_0,t_1]} \qquad \text{式（11）}$$

由于市场系统性风险必是最能解释个股波动的因素，我们直接将系统性风险在等式左侧进行处理。式（11）右侧，日频超额收益序列$\Delta r^{B}_{i,[t_0,t_1]} = r_{i,t\le t_0} - r_{B,t\le t_0}$。由$\Delta r^{B}_{i,[t_0,t_1]}$和市场组合序列$r_{M,t\le t_0}$，可直接估计$t_0$截面的超额系统性风险暴露$\Delta\tilde{\beta}^{M}_{i,t_0}$以及随机项估计误差$\tilde{\sigma}[\varepsilon_{i,t\le t_0}]$。以后者确认被解释变量的截面异方差参数取值为$\tilde{\sigma}[\varepsilon_{i,t\le t_0}]\cdot\sqrt{T_i(t_0,t_1)}$，其中，$T_i(\cdot)$为区间实际可交易天数。

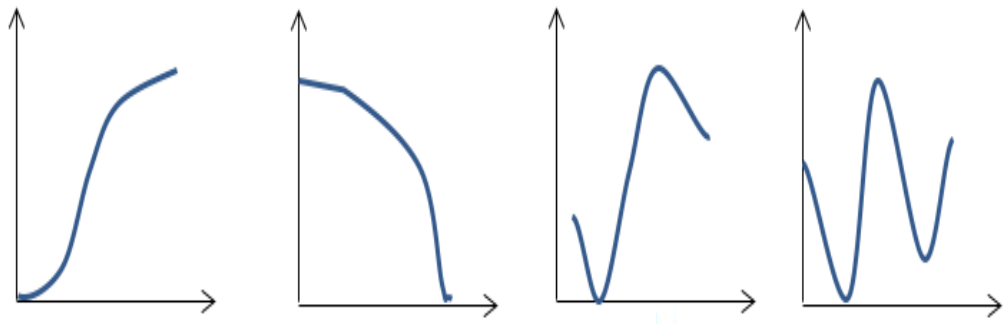
单因子检验

通过单因子检验可以对因子进行初步筛选，对于动能较弱、或者对个股没有区分能力的因子，应当事先予以淘汰，而不参与到后续的多因子组合中。我们的因子模型始终

以个股因子暴露对基准的偏离来解释超额收益，因而需先对因子的暴露原值进行处理。计算因子超额暴露原值$\Delta\tau_{i,t_0}^{(j)} = \tau_{i,t_0}^{(j)} - \tau_{B,t_0}^{(j)}$，其中$\tau_{B,t_0}^{(j)} = \bar{\tau}_{i\in B,t_0}^{(j)}$，即因子暴露对标投资基准$B$的结果，相对于投资基准的中性化处理，而后用 Boxplot + Interval Censoring 的方法删除异常值，并调整偏度和峰度。

对数据进行初步处理后，再对被解释变量与超额因子暴露的单调性检验。实际情形下，无法奢求被解释变量和解释变量间存在十分理想的线性关联，因此我们降低这一标准至两者之间存在近似的单调性关联，并可以采用双变量分布的 P - P 图或双变量之间的卡方独立性检验来测度单调性关联，并确定满足单调性的要求的边界，区间删失不满足单调性要求的样本区间。

图 9 双变量分布的 P-P 图可能情形



资料来源：招商证券

在进行单因子收益估计之前，还需要对因子进行标准化赋值。标准化赋值应在不违背单调性的条件下使得各因子超额暴露的“量纲”相同，即同时满足单位化与对称性的要求，确保因子收益及波动具有可比性。设定标准化赋值的取值范围为$[-0.5, +0.5]$将前序步骤处理后的因子超额暴露的取值区间切分为 n 个等长的子区间，对取值落在第l个子区间的数值进行标准化赋值$\Delta\beta_{i,t_0}^{(j)} = -0.5 + (l-1)/(n-1)$。

传统的标准化赋值有两种，一种是中心化处理（$x' = \frac{x-\mu}{\sigma}$），这中方法是对个股之间因子暴露差异进行了比例映射，完全保留了个股因子暴露之间的差异，只是对这种差异按一定比例进行了缩小或者放大；另一种则是“原值排序值打分”，这种方法对因子暴露差异进行了修正，将所有的原值相邻取值都做等间隔处理，减少了因子差异的变异度。而我们所采用的标准化赋值方法是中心化处理与以“原值排序值打分”标准化赋值方法的折衷，即对因子暴露中较小的差异进行了抹平，而对较大差异进行等比例映射。

对数据进行上述处理后，按设定间隔构造截面，在各横截面上估计单因子收益。以加权最小二乘法（WLS）估计被解释变量$\Delta r_{i,[t_0,t_1]}^{B,M}$与单一因子超额暴露的标准化赋值$\Delta\beta_{i,t_0}^{(j)}$的横截面回归，样本权重$\omega_{i,t_0}^{WLS} = \left(\tilde{\sigma}\left[\varepsilon_{i,t\le t_0}\right]\sqrt{T_i(t_0,t_1)}\right)^{-1}$。横截面回归中需剔除在 WLS 估计中样本权重$\omega_{i,t_0}^{WLS}$异常偏大的样本。基于截面回归显著性，结合同期市场振幅判别因子收益的强度与动量持续性等历史序列数据与指标，形成单因子的有效性排序。

因子逐层增量解释

我们采用的因子模型是通过逐层增量解释来最终确定入选模型的多因子。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$y^{(1)} = \Delta r_{i,[t_0,t_1]}^{B,M}$，即因子模型的被解释

变量；解释变量$x^{(1)}=\Delta\beta_{i,t_0}^{(1)}$，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$\Delta\tau_{i,t_0}^{(2)}$与$\Delta\tau_{i,t_0}^{(1)}$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$\Delta\tau_{i,t_0}^{(2)}$生成对应的标准化赋值$\Delta\beta_{i,t_0}^{(2)}$；

3、 残差项$Net^{(2)}=\Delta\tau_{i,t_0}^{(2)}-\tilde{c}^{(1)}-\tilde{d}^{(1)}\cdot\Delta\tau_{i,t_0}^{(1)}$并生成对应的标准化赋值$\Delta\beta_{i,t_0}^{(2)}$。

输出结果$\Delta\beta_{i,t_0}^{(2)}$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$x^{(2)}=\Delta\beta_{i,t_0}^{(2)}$，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$\gamma=1-\left(1-\gamma^{(1)}\right)\left(1-\gamma^{(2)}\right)$。

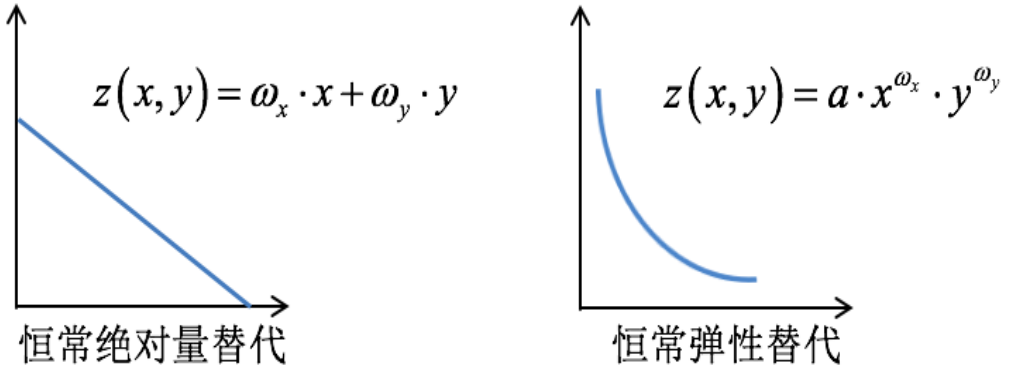
依次类推，第(k)层横截面回归用于挑选优先排序(k)因子。被解释变量$y^{(k)}=Res^{(k-1)}=\Delta r_{i,[t_0,t_1]}^{B,M}-\sum_{j=1}^{k-1}\tilde{a}^{(j)}-\sum_{j=1}^{k-1}\tilde{b}^{(j)}\cdot\Delta\beta_{i,t_0}^{(j)}$，代表所有前序优先因子$(1,2,\cdots,k-1)$未能解释的横截面股价表现差异。解释变量$x^{(k)}=\Delta\beta_{i,t_0}^{(k)}$为第$(k-1)$层辅助横截面回归残差项$Net^{(k)}$对应的标准化赋值变量，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$\gamma=1-\prod_{j=1}^{k-1}\left(1-\gamma^{(j)}\right)\gamma^{(k)}$，因子$(1,2,\cdots,k)$协同形成的因子模型整体解释度为$\gamma=1-\prod_{j=1}^{k}\left(1-\gamma^{(j)}\right)$。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：$Net^{(0)}=(0)_{n\times1}$，$Net^{(1)}=\Delta\tau_{i,t_0}^{(1)}$，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。因子间逐层的辅助横截面回归也是对新纳入因子对于所有前序优先因子解释信息逐层的中性化处理，从模型结构稳定性的角度考虑，例如行业分类、规模市值等截面排序稳定性较高的因子更适合作为优先排序靠前的入选因子。因子间逐层的辅助横截面回归也是反映了因子间解释信息的替代关系，类似打分模型中的因子加权，但区别在于后者是以“权重”来设定因子间解释信息的替代率。

图 10 因子间解释信息的替代关系



资料来源：招商证券

用逐层增量解释的方法可以消除因子之间可能存在的多重共线性，使得线性回归得到的估计值最大限度接近最优线性无偏估计量。相比于主成分分析法，它的优点在于保留了因子明确的经济学意义，能更清晰地辨别每个因子对总收益的贡献，后续对因子进行调整时，也能更有针对性；而主成分分析法通过因子旋转之后得到的测度项的经济学意义会相对模糊，不利于对回报进行归因。

结论

因子模型由于其扎实的经济学基础和良好的可扩展性，目前被国内投资者们普遍接受。本报告在学术理论的基础上，试图以新的视角审视因子模型，并提出独具特色的因子模型构建方法。

在我们看来，因子模型不应直接用于选股，而应看重实际带来回报的收益因子，将个股仅仅看作是因子视角下携带因子暴露配置信息的基础可交易载体。通过单因子测试、多因子排列来选出最优的因子配置方案，而能配置出最优因子方案的个股组合并非只有一种。

在多因子组合排序阶段，报告提出了逐层增量解释的方法，该方法为多因子组合提供了新的思路。采用逐层增量解释的意义在于明确超额收益的来源，通过因子对组合超额收益（一阶）、股价差异整体解释度（二阶）形成对模型可纳入因子明确归因的可加形式。

本报告是我们因子系列报告的第一篇，提纲挈领地展示了我们因子模型搭建的大致流程。在后续的报告中，我们将会按照本报告中提到的构建方法构建因子模型，并根据实际的数据测算的情况，可能会在某些方面做出必要的调整，目的在于形成一套较为完整的因子策略。

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