知乎 首发于 川流不息

出色不如走运 (III)?

石川
量化交易 话题的优秀回答者

已关注

32 人赞同了该文章

摘要

本文使用随机因子的实证结果定量说明了仅靠运气就能够达到的选股效果，帮助判断选股因子是否真正有效。

1 引言

使用因子选股的逻辑是因子 —— 无论是来自基本面、量价还是宏观经济等 —— 都对股票未来的收益率有预测性。在定量评价一个因子是否有效时，主要的考察方式之一是计算该因子的收益率是否显著不为零（原假设）。

假设因子的预期收益率和该预期收益率的 standard error 分别为 $\mathbf{E}[f]$ 和 $\mathbf{s.e.}(\mathbf{E}[f])$ ，则假设检验的 t-statistic 为：

$$\mathbf{t\text{-}statistic} = \frac{\mathbf{E}[f]}{\mathbf{s.e.}(\mathbf{E}[f])}$$

multiple testing）并从里面挑出来最好的，由于 data mining 的问题（即运气），即便最好因子的 t-statistic 大于 2，也不能认为它是有效的。之前的两篇文章《出色不如走运?》以及《出色不如走运(II)?》对这个问题进行了探讨。

今天这篇是《出色不如走运(III)?》。

假设同时考察 n 个因子、这些因子对于股票收益率的预测能力满足 Uniform distribution。如果从这 n 个因子中挑出效果最好的，这个“最好的”因子的 t-statistic 和 p-value 有哪些性质呢？我们想要回答的问题是：**在多重检验的 n 选 1 问题中，对于给定的显著性水平 p （比如 5%），单一因子的 p-value 或 t-statistic 应满足什么条件才能拒绝原假设。**

根据 order statistic 的概率知识可知，这 n 个因子中第 i 好的满足 Beta distribution：

$$U_{(i)} \sim \mathrm{Beta}(i, n+1-i)$$

从 n 个里面挑出最好的相当于令 i = n。根据 Beta distribution 的定义和简单计算有：

$$\mathrm{prob}(U_{(n)} < x) = x^n$$

令 $x = (1-p)^{1/n}$ 并利用 $\mathrm{prob}(U < x) = 1 - \mathrm{prob}(U \geq x)$ 可知：

$$\mathrm{prob}\left(U_{(n)} \geq (1-p)^{1/n}\right) = p$$

在因子分析中，通常关注的是因子收益率是否显著不为零 —— 可正可负 —— 因此一般使用双边检验。对于给定的 p-value（单边 $p/2$ ），由上式可知（将 p 换成 $p/2$ ），这 n 个因子中最好的那个的 t-statistic 的绝对值需不小于以下阈值才能拒绝原假设：

$$\text{t-statistic}^{\star} = \mathcal{N}^{-1}\left(\left(1-\frac{p}{2}\right)^{1/n}\right)$$

当 n 很大时，从上式可进一步推导出单一因子的 p-value 需要小于 p/n 才能在 n 选 1 的 multiple testing 下拒绝原假设。举例来说，我们考察 10 个因子并希望以 5% 的显著性水平找到真正有效的因子，则这些因子各自的 p-value 只有小于 5%/10 = 0.5% 才能拒绝原假设。这正是大名鼎鼎的 **Bonferroni correction（邦费罗尼校正）**。

实际因子选股面临更复杂的问题：如何从 n 个因子中选出最好的 k 个，而非 1 个；如何配置选出来的这 k 个因子 —— 等权配置还是按照它们**样本内**的表现好坏配置。如果不妥善解决 multiple testing 的问题，上述这些做法会导致**选择偏差（selection bias）**以及**过拟合偏差（overfitting bias）**。

在选择因子时，通常的做法是在回测中使用因子定期构建投资组合，然后分析因子预期收益率的 t-statistic。如果该 t 值小于零（且显著为负）则把该因子反过来使用。假设同时考察 n 个因子，并根据因子 t-statistic 绝对值的大小采用下列做法之一：

1. 按照样本内 n 个因子 t-statistics 的正负同时使用全部因子，按照等权或者样本内因子效果赋权来选股（n 选 n 问题） —— 这种做法引入 overfitting bias；

2. 从这 n 个因子中挑出样本内 t-statistic 绝对值最大的 1 个（n 选 1 问题），使用该因子选股 —— 这种做法引入 selection bias；

3. 从这 n 个因子中挑出样本内 t-statistic 绝对值最大的 k 个（n 选 k 问题），并按等权或样本内效果赋权选股 —— 这种做法同时引入 selection bias 和 overfitting bias。

Novy-Marx (2015) 研究了多因子选股回测中的 selection bias 和 overfitting bias 问题。本文第一节中的数学推导正是来自 Novy-Marx (2015)，而它仅仅是 n 选 1 的一种简化情况。在投资实务中，更常见的是上述第二种 n 选 k 的问题，它面临**“因子怎么选”**和**“因子如何配”**这两个严峻的问题，一不小心就会引入大量的噪音。

毫无疑问，multiple testing 下的数据挖掘是因子选股的大敌。通过 data mining，仅仅依靠运气，挑出来的因子 —— 哪怕再没有业务含义 —— 也会在样本内获得显著不为零的选股收益率，但它只是过拟合而已。

针对 multiple testing 中的 data mining，Novy-Marx (2015) 从理论和实证经验两个角度分析了上述 n 选 n、n 选 1、n 选 k 三个问题中，多因子策略的 t-statistic 的分布问题。该文使用纯随机产生的因子 —— 理论上没有任何预测性 —— 在美股上选股、以美股的真实收益率计算这些随机因子的因子收益率，从而定量分析上述三个问题中多因子选股策略的 t-statistic 的阈值。

这些 t-statistic 的阈值远超单因子检验中的 2.0，而如此之高的阈值更是仅仅来自于运气和 data mining。在实际选股中，使用多因子构建的策略的 t-statistic 唯有超过这些阈值才意味着它们真的对收益率有统计上非显著为零的预测性。

本文下一节借鉴 Novy-Marx (2015) 的思路产生随机因子，使用中证 500 指数的成分股进行实证分析。

3 来自中证 500 的实证

实证中的回测期从 2010 年 1 月到 2019 年 1 月，考察 n 个随机因子的选股能力。具体的：

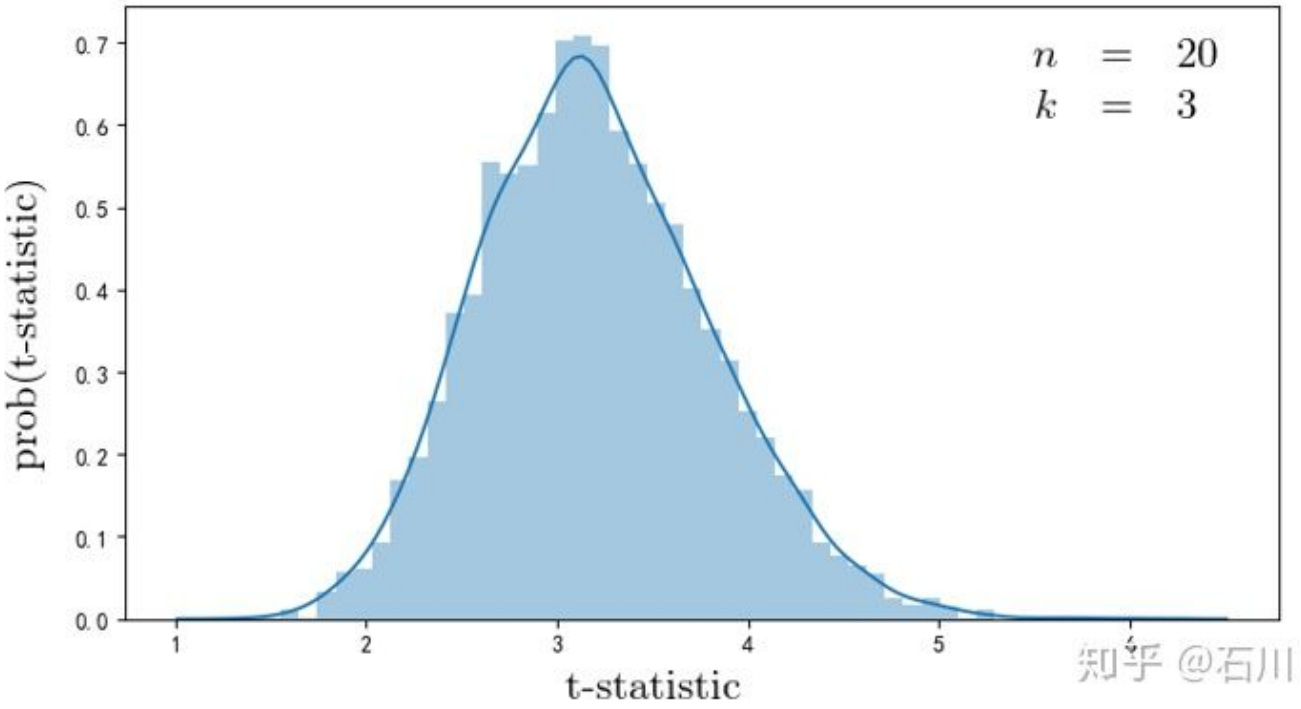
1. 对于每一个因子，在每月末，随机生成 500 支成分股在该因子上的取值并从高到低排列，选择取值最高的 10% 做多、取值最低的 10% 做空，以该多空组合的收益率作为该期因子的收益率

3. 按照每个随机因子 t-statistic 绝对值的大小，挑选绝对值最大的 k (≤ n) 个因子，并按照等权或者正比于它们 t-statistics 的绝对值大小配置因子；

4. 以最终多个因子的配置结果作为最终的选股结果，计算该策略在整个回测期内的 t-statistic；

5. 上述 1 – 4 步完成了 n 选 k（当 k = 1 和 n 时，问题分别变为 n 选 1 和 n 选 n）的一次实验。为了得到 n 选 k 问题中 t-statistic 的经验分布并计算 5% 显著性水平下的 t-statistic 的阈值，对于每一组 n 和 k，将上述 4 步进行 5000 次仿真，从而计算 t-statistic 的阈值。

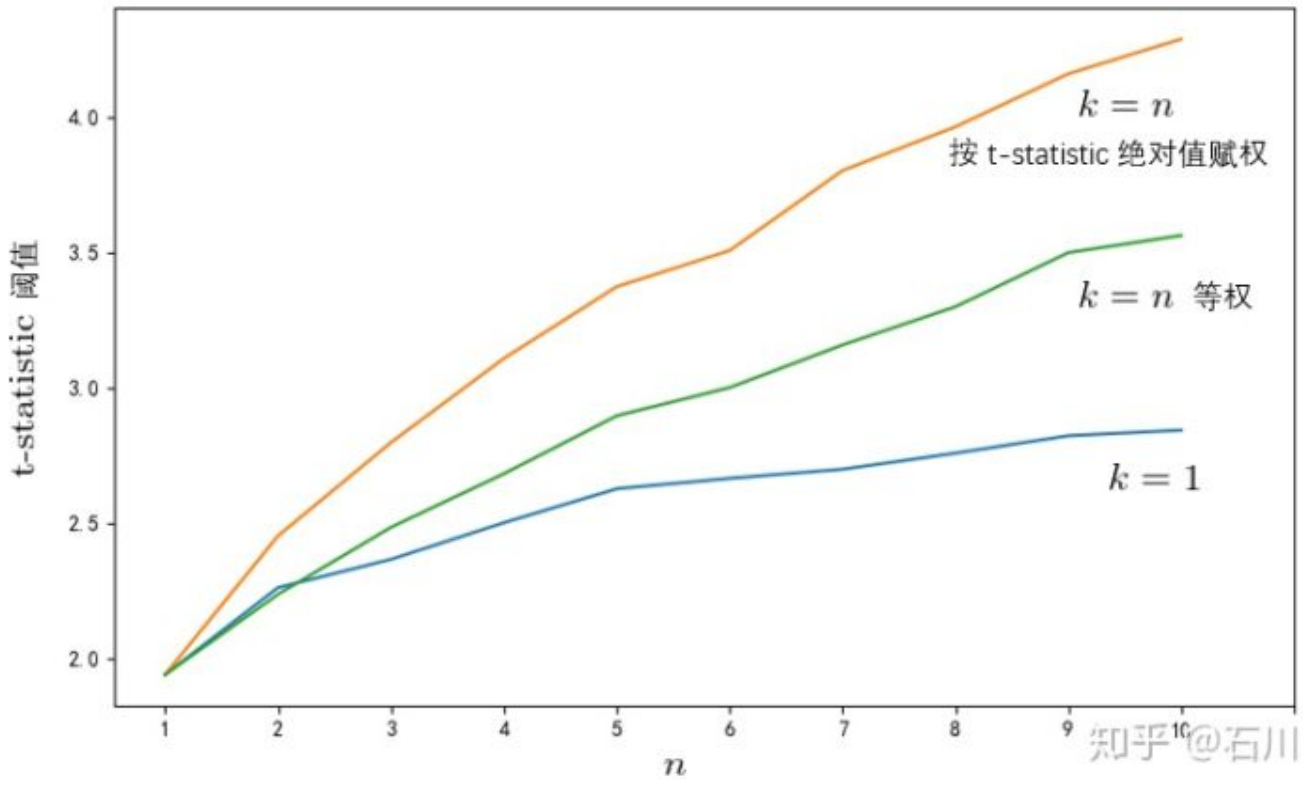
举个例子。下图是当 n = 20，k = 3（即从 20 个随机因子中选出样本内 t-statistic 绝对值最大的 3 个，并按 t-statistic 绝对值大小配置）时，5000 次仿真得到的该策略的 t-statistic 的经验分布，其均值为 3.2，其 5% 显著性水平下对应的 t-statistic（即该分布中 95% 分位数）高达 4.16。



该结果表明，如果我们从 20 个源于业务逻辑（或者很多人乐此不疲的 data mining）的因子中选择 3 个最好的来选股时，该策略的 t-statistic 要超过 4.16 才能认为这 3 个因子的选股效果不仅仅是运气。

接下来看看不同 n 和 k 的取值下，5% 显著性水平对应的 t-statistic 的阈值的情况。下图比较了不同 n 取值下，n 选 1 和 n 选 n 两种极端情况 —— n 选 1 代表仅有 selection bias；n 选 n 代表仅有 overfitting bias。下图传递出以下信息：

1. 随着考察的随机因子个数（n）的增加，策略 t-statistic 的阈值逐渐递增；

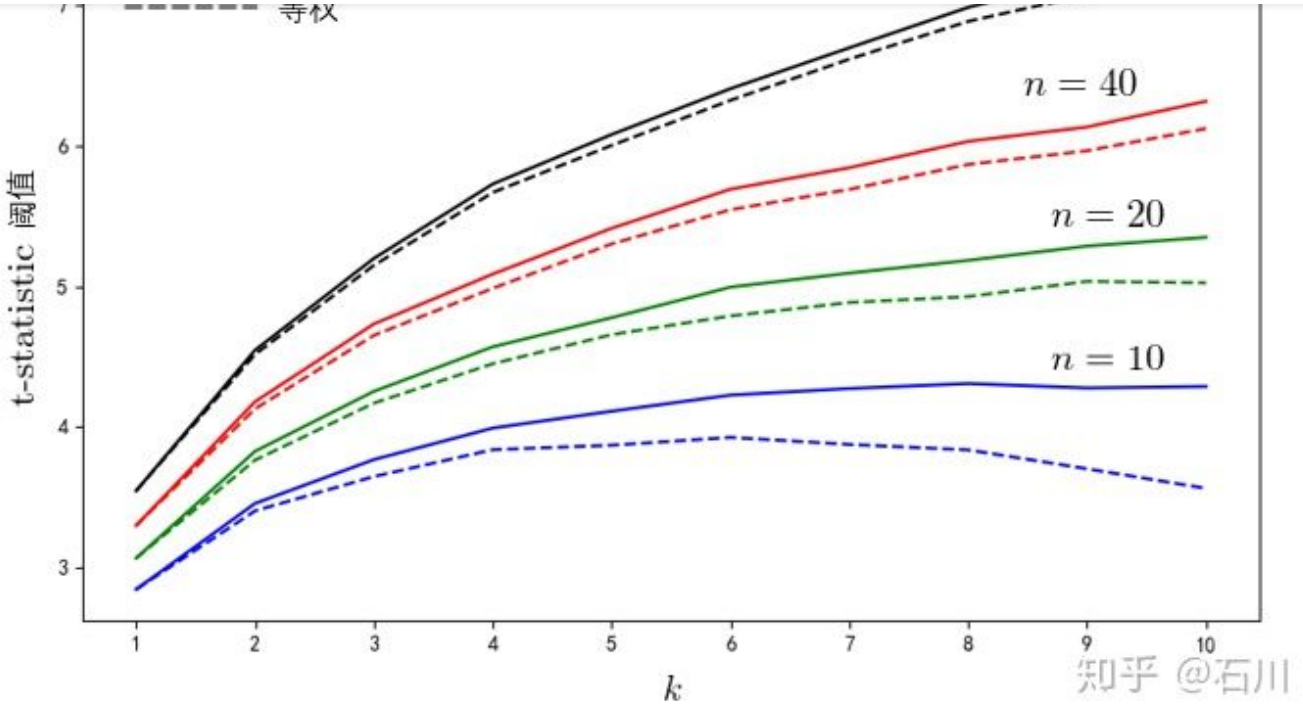


再来看看更一般的 n 选 k 的情况。下图显示了 n = 10、20、40 和 100 时，不同 k 取值下的选股策略的 t-statistic 阈值。在一般的投资实务中，尝试 100 甚至几百个因子并选择其中某些好的是十分常见的。从该实证结果中可以观察到：

1. 随着 n 和 k 的增加，对于按照随机因子 t-statistic 绝对值赋权配置的策略，它们的 t-statistic 阈值递增；

2. 随着 n 的增加，等权配置和按因子样本内表现配置的效果越来越接近；

3. 对于等权配置因子的情况，能够观察到策略的效果并不随 k 递增；比如当 n = 10 时，t-statistic 的阈值随 k 先增大后减小。



上述实证结果中，最有意思的大概是第三条。**对于等权配置的情况，在一开始，使用更多的因子可以降低策略的波动率，从而提升 t-statistic 的阈值；而一旦因子个数超过最优值，越来越多排名靠后的因子被选入，它们会降低策略的收益率，从而降低 t-statistic 的阈值。**这是在因子投资实务中需要考虑的问题。

从图中可以看到，对于实证中考察的最极端情况，即“从 100 个因子选 10 个最好的”，仅仅靠运气，以随机因子构建的策略在中证 500 成分股的样本内回测中就能取得高达 7 以上的 t-statistic 阈值。Data mining 造成的 selection bias 和 overfitting bias 不容小视。

4 结语

近年来，海外学术界越来越意识到 multiple testing 造成的因子分析中 data mining 的问题。一些先进的统计手段被提出以帮助鉴别哪些是真正有效的因子，哪些仅仅是运气。这些文献包括《出色不如走运(II)?》中介绍的那些，以及本文提及的 Novy-Marx (2015)。

在 empirical asset pricing 和 factor allocation 方面，我们都是 data mining 的好手。拿来一个因子，如果不好使，可以对它进行差分 —— 美其名曰增长率；再不好使，二阶差分 —— 美其名曰加速度；还不好使，行业中性、市值中性试一下、用各种其他因子回归得到残差再试一下；对于选出的因子，等权配如果效果不理想，可以按照事后夏普率配一下；还不理想？使用滚动窗口进行动态因子择时……

诚然，对于有严谨金融逻辑的因子 —— 比如 ROE —— 我们没有必要把它和一帮其他“邪门”因子一起比较，然后要求 ROE 也有非常高的 t-statistic，这是对统计手段的走火入魔。但是，**对于**

> *While one should combine multiple signals they believe in, one should not believe in a combination of signals simply because they backtest well together.*

感谢阅读，祝各位新春快乐！节后见。

参考文献

- Novy-Marx, R. (2015). *Backtesting strategies based on multiple signals*. NBER Working Paper, No. 21329.

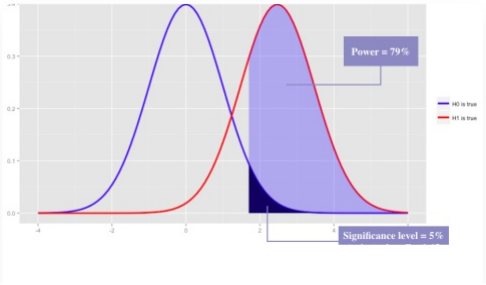
免责声明：文章内容不可视为投资意见。市场有风险，入市需谨慎。

编辑于 2019-07-03

多因子模型 过拟合 数据挖掘

文章被以下专栏收录

推荐阅读



知乎 首发于 川流不息

姚岑卓 发表于数据科学修... Vinjn... 发表于黑客与画家 东山宅

1 条评论 ⇌ 切换为时间排序

写下你的评论...

lakyblu 6 个月前

multiple testing！哈哈，仅仅是marginal screen，还没有算上test statistics是correlated的情况，也没有算上factor是不是unit root... inference是很难的问题，large scale inference则是难上加难

1