机器学习时代的回测规程

原创：石川 川总写量化 5月23日

作者：石川，北京量信投资管理有限公司创始合伙人，清华大学学士、硕士，麻省理工学院博士。知乎专栏：
https://zhuanlan.zhihu.com/mitcshi。

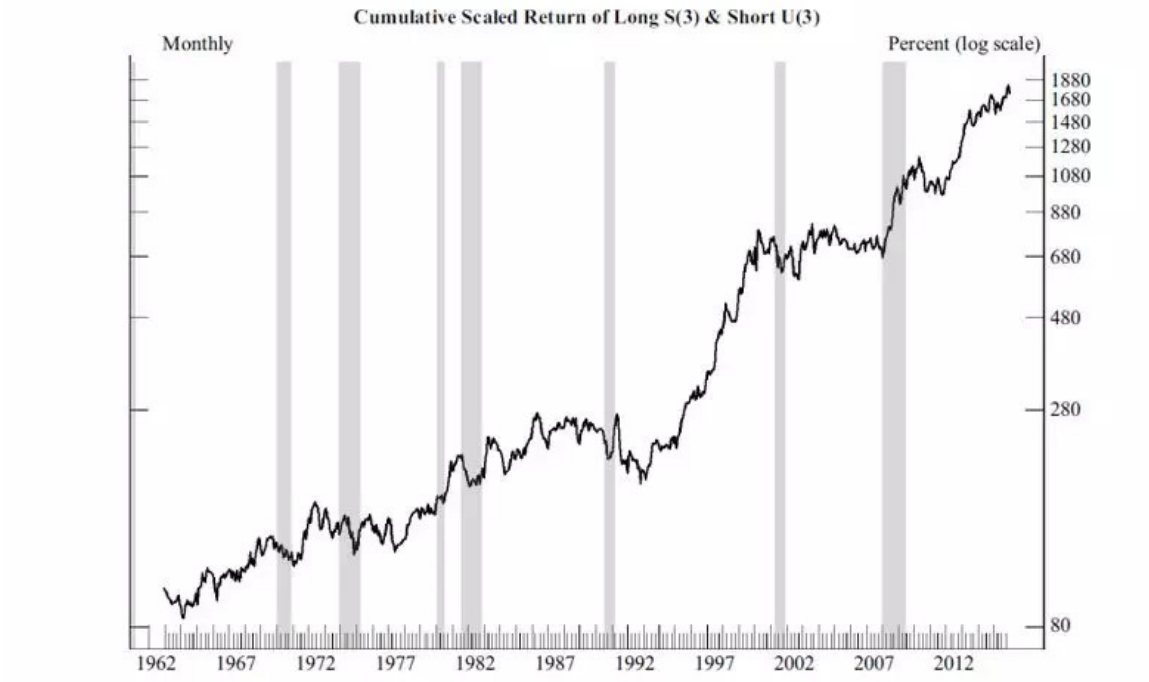
摘要 在回测中牢记并遵守这些准则可以有效降低过拟合的风险、避开噪音、找到真正在样本外可持续的因果关系，获取更高的收益。

0 引言

让我们从下图这个令人欣喜的回测（backtesting）说起。

EXHIBIT 1

Long–Short Market-Neutral Strategy Based on NYSE Stocks, January 1963 to December 2015



Notes: Gray areas denote NBER recessions. Strategy returns scaled to match S&P 500 T-bill volatility during this period.
Source: Campbell Harvey, using data from CRSP.

出处：Arnott, Harvey, and Markowitz (2019)

上图是某针对美股的选股策略在长达 50 年的回测内的净值曲线。该策略采用多空对冲、市值中性的方法构建。该策略表现出了五大优秀量化策略的**必要不充分**（呵呵）特征：

1. 因子计算的方法在回测期内完全一致，没有任何变化；

2. 该策略的表现在近期并没有变差的迹象，说明在该因子上并没有发生“拥挤”；

3. 该因子穿越牛熊，在金融危机时代甚至出现了上涨（在可以做空的假设下）；

4. 该因子和其他主流因子（包括市场、Size、Value、Momentum 等）的相关度极低；

5. 该因子的年换手率仅为 10%，交易成本可以忽略不计。

Too good to be true?

没错，它正是 data mining 的产物。该因子的构建完全没有使用任何基本面或者交易数据，而仅仅依赖美股上市公司股票代码上的字母。比如苹果公司的股票代码是 AAPL，该代码上的第 1 至 4 位上的字母分别为 A、A、P 以及 L。该因子的构建方法是做多股票代码第三位字母为 S 的股票、做空股票代码第三位字母为 U 的股票（记为 S(3) – U(3)）。

在实验中，考虑股票代码的前 3 位字母；考虑到全部可能的 26 个字母，以及每个字母可以出现在多、空两头，因此实验中有成千上万种组合方式。而 S(3) – U(3) 这种组合正是从这些组合中脱颖而出的、具备了上述五大优秀特征的、仅仅来自 data mining 的虚假策略。

上面这个策略是靠蛮力（brute force）找到的，并不能说是机器学习（Machine Learning）的产物。机器学习会进行仔细的交叉验证（cross-validation）以确保我们在训练集和测试集上看到相似的结果。不幸的是，上述策略在整个回测期内的稳定表现大概率会让它通过交叉验证。这背后的原因是股票市场的数据容易出现路径依赖，造成训练集和测试集之间并不独立。

这个例子说明，**量化投资的小伙伴在回测基于机器学习的策略时将面临很大的挑战。**回测的目的是去伪存真，排除噪音、发现预测指标和资产收益率之间真正的因果关系，从而在样本外的实盘交易中获得收益。如果回测不靠谱、落入各种陷阱，那么实盘的结果则可想而知。这个问题在机器学习如此普及的今天显得更加严重。

为了帮助量化交易者更好的杜绝样本内的过拟合，提高发现真正有效策略的概率，三位大咖站了出来：来自 Research Affiliates 的 Robert Arnott，杜克大学教授、前 AFA 主席 Campbell Harvey，以及诺贝尔经济学奖获得者 Harry Markowitz 在 IPR Journals 的最新成员 Journal of Financial Data Science 的处女刊上发表了一篇题为 **A Backtesting Protocol in the Era of Machine Learning** 的文章（Arnott, Harvey, and Markowitz 2019）。

本文中我用“规程”来对应 Protocol 一词，它也可以被译作“协议”或者“清单”，其目的就是通过逐步遵循这些准则来减少样本内过拟合的可能性。这个 protocol 之于回测可靠性的作用就好比飞行员的 checklist 之于飞行安全的作用。Arnott, Harvey, and Markowitz (2019) 一文提出的 protocol 一共包括七部分，它们是：

1. 研究动机；
2. 多重检验；
3. 样本选择和数据；
4. 交叉验证；
5. 模型动力学；
6. 模型复杂度；
7. 研究文化。

它们构成了一个完整且可操作的体系，能够帮助我们更好的规避样本内的虚假信号、找出能在样本外更有效的交易策略。前文《所有样本数据都是样本内》曾论述过 protocol 中的第四部分。不过，鉴于它的系统性，我想用今天这篇文章把这七个角度全部梳理一下。

以下行文并不会逐字逐句的转述 Arnott, Harvey, and Markowitz (2019) 提出的每一个 bullet point，而是会结合我有限的经验和粗浅的认识解读我认为最重要的一些内容。浏览本文并不能 100% 代替阅读原作，因此强烈建议感兴趣的小伙伴找来 Arnott, Harvey, and Markowitz (2019) 看一看。

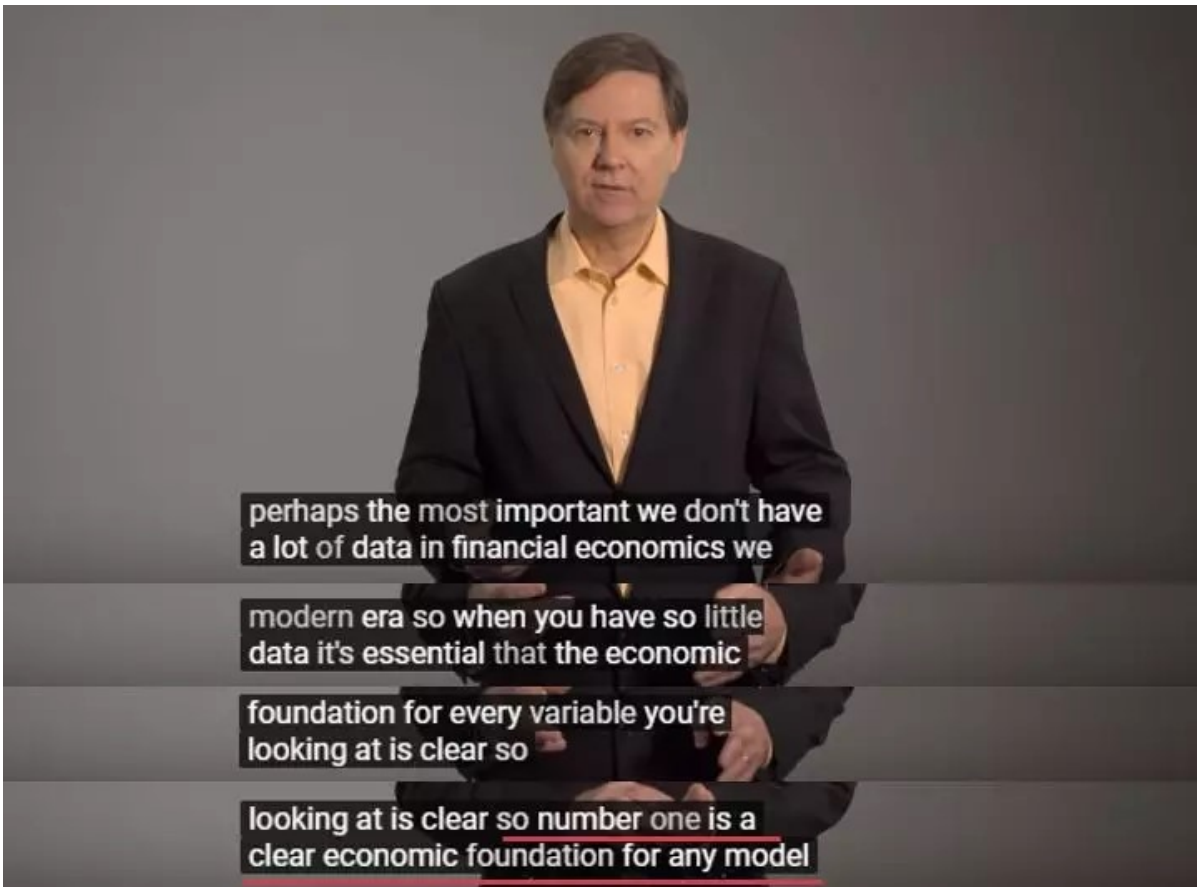
此外，由于公众号之前在倡导科学回测和防止过拟合方面也做过许多努力，很多文章都能很好的 fit 进这个 protocol，所以会在行文中把它们串联起来。

下文第 1 到第 7 节将分别论述这个 protocol 的七个方面。第 8 节总结全文。

1 研究动机

回测规程的第一个方面是**研究动机（Research Motivation）**。

Harvey 教授直言，**金融领域的数据样本太少了**（也许超高频除外）。以美股为例，现代金融时代的股票月频数据大概只有 700 期（相当于 60 年），这对于机器学习应用来说太少了（回想一下 A 股，通常单因子评测的回测期只有区区 10 年，真是太短了）。因此，这个 protocol 中第一也是最重要的一点就是 **a clear economic foundation for any model** —— **任何策略都应该有一个理论先验**。注意，是先验，而不是看到数之后再“真香”编故事。



Chordia, Goyal, and Saretto (2017) 使用基本面指标的不同组合方法构建了两百万个针对美股的因子策略。在实验设计中，他们对 data mining 进行了必要的惩罚，并最终找到 17 个在统计上和经济上都显著的因子。

其中一个因子的构建方法为：分子是 long-term debt issuance 和 preferred stock redeemable 之差；分母是 minimum rental commitments four years into the future。这个因子使用了三个财务指标，但是该组合却毫无业务含义。而上述其他 16 个“显著”的因子都具有类似的结构，它们都是 data mining 的结果。

在现实中，人们往往站在“任何策略都应该有一个理论先验”的对立面上，即先看数据再找理由。比如对于前面那个 S(3) – U(3) 的例子。它的那些优秀特征会让人去寻找虚假的理论依据来说服自己。当一个人能够为 S(3) – U(3) 找到理由，那么如果回测的结果显示相反的结果，即 U(3) – S(3)，相信 TA 也能够找到理由。

> *Any suspicion that the hypothesis was developed **after** looking at the data is an obvious red flag.*

2

多重检验

Protocol 的第二方面是**当心多重检验（Multiple Testing and Statistical Methods）**。公众号的小伙伴对它一定不陌生，之前的文章《出色不如走运？》、《出色不如走运 (II)？》、《出色不如走运 (III)？》谈的全是它。

多重检验指的是：**当我们测试一个策略的许多组参数，或者很多选个因子时，仅仅依靠运气，这些参数或者因子中效果最好的那个就能在样本内获得很高的夏普率（这也被称作 inflated Sharpe Ratio）**。在回测时必须时刻考虑多重检验的影响。

用白话的理解就是：如果我以某个金融学或经济学原理为先验，构建了一个因子并测试有效，那么它大概是真有效；然而，如果我两眼一抹黑试了 100 个因子，然后只挑出了最好的那一个，那么这个因子很可能只是个 lucky factor。

Bailey and Lopez de Prado (2012, 2014) 专门就 inflated Sharpe Ratio 进行了探讨。他们假设不同参数的策略的夏普率满足均值为 E[SR]、方差为 V(SR) 的正态分布。在上述假设下，N 组不同参数中样本内最大的夏普率的期望满足（式中 γ 是欧拉-马斯刻若尼常数）：

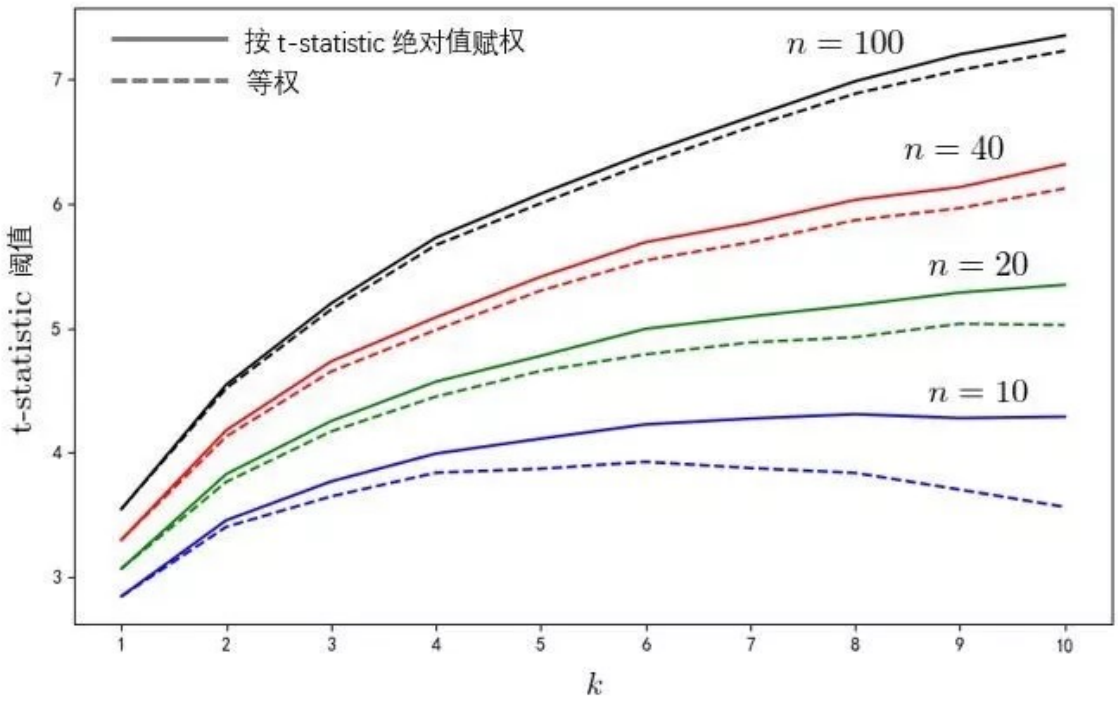
$$\begin{aligned} \mathrm{E}[\mathrm{SR}_{\max}] &\approx \mathrm{E}[\mathrm{SR}] \\ &+ \sqrt{\mathrm{V}(\mathrm{SR})}\left((1-\gamma)Z^{-1}\left[1-\frac{1}{N}\right]+\gamma Z^{-1}\left[1-\frac{1}{N}e^{-1}\right]\right) \end{aligned}$$

该关系式表明，样本内的最大夏普率随 N 增大和 V(SR) 增大。假设 V(SR) = 1，则我们只需要测试 100 组设定，样本内的虚高夏普率就高达 2.5，尽管它对应的 null hypothesis 是该策略真实夏普率为 0。这就是不考虑多重检验的危害。

在《出色不如走运 (III)?》一文中，我们根据 Novy-Marx (2015) 的方法、使用中证 500 的成分股做了随机因子的实证。在实证中，纯随机的产生对收益率毫无预测性的 n 个因子，然后根据它们的表现选出其中最好的 k 个，再把和 k 个因子配置在一起，考察它们在样本内上述 k 个因子构成的投资组合收益率的 t-statistic 到底能有多高（由于这些随机因子毫无预测

性，因此 null hypothesis 是它们的预期收益率为零；评价标准为投资组合收益率 t-statistic 经验分布的 95% 分位数阈值）。

下图给出了实证结果。从中不难观察到以下三点：（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；这是因为当 k 逐渐增大时，使用更多的因子可以降低组合的波动率、提升 t-statistic 的阈值；一旦 k 超过最优值，越来越多排名靠后的因子被选入，降低组合的收益率以及 t-statistic 阈值。



为了在实证研究中发现样本内更好的策略或者更显著的因子 —— 无论是为了讨好基金经理还是为了在顶刊上发文 —— multiple testing 的不正之风早已席卷了学术界和业界。

Harvey, Liu, and Zhu (2016) 研究了学术界发表的 316 个选股因子。他们通过考虑不同因子之间相关性提出了一个全新检验框架。该方法可以排除 multiple testing 的影响。该研究表明，只有在 single testing 中 t-statistic 超过 3.0（而非人们传统认为的 5% 的显著性水平对应的 2.0）的因子才有可能在排除了 multiple testing 的影响之后，而非来自运气。不过，Harvey 同时也指出，3.0 其实都是非常保守的。

我们自己在回测时应时刻谨记 multiple testing 的影响；此外，在学习别人的发现时也要保持着一颗怀疑之心，因为没有多少人告诉我们，在 TA 提出的这个样本内显著因子之前有过多少次失败的尝试。

3

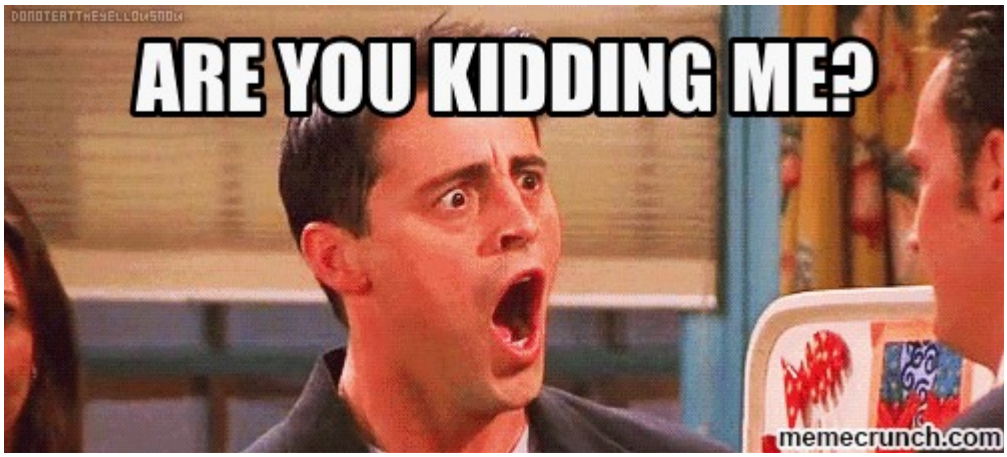
样本选择和数据

Protocol 的第三部分是**样本选择和数据（Sample Choice and Data）**。它的核心要素包括：**（1）回测前就要确定回测区间，而非事后调整；（2）确保数据质量；（3）小心处理异常值（outliers）—— 不要凡事都想当然；（4）认真记录进行的数据变形处理。**

所有的这些努力其实都是为了**避免 p-hacking**。

Harvey 教授在介绍 Arnott, Harvey, and Markowitz (2019) 这篇文章的短片中讲了一个故事。一个量化研究员给他展示了一个股票策略，该策略在回测期内的表现非常好；只不过该回测有一个致命的问题：它的回测窗口不包含 2008 年的金融危机。当 Harvey 教授问他为什么排除这段时期，得到了令人无语的答复："因为策略在这段时间内失效了"。

Excuse Me???



这就是先看结果再调整回测区间，妥妥的 p-hacking 反例。

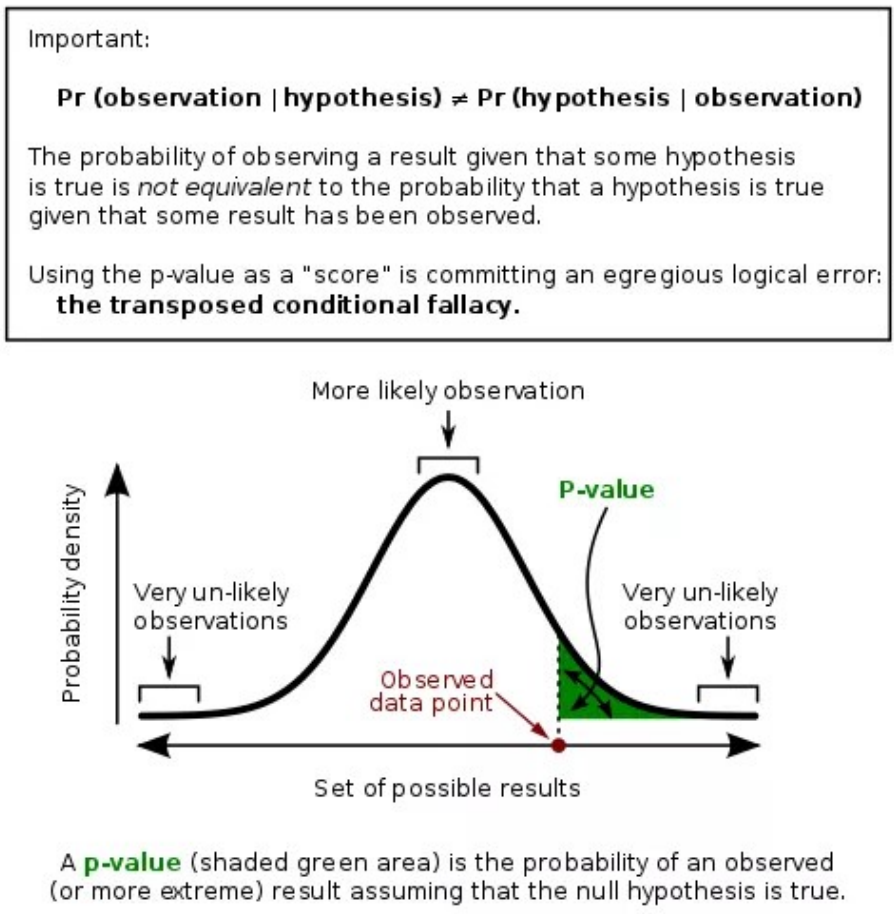
法国哲学家孔德将科学分成不同的等级（Comte 1856）。像数学、物理这类“硬科学”位于等级的上方，而社会学、经济学这些“软科学”位于等级的下方。“硬”和“软”本身并无“好”与“坏”之分。

硬科学可以从数据可以直接得到结论、无需任何人工解释，且结论是高度可归纳的。比如数学上的四色问题，一旦证明成立那就是成立；又如物理上的引力波，一旦发现那就是说明它的存在，这些都是确切的。**反观软科学，研究成果依赖于提出怎样的假设，如何处理数据，以及如何分析、解释结果，总之“事在人为”。**金融学是软科学，很多实证分析结果都会因人而异。

比如在股票研究中“使用过去 50 年的数据还是过去 30 年的数据？”“使用美股还是其他国家的股票？”“使用日收益率还是周收益率？”“使用百分比收益率还是对数收益率？”“是否以及如何剔除异常值？”“使用 OLS 还是 GLS？”……这些看似自然的选择背后其实都以追求样本内更显著的 p-value 为动机，一切阻碍获得超低 p-value 的数据都会被巧妙的避开。**这种为了获得超低 p-value 而在研究中刻意选取的数据处理方法就是 p-hacking。**

人们对于 p-hacking 的狂热源于对 p-value 的错误解读。

在统计学中，如果 H0 和 H1 分别表示 null hypothesis 和 alternative hypothesis，则 p-value = prob(D|H0)，即在 H0 成立下观测到数据 D 的概率。**从该定义出发，p-value 不代表原假设或者备择假设是否为真实的，即 p-value ≠ prob(H0|D) 以及 p-value ≠ prob(H1|D)。**



A **p-value** (shaded green area) is the probability of an observed (or more extreme) result assuming that the null hypothesis is true.

在检验一个策略或者因子是否有显著收益时，我们需要的是 prob(H0|D)，即在观察到 D 的条件下，原假设为真的概率是多少。这个问题仅依靠 p-value 自身无法回答的。为此，Harvey (2017) 提出了一个基于贝叶斯的框架，它可以正确求解我们关注的问题。

关于 p-hacking 和上述贝叶斯框架，《在追逐 p-value 的道路上狂奔，却在科学的道路上渐行渐远》一文曾有过非常详细的论述，在此不再赘述。

4 交叉验证

回测规程的第四部分是**交叉验证（Cross-Validation）**，这部分包括以下两个要素：

1. Out of Sample is Not Really Out of Sample；

2. Iterated Out of Sample is Not Out of Sample。

前文《所有数据都是样本内》曾对上述两点分别做过详细阐述，本文就不在重复之前的内容。这一条想要强调的是：**由于历史数据都是已经发生过的，它们都是样本内数据，因此必须小心解读交叉验证的结果，即便通过了交叉验证，也不能无脑的相信完全排除了过拟合的问题。**

关于更合理的使用交叉验证，Bailey et al.（2017）的研究成果值得借鉴。他们提出了一个**Combinatorially-Symmetric Cross-Validation（组合对称交叉验证，简称 CSCV）**方法，它可以定量的计算样本内过拟合的概率。《美丽的回测 —— 教你定量计算过拟合概率》一文详细的介绍了该方法。它的优势在于：

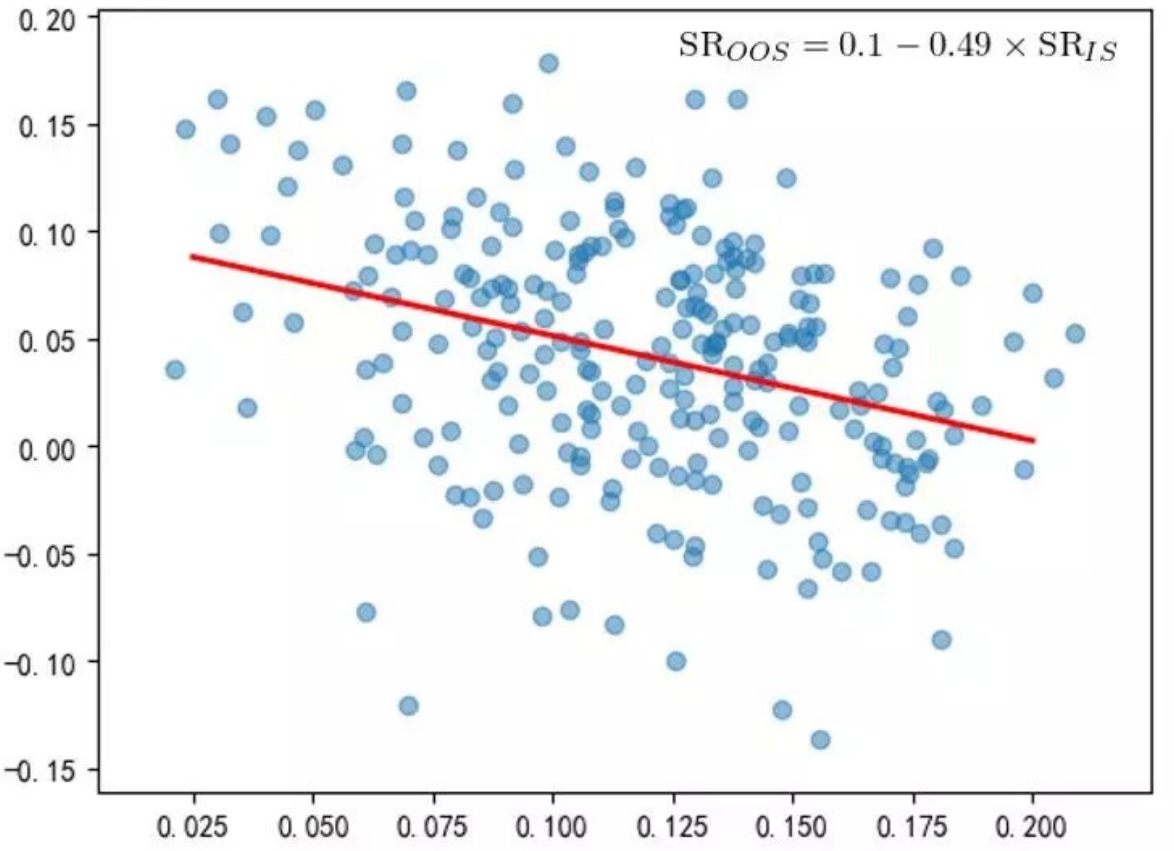
1. 保证了训练集和测试集同样大小，使得样本内外的夏普率具有可比性；

2. 保证了训练集和测试集的数据是对称的，因此夏普率在样本外的降低只可能来自过拟合；

3. 保留了收益率序列的时序相关性；

4. 利用 Bootstrap 理念求解过拟合的概率，不需要对过拟合的随机模型或者参数做任何假设。

举个例子。按照 CSCV 方法，下图描述了某趋势追踪策略在不同参数下，其样本内夏普率（SR_IS）和同参数在样本外夏普率（SR_OOS）的负相关关系，意味着验本内效果越好对应着样本外表现越差。该策略的样本内过拟合概率高达 0.572。一个真正有效的策略在样本内的过拟合概率不应如此之高。



无论从独立性还是可交易特征而言，交易数据其实都十分匮乏。它们对传统的交叉验证造成了极大的挑战，在使用机器学习时应牢记这一点，理性看待交叉验证结果。

5

模型动力学

模型动力学（Model Dynamics）是回测规程的第五部分，**它关注的是量化策略在样本外的表现逐渐变差的问题。**而这背后可能存在两个原因：**（1）市场结构发生变化导致策略失效，比如越来越多的人开始使用某个策略或者因子，使得它变得拥挤。（2）策略使用者自身的行为偏差导致一个好模型最终沦为一个失效模型。**

我在之前的文章中多次表达过一个观点：**任何策略能赚钱都是利用了市场的某种非有效性；一旦使用该策略的人越来越多，市场在这方面就变得更加有效，从而削弱策略的盈利能力。**

在技术分析领域，上述观点的最好例证之一是布林带（Bollinger bands）。毫无疑问，布林带是几十年前最盛行、最管用的技术分析策略之一。然而，人们越来越发现该方法挣钱的能力

越来越差。对此，Fang, Jacobsen, and Qin (2017) 针对全球十几个主要市场进行了实证分析。

他们的研究发现，1983 和 2001 这两个重要时间节点对于布林带的效果影响巨大。1983 年，John Bollinger首次在电视广播中介绍了布林带，使得这个之前神秘的方法开始走进大众视野。而 2001 年，John Bollinger 更是发表了Bollinger on Bollinger Bands 这本红极一时的技术流圣经；在随后的 4 年内，这本书被翻译成其他 12 种语言在全世界范围内迅速传播，这使得布林带一下变得家喻户晓。Fang, Jacobsen, and Qin (2017) 发现，布林带的流行和普及（特别是 2001 年之后）直接造成了该策略的失效。

这样的例子在股票因子投资中也不胜枚举。**一个新因子被提出后，随着越来越多人使用，它在 post-publication 样本外的效果势必会打折扣。**McLean and Pontiff (2016) 研究了 97 个因子在被发表之后的表现，发现因子的收益率比论文中的 in-sample 降低 50% 以上。

有时，策略并没有变得拥挤，但它在样本外还是持续变差。这背后的另一个原因是使用者的非理性行为偏差。

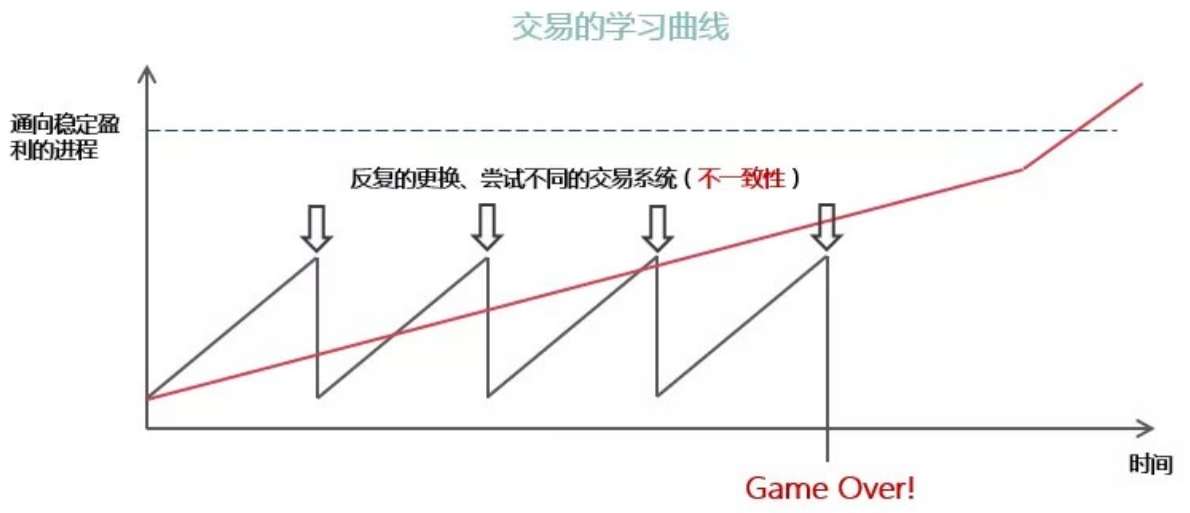
任何一个策略或者交易系统，都是基于对市场的某个假设。然而市场充满着不确定性，因此它必然会在一些时候背离这个假设，这时该交易系统就会出现亏损。一个优秀的交易系统是一个长期来看能够盈利的系统，而非一个能够每笔交易都赚钱的系统。

随着交易的进行，由于小数定律造成的偏误，很多人在几次亏损后就开始“怀疑人生”了，认为“this time is different”、开始要对策略动刀子。这种想法非常危险。如果你真的这么做的了，为了每一笔的亏损都对你的系统进行了修补，便走上了“处处精准过拟合”的快车道，策略最终将会对市场未来的变化无能为力。

> *Most traders take a good system and destroy it by trying to make it into a perfect system. –– Robert Prechter*

改造一个长期来看可以赚钱的优秀系统必须要非常小心。对哪怕是一个参数的哪怕是一丁点的调节都会改变该系统的效果。这么做是以改动后的系统对最新的交易数据表现更佳为前提；但是如果不能证明它在未来的样本外更有效，那么如此“改进”仍然是徒劳的。

量化投资背后的核心是单次优势 + 大数定律。这二者中大数定律又更加重要，它要求我们在交易中尽一切努力做到一致性。一般交易者的学习曲线如下面图中的黑色曲线：无法做到严格遵循一个交易系统，总是带着个人情感进行交易，将自己行为带来的不确定性错误地强加于系统的表现之上。这些交易者无法持之以恒，三天两头更换系统，最终输光本金。与之相反的，一个优秀的交易者会专注于一致性，这会让他在通往盈利的进程中越走越远，最终到达胜利的彼岸。



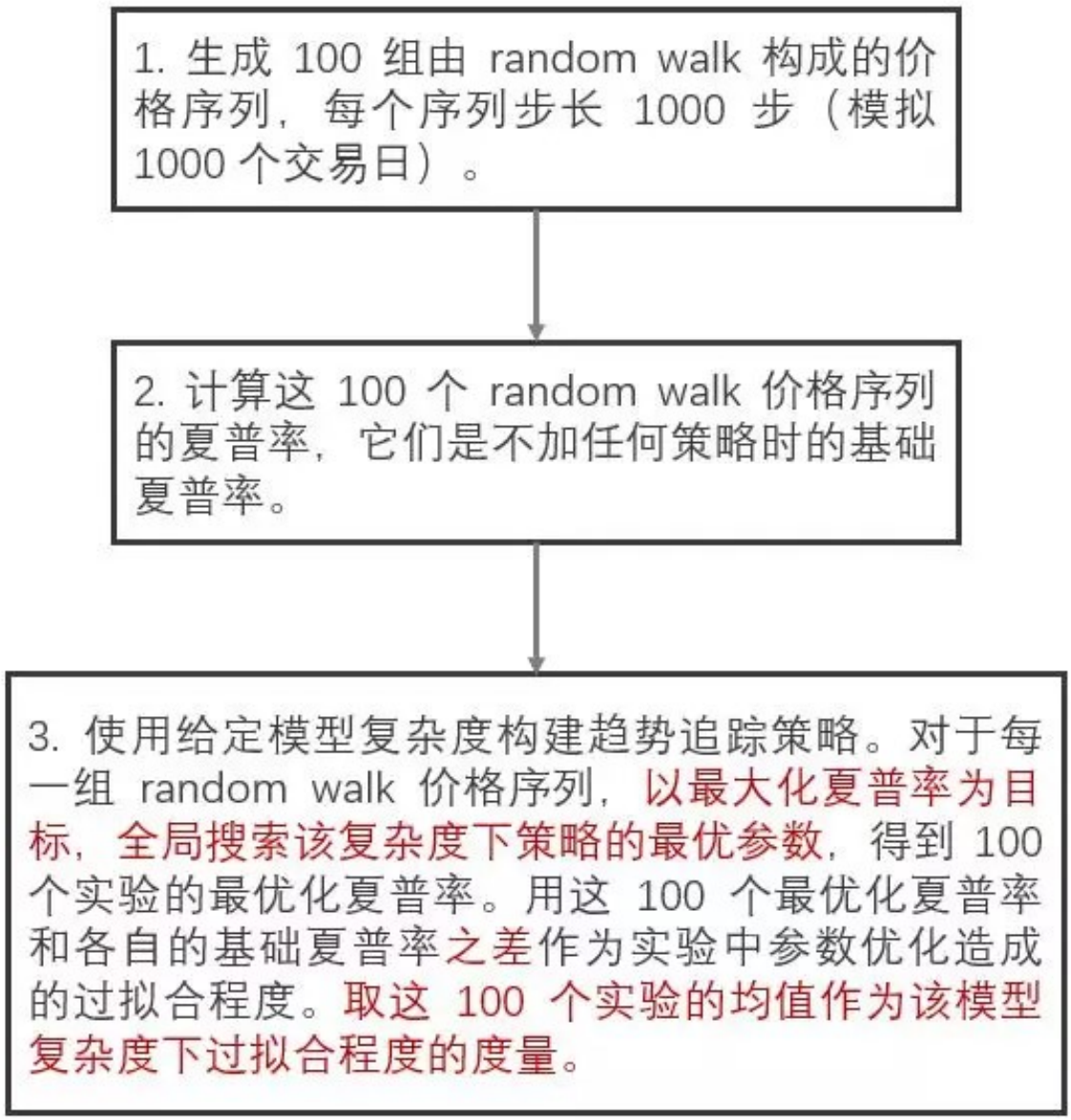
6 模型复杂度

回测规程的第六部分是**模型复杂度（Model Complexity），主张我们应该追求策略的而简单性和可解释性。**

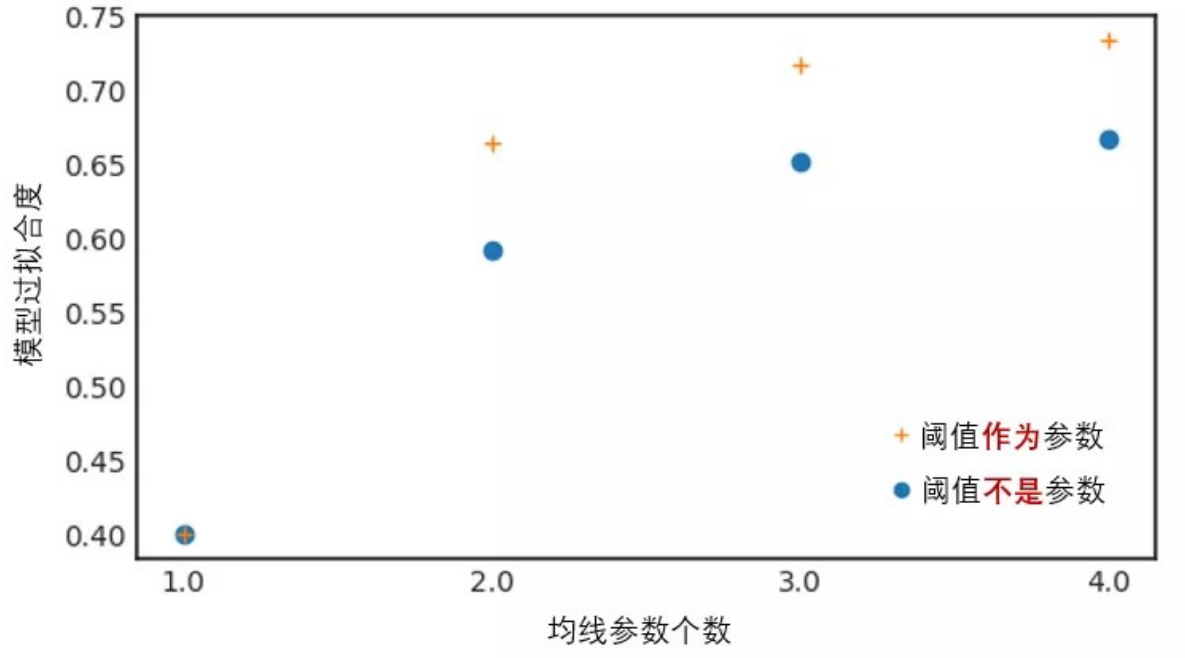
我们大概都有下面这样的经验：一个策略的夏普率不够亮眼，那么可以通过加入止盈、止损，中性化处理、甚至是对投资标的进行筛选来进一步提高其在样本内的表现。此外，对上面的每一个处理方法，我们似乎都能找到合理的解释和来自其他文献的理论和实证支持。在确认偏误下，我们非常愿意相信这些处理都是合理的、并没有引入过拟合。

任何通过增加参数维度来提高样本内的表现 —— 无论这些理由听上去多么合理 —— 都实打实的提高了模型的复杂度；更高的模型复杂度则更容易出现过拟合。

前文《模型复杂度随想》曾对上述观点做过一个简单实验。该文提出了如下图所示的流程来定量计算模型复杂度造成的过拟合程度。



考虑一个基于均线多头排序的简单多头趋势追踪策略。模型复杂度的两个维度是：（1）均线多头排序中用到的不同周期均线的个数；（2）这些均线秩相关系数的阈值（用来决定是否开仓、空仓）。使用纯随机游走产生的假想资产价格曲线，按不同复杂度构建趋势追踪策略。模型的过拟合度和复杂度之间的关系如下图所示，**说明模型过拟合度随模型复杂度递增。**

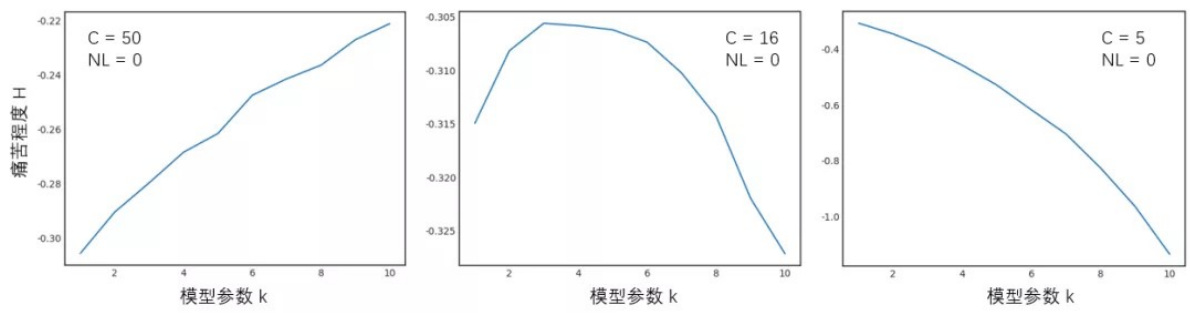


在第六方面，Arnott, Harvey, and Markowitz (2019) 倡导的第二点是追求可解释的机器学习（seek interpretable machine learning）。量化策略，尤其是使用了机器学习算法的量化策略不应该是黑箱。任何使用者都应该了解这个算法到底干了什么。最近几年，计算机领域的一个细分学科逐渐受到世人关注，它研究的对象是 interpretable classification 和 interpretable policy design（一个例子见 Wang et al. 2017）。相信在未来，可解释的机器学习在金融领域能够大有可为。

关于模型复杂度，我想补充一点 Arnott, Harvey, and Markowitz (2019) 没有的内容，同样来自《模型复杂度随想》，那就是**相较于简单的模型，复杂度更高的模型可能会在亏损时给人更痛苦的主观感受。**在这方面，我做了一些探索性的研究，指出了模型复杂度和实盘痛苦程度之间的非线性关系：

> **1.** 当模型复杂度逐渐提升时，由于它更好的捕捉了收益率和信号之间的（非线性）关系，这是能带来样本外效果的提升的，减少亏损的痛苦；
>
> **2.** 当模型过于复杂时，由于样本内过拟合可能性上升；模型复杂度会非线性的放大同等程度亏损（比如最大回撤）给人们造成的痛苦。

根据以上描述，模型复杂度和实盘的痛苦程度大概如下图所示（具体请看《模型复杂度随想》）。



在当下，我们越来越崇尚各种复杂的模型。以上探索仅仅希望提出一些思考：**我们在样本外是否 100% 做好了准备接受复杂模型？交易中存在各种认知偏差，如果我们连最简单的按一根均线做趋势追踪都无法坚决的执行，那又有什么来保证我们在面对实盘亏损时能够坚守复杂模型呢？如果我们不能坚守复杂模型，那么开发复杂模型所付出的心血和努力是否付之东流呢？**

7

研究文化

回测规程的最后一部分是**研究文化（Research Culture）**，它包括以下两点：

1. Establish a research culture that rewards quality;

2. Be careful with delegated research.

上面第一条说的是，**在开发量化策略或者因子时，比起追求样本内的惊艳效果，我们更应该看中研究的质量**，例如研究是否避免了各种偏差、尽最大努力的排除了过拟合、是否存在先验理论、是否足够独立等。**一个因子或指标，无论有用没有，只要能够被复现，都是有益的发现，都为帮助我们更好的理解市场起到了巨大贡献。**

在学术界，为了提升期刊的声望，编辑们都更倾向于录用低 p-value 的文章；为了在更高水平的期刊上发文，学者们更倾向于找到低 p-value 的因子。在美国绝大多数学校里，如果能在 Journal of Finance 发表一篇文章，一个教授就有可能得到终身教职。这一环扣一环的错误关系导致了严重的 publication bias，我们被大量依靠样本内 data mining 和 p-hacking 获得的虚假因子蒙蔽了双眼，而高研究质量却低显著性的因子在顶级期刊上则难有容身之处。

这部分的第二条说的是，很多时候由于基金经理的精力有限，无法亲力亲为研究每个策略。因此会把研究分发给不同的研究员。**研究员应该保持独立性、进行高质量的研究，而不是通过寻找虚假的显著性来取悦基金经理。**

任何策略都最终会失效，而客观、严谨的研究文化才是能够源远流长的，才是我们应该努力追寻的。

在美国，要论业界的"学术天团"，一般人大概首先会想到 AQR。然而，还有个更老牌、更大牌的管理人，它就是 Dimensional Fund Advisors L.P.，它的 Directors 中不乏 Eugene Fama、Ken French、Myron Scholes 这些赫赫有名的学者。在 Dimensional 的官网上记录着 Ken French 下面这句话，一语道破了研究文化的真谛 —— **任何时候我们都要努力探寻真谛、做对的事情。**

8 结语

好了，上面七小节介绍了回测规程中的七方面内容。接下来我们可以"召唤神龙"了。下图给出了 Arnott, Harvey, and Markowitz (2019) 自己总结的七方面，每一个 bullet point 都值得好好体会。

EXHIBIT 2

Seven-Point Protocol for Research in Quantitative Finance

1. Research Motivation
- a. Does the model have a solid economic foundation?
- b. Did the economic foundation or hypothesis exist before the research was conducted?

2. Multiple Testing and Statistical Methods
- a. Did the researcher keep track of all models and variables that were tried (both successful and unsuccessful), and are the researchers aware of the multiple-testing issue?
- b. Is there a full accounting of all possible interaction variables if interaction variables are used?
- c. Did the researchers investigate all variables set out in the research agenda, or did they cut the research as soon as they found a good model?

3. Data and Sample Choice
- a. Do the data chosen for examination make sense? And, if other data are available, is it reasonable to exclude these data?
- b. Did the researchers take steps to ensure the integrity of the data?
- c. Do the data transformations, such as scaling, make sense? Were they selected in advance? And are the results robust to minor changes in these transformations?
- d. If outliers are excluded, are the exclusion rules reasonable?
- e. If the data are winsorized, was there a good reason to do it? Was the winsorization rule chosen before the research was started? Was only one winsorization rule tried (as opposed to many)?

4. Cross-Validation
- a. Are the researchers aware that true out-of-sample tests are only possible in live trading?
- b. Are steps in place to eliminate the risk of out-of-sample iterations (i.e., an in-sample model that is later modified to fit out-of-sample data)?
- c. Is the out-of-sample analysis representative of live trading? For example, are trading costs and data revisions taken into account?

5. Model Dynamics
- a. Is the model resilient to structural change, and have the researchers taken steps to minimize the overfitting of the model dynamics?
- b. Does the analysis take into account the risk/likelihood of overcrowding in live trading?
- c. Do researchers take steps to minimize the tweaking of a live model?

6. Complexity
- a. Is the model designed to minimize the curse of dimensionality?
- b. Have the researchers taken steps to produce the simplest practicable model specification?
- c. Has an attempt been made to interpret the predictions of the machine learning model rather than using it as a black box?

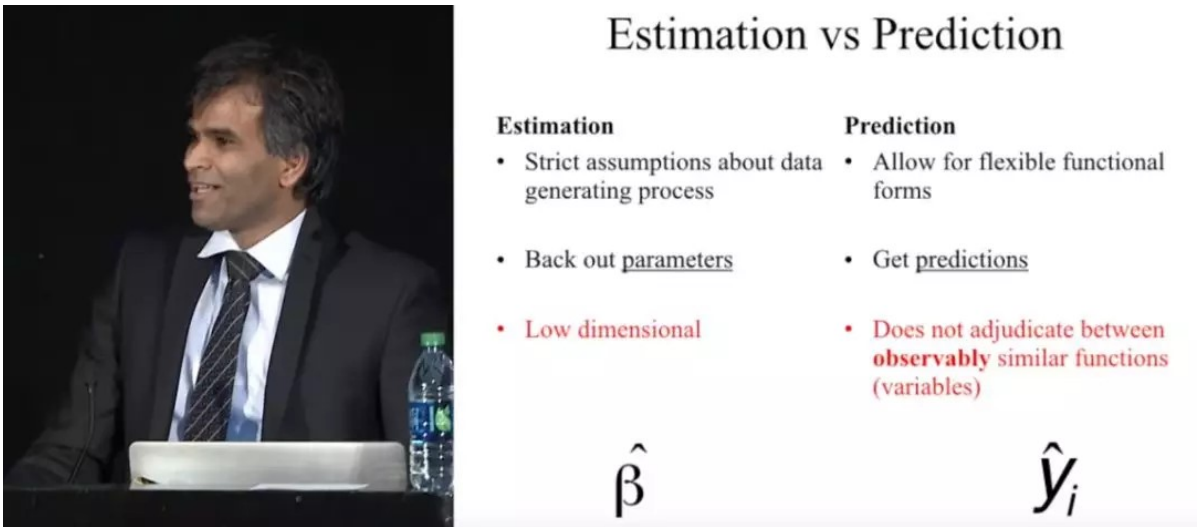
7. Research Culture
- a. Does the research culture reward the quality of the science rather than the finding of a winning strategy?
- b. Do the researchers and management understand that most tests will fail?
- c. Are expectations clear (that researchers should seek the truth, not just something that works) when research is delegated?

出处：Arnott, Harvey, and Markowitz (2019)

最后想强调的是，Arnott, Harvey, and Markowitz (2019) 并不是为了否定机器学习在投资中越来越重要的作用。恰恰相反的是，他们提出这个框架就是为了让我们更好的享受机器学习的成果。

对投资来说，我们最关心的是 prediction 是否准确，而非参数的 adjudication。它的意思是只要能提高样本外的预测性，我们可以牺牲参数估计的准确性。公允的说，从探寻市场真谛的角度来说，我们当然关心 β 的估计是否准确；然而，从投资实际效果的角度来看，我们更应关注样本外 y 预测值是否靠谱。

预测的目标是最小化 loss function；而传统计量经济学中 estimation 的目标是参数的 unbiasedness。参数估计准了不一定意味着样本外的预测性一定更好。关于这方面的论述，我推荐各位看看 Sendhil Mullainathan 教授在 AFA Lecture 上做的 Machine Learning and Prediction in Economics and Finance 主题演讲。



客观的说，由于金融数据的一些特殊性（非结构化、高维度、稀疏、信噪比低等），传统计量经济学在很多时候确实难有作为，而机器学习算法则更有前景。关于这点，Lopez de Prado 做过一篇题为 《The 7 Reasons Most Econometric Investments Fail》的报告。[量化投资与机器学习] 公众号曾对这篇报告进行过解读（见《AQR最最最新｜计量经济学应用投资失败的7个原因》），感兴趣的朋友不妨看一看。当然，这并不意味着我们就应该轻易摒弃计量经济学模型、毫无顾忌的投身到机器学习的怀抱。

> *It is naïve to think we no longer need economic models in the era of machine learning. Given that the quantity and quality of data is relatively limited in finance, machine learning applications face many of the same issues quantitative finance researchers have struggled with for decades.*

本文介绍的回测规程乍一看完虽然没有太多惊艳之处，但它却能产生非常积极的效果。正如飞机驾驶舱里面的 checklist 能极大的提升飞行安全一样，在回测中牢记并遵守这些准则可以有效降低过拟合的风险、避开噪音、找到真正在样本外可持续的因果关系，获取更高的收益。

参考文献

Arnott, R., C. R. Harvey, and H. Markowitz (2019). A backtesting protocol in the era of machine learning. *Journal of Financial Data Science*, Vol. 1(1), 64 – 74.

Bailey, D. H. and M. Lopez de Prado (2012). The Sharpe ratio efficient frontier. *Journal of Risk*, Vol. 15(2), 3 – 44.

Bailey, D. H. and M. Lopez de Prado (2014). The deflated Sharpe ratio: correcting for selection bias, backtest overfitting, and non-Normality. *The Journal of Portfolio Management*, Vol. 40(5), 94 – 107.

Bailey, D. H., J. M. Borwein, M. Lopez de Prado, and Q. J. Zhu (2017). The probability of backtest overfitting. *Journal of Computational Finance*, Vol. 20(4), 39 – 69.

Chordia, T., A. Goyal, and A. Saretto (2017). p-Hacking: evidence from two million trading strategies. Swiss Finance Institute Research Paper No. 17-37, SSRN.

Comte (1856). The Positive Philosophy of Auguste Comte, translated by Harriett Marineau (Calvin Blanchard, New York). Vol. II.

Fang, J., B. Jacobsen, and Y. Qin (2017). Popularity versus profitability: evidence from Bollinger bands. *The Journal of Portfolio Management*, Vol. 43(4), 152 – 159.

Harvey, C. R. (2017). Presidential address: The scientific outlook in financial economics. *The Journal of Finance*, Vol. 72(4), 1399 – 1440.

Harvey, C. R., Y. Liu, and H. Zhu (2016). … and the cross-section of expected returns. *Review of Financial Studies*, Vol. 29(1), 5 – 68.

Lopez de Prado, M. (2018). *Advances in financial machine learning*. Hoboken, NJ: John Wiley & Sons.

McLean, R.D. and J. Pontiff (2016). Does academic research destroy stock return predictability? *The Journal of Finance*, Vol. 71(1), 5 – 32.

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

Wang, T., C. Rudin, F. Doshi-Velez, Y. Liu, E. Klampfl, and P. MacNeille (2017). A Bayesian framework for learning rule sets for interpretable classification. *Journal of Machine Learning Research*, Vol. 18, 1 – 37.