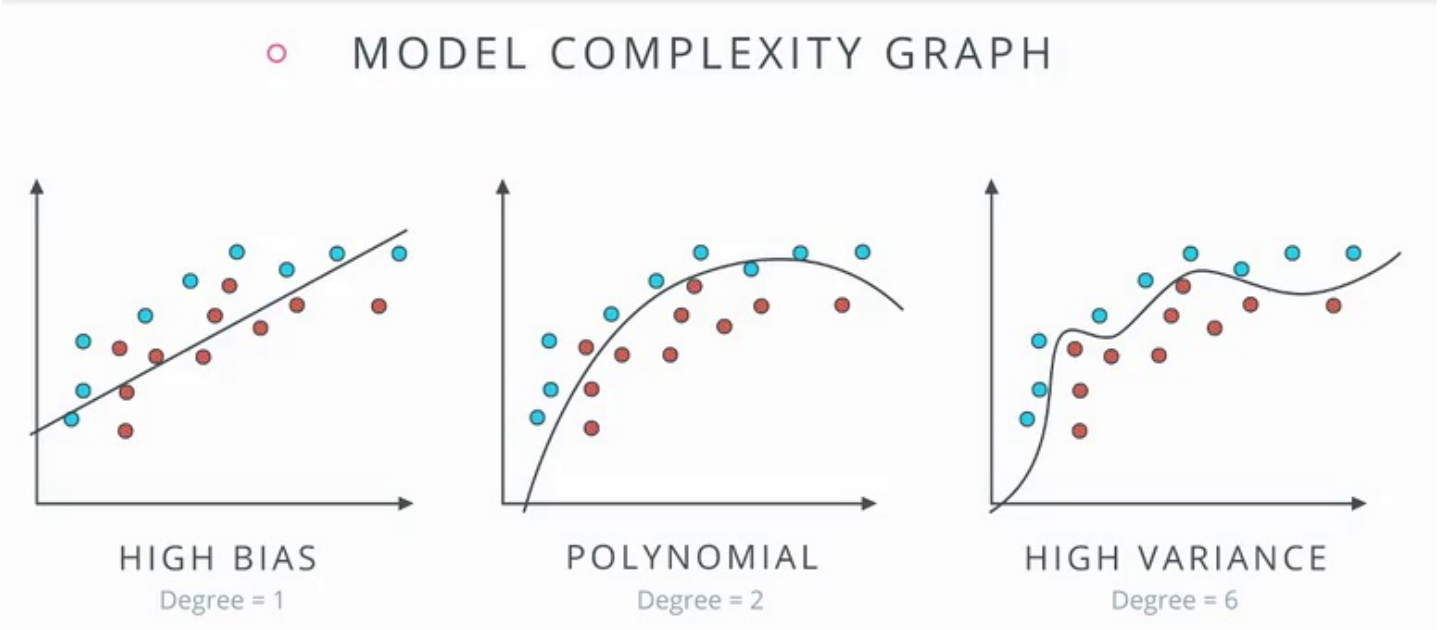
知乎 首发于 川流不息



模型复杂度随想

石川
量化交易 话题的优秀回答者

已关注

王坚等 30 人赞同了该文章

摘要

当模型复杂度一样时，人们偏好风险收益特性更高的策略；当风险收益特性一样时，人们偏好模型复杂度更低的策略。各种复杂模型带来的边际超额收益能否 justify 它们的复杂度呢？拭目以待。

1 引言

一个初入量化投资的分析师经过了一个月的奋斗开发出了一个双均线趋势追踪模型后，兴冲冲的跑来和他的基金经理汇报，于是便有了下面这段对话。

- **分析师（一脸兴奋）**：我开发出了一个双均线系统，绝对没有数据挖掘，只有计算均线的两个参数，该参数对绝大多数商品期货都有效、适应性极强。
- **基金经理**：做趋势追踪还有其他的方法，比如时间序列分析、其他技术分析手段、以及机器学习里面的各种复杂算法。你的系统和这些比较过吗？
- **分析师（心说“一猜你就会问这个”）**：常见的这些方法我都仔细试过了，它们的效果都没有双均线系统好。
- **基金经理**：……
- **分析师**：真的！我做了非常详细的对比，逐笔分析了各种不同策略的交易记录，双均线是最
的。
- **基金经理：从你排除其他策略、挑出双均线系统的那一刻，你就已经过拟合了。**

知乎 首发于 川流不息

上面这段对话当然是我杜撰的。想通过它表明的观点是，**我们将不同的量化技术应用到同样的数据上构建某一类（比如趋势追踪、反转、套利）策略时，最终会挑出来表现最好的量化技术，无论这个技术复杂与否（线性的、非线性的），这个过程本身就是在过拟合。**

最终被挑出来的，注定是因为在样本内战胜了其他的。从“超参数”（见《科学回测中的大学问》）的意义上说，这个模型难逃 data mining 之嫌，因为**它比别的模型更好很可能是因为它对样本数据内的噪音刻画的更精准，而非发现了一些被其他策略忽视到的真实存在于数据之间的因果关系。**

以上这点粗浅的认识当然不是鼓励大家放弃回测中表现好的、使用表现差的量化技术。就我自己有限的经验来看，任何策略都或多或少存在数据挖掘的问题，而这个问题随着模型复杂度的增加更加突出。

今天就简单聊聊模型复杂度。讨论主要从以下两个角度展开：

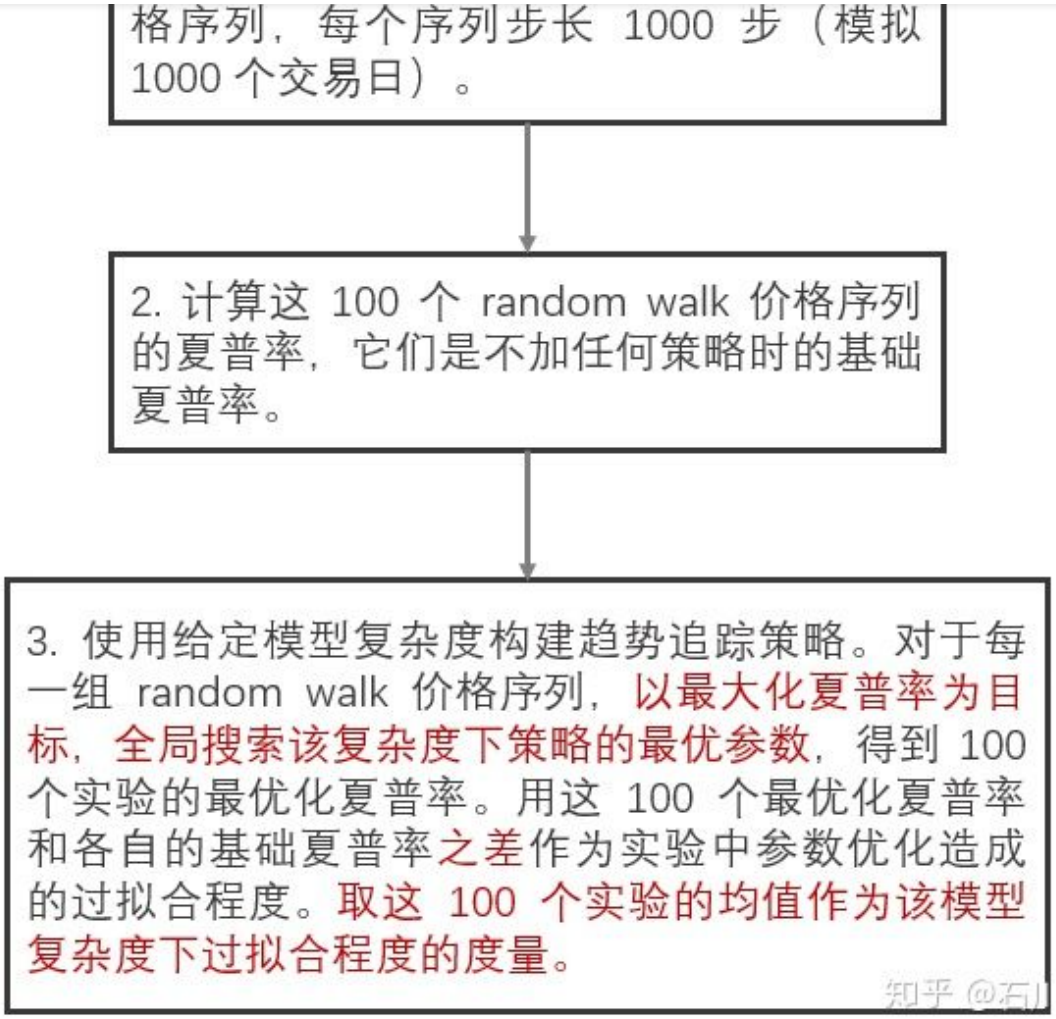
1. **模型复杂度和过拟合程度**：定量分析模型复杂度和构建策略时 data mining 的程度。
2. **模型复杂度和损失带来的主观感受**：回答诸如“面对实盘中同等大小 —— 比如 -10% ——的回撤，不同复杂度的模型是否能给我们带来同样的主观感受”这样的问题。

这两个角度的研究都是很大的课题，本文仅仅是做一点抛砖引玉的探讨。

2 模型复杂度和过拟合程度

在构建一个量化投资策略时，一旦确定了模型复杂度，就要进行参数优化。只要是参数优化，无论再怎么小心，都会存在过拟合。本节使用趋势策略阐述**在给定的模型复杂度进行参数优化**和**过拟合程度**之间的关系。分析流程如下：

知乎 首发于 川流不息



分析中采用的趋势追踪策略是均线多头排列策略。它的定义和模型复杂度介绍如下。

在市场有大趋势的时候，均线一般呈现多头或者空头结构，即不同周期 T 的均线排序和 T 的排序非常一致（比如上涨时，通常有 MA5 > MA15 > MA30）。当投资品从上涨向下跌转换、或由下跌向上涨转换时，短周期均线会先于长周期均线发生变化。在前者发生时，短周期均线开始逐步下穿长周期均线；在后者发生时，短周期均线开始逐步上穿长周期均线。在发生由涨转跌或由跌转涨时，不同周期均线的排序和时间窗口 T 大小的排序关系被打乱，不再完全一致。

使用秩相关系数计算均线排序和时间窗口 T 排序之间的一致性，并使用它择时、构建趋势追踪策略（这里只考虑多头策略）。当均线多头排列时，均线和 T 之间的秩相关性为 1；当均线空头排列时，均线和 T 之间的秩相关性为 -1。由涨转跌时，短期均线开始下穿，秩相关性从 1 开始下降；由跌转涨时，短期均线开始上穿，秩相关性从 -1 开始上升。由此，可以构建策略如下：

使用给定的均线参数周期，各自计算指数平均，进而计算均线排序和参数排序的秩相关系数。空仓时，如果秩相关系数上穿 -TH 则满仓；满仓时，如果秩相关系数下穿 TH 则空仓。不考虑任何成本。

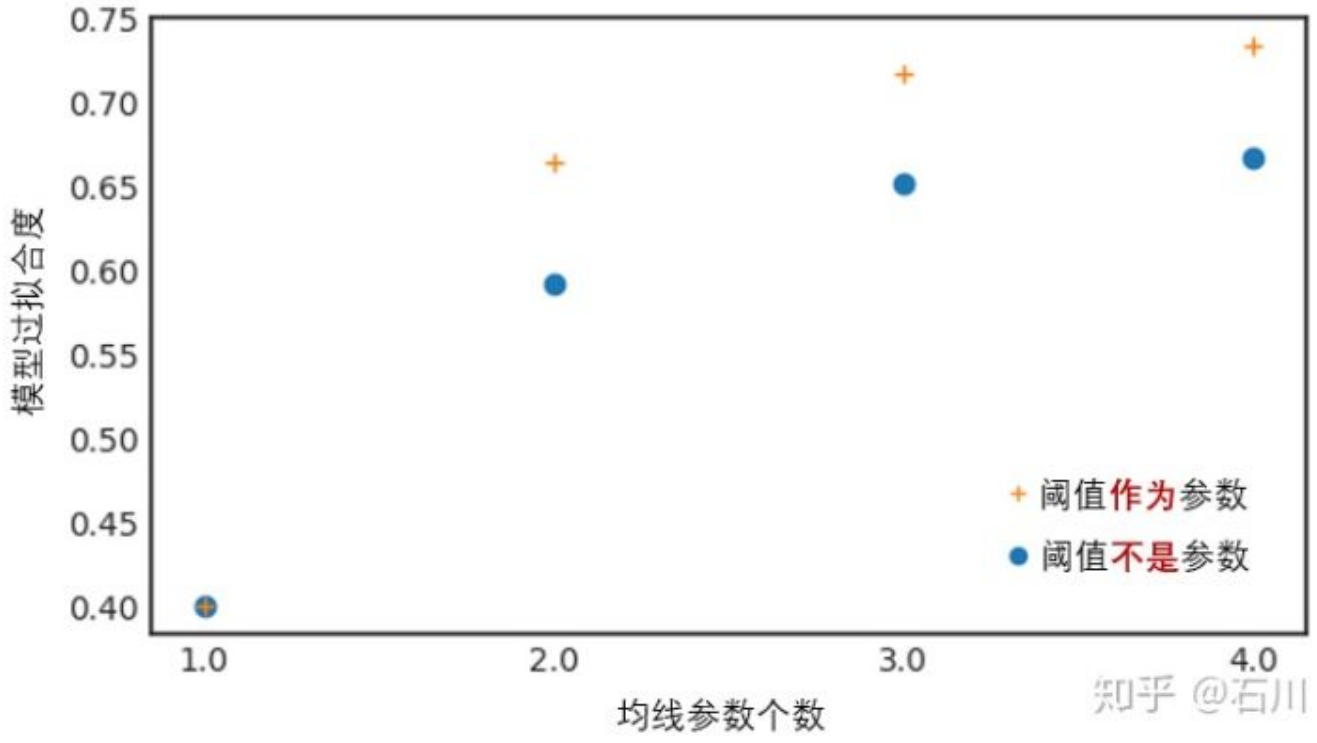
在这个策略中，模型复杂度由如下两组参数刻画：

知乎 首发于 川流不息

这两组参数各自从不同层面增加了模型的复杂程度。在分析中，它们的取值如下：

1. 均线参数的个数从 2 到 5 递增，依次增加模型的复杂度。第一个均线周期取值范围是 10 到 100，步长 10；从第二个均线周期开始，在搜索参数时，其取值范围似是前一个均线的取值与 100 之间，步长 10。此外，允许新加入的均线对策略不产生作用。这保证了随着均线个数增加，求解的空间是递增的，从而保证了最优目标函数的单调性。
2. 在分析时，首先仅考虑均线参数个数造成的影响，因此假设阈值为 TH = 0.5 恒定。之后，为了同时考察阈值对过拟合程度的影响，允许阈值 TH 从 0.1 到 0.9 之间（步长 0.1）选择。

依照上述描述进行实验，得到的模型复杂度和过拟合程度的关系如下图所示。其中蓝色圆圈表示仅考虑均线参数个数这一种模型复杂度时的情况，而黄色十字表示同时考虑阈值作为模型复杂度的情况。



当我们使用真实的交易数据进行策略的参数优化时，尽管使用了训练集和测试集、考虑了参数平原、从各种业务层面解释了参数的选择，依然无法消除参数优化中过拟合的影响。**更不幸的是，对于真实交易数据，由于不知道它其中哪些是因果关系、哪些是噪音，因此我们甚至无法评价参数优化造成的过拟合程度。**

然而在上述实验中，由于价格序列由随机游走生成，因此随着实验个数的增加，我们预期它们的基础夏普率均值是 0。这正是使用 random walks 来验证策略的好处，因为它的“正确答案”是已知的 —— **一个不存在过拟合的策略在随机游走价格序列上不应该能持续的赚到钱。**

知乎 首发于 川流不息

模型的过拟合程度（100 个策略夏普率均值于基础家谱率均值之差）也在上升；而随着模型复杂度从多维度的提升（即加入阈值参数），模型的过拟合程度产生了跳变。

上述结果说明模型的过拟合程度随模型的复杂度递增。

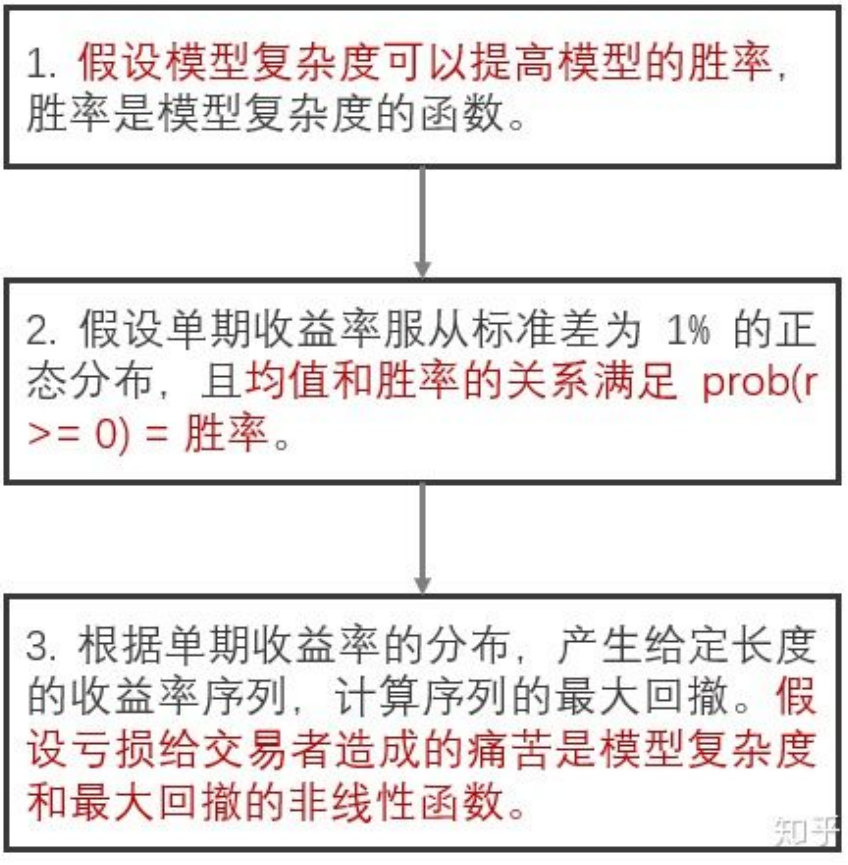
3 模型复杂度和损失带来的主观感受

本节来看看模型复杂度和策略损失带来的主观感受之间的关系。

《追求卓越，但接受交易中的不完美》一文曾阐述了如下观点：一个策略投放到实盘时最大的敌人是交易者的心理关。**这个心理关指的是交易者能否克服实盘中的心理压力从而坚持使用这个策略。**对于任何一个量化投资策略，几乎可以确定的是它在回测中的表现是其在实盘中表现的上限。在实际交易中，价格时刻在波动，充斥着噪音的各路消息以远超过我们能够接受的速度涌来，使人快步踏入行为金融学中的各种认知偏差陷阱、丧失冷静；**面对真金白银的亏损，交易者会比想象的更脆弱、更容易怀疑策略的开发中是否存在没有考虑到的问题（对于复杂策略更是如此）、自我动摇想要放弃这个系统 —— 这就是损失带来的主观感受。**

当一个策略持续出现回撤，亏损超过回测中最大回撤时，复杂度是否对亏损带给我们的痛苦程度（以及对策略不自信的程度）造成影响呢？

为了回答这个问题，自然要建模。建模的流程如下图所示。



知乎 首发于 川流不息

假设模型复杂度和胜率的关系如下：

$$w = w_0 + 0.003k + 0.01\mathrm{NL}$$

其中 w_0 是基础胜率（假设等于 0.5），k 代表模型中参数的个数，NL 为 binary 变量，取值 0 或者 1，代表模型是否为非线性的（NL = 1 表示非线性）。Disclaimer：本模型没有任何 reference，只是我为了得到量化分析结果选用的一个简单模型。

假设单期收益率满足标准差为 1% 的正态分布，均值则和胜率有关。胜率代表着单期收益率大于等于零的概率，因此我们必须选择均值以满足 $\mathrm{prob}(r \geq 0) = w$ 。根据这个关系，可以求出均值为：

$$\mu = -\mathrm{ISF}(w) \times \sigma$$

其中 ISF 表示标准正态分布的 inverse survival function。

得到单期收益率的分布之后，就可以构建任意长度的收益率序列。分析中，我们构建长度为 1000 的序列，以此作为该复杂度下假想策略的收益曲率序列的一个实现，并计算出它的 NAV。有了 NAV 就可以计算出它的最大回撤（max drawdown，MDD）。假设亏损造成的痛苦（记为 H ）和最大回撤以及模型复杂度的关系如下：

$$H = e^{(k-1+10\mathrm{NL})/C} \times \mathrm{MDD}$$

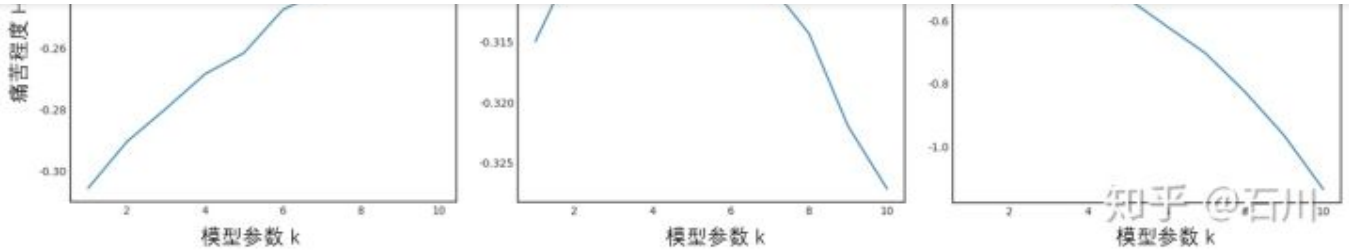
上述模型（disclaimer：同样没有任何 reference）说明 H 由两部分组成：模型复杂度和最大回撤。由该模型的表达式可知，在同样的最大回撤下，不同的模型复杂度给人的主观感受是不一样的，**模型复杂度非线性的放大了亏损造成的痛苦**。当 $k = 1$ （模型至少有一个参数）且模型为非线性（NL = 0）时，H 的第一项为 1，因此它仅由最大回撤决定。当模型复杂度上升时，对复杂模型的惩罚程度由参数 C （非负实数）控制。C 越小说明对模型复杂度的惩罚越高（即复杂模型会显著放大最大回撤造成的痛苦程度）。

结合上述胜率和痛苦程度的模型可知，模型复杂度可以增加胜率（hopefully），但它是以提高亏损造成的主观痛苦为代价的。因此，在这二者之间存在一个平衡。

下面来看一些实验结果。对于每一个给定的模型复杂度，随机产生 2000 个长度各为 1000 的收益率序列，并计算它们的最大回撤以及痛苦程度 H ，取这 2000 个实验的均值作为该模型复杂度下损失造成的痛苦程度的度量。

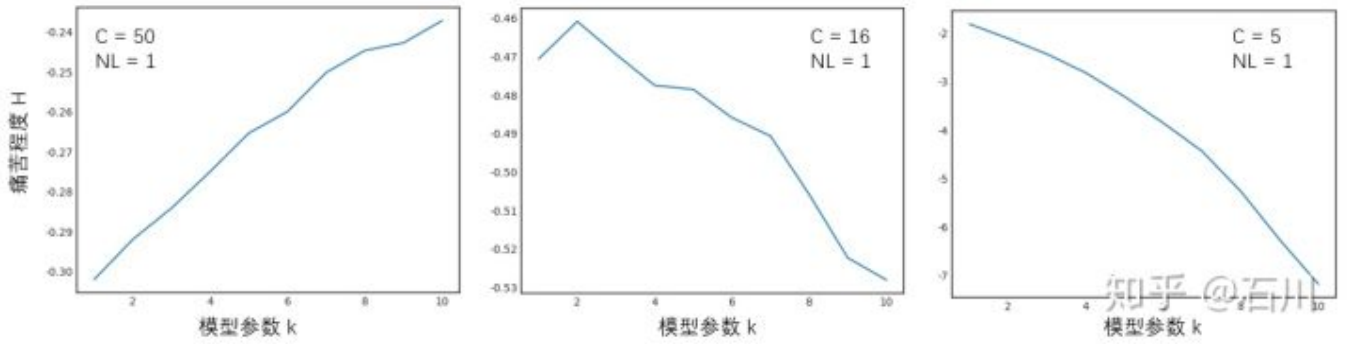
首先考虑线性模型，即 NL = 0 的情况。下面三张图分别显示了 C 取不同数值时，参数个数和 H 的关系：

知乎 首发于 川流不息



当 C 很大时，我们对模型复杂度的惩罚很低，模型复杂度的作用单边体现在提高胜率上。更高的胜率意味着更低的最大回撤，因此随着模型参数的增加，痛苦程度逐渐降低。当 C 很小时，情况正好相反。模型每增加一个参数，造成的痛苦程度非线性急速攀升，大大的抵消掉高胜率造成的低回撤的影响，痛苦程度随模型复杂度单调上升。当 C 取值中规中矩时，从上面中间的图中能够观察到胜率和痛苦程度之间的取舍，在理论上存在最佳的模型复杂度。

当 NL = 1 时，可以观察到和前面类似的结果（下图）。由于在 H 的建模中，我们对 NL 的惩罚较高（系数为 10），因此对于同样的 C 和 k，NL = 1 比 NL = 0 意味着更大的亏损痛苦。



上面的分析都是探索性的，并没有实证数据作为依据（难以找到使用不同模型复杂度策略的投资者并统计它们面对亏损时的不同感受）。我分析的初衷是，**在构建投资策略时，任何决定都要在得与失之间取舍。**复杂模型在提高胜率的同时，也一定在某种程度上有它的弊端。从我有限的经验来说，在实盘中出现同样程度的亏损时，复杂的模型比简单的模型更让人不安。

在当下，我们越来越崇尚各种复杂的模型。本小节仅仅希望从一个完全不同的角度来提出一些思考：**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？**

4 结语

前不久我听了 Vanguard 题为《先锋领航多资产 FOF 策略及外部管理人选聘概览》的报告。感最深的是当谈到对策略的看法时，先锋的观点是**策略的理念一定要简单 —— 能用一句话说清楚策略赚的什么钱，就不要用两句描述；策略的程序一定要可理解、完全透明。**

知乎 首发于 川流不息

借用老罗的一句话那就是：

> *Simplicity is the hidden complexity.*

- **当模型复杂度一样时，人们偏好风险收益特性更高的策略；**
- **当风险收益特性一样时，人们偏好模型复杂度更低的策略。**

各种复杂模型带来的边际超额收益能否 justify 它们的复杂度呢？拭目以待。

（全文完）

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

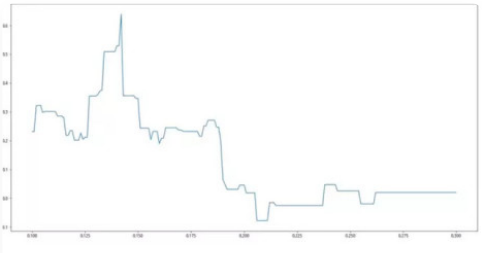
量化交易 交易心理 算法交易

▲ 赞同 30 ▼ 12 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



风险收益一致性择时模型

如果K线走势未来不可预知，那么交易凭什么获利?

想要洞察到投机交易的真相，至少需要2层突破。第一层突破：洞察到走势的不确定性，即未来走势是不可提前预知的。这一层其实很好理

全球十大

http://code
交易 原
易系 先
Ma 》
布的 系
Catscan D

知乎 首发于 川流不息

12 条评论 ⇌ 切换为时间排序

写下你的评论...

雪在烧 1 年前

我有一个策略EA， 基于 有涨有跌的基本哲学，无任何指标，加上 对冲，加码， 仓位管理等风险和利润控制，有兴趣吗?

赞

BugCreater 回复 雪在烧 1 年前

你要是真赚钱还要拉别人？别逗了

赞

雪在烧 回复 BugCreater 1 年前

我这是拉人吗？ 明显是想卖这个EA。

赞

卢维 1 年前

Vanguard那句话其实也有偏颇。有很多事情可以用一句话说得清楚，但绝大部分人听不懂。

赞

石川 (作者) 回复 卢维 1 年前

有道理……

赞

ALPHA16 回复 卢维 1 年前

但也不要忘了，真理通常不是一两句话就能说得清，特别是在一个充斥着谎言与谬误的市场——真理，仿佛秀才与兵。

赞

大亏货 1 年前

当个人水平一样时，越简单的模型研究出错率越低。但是当出错率一样时，就是越复杂的模型越好了。所以说到底还是水平问题。

赞

石川 (作者) 回复 大亏货 1 年前

赞

mockingbird 1 年前

无意中看见这篇文章，谢谢作者无私的分享，谈几点自己的感觉。

首先，看到‘偏好’，莫名的熟悉感一下子就忍俊不禁了，继而对‘当模型复杂度一样时，人们偏好风险收益特性更高的策略；当风险收益特性一样时，人们偏好模型复杂度更低的策略’得出第一想法，绝大多数这句话只能作为假设前提；

其次，复杂度这个问题，大家平时都会时不时的聊一下。如果复杂度用模型的参数、阈值数量来衡量，这些参数或者阈值对模型越是重要，其变动对模型的结果越是强烈，尤其是采用单一的风险收益指标去衡量，这种强烈的影响意味着参数、阈值敏感性较高，当然由参数、阈值取值带来的过拟合的情况会随着数量的增加而更加严重。

最后，除非一个模型能够解释那个视角下所有的价格表现，否则过拟合基本上是任何模型都不可避免的。事实上，过拟合作为一个中间结果，会对模型最终的交易结果产生影响，识别过拟合带来的影响，并对这些交易结果进行分析，再调整整个策略的应对措施是解决过拟合的方法之一。

赞

石川 (作者) 回复 mockingbird 1 年前

谢谢！非常有益的评价。我们也一直在降低过拟合的道路上摸索，对您说的感受很深。

1 回复 踩 举报

呼雷 11 个月前

策略是否复杂，要与需求相匹配，逻辑要完整，不要有缺陷，不是说越复杂就越会有问题，不成正比的。好的策略，不是一根均线那么简单就能说清楚的，要不然，谁还输钱！

赞

呼雷 11 个月前

过度拟合方面的复杂，是不好的，好比树叶，就同一棵树上的树叶，看似相似，但没有两片是相同的。如若过度拟合，执必过于复杂导致不能容括，而造成不能复盖的回测损失。

赞