实习汇报

张南怡

2022 年 11 月 7 日

摘要

联海实习第一周，阅读至少九篇研报，初步掌握市面上常用的因子选股模型，形成自己的一套问题分析框架，然后呢，我也希望能够在实习的过程中，尽可能的多用数学理论去解释我们的行为，所以本讲义也会涉及一些数学推导。

联海实习的第二周，阅读波动率相关研报，

目录

插图

1 导论

1.1 定义 量化投资

利用数学，金融学，统计学，计算机科学的知识投入当前资金或者其他资源以期望在未来获得收益的行为。

本节写一些对于多因子模型的浅薄理解。多因子模型的理论基础来源于套利定理 (APT), 即认为我们的收益变量可以由一组公共因子的线性组合加上一个随机扰动项线性表出, 即

$$R - E(R) = B(F - E(F)) + \epsilon$$

1.2 定义 套利

一个投资组合 ω 称作套利，如果满足以下四条:
1. $\omega'\iota = 0$, 初始资金为 0，一个多空组合
2. $\omega' B = 0$，因子暴露为 0，屏蔽了因子的风险
3. $\omega' R \geq 0$, 策略不会亏钱
4. $P(\omega' R > 0) > 0$, 以正概率赚钱

而套利定理指出，$E(R)$ 属于矩阵 B 的列空间，也就是说单只股票期望收益可由因子暴露线性表出。从均值方差分析到 alpha 对冲策略，在均值方差分析中，我们认为未来的股票的期望收益是可以提前预知的，但事实上这在实际中并不可行，所以我们根据套利定理，我们需要找到一些能对预测未来收益率的因素（也称作 alpha 因子，注意区分因子暴露，因子和因子收益率）。此外在均值方差分析中，我们的目标函数是最大化二次效用，也就是

$$\max_{\omega} \omega' E(R) - A\omega' \Sigma \omega.$$

自然的，目标函数也应该发生相应的改变。我们认为目标函数由三部分组成，alpha 模型，风险模型，成本模型。alpha 模型的构建要求我们去寻找 alpha 有效的 alpha 因子，这涉及到了 alpha 因子的构建，分类，有效性检验问题，最后还得利用各种办法加权办法去合成一个 alpha 因子。第二部分是风险模型，主要目的是屏蔽一些 xxx 风险，最简单的形式就是二次效用函数的第二项。最后呢，成本模型，emm，不太熟悉交易流程，只能参考一下别人怎么写的了。

2 定义

有很多行业术语不太明白的，在这个部分做一个汇总。

2.1 定义

- 同向显著：本次显著 * 上次显著 =1
- 状态切换：本次显著 * 上次显著 =-1
- 指数增强：在市场指数本身走势的基础上，获取比市场指数收益更高的一些收益。问题是，胜率高于多少才叫指数增强呢?
- A 股的主要风格：行业和市值
- 信息系数：本期因子和下期收益率的 Pearson 相关系数 (n=300?)
- ST. ST*:
- 因子，因子收益，因子暴露：因子这一称法来源于多元统计的因子分析，我们认为股票收益率可以由几个“因子”的线性组合加上随机扰动项表示，因此把对股票收益率有显著影响的变量称作” 因子”，因子收益的定义稍复杂，后续再说，另外，值得注意的是，吴岚老师说，某些因子值可以直接作因子收益或者因子暴露，这个也得找个机会讨论讨论。
- 风险敞口：Risk exposure
- 信息比率 (IR): $IR = \frac{\alpha}{\sigma}$，还看到了一个 IR，定位为 IC 的均值除以 IC 的标准差
- 行业中性：回归中加入行业虚拟变量
- 风格中性：回归中加入对数市值，beta。这里的 beta 是什么?
- 生存偏误：股票池应该是 Dynamic
- 下月收益：1-月末/月初?
- 风险模型：用于控制策略风险，通常是一个 $n \times n$ 的矩阵，n 是股票池中的股票数
- alpha 模型：优化问题中的第一项，alpha 定义很多，后文也会详细讨论。
- alpha 模型的因子权重:
- 交易成本模型：优化问题中第三项，用于控制交易成本。
- 定价因子：通过 Fama-Macbeth 回归系数显著性检验的因子
- alpha 因子：IC 序列均值显著异于 0 且稳定性较强的因子 (IC_IR) 较高
- 风险因子:
- 跟踪误差：衡量基金经理主动投资的风险，后文有详细介绍

-
- Fama-Macbeth 回归：用于估计 lamdba，自己想想 lambda 是什么，在 APT 定理里。
- 成份股，全市场：通常成份股是指指数，例如中证 500，沪深 300 这些叫做成份股。全市场，很容易理解，差不多有 5000 只股票。
- 概念，板块
- 轮动
-

2.2 定义 跟踪误差

1. 跟踪偏离度 (Tracking Difference)：

$$TD_{i,t} = R_{i,t} - R_{m,t}$$

其中 R_i, t 代表基金在 t 时刻的收益率，R_m, t 代表市场组合的收益率。而所谓基金，也就是一个动态的资产组合。

2. 跟踪误差 (Tracking Error)：

$$TE_i = \sqrt{\frac{1}{n-1}\sum_{t=1}^{n}\left(TD_{ti} - \overline{TD_i}\right)^2}$$

跟踪误差越大，说明基金的净值率与基准组合收益率之间的差异越大，并且基金经理主动投资的风险越大。通常认为跟踪误差在 2% 以上意味着差异比较显著。

我们经常谈到 alpha，那什么是 alpha 呢，alpha 的定义很多，在不同场景下的定义不尽相同

2.3 定义 *Alpha*

2.4 定义 好的多因子模型

好的多因子模型，应该具备

1. 完备性：模型应该尽可能解释各个股票的收益率来源
2. 无关性：因子的相关性较低

3. 可解释性：因子应当具备统计学和金融学上的意义

Example 2.5:

$$r_{i,t} = \alpha_{i,t} + \beta_i MKT_t + s_i SMB_t + h_i HML_t + \varepsilon_{i,t}$$

MKT_t, SMB_t, HML_t 分别代表市场收益率，市值因子收益率，估值因子收益率

Remark:

- 因子分层 v.s 为什么不将该因子纳入回归呢?
- 为什么不考虑上市不满 6 个月的股票?
- 为什么我们要考虑因子间的 IC 系数，而不是同期的相关系数?
- 既然大部分量化策略是同质的，为什么不利用这一点呢?

3 量化指标

本节列示了一些评判策略好坏的指标及其计算方式，并给出这些指标的理想范围。

Beta:

$$\beta_p = \frac{cov(r_p, r_M)}{var(r_M)}.$$

beta 衡量了现有投资组合对于市场的反应程度，beta=1 代表市场中性策略，不是 beta 越大越好，只是一个相关性的度量. Sharpe ratio:

$$SP = \frac{E(r_p) - r_f}{\sigma_p}$$

SP 承担一单位风险所获得的收益，市场组合的 SP-ratio 称为风险的市场价格。

information ratio:

$$\frac{\alpha}{\sigma_\epsilon}$$

有点信噪比的意思

Max Drawdown:

$$MDD_t = \max_{i<j\leq t}(1 - \frac{P_j}{P_i})$$

turnover rate: 假设投资者投资于 n 个资产, 在 t 时刻进行调仓操作 (即买卖资产, 调整各个资产持有量), 调仓前持有资产总价值为 V_t, 其中各个资产价值分别为 $V_{t-}^1, V_{t-}^2, \ldots, V_{t-}^n$; 调仓后持有的各个资产价值分别变为 $V_{t+}^1, V_{t+}^2, \ldots, V_{t+}^n$ 。则 t 时刻的换手率被定义为

$$\text{换手率}_t = \frac{\sum_{i=1}^{n} \left| V_{t-}^i - V_{t+}^i \right|}{2V_t}.$$

4 数据处理

数据处理是我们进行计量分析的关键，在研报中的不同部分，数据处理也完全不同，什么时候应该标准化，什么时候应该归一化，在时间序列上做还是在横截面上做，什么时候应该剔除风格 (A 股市场中主要是行业和市值) 的影响，

4.1 去异常 (Winsorize)

三种去异常值的办法: 3sigma 原则，分位数原则 (保留百分之 95) 的数据，中位数去极值。大部分采用中位数去极值的办法对数据预处理（但我很奇怪，因子删去的话，对应的股票不也不在考虑范围内了吗），具体的流程是

1. 首先计算样本的中位数 (Fmedian)，
2. 然后计算离差的中位数（MAD)，
3. 删去中位数 3 倍 MAD 之外的数据。

4.2 去量纲 (Measurement Unit)

横截面上归一化: 因子的功效分析
标准化: 因子显著性检验，

4.3 中性化 (Neutralize)

A 股市场存在中存在明显的市值效应和行业效应，不少 alpha 因子的超额收益来源于此，所以需要剔除这些因素，检验因子能否获得行业，风格因素之外的 alpha。因子分层，纳入回归，

行业中性化：z-score 对行业虚拟变量回归，残差作为中性化后的因子值，采用申万一级行业，

风格中性化：z-score 对市值对数，beta 进行回归，残差作为中性化后的因子值，

财务因子采用行业中性化再风格中性化，对于技术类和风险类因子，采用风格中性处理，

4.4 其它

流通市值取对数后符合正态分布

5 因子分类

我采用吴岚老师的教材上的分类方式对因子进行分类，

5.1 按照因子来源

宏观经济因子，基本面因子，技术因子

5.2 按照风险来源

市场因子，行业因子，风格因子

5.3 基于财务指标

技术因子：反转因子，换手率因子，
反转因子: 价差偏离度
风险因子：市值因子，流动性风险因子，特质风险因子，特异度因子，

6 因子挖掘 (因子库构建)

6.1 一些被抛弃的因子

6.2 特异度

6.3 市值调整换手

6.4 因子挖掘交易热度

A 股市场投机氛围浓厚，规律是被投机的股票后期大概率跌，被

6.4.1 Basic Forms

6.4.2 Imporved

7 因子检验

我们通过统计学规律，经济学理论构建了因子后，首要的是检验因子的有效性，之后我们要判断新因子能否被之前的因子解释，从而需要相关性分析。

显著性检验，Pearson IC, Spearman IC 的正负显著度，同向显著，切换比例, 胜率，分组单调性，多空组合 (做多优势，做空劣势) 是否显著，多空组合的年化波动率，夏普比，最大回撤

相关性检验，通常套路是，我们构造了一个新的因子 A，他和已有的某些因子 B 具有强相关性，所以我们先以 B 分层，A 分组，检验有效性，发现 A 有 B 之外的超额收益，但如果以 A 分层，B 分组，发现 B 没有超额收益了，所以 A 这个因子是可行的。当然还有一种可能是，A,B 因子间有少量共有信息源，但仍相对独立。

对于因子间的两两间解释作用，因子分层，纳入回归，但考虑多因子间的话，可以用 Fama-Macbeth，1. First regress each of n asset returns against m proposed risk factors to determine each asset's beta exposures.

$$
\begin{aligned}
R_{1,t} &= \alpha_1 + \beta_{1,F_1} F_{1,t} + \beta_{1,F_2} F_{2,t} + \cdots + \beta_{1,F_m} F_{m,t} + \epsilon_{1,t} \\
R_{2,t} &= \alpha_2 + \beta_{2,F_1} F_{1,t} + \beta_{2,F_2} F_{2,t} + \cdots + \beta_{2,F_m} F_{m,t} + \epsilon_{2,t} \\
&\vdots \\
R_{n,t} &= \alpha_n + \beta_{n,F_1} F_{1,t} + \beta_{n,F_2} F_{2,t} + \cdots + \beta_{n,F_m} F_{m,t} + \epsilon_{n,t}
\end{aligned}
$$

2. Then regress all asset returns for each of T time periods against the previously estimated betas to determine the risk premium for

$$
\begin{aligned}
R_{i,1} &= \gamma_{1,0} + \gamma_{1,1} \hat{\beta}_{i,F_1} + \gamma_{1,2} \hat{\beta}_{i,F_2} + \cdots + \gamma_{1,m} \hat{\beta}_{i,F_m} + \epsilon_{i,1} \\
R_{i,2} &= \gamma_{2,0} + \gamma_{2,1} \hat{\beta}_{i,F_1} + \gamma_{2,2} \hat{\beta}_{i,F_2} + \cdots + \gamma_{2,m} \hat{\beta}_{i,F_m} + \epsilon_{i,2} \\
&\vdots \\
R_{i,T} &= \gamma_{T,0} + \gamma_{T,1} \hat{\beta}_{i,F_1} + \gamma_{T,2} \hat{\beta}_{i,F_2} + \cdots + \gamma_{T,m} \hat{\beta}_{i,F_m} + \epsilon_{i,T}
\end{aligned}
$$

Fama-Macbeth 回归的流程，t 上的一个样本，作回归，得到因子暴露的估计 (n*m)，然后固定 t，利用收益率对 n 个进行回归，得到 gamma。

因子稳定性如何衡量呢?

7.1 单因子检验

对于一个因子，我们可以通过下述办法检验其有效性，检验因子大小是否和未来股价有显著相关性。注意开始检验前，winsorize，standardize 一下。

7.1 定义 单因子检验-*IC*

可以使用 IC 和 Rank IC,

$$
\mathrm{IC}_t = \frac{\sum_{i=1}^{300} \left(x_{i,t} - \bar{x}_t\right)\left(R_{i,t+1} - \bar{R}_{t+1}\right)}{\sqrt{\sum_{i=1}^{300} \left(x_{i,t} - \bar{x}_t\right)^2 \sum_{i=1}^{300} \left(R_{i,t+1} - \bar{R}_{t+1}\right)^2}}
$$

Remark:
1. 月度回归 v.s. 整体回归，样本量较大时容易显著，同时月度回归可以研究因子风格持续或者反转的时间。
2. Pearson v.s. Spearman: 后者可以刻画非线性关系，但会损失边缘分布的信息，结论是具有同一个 copula 函数的联合分布的 Spearman rho 和 kendall tau 系数相同。

看到研报里在计算 IC 是选择了标准化，这有必要吗？算出来结果不一样？

Remark: 检验后的可以解读的信息
对于一个因子，我们共计有 T-1 个 IC 值，可以计算
1. 正负向显著比例：(选择比例较高的预测，如果持续性较强，则可以最近一次显著作为下次预测)
2. 同向显著比例/状态切换比例：同向显著比例大，风格延续性强，使用动态权重，反之，静态权重可能较好。
3. 正负显著比例至少一项大于 0.35, 或者加总大于 0.6，则可以初步认为因子显著
4. 因子的稳定性：同向显著比例大于 max(正，负)，则后续多因子模型构建中，动态的决定指标的参数可能胜率更高。

7.2 定义 单因子检验—分层检验

在月末 (调仓日)，按照因子大小排序，将样本分为 10 组，这样我们认为在各组之间，因子值就没什么差别，然后在组内等权重构造投资组合，比较各组的表现，比如，年化收益，超额收益，信息比，月胜率，最大回撤等。

7.3 定义 单因子检验—多空组合

做多优势组，做空劣势组构造投资组合，此时权重向量满足

$$\omega'\iota = 0.$$

吗?

多空组合也是一个投资组合，我们可以利用之前的一些指标去评判。

7.4 定义 单因子检验—行业中性组合

根据流通市值加权得到 31 个行业的权重向量 $R'\omega$，在不同行业间利用因子进行分组 (等分为 10 组)，各组之间等权重，把每个行业的第一组拿出来作为我们的优势组，依次类推。

图 1: IC 序列

Remark: 我对这个行业分层的理解就是呢，把因子对行业虚拟变量做了一个回归，然后取残差作为因子重新排序，效果应该差不多，试下推导。

7.2 单因子检验 II—剔除行业，风格后

本节比上一节多了一个流程，也就是 neutralize，为什么要 Neutralize，因为我国 A 股市场具有明显的行业和市值效应，这也是很多因子的 alpha 来源。我们想要关注因子除去市值和行业后，还能够有产生超额收益的信息。

7.5 定义 单因子检验—风险调整 *IC*

因子经过 winsorize，standardize，neutralize 后计算的 IC 值。

7.3 因子间的相关性结构

因子之间存在高相关性会对我们的模型产生灾难性的影响，所以我们需要剥离一个因子对另一个因子的影响，目前主要看到了分层 + 分组，将因子纳入回归，Fama-Macbeth 这些

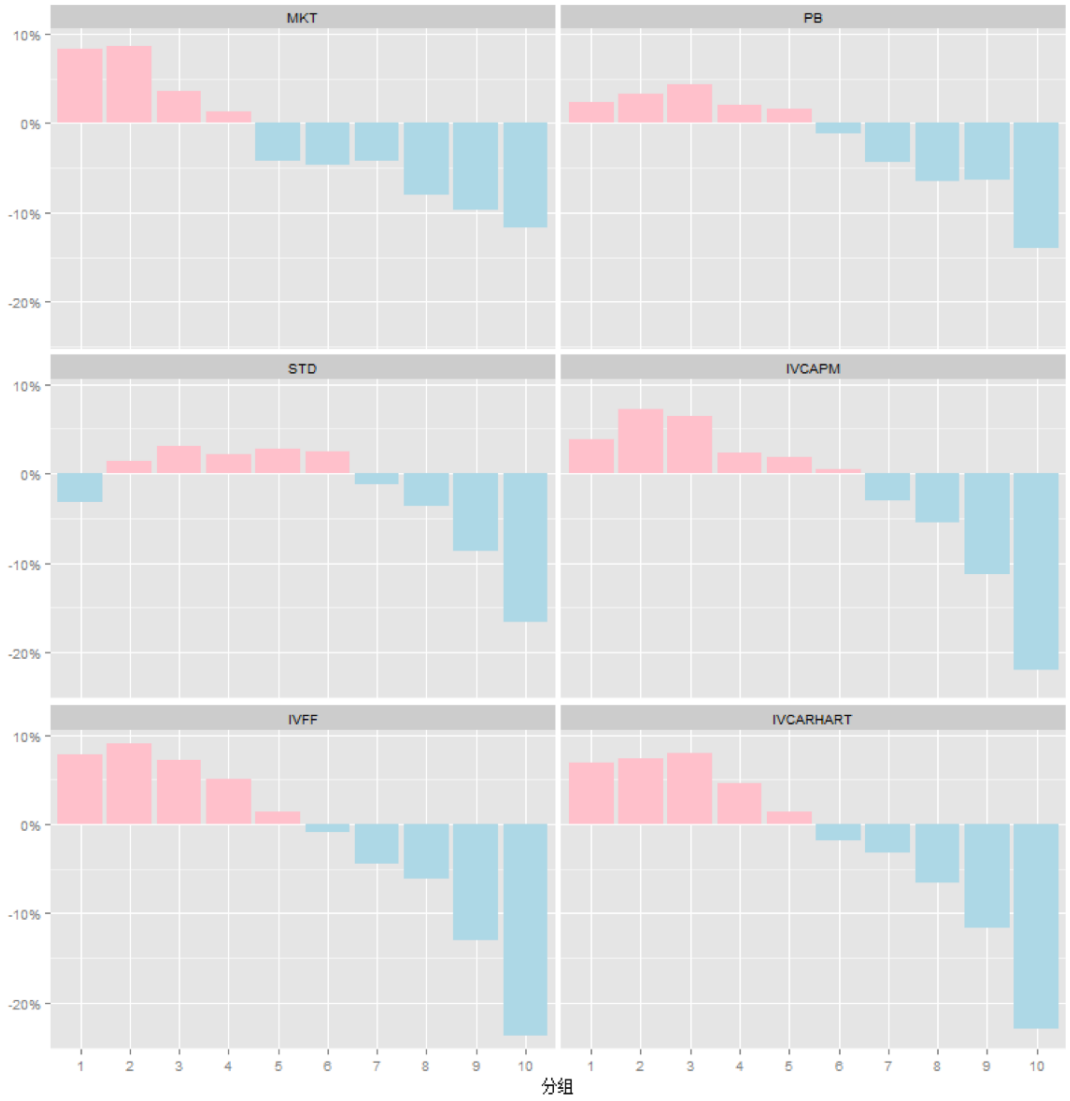

图 2: 比较理想的因子应当具有的效果 (当然也有几个不理想的)

办法，我们将逐一介绍。值得注意的是，我们也关注因子的收益风险特征，所以也需要考察因子收益间的相关性。

7.6 定义 因子相关性—分层 + 分组

可以用于剔除分层因子 F 对于分组因子对于收益率的影响

因子相关性分析可以让我们对超额收益归因，更好的构建多因子模型。

Remark: 假如我们有两个因子 F_1, F_2, F_1 分层，F_2 分组后，构建多空组合，如果收益率变得不显著，那么 F_2 带来的超额收益主要源于 F_1，就这么个逻辑，如果还是显著的，那么就可以认为 F_2 有除 F_1 之外的超额收益来源。

7.7 定义 因子相关性—因子收益率回归

这里举一个例子吧，为什么这个部分和上个部分不同呢，做回归我们可以构造新的因子，而分层呢，不行（但可以加入固定效应？在换手率和市值的相关性分析中，我们发现经过市值分层后，低换手率股票仍然带有显著正向收益，因此将对数换手率，对对数流通市值回归后，残差项即可作为剔除市值影响后的换手率代理变量）

另外呢，为什么要用因子收益率呢？这里就涉及到了多因子模型，做回归的目的在于，检验因子暴露是否显著以及检验 alpha 是否显著

Remark: 这个回归的因变量应该是股票收益率了，想想应该是面板，截面，时序中，哪个层面上的回归。

Remark: 然后还得注意，如果我们需要剥离因子的影响，这时候是用的因子值对因子值的回归。

7.4 因子的 IC 的相关系数

可以用于提高多因子模型的稳定程度

7.5 疑问

作 IC 的时候，是用月末的因子和下个月的月均收益率呢，还是月末当天的收益率。答：一般是月末的因子和次月第一天的因子做相关系数，但也可以增加间隔的时间。

8 alpha 对冲

为什么要叫 alpha 对冲呢？
在开始本章前，再次强调多因子模型的逻辑，首先是根据一套具有经济含义的投资逻辑，比如说购买具有相对低估值低高成长公司，而且这些公司没有被过度炒作。然后根据流程去挑选合适的因子合成为一个 alpha，再通过一些约束，完成我们的优化问题。这里先给出优化的目标函数，

$$\max_{\omega} \ \alpha'\omega - A\omega'\Sigma\omega - \lambda\tau'|\omega - \omega_0|$$

这样咱们就可以知道如何着手建模了。

8.1 目标函数构建

alpha 模型，风险模型，成本模型，

8.1.1 alpha 模型

在多因子择股量化投资流程中，alpha 模型直接决定了投资组合能否获得稳健的超额收益。我认为构建 alpha 模型的要素不是从统计学意义上挑选因子，反而应当先构建一套符合经济学直觉的投资理论，比如我们想要挑选不那么被过度投机的股票，该过程称为投资逻辑构建。投资思路的确定缩小了因子的范围，然后我们再将这些因子筛选出来进行前文所述的检验，比如 RANK IC，RANK IC 的 t 统计量，IR，IC 正负显著比例，多空组合的各项评判指标（收益率，波动率，夏普比，最大回撤等），条件允许的话还可以进行相关性分析等，最终挑选出少于 10 个优质因子值，我们称之为alpha 因子。关于叫法来源，我觉得可能是这些因子很可能带来不错的 alpha 收益。

Remark: 吴老师说

有较为稳定的 IC 的因子叫做 alpha 因子，他们可以用于预测期望收益，因子我们常用 alpha 因子估计期望收益向量 μ.

如果只有一个因子，那么经过处理后直接作为 μ, 如果有多个因子，那么可以用下述整合办法去得到一个新的 μ。

Remark: 吴老师说

需要值得注意的是，上述办法估计 μ，那么 mu 的含义已经不再是股票的期望收益，更多可以理解为借助 alpha 因子得到对这些股票的心理预期。

这就解释了为什么目标函数第一项是 alpha。接下来是利用这些 alpha 因子去构建选股策略，首先先把因子处理一下，winsorize，standardize，neutralize 就行了，然后我们想办法把他们加总，这一步叫做因子合成。我们可以采用下述办法进行加权，

8.1 定义 等权重合成

现在不同类别的因子上分配相同的权重，然后在同类因子的内部再分配等权重。

8.2 定义 历史滚动 IR 加权

当各因子之间相关性较低时，我们可以使用这个办法，具体而言，先假设

$$cv_t = \left[cv(IC_{1,t}), \cdots, cv(IC_{k,t}).\right]$$

计算 IC 的 cv 时可以选用不同的滞后期数，最后我们因子的权重为

$$\omega_F = \frac{cv_t}{cv_t' \iota}$$

啊，上边的有一些 typo，应该是变异的倒数。

8.3 定义 历史面板数据进行回归

被解释变量为股票的月度收益率，自变量是处理后的因子值，得到回归方程。利用最新一期的因子值更新数据，预测值作为”alpha”。

当然，在研报八—动态情景的多因子 Alpha 模型中会介绍更复杂的加权办法，当然我们这里得到的叫因子综合得分，究竟是不是这里的 alpha，还有待而论。

8.4 定义 动态情景 + 历史滚动 IR

研报八里边那个，相对于一般的历史滚动，加入了情景因子。但我有点没看明白他怎么给情景因子赋的权重。

Remark: 当然，如果对于既定目的下，经济学含义特别明显的因子，我们还有其它的加权方法，比如加总 Copula，这个说法不太严谨，但我自己能明白是啥意思，实在明白不了？那想想交易热度怎么构建的吧。

8.1.2 风险模型

我目前对于风险模型的理解是，目标函数中的第二项，控制投资组合的风险。而风险有多种度量方式，比如协方差矩阵，当然，如果直接用协方差矩阵的话，一个是咱们的这个 T 必须大于股票数量 N，第二呢，可能存在一些噪声，所以我们也有一系列改进方案，当然，还是按照流程来

8.5 定义 风险模型-协方差矩阵

我们利用协方差矩阵刻画整体风险，具体计算办法为:

$$S = \frac{1}{T}X'(I - \frac{1}{n}\iota\iota')X.$$

Remark: 有一些问题，比如我们采用月度调仓，股票池中有 2000 只股票，此时需要时序上大于 2000 次观测，这可行吗？所以改进一下

8.6 定义 风险模型-因子模型

$$S_t = B_t F_t B_t' + E$$

其中 B 为因子暴露矩阵

利用 BFB' 对风险进行衡量。

Remark: 因子模型有，基本面因子模型，宏观因子模型，统计因子模型 (把主成分作为因子)，这里可以来一个详细推导。

8.7 定义 风险模型-协方差矩阵压缩估计量

$$\Sigma_{shrink} = (1-\beta)F + \beta S$$

Remark: 会不会有 Bayes 信度模型。

8.1.3 成本模型

目前先不看，目前用到的主要是常数交易费用。

8.2 优化问题

8.2.1 Basic Form

假设我们通过因子检验，合成，已经通过加权，构造了因子向量 alpha(在一个时间截面上，一只股票对应一个 alpha)，假定股票池有 n 只股票，我们的目标是在控制风险和成本的情况下最大化 alpha（为什么不是最大化收益，我猜测是收益率是随机游走，基本不存在自相关性，现在收益高不代表未来收益高，所以最大化收益向量不如最大化我们这里的 alpha，至少这里的 alpha 都是比较能预测未来收益率的）

$$\max_{\omega} \ \alpha'\omega - A\omega'\Sigma\omega - \lambda\tau'|\omega - \omega_0|$$

s.t.

$$\begin{aligned}
&R'\omega = R'\omega_{\text{bench}} \\
&\omega'\iota = 1 \\
&0 \le \omega_i \le \min(maxposition, \omega_{0,i} + maxtradesize_i/booksize_i)
\end{aligned}$$

值得注意的是，α 应该是合成的 alpha 因子，A 代表风险厌恶系数，λ 用于控制交易成本。关于目标函数的解读，首先是要最大化 alpha 因子，也就是目标函数的第一项，第二项用于控制投资组合的风险，最后一项绝对值是逐项运算，用于控制成本。

关于约束条件的解读，R 是一个 31×300 的矩阵，每一列之和为 1，用于存储股票的行业信息。因此我们的第一个条件含义在于，投资组合中资金分配在一个行业的比例，应该等于行业

内动态市值加权的占比。第二，第三个条件比较容易理解。一个变种是加入显性跟踪误差约束条件

$$\max_{w} \quad \alpha' w - \lambda \tau' |w - w_0|$$

s.t.

$$\begin{aligned}
&\mathrm{R}' w = R' w_{\text{bench}} \\
&(w - w_{\text{bench}})' \Sigma (w - w_{\text{bench}}) \leq \frac{TE^2}{252} \\
&\omega' \iota = 1 \\
&0 \leq w_i \leq \min(\text{maxposition}, w_{0,i} + \text{maxtradesize}_i / \text{booksize}_t)
\end{aligned}$$

Remark: 对于新加的这个约束，我的理解是相当于控制一个多空组合的风险，因为 $(\omega - \omega_{bench})' \iota = 0$

规划求解问题：

8.3 归因分析

9 调仓频率

调仓频率和因子的信息衰减速度需要做一个匹配，以充分利用因子的信息。

10 一些有待解决的问题

10.1 行业中性化的相关问题

我们在优化问题中提到了，

$$R' \omega = R' \omega_{bench}$$

这个条件和行业中性化有些像。所以再仔细想想呢？

11 因子挖掘

12 因子挖掘—交易行为类

在构造一个因子前，我们需要去说明一套能赚钱的投资理论，然后结合理论去寻找量化指标。

12.1 投资逻辑

投资逻辑，A 股市场投机氛围浓厚，规律是被投机的股票后期大概率跌，而相对正常的股票，反而会涨，所以我们的逻辑是卖投机程度高的，买投机程度不足的。逻辑虽然简单，但投机程度无法观测，我们需要去构造能够衡量投机程度的指标。

12.2 指标构建

过度投机的股票会有，高波动性，风格独立，价格时滞，高换手的特点，所以我们分别采用下述四个指标对其进行衡量

特质波动率

我们用特质波动率来衡量个股的波动程度，特质波动率实质是条件方差，所以我们用条件方差的均值进行刻画。

$$\text{特质波动率}^2 = \frac{SSR_{Fama}}{n}$$

被投机的股票多空分歧大，进而展现出大的波动。

特异度

被过度投机的股票会过分利用自身信息，忽视市场规律，从而我们的 $Fama - French$ 模型对股价的解释力度较差，这说明 R^2 较低. 因此我们定义

$$IVR = \frac{SSR}{SST} = \frac{\text{特质方差}}{\text{总方差}}$$

价格时滞

过度投机的股票不能及时反应价格信息，但这并不意味着过去的信息就有用，所以我觉得讨论这个指标纯粹是浪费时间。

高换手

过度投机的股票换手率较高，这个很容易理解，但市值和换手率相关度较高，所以我们先取对数 (让数据服从正态)，然后将对数换手率关于对数市值作一个正交分解，将残差作为市值调整的收益率

12.3 交易热度 I

经过相关性分析，我们发现特异度和市值调整换手能够包含其他交易行为类指标的所有信息，我们利用加总分位数的方式合成了一个新的因子，叫做交易热度

$$\text{BehaviorIndex}_{i,t} = \frac{1}{2}\left[Q\left(\text{IVR}_{i,t}\right) + Q\left(adjTurnover_{i,t}\right)\right]$$

其中，BehaviorIndex $x_{i,t}$ 为股票 i 在时刻 t 的交易热度，IVR $R_{i,t}$ 为股票 i 在时刻 t 的特异度，adjTurnover $_{i,t}$ 为股票 i 在时刻 t 的市值调整换手。$Q\left(I_{i,t}\right)$ 表示股票 i 的指标 $I_{i,t}$ 在时刻 t 样本空间内所有股票中所对应的分位数 (累计分布概率)。交易热度在 0，1 中取值，而且我们取分位数的操作 (其实也是嵌入广义逆) 是一个单调变换，不会改变秩相关系数，因

此可以预计到，他基本和对数流通市值不相关。

Remark: 这个 Q(X)，代表 X 的分位数，其实差不多就是 F(X), 其中 F 是经验累积分布函数，所以差不多 Q(X) 就可以看作一个均匀随机变量 (的观测)。

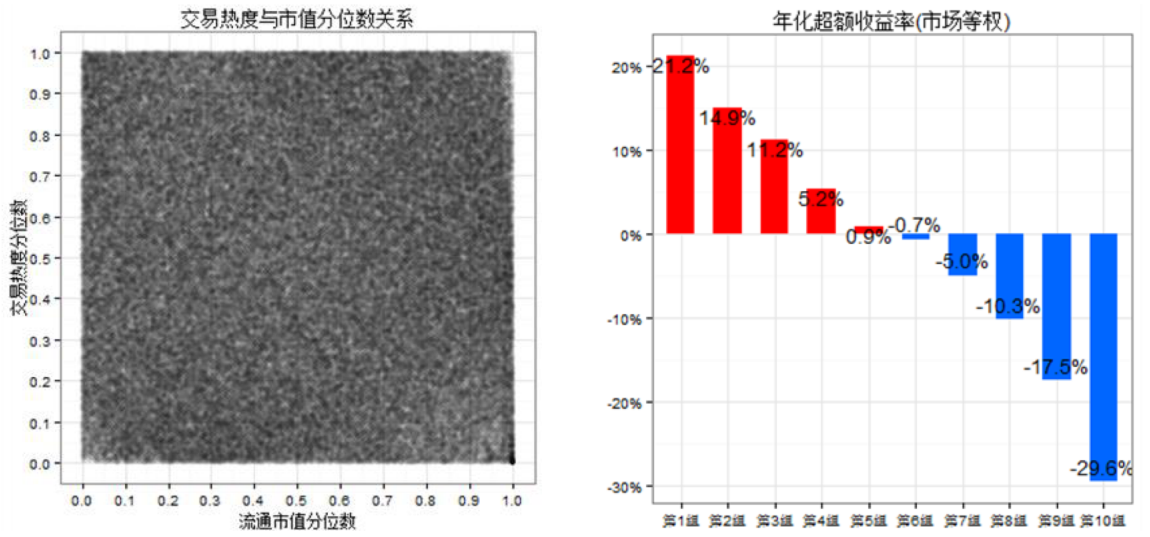

图 3: 交易热度和流通市值的关系以及分组表现 (看起来像独立 copula)

12.4 交易热度 II

我们在下一节定义了一个价差偏移度，不同于特异度和市值调整换手，价差偏移度直接从股票价格与其相似股票的偏离程度出发，以衡量股票的投机程度。由于三个指标有明显的超额收益，而且信息来源相对独立，所以综合起来更新交易热度，

$$\text{BehaviorIndex}_{i,t} = \frac{1}{3}\left[Q\left(\text{IVR}_{i,t}\right) + Q\left(adjTurnover_{i,t}\right) + Q(SpreadBias_{i,t})\right]$$

可以发现交易热度市值是 3 个均匀随机变量之和，交易热度越高，股票相对高估，后期预期收益率较低。

13 因子挖掘—价差偏移度

我们利用了特异度和市值调整换手这两个交易行为类指标去衡量个股被投机的程度。但他们直接没有从股票的价格考虑股票被相对高估低的情况。

13.1 投资逻辑

股票的涨跌以概念，板块的方式轮动炒作。事实上，某一板块大部分股票都涨时，其它没涨的股票就有补涨的需求，反之也成立。因此，当类别 A 的某只股票表现弱于其相似的股

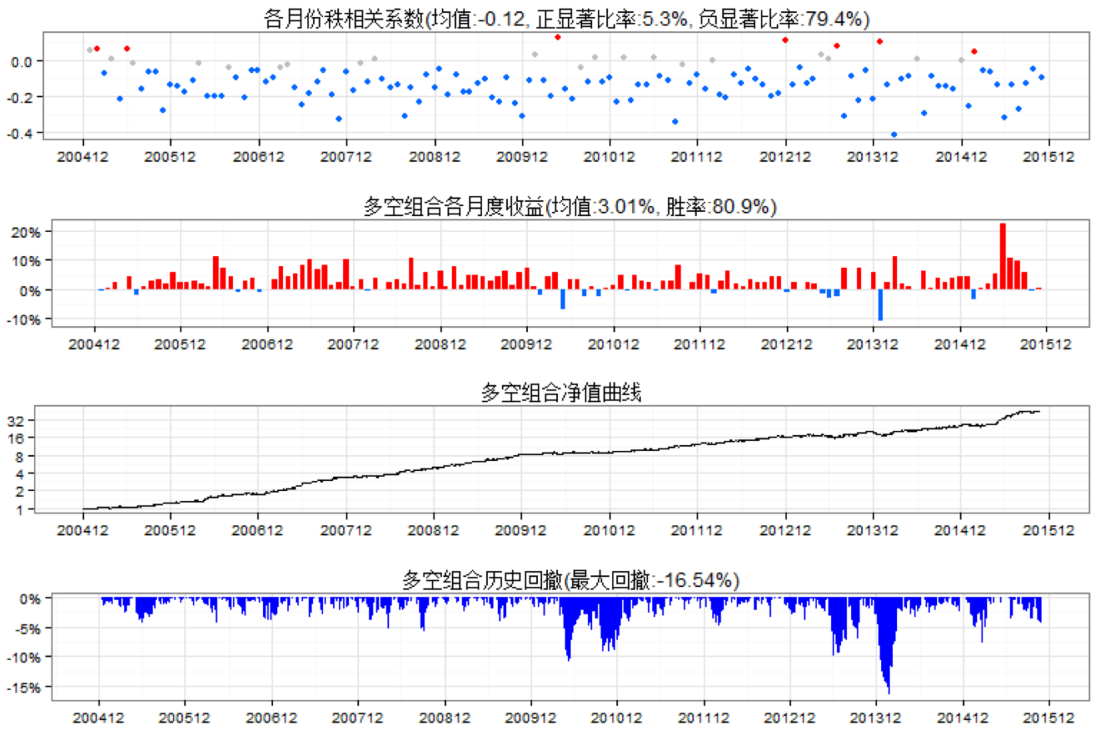

图 4: 交易热度历史表现

票时，买入则可能获得超额收益。

13.2 指标构建

借助统计讨论的思想，我们提出价差偏移度，试图捕捉股票相对同类型股票的相对高低估程度。

13.1 定义 股票的距离

$d(S_i, S_j) = 1 - corr(S_i$的过去 250 交易的日涨跌幅度, S_j的过去 250 交易的日涨跌幅度$)$

当然，上边的这个距离不满足三角不等式 (满足吗？)，从而并不满足距离的公理化定义，但对于我们的问题，这个距离已经够用了。

13.2 价差偏移度的构建流程

1. 固定股票 i，提取 i 距离最近 (d(i,j) 最小的 10 个 j) 的 10 只股票，等权重构成参考价格 (ref)
2. 定义对数价差

$$\text{PriceSpread}_{i,t} = \ln S_{i,t} - \ln ref_{i,t}$$

	年化收益率	年化超额收益	夏普比	信息比	月胜率	最大回撤	月均换手
中证全指	17.5%	-13.5%	0.68	-1.33	36.6%	-71.5%	
市场等权（基准）	31.1%	0.0%	0.98		0.0%	-70.5%	5.8%
第1组	58.8%	27.7%	1.56	2.98	78.0%	-63.8%	76.6%
第2组	46.7%	15.7%	1.27	2.24	69.9%	-66.0%	85.5%
第3组	42.1%	11.0%	1.18	1.89	77.2%	-70.2%	87.7%
第4组	37.6%	6.5%	1.09	1.30	65.0%	-68.6%	89.0%
第5组	35.0%	4.0%	1.04	0.91	56.1%	-72.2%	89.8%
第6组	29.8%	-1.3%	0.92	-0.14	44.7%	-70.2%	89.7%
第7组	24.4%	-6.6%	0.80	-1.19	39.8%	-72.3%	89.0%
第8组	19.7%	-11.3%	0.69	-1.87	30.9%	-72.6%	88.0%
第9组	12.7%	-18.4%	0.51	-2.59	22.8%	-74.4%	85.8%
第10组	-0.7%	-31.7%	0.15	-3.15	10.6%	-83.5%	78.3%

图 5: 交易热度 3 因素，历史表现

3. 计算过去 60 日对数价差的 Z-score(时间序列上)，作为价差偏移度.

Remark: 价差偏移度是一个相对意义上的反转因子，价差偏移度低，股票跑输组合，估值较低，未来会涨，买入将获得超额收益，反之成立。值得注意的是，上述成立的一个必要条件是，股票基本面没有发生重大变换。

13.3 因子检验

13.3.1 分组

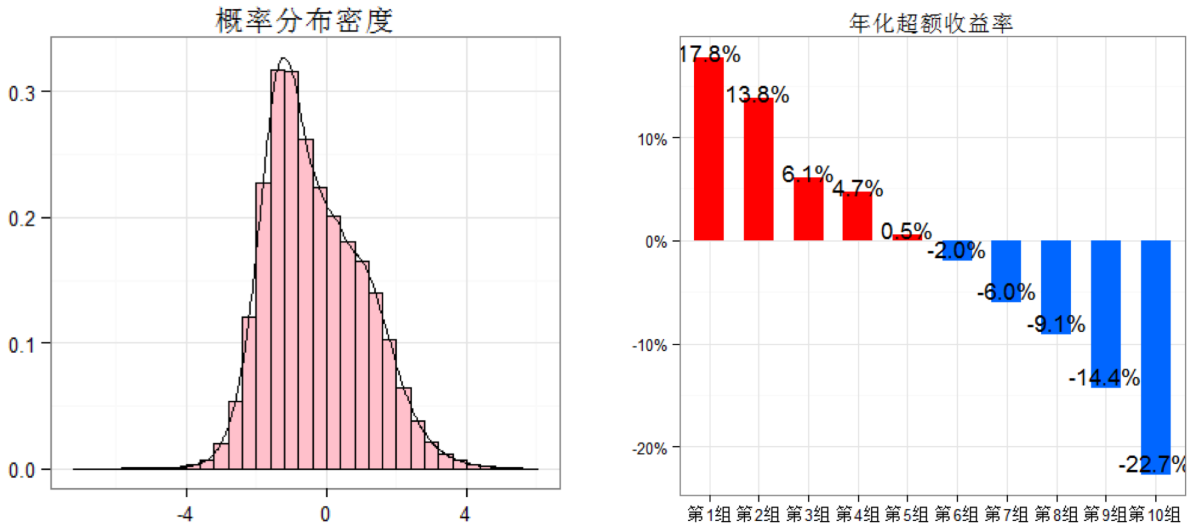

图 6: 价差偏移度的分组收益

满足单调性，然后分布呢，不知道怎么形容。看看历史 IC 和多空组合的历史表现吧，

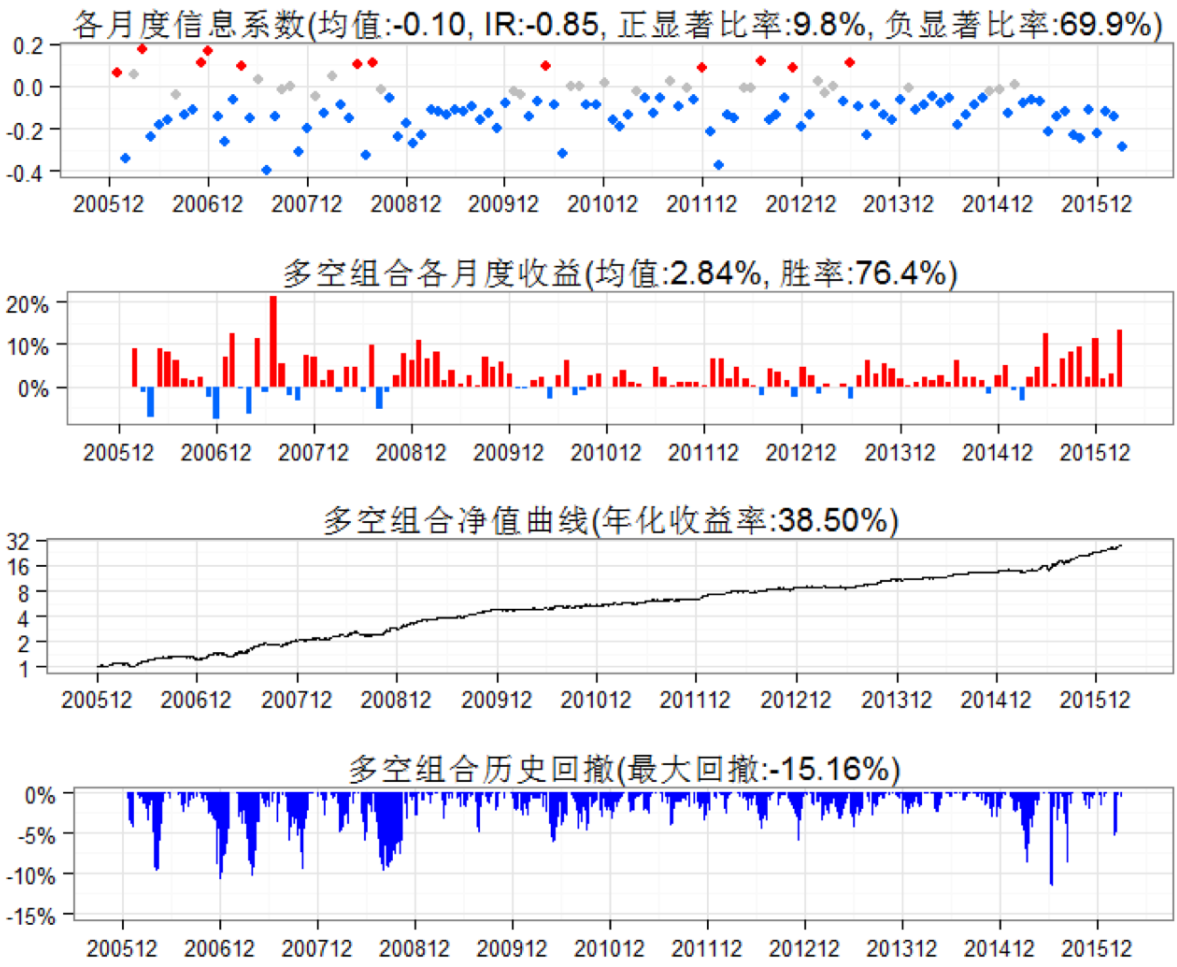

图 7: 价差偏移度的历史表现

13.3.2 相关性分析

价差偏移度是一个相对的反转因子，所以和 1 个月反转，3 个月反转有相关性，然后我们这里利用，分层 + 分组构建多空组合，Fama-Machbeth 对收益来源进行归因。

14 因子挖掘—日内残差高阶矩

15 研报阅读一

[?]

15.1 单因子有效性检验

量化选股主要涉及到两个核心的问题。第一，如何选出有逻辑意义并且能够有效的区分个股的因子，使得因子值对于个股未来收益有一定的预测能力。第二，影响市场的因子众多，

且市场风格并非一成不变，我们如何构建一个能够适应市场变化的多因子模型，筛选出大概率能够战胜市场的股票组合

15.1

-
- 市值因子，反转因子，换手率指标，各类业绩增速指标和估值指标表现较好
- 资产负债率，周转天数，ROE，ROA，销售利润对股价有一定的预测性

15.2 单因子检验

我们的目标在于检验因子与未来收益率是否存在显著相关性。主要流程是，按照每一期指标值大小对股票分组，这样各组的指标可以看作一致，然后比较各组的表现，比如累计收益，信息比率，最大回撤单调性越强，优势组的胜率越高。

15.2.1 OLS v.s. Robust

Robust 能够容忍异常值

15.2.2 整体回归 v.s. 按月度回归

按月度回归的好处：减少样本量（样本量过大会导致相关系数显著），有利于观察指标的历史变化情况，比如正负月份的具体比例，因子的持续周期和反转频率

15.2.3 相关系数的度量

Pearson，Spearman rho，kendall tau 对比，后面俩可以度量非线性关系，但随时了边缘分布的信息，经验法则是 Spearman rho 的显著比例高于 Pearson rho。

15.2.4 是否应当行业中性化

基本面因子，如PB，资产周转率，市值，存在行业固定效应，所以全市场分组不合理。，行业中性化相当于，我们在每个行业按照 PB 进行分组，在根据流通市值加权，这样就可以认为每组的不存在行业固定效应了。

15.2.5 同向显著比例，状态切换比例

用当月因子值与次月收益率算出来的回归系数，可能随时时间发生改变，实际预测的时候怎么办呢？
1. 选择正负比例中较高的那个，2. 如果因子的特性延续性较好，则可以最近一期的数据情况对下一期进行预测。

15.3 结果展示及业界通用做法

15.2

- 衡量因子是否显著：秩相关系数正/负比例至少一项大于 0.35, 或者之和大于 0.6
- 是否行业中性化：行业内分组的优胜组年化超额收益大于全市场分组
- 因子稳定度的衡量：1. 同向显著次数大于状态切换次数，说明因子有一定趋势 2. 当同向显著的比例大于 max（正显著相关，负显著相关），则在后续的多因子模型建立中，动态的决定指标的参数可能胜率更

一些疑问是，本月因子值与下月收益率做回归，但问题是数据的长度可能不匹配，而且对应规则如何呢？是当前月第一天对应次月第一天吗？还有就是所有因子每天都有数据吗？会不会有某些因为长时间保持不变呢？或者说，我们做的是截面上的回归，那这和样本量有什么关系呢？

是否只有单调性好的指标才是好的因子？并不是，某些因子的特性，或者要达到一定阈值后因子的效果才会体现出来。

16 研报阅读二—特质波动率

16.1 现象及结论

低特质波动率，未来预期收益更高。首先是特质波动率，严格定义一下，在因子模型中，我们认为个股 i 的超额收益率为

16.1 Theorem *方差分解*

$$R_i = \alpha + \langle B, F \rangle + \epsilon_i$$

$$Var(R_i) = Var(E(R_i|F)) + E(Var(R_i|F))$$

我们认为 $Var(R_i|F)$ 就是 R_i 的特异风险的平方。

16.2

- FF 三因子模型残差的波动率能更好捕捉股票的特质风险，记作 IVFF.（好体现在哪里）
- 特质风险与未来截面收益率存在负相关关系，IVFF 预测能力最强
-

17 研报阅读三—交易行为与股票收益

17.1

- 投机程度弱的股票未来大概率涨，所以应当买入
- 通过交易行为指标度量投机程度，具体有:
  特征波动率，特异度，价格时滞，市值调整换手
- 这些交易行为类指标均表现出较强的收益预测能力
- 特异度和市值调整换手在控制其他交易行为类指标后还能带来超额收益。

17.1 交易行为类指标

过度投机的股票会有，高波动性，风格独立，价格时滞，高换手的特点。

17.1.1 特质波动率

我们用特质波动率来衡量个股的波动程度，特质波动率实质是条件方差，所以我们用条件方差的均值进行刻画。

$$\text{特质波动率}^2 = \frac{SSR_{Fama}}{n}$$

被投机的股票多空分歧大，进而展现出大的波动。

17.1.2 特异度

被过度投机的股票会过分利用自身信息，忽视市场规律，从而我们的 $Fama-French$ 模型对股价的解释力度较差，这说明 R^2 较低. 因此我们定义

$$IVR = \frac{SSR}{SST} = \frac{\text{特质方差}}{\text{总方差}}$$

17.1.3 价格时滞

过度投机的股票不能及时反应价格信息，但这并不意味着过去的信息就有用，所以我觉得讨论这个指标纯粹是浪费时间。

17.1.4 高换手

过度投机的股票换手率较高，这个很容易理解，但市值和换手率相关度较高，所以我们先取对数 (让数据服从正态)，然后将对数换手率关于对数市值作一个正交分解，将残差作为市值调整的收益率

18 研报阅读八—动态情景多因子 Alpha 模型

18.1 Alpha 模型的构建

什么是 alpha 模型? alpha 模型的最终目标在于给股票未来收益率排序，核心是因子的挑选和不同因子间的权重配比。动态情景 alpha 模型，Dynamic Contextual Alpha(DCA)，

19 研报阅读九—日内残差高阶矩与股票收益

高频数据，分笔数据 (tick)，快照数据（quote），分钟数据，资金流量数据等。日内价格行为特征和股票未来收益率之间的关系

20 研报—波动率

21 123