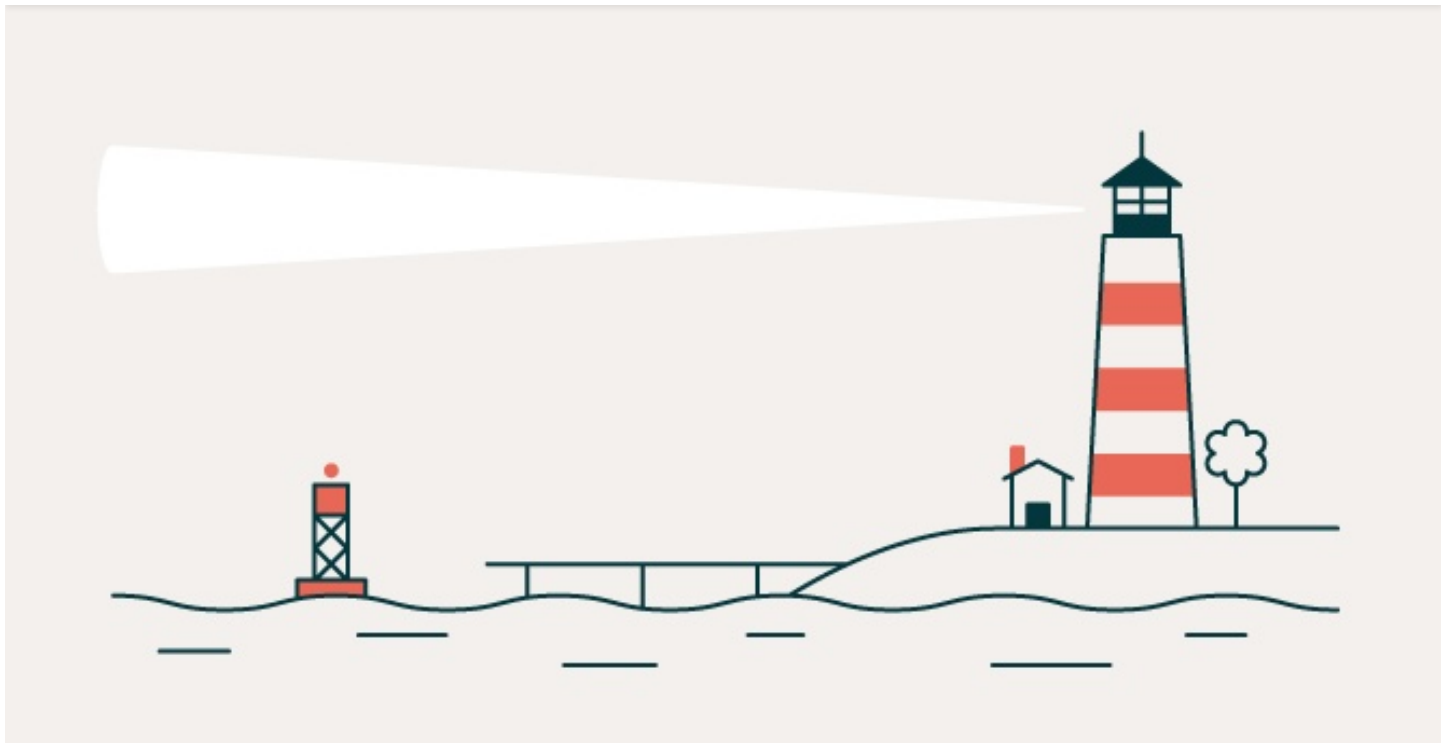


知乎

首发于
川流不息

如果我们能正确预测基本面因子



石川

量化交易 话题的优秀回答者

已关注

47 人赞同了该文章

摘要

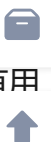
使用历史基本面信息来预测未来的基本面，并基于预测值选股；如果能够预测准确，便可在价格修正以反映最新的基本面时获得超额收益。

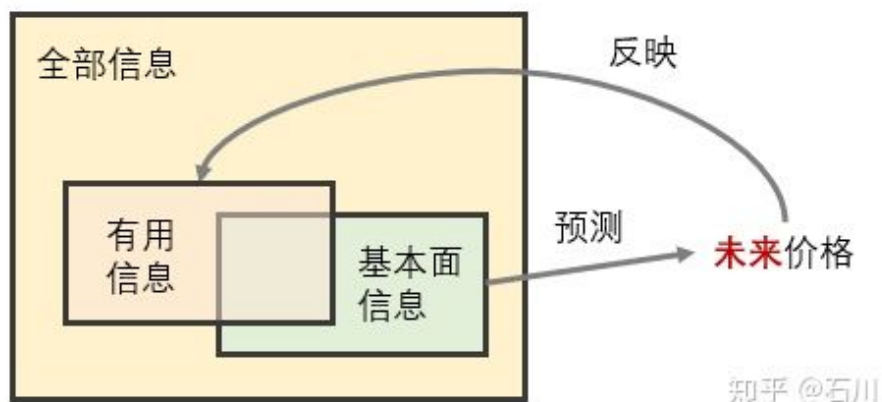
1 引言

基于基本面的量化多因子选股一直是市场中研究和实践的重点，其核心逻辑是找到并配置能够预测股票未来收益率的因子。随着机器学习技术的飞速发展，越来越多的先进算法（比如各种神经网络）被用于挖掘基本面因子和股票未来收益率之间的非线性关系。

诚然，股票的价格和上市公司的基本面密切相关。但是在选股时，我们很难准确判断模型的有效输入。对于投资这种开放性问题，大千世界的任何信息都有可能对决策产生影响，信息量巨大而信噪比极低。

如下图所示，基本面信息仅仅是现实世界全部信息的一小部分，其中还包含了大量的无用噪音；是虚假、错误信息。股票未来的价格反映了全部信息中的一小撮有用信息；然而具体哪些信息有用

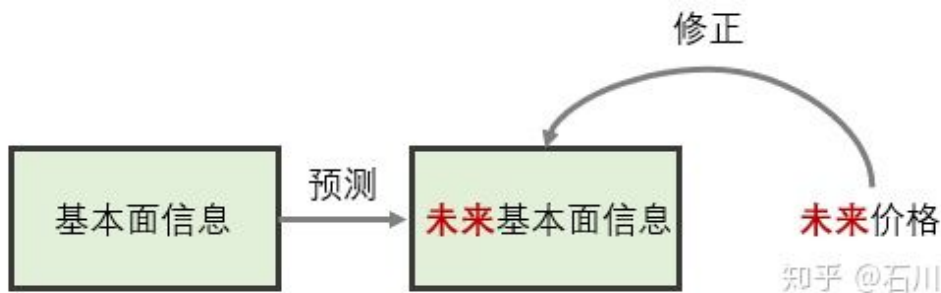




知乎 @石川

此外，虽然市场上存在一些异象无法被有效市场假说（EMH）解释（它们源自行为金融学中的各种认知偏差），但 EMH 仍然在大部分时间相对有效，是主宰市场的第一范式。这就意味着最新的基本面信息已经一定程度上反映在价格之中了；再使用它们来预测未来的价格，其效果会打一定的折扣。

为了应对上述种种困难，学术界和业界在最近几年提出了一个新的思路（如下图所示）：**使用历史基本面信息来预测未来的基本面，并基于预测值选股；如果能够预测准确，当最新财报被公布的时候，价格发生修正以反映最新的基本面信息，使得那些提前布局的投资者获得超额收益。**



知乎 @石川

这种做法背后的假设是能够通过历史基本面信息预测出未来的基本面。我对此的看法是，**这绝不容易**，但与使用历史基本面信息预测未来收益率相比，它背后的逻辑更直接、纯粹，而不像基本面和收益率之间还隔了一层。**由此可以猜想，过去和未来基本面之间的信噪比比过去基本面和未来价格之间的信噪比更高。**

本文的目的有如下两个：

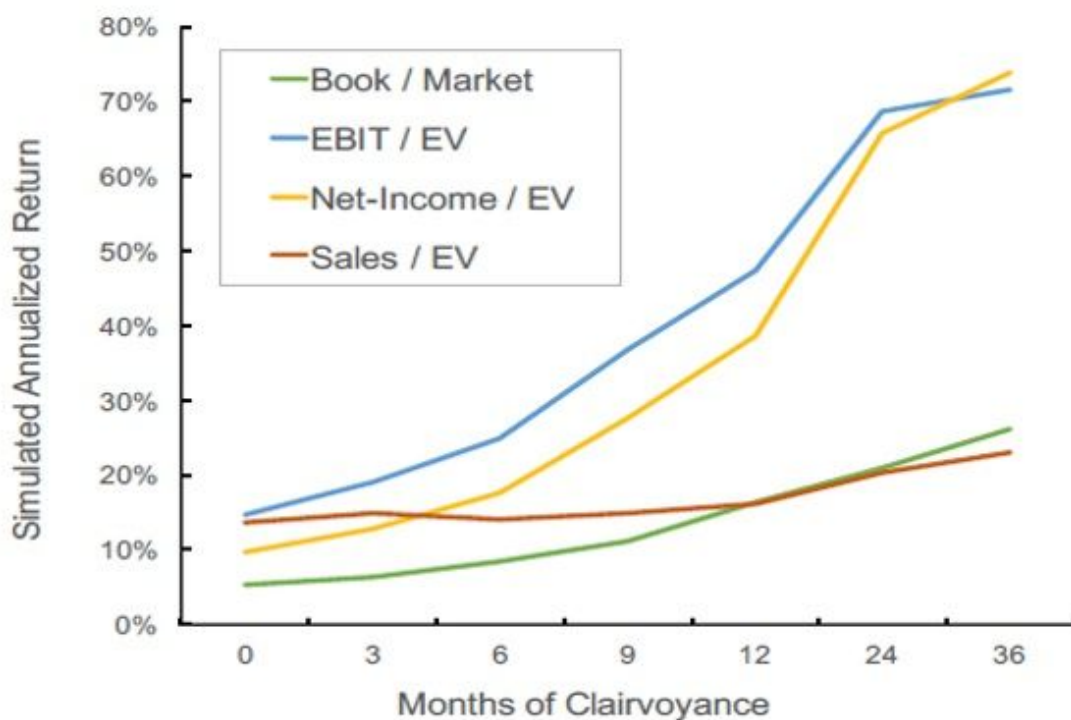
1. 海外的业界基于上述思路，通过机器学习算法预测未来基本面信息、提出了**前视因子模型 (Lookahead Factor Model, LFM)** 的概念。本文对这个方法进行简要介绍。
2. 本文的题目中用了“如果” (if)，而不是“如何” (how)。我希望首先定量探讨通过**预测**来——完美以及不完美的情况下——能否带来超额收益。**只有这个 if 的前提成立了，以 how 才有意义**（而且这个 how 也并不一定非要通过机器学习实现）。在这方面，本文基于中证 500 做一些简单实证。

2 Lookahead Factor Model

2017 年, Alberg and Lipton (2017) 在 NIPS Time Series Workshop 上报告了一篇题为《Improving factor-based quantitative investing by forecasting company fundamentals》的文章。本节对它做简要介绍。

之所以不花太多的篇幅, 是因为这毕竟不是发表在 NIPS 正会上的, 而且它仅仅是一个 poster session, 文章本身也很短, 所以其效果的可靠性仍有待检验。但我仍希望它能给小伙伴们带来一些启发。

Alberg and Lipton (2017) 的研究动机是观察到, 如果能够开天眼知道未来一段时间的基本面信息 (包括 Book-to-Market、EBIT/EV、Net Income/EV 以及 Sales/EV 这些), 并通过它们来选股, 则可以在美股上获得更高的收益, 且预知未来的时间越长 (用 months of clairvoyance 表示) 收益越高 (下图)。



出处: Alberg and Lipton (2017)

知乎 @石川

以此为动机, 他们使用了 Multilayer Perceptrons (MLPs) 以及 Recurrent Neural Networks (RNNs) 两种算法构建了预测模型, 使用历史基本面信息来预测未来的基本面信息。

在回测中, 二位作者使用了 1970 年 1 月至 2017 年 9 月 NYSE、NASDAQ 以及 AMEX 上的股票 (排除了非美国的公司、金融行业公司以及超小市值公司)。模型的输入特征包括 16 个基本面变量和 4 个价格时序变量 (全部是动量类的)。这 16 个基本面特征包括:

3. SG&A (TTM);
4. EBIT (TTM);
5. Net income (TTM);
6. Cash and cash equivalents (当季);
7. Receivables (当季);
8. Inventories (当季);
9. Other current assets (当季);
10. PP&E (当季);
11. Other assets (当季);
12. Debt in current liabilities (当季);
13. Accounts payable (当季);
14. Taxes payable (当季);
15. Other current liabilities (当季);
16. Total liabilities (当季)。

在建模时，所有特征均经过必要的标准化处理；整个回测期被分为样本内（1970 - 1999）和样本外（2000 - 2017）两部分；样本内被进一步分为 training set 和 validation set 两部分。Alberg and Lipton (2017) 使用样本内的 validation set 确定模型的超参数 —— 包括 learning rate、model architecture、objective function weighting —— 以及 early stopping criteria。

为了解决 RNN 容易过拟合的问题，Alberg and Lipton (2017) 特意指出他们的模型是多任务学习，同时预测 16 个基本面指标；损失函数为预测值和实际值之间的均方误差（MSE）。不过作者也指出，在全部 16 个基本面指标中，**他们最终用来选股的是 EBIT/EV，因此在损失函数中对该项赋予了更高的权重。**

实证结果显示，基于神经网络的复杂算法在样本外取得了比 naïve 算法（即使用当期数据猜下一期）更低的MSE。相比 S&P500 指数本身以及传统的因子选股（基于当期 EBIT/EV 选股），两个 lookahead factor models 都取得了更高的年化收益率和夏普率（下表）。

	S&P 500	传统因子选股	LFM-MLP	LFM-RNN
年化收益率	4.5 %	14.4 %	17.1 %	16.7 %
夏普率	0.19	0.55	0.68	0.67
均方误差	N/A	0.62	0.47	0.47

出处：Alberg and Lipton (2017)

知乎 @石川

以上就是关于 Alberg and Lipton (2017) 这篇文章的简要介绍。国内的一些量化论坛基于该方在 A 股上做了实证，感兴趣的朋友可以找来看看。

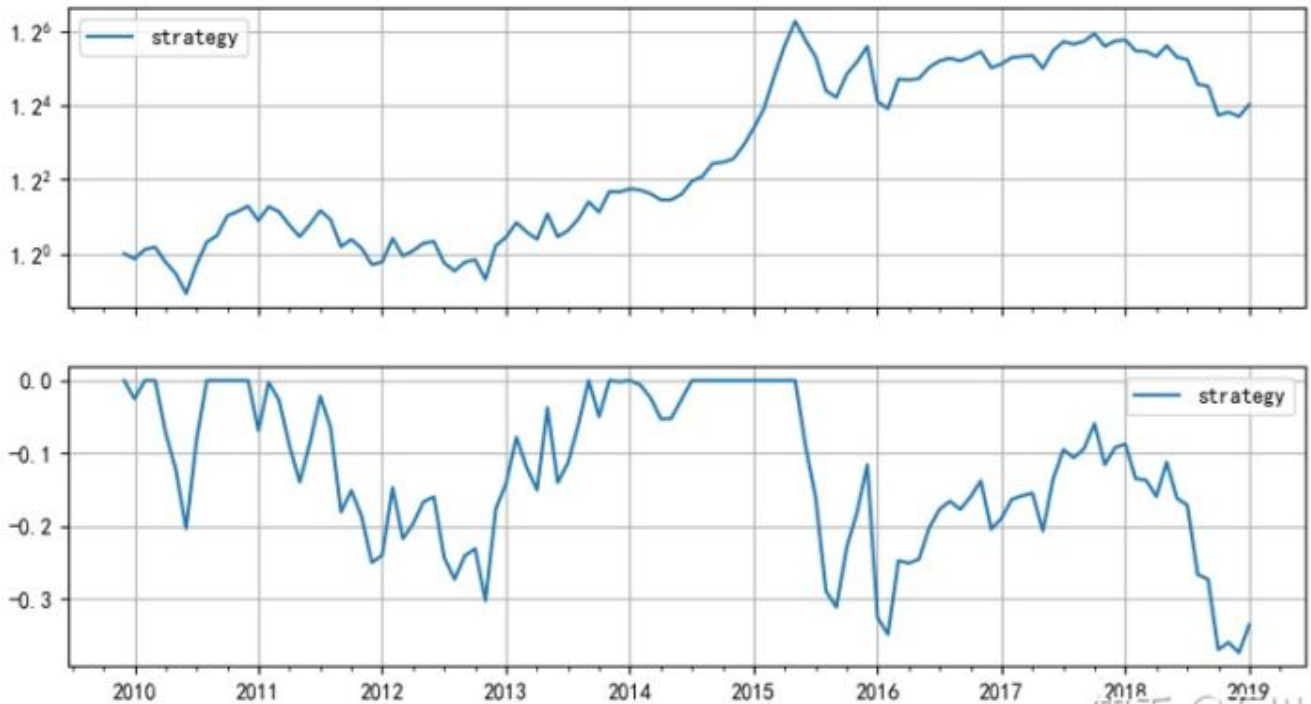
知乎

首发于
川流不息

本节和下一节来回答本文关注的 if 问题。我们以中证 500 成分股为选股池，回测期为 2009 年 12 月到 2018 年 12 月，选择 Earnings per Share (EPS) 作为目标基本面因子来考察成功预测未来 EPS 能否获得超额收益。

首先，来看看常规做法：每月末按最新 EPS 数据选取该指标最高的 20 支股票等权配置，按月调仓，不考虑任何交易成本。该选股的效果如下。

基于历史 EPS 选股



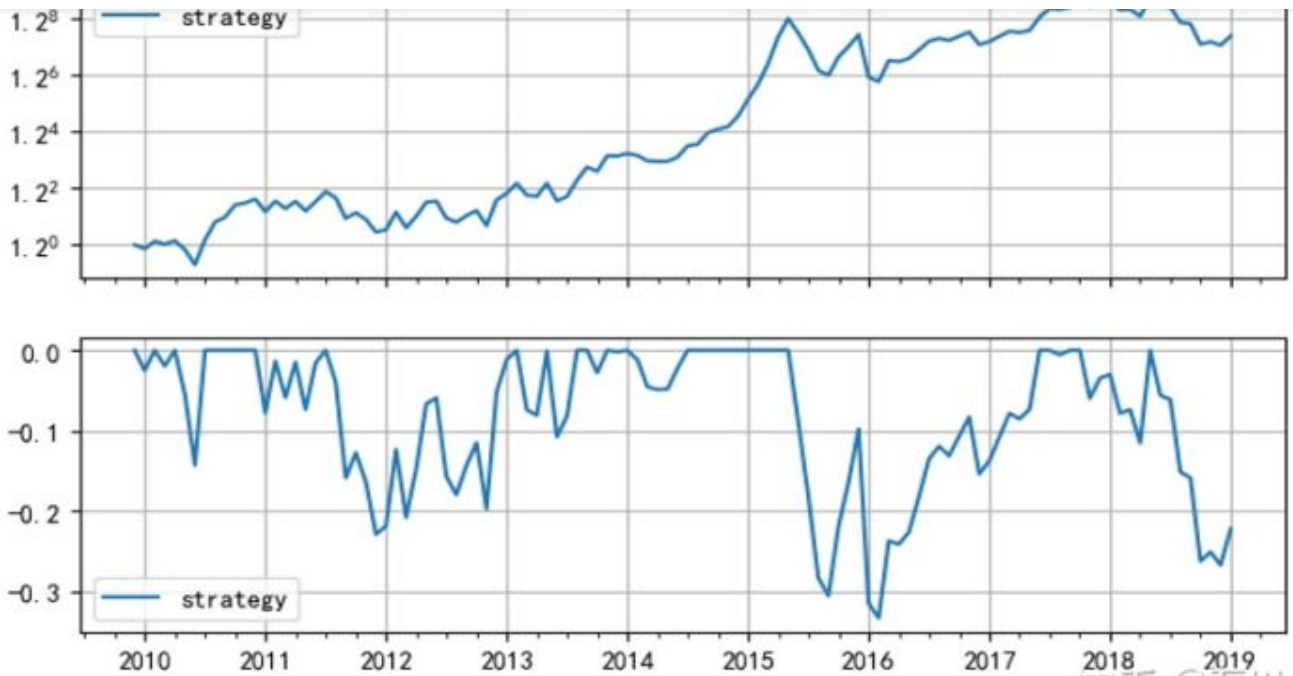
数据支持: RiceQuant

知乎 @石川

接下来，假设开天眼已知未来一个月的 EPS，并基于此选择 20 支 EPS 最大的股票等权配置、按月调仓。该策略的表现如下。



知乎

首发于
川流不息

数据支持: RiceQuant

知乎 @石川

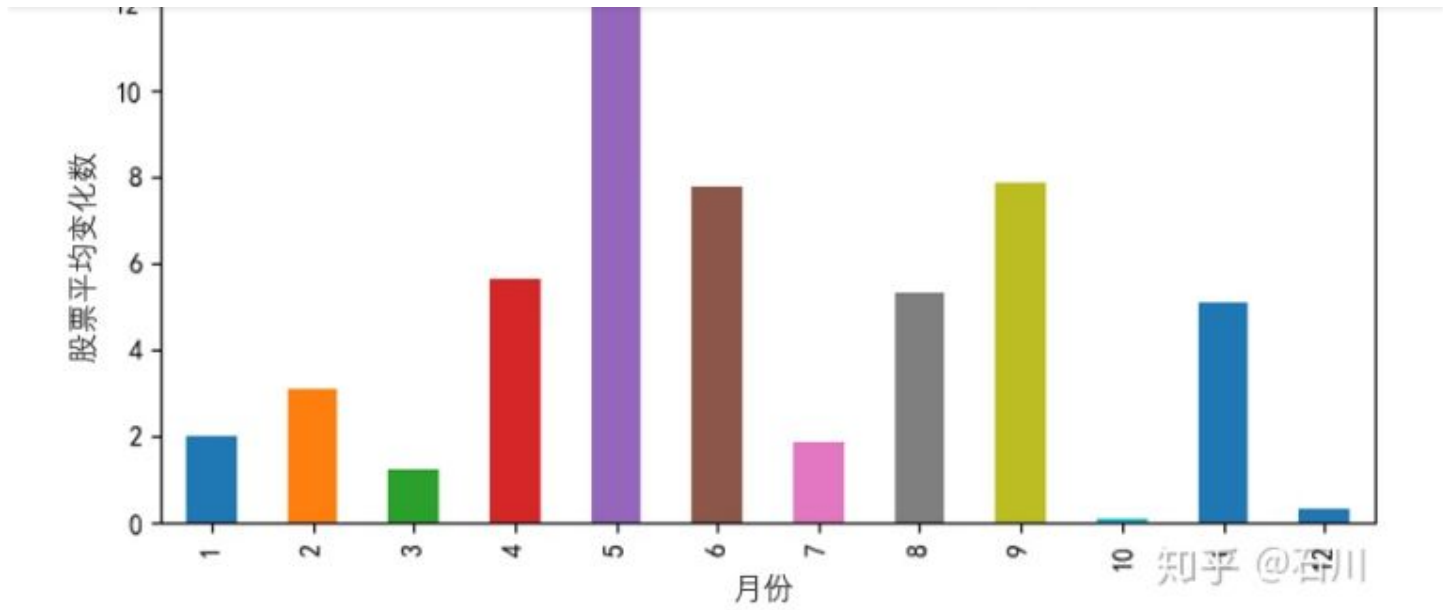
将这两个选股策略和中证 500 指数本身放在一起比较，高下立判。**无疑，EPS 本身是一个非常优秀的选股因子，而如果能预知未来 EPS 则可以获得更大的优势、更高的收益。**



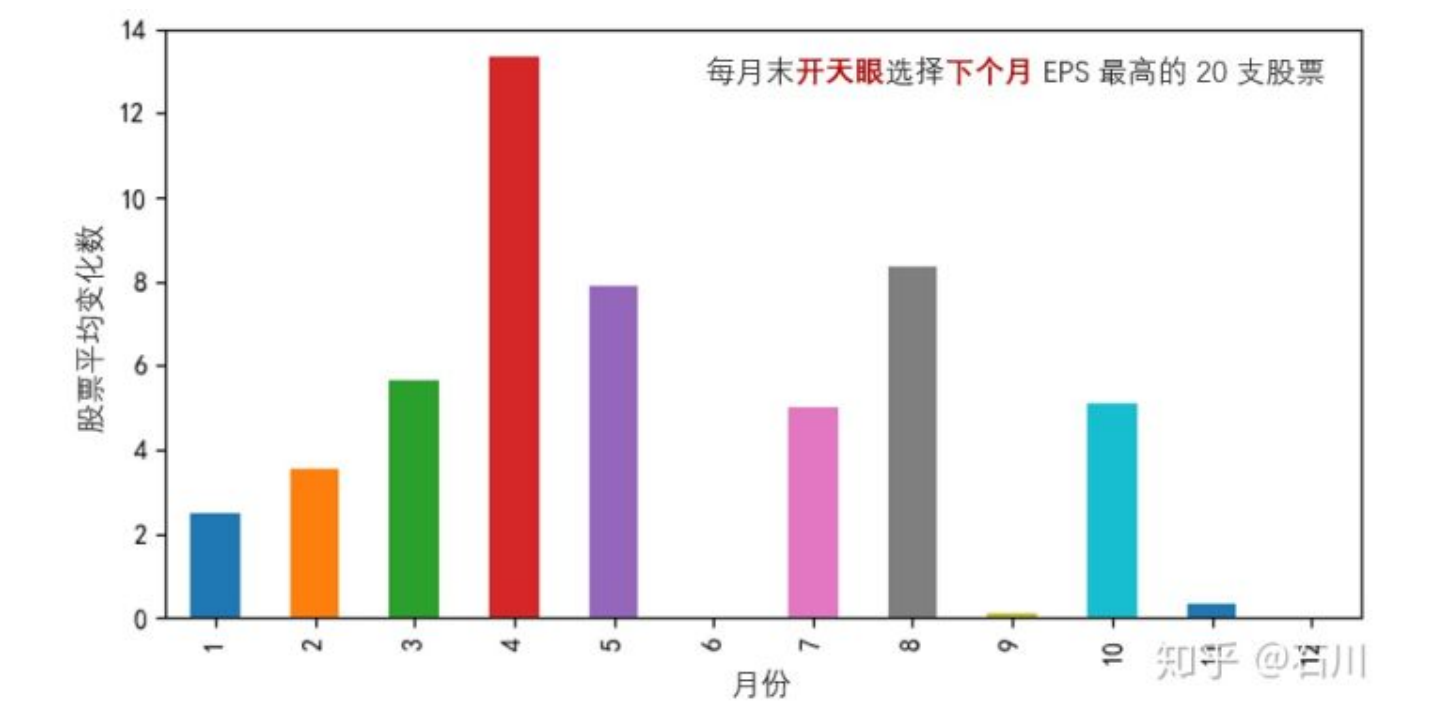
知乎 @石川

发生这种现象背后的原因是什么呢？

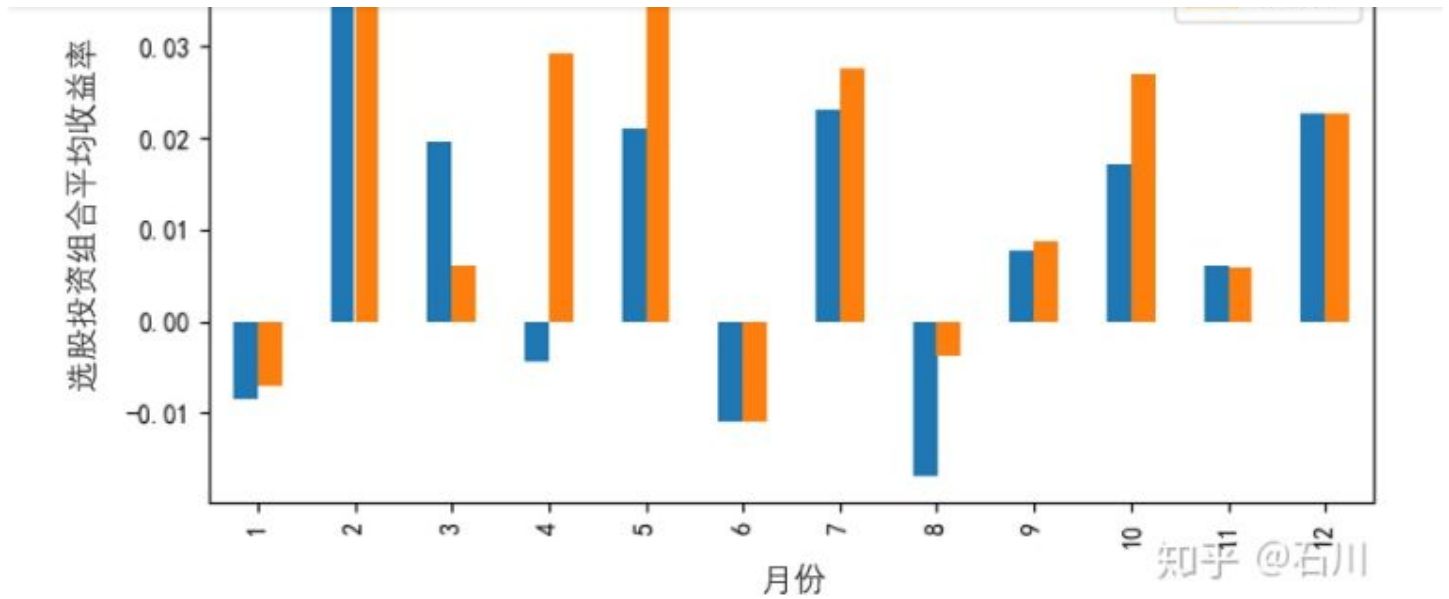
下图显示了使用当期 EPS 选股时，回测期内每月平均换股的数量。其中变化最多的是五月份。这是因为上市公司需要在每年 4 月 30 日之前披露上一年的年报。因此在每年四月底，使用最新的 EPS 选出的股票往往较之前的选择有较大变化，这便解释了五月份的持股较四月份的持股变化最大。其他月份的股票变化和季报、中报披露以及中证 500 调整成分股有关。



类似的，下图显示了使用下个月 EPS 选股时，每月平均换股的数量。比较这前后两张图，其中最大的区别就是正常情况下五月份的变化提前在四月发生（因为我们假设在三月底就知道四月最新披露的 EPS 中最大的股票）。



对比使用当期 EPS 选股和使用未来 EPS 选股的每月平均收益率，可以看到前面提到的这种前移带来的巨大差异。正如下图所示，开天眼（图中 Oracle）选股的四月份收益率较正常（图中 Regular）选股的收益率有巨大提升。此外，开天眼选股在八月份也有巨大的提升（8 月 30 日是中报披露的截止日期）。

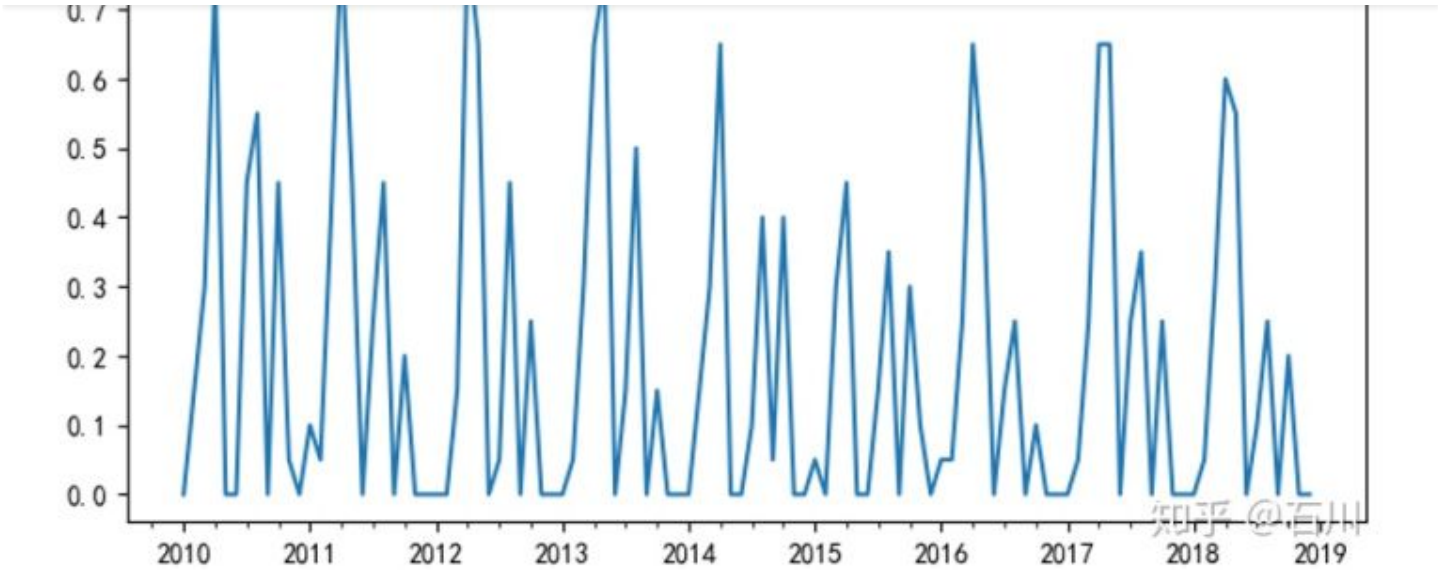


以上结果说明，如果能准确的预测并使用未来的 EPS 选股，确实能够获得更高的收益。不幸的是，现实中我们不能开天眼。

4 不完美预测能否带来提高？

第三节的结果仅仅是一个美好的愿景。在实际中，即便能够在一定程度上预测未来的 EPS（或其他基本面信息），其准确性也难以保证。**在不完美的预测下，这种努力是否能够带来超额收益呢？**这就是本节探讨的问题。

以开天眼的结果作为 EPS 选股的标准答案，可以计算出使用历史 EPS 选股的错误率。从下图不难看出，由于基本面指标变动的频率较低，**在很多没有新信息披露的月份里，选股的错误率也很低，但是这并没有什么作用（这一点在构建损失函数的时候值得考虑）。起决定性作用的是错误率高的那些月份。**



上一节的分析指出，使用未来 EPS 选股的优势在于提前知道年报和中报（特别是前者）中最新的 EPS，因此四月和八月贡献了巨大的超额收益。观察上图不难发现，使用历史 EPS 选股时，四月份的错误率也是全部十二个月份中最高的。下表统计了回测期内每年四月和八月使用历史 EPS 选股的错误率。

月份	错误率	月份	错误率
2010 年 4 月	75 %	2010 年 8 月	55 %
2011 年 4 月	80 %	2011 年 8 月	45 %
2012 年 4 月	80 %	2012 年 8 月	45 %
2013 年 4 月	65 %	2013 年 8 月	50 %
2014 年 4 月	65 %	2014 年 8 月	40 %
2015 年 4 月	45 %	2015 年 8 月	35 %
2016 年 4 月	65 %	2016 年 8 月	25 %
2017 年 4 月	65 %	2017 年 8 月	35 %
2018 年 4 月	60 %	2018 年 8 月	25 %
平均	66.7 %	平均	39.4 %

由此可见，预测 EPS 的努力可以考虑以降低四月和（或）八月选股的错误率为目标。为了回答“不完美预测能否带来提高”这个问题，首先考虑四月为目标并按照以下逻辑进行仿真实验。

对于给定的正确率 c ，在回测中每年的三月底从四月 EPS 最高的 20 支股票中随机选出 $20 \times c$ 支，作为预测模型正确的部分；从剩余 480 支股票中随机选出 $20 \times (1 - c)$ 支作为预测模型错误的部分；如此便完成一次选取（这是一个非常粗糙的处理方式，因为从剩余 480 支股票中随机

知乎

首发于
川流不息

为正确率 c 下该预测模型的选股结果，以此考察 EPS 预测正确率和选股效果的关系。

下表中 Panel A 展示了四月份 EPS 预测不同正确率下选股的效果。结果表明，随着正确率的提升，收益率和夏普率逐渐增加，选股效果远强于使用当前 EPS 的表现（使用当前 EPS 的年化收益率为 6.33%；夏普率为 0.37）。同时也看到，即便是把正确率提高到 90%，其选股效果也显著弱于全部使用未来 EPS 的情况，这说明仅仅改进四月份的正确率是不够的。

Panel A: 改进四月 EPS 预测正确率后的选股效果

正确率	0.5	0.6	0.7	0.8	0.9
年化收益率 (%)	9.74	10.09	10.57	10.84	11.11
最大回撤 (%)	-40.78	-40.68	-39.95	-39.95	-39.61
夏普率	0.49	0.50	0.52	0.53	0.54

Panel B: 改进四月和八月 EPS 预测正确率后的选股效果

正确率	0.5	0.6	0.7	0.8	0.9
年化收益率 (%)	11.73	11.89	12.31	12.48	12.65
最大回撤 (%)	-34.92	-34.97	-34.80	-35.19	-35.40
夏普率	0.57	0.57	0.59	0.59	0.60

上表中 Panel B 展示了同时提高四月和八月正确率的结果。随着年报和中报披露月份的同时改进，选股效果较 Panel A 中的结果进一步显著提升。当然，我们也必须注意到，**正确率的边际效应在逐渐减小**。上述结果留给我们的启发有以下两点：

1. 以月频进行基本面预测时，应该着重考虑年报和季报集中披露的月份的正确率；
2. 正确率对选股效果提升的边际效应逐渐减弱，当对重点月份预测的正确率提高到一定水平后，可考虑提升其他月份预测的正确性。

从实证结果来看，即便是不完美的预测也是值得尝试的。

5 结语

上周的文章中提到了“预期差”的概念——如果我们能找到市场中的预期差，便可以利用它来获得超额收益。本文提到的准确预测 EPS 其实质也是提前捕捉预期差，等待价格向价值修正。

想再次强调的是，针对预测基本面这个话题，本文探讨的是 if 的问题，而不是 how 的问题，并过实证给出了预测中一些可能需要关注的地方。在 how 的问题上，希望文中介绍的 Alberg and

知乎

首发于
川流不息

在基本面的预测方面，无论采用什么方法，预测准确才是最关键的。这就是为什么优秀分析师的盈利预测（一致预期数据）是很值钱的，因为这些数据确实能够带来 α 。只不过很多时候，这些 α 的成本太高了。这就给了机器学习契机。希望在未来能够看到学术界和业界在这方面的更多突破，将先进机器学习算法运用到上市公司财务数据的预测中，为低成本的获取 α 提供新的可能性。

参考文献

- Alberg, J. and Z. C. Lipton (2017). *Improving factor-based quantitative investing by forecasting company fundamentals*. NIPS Time Series Workshop 2017 (sites.google.com/site/nips-tsw/), available at arxiv.org/pdf/1711.0483....

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。

原创不易，请保护版权。如需转载，请联系获得授权，并注明出处，谢谢。已委托“维权骑士”（[维权骑士 免费版权监测/版权保护/版权分发](#)）为进行维权行动。

编辑于 2019-07-03

多因子模型 基本面研究 量化交易

▲ 赞同 47 ▼ ● 16 条评论 ➤ 分享 ★ 收藏 ...

文章被以下专栏收录

**川流不息**

北京量信投资管理有限公司是一家在中国基金业协会备案登记的专业私募基金管理人...

[关注专栏](#)

推荐阅读

因子对称正交

可关注个人公众号：信号处理与机器学习 wave_cottage 由于金融数据信噪比极低，为了鲁棒性和可解释性，传统选股都采用的线性模型

带权重的多因子策略1.0

为大家带来社区一个多因子策略，适合入门进行观看哦。1.初始股票池：沪深300 `context.stocks_set = index_components('#39;000300`





淮浩

发表于信息论与信...

johnsonchak

王克勤

16 条评论

切换为时间排序

写下你的评论...



60荷兰盾

7 个月前

how的问题还是挺难的，财务指标和个股，行业周期，宏观经济都相关，不过还是值得花时间研究一下的，感谢~

赞



石川 (作者)

7 个月前

完全同意.....

赞



云中客

7 个月前

我想过一种因子分析方法，就叫广义宏观因子分析吧，方式是不分析财务经济等价格走势直接相关数据，而是分析人类社会宏观运转规律，和金融走势比较，寻找些微相关，因为人类社会运转规律性很强，所以这种预测方法可以非常稳健持久。但不适合股票这种比较狭隘的品种，对外汇，期权等全球性品种效果应该很好，您觉得这个想法如何

赞



云中客

7 个月前

我查了下好像目前没有用这种方式预测的，偶尔有类似的，如高盛一个实习生发明了一种使用电商网站某种特定物品销售额预测金融走势的例子，这个间接相关例子和我说的有点相似，但我的想法是预测因子要足够宏观稳健就像一年四季一样永远持续。您觉得可行吗

赞



石川 (作者) 回复 云中客

7 个月前

华泰金工最近一两年弄了个频域分析指出跨市场跨国家存在三大周期.....虽然我完全不相信频域分析（因为在时序的末端是失真的），但也推荐你找来看看。：）

赞



云中客

7 个月前

为什么放着直接相关的经济财务指标不分析要分析间接宏观相关因子呢，就像文章说的，事物之间的高低差异，才是预测的核心，财务因子研究的人太多，没有什么差异价值了，间接宏

知乎

首发于
川流不息

石川 (作者) 回复 云中客

7 个月前

关于宏观经济对股市的预测作用，海外有一些研究。我简单在A股试了试，效果一般，可看[这里mp.weixin.qq.com/s/AQsk...](https://mp.weixin.qq.com/s/AQsk...)

赞



赵志强

7 个月前

如果这个思路有效的话，比尔盖茨和马化腾就会一直增持自家股票了。

赞



石川 (作者) 回复 赵志强

7 个月前

哈哈。以下我的朴素想法哈，不知道您如何定义“有效”。从开天眼效果显著超过用历史数据来说，if的问题是有效的。下面是how的问题，这肯定很难，当然如果最终证明咋都预测不了，那确实how也无效。至于您说的大咖增持自己股票的问题，我的看法是当财富到达那种级别后，他们首要的目标是保值和稳健增长了。中美实证都显示，即便是开天眼，纯多头选股的波动和回撤也是非常大的，这估计不是大咖想要的。第二，即便高管提前知道最新财报，利用内部信息提前交易也是严重不合规的。

赞



郭嘉公元

7 个月前

为什么图上显示 ricequant?

赞



石川 (作者) 回复 郭嘉公元

7 个月前

我用了米筐的数据.....

赞



吴二真

5 个月前

你拿分析师预测的eps数据再做一次呗？看看是不是在naive和开天眼之间？

赞



石川 (作者) 回复 吴二真

5 个月前

分析师一致预期eps还是有用的:)

赞



吴二真 回复 石川 (作者)

5 个月前

我知道这个数据有用，但是不一定是文中那个用法。我的意思是，按文中那种用法回溯一下看看，结果如何？



知乎

首发于
川流不息

财政保守马妒骡

1 个月前

问个可能有点冒犯性的问题，经常看你们做量化的说alpha，是不是说beta之类的不可能大于这个百分之一二十？我之所以对这个问题感到疑惑，是因为我自己做交易的时候就只用了K线组合和几个均线，在外汇上就做到了很高的收益，难道仅凭K线组合和几个均线组成的系统就获得了很厉害的alpha？

👍 赞 ↩ 回复 🗑 踩 🚩 举报



Deva Fang 回复 财政保守马妒骡

11 天前

“仅凭K线组合和几个均线组成的系统”个人认为是基于投资者行为信息，是获得了alpha。这里有个资金容量的问题，你的方法获得了很高回报，但是容量可能不是很大。如果管理一个亿资产，你可能无法维持原来的高回报。还有你的估计是主观交易方法，很难量化，即使量化了效果也会大打折扣。

👍 赞

