北京大学量化交易协会2021级培训

多因子模型

李炜杭 刘宇林 耿岱琳 许鹏程 宁磊 陈傲霜

2021-10-11

多因子模型

1 多因子概述

2 因子评价

3 多因子模型构建

4 投资组合优化

5 经典因子与模型简述

多因子模型

概 念

- 什么是因子？因子就是影响股票收益的因素，比如宏观因素，基本面因素，统计性因素等。
- 什么是多因子选股？多因子选股采用一系列的因子（主要考虑使用价值、成长、质量以及市场等四大类因子）作为选股标准，将多个具有逻辑背景的因子策略相结合，选取在各个因子上综合得分较高的股票构建投资组合。
- 多因子模型的类型有哪些？
- **宏观因子模型：**因子表示能够显著影响收益的宏观经济变量的非预期变动，例如利率、通胀风险、经济周期以及信用利差等
- **基本面因子模型：**因子主要表示能够解释证券横截面差异的主要因素。例如，账面市值比、市盈率以及杠杆率。
- **统计因子模型：**通过对证券的历史收益表现进行统计并提取出影响收益的主要因子。主要的因子统计模型有因子分析模型与主成分分析模型。
- 有哪些因子？

多因子模型

因子类型	具体指标
规模类因子	总市值，流通市值，自由流通实质
估值类因子	市盈率，市净率，市销率，市现率，企业估值倍数
成长类因子	营业收入同比增长率、营业利润同比增长率
盈利类因子	净资产收益率ROE、总资产报酬率ROA、销售毛利率、净利率
动量反转因子	前1，2，3，6个月涨跌幅
交投因子	前一个月日均换手率
波动因子	前一个月波动率，前一个月振幅
股东因子	户均持股比例、户均持股比例变化、机构持股比例变化
分析师因子	预测当年净利润增长率、主营业务增长率

多因子模型-发展历史

1900年 盲目跟风炒股

- 当时美股市场是投机市场，人们只会关注股票的购买数量以及股票的涨速，属于盲目跟风炒股

1929年 股市大崩盘

- 1929年股市大崩盘，几乎所有的股票下跌都超过了80%，很多人都因此破产，经历了大崩盘后，人们开始思考股票价格到底是由什么决定的。

1934年史密斯提出了“股票价值决定于其未来收益”的重要思想

- 经历了大崩盘之后，人们逐渐意识到基本面分析的重要性。价值投资原理也逐步诞生。

多因子模型-发展历史

1952年 Markowitz投资组合

- 不仅仅着眼于单个资产，开始研究有效的资产组合，引入均值和方差刻画股票投资的收益和风险

1963年 资本资产定价模型（CAPM）

- 构建投资组合时，可以通过分散投资消除非系统性风险，因此只需要考虑系统性风险
- 决定资产期望回报率的是资产回报与市场波动的相关性β

1976年Ross套利定价模型（APT）

- 认为证券收益率与一组因子有着线性关系，是多因子模型的基础

1992年Fama和French的三因子模型

多因子模型-理论基础

Markowitz 投资组合理论

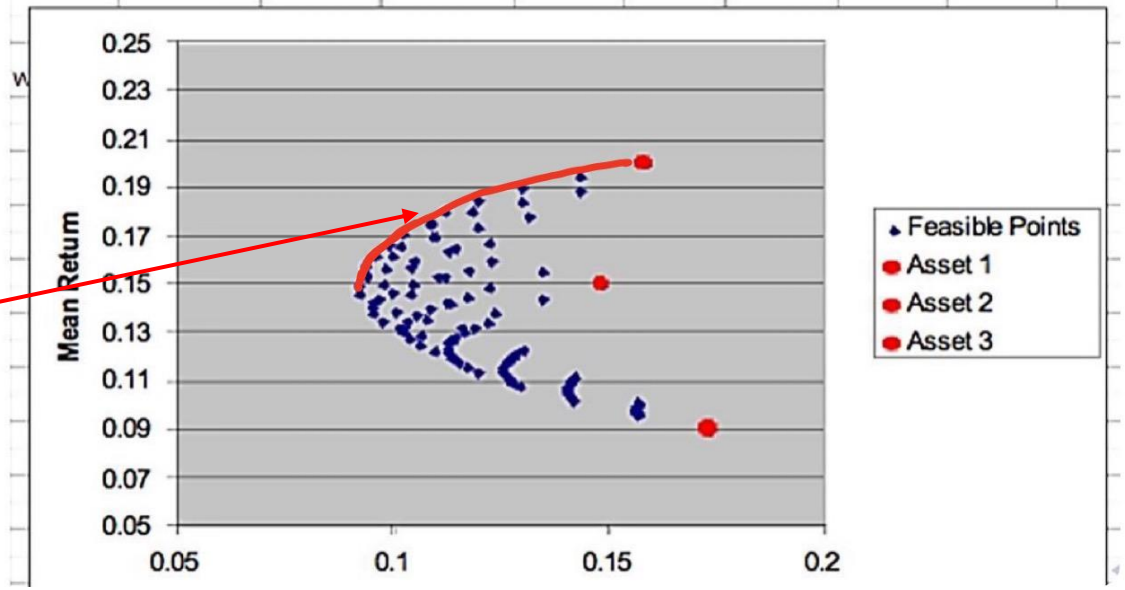
- 最小化方差的最优化模型
- 假设存在N个风险资产，每个风险资产的随机回报率为R_i,资产投资的权重为$w = (w_1, w_2, \dots, w_n)'$
- $R = (R_1, R_2, \dots, R_n)'$,协方差矩阵为$\Sigma$, $\mu = E(R)$, 投资组合的最优化模型如下：

$$min\, w'\Sigma w$$
$$s.t. w'u = r$$
$$w'e = 1$$

有效投资组合

The Feasible Set

Asset	Return	Std. Dev.	Covariance	1	2	3
1	9%	0.1732051	1	0.03	0	0
2	15%	0.148324	2	0	0.022	0
3	20%	0.1581139	3	0	0	0.025



多因子模型-理论基础

CAPM模型

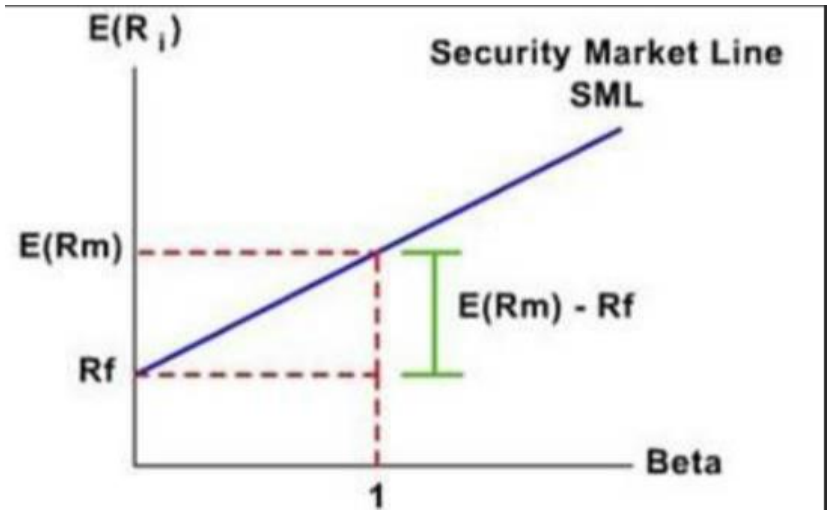
- 模型设定如下：

$$E(R_i) = R_f + \beta_i[E(R_m) - R_f]$$

其中R_i表示资产i的收益率，R_f表示无风险收益率，R_m表示市场基准收益率，β_i表示资产i的收益率变化对市场组合收益率变化的敏感程度，其值为资产i的收益率与市场组合收益率的协方差与市场收益率的方差的比值：

$$\beta_i = \frac{Cov(R_i, R_m)}{Var(R_m)}$$

- **模型假设如下：**
- 1. 效用函数可视为收益率的函数
- 2. 用方差或标准差衡量投资收益率的风险系数
- 3. 任何投资者都偏好更高的期望收益和更低的风险系数
- 4. 所有投资可以无限细分且任意交易
- 5. 没有税负以及交易成本
- 6. 投资者对不同投资收益和风险又相同预期
- 7. 不存在通货膨胀且折现率不变



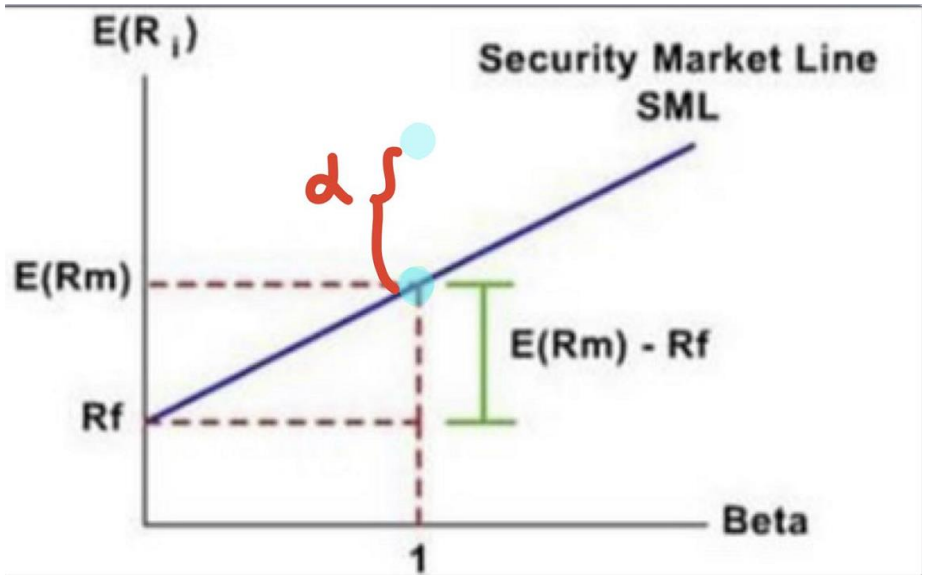
多因子模型-理论基础

CAPM模型

- 超额收益率α用来衡量投资证券组合的实际收益与市场期望收益的差值

$$\alpha = \left(E(R_i) - R_f\right) - \beta_i\left[E(R_m) - R_f\right]$$

$\alpha > 0$, 说明证券组合的实际收益位于证券市场线(SML)之上，被低估，推荐买入
$\alpha < 0$, 说明证券组合的实际收益位于证券市场线(SML)之下，被高估，推荐卖出
$\alpha = 0$, 说明证券组合定价合理
$\beta > 1$, 说明证券组合的波动性高于市场；$\beta < 1$, 说明证券组合的波动性低于市场；$\beta = 1$, 说明证券组合的波动性与市场相同



多因子模型-理论基础

APT模型

- 假设市场中的资产收益率由K个因素决定，即：

$$R_i = E(R_i) + \beta_{i1}\lambda_1 + \beta_{i2}\lambda_2 + \cdots + \beta_{iK}\lambda_K + \epsilon_i$$

λ_k表示因子收益率，一个因子描述了众多资产共同暴露的某种系统性风险，该风险是资产收益率背后的驱动力；因子收益率正是这种系统性风险的风险溢价或风险补偿，它是这些资产的共性收益。即由市场共同因子所描述的风险。

β是因子暴露，表示资产组合对不同因素（系统性风险）的敏感程度

ϵ是特质收益率，表示与因子无关的风险，即非系统风险

$E(R_i)$表示所有决定资产收益率的因素对应风险收益为0的时候，证券组合的期望收益

- 模型补充：
- APT模型是多因子模型的基础，其假设某种资产的收益率与市场中的多个因素（即多因子）有关，同时还包含了一个与市场因子无关的非系统性风险，可以证明当市场中的证券数量趋于无穷的时候，非系统性风险的风险系数趋于0。
- CAPM模型是APT模型的特例，CAPM模型仅考虑了系统风险中的一个因子，即市场组合的回报率。

多因子模型-理论基础

Fama-French三因子模型

- 1992年， Fama和French 1992年对美国股票市场决定不同股票回报率差异的因素的研究发现，股票的市场的beta值不能解释不同股票回报率的差异，而上市公司的市值、账面市值比、市盈率可以解释股票回报率的差异。因此提出了著名的三因子模型：

$$R_i - R_f = \alpha + \beta_{im}(R_m - R_f) + \beta_{iSMB} * SMB + \beta_{iHML} * HML + \epsilon_i$$

R_i表示证券组合的实际平均收益率
α表示超额收益率
β表示资产组合对不同因素的敏感程度，简称因子载荷
SMB表示小市值的股票与大市值的股票的收益率之差
HML表示高账面市值比股票与低账面市值比股票的收益率之差

多因子模型-理论基础

多因子模型的发展

多因子模型	提出者	所含因子
Fama-French三因子	Fama and French(1993)	市场、规模、价值
Carhart四因子	Carhart(1997)	市场、规模、价值、动量
Novy-Marx四因子	Novy-Marx(2013)	市场、规模、价值、盈利
Fama-French五因子	Fama and French(2015)	市场、规模、价值、盈利、投资
Hou-Xue-Zhang四因子	Hou et al	市场、规模、盈利、投资
Stambaugh-Yuan四因子	Stambaugh and Yuan(2017)	市场、规模、管理、表现
Daniel-Hirshleifer-Sun三因子	Daniel et al(2020)	市场、长周期行为、短周期行为

多因子模型-选股步骤

数据处理

- 数据采集与提取→数据标准化与预处理

因子评价

- 逻辑分析→多空回测→IC/IR

模型构建

- 因子筛选→因子赋权→模型构建

组合优化

- 风险预测→确定组合收益目标/风险目标→权重约束→因子暴露约束→个股上下限约束→二次规划求解组合权重最优分配→模拟业绩回溯

多因子模型

1 多因子概述

2 **因子评价**

3 多因子模型构建

4 投资组合优化

5 经典因子与模型简述

因子评价

1 **逻辑解释**

2 多空回测

3 IC/IR

在开始测试之前

概 念

- 对单因子的测试始于对因子影响股票预期收益率的逻辑的分析。
- 什么是在实践中能够放心被投资者使用的因子？--为何确信未来可用？
- 并非单纯的统计现象--需要金融、经济、会计等知识分析

图 15、EP_TTM 因子行业中性等权多空组合表现

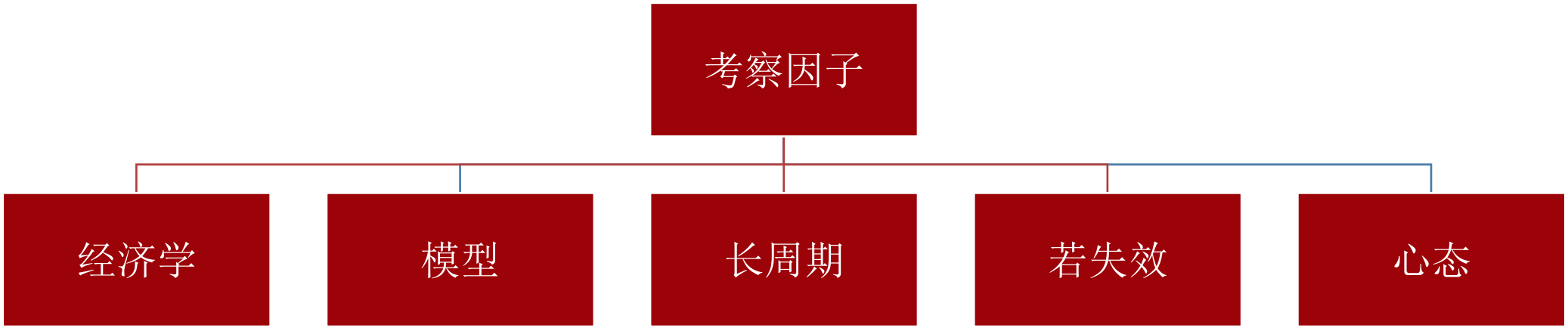
- EP_TTM是归母净利润/市值。研究认为估值越低，预期收益率越高的逻辑是被广泛认可的。那么可以像左图一样构建多空组合（具体操作马上讲）

- 09年到12年--失效
- 13年到15年--失效甚至相反

低估值逻辑失效？！

看逻辑还是只看实证？

逻辑解释



社会科学的属性：非黑即白的想法也许在这里并不适用，将可解释性与数据实证相结合。

1.全面搜集因子起效逻辑相关文献，根据个人的理解和知识背景，生成自己的模型
2.模型无对错，但有好坏之分。

有了模型后用统计方法，对该因子 的预测能力进行长周期的评价检验，并持续跟踪。

1.思考逻辑有无本质变化(背后依赖的经济、金融模型有没有出现本质改变)
2.若否定，则坚定地持有(不能反向使用因子，因为没有从逻辑上否认有效性。

1.因子长期对股票价格具有预测能力的原因，正是源于他并不是时刻都有效!
2.重视因子逻辑，可以帮助投资者规避数据挖掘的陷阱。
3.不断思考，保持对逻辑解释的审慎态度!

因子评价

1 逻辑解释

2 **多空回测**

3 IC/IR

多空回测之前--数据预处理

中位数标准化

- 极端值会有很大干扰，所以先处理极端值常见办法是中位数去极，对于某序列x：
- $$\tilde{x}_i = \begin{cases} x_m + n * D_m, & if\ x_i > x_m + n * D_m \\ x_m - n * D_m, & if\ x_i < x_m - n * D_m \\ x_i, & else \end{cases}$$
- 其中x_m为序列x的中位数，D_m为序列$|x_i - x_m|$的中位数，对新的序列标准化（z-score）

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。**
- 也即令$\tilde{x}_i = rank(x_i)$，对新的序列标准化即可

第一种方式的好处在于能够更多 保留因子暴露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

第二种方式好处是由于经过排序，容易看出因子暴露和收益率之间的相关性的方向。

注：业界中$n = 3 * 1.4826$

多空回测之前--数据预处理

缩尾法/固定比例法

- 最简单也最常见的方法，是学术论文中的标准办法。
- 将变量从小到大进行排序，将小于$p\%$分位数（主观选择）和大于$1-p\%$分位数的指标剔除。
- 缺点：必然删一些数据，可能删掉了合理的；对分布很不对称的指标用该法也不太好。

三倍标准差法

- 认为落在3σ以外的是小概率事件（参考正态分布）视为异常值。
- 好处是考虑了数据的波动情况，在分布接近正态的时候效果非常好
- 缺点是数据确实偏差较大的情况下偏差大，此外，部分过大的异常值会让3σ偏大，少识别了异常值

预处理还可能出现的问题是**数据缺失**，解决办法有：
直接删除
财务因子沿用上期，别的直接删除
滑动平均补充
标准化后赋值为0（也即几乎不考虑这些值的影响）
……

多空回测之前--多重排序

独立二重排序

- 使用两个排序变量分别独立地把股票划分为a、b个组，两两取交集得到ab个投资组合
- 右图为2*3独立双重排序

- 为何进行独立双重排序？
- **纯**因子收益率

		BM		
		High（高组）	Middle（中间组）	Low（低组）
市值	Small（小市值）	S/H	S/M	S/L
	Big（大市值）	B/H	B/M	B/L

市值分组依据：NYSE 中位数；
BM 分组依据：NYSE 30% 和 70% 分位数。

独立多重排序

- 同理可以扩展到多重，类似字典排序
- Hou, Xue, and Zhang (2015) 使用市值、单季度 ROE 和总资产变化率进行 2 × 3 × 3 独立三重排序
- 比如S/H/H 代表由小市值、高 ROE 和高总资产变化率股票构成的分组

$$\begin{aligned} I/A = & \frac{1}{6}(S/L/L + S/M/L + S/H/L + B/L/L + B/M/L + B/H/L) \\ & - \frac{1}{6}(S/L/H + S/M/H + S/H/H + B/L/H + B/M/H + B/H/H) \end{aligned}$$

为了检验一个新的因子是否真的可以获得超额收益，拿它和已有因子进行双重排序，排除已有因子的影响
关于排序可参考：《因子投资方法与实践》page28

多空回测之前--多重排序

独立排序分析

- 简单
- 两个变量的地位是完全对称的，随便对哪个因子构建排序组合都可
- 取交集后某些块数据可能太少（若高度正相关，集中在对角线上）

		BM		
		High（高组）	Middle（中间组）	Low（低组）
市值	Small（小市值）	S/H	S/M	S/L
	Big（大市值）	B/H	B/M	B/L

市值分组依据：NYSE 中位数；
BM 分组依据：NYSE 30% 和 70% 分位数。

条件多重排序

- 先按市值分为两组S、B
- 每组内按BM高低进行排序
- 换言之S/H里的BM与B/H里的BM可能差的很多，而独立排序法差别不大
- 从排序目的上看：条件排序是更好的选择，但是老的模型一般就只是独立排序

一方面，排序法后通过回测拿到了可用于回归的因子收益率
另一方面，多重排序后降低了别的因子对想研究因子的影响
　　消除影响的另一方法：**中性化处理**（常用于消除在行业、市场上的暴露）回归取残差
$F_i = \beta_M * ln(MktVal_i) + \sum_{j=1}^{N} \beta_j d_j + \varepsilon_i$（$d_j$为行业虚拟二值变量，为何市值取对数？）
剔除后选股更分散，可能效果更好

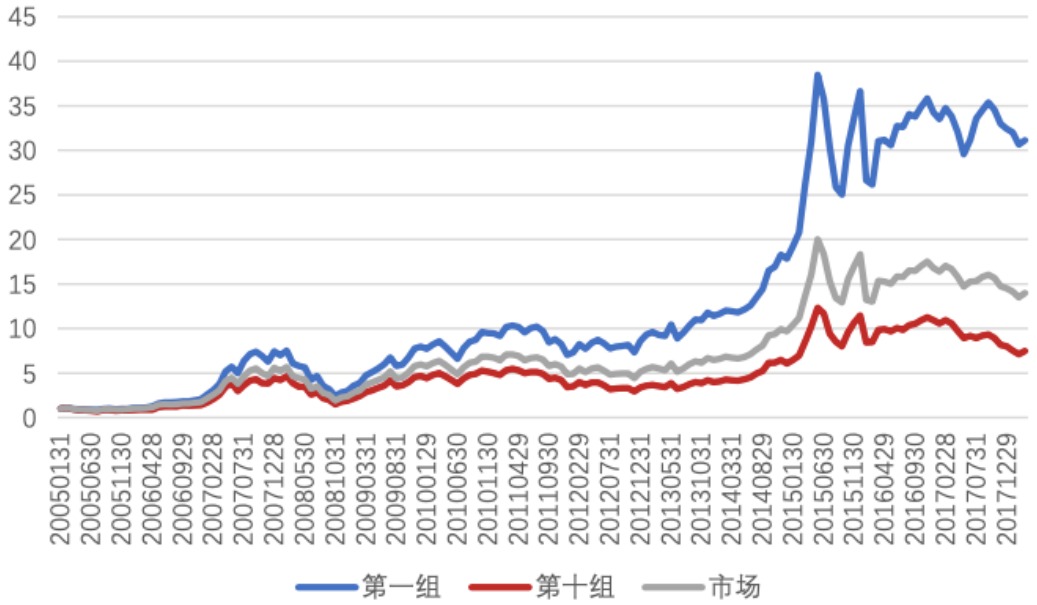
模拟因子的表现--分位数+多空回测

策略步骤

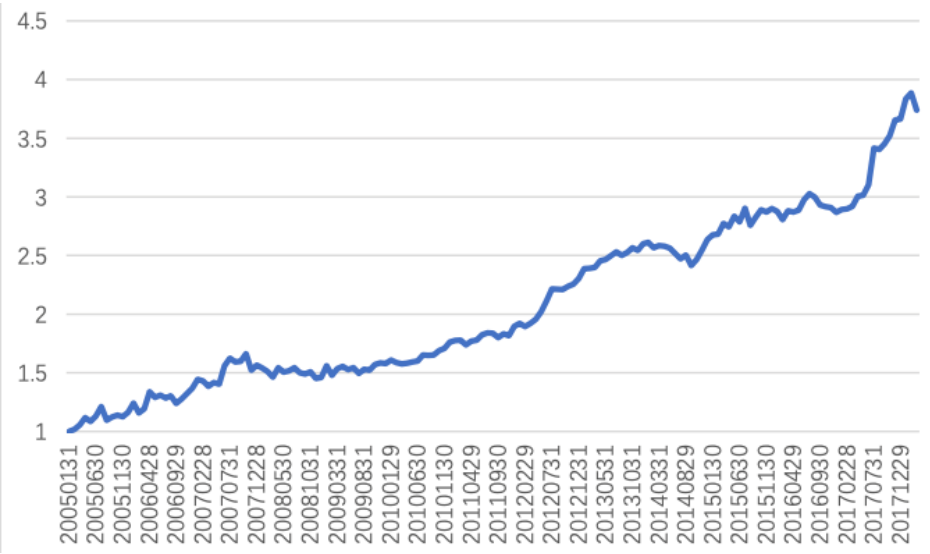
- Step1:确定需要研究的因子并将可投资股票池中的股票依据因子暴露β/因子值大小排序;
- Step2:将排序后的股票池依据因子的 L 分位点切分成股票数量大体相等的L 个组合，每个组合内的股票则根据流通市值加权或等权加权;
- Step3:做多第一分位组合，做空第 L 分位组合，构建多空对冲组合来描述因子表现。
- 如果我们测试 的周期是月频，那么每个月底都要重复一次（换手也是研究指标之一）

分层法：观察每组收益，应该有单调的规律

研究指标：波动率、夏普比率、平均换手率.....



单季度营业利润同比增速因子



多空回测

缺陷

- 隐含了对因子和预期收益之间是线性关系(或者至少是单调关系)的假设，而这在现实中不一定成立。
- 多空对冲组合的构建需要做空，这一行为在很多市场被禁止或成本很高，而且在测试多空因子组合时我们一般也不会考虑交易成本、流动性和其他限制条件，因此所得到的结果总是偏乐观的。
- “纯因子”问题：多数因子都存在和其他因子的相关性，而这种相关性会使得我们估计 的因子收益中包含一部分其他因子的表现。

第三点举例：市值与ROE

针对最后一点缺陷的简单改进

- 将全体可投资股票池按照某个重要的风险维度(例如行业、市值等)进行分块。
- 再在每个分块内将股票按照目标因子暴露均分成 L 组，把每个分块内的第一组 合并为第一分位组合，……，第 L 组合并为第 L 分位组合。
- 所考虑的维度较多时，就不再适用了。

使用纯因子多空组合测试法！

纯因子多空回测

概述

- 一共有 M 个因子，研究对象是其中的因子 m
- 利用时点 $t-1$ 的信息构建一个表征因子 m 的多空组合ω_m^{t-1},调整它来得到纯因子收益率

多空组合要求：

$\sum_{i=1}^{N} \omega_{i,m}^{t-1}$=0(无现金敞口)

$\sum_{i=1}^{N} \omega_{i,m}^{t-1}\ \beta_{i,m}^{t-1}$=1(对因子m给一个单位暴露)

$\sum_{i=1}^{N} \omega_{i,m}^{t-1}\ \beta_{i,j}^{t-1}$=0 ∀j≠ m (其他因子0暴露)

与APT非常像

在下一期做横截面回归

$$r_i^t = \lambda_m^t \beta_{i,m}^{t-1} + (\lambda_1^t \beta_{i,1}^{t-1} + \ldots + \lambda_{m-1}^t \beta_{i,m-1}^{t-1} + \lambda_{m+1}^t \beta_{i,m+1}^{t-1} + \ldots) + u_i^t$$

算收益率λ^t：$\lambda^t = [((\beta^{t-1})'\beta^{t-1})^{-1}((\beta^{t-1})']r^t$

记:$((\beta^{t-1})'\beta^{t-1})^{-1}((\beta^{t-1})' \triangleq [\omega_1^{t-1}, \omega_2^{t-1}, \ldots \omega_K^{t-1}]$

从而：$\lambda_m^t = \omega_m^{t-1} r^t$；可验证$\omega_m^{t-1}$符合条件

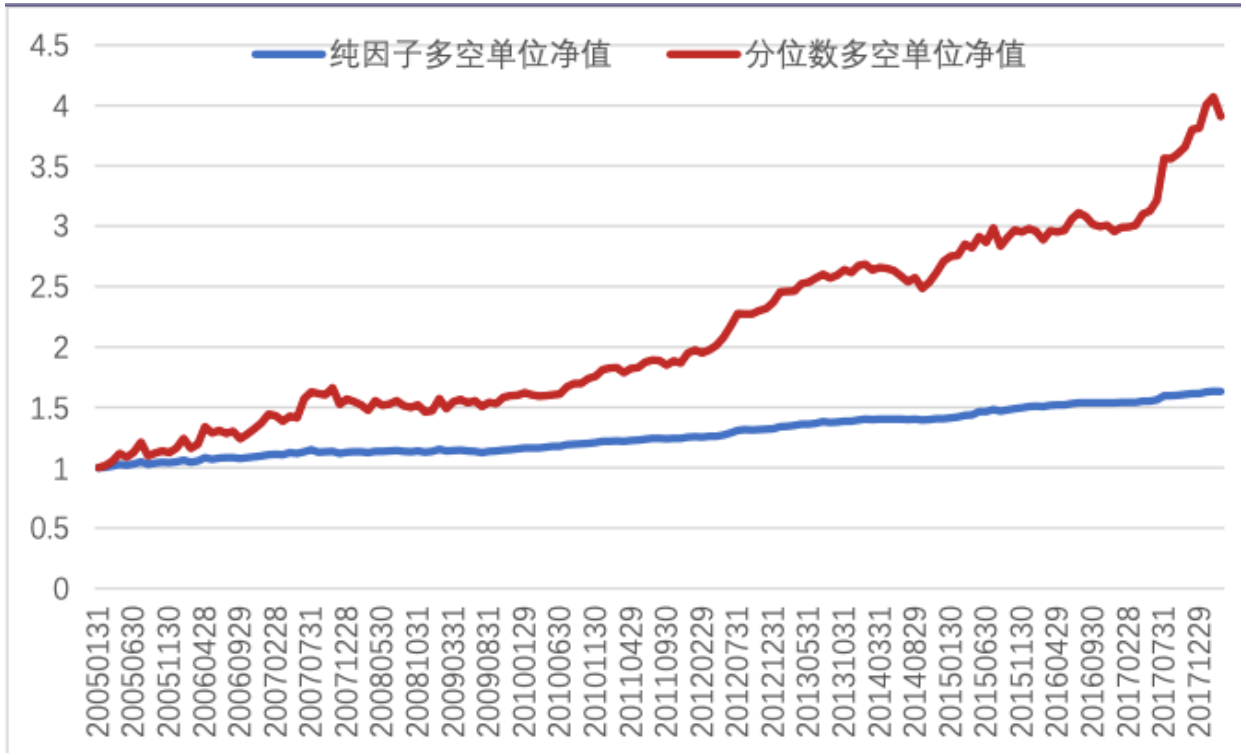
从而得到λ_m^t时间序列（$t = 2,3..T$）

为何要用截面回归/Fama-MacBeth回归
Barra CNE5

纯因子多空回测

结果

- 右图为**单季度营业利润同比增速因子分位数多空净值与控制了行业和市值暴露的纯因子多空净值的走势对比**
- 控制行业和市值因子，既降低了目标因子的收益，也同时大幅降低了波动率。这说明由季度增速因子所导致的行业与市值的被动暴露对风险、收益均有很大比例的贡献
- 相较于分位数多空组合而言纯因子组合的夏普比率和月度胜率都更高，说明因子组合的稳定性得到了明显提升。



纯因子多空法对股票横截面信息的利用更加充分，且规避了其他因子对目标因子风险收益特征的影响，但可投资性较差依然是主要弱点。

统计结果比较

统计指标	分位数多空	纯因子多空
年化收益率	10.91%	3.79%
年化波动率	10.67%	2.42%
夏普比率	1.02	1.56
胜率	65%	74%

回归的“进化历程”

$$r_i^t = \alpha_i + \beta_i \lambda_t (or\ \beta_i f_t) + \epsilon_i^t$$
$$i = 1 \ldots N, t = 1 \ldots T$$

时间序列回归 因子收益-->因子暴露

- 可参照三因子模型、CAPM模型使用过程。
- **排序法**构建**风格因子**的因子模拟投资组合，拿到λ，对每个资产i，OLS出因子暴露β_i，α_i（一个资产，T个时间）
- 像GDP这种因子就没法做，因为拿不到λ

截面回归 因子取值-->因子暴露-->因子收益

- 使用因子取值f，重复时间序列回归的做法，拿到β_i（此时截距不是α_i，因为因子取值不一定是因子收益）
- r_i^t在时间序列上（一个资产，T个时间）取均值得到$E(r_i)$，结合上面的β_iOLS出λ（对N个资产的回归），这一步回归才拿到了α_i（残差）

Fama-MacBeth 截面回归最后一步是先平均在回归，这里是先回归再平均。

- 使用因子取值f，重复时间序列回归的做法，拿到β_i（此时截距不是α_i，因为因子取值不一定是因子收益）
- 对于每个t，结合上面的β_i用r_i^tOLS出λ_t，α_t^i，（要做T次）然后求平均值拿到λ与α_i

注意：公司特征法（barra说明书等）直接拿到β_i，一般比时间序列效果更好，截面回归也不必进行第一步了
eg. BM标准化等处理之后直接作为暴露，比时序回归出的暴露表现好些
参考与深入：《因子投资方法与实践》page39

因子评价

1 逻辑解释

2 多空回测

3 **IC/IR**

IC

标准化

- 自然希望更正式、更简单的评 价因子预测能力强弱的指标。广泛用信息系数（Information Coefficient）
- 信息系数指的是当期因子值与下期股票收益率之间的横截面线性相关系数，传统的基于 Pearson 相关系数的 IC 数学定义，对因子m有：$PersonIC_m^t = corr(\beta_m^{t-1}, r^t)$
- 仅考虑因子值和收益率的排名，此时的信息系数也被称作 Spearman Rank IC，简称 Rank IC，$RankIC_m^t = corr(rank(\beta_m^{t-1}), rank(r^t))$

本质就是相关系数
性质？--能告诉我们什么？

- 优劣： RankIC不受异常值的影响(虽然我们在前期已经对因子值中的异 常值进行了处理，但收益率中的异常值也是普遍存在的)，结果更为稳健。
- 当然也可能缺失了关键信息

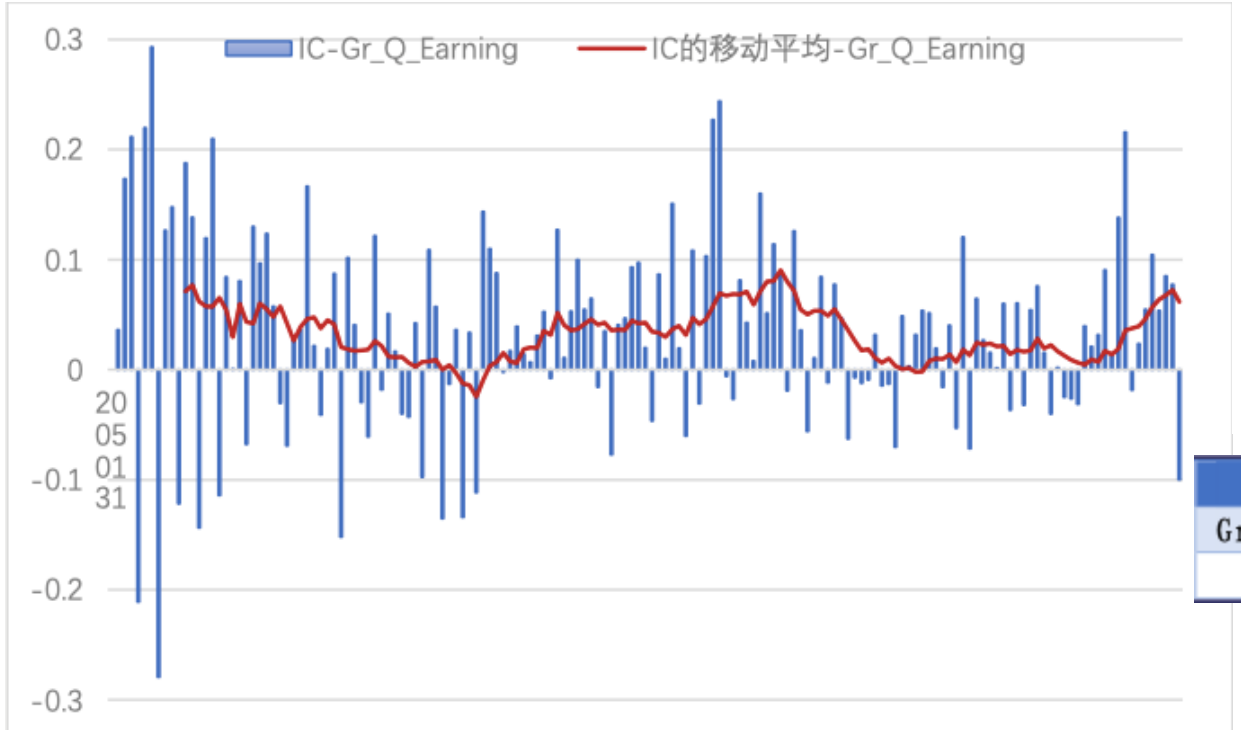
Stock	Factor Score	Subsequent Return	Factor Score Rank	Return Rank
A	(1.45)	(3.00%)	9	8
B	(1.16)	(0.60%)	8	7
C	(0.60)	(0.50%)	7	6
D	(0.40)	(0.48%)	6	5
E	0.00	1.20%	5	4
F	0.40	3.00%	4	3
G	0.60	3.02%	3	2
H	1.16	3.05%	2	1
I	1.45	(8.50%)	1	9
Mean	0.00	(0.31%)		
Standard Deivation	1.00	3.71%		
Pearson IC		(0.80%)		
Spearman Rank IC				40.00%

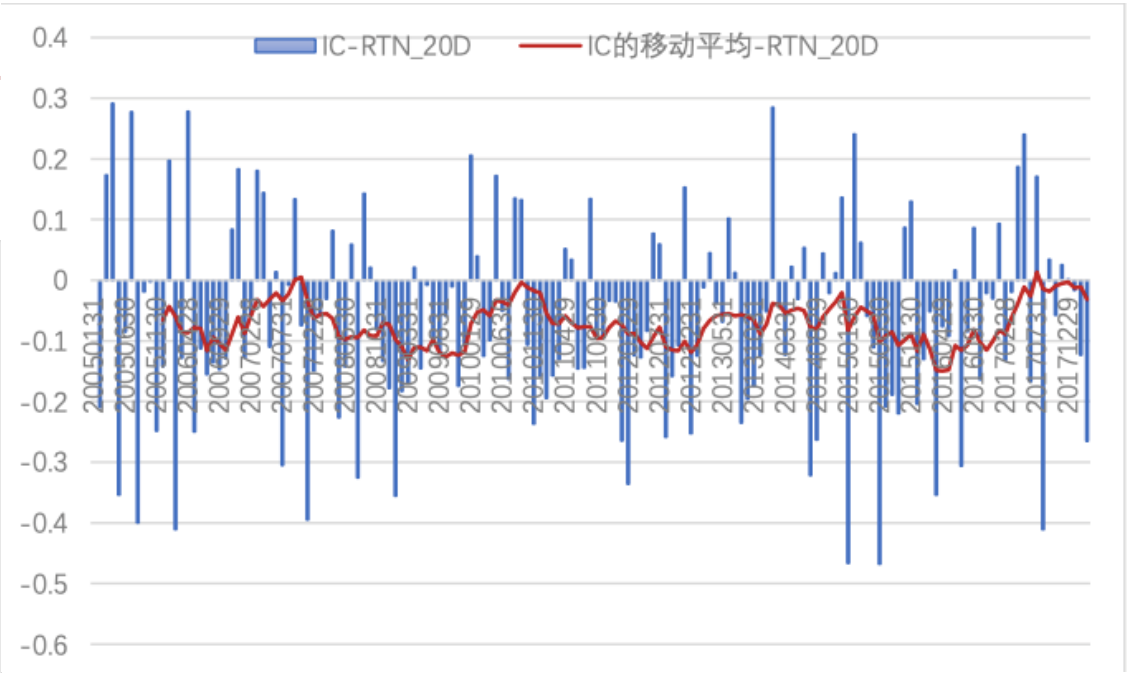
IC--作为时间序列

其他度量

- 作为时间序列，有最大最小值均值方差
- 定义IC_IR $= \frac{avg(IC_m^t)}{std(IC_m^t)}$，风险收益比。区别于IR $= \frac{R_P - R_B}{Var(R_P - R_B)}$（相对于基准组合的超额收益稳定性）
- IC的T统计量：$\frac{avg(IC_m^t)}{std(IC_m^t)}\sqrt{T-1}$

IC、IC_IR越高越受欢迎



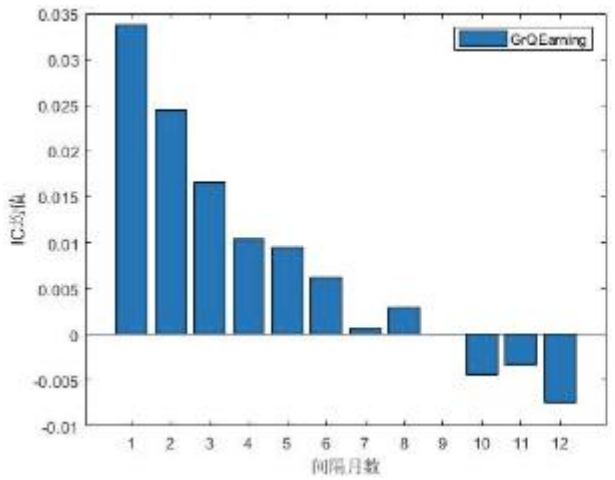


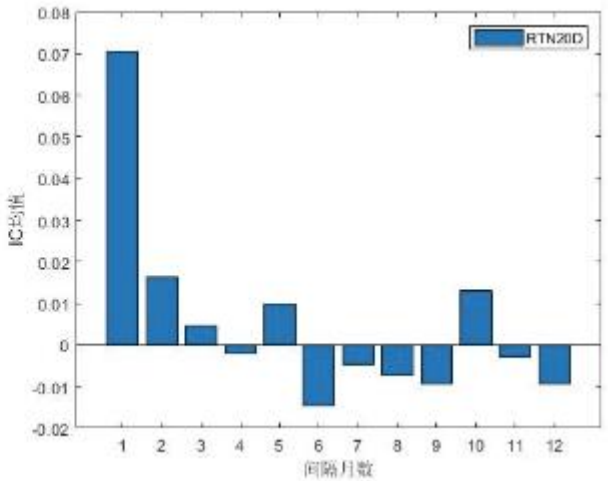
因子	平均值	标准差	最小值	最大值	IC_IR	t 统计量
Gr_Q_Earning	3.38%	8.62%	-27.94%	29.32%	0.39	4.92
RTN_20D	-7.05%	15.98%	-46.78%	29.15%	-0.44	-5.55

IC衍生指标

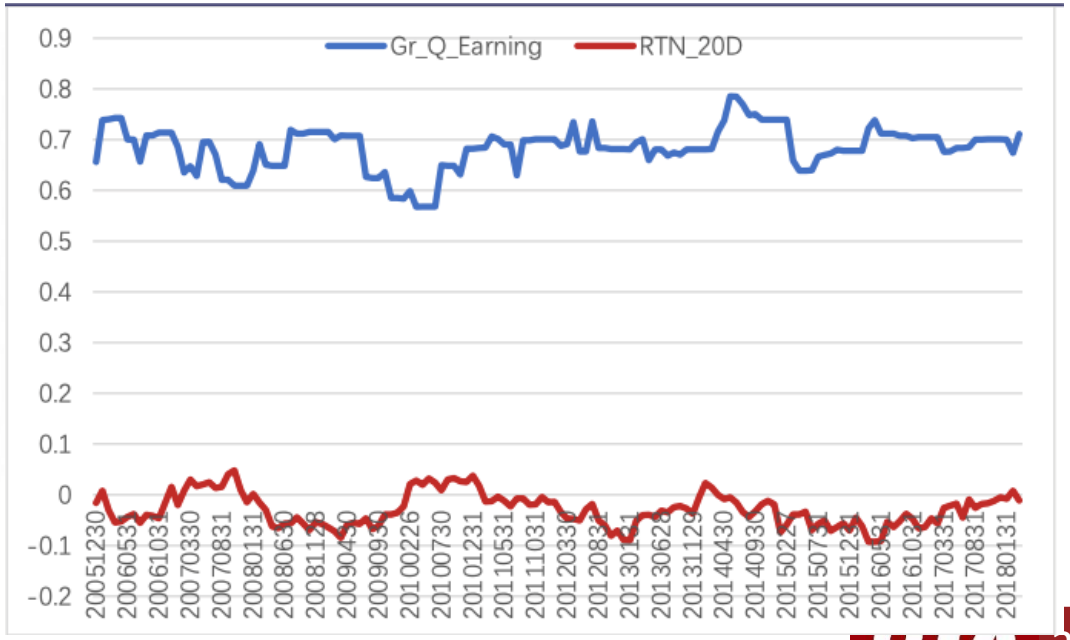
定义

- 观察时点与预测时段之间的间隔时间越来越长，因子的预测能力会出现什么变化呢?--IC衰减。$RankIC_m^{t,Q} = corr(rank(\beta_m^{t-1}), rank(r^{t+Q}))$
- 与因子信息衰减紧密相连的一个重要概念就是所谓的因子换手率。投资组 合的换手率不仅依赖于因子值的变化，也与投资组合的构建过程密切相关。$FactorAutocorrelation = corr(\beta_m^{t-1}, \beta_m^t)$
- 因子间相关性：$FactorScoreCorr_{m,k}^t = corr(rank(\beta_{i,m}^t), rank(\beta_{i,k}^t))$。处理高相关性问题





$FactorAutocorrelation$走势图



多因子模型

1 多因子概述

2 因子评价

3 **多因子模型构建**

4 投资组合优化

5 经典因子与模型简述

线性多因子Alpha模型构建

问题定义

- 经典的线性多因子Alpha模型的矩阵形式如下，

$$R^t = \beta^{t-1}\lambda^t \quad (1)$$

- 在$t-1$时刻，预测时段$[t-1, t]$的收益率，就需要预测相应时段的因子收益率λ^t
- 记$\hat{R}^t$为对R^t的预测值，w^{t-1}为对λ^t的预测，那么以预测为目的的线形多因子模型如下，

$$\hat{R}^t = \beta^{t-1}w^{t-1} \ (2)$$

- 预测因子收益就等价于在期初确定不同因子的权重

构建步骤

- 一般来说，因子m的权重w_m^{t-1}可以看作是所有因子过往收益率$\{\lambda^2, \dots, \lambda^{t-1}\}$和对因子m表现有影响的一系列外部变量$\{Y_1, \dots, Y_k\}$的多元函数$F_m(\lambda^2, \dots, \lambda^{t-1}, Y_1, \dots, Y_k)$。搭建线性多因子模型的核心就是确定函数$F_m(\cdot)$的具体形式。函数$F_m(\cdot)$形式的确定主要分两步：

1. **因子筛选**：确定因子m是否纳入当期多因子模型，即$F_m(\cdot)$是否恒等于0；
2. **因子赋权**：在确定因子m是否纳入当期多因子模型的基础上，给出函数$F_m(\cdot)$的具体形式已确定权重w_m^{t-1}的取值；

多因子模型构建

1 因子筛选

2 因子赋权

3 非线性多因子模型

因子筛选

可能遇到的问题

- **维度高**：太多因子可供选择
- **多重共线性**：因子之间彼此相关联是常态，大多数基于线性回归的框架对于多重共线性问题的应对能力比较弱
- **低信噪比**：金融市场数据信噪比相较于科研和工程问题很低

方法分类

- **主观因子选择**：最大程度强调了经济逻辑对因子选择的意义，有效规避了数据挖掘带来的潜在风险
- **系统化因子选择**：
  - 在业界大多起始于一个初始因子库，初始因子库应尽可能保证因子多样性和一定的数量；
  - 希望筛选出逻辑清晰、统计意义下能够支持长期配置且具有信息增量的Alpha因子
  - 方法非常多样，这里介绍一种以Fama-MacBeth回归为主要工具的“因子正交化”方法

因子筛选

步骤

考虑样本时段为$\{t, t = 1, \dots, T\}$，共有M个备选因子$\{X_j, j = 1, \dots, M\}$

- Y_s：经过s次筛选处理后已选择的因子集合，已选择因子的下标集合为Φ_s，$\Phi_s \subseteq \{1,2, \dots, M\}$；
- Z_s：经过s次筛选处理后被剔除的因子的集合，被剔除的因子的下标集合为Ψ_s，$\Psi_s \subseteq \{1,2, \dots, M\}$。

当$s > 0$时，我们进行第$s + 1$次筛选时的流程如下，

1) 在每个时点t，将剩余的备选因子逐一与已选择的因子集合Y_s中的所有因子一起作横截面多元线性回归，得到回归残差项$\{\varepsilon_j, \mathrm{j} \in \Omega_s\}$，其中 $\Omega_s = \{1,2, \dots, M\} - \Phi_s - \Psi_s$；
2) 分别把每个残差因子$\{\varepsilon_j, \mathrm{j} \in \Omega_s\}$和所有当前已选择因子一起作为自变量，进行 Fama-MacBeth 回归测试，得到每个残差因子的回归系数(因子收益率)的t统计量T_{ε_j}，和截面回归的调整R平方的时间序列均值 $\bar{R}^2_{adj,\varepsilon_\mathrm{j}}$；
3) 删除$\left|T_{\varepsilon_j}\right| < 1.96$的因子，并更新第$s + 1$次筛选后被剔除因子的下标集合为$\Psi_{s+1}$；
4) 在所有$\left|T_{\varepsilon_j}\right| \geq 1.96$的因子中，选择$\bar{R}^2_{adj,\varepsilon_\mathrm{j}}$最大的因子$\varepsilon_\mathrm{h}$作为第$s + 1$次筛选所选择的因子，并更新已选择因子下标的集合为$\Phi_{s+1} = \Phi_s \cup \{h\}$，更新已选择因子的集合为$Y_{s+1} = Y_s \cup \{\varepsilon_\mathrm{h}\}$，然后进入下一次筛选；
3) 若对$\forall j \in \Omega_s$，都有$\left|T_{\varepsilon_j}\right| < 2$，那么这表明所有备选因子的回归系数(因子收益率)都不显著，则筛选过程停止。

多因子模型构建

1 因子筛选

2 **因子赋权**

3 非线性多因子模型

因子赋权

方法

在做因子赋权前，一般先对因子进行标准化，以统一量纲。

1. 等权加权
2. IC, ICIR加权
3. 全局方差最小化加权
4. Grinold&Kahn加权

…

因子赋权

等权加权法

- 因子等权加权法即赋予每个因子相等的权重，几乎没有任何参数，也基本上没有对因子预期收益率的分布做出任何假设。

$$w_m^{t-1} = \frac{1}{M}$$

因子赋权

IC, ICIR加权

- **IC**是每个时间截点上因子暴露值和股票下期收益的相关系数。IC值越高意味着该因子的暴露度与未来收益值存在较明显的相关关系。|IC|>0.03，认为因子有效，可以用于区分股票。
- **IR**即信息比率(Information Ratio)，是超额收益的均值与标准差之比，可以根据 IC 近似计算。
  IR = IC的多周期均值/IC的标准方差。

- IC, ICIR加权用过去一段时间的因子 IC 时间序列来计算 IC 均值或 ICIR，以此作为当期的因子权重。
- 假设因子近期的收益或风险调整收益会在短期未来获得延续，也就是所谓的因子动量效应

IC加权

$$w_m^{t-1} = \frac{1}{L}\sum_{l=0}^{L-1} IC_m^{t-1-l}$$

ICIR加权

$$w_m^{t-1} = \frac{avg\left(IC_m^{t-1-l}\right)}{std\left(IC_m^{t-1-l}\right)}\Bigg|_{l=0}^{L-1}$$

L：观察期，过去L期IC值的平均

- L越大，因子权重就越反映因子长期表现，从而也越稳定
- L越小，因子权重就越反映因子短期表现，变化也更频繁

因子赋权

全局方差最小化加权

- 全局方差最小化(Global Minimum Variance, GMV)加权法完全依赖于对因子收益率协方差矩阵的估计，而并不需要对因子收益率进行预测
- 在 GMV 优化严格要求权重非负，也就是说在依据因子逻辑调整过因子方向的前提下，做空因子是不允许的
- 其基本问题形式如下:

$$\arg\min_{\omega} \frac{1}{2}\omega'\Sigma_{IC}\omega,$$

$$s.t.\quad \omega' i = 1, \omega \geq 0$$

- $\sum IC$：IC协方差矩阵
- 估计$\sum IC$的方法：
  1. 样本协方差矩阵$\widehat{\sum IC}$：无偏估计，且是正态假设下的极大似然估计，但是方差较大
  2. 压缩估计量法*：基本思想是使用一个方差小但偏差大的协方差矩阵估计量$\hat{\Phi}$作为目标估计量，和样本协方差矩阵$\widehat{\sum IC}$做一个调和，牺牲部分偏差来获得更稳健的估计量，表达方式如下，参数λ可以通过最小化估计量的二次偏差得到

$$\hat{\Sigma}_{IC,shrink} = \lambda\hat{\Phi} + (1-\lambda)\hat{\Sigma}_{IC}$$

*Ledoit, O., & Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of multivariate analysis*, *88*(2), 365-411.

因子赋权

Grinold&Kahn加权*

- 之前介绍的等权加权完全忽视了因子收益率的分布特 征，IC 加权法聚焦于因子历史收益率，而 GMV 加权法则只关注风险。虽然 ICIR 加权法有同时考虑风险和收益，但它也忽略了因子之间广泛存在的相关性
- Grinold & Kahn 加权是均值方差优化在因子空间的一个标准实践。从理论上讲，G&K 加权法就是在追求因子组合的预期 IR 最大化
- 其基本问题形式如下:

$$\arg\max_{\omega} IR = \frac{\omega' \overline{IC}}{\sqrt{\omega' \Sigma_{IC} \omega}},$$

$$s.t. \quad \omega' i = 1, \omega \geq 0$$

优化问题的两个关键输入：

- $\sum IC$：IC协方差矩阵
- $\overline{IC}$：因子IC的数学期望；
  - 估计起来有挑战，短期因子收益率的时间序列预测的可行性还有待论证，实现起来也十分困难
  - 一般来说，在投资实践中，许多研究人员会直接使用长期的 IC 样本均值来替代对 *IC* 的预测

*Kahn, R., & Grinold, R. (1999). Active Portfolio Management. New York, NY: McGraw-Hil1

多因子模型构建

1 因子筛选

2 因子赋权

3 **非线性多因子模型**

非线形多因子模型初步

线形模型的问题

1. 股票横截面预期收益率与阿尔法因子之间的关系并不一定是线性的，非线性相关性的存在可能会减弱线性多因子模型的预测能力；
2. 线性模型中一般会假设因子收益率 λ^t 与具体个股无关，但我们确实观察到有些因子在不同的股票板块内会体现出完全不同的预测能力和收益特征；
3. 由于线性模型的大量使用，再加上因子的选择也是大同小异，许多机构的持仓都呈现出高度同质化的特征，这一方面降低了策略获取超额收益的有效性，另一方面也增加了模型面临流动性风险的概率。

所以，机器学习……

多因子模型

1 多因子概述

2 因子评价

3 多因子模型构建

4 **投资组合优化**

5 经典因子与模型简述

投资组合优化

1 风险模型

2 组合优化

风险模型

N个资产的时序多元回归模型：

$$R_t^e = \alpha + \beta\lambda_t + \epsilon_t \qquad (1)$$

其中：

$$R_t^e = [R_{1t}^e, R_{2t}^e, \dots, R_{Nt}^e]', \quad \alpha = [\alpha_1, \alpha_1, \dots, \alpha_N,]', \quad \beta = [\beta_1, \beta_2, \dots, \beta_N]',$$
$$\lambda_t = [\lambda_{1t}, \lambda_{2t}, \dots, \lambda_{Kt}]', \quad \epsilon_t = [\epsilon_{1t}, \epsilon_{2t}, \dots, \epsilon_{Nt}]'$$

资产特质性收益率满足：

$$E[\epsilon_t] = 0, cov(\lambda_t, \epsilon_t) = 0$$

则对(1)式两侧求协方差矩阵有：

$$\Sigma = \beta\Sigma_\lambda\beta' + \Sigma_\epsilon \qquad (2)$$

此处ϵ_{it}相互独立，故Σ_ϵ是对角阵

假设A股有4000只股票，那么估计Σ需要$\frac{N(N+1)}{2} = 8{,}002{,}000$个参数

如果使用多因子模型，假设因子数量K=50，那么只需要估计$\frac{K(K+1)}{2} + N = 5{,}275$个参数

Barra多因子模型——CNE5

CNE5包含1个国家因子，P个行业因子，以及Q个风格因子：

$$R_t^e = \lambda_{ct} + \beta_{t-1}^{I_1}\lambda_{I_1 t} + \cdots + \beta_{t-1}^{I_P}\lambda_{I_P t} + \beta_{t-1}^{S_1}\lambda_{S_1 t} + \cdots + \beta_{t-1}^{S_Q}\lambda_{S_Q t} + u_t \qquad (3)$$

所有股票在国家因子λ_{ct}上的暴露为1

$\beta_{t-1}^{I_i}$：行业因子暴露，哑变量，属于该行业为1，否则为0

$\beta_{t-1}^{S_q}$：风格因子暴露

国家因子可以表示为P个行业因子的线性组合，为了保证解的唯一性，对行业因子收益率限制:

$$\sum_{i=1}^{P} s_{I_i}\lambda_{I_i t} = 0 \qquad (4)$$

其中s_{I_i}为属于行业I_i的股票的市值权重之和

风格因子暴露确定：

1. 以公司特征直接作为因子暴露的原始值
2. 对因子暴露进行标准化，减市值加权的均值，再除以标准差

Barra多因子模型——CNE5

假设市场组合对任一风格因子都是中性的，则去均值后，因子暴露满足:

$$\sum_{i=1}^{N} s_i \beta_{it-1}^{S_q} = 0 \qquad (5)$$

其中s_i为属于股票i的市值权重

国家因子实质上就是按市值加权的市场组合，即$R_{Mt}^{e} \approx \lambda_{ct}$：

$$\begin{aligned}
R_{Mt}^{e} &= \lambda_{ct} + \sum_{p=1}^{P} \beta_{Mt-1}^{I_p} \lambda_{I_p t} + \sum_{q=1}^{Q} \beta_{Mt-1}^{S_q} \lambda_{S_q t} + u_{Mt} \\
&= \lambda_{ct} + \sum_{p=1}^{P} \left(\sum_{i=1}^{N} s_i \beta_{it-1}^{I_p} \right) \lambda_{I_p t} + \sum_{q=1}^{Q} \left(\sum_{i=1}^{N} s_i \beta_{it-1}^{S_q} \right) \lambda_{S_q t} + \sum_{i=1}^{N} s_i u_{it} \\
&= \lambda_{ct} + \sum_{p=1}^{P} s_{I_p} \lambda_{I_p t} + 0 + \sum_{i=1}^{N} s_i u_{it} \\
&= \lambda_{ct} + 0 + 0 + \sum_{i=1}^{N} s_i u_{it} \\
&\approx \lambda_{ct}
\end{aligned}$$

CNE5模型求解

(3)式省略下标：

$$R^e = \beta\lambda + u \qquad (6)$$

使用加权最小二乘法(WLS)对(6)式进行求解，假设个股特质性收益率的方差与其市值平方成反比，选择回归权重矩阵W如下：

$$W = \begin{bmatrix} \frac{\sqrt{s_1}}{\sum_{i=1}^{N}\sqrt{s_i}} & 0 & \cdots & 0 \\ 0 & \frac{\sqrt{s_2}}{\sum_{i=1}^{N}\sqrt{s_i}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{\sqrt{s_N}}{\sum_{i=1}^{N}\sqrt{s_i}} \end{bmatrix}$$

CNE5模型求解

构建代表(4)式的约束矩阵C：

$$\begin{bmatrix} \lambda_c \\ \lambda_{I_1} \\ \vdots \\ \lambda_{I_p} \\ \lambda_{S_1} \\ \vdots \\ \lambda_{S_Q} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & -\frac{s_{I_1}}{s_{I_P}} & -\frac{s_{I_2}}{s_{I_P}} & \cdots & -\frac{s_{I_{P-1}}}{s_{I_P}} & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 0 & \cdots & 1 \end{bmatrix} \begin{bmatrix} \lambda_c \\ \lambda_{I_1} \\ \vdots \\ \lambda_{I_{p-1}} \\ \lambda_{S_1} \\ \vdots \\ \lambda_{S_Q} \end{bmatrix}$$

WLS求解可得：

$$\lambda = \Omega R^e$$

其中Ω为纯因子投资组合权重矩阵：

$$\Omega = C(C'\beta'W\beta C)^{-1}C'\beta'W$$

此时因子收益率序列λ_t和特质性收益率序列u_t已知，据此便可估计$\widehat{\Sigma_\lambda}$和$\widehat{\Sigma_\epsilon}$，由此我可以得到：

$$\hat{\Sigma} = \beta\widehat{\Sigma_\lambda}\beta' + \widehat{\Sigma_\epsilon}$$

下面我们将使用$\hat{\Sigma}$进行组合优化。

参考与深入：《因子投资方法与实践》page307

投资组合优化

1 风险模型

2 **组合优化**

组合优化：组合优化的概念和均值方差最优化（MVO）

组合优化的基本概念

组合优化的基本概念：在一定的约束条件下，寻求目标函数的最大化或最小化。

Max f(x)
s.t. conditon1
condition2
....

在构建具体的数学模型之前我们需要刻画出组合的收益与风险

均值方差最优化

均值方差最优化（Mean-Variance optimization）：求在既定风险约束和其它投资约束条件下，使得组合的收益最大化的股票权重。或者求既定收益和其它投资约束下，使得组合的风险最小化的股票权重。

组合优化：投资组合的期望收益和组合风险的表示

投资组合的期望收益

投资组合的t+1时期的期望收益: $R^e_{P(t+1)} = h_P \cdot R^e_{t+1} = h_P \cdot \beta_{t+1} \cdot \lambda^e_{t+1}$

投资组合的主动收益的期望值: $R^e_{PA(t+1)} = R^e_{P(t+1)} - R^e_{B(t+1)} = h_{PA} \cdot R^e_{t+1} = h_{PA} \cdot \beta_{t+1} \cdot \lambda^e_{t+1}$， $h_{PA} = h_P - h_B$，其中h代表权重向量，B代表基准组合。

λ^e_{t+1}的计算方法

- **历史均值法**：用前M期因子历史收益率的均值作为t+ 1期因子的预期收益率： $\lambda^e_{t+1} = \frac{\sum_{i=t-M+1}^{t} \lambda_i}{M}$
- **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。

$$\lambda^e_{t+1} = EWMA_t = w * \lambda_t + (1 - w) * EWMA_{t-1}$$

0 < w< 1，w越接近1，则当前观察值权重越大，之前的历史值权重越小。w数值越小，则估计值约平稳，因为如果当前数据λ_t发生突变，那么受到数据突变的影响越小。

更多有关λ^e_{t+1}的计算参考华泰研报《华泰多因子模型体系初探华泰多因子系列之一》第23-24页

组合优化：投资组合的期望收益和组合风险的表示

投资组合的风险

组合的方差： Var($R^e_{P(t+1)}$)= Var($h_P \cdot R^e_{t+1}$)= $h_p\ Var(R^e_{t+1})h'_p$=$h_P \Sigma h'_p$，$Var(R^e_{t+1}) = \Sigma = \beta\Sigma_\lambda\beta' + \Sigma_\epsilon$，为之前风险模型中所估计出的协方差矩阵

组合的跟踪误差： 投资组合的跟踪误差为组合主动收益的标准差，记为σ_{PA}，σ_{PA}^2=Var($R_P - R_B$)= $h_{PA}\Sigma h'_{PA}$

信息比率： IR= R^e_{PA}/σ_{PA}，表示相较于基准组合额外承担一单位风险所带来的收益，可以用于业绩归因，信息比率越大越好。

组合优化：均值方差最优化的基本模型

前面提到最优化问题可以为两种思路：一种是既定风险约束和其它投资约束条件下，使得组合的收益最大化。另一种是既定收益和其它投资约束下，使得组合的风险最小化。这两种思路对应的基本数学模型如下：

既定风险约束下投资收益最大化：

$$\text{Max } h_P \cdot R_{t+1}^e$$

$$\text{s.t. } \sum_{j=1}^{n} h_{pj} = 1,\ h_P \Sigma h_p' \leq \sigma^2$$

$$hpj \geq 0, j=1,2,3\ldots\ldots n$$

既定收益约束下风险最小化：

$$\text{Min } \frac{1}{2} h_P \Sigma h_p'$$

$$\text{s.t. } \sum_{j=1}^{n} h_{pj} = 1,\ h_P \cdot R_{t+1}^e \geq r$$

$$h_{pj} \geq 0, j=1,2,3\ldots\ldots n$$

组合优化：其它约束

我们选用既定风险约束下使组合收益最大化的模型形式，探讨一些除了风险约束以外的其它约束，即个股权重上下限约束，行业权重约束，因子暴露约束。在具体的投资中可以加入其它约束，不限于所提及的这三个，根据所面对的具体投资约束而定。

个股权重上下限约束

对于纯多头头寸，无基准组合，权重约束为： $0 \le h_{Pj} \le h_j^{upper}$

对于纯多头头寸，存在基准组合，权重约束为： $\sum h_{PAj}=0$，$-h_{Bj} \le h_{PAj} \le h_j^{upper}$，

对于存在基准组合的约束条件可由下面简单过程得出：

由$\sum h_{PAj}=\sum (h_{Pj}-h_{Bj}) = \sum h_{Pj} - \sum h_{Bj} = 1-1=0$。（有超配就有低配，最终主动权重和为0）

由$0 \le h_{Pj} \le h_j^{upper}$,可知$-h_{Bj} \le h_{PAj}=h_{Pj}-h_{Bj} \le h_j^{upper} - h_{Bj} \le h_j^{upper}$

组合优化：其它约束

行业权重约束

由于多因子模型本质上是一个统计套利模型，不适合对市场因子和行业因子进行收益预测和风险管理。因此目前国内市场上多因子模型最流行的用法是：

1. 通过股指期货对冲组合的市场风险（市值对冲）；
2. 通过行业中性对冲组合的行业风险（以业绩基准的行业权重为基准进行对齐，即组合 在每个行业上的权重分配与业绩基准一致）。

对于任意一只股票，其行业哑变量$(0,0,\cdots,1,\cdots 0)$，对于所有股票组成的哑变量矩阵S：

$$S=\begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1S} \\ s_{21} & s_{22} & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ s_{N1} & \cdots & \cdots & s_{NS} \end{bmatrix}$$

所加入的行业权重约束为：$\sum_{j=1}^{N} h_{PAj} * s_{ji} = 0$　　即主动的行业暴露为0

组合优化：其它约束

组合因子暴露

实际操作中，需要对因子暴露进行约束，避免在单个因子上暴露过多的风险。

组合P在因子k上的暴露为： $\sum_{j=1}^{N} h_{pj}\,\beta_{jK}$

主动的因子k上的暴露为： $\sum_{j=1}^{N} h_{PAj}\,\beta_{jK}$

该结果可由下面过程得到：

因子荷载矩阵为：β_{t+1} =（β_{jk}）$_{NK}$，j=1,2,3..N,k=1,23,….K

组合的收益:$R_P = h_P \cdot R^e_{t+1} = h_P \cdot \beta_{t+1} \cdot \lambda^e_{t+1}$ =（ h_{P1}， h_{P2}, … h_{PN}）$\cdot \begin{pmatrix} \beta_{11} & \cdots & \beta_{1K} \\ \vdots & \ddots & \vdots \\ \beta_{N1} & \cdots & \beta_{NK} \end{pmatrix} \cdot \lambda^e_{t+1}$ =($\sum_{j=1}^{N} h_{pj}\,\beta_{j1}$, $\sum_{j=1}^{N} h_{pj}\,\beta_{j2}$,…

$\sum_{j=1}^{N} h_{pj}\,\beta_{jK}) \cdot \lambda^e_{t+1}$

组合因子暴露约束

如果对于因子k的暴露上限为U_k，则要求：

不存在基准组合：$\left|\sum_{j=1}^{N} h_{pj}\,\beta_{jK}\right| \le U_k$

存在基准组合：$\left|\sum_{j=1}^{N} h_{PAj}\beta_{jK}\right| \le U_k$

组合优化：其它约束

同时考虑到上面的三个约束，且存在基准组合，最终的最优化问题为：

$$\text{Max } h_P \cdot R_{t+1}^e$$

$$\text{s.t. } \sum_{j=1}^{n} h_{pj} = 1,\ h_P \Sigma h_p' \le \sigma^2$$

$$hpj \ge 0,\ \text{j=1,2,3……N}$$

$$\sum \text{h}_{PAj} = 0,\ -h_{Bj} \le h_{PAj} \le h_j^{upper}$$

$$\sum_{j=1}^{N} h_{\text{PA}j}\, \text{s}_{\text{ji}} = 0,\ \text{i=1,2,3…..I(表示行业个数）}$$

$$\left| \sum_{j=1}^{N} h_{\text{PA}j} \beta_{\text{jk}} \right| \le \text{U}_k\ \text{k=1,2,3…K}$$

组合优化：均值方差最优化方法存在的一些问题

传统MVO的缺陷

传统的MVO最优化在运用时有五大缺陷：

- **参数估计误差大，带来垃圾进垃圾出的后果。**传统的MVO是基于对股票期望收益率和协方差矩阵的估计进行最优化的，由于股票数量巨大，容易导致较大的估计误差。
- **结果对参数输入非常敏感。**而且相较于估计的协方差矩阵，最优化结果对输入的均值更为敏感。
- **优化结果可能过于集中，使得权重集中于个别证券。**
- **换手率高，交易成本太大。**由于对输入较为敏感，随着时间变化，输入的均值和协方差矩阵发生微小变化，结果的权重发生很大的变化，导致换手率高。
- **较差的样本外表现。**DeMiguel et al.（2009）的结果表明，基于历史数据的均值方差组合，由于估计误差的存在，在样本外表现很难超越等权重组合。

组合优化：对MVO的改进

多因子模型对传统MVO的改进

- **多因子模型减少输入参数的估计误差：**多因子模型用因子的预期收益，因子协方差矩阵，特异收益协方差矩阵，大大减少了所需估计的参数数量，在一定程度上已经减少了估计误差，
- **防止优化结果过于集中：**在约束中加入股票权重约束，因子暴露约束可以解决MVO得到的结果过于集中的问题。

对多因子模型MVO的改进

改进MVO的一个办法是改进模型输入的参数，即因子的预期收益，因子协方差矩阵，特异收益协方差矩阵。
一个可行的办法是压缩估计法。
压缩估计法：用样本所估计出的参数值和预先设定的一个参数值取加权平均，作为最终输入参数值的估计。
以估计因子预期收益为例：
先找到市场上一个指数的权重，利用样本的估计的协方差矩阵，假定该权重满足MVO，那么能倒推出一个该权重下对应的因子预期收益（λ_{t+1}^{g}），再和样本所得的因子预期收益(λ_{t+1}^{s})做加权平均得出最终因子预期收益(λ_{t+1}^{e})作为模型输入参数。

$$\lambda_{t+1}^{e}=\mathrm{k}\,\lambda_{t+1}^{g}+(1-\mathrm{k})\,\lambda_{t+1}^{s}$$

更多关于MVO的缺陷和参数估计的改进方法参考https://www.factorwar.com/research/asset-allocation/ **《如何分配资金？组合优化的是是非非》**

多因子模型

1 多因子概述

2 因子评价

3 多因子模型构建

4 投资组合优化

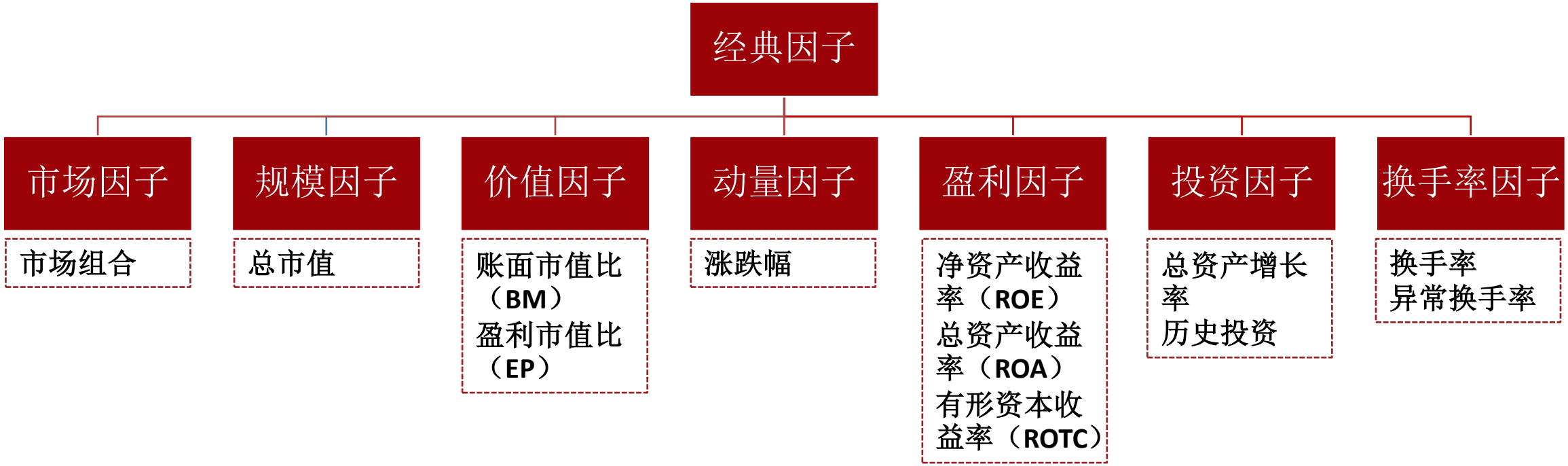
5 **经典因子与模型简述**

经典因子与模型简述

1 经典因子

2 主流多因子模型

经典因子


经典因子
市场因子
市场组合
规模因子
总市值
价值因子
账面市值比（BM）
盈利市值比（EP）
动量因子
涨跌幅
盈利因子
净资产收益率（ROE）
总资产收益率（ROA）
有形资本收益率（ROTC）
投资因子
总资产增长率
历史投资
换手率因子
换手率
异常换手率

经典因子

- 概念
- 常用构建方法
- 成因：因子背后的原因是研究热点之一
- 实证分析：
  - 时间跨度：2000.1.1-2019.12.31；
  - 股票池：除科创版外的所有A股；
  - 调仓频率：每月末
  - 股票权重：等权重和总市值加权
  - ……
- 因子评价

项目	说明
数据来源	Wind 和 Tushare
收益计算	后复权收盘价
长期停牌股复牌首日收益率	压缩处理
缺失值处理	向前填充或者不填充
最少交易日	不少于计算窗口 2/3 的数据
财务数据处理	遵循 point-in-time 原则
股票范围	全部 A 股，不剔除金融股
时间区间	2000 年 1 月 1 日至 2019 年 12 月 31 日
黑名单	待退市股、净资产为负股、次新股和风险警示股
不可交易股	停牌股、一字涨停股、一字跌停股
异常值处理	缩尾法
单变量排序	按变量取值从小到大将股票分成 10 组
双重排序	独立双重排序，即分别使用两个变量各自将股票分为 5 组，然后交叉得到 25 个投资组合[3]
投资组合内股票权重	等权重或按总市值加权
调仓频率	每月末
交易成本	设为 0

续表

来源：石川等，《因子投资》

市场因子

概 念

- 多因子模型源于资本资产定价模型CAPM，CAPM是仅包括**市场因子**的单因子模型，其数学表达式为：

$$E[R_i] - R_f = \beta_i\left(E[R_M] - R_f\right)$$

- R_{M}为**市场组合**的收益率，市场组合也被称为市场因子；
- 由于市场因子的预期超额收益在长期内为正，所以CAPM 模型意味着“高风险，高收益”；

因子评价

- 总体来看，市场因子对于解释股票收益的时间序列变化很有帮助，但诸多证据也显示单凭市场因子不能很好地解释股票收益的横截面差异，这使得引入其他因子、构建多因子模型成为必要的选项。

市场因子

实证分析

- （1）检验市场组合的超额收益率对个股超额收益在时序上的波动的解释：
  - 使用每支个股的全部超额收益率和市场组合同期超额收益率，通过时序回归求出R^2。
- （2）检验CAPM（高风险，高收益）；
  - T月末，采用过去252个交易日中个股超额收益和市场组合超额收益，通过时序回归计算出个股市场的β，依照β大小把所有股票分成10档；
  - 如果CAPM成立，那么这10档收益率应该和每组β正相关。

(1) 市场因子能够在一定程度上解释个股收益率的时序波动

表 3.12　R^2 的描述性统计

均值	标准差	最小值	25% 分位数	中位数	75% 分位数	最大值
29.09%	13.85%	0.00%	19.83%	30.52%	38.58%	85.25%

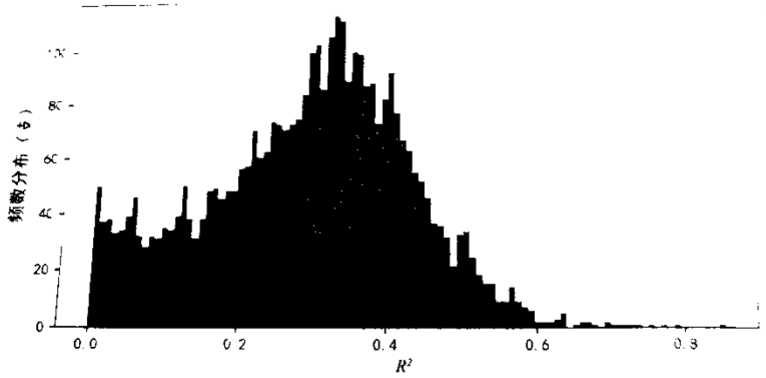
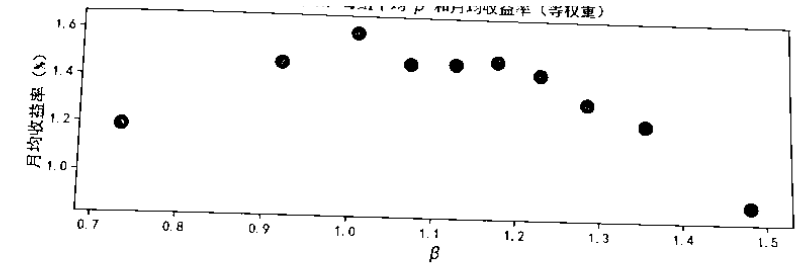
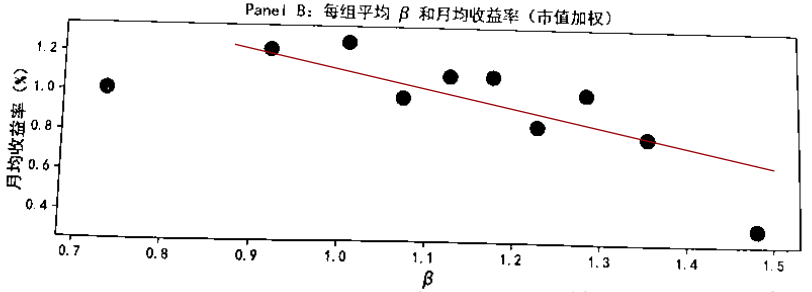


图 3.11　时序回归 R^2 分布

表 3.13　10 个投资组合的市场 β（实证期内时间平均）

Low	2	3	4	5	6	7	8	9	High
0.73	0.92	1.01	1.07	1.13	1.18	1.23	1.29	1.36	1.48





来源：石川等，《因子投资》

(2) 收益率和β没有显著的正相关关系，CAPM被A股的数据拒绝：

- 在CAPM的基础上加入其他风格因子，找到适合A股市场的多因子模型；
- A股市场可能存在著名的低β异象

规模因子

概 念

- **规模效应**：小市值股票有着比大市值股票更好的表现；
- **规模因子**：反映规模效应的因子即为规模因子，因著名的Fama-French三因子模型而家喻户晓，常用总市值构建规模因子；
- **成因**：
  - 风险补偿说：小市值公司普遍在过去遭遇过困境使得市值大幅下滑，即所谓“堕落的天使”；
  - 投资者行为；退市偏差；少数股票的极端收益；季节效应；模型设定偏误.....

因子评价

- 早期实证证据普遍支持股票市场中存在显著的规模效应，而新近的研究则表明，随着投资者熟知规模效应，规模因子表现也趋于平庸，甚至其长期超额收益率逐渐变得不再显著。

规模因子

实证分析

- **规模因子构建：**总市值 = 股票收盘价 × 总股本
- **步骤：**
  - 每月末将股票总市值从低到高分成10组，并进行描述性统计以发现不同组合在常见指标（如ROE、年化波动率、P/B等）上是否存在异常；
  - 计算不同投资组合的月均收益率及t值；
  - 因子收益率：做多小市值small组、做空大市值Big组（月均收益率：1.41%（等权重）；1.40%（市值加权）

小市值效应在A股上显著，即投资组合的收益率随着市值的增加而下降

表 3.15　市值单变量排序月均收益率（%）

Panel A：等权重										
Small	2	3	4	5	6	7	8	9	Big	Small–Big
2.35	1.81	1.52	1.36	1.21	1.16	0.96	1.02	0.90	0.94	1.41
(3.02)	(2.33)	(2.04)	(1.84)	(1.65)	(1.58)	(1.34)	(1.50)	(1.33)	(1.41)	(3.11)
Panel B：市值加权										
Small	2	3	4	5	6	7	8	9	High	Small–Big
2.34	1.81	1.52	1.36	1.21	1.15	0.97	1.01	0.90	0.94	1.40
(2.99)	(2.33)	(2.04)	(1.84)	(1.65)	(1.58)	(1.36)	(1.48)	(1.34)	(1.41)	(2.90)

括号内为经 Newey–West 调整的 t-值。

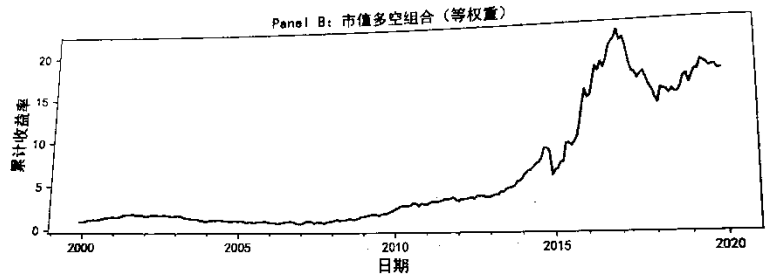


图 3.15　规模因子累计收益率（等权重）

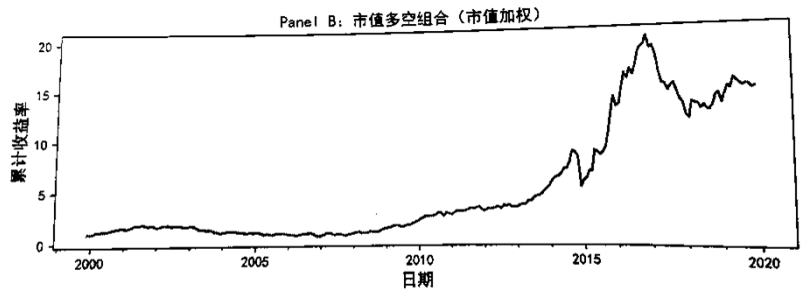


图 3.16　规模因子累计收益率（市值加权）

来源：石川等，《因子投资》

做多小市值股票在A股上能够获得非常丰富的回报：

- 没有考虑交易费用或滑点；
- 年化波动率大；
- 随着A股市场变得更加有效，大市值蓝筹优质股自2017年走出了一波上涨行情，规模因子遭到大幅回撤

价值因子

概 念

- **价值因子**：相比估值较高的股票，估值较低的股票有着更高的预期收益率；
- **常见估值指标**：

$$\text{账面市值比(BM)} = \frac{\text{每股账面价值}}{\text{每股价格}}$$

$$\text{盈利市值比}(EP) = \frac{\text{每股净利润}}{\text{每股价格}}$$

- **成因**：
  - 财务风险困境假说：高BM很可能反映着更高的财务困境风险；
  - 投资者行为偏差

因子评价

- 虽然价值因子历史悠久，但价值因子近年来遭遇了不少挑战，并且自2008年以来的后金融危机时代表现并不好。由于价值因子在最近十年的惨淡表现，很多学者试图从金融学和会计学原理出发，通过改造BM来提升价值因子的表现。

价值因子

实证分析

- **规模因子构建：**BM = 归股东权益合计（不含少数股东权益） ÷ 总市值
- **步骤：**
  - 每月末将股票BM从低到高分成10组，并进行描述性统计以发现不同组合在常见指标（如ROE、年化波动率、P/B等）上是否存在异常；
  - 运用单变量排序法计算不同投资组合的月均收益率及规模因子收益率；
  - 使用BM和市值进行双重排序法计算不同投资组合的月均收益率及规模因子收益率（降低市值对BM 的影响）。

表 3.18　BM 和市值独立双重排序月均收益率（%）

Panel A：等权重

	Low	2	3	4	High	High–Low
Small	1.42	2.02	2.27	2.24	2.35	0.93
	(1.81)	(2.60)	(2.94)	(2.87)	(3.03)	(4.76)
2	0.92	1.16	1.52	1.74	1.74	0.82
	(1.27)	(1.57)	(2.02)	(2.31)	(2.32)	(3.38)
3	0.90	1.08	1.08	1.36	1.45	0.56
	(1.20)	(1.51)	(1.51)	(1.81)	(1.91)	(2.12)
4	0.69	0.84	0.98	1.24	1.36	0.68
	(1.02)	(1.24)	(1.37)	(1.65)	(1.79)	(2.07)
Large	0.76	0.61	0.95	1.13	1.21	0.45
	(1.15)	(0.90)	(1.34)	(1.59)	(1.70)	(1.22)
平均	0.94	1.14	1.36	1.54	1.62	0.69
	(1.34)	(1.63)	(1.90)	(2.10)	(2.20)	(2.89)

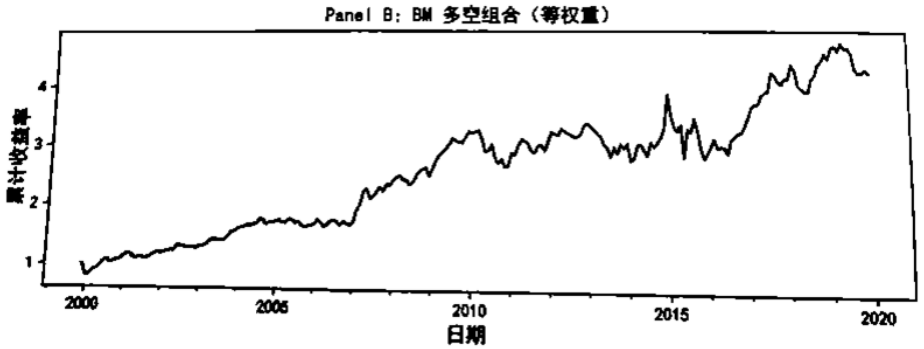


图 3.18　双重排序构建的价值因子累计收益率（等权重）

总体上看，BM与收益率存在正相关关系，但BM在小市值中的作用要远远高于在大市值中的作用。

来源：石川等，《因子投资》

动量因子

概 念

- **动量因子**：股票间的相对强弱趋势会延续，“强者恒强，弱智恒弱”；
- **构建方法**：做多过去一段时间表现最好的股票（赢家组合）、同时做空这段时间表现最差的股票（输家组合）；
- **成因**：
  - 投资者行为偏差：投资者对私有信息的过度自信及其有偏的业绩自我归因；
  - 系统性风险敞口。

因子评价

- 动量因子是一个颇受争议的因子。有研究提供了动量因子在全球多个市场中广泛存在的证据，但在日本和中国A股市场中，动量效应的表现却惨不忍睹。其背后的原因还值得进一步探索。

动量因子

实证分析

- **动量因子构建：**在t月末将t-12到t-1之间的11个月内的总收益作为动量因子的排序变量
- **步骤：**
  - 每月末将股票按照动量变量的取值从低到高分成10组，并进行描述性统计以发现不同组合在常见指标（如ROE、年化波动率、P/B等）上是否存在异常；
  - 运用单变量排序法计算不同投资组合的月均收益率及动量因子收益率；
  - 使用动量和市值进行双重排序法计算不同投资组合的月均收益率及规模因子收益率（降低市值对动量因子的影响）。

表 3.21　动量和市值独立双重排序月均收益率（%）

Panel A：等权重

	Low	2	3	4	High	High–Low
Small	1.94 (2.56)	2.31 (2.85)	2.22 (2.85)	1.93 (2.42)	1.70 (2.17)	−0.23 (−0.91)
2	1.48 (1.92)	1.48 (1.93)	1.51 (2.08)	1.48 (2.04)	1.39 (1.91)	−0.09 (−0.39)
3	1.14 (1.50)	1.29 (1.71)	1.21 (1.68)	1.14 (1.58)	1.14 (1.55)	−0.01 (−0.03)
4	0.87 (1.19)	1.04 (1.40)	1.05 (1.50)	0.90 (1.30)	1.09 (1.60)	0.23 (0.89)
Large	0.51 (0.70)	0.82 (1.20)	1.05 (1.50)	0.89 (1.34)	0.84 (1.29)	0.34 (0.96)
平均	1.19 (1.61)	1.39 (1.88)	1.41 (1.98)	1.27 (1.79)	1.23 (1.78)	0.05 (0.22)

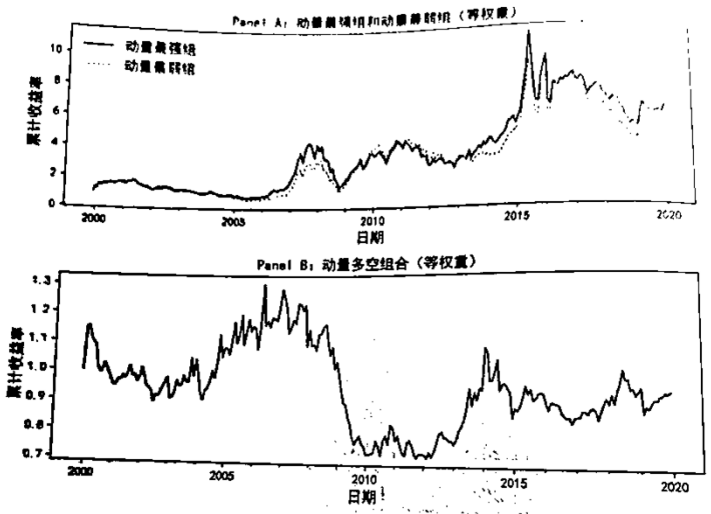


图 3.21　双重排序构建的动量因子累计收益率（等权重）

来源：石川等，《因子投资》

动量效应仅微弱地存在于A股大市值股票中，而对于小市值的股票则更多地表现出反转，说明以学术界中常见方法构建的动量效应并不存在于A股市场中

盈利因子

概 念

- **盈利能力**：企业利用已有资源获取利润的能力；
- **常见盈利因子**：

$$\text{净资产收益率(ROE)} = \frac{\text{归属于股东的净利润}}{\text{净资产}}$$

$$\text{总资产收益率(ROA)} = \frac{\text{归属于股东和债权人的净利润}}{\text{总资产}}$$

$$\text{有形资本收益率(ROTC)} = \frac{\text{息税前利润}}{\text{有形资本}}$$

- **理论基础**：
  - 股利贴现模型：Fama and French从股利贴现模型出发推导了预期盈利和股票预期收益率之间的关系
  - 实体投资经济学理论。

因子评价

- 和其他依赖经验研究的因子相比，盈利因子的理论基础要扎实很多。除盈利水平外，不少研究还探索了盈利质量、盈利波动和盈利增长等维度，这极大地拓展了盈利因子的内涵。

盈利因子

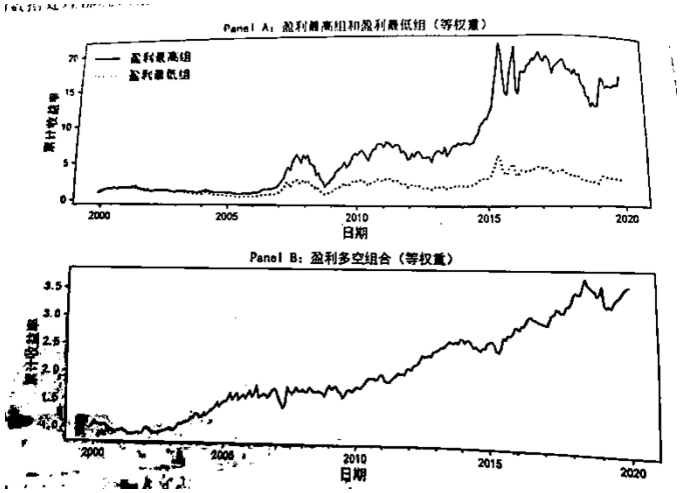
实证分析

- **盈利因子构建：ROE** = 最近**12**个月营业利润 ÷ 股东权益（不含少数股东权益）最近四个报告期均值
- **步骤：**
  - 每月末将股票按照ROE的取值从低到高分成10组，并进行描述性统计以发现不同组合在常见指标（如ROE、年化波动率、P/B等）上是否存在异常；
  - 运用单变量排序法计算不同投资组合的月均收益率及ROE收益率；
  - 使用ROE和市值进行双重排序法计算不同投资组合的月均收益率及规模因子收益率（降低市值对盈利因子的影响）。

表 3.24 ROE(TTM) 和市值独立双重排序月均收益率（%）

Panel A：等权重

	Low	2	3	4	High	High-Low
Small	1.95 (2.43)	2.09 (2.65)	2.20 (2.90)	2.18 (2.94)	2.17 (2.77)	0.22 (1.03)
2	1.18 (1.51)	1.28 (1.72)	1.54 (2.07)	1.85 (2.59)	1.69 (2.29)	0.52 (2.36)
3	0.81 (1.06)	1.07 (1.38)	1.15 (1.58)	1.51 (2.12)	1.40 (2.00)	0.58 (2.73)
4	0.44 (0.60)	0.81 (1.09)	0.90 (1.25)	1.17 (1.76)	1.34 (1.97)	0.90 (4.14)
Large	0.57 (0.77)	0.52 (0.71)	0.66 (0.98)	0.74 (1.14)	1.27 (1.91)	0.70 (2.28)
平均	0.99 (1.32)	1.15 (1.55)	1.29 (1.81)	1.49 (2.18)	1.58 (2.27)	0.58 (3.27)



来源：石川等，《因子投资》

盈利因子在中间市值档位内更加显著，在大市值和小市值中被削弱。总体而言，ROE除对small 组内的其余股票都有很好地解释他们预期收益率的截面差异的能力。

投资因子

概 念

- **投资效应**：当期投资较多的公司相比于投资较少的公司，在未来的预期收益率更低，即投资和预期收益之间呈现负相关；
- **常见投资因子**：历史投资；异常资本投资；总资产增长率等；
- **成因**：
- 行为偏差：有着高额投资的公司往往有过度投资的倾向；
- 系统性风险溢价；
- 错误定价。

因子评价

- 投资因子是近几年被逐渐认可的一个因子，大量学术研究成果表明投资因子存在于多个国家的股票市场中，但已有研究大多认为A股市场不存在显著的投资效应。

投资因子

实证分析

- **投资因子构建——总资产增长率：**上市公司年报中披露的总资产计算总资产增长率
- **步骤：**
  - 每月末将股票按照总资产增长率的取值从低到高分成10组，并进行描述性统计以发现不同组合在常见指标（如ROE、年化波动率、P/B等）上是否存在异常；
  - 运用单变量排序法计算不同投资组合的月均收益率及总资产增长率的收益率；
  - 使用总资产增长率和市值进行双重排序法计算不同投资组合的月均收益率及投资因子收益率（降低市值对投资因子的影响）；（低投资的月均收益率低于高投资）；
  - 使用总资产增长率和ROE（盈利因子）进行双重排序法计算不同投资组合的月均收益率及投资因子收益率（降低ROE对投资因子的影响）；

表 3.28　ROA 和总资产增长率条件双重排序月均收益率（%）

Panel A：等权重

	Low	2	3	4	High	Low-High
ROA Low	1.27 (1.60)	1.25 (1.68)	1.32 (1.80)	1.35 (1.77)	1.26 (1.69)	0.01 (0.07)
2	1.47 (1.94)	1.51 (1.98)	1.33 (1.83)	1.20 (1.62)	1.23 (1.65)	0.24 (1.15)
3	1.37 (1.83)	1.35 (1.83)	1.34 (1.90)	1.23 (1.79)	1.33 (1.82)	0.04 (0.20)
4	1.27 (1.77)	1.43 (2.02)	1.48 (2.17)	1.29 (1.89)	1.48 (2.05)	−0.21 (−1.24)
ROA High	1.31 (1.95)	1.35 (2.05)	1.47 (2.29)	1.40 (2.16)	1.25 (1.85)	0.06 (0.29)
平均	1.31 (1.83)	1.38 (1.92)	1.39 (2.01)	1.30 (1.86)	1.31 (1.83)	0.03 (0.19)

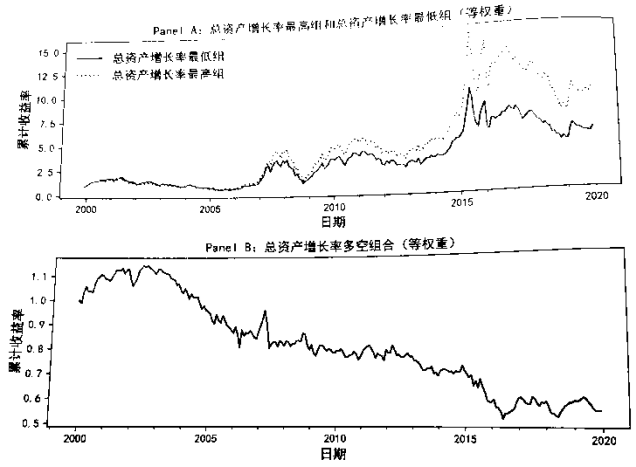


图 3.27　双重排序构建的投资因子累计收益率（等权重）

来源：石川等，《因子投资》

无法在A股上观察到显著的投资因子

换手率因子

概 念

- **换手率因子起源**：成交量对股票未来表现有显著影响，换手率作为标准化的成交量指标受到关注；
- **共识**：换手率与预期收益之间呈负相关；
- **成因**：
  - 股价崩盘风险：换手率相对于过去6个月的趋势显著增长且在过去36个月又显著正收益的股票面临最高的崩盘风险；
  - 行为偏差；
  - 套利限制。

因子评价

- 早期相关研究大多将成交量和换手率当作流动性的一种度量方法，但随着Amihud（2002）将非流动性定义为单位成交金额对应的平均收益变化，换手率便不再适合作为流动性的代理变量，而应该将其视作趋势是否能延续的表征。虽然学术界与业界的共识是换手率与预期收益之间呈负相关，但实证结果也有例外，如在A股市场中发现更高的换手率拥有更高的价格。

换手率因子

实证分析

- **换手率因子构建——异常换手率：**过去21个交易日的平均换手率和过去252个交易日的平均换手率的比值；
- **步骤：**
  - 每月末将股票按照异常换手率的取值从低到高分成10组，并进行描述性统计以发现不同组合在常见指标（如ROE、年化波动率、P/B等）上是否存在异常；（换手率与其他变量的相关性非常低，这是一个非常优秀的性质）
  - 运用单变量排序法计算不同投资组合的月均收益率及异常换手率的收益率；
  - 使用异常换手率和市值进行双重排序法计算不同投资组合的月均收益率及换手率因子收益率（降低市值对换手率因子的影响）；

表 3.31　异常换手率和市值独立双重排序月均收益率（%）

Panel A：等权重

	Low	2	3	4	High	Low-High
Small	2.69 (3.45)	2.39 (3.04)	2.16 (2.73)	1.97 (2.53)	0.99 (1.28)	1.70 (8.08)
2	1.97 (2.68)	1.77 (2.44)	1.64 (2.20)	1.38 (1.81)	0.47 (0.62)	1.50 (6.68)
3	1.74 (2.34)	1.56 (2.06)	1.42 (1.97)	1.01 (1.36)	0.43 (0.58)	1.32 (5.97)
4	1.32 (1.92)	1.45 (2.04)	1.21 (1.71)	0.96 (1.35)	0.28 (0.39)	1.04 (5.04)
Large	1.03 (1.55)	1.21 (1.78)	1.20 (1.74)	0.95 (1.39)	0.41 (0.57)	0.62 (2.23)
平均	1.75 (2.47)	1.68 (2.34)	1.53 (2.13)	1.25 (1.74)	0.52 (0.71)	1.24 (6.67)

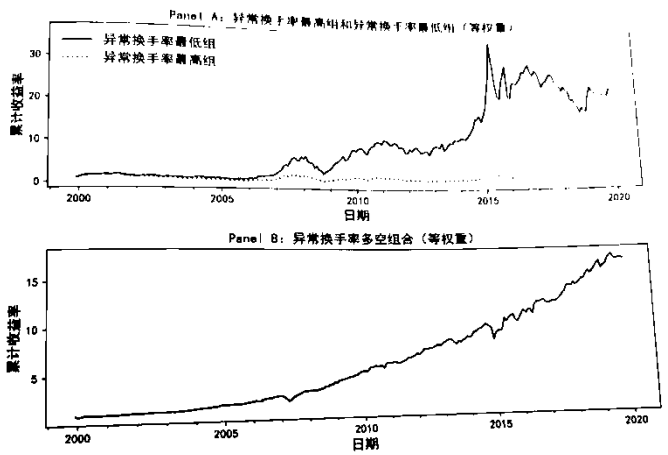


图 3.30　双重排序构建的换手率因子累计收益率（等权重）

来源：石川等，《因子投资》

换手率因子的月均收益率非常显著，并且低换手率异象在小市值中更加显著；换手率因子的累计收益率曲线也十分平顺：

- 未考虑交易费用；
- 空头限制

经典因子与模型简述

1 经典因子

2 **主流多因子模型**

主流多因子模型

- 优秀多因子模型回答两个问题：
  - 资产收益率背后的驱动因素（模型中包括哪些因子）；
  - 每个驱动因素背后的原因（因子解释）。

——《因子投资》

Fama-French 三因子模型

- 1993年，Fama and French在CAPM的基础上，加入了价值（High-Minus-Low, HML）和规模（Small-Minus-Big, SMB）两个因子，提出了三因子模型，它也是多因子模型的开山鼻祖：

$$E[R_i] - R_f = \beta_{i,MKT}\left(E[R_M] - R_f\right) + \beta_{i,SMB}E[R_{SMB}] + \beta_{i,HML}E[R_{HML}]$$

- $E[R_M]$：**市场组合**预期收益率；
- $E[R_{SMB}]$：**规模因子**（SMB）的预期收益率（市值）；
- $E[R_{HML}]$：**价值因子**（HML）的预期收益率（BM）；
- $\beta_{i,MKT}, \beta_{i,SMB}, \beta_{i,HML}$：个股$i$在相应因子上的暴露。

		BM		
		High（高组）	Middle（中间组）	Low（低组）
市值	Small（小市值）	S/H	S/M	S/L
	Big（大市值）	B/H	B/M	B/L

市值分组依据：NYSE 中位数；
BM 分组依据：NYSE 30% 和 70% 分位数。

出处：川总写量化

- Fama-French 三因子模型被提出后逐步取代了 CAPM 成为资产定价的第一范式。

$$\begin{aligned}
\text{SMB} &= \frac{1}{3}(\text{S/H} + \text{S/M} + \text{S/L}) - \frac{1}{3}(\text{B/H} + \text{B/M} + \text{B/L}) \\
\text{HML} &= \frac{1}{2}(\text{S/H} + \text{B/H}) - \frac{1}{2}(\text{S/L} + \text{B/L})
\end{aligned}$$

Carhart四因子模型

- Fama-French 三因子模型虽然有足够的开创性，但是“适用性”却有限，有很多其无法解释的异象。在众多异象中，最显著的当属截面动量异象。1997年，Carhart 在 Fama-French 三因子模型中加入了截面动量因子（取动量英文单词前三个字母，记为 MOM）并提出了 Carhart 四因子模型：

$$E[R_i] - R_f = \beta_{i,MKT}\left(E[R_M] - R_f\right) + \beta_{i,SMB}E[R_{SMB}] + \beta_{i,HML}E[R_{HML}] + \beta_{i,MOM}E[R_{MOM}]$$

- $E[R_M]$：**市场组合**预期收益率；
- $E[R_{SMB}]$：**规模因子**（SMB）的预期收益率（市值）；
- $E[R_{HML}]$：**价值因子**（HML）的预期收益率（BM）；
- $E[R_{MOM}]$：**动量因子**（MOM）的预期收益率（历史收益率）；
- $\beta_{i,MKT}, \beta_{i,SMB}, \beta_{i,HML}, \beta_{i,MOM}$：个股$i$在相应因子上的暴露。

动 量

- t月末将所有股票按 t - 12 到 t - 1 这 11 个月的总收益率排序，并通过做多排名前 30% 同时做空排名后 30% 的股票构建动量因子;
- 值得注意的是，Carhart 并未使用动量和市值进行双重排序。

Novy-Marx四因子模型

- 2013 年，Novy-Marx 指出盈利能力和未来预期收益率密切相关，并由此提出了一个四因子模型：

$$E[R_i] - R_f = \beta_{i,MKT}\left(E[R_M] - R_f\right) + \beta_{i,HML}E[R_{HML}] + \beta_{i,UMD}E[R_{UMD}] + \beta_{i,PMU}E[R_{PMU}]$$

- $E[R_M]$：**市场组合**预期收益率；
- $E[R_{HML}]$：**价值因子**（HML）的预期收益率（BM）；
- $E[R_{UMD}]$：**动量因子**（UMD）的预期收益率（历史收益率）；
- $E[R_{PMU}]$：**盈利因子**（PMU）的预期收益率（毛利润）；
- $\beta_{i,MKT}, \beta_{i,HML}, \beta_{i,UMD}, \beta_{i,PMU}$：个股$i$在相应因子上的暴露。

		Gross Profitability		
		Profitable（盈利）	Neutral（中性）	Unprofitable（不盈利）
市值	Small（小市值）	S/P	S/N	S/U
	Big（大市值）	B/P	B/N	B/U

市值分组依据：NYSE 中位数；
GP 分组依据：NYSE 30% 和 70% 分位数。

出处：川总写量化

- 行业中性处理：做多一支股票的同时按同等权重做空该股票所属的行业指数，从而得到行业中性化后的投资组合。

$$\mathrm{PMU} = \frac{1}{2}(\mathrm{S/P} + \mathrm{B/P}) - \frac{1}{2}(\mathrm{S/U} + \mathrm{B/U})$$

Fama-French五因子模型

- 2015 年，Fama and French 在他们的三因子模型基础上添加了盈利和投资两个因子，提出了新的五因子模型：

$$E[R_i] - R_f = \beta_{i,MKT}\left(E[R_M] - R_f\right) + \beta_{i,SMB}E[R_{SMB}] + \beta_{i,HML}E[R_{HML}] + \beta_{i,RNW}E[R_{RNW}] + \beta_{i,CMA}E[R_{CMA}]$$

- $E[R_M]$：**市场组合**预期收益率；
- $E[R_{SMB}]$：**规模因子**（SMB）的预期收益率（市值）；
- $E[R_{HML}]$：**价值因子**（HML）的预期收益率（BM）；
- $E[R_{RNW}]$：**盈利因子**（RNW）的预期收益率（ROE）；
- $E[R_{CMA}]$：**投资因子**（CMA）的预期收益率（总资产变化率）；
- $\beta_{i,MKT}$，$\beta_{i,SMB}$，$\beta_{i,HML}$，$\beta_{i,RNW}$，$\beta_{i,CMA}$：个股i在相应因子上的暴露。

ROE

市值	Robust（稳健）	Neutral（中性）	Weak（疲软）
Small（小市值）	S/R	S/N	S/W
Big（大市值）	B/R	B/N	B/W

总资产变化率

市值	Aggresive（激进）	Neutral（中性）	Conservative（保守）
Small（小市值）	S/A	S/N	S/C
Big（大市值）	B/A	B/N	B/C

市值分组依据：NYSE 中位数；
ROE 和总资产变化率分组依据：NYSE 30% 和 70% 分位数。

出处：川总写量化

$$\mathrm{RMW} = \frac{1}{2}(\mathrm{S/R} + \mathrm{B/R}) - \frac{1}{2}(\mathrm{S/W} + \mathrm{B/W})$$

$$\mathrm{CMA} = \frac{1}{2}(\mathrm{S/C} + \mathrm{B/C}) - \frac{1}{2}(\mathrm{S/A} + \mathrm{B/A})$$

$$\begin{aligned}
\mathrm{SMB} &= \frac{1}{3}(\mathrm{SMB_{BM}} + \mathrm{SMB_{ROE}} + \mathrm{SMB_{INV}}) \\
\text{where } \mathrm{SMB_{BM}} &= \frac{1}{3}(\mathrm{S/H} + \mathrm{S/M} + \mathrm{S/L}) - \frac{1}{3}(\mathrm{B/H} + \mathrm{B/M} + \mathrm{B/L}) \\
\mathrm{SMB_{ROE}} &= \frac{1}{3}(\mathrm{S/R} + \mathrm{S/N} + \mathrm{S/W}) - \frac{1}{3}(\mathrm{B/R} + \mathrm{B/N} + \mathrm{B/W}) \\
\mathrm{SMB_{INV}} &= \frac{1}{3}(\mathrm{S/C} + \mathrm{S/N} + \mathrm{S/A}) - \frac{1}{3}(\mathrm{B/C} + \mathrm{B/N} + \mathrm{B/A})
\end{aligned}$$

从某种程度上说，Fama-French 五因子模型是他们向学界各种异象妥协的结果。随着诸多无法被三因子模型解释的异象相继被提出，他们意识到了在定价模型中加入新因子的必要性。五因子模型正是这个背景下的产物，

Hou-Xue-Zhang四因子模型

- 2015年，Hou, Xue, and Zhang从实体投资经济学理论出发提出了一个四因子模型：

$$E[R_i] - R_f = \beta_{i,MKT}\left(E[R_M] - R_f\right) + \beta_{i,ME}E[R_{ME}] + \beta_{i,I/A}E\left[R_{I/A}\right] + \beta_{i,ROE}E[R_{ROE}]$$

- $E[R_M]$：**市场组合**预期收益率；
- $E[R_{ME}]$：**规模因子**（ME）的预期收益率（市值）；
- $E\left[R_{I/A}\right]$：**投资因子**（I/A）的预期收益率（总资产变化率）；
- $E[R_{ROE}]$：**盈利因子**（ROE）的预期收益率（ROE）；
- $\beta_{i,MKT}, \beta_{i,ME}, \beta_{i,I/A}, \beta_{i,ROE}$：个股$i$在相应因子上的暴露。

因 子

- 在实证研究中，Hou, Xue, and Zhang 使用 ROE 和总资产变化率作为代表盈利和投资的指标；
- 在构建因子时，为了体现上述条件预期收益率的关系，他们使用市值、单季度 ROE 和总资产变化率进行 2 × 3 × 3 独立三重排序，其中市值按纽交所中位数划分、ROE 和总资产变化率按纽交所 30% 和 70% 分位数进行划分。

$$\begin{aligned}\text{ME} &= \tfrac{1}{9}(\text{S/L/L}+\text{S/M/L}+\text{S/H/L}+\text{S/L/M}+\text{S/M/M}+\text{S/H/M}\\ &\quad +\text{S/L/H}+\text{S/M/H}+\text{S/H/H})\\ &\quad -\tfrac{1}{9}(\text{B/L/L}+\text{B/M/L}+\text{B/H/L}+\text{B/L/M}+\text{B/M/M}+\text{B/H/M}\\ &\quad +\text{B/L/H}+\text{B/M/H}+\text{B/H/H})\end{aligned}$$

$$\begin{aligned}\text{ROE} &= \tfrac{1}{6}(\text{S/H/L}+\text{S/H/M}+\text{S/H/H}+\text{B/H/L}+\text{B/H/M}+\text{B/H/H})\\ &\quad -\tfrac{1}{6}(\text{S/L/L}+\text{S/L/M}+\text{S/L/H}+\text{B/L/L}+\text{B/L/M}+\text{B/L/H})\end{aligned}$$

$$\begin{aligned}\text{I/A} &= \tfrac{1}{6}(\text{S/L/L}+\text{S/M/L}+\text{S/H/L}+\text{B/L/L}+\text{B/M/L}+\text{B/H/L})\\ &\quad -\tfrac{1}{6}(\text{S/L/H}+\text{S/M/H}+\text{S/H/H}+\text{B/L/H}+\text{B/M/H}+\text{B/H/H})\end{aligned}$$

参考资料

1. 因子投资：发展与实践——石川
2. 华泰证券——《华泰多因子模型体系初探》
3. 多因子模型发展历史: https://www.douban.com/note/552529553/
4. 多因子发展知乎：https://zhuanlan.zhihu.com/p/335010915
5. 多因子概述知乎：https://zhuanlan.zhihu.com/p/197803793
6. CAPM模型：https://wiki.mbalib.com/wiki/资本资产定价模型
7. Fama-French三因子模型：https://zhuanlan.zhihu.com/p/341902943
8. APT模型：https://www.zhihu.com/question/37191834?sort=created
9. BetaPlus小组——资产配置理论：https://www.factorwar.com/research/asset-allocation/
10. Ledoit, O., & Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. Journal of multivariate analysis, 88(2), 365-411.
11. Kahn, R., & Grinold, R. (1999). Active Portfolio Management. New York, NY: McGraw-Hil1

北京大学量化交易协会