# 多因子选股策略的理论基础与模型构建



## 一、多因子模型的基本原理

### 1. 历史背景

在没有价格指数的时代，人们把投资收益看成一个整体，认为所有收益都是凭借自身投资技巧获得的。那时候的投资者只能将现在的业绩和过去的业绩相比，将投在不同资产上的收益率进行大小的计算与评判，或把自己的收益同他人比较，根本意识不到收益可以被分解成若干部分。

随着越来越多的指数被编制出来（运用市值加权、价格加权等方法），投资者终于有了 Benchmark 的概念——哦，原来可以把自己的

收益和市场基准相比啊!上世纪 60 年代,经济学家威廉·夏普、杰克·特雷诺和简·莫森等人提出了资本资产定价模型（Capital Asset Pricing Model，CAPM）。七十年代，投资者意识到具有某些相似特征的股票在市场会有相似的走势，利用 CAPM 模型仅通过单因子解释市场存在不足，套利定价模型（Arbitrage Pricing Theory，APT）被提出来了。APT 模型认为，套利行为是现代有效市场（即市场均衡价格）形成的一个决定因素，如果市场未达到均衡状态的话，市场上就会存在无风险套利机会，套利行为会使得市场重新回到均衡状态。APT 模型用多个因素来解释风险资产的收益，并根据无套利原则，得到风险资产均衡收益与多个因素之间存在（近似的）线性关系。也就是说，股票或者组合的预期收益率是与一组影响它们的系统性因素的预期收益率线性相关的，影响股票预期收益率的因素从 CAPM 中的单一因素扩展到多个因素。多因子模型（Multiple-Factor Model, MFM）正是基于 APT 模型的思想发展出来的完整的收益分解模型。

## 2. 理论介绍

多因子模型被分为两部分：收益模型和风险模型。多因子收益模型定量刻画了股票预期收益率与股票在每个因子上的因子载荷（风险敞口），以及每个因子每单位因子载荷（风险敞口）的因子收益率之间的线性关系。多因子收益模型也同样可以理解为将 N 只股票的收益率分解为 M 个因子的线性组合与未被因子解释的残差项。其一般表达式为：

$$r_i = \sum_{k=1}^{K} X_{jk} * f_k + \varepsilon_i$$

其中，$X_{jk}$为股票 j 在因子 k 上的因子暴露（因子载荷、因子值）；$f_k$为因子 k 的因子收益（回归系数）；$\varepsilon_i$为股票 i 的残差收益率。多因子收益模型的本质是将对 N 只股票的收益预测转变成对 k 个因子的收益预测，极大地降低了预测的工作量，提高预测精度。多因子收益模型并不是一个因果关系的模型，即所谓的因子只是在统计上和收益率存在相关关系，是试图解释收益风险的维度，多因子收益模型并不关心他们是否存在因果关系。

多因子风险模型的基本思路是，通过估计因子的协方差矩阵，刻画股票池未来的波动风险。而后对选股结果以及股票配置仓位进行二次优化，一般表达式为：

$$\max : w'\mu$$
$$st.\sum w = 1$$
$$w' \wedge w \le \sigma^2$$

其中，w 为股票池中的股票权重，$\mu$为个股票根据收益模型计算出来的预期收益率，$\wedge$为个股票根据风险模型计算出来的协方差矩阵，$\sigma$为风险限制条件，常数。

本报告主要介绍多因子模型中的收益模型，对于多因子风险模型，我们会在后续报告中予以探讨。

## 二、构建多因子模型的技术路线图

股票多因子模型从逻辑和算法上来看并不复杂，关键在于因子的挖掘与权重确定。多因子模型的构建相对标准化，具体包括：股票样本筛选、选取候选因子及数据预处理、检验候选因子的有效性、建立多因子收益模型、多因子风险模型的建立与二次规划（如图 1 所示）。

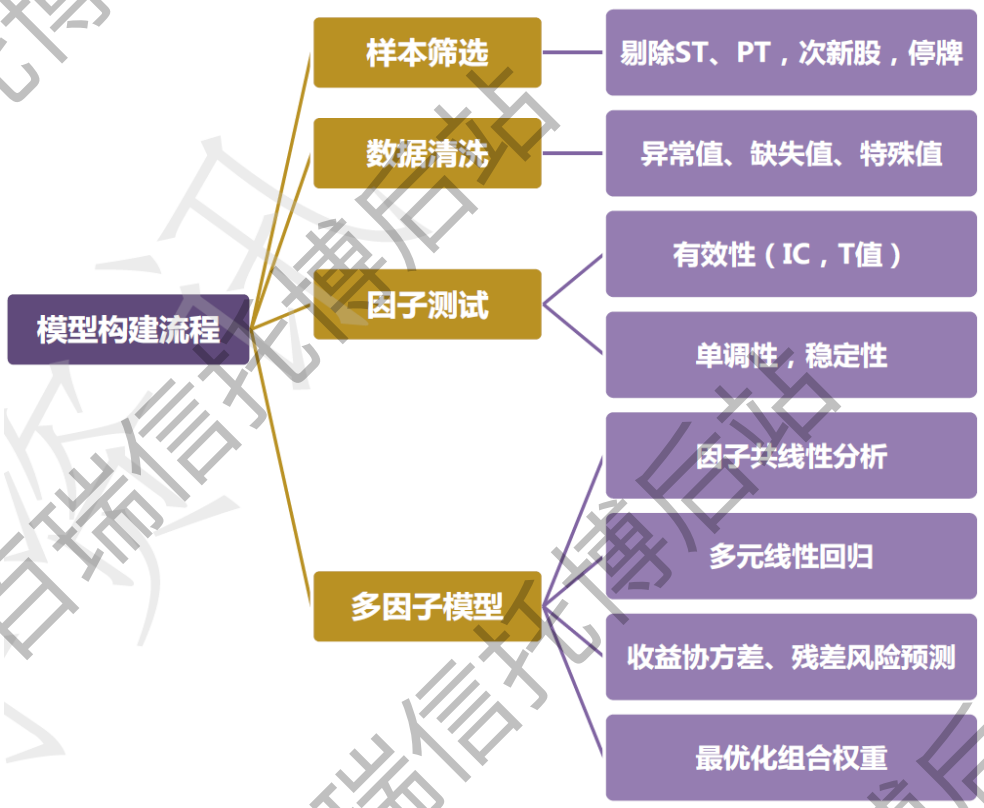


图 1 构建多因子模型的技术路线图

## 三、候选因子选取及数据预处理

首先需要确定一下股票范围：全体 A 股，也可以是中证 500 成份股。为了使测试结果更符合投资逻辑，我们设定了三条样本筛选规则：剔除选股日的 ST/PT 股票、剔除上市不满两年的股票、剔除选股日由于停牌等原因而无法买入的股票。

### 1. 选取候选因子

确定备选因子池，其次是确定因子的具体计算方法。通常研究所有的数据来源于 Wind 数据库，初始股票池为全部 A 股。报告选取十二大类因子：估值因子（Value Factor）、成长因子（Growth Factor）、财务质量因子（Financial Quality Factor）、杠杆因子（Leverage Factor）、规模因子（SizeFactor）、动量因子（Momentum Factor）、波动率因

子（Volatility Factor）、换手率因子（Turnover Factor）、改进的动量因子（Modified Momentum Factor）、分析师情绪因子（Sentiment Factor）、股东因子（Shareholder Factor）和技术因子（Technical Factor）。因子库是多因子模型的重要组成部分，我们持续探索，力求发现新的有效因子。以估值因子和成长因子为例，可以分为以下多个具体因子，如表 1 所示。

**表 1　大类因子及其描述**

| 大类因子 | 具体因子 | 因子描述 |
|---|---|---|
| 估值因子 | EP | 净利润（TTM，过去 12 个月）/总市值 |
| | EPcut | 扣除非经常性损益后净利润/总市值 |
| | BP | 净资产/总市值 |
| | SP | 营业收入/总市值 |
| | NCFP | 净现金流/总市值 |
| | OCFP | 经营性现金流/总市值 |
| | FCFP | 自由现金流/总市值 |
| | DP | 分红/总市值 |
| 成长因子 | sales_growth_q | 营业收入增长率_当季同比 |
| | sales_growth_ttm | 营业收入增长率_TTM 同比 |
| | sales_growth_3y | 营业收入增长率_三年复合增长率 |
| | profit_growth_q | 扣非后净利润增长率_当季同比 |
| | profit_growth_ttm | 扣非后净利润增长率_TTM 同比 |
| | profit_growth_3y | 扣非后净利润增长率_三年复合增长率 |
| | operationcashflow_growth_q | 经营性现金流增长率_当季同比 |
| | operationcashflow_growth_ttm | 经营性现金流增长率_TTM 同比 |
| | operationcashflow_growth_3y | 经营性现金流增长率_三年复合增长率 |

本报告的数据均来自 wind 数据库，针对每个单独的因子，我们取日线或月度或季度的数据。考虑到实际操作时涉及报告的截止日期，实际的报告影响日期范围如下：

年报：4 月 30 日截止，实际参考一季报最新数据；

一季报：4 月 30 日截止，影响范围 5 月、6 月、7 月、8 月；

半年报：8 月 31 日截止，影响范围 9 月、10 月；

三季报：10 月 31 日截止，影响范围 11 月、12 月、1 月、2 月、3 月、4 月。

**2. 数据预处理**

（1）因子中性化处理

在数据处理过程中，有很多因子受其他因子如市值因子、行业因子的影响。为了解决不同市值范围、不同行业间股票的可比性，在因子排序前，可以对选股因子进行市值和行业的风格中性化处理。对于部分因子，我们应该剥离出他们受市值、行业等因素影响的部分，还原出因子的本来面目，然后进行分组排序。

$$\alpha_i = c + \beta_{MV} * \ln(MV_i) + \sum_{j=1}^{N} \beta_{ij} * Ind_{j,i} + \varepsilon_i$$

其中，$\alpha_i$ 为股票 i 的因子值，$MV_i$ 为股票 i 的总市值，$Ind_{ji}$ 为行业虚拟变量，$\varepsilon_i$ 为我们所需市值、行业中性化后的因子值。

以主营业务收入为例，未市值中性化时的多空收益累计收益曲线与总资产的多空收益曲线吻合度很高，相关系数高达 0.965。同时主营业务收入越高，未来股票收益越低也不符合一般的市场逻辑。通过市值中性化处理，市值因素被剔除，不同市值范围的股票可以通过主营业务收入这一因子进行比较，而市值中性化后主营业务收入可以提供正向多空累计收益，成为因子池的一个备选项，如图 2 所示。


**图 2　中性化处理前后主营业务收入多空累计收益净值曲线**

（2）数据标准化处理

由于各个因子的量纲不一致，为方便进行比较和回归，需要对因子进行标准化处理。对因子进行标准化处理主要有两种方式：直接对因子载荷原始值进行标准化；首先将因子载荷原始值转换为排序值，然后再进行标准化。第一种方式的好处在于能够更多保留因子载荷之间原始的分布关系，但是进行回归的时候会受到极端值的影响；第二种方式的好处在于标准化之后的分布是标准正态分布，容易看出因子载荷和收益率之间的相关性的方向。

方法一：因子载荷原始值标准化

由于少数极端值会因子和收益率之间的相关关系估计造成严重干扰，而多因子模型本身是一个追求投资宽度的模型，所以在进行因子载荷标准化之前，我们需要对极端值进行处理。由于常见的 3σ 去极值法是基于样本服从正态分布这个假设的，但往往我们发现大部分因子值的分布都并不服从正态分布，厚尾分布的情况较为普遍。因此我们采用更加稳健的 MAD（Median Absolute Deviation 绝对中位数

法）：

$$\overline{x_i} = \begin{cases} x_M + n * D_{MAD} & if \quad x_i > x_M + n * D_{MAD} \\ x_M - n * D_{MAD} & if \quad x_i < x_M - n * D_{MAD} \\ x_i & else \end{cases}$$

其中，$x_M$ 为序列 $x_i$ 的中位数，$D_{MAD}$ 为序列 $|x_i - x_M|$ 的中位数，$\overline{x_i}$ 为去极值修正后的值。数据去极值修正后再进行标准化：

$$\overline{x_i} = \frac{x_i - u}{\sigma}$$

其中，u 为序列 $x_i$ 的均值，$\sigma$ 为序列 $x_i$ 的标准差，$\overline{x_i}$ 为序列 $x_i$ 标准化后的值。

方法二：因子载荷排序值标准化

排序标准化只关注原始序列的序关系，在做相关性分析时也只是关注排序之间的相关性，对原始变量的分布不作要求，属于非参数统计方法，适用范围相对广。第一步将原始序列转换为序关系序列：$\overline{x_i} = rank(x_i)$，其中，$\overline{x_i}$ 为在序列 $x_i$ 中的排序。第二步标准化方法与前面的标准化方法一致。

## 四、候选因子的有效性检验

原始因子集合是在逻辑上被认为与股票收益率存在关联性的因素，实证中并不是每个原始因子和股票收益率都存在相关性，因此需要对原始因子进行有效性检验，排除跟收益率相关性不高的因子。因子的有效性是多因子模型成功的基石。

通过多角度、更严格的方法度量因子的有效性和稳健性，可以确保分析结果不受数据的偶然性巧合的影响。当前主流的因子有效性测试方法包括三种，分别是因子 IC 指标、因子收益率序列以及因子值

分档计量法。

1. 因子 IC（信息系数）

因子 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相关系数。

因子 IC 值反映的是个股下棋收益率与本期因子暴露度之间的线性相关程度，是使用该因子进行收益率预测的稳健性；而回归法中计算出来的因子收益率本质上是一个斜率，反映的是从该因子可能获得的收益的大小，这并不能代表任何关于稳健性的信息。举个例子，股票池中 5 只个股第 T 期在动量因子上的暴露度为-2、-1、0、1、2，假设他们第 T+1 期收益率为-0.2、-0.1、0、0.1、0.2，，则因子 IC 值（相关系数）为 1，因子收益率（回归系数）为 0.1；假设他们第 T+1 期收益率为-0.4、-0.2、0、0.2、0.4，则因子 IC 值（相关系数）为 1，因子收益率（回归系数）为 0.2。

IC 为正表示该因子与股票的未来收益有正相关关系，应做多因子暴露值高的股票。常见的 IC 值有两种，一种是 Normal IC，另一种是 Rank IC（如图 3 所示）。

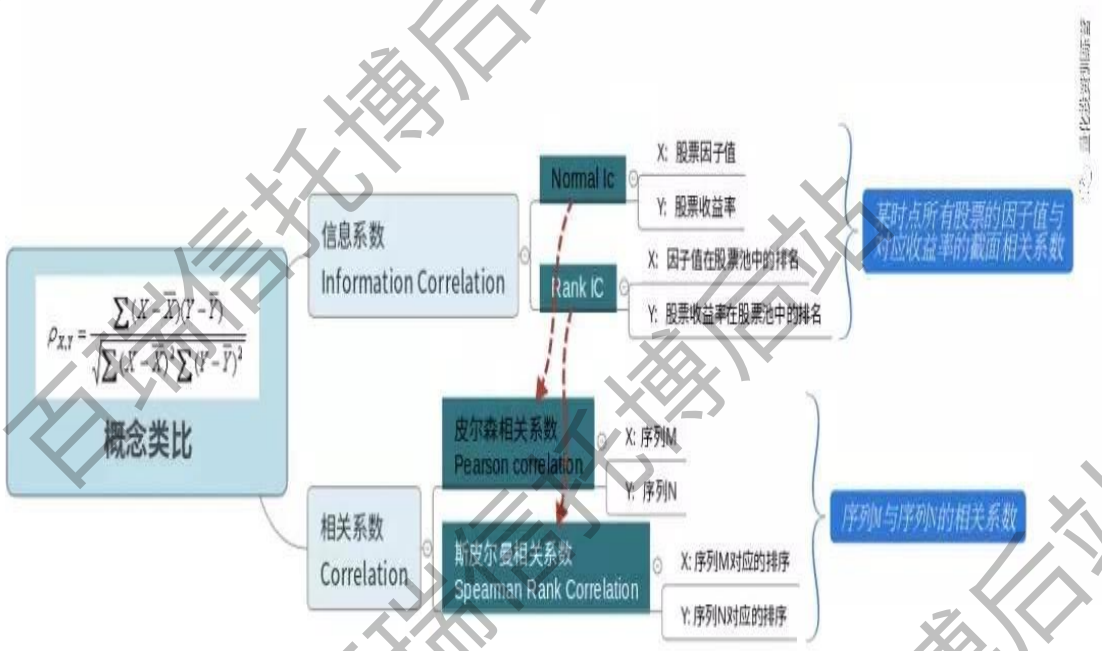


**图 3　因子 IC 值的基本原理**

Normal IC 即某时点某因子在全部股票的暴露值与其下期回报的截面相关系数：Normal IC = corr（$f_{t-1}$ , $r_t$）。其中，$f_{t-1}$ 为 t-1 其股票的因子值，$r_t$ 为 t 期的股票收益率。Rank IC 即某时点某因子在全部股票暴露值排名与其下期回报排名的截面相关系数：Rank IC = corr($order_{f,t-1}$ , $order_{r,t-1}$)。其中，$order_{f,t-1}$ 为 t-1 期个股票收益排名，$order_{r,t-1}$ 为 t 期各股票收益率排名。

当得到各因子 IC 值序列后，我们可以从以下几个方面衡量因子的有效性：

第一，IC 值序列的均值及绝对值均值——衡量因子预测能力的指标；

第二，IC 值序列的标准差——衡量因子预测能力稳定性的指标；

第三，IC 值序列大于零（或小于零）的占比——判断因子效果的一致性的指标；

第四，IR（IR=IC/IC 的标准差）。

在计算因子的 IC 值时，需要注意一下几个问题：

第一，因子提纯。在利用 IC 值评价因子有效性时，可以预先对因子进行提纯，排除行业、市值等重要因素的影响，使结果更明晰。具体来说，就是在因子标准化处理之后，在每个截面期上用其做因变量对市值因子及行业因子等做线性回归，取残差作为因子值的一个替代，这种做法可以消除因子在行业、板块、市值等方面的偏离。例如，股息率因子较高的个股可能较多分布在电力及公用事业、汽车、商贸零售等行业以及大市值板块，经过因子提纯之后，股息率因子较高的个股就会平均分布在各行业及板块了。

第二，IC 衰退。在不同时间段，每个因子的有效性是不一样的，呈现出轮动的特征。因子在过去独占鳌头，不代表未来仍然出类拔萃；曾经默默无闻的因子，可能在下一阶段脱颖而出。每一时期，都有“当红”因子和“过气”因子。也就是说，因子有效性是具有时效性的，IC 作为度量因子有效性的指标，其稳定性值得关注。因子 IC 衰退，是通过观察随着滞后时间的延长，因子有效性降低的速度。研究发现，很多因子具有相对稳定的半衰期，即因子有效性降低为一半所需要的

时间，因而可以通过观察半衰期的长短判断该因子的稳定情况。

与上文提到的 IC 指标计算方法类似，IC 衰退的计算，只不过数据用的是所有股票当期的因子暴露值与滞后 i 期的收益率数据。首先，计算每期的因子暴露值和滞后 i 期的收益率间的 IC 信息系数，其中 i=1,2,...,12；其次，分别对因子每隔 i 期的信息系数计算均值。

第三，因子 IR（信息比）。信息比率（IR）,即超额收益的均值与标准差之比。在多因子模型中，计算信息比率是用 IC 值的均值与其标准差的比值：$IR \approx \frac{\overline{IC_t}}{std(IC_t)}$。

在实际运用中，各因子的 IC 或 Rank-ic 统计量随时间可能有较大的波动，因此为考虑因子有效性的波动性，较多研究还进一步采用因子 IC 或 Rank-ic 统计量的信息比率（IR）指标来全面衡量因子有效性，即使用 IC 或 Rank-ic 的均值统计量除以方差统计量。IC_IR 的绝对值越高，该因子的选股效果越好。

**2. 根据因子收益率序列确定有效性**

截面回归（Cross-Section Regression）是目前业界较常用于因子测试的方法。相比全样本面板回归（Panel Data Regression）的方法，截面回归更有利于对因子变化趋势的捕捉。

（1）单因素回归确定每个因子每期的收益

我们选择每期针对全体样本做一次回归，回归是因子暴露（因子载荷、因子值）为已知变量，回归得到的每期的一个因子收益值 $f_j$。

进行截面回归判断每个单因子的收益情况和显著性时，需要特别关注 A 股市场中在过去的很长一段时间内显著影响个股收益率的因

素，例如行业因素和市值因素。为了能够在单因子测试时得到因子真正收益情况，我们在回归测试时对市值因子、行业因子做了剔除。针对因子 k，单因子的回归模型如下（横截面数据进行回归）：

$$r_j^t = \sum_{s=1}^{S} X_{js}^t * f_s^t + X_{jl}^t * f_l^t + X_{jk}^t * f_k^t + \varepsilon_j^t$$

其中，$r_j^t$ 为股票 j 在 t 期的收益率；$X_{js}^t$ 为股票 j 在第 t 期在行业 s 上的暴露；$f_s^t$ 为行业 s 在第 t 期的收益率；$X_{jl}^t$ 为股票 j 在第 t 期在市值 l 上的暴露；$f_l^t$ 为市值 l 在第 t 期的收益率；$X_{jk}^t$ 为股票 j 在第 t 期在因子 k 上的暴露；$f_k^t$ 为因子 k 在第 t 期的收益率。$X_{js}^t$ 是一个 0-1 的哑变量，即如果股票 j 属于行业 s，则暴露度为 1，否则为 0。在有的模型中，会对公司所在属性进行拆分，比如公司 j 的业务 50%属于行业 a，30%属于行业 b，20%的业务属于行业 c，则股票 j 在行业 a 的暴露度为 0.5，在行业 b 的暴露度为 0.3，在行业 c 的暴露度为 0.2。

A 股的行业分类，主要存在两种方式，一种是外来的 GICS 风格的行业分类，一种是本土的行业分类。GICS 风格的行业分类，我们参考中证指数公司发布的中证行业指数系列：中证能源、中证材料、中证工业、中证可选、中证消费、中证医药、中证金融、中证信息、中证电信、中证公用。本土的行业分类，我们参考中信行业指数系列：石油石化、煤炭、有色金属、电力及公用事业、钢铁、基础化工、建筑、建材、轻工制造、机械、电力设备、国防军工、汽车、商贸零售、餐饮旅游、家电、纺织服装、医药、食品饮料、农林牧渔、银行、非银行金融、房地产、交通运输、电子元器件、通信、计算机、传媒、综合。

在每一个横截面上使用上述模型进行加权最小二乘回归（WLS），权重采用流通市值的平方根，一定程度上消除了异方差性。

（2）因子收益率序列及其 t 检验

在通过多期截面回归之后，我们可以得到因子收益序列，$f_k^t$ 是因子 k 在第 t 期的因子收益（回归得到的系数）。对于因子有效性，可以从以下四个方面的分析：

第一，t 值绝对值序列的均值——衡量因子整体显著性的指标。对于每一期的截面回归，都可以得到一个收益率因子 $f_k^t$（截面回归得到的系数）的 t 值。对于 t 值序列，首先取绝对值，然后计算 $|t|$ 的均值，$|t|$ 是判断因子是否为有效因子的重要指标。之所以要取绝对值，是因为只要 t 值显著不等于 0，即可认为在当期因子和收益率存在显著的相关性。但这种相关性有时候为正，有时候为负，如果不取绝对值，则很多正负相抵消，会影响因子的有效性。

第二，t 值绝对值系列大于 2 的比例——衡量因子显著性是否稳定的指标。检验 $|t|>2$ 的比例是为了保证 $|t|$ 平均值的未定型，避免出现少数数值特别大的样本值拉高均值。

第三，因子收益 $f_k^t$ 序列的平均值——衡量因子收益能力大小的指标。

第四，因子收益 $f_k^t$ 序列的标准差——衡量因子收益能力波动率的指标。

第五，因子收益 $f_k^t$ 序列的 t 值检验——衡量因子收益统计上是否显著不为 0 的指标。对于每一期的截面回归，都可以得到一个因子收

益率 $f_k^t$，对于 $f_k^t$ 序列同样需要进行 t 检验。为得到因子 k 在第 t 期是否和股票收益率显著相关，即 $f_k^t$ 是否显著不为 0，我们需要对其进行 t 检验：

$$t = \frac{\bar{x} - u}{\sigma / \sqrt{n-1}}$$

t 为 x 的 t 统计量；$\bar{x}$ 为样本的均值；u 为总体的均值；$\sigma$ 为样本的标准差；n 为样本的容量。

第六，因子收益 $f_k^t$ 序列大于 0 的概率——衡量因子收益率方向是否一致的指标。

**3. 因子值分档计量法**

首先，计算股票池中样本股票各个因子的指标；然后，根据因子指标的计算结果，从小到大对样本股票进行排序，并分为 5 个等份，从而在整个样本期内形成不同因子下的 5 个排序组合；最后，用一定的方法度量因子的有效性。

将股票分为 5 档之后，计算两次月平均收益率差：Top20%(第一档)-Bottom20%（第五档）、Top40%(第一、二档)-Bottom40%（第四、五档）。计算上述收益率的平均值，如果平均值为正，则该因子的影响总体上是正向的，如果平均值为负则反之。正向（负向）因子的有效性：月均收益率差为正（负）的月份数在总月份数中的占比。此外，还可以根据每个月份五档的因子排名与对应的平均收益率排名，计算其相关系数以及显著性检验的 P 值，越显著月说明因子对收益的影响越稳定。

## 五、建立多因子模型

在多因子策略的构造过程中，各因子的权重设置是研究的重点。筛选出有效因子之后，可以利用因子来构建模型筛选股票，主要包括打分法和回归法。

### 1. 打分法

在模型运行期的某个时间开始，例如每个月初，对市场中正常交易的个股计算每个因子的最新得分并按照一定的权重求得所有因子的平均分。最后，根据模型所得出的综合平均分对股票进行排序，然后根据需要选择排名靠前的股票。例如，选取得分最高的前 20%股票，或者选取得分最高的 50 到 100 只股票等。

$$S = \sum_{i=1}^{n} w_i * score_i$$

常见的因子加权方式有等权，IC 加权，IC_IR 加权等，具体如下：

**表 2　常见因子加权方式**

| 权重分配方式 | 优缺点 |
| --- | --- |
| 等权：各因子取相同的权重<br>$W = \left(\frac{1}{n}, \frac{1}{n} \cdots \frac{1}{n}\right)$ | 优点：简单方便 |
| | 缺点：未考虑因子有效性、稳定性及相关性的差异 |
| IC 值均值加权：取因子过去一段时间的 IC 值均值为权重<br>$W = \left(\overline{IR_{f1}}, \overline{IR_{f2}} \cdots \overline{IR_{fn}}\right)$ | 优点：考虑了因子有效性的差异 |
| | 缺点：未考虑因子稳定性及因子相关性 |
| IC_IR 加权：取过去一段时间的 IC_IR 为权重，IC_IR=IC 均值/IC 标准差<br>$W = \left(IR_{f1}, IR_{f2} \cdots IR_{fn}\right)$ | 优点：考虑因子稳定性和有效性 |
| | 缺点：未考虑因子相关性 |

## 2. 回归过程中需要注意的关键问题

（1）多重共线性分析

在因子搜集的时候就会根据因子的具体经济含义对因子进行大类划分，但是同类型的因子可能存在较强的相关性，多元线性回归的时候会造成多重共线性（Multicollinearity），多重共线性是指回归模型中的解释变量之间由于存在精确相关关系或高度相关性而使模型估计失真或者难以估计准确。

多重共线性的判断。处理多重共线性，首先需要判断因子间是否存在多重共线性。判断多重共线性的方 法主要有以下两种：第一，相关性矩阵。计算各因子历史序列的两两相关性，取平均值，得到相关性矩阵，辨别出相关性较高因子。这是最直观判断多重共线性的方法。第二，VIF 检验。相关性矩阵给出了两两因子的相关性，但如果某因子与其他多个因子间存在相互表示的线性关系，相关性矩阵则有可能无法检测出来，这就需要引入 VIF 检验。VIF 是方差膨胀因子（Variance Inflation Factors）的英文缩写，是统计学中常用的一种多重共线性检测手段。该方法通过检查指定因子能够被回归方程中其他全部因子所解释的程度来检测多重共线性。通过计算，得到方程中的每个因子的 VIF 值，过高的 VIF 值表明该因子的引入增大了整个系统的多重共线性。值得强调的是，在一般统计教科书中，会建议把 VIF>10 作为存在多重共线性的标志。但在多因子模型中，因子间的解释能力本就较弱，VIF 普遍偏低，一般 VIF>4 的时候，因子的多重共线性已经比较显著了。所以最终判断时仍应结合实际问题具体分

析。

多重共线性的处理。对于确定存在共线性的因子，一般有如下三种处理方法：第一，直接剔除。对于和其他因子表现出较高相关性，但又不能提供更多信息的因子，一般采用直接剔除处理（剔除冗余因子的步骤：第一，对不同因子下各个时期各股票的分组号码求相关系数；第二，求样本期各因子间的相关系数算术平均并构建相关性矩阵；第三，设定一个相关性系数阈值，将得分相关性矩阵中大于该阀值的相关系数所对应的因子只保留与其他因子相关性较小、有效性更强的因子，而其它因子则作为冗余因子剔除）。第二，因子合成。对于大类内因子，表现出一定相关性，又不能直接剔除的，可以采用将小因子合成大因子的方式。如月度换手率、季度换手率、半年换手率这三个因子相关性较强，可合并为流动性因子 LIQ。对于因子合成，主要的方法有等权法、历史收益率加权法、历史信息比率加权法以及主成份分析法。第三，因子正交。大类间因子如表现出一定相关性，无论从直观理解，还是经济学解释的角度都不适宜采用因子合并的方式，这时可以使用因子正交的手法，将相关性较高的因子之一相对另一因子做回归，取回归残差项代替因子值。

如果是经济含义类似的同类型因子，存在明显相关性，为尽可能多的保留因子信息，我们可以将因子进行合并；如果是经济含义不同的因子，存在明显相关性，我们只能有所取舍，保留更加显著的因子，而舍弃相对不显著的因子。

（2）残差异方差分析

异方差性（Heteroscedasticity）是相对于同方差而言的。所谓同方差，是为了保证回归参数估计量具有良好的统计性质，经典线性回归模型的一个重要假定：总体回归函数中的随机误差项满足同方差性，即它们都有相同的方差。如果这一假定不满足，即：随机误差项具有不同的方差，则称线性回归模型存在异方差性。

因此需要对模型的残差是否存在异方差进行检验。对于存在异方差的模型，在进行回归分析的时候需要采用加权最小二乘法（WLS）。

（3）回归方法的选择

最小二乘法 OLS。OLS 是最常用和最简单的方法，但该方法的缺点是 OLS 需要假设回归方程的残差均具有相同的方差，但由于股票收益率常常存在异常值，同时不同股票之间的收益率波动性也不尽相同。使用 OLS 时，异常值会对回归结果和回归测试的显著性检验带来较明显的偏差。RLM (Robust Linear Model)。Robust Regression 稳健回归同样常见于单因子回归测试，RLM 通过迭代的赋权回归可以有效的减小异常值（outliers）对参数估计结果有效性和稳定性的影响。

在独立同分布正态误差的线性模型中，OLS 是有效无偏估计。然而当误差服从非正态分布时，OLS 就很容易给异常值 outliers 赋予较高的权重，从而导致模型结果失真。RLM 中常用的 M-estimator 方法则是采用迭代加权最小二乘估计回归系数，根据回归残差的大小确定各点的权重，以达到稳健的目的。为减少“异常值”作用，RLM 可以对不同的样本点赋予不同的权重，即对残差小的点给予较大的权

重，而对残差较大的点给予较小的权重，根据残差大小确定权重，并据此建立加权的最小二乘估计，反复迭代以改进权重系数，直至权重系数的改变小于一定的允许误差(tolerance)内。

RLM 方法可以更好的处理异常值的影响，从而提高回归分析的有效性。因此我们在单因子测试时将采取这种更为稳健的回归方法。

3. 回归法

具体的做法是收集最原始的因子集 F1 的相关数据，对 F1 进行有效因子筛选，得到因子集 F2，对 F2 进行多重共线性分析，得到因子集 F3；将股票多因子的取值与下期的股票收益率进行回归分析，得到一个回归方程，然后再把最新的因子集 F3 的因子值代入回归方程得到一个对未来股票收益的预判，然后再以此为依据进行选股。

$$r_j^t = \sum_{s=1}^{S} X_{js}^t * f_s^t + \sum_{k=1}^{K} X_{jk}^t * f_k^t + \varepsilon_j^t$$

其中，$r_j^t$ 为股票 j 在 t 期的收益率；$X_{js}^t$ 为股票 j 在第 t 期在行业 s 上的暴露；$f_s^t$ 为行业 s 在第 t 期的收益率；$X_{jk}^t$ 为股票 j 在第 t 期在因子 k 上的暴露；$f_k^t$ 为因子 k 在第 t 期的收益率。$X_{js}^t$ 是一个 0-1 的哑变量，即如果股票 j 属于行业 s，则暴露度为 1，否则为 0。

## 六、根据模型估计股票预期收益率

### 1. 因子收益率向量的计算

通过多元线性回归，可以得到所有因子的历史收益率序列：

$$F = \begin{bmatrix} f_1^1 & f_2^1 & \cdots & f_K^1 \\ f_1^2 & f_2^2 & \cdots & \cdots \\ \vdots & \vdots & \vdots & \vdots \\ f_1^T & f_2^T & \cdots & f_K^T \end{bmatrix}$$

由于因子每期收益或多或少存在不稳定性，为保证模型的稳定性，需要对因子历史收益序列进行分析，给出下一期因子收益的合理预期值。对于 T+1 期因子的预期收益率的估计，可以采用以下几种方式：

（1）历史均值法：用前 N 期的因子收益率的均值作为 T+1 其因子的预期收益率：

$$f_k^{T+1} = \frac{\sum_{t=T-N+1}^{T} f_k^t}{N}$$

一般情况下，N 取 36 或者 60，即前 36 个月或者 60 个月的均值。

（2）指数加权移动平均法（Exponentially Weighted Moving Average，EWMA）：由于因子收益率包含的信息有可能也是存在衰减，所以离当前越近的观测值权重越重，越远的观测值权重越轻。

（3）时间系列预测法：根据 AR（自回归模型）、MA（移动平均模型）、ARMA（自回归移动平均模型）、ARIMA（自回归积分移动平均模型）对未来之进行预测。

（4）滤波法提取趋势项：由于因子收益率存在较大波动性，我们创新地通过 HP 滤波法提取因子历史累积收益率的趋势项，以滤波曲线的终值除以样本期的长度作为因子的预期收益率。此种方法人工预设的参数很少，在获取因子收益率长期变化规律的同时能够尽量消除噪声的影响，经实证检验效果不错。

2. 个股收益率的计算

我们可以根据因子收益和每个股票的因子载荷计算出个股的预

期收益率。在估算出 T+1 期的因子收益率向量$\left(f_1^{T+1}, f_2^{T+1} \cdots f_k^{T+1}\right)$之后，以及计算出 T+1 期的因子载荷矩阵 $X^{T+1}=\begin{bmatrix} X_{11}^{T+1} & X_{12}^{T+1} & \cdots & X_{1K}^{T+1} \\ X_{21}^{T+1} & X_{22}^{T+1} & \cdots & \cdots \\ \vdots & \vdots & \vdots & \vdots \\ X_{N1}^{T+1} & \cdots & \cdots & X_{NK}^{T+1} \end{bmatrix}$。通过公式 $r_j^{t+1}=\sum_{k=1}^{K} X_{jk}^{T+1} * f_k^{T+1}$ 就可以计算出 T+1 期每只股票的预期收益率向量$\left(r_1^{T+1}, r_2^{T+1} \cdots r_N^{T+1}\right)$。

## 七、多因子模型有效与否的几个关键点

第一，初始因子库的构建

从多因子策略的名称中就能看出，因子的选取是构建模型的基础，也是非常重要和关键的一步。由于该策略是建立在市场无效或者弱有效的前提之下，随着使用多因子选股模型的投资者数量不断增加，有的因子会逐渐失效，而另外一些新的因子可能被验证有效而加入到模型中。此外，一些因子可能在过去的市场环境中比较有效，而随着市场风格的改变，这些因子可能短期失效，另外一些以前无效的因子会在当前的市场环境下表现较好。因此，在初始因子选取过程中，需要尽可能多地扩充因子库中的因子。

第二，因子有效性分析

构建完成因子库之后，就需要对其中因子的有效性进行判断。因子有效性的判断方法非常多，但如何对这些判断方法进行综合运用却没有公认步骤。准确、合理地度量各个因子的有效性，考验着策略构建者的能力。

第三，多元回归模型的构建

单纯的多元回归模型的构建并不难，难的是在进行多元回归的时候需要处理的多个问题：多重共线性、异方差性、回归方法的选择等等。每个问题处理不好都会影响回归结果的有效性，但这些问题又都是计量经济学中普遍存在的较难处理的问题。如何将这些问题对回归结果的影响降到最低，考验着策略构建者在计量经济学领域的功力。

第四，因子收益率向量的预测

即便是将上面的所有问题都解决之后，即找到市场上的绝大部分因子，并逐一对其进行合理的有效性分析；在解决多重共线性、异方差性、回归方法选择等多个问题的基础上，通过进行多元回归分析，得到所有因子的历史收益率序列，此时还需要对因子历史收益序列进行分析，预测出下一期因子收益的合理预期值。预测方法的选择并没有固定之术，同样考验着策略构建者在计量经济学领域的功力。