# 第一章 Algorithmic Trading Overview（量化交易概述）

## 1. 量化交易是什么？

## 2. 量化交易的理论发展

### 先验理论

> 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。

让我们先看一个策略的“美丽回测结果”：

![图一](https://pic2.zhimg.com/80/v2-bb0d9f0acf8ee8655ddcc7cac088c55d\_hd.jpg)

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：

- 因子计算的方法在回测期内完全一致，没有任何变化；

- 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”；

- 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）；

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

- 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。

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。

因此，对于从事量化交易的人来说，特别容易掉入“数字游戏”的圈子之中，提出一个数字上看似美好的模型而无法真正应用于实际交易之中。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。很多人都会认为：量化交易是计算机视角下的数字结果；而事实恰恰相反，量化交易扎根于理论之中。

### CAPM与因子动物园

> I thought that if Sharpe was going to publish, what’s the point of my publishing my paper? -by Treynor

资本资产定价模型（CAPM）是现代金融学的奠基石，它以马柯威茨证券组合理论为基础，研究如果投资者都按照分散化的理念去投资，最终证券市场达到均衡时，价格和收益率如何决定的问题。

该模型对资产风险与其期望收益率之间的关系给出了精确的预测。这一关系发挥着两个重要作用：

- 它为评估各项资产提供了一个基准收益率

- 该模型帮助我们对还没有上市交易资产的期望收益率做出合理估计

对于绝大多数像你、我一样的读者，谈到 CAPM，恐怕条件反射直接想到的只有 William Sharpe（毕竟他太出名，而且也因此获得了诺贝尔奖）。真正深耕金融领域的小伙伴可能除了 Sharpe 之外会知道 Lintner 和 Mossin（尤其应该知道 Lintner，因为他推导 CAPM 的出发点和 Sharpe 不同）。但是由于 Treynor 太低调，恐怕很少有人知道他也是 CAPM 的发明者之一（事实上，他应该是最早的发明者，但是他早期的论述并没有公开发表）。

1958 年，Modigliani 和 Miller 提出了著名的 MM 定理（Modigliani and Miller 1958），它也称为资本结构无关原理。该定理认为，在不考虑税，破产成本，信息不对称并且假设在有效市场里面，企业价值不会因为企业融资方式改变而改变。

MM定理如此吸引人，以至于直接影响了六十年代的CAPM理论的提出。按时间先后顺序为：

- Treynor, J. L. (1961). Market Value, Time, and Risk. Unpublished manuscript. “Rough Draft” dated 8/8/61, 95-209.

- Sharpe, W. F. (1964). Capital Asset Prices: A Theory of Market Equilibrium Under Conditions of Risk. Journal of Finance, Vol. 19(3), 425 – 442.

- Lintner, J. (1965a). The Valuation of Risk Assets and the Selection of Risky Investments in Stock Portfolios and Capital Budgets. The Review of Economics and Statistics, Vol. 47, 13 – 37.

- Mossin, J. (1966). Equilibrium in a Capital Asset Market. Econometrica, Vol. 34(4), 768 – 783.

事实上，早起的CAPM各不相同，甚至没有被各自的提出者放到一个论文的显著位置上。直到2003年，French的一篇文章完整的揭示了CAPM的四位提出者的共性，如下图所示：

![图一](https://pic4.zhimg.com/80/v2-55db6e23ee320b2e081d704afc3f9973\_hd.jpg)

CAPM是如此重要，以至于从进入大众视野中一开始便牢牢占据了之后五十年中的量化C位。然而，虽然理论的假设非常完美，投资者在实际使用过程中却不得不对其不断的修正，于是，在这领域的研究如雨后春笋般层出不穷，从单因子，到双因子、三因子，乃至多因子。第一个尝试这项工作的是Fischer Black。 Black et al. (1972) 观察到，实际的市场数据证明，真实的股票收益率和它们的 β 画出来的资本市场线（SML）远比 CAPM 预测的要更平缓。为了解释两者的β的差距，Black抛弃了 CAPM 中存在无风险利率这个假设并提出了一个双因子模型：

![图一](https://www.zhihu.com/equation?tex=%5Cmbox%7BE%7D%5Br\_j%5D%3D%5Cmbox%7BE%7D%5Br\_z%5D%281-%5Cbeta\_j%29%2B%5Cmbox%7BE%7D%5Br\_m%5D%5Cbeta\_j)

另一位解释β的集大成者，既是Fama, Fama 和 French 提出了大名鼎鼎的三因子模型，它在市场因子的基础上加入 HML 和 SMB 两个因子。这些努力都是为了能够更好的解释我们在股票收益率数据中观察到的β在截面上的差别。在之后数十年的研究生涯中，三因子、四因子与五因子模型被不断的提出；而其跟随者则利用数理统计的方法，提出了成百上千种的因子。一时间，factor zoo成了投资市场的主要研究方式，每一家公司都在试图使用更多的因子来完善自己的策略

（插一段factor zoo 到factor war的事）

在 CAPM 出现之前，人们不了解风险和预期收益之前的关系。而 CAPM 说明，不是所有的风险都能带来回报 —— 只有那些无法通过多样化而消除的系统性（市场）风险才能带来回报。从最早的 CAPM，再到 Black zero-beta CAPM，再到上面的这个 betting against beta。举这个例子是想说明无数学者前赴后继的投身于理解市场真谛的努力中，让我们逐渐看清什么是正确的。

### 动量理论

动量策略最初进入人们的视线是在 1967 年。当年，Robert Levy 在 Journal of Finance 上发表了一篇题为 Relative Strength as a Criterion for Investment Selection 的文章（当年动量 —— momentum —— 一词还没有被造出来）。在该文中，Levy 给出的结论是：通过买入过去一段时间最强势的股票要比随机选股获得更高的收益。这个发现令人惊喜，但关于动量的研究并没有因此而拉开大幕。因为在同时期，有效市场假说（efficient-market hypothesis，EMH）诞生了。

在 19 世纪 60 到 70 年代，在金融领域占有举足轻重地位的芝加哥大学（具体的，Eugene Fama）提出了有效市场假说。从 EMH 的观点出发，Levy 的发现根本不值一提，一定是搞错了。在 EMH 被提出之后的 25 多年里，该理论被学术界广泛认可、获得了飞速发展。而这段 EMH 的繁荣时期恰恰是动量策略的一段黑暗时刻。在那 25 年里，学术界的顶级期刊上的所有篇幅都预留给了 EMH 的追随者；任何挑战 EMH 的发现都被禁止。在对动量策略的看法上，价值投资的先知们和 EMH 的拥趸持有完全相同的看法 —— 动量投资是“一种黑色艺术，一种巫术魔力，只有愚人和异端者才会尝试”。这其实不难理解，在价值投资者看来，价值投资是对基本面的充分挖掘、从而找到价值和价格的背离，这需要对行业和公司的深入理解、这是一门科学；而动量呢？只需要傻买过去涨的猛的就行了，毫无技术含量。

1993 年 Narasimhan Jegadeesh 和 Sheridan Titman 在 Journal of Finance 上发表了一篇对于动量策略来说是里程碑式的文章，题为 Returns to Buying Winners and Selling Losers: Implications for Stock Market Efficiency。在大量的实证数据分析之下，动量策略终于走上台面；所有否定者都明白，如果再忽视它就只能说明自己的无知与傲慢。

如今，动量选股广泛被人接受。就连 EMH 之父 Eugene Fama 也承认动量的存在。在 2008 年，美国金融协会的金融大师系列访谈节目中，Fama 坦言“动量策略确实存在于世界各国，除了日本”（下图）。

![图一](https://pic3.zhimg.com/80/v2-33ebfb98db311bd35e28e5e0d55103d6\_hd.jpg)

别笑，中国也没有。

不过后来中国和日本的投资市场也被证明存在动量效应，并且是由Fama的得意门生Asness证明的。2013 年，Asness、Moskowitz 以及 Pedersen 在著名的 Journal of Finance 上发表了一篇影响深远的文章，题为 Value and Momentum Everywhere（无处不在的价值和动量，Asness et al. 2013）。这篇长度为 57 页的文章从最初提交到最终发表历时 4 年，通过大量实证指出价值和动量存在于全球所有市场中（股票、外汇、固定收益、商品期货等）。

至此，投资市场对于动量效应的存在已经没有异议，大量的机构投资者开始了对动量的研究。然而，笔者在此想之处的是：虽然学界与业界都已经对动量给与了很高的重视（动量和价值是 Barra 因子库中的两个雷打不动因子）， 但这种重视依旧无法匹配动量效应在量化交易中应有的权重，而是更多的将其等权的归为一个因子。

### 行为金融理论

> Markets can remain irrational longer than you can remain solvent.

> -by Keynes

作为行为经济学的一个分支，行为金融学研究投资者行为。它认为证券的市场价格并不只由证券内在价值所决定，还在很大程度上受到投资者主体行为的影响，即投资者心理与行为对证券市场的价格决定及其变动具有重大影响。2013 年，诺贝尔经济学奖同时颁发给了对市场有效性持完全对立立场的 Eugene Fama 和 Robert Shiller（另外一位是提出 GMM 的 Lars Peter Hansen）。Fama 和 Shiller 的同时获奖颇具“讽刺意味”。在颁发诺贝尔奖时，瑞典皇家科学院指出，这三位学者的发现表明“市场价格的波动受到理性和人性行为共同影响”。这是人们在探究市场真相道路上坚实的一步，也是行为金融学发展历程中的重要里程碑。

由于投资品收益的不确定性，投资者事实上都是在风险下做决策（decision-making under risk），这也是行为金融学主要研究内容。在这方面，最著名的模型包括期望效用理论（Expected Utility Theory, Bernoulli 1954）和前景理论（Kahneman and Tverskey 1979）。两者相较，后者可以解释所有前者能解释的现象，但反之则不然。可见，前景理论优于前者。

行为金融学的代表人物当属 Robert Shiller，而奠定其地位的正是 Shiller (1984) 这篇提出了噪音交易者模型的论文。这篇论文成为了日后日益增长的行为金融学文献的起点。在 Shiller (1984) 的模型中，聪明投资者依基本面价值进行投资；而噪音交易者的存在造成了价格和内在价值出现了偏离；价格的过度波动来源于人们非理性行为造成的对基本面信息的过度反应。反观聪明交易者，虽然能对预期回报做出理性反映，但这种反映因受到自身财富的限制而并不充分。

自 Shiller (1984) 发表之后，大量相关的研究结果被提出，这其中有很多是从行为偏差的角度来修正 CCAPM 中理性偏好的假设，这无疑有着非凡的意义，它意味着基于理性和行为的模型正在结合。另一方面，更多的学者开始用心理学的发现来研究个人的行为和偏误，这些心理学发现包括 prospect theory、overconfidence 以及 mental accounting 等。

![图](https://pic4.zhimg.com/80/v2-0dbd33705be090db825780a622d87547\_hd.jpg)

Shiller (1984) 的另一个功绩是回应了长久以来“理性投资者”派对行为金融学的最大质疑 —— 如果价格偏离了价值，即能被预测，为什么没有因套利而消失？Shiller (1984) 提出了 limits to arbitrage（有限套利），排除了这个阻碍行为金融学发展的障碍。

在行为金融学的进一步发展中，大量的研究表明资产的价格和投资人非理性的情绪有密切的关系（比如泡沫）。在 Wikipedia 上列出的认知偏差超过 110 种，这里列举了其中最广为人知的一些：

- Anchoring（锚定效应）

- Availability Heuristic（可得性启发法）

- Bandwagon Effect（从众效应）

- Confirmation Bias（确认偏误）

- Framing Effect（框架效应）

- Gambler’s Fallacy（赌徒谬误）

- Hindsight（后见之明）

- Insensitivity to Sample Size（对样本数不敏感）

- Outcome Bias（结果偏误）

- Overconfidence（过度自信）

- Peak-End Rule（峰终定律）

- Selective Perception（选择性知觉）

作为一名量化交易从业者，看到这个单子应该是很激动的：因为它向我们展示了量化的研究方向。任何一种认知偏差都会直接导致市场交易者的非理性行为，而这些非理性行为进一步的展现为市场的波动。无论这种波动是长期的还是短期，都带来了一种 pattern， 一种可预测的、不断重复的市场特征。通常来讲，我们认为认知偏差-行为偏差-市场非理性波动这一传导链会在市场微观结构上有着更好的表现，因此已经有大量的研究员开始了相关的研究。

### 机器学习交易

虽然我建议投资者更冷静的看待市场中对于“人工智能选股”的追捧，但我也承认，机器学习是一种非常优秀的研究工具。

## 3. 量化交易系统基本架构

- Parameter optimization（参数优化）

- Overfitting and cross-validation（过度拟合和交叉验证）

## 4. 宽客们依旧会陷入的心理偏差

### 数字陷阱

### 经验陷阱

# 第二章 Python for Finance 常用packages 学习

> Life is short，you need Python。

> by Bruce Eckel

为什么Python已经成为了量化研究中最主流的语言？

事实上，在使用Python很长一段时间内，我都认为它是一种“可执行的伪代码”，因为Python的代码实在太简洁了。Python是一种为提高开发速度而设计的高级语言。它拥有大量的库，可用于几乎任何可以想象的计算任务。Python在量化研究方面的特点是显而易见的：

- 学习 - 与C ++等其他语言相比，Python非常容易学习。仅仅几周之后，你就可以非常高效地使用Python用法。

- 库 - 使用Python的主要原因是它带有一个惊人的数组library。特别是，我们将利用NumPy（向量化操作），SciPy（优化算法），熊猫（时间序列分析），statsmodel（统计建模），scikit-learn（统计/机器学习），IPython（交互式开发）和matplotlib（可视化）。

- 发展速度 - Python在某种程度上擅长开发速度评论说它就像写在“伪代码”中一样。工具的交互性IPython使战略研究非常迅速，而不会牺牲稳健性。

- 执行速度 - 虽然速度不如C++快，但Python提供了科学计算经过严格优化的组件（通过矢量化）。如果执行速度成为一个问题，可以利用Cython并获得类似于C的执行速度代码复杂度小幅增加。

- 成本/许可证 - Python是免费的，开源和跨平台的，可运行Windows，Mac OSX或Linux。

然而，这不等于Python是唯一的语言：R在时间序列的处理上依旧是由于Python的；大量的研究人员已经使用MATLAB或SAS来高效地进行数据分析；而对于高频交易来说，C++无疑是最优的选择。从我身边的例子来看，很少的量化从业者只使用一种语言；但几乎每个人都讲Python作为常用语言之一。因此，本章将简单的介绍下一个合格的量化从业者应掌握的Python内容，以库的方式逐一呈现。

## 1. 学习数据分析基础 library （库） -- NumPy:

Numpy相对于Python的定位，类似于STL相对于C++：提供了更为高效的数据结构来存储和处理大型矩阵，比Python自身的嵌套列表（nested list structure)结构要高效的多。而且，Numpy本身是由C语言开发。这个是很基础的扩展，其余的扩展都是以此为基础。

- Creating Arrays（创建数组）

- Using Arrays and Scalars（使用数组和标量）

- Indexing Arrays（索引数组）

- Array Manipulation（数组操作）

- Array Functions（数组函数）

## 2. 学习数据分析高阶 library – Pandas:

- DataFrames and file reading（DataFrames和文件阅读导入）

- Index and Reindex Objects, Index Hierarchy（索引和索引命令对象,索引的层次结构）

- Select/Drop Entry（选择/删除条目）

- Data Alignment, Rank and Sort、Handling missing data（数据对齐、等级和排序，处理缺失数据 ）

- Summary Statistics（汇总统计）

## 3. 统计分析和最优化 library—scipy

- Optimization（优化）

- Statistical test（统计检验）

- Linear algebra-linalg （线性代数）

## 4. 画图 library—matplotlib

- How to plot basic graphs for different types（如何绘制基本图形为不同的类型）

- How to plot multiple graphs and do arrangement（如何绘制多个图形并进行排列）

- Advanced plotting （高级绘图/数据可视化）

## 5. 统计模型library--statsmodel

- Regression and generalized regression models（回归和广义回归模型）

- Time series analysis （时间序列分析）

- Statistical test（统计检验）

- Distributions （分布）

## 6. 金融数据处理

- Frequency of data（数据的频率）

- How to source data from Bloomberg、Yahoo Finance and so on（如何得到源数据）

- Data quality check and cleaning(smooth, seasonality adjustment, fill-forward and so on)（数据质量检查和清理）

# 第三章 量化交易中的数学知识

## 1.Forecasting measures and techniques overview

（预测措施和技术概述）

## 2. Statistical learning and techniques

（统计学习和技术）

## 3. Financial time series analysis（金融时间序列分析）

- 序列相关系和random walk

（随机游走）

- 平稳时间序列模型-AR/MA/ARMA

（波动率预测模型）

- 非平稳时间序列模型-ARIMA/异方差模型-GARCH

## 4. 贝叶斯模型

- Advance algorithmic trading overview

（高级算法交易概述）

- What is Bayesian statistics

（什么是贝叶斯统计）

- Bayesian Inference methods

（贝叶斯推理方法）

- Markov Chain Monte Carlo

（MCMC 马科夫链门特卡罗）

- Linear regression model based on Bayes

（基于贝叶斯的线性回归模型）

- Bayesian stochastic volatility model

（贝叶斯随机波动模型）

## 5. State-model and Kalman filter（状态模型和卡尔曼滤波 ）

- Kalman filter theory （卡尔曼滤波器理论）

- Application to regression and pair trading in Python （卡曼滤波器在回归及配对交易方面的应用）

## 6. Hidden Markov Models （隐式马科夫模型）

- HMM theory （HMM理论）

- Application to market regime detection in Python（HMM在市场机制判定/探测的应用）

## 7. Python for ODE PDE numerical methods (Python for 偏微分方程数值解)

- ODE examples in Finance

(常微分方程金融例子)

- Forward Backward Crank-Nicholson Methods for ODE

（向前向后CN方法）

- Explicit Implicit and CN methods for PDE

(显式隐式CN方法)

- Option pricing examples for PDE

(偏微分方程期权定价例子)

# 第四章 传统量化交易策略和Python实现

> 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。

事实上，我不是很想写这一章，我担心这一章节会成为善于偷懒的读者的主要阅读部分。恰恰相反，这一章节的内容恰恰是一名合格的量化交易员应该跳过的：这一章仅仅是为了展示一些传统的、简单的策略的实现方法。无一例外，虽然市面上绝大多数关于量化的教程和书都会教你如何实现，这些策略在各种交易对象上都无法证明是持续有效的，读者更应该将本章内容当做一个初学者的作业来看。

因此，在阅读时请牢记这句几乎出现在每份研报上的话：

## 1. 第一个简单的策略

策略逻辑

模型实现

策略分析结果

## 2. Moving-average trade（移动平均交易）

策略逻辑

模型实现

策略分析结果

## 3. 反转策略

策略逻辑

模型实现

策略分析结果

## 4. 海龟交易法则

策略逻辑

模型实现

策略分析结果

## 5. 均值回归策略

策略逻辑

模型实现

策略分析结果

## 6. 羊驼交易法则

策略逻辑

模型实现

策略分析结果

## 7. PEG策略

策略逻辑

模型实现

策略分析结果

## 8. 鳄鱼交易法则

策略逻辑

模型实现

策略分析结果

# 第五章 股票市场量化交易常见策略

## 1. 因子与异象

### 异象

多因子模型是 empirical asset pricing 的一种常见方法；其研究的核心问题是找到一组能够解释股票预期收益率截面差异的因子（见《股票多因子模型的回归检验》）。假使我们根据基本面特征或量价指标（或 whatever，下面统称为特征）挑选出一揽子股票并构建多空投资组合；如果该组合的收益率无法被用于 asset pricing 的多因子模型解释，则称该特征为一个异象（anomaly）。

在数学上，这意味着该组合有模型无法解释的 α 收益率：以使用该特征构建的多空组合收益率为被解释变量放在回归方程的左侧，以多因子模型中因子收益率为解释变量放在回归方程的右边，进行时序回归，回归的截距项就是 α 收益率；如果 α 显著不为零，则说明该特征是一个异象。

从有效市场假说的观点出发，市场中不应该存在很多异象。当然，有效市场假说并不完美，再加上学术界几十年来的“不懈努力”，针对美股挖掘出了 400+ 个异象，这些异象在样本内的统计检验中都获得了很高的 t-statistics。

产生如此多的异象主要有两个原因：

第一个也是最主要的一个原因是数据挖掘。在 p-hacking 的激励和 multiple testing 的盛行下，大量所谓的异象在样本内被挖出。Harvey, Liu and Zhu (2016) 研究了学术界发表的 316 个所谓显著异象，并指出再考虑了 multiple testing 的影响后，异象收益率的 t-statistic 至少要超过 3.0（而非人们传统认为的 5% 的显著性水平对应的 2.0）才有可能是真正有效、而非来自运气。

第二个原因和回归方程右侧的定价模型有关。比如，如果仅以 CAPM 为定价模型，那么很多异象都能获得 CAPM 无法解释的 α 收益率；随着定价模型中因子个数的增加，更多的异象变得不再显著。然而，真正的定价模型是未知的。

谈及对异象的研究，不能不提的一篇文章是 Hou, Xue and Zhang (2017)。这篇长达 146 页的文章惊人的复现了学术界提出的 447 个异象，涵盖动量（57个）、价值/成长（68个）、投资（38个）、盈利（79个）、无形资产（103个）、以及交易摩擦（102个）六大类。下图节选了少量动量类异象说明，感受一下。

![图一](https://pic3.zhimg.com/80/v2-b0faff0a4828608a0306b2d3ff94c2ae\_hd.jpg)

对于这 447 个异象，当排除了微小市值股票的影响后，其中 286 个（64%）不再显著（在 5% 的显著性水平下，下同）；如果按照 Harvey, Liu and Zhu (2016) 的建议把 t-statistic 阈值提升到 3.0，则其中 380 个（85%）异象不再显著；最后，如果使用 Hou, Xue and Zhang (2015) 提出的 4 因子模型作为定价模型，那么其中 436 个（98%）异象不再显著，剩余存活的仅有 11 个。

好一个数据挖掘！

除此之外，Linnainmaa and Roberts (2018) 花费了很大的经历构建了全新的样本外数据，研究了美股中源于会计数据的 36 个异象在样本内、外的表现的差异。分析表明，绝大部分异象在样本外明显失效，它们的失效说明这些异象并非来自未知风险以及错误定价这两种解释，而更有可能仅仅是数据挖掘的产物。

### 因子

上一节介绍了异象，本节就来看看什么是因子（factors）。一个异象是可能成为一个优秀因子的；然而由于异象之间的相关性，并不是所有异象都是因子。

一个因子应该能够对解释资产（可以是个股也可以是个股组成的投资组合）预期收益率的截面差异有显著的增量贡献。如果异象满足上述条件，它就可以被称之为一个因子。在这个定义中，有两个关键词值得解读，它们是“解释”和“增量贡献”：

- “解释”说明这个异象（或者潜在因子）已经从回归方程的左侧移到了回归方程的右侧，它被用来当作解释变量来对资产的收益率做回归，考察它是否能够解释预期收益率的截面差异。

- “增量贡献”暗示着同时考虑多个异象（因子）时，由于它们之间不完全独立，需要排除相关性的影响。

举个例子。我们知道价值因子是一个靠谱的选股因子。然而，很多指标 —— 比如 E/P 或 B/P 都可以用来构建价值因子的 High-Minus-Low 组合。如果同时基于 E/P 和 B/P 构建了 HML\_EP 和 HML\_BP 两个因子，它们之间的相关性注定是非常高的。一旦选择了其中之一作为价值因子，另一个对于资产预期收益率截面差异解释能力的增量贡献就不再显著、无法成为因子。

从资产定价的理论角度来说，多因子模型中的因子之间应尽可能独立；但是从投资实践来说，上面例子中的 E/P 和 B/P 可以被同时使用构建一个 HML 价值因子，这有助于降低波动且增加因子的鲁棒性。

在从一揽子异象中筛选因子时，常见的做法是将它们同时作为回归分析中的解释变量，采用 Fama-MacBeth Regression（Fama and MacBeth 1973）来分析这些异象的收益率是否显著。在这方面，Green, Hand and Zhang (2017) 是一个很好的例子。

Green, Hand and Zhang (2017) 使用 Fama-MacBeth Regression 同时检验 94 个异象，并考虑了 multiple testing 对 t-statistic 以及 p-value 造成的影响，最终发现仅有 12 个异象可能成为潜在的因子：1. 账面市值比；2. 现金；3. 分析师数量的变化；4. 盈余公告宣告收益；5. 一个月的动量；6. 六个月动量的变化；7. 盈利同比增长的季度数量；8. 年度研发支出占市值的比重；9. 收益波动性；10. 股票换手率；11. 股票换手率的波动性；12. 零交易的天数。

上述结果告诉我们：在修正 multiple testing 的数据挖掘、以及考察了不同异象的相关性之后，真正能够解释资产预期收益率截面差异的独立因子少之又少。

### 多因子检验

现在我们已经了解了异象，并通过回归分析从异象中找出了因子，接下来就是挑选因子构建多因子模型了。在构建多因子模型时，两个必须要回答的问题是：（1）选择多少个因子合适？（2）选择哪些因子更好？学术界对于第一个问题的共识为主流因子模型奠定了基调；而第二个问题则涉及不同多因子模型之间的比较。

对于第一个问题，我们总可以仅使用一个因子，比如 CAPM 模型仅使用了市场因子；又或者我们可以使用许多因子。一个极端的情况是把每个上市公司作为一个因子，每个公司的股票只在自己公司的因子上有暴露，在代表其他公司的因子上零暴露。这样的模型显然能完美解释股票预期收益率的截面差异，但在现实中没人会那么用。以样本内过拟合为代价，更多的因子总能更好的解释收益率的截面差异。

在学术界，“几个因子合适”这个问题所遵循的准则是 The Law of Parsimony（简约法则），它还有一个更为人熟知的名字 —— Occam’s razor（奥卡姆剃刀）。如果从 ICAPM（Intertemporal CAPM）的角度来理解多因子模型，每个因子代表某种 state variable；而 state variable 是投资者想要对冲的某种风险。从这个意义上说，因子的个数应该是有限的。

依据简约法则，学术界主流的多因子模型包括以下几个（按时间顺序、排名不分先后，不完整 list），它们的因子个数均在 3 到 5 个之间：

- Fama-French 三因子模型（Fama and French 1993）：多因子模型的开山鼻祖，包括 MKT，HML 以及 SMB 三因子。

- Carhart 四因子模型（Carhart 1997）：在 Fama-French 三因子模型上加上了动量 MOM 因子。

- Novy-Marx 四因子模型（Novy-Marx 2013）：包含 MKT，HML，MOM 以及 PMU 四个因子；其中 PMU 所用的财务指标是 Gross Profit-to-Asset，代表 profitability 维度。

- Fama-French 五因子模型（Fama and French 2015）：Fama 和 French 在其三因子模型的基础上加入了 CMA 和 RMW 两个因子，分别代表 investment 和 profitability 两个维度。

- Hou-Xue-Zhang 四因子模型（Hou, Xue and Zhang 2015）：包含 MKT，SMB，IVA 以及 ROE；其中 IVA 是 total assets 的年增长率，代表 investment 维度。

- Stambaugh-Yuan 四因子模型（Stambaugh and Yuan 2016）：包含 MKT，SMB，MGMT 和 PERF 四个因子。MGMT 和 PERF 分别使用了 6 个和 5 个指标，代表和 management 以及 performance 相关的两个 mispricing 因子。虽然该模型只有四个因子，但它用到的基本面和量价指标多达 12 个。

- Daniel-Hirshleifer-Sun 三因子模型（Daniel, Hirshleifer and Sun 2018）：在 MKT 的基础上，使用 PEAD 和 FIN 两个指标作为短期和长期行为因子（behavioral factors）的代理指标，构建了三因子模型。该模型由于包括了传统的 MKT 风险因子，又包括行为因子，故称为复合模型。

如何比较不同的多因子模型呢？学术界主要有以下三种方法：

- GRS tests；

- Mean-Variance Spanning tests；

- Bayesian approach。

GRS tests（Gibbons, Ross and Shanken 1989）检验 n 个资产在给定因子模型下的定价错误（pricing error）—— 即 α —— 是否在统计上联合为零（jointly equal to zero）。在比较两个多因子模型时，使用两个模型的因子互为资产和定价模型进行检验。

## 2. 基本面量化

## 3. 股票α策略模型

α的来源

## 4. 财务报表评分卡

## 5. 理想与现实：中国股市量化交易面临的问题

股市时间太短、数据挖掘无法避免

任何量化策略都是“以史为鉴”；所有被视为优秀的因子无疑不在使用历史数据回测时展现了优秀的收益率。然而，学术界和业界根据历史回测找到的因子一定存在选择性偏差：就是因为表现好才被选出来，如果表现不好根本不会被选出来拿来用。这些因子都不可避免的、或多或少的带着数据挖掘的弊端；它们偏重拟合历史，并假设历史在未来能够重复。

然而，对于国内 A 股市场。股市从上世纪 90 年代初建立以来，虽说也有近三十年，但实际是经历了如下五个阶段：

- 1990 年 – 1991 年：股市初创阶段

- 1992 年 – 1997 年：股市试验阶段

- 1998 年 – 2001 年：股市规范阶段

- 2002 年 – 2004 年：股市转轨阶段

- 2005 年 – 今： 股市重塑阶段

这其中，真正能被用来做量化投资分析的也就是从 2005 年开始，至今不过短短十二年，远不及美股中丰富的 30 年、50 年甚至是更长的回测期。如果一个因子在历经了美股的三、五十年回测都不敢说在样本外一定有效，我们又怎敢拿着区区 20 年数据挖掘出来的因子来使用呢？

投机者比重高，市场理性程度不够

由于巴菲特的巨大成功，价值投资对量化因子选股影响深远。价值投资背后的核心逻辑是股票的价格围绕着其内在价值波动。Shiller (1984) 的研究提出市场由 smart-money 投资者和 noise 交易者构成。前者依股票基本面价值进行投资；而后者的存在造成了股票价格围绕价值投资的原因。在一个理性程度高的市场，前者主导，价格围绕价值的波动不大；而在一个理性程度低的市场，后者主导，价格往往大幅偏离股票基本面显示的内在价值。

国内 A 股市场显然属于后者。对于 A 股市场的价值投资者来说，即便他们能够准确的评价股票的价值，但如果价格迟迟不向价值回归，那这无疑是一种煎熬；而对于投机者来说，他们往往换手率很高，个股持仓时间短，压根就不关注价格和价值之间的关系。大量国内的因子投资实证显示，因子投资组合要想赚钱，其持仓周期仅仅是三到四周，这较海外的实证几乎少了一个数量级。有多少股票的价格能在短短数周之内回归到基本面内在价值？因此，我们只能对因子有效给出一个勉强的猜测，即我们不是在等待股票的价格向价值回归，而是在个股的相对价格价值偏离之间轮动！每次短周期的调仓，我们大抵只是把价格偏离价值的较少的股票换成了价格偏离价值更大的股票上。无论这个猜测是否正确，它早已和价值投资理念所衍生出来的因子的作用毫无半点关系了。

数据可信度和效用不高

独门数据源永远是因子选股中最核心的优势。如果你有别人没有的靠谱数据，你一定能利用它练就出“独门武功”。不幸的是，我们绝大多数人没有这样的资源。国内的量化因子选股主要利用三大块数据源：（1）股票交易数据（价格、交易量等）；（2）会计财务报表数据；（3）分析师对股票的分析评级数据。出于不同的原因，这三个数据源的可信度和效用都有一定的问题。

交易数据应该是最客观的数据。但是对于一些小市值股票，“庄家”的存在往往使价格出现失真，从而造成失真的收益率。以此挑选因子势必南辕北辙。

卖方分析师是股票投资的一个重要组成部分。卖方通过对行业和公司进行调研分析，对股票评级并给出投资建议；这对于引导市场一致预期以及买方机构投资决策将产生一定影响，进而影响股价变化。优秀分析师的真知灼见可以为我们了解行业和公司提供宝贵的依据。站在量化投资的角度，很多学者将分析师对股票的判断作为一个挖掘超额收益的因子使用。然而实证发现，在排除了市值和估值的暴露后，分析师评级因子并无额外获得超额收益的能力。

最后剩下的就是财报因子。财务报表是上市公司最重要的信息披露渠道，是投资者了解上市公司财务状况和经营情况的基本途径。巴菲特就是靠严谨的财务分析取得了连续 32 年战胜市场的纪录。在美股中，投资者对于财报的反应迅速而到位。然而在 A 股中，由于内幕交易和财报数据造假难度低，导致财报的效用远低于美股。

畸形的IPO制度带来了壳价值

对于 A 股的上市发审制度，造成上市公司的壳资源价值非常高。由于 A 股上市成本高、时间长（发审制度下需要排队两年左右），使得通过收购已经上市、但市值不大的股票实现借壳上市，是一件性价比不低的生意。同时 A 股小市值成长股估值偏高、可以以更高市盈率发行股票、以极低的成本融资（30 倍市盈率发行股票就相当于 3.3% 的利率借款、还不用还、还可以进一步去作为资本金上债务杠杆），非常合算。当然，还可以通过上市膨胀企业家身家顺便大幅套现，总之对非上市公司来说上市是一个收益颇丰的事情。

这就是为什么A 股的 ST 股票是一道独立的风景，经常在业务萎靡、亏损放大和面临退市的不利局面下上演反转剧情：业绩扭亏保壳成功（这是一个浪子回头的故事）、或者成为壳资源被优秀企业借壳上市（这是一个屌丝变身高富帅的故事），股价进而逆转飙升。

这里推荐下

> Lee, C. M. C., Y. Qu, and T. Shen (2017). Reverse mergers, shell value, and regulation risk in Chinese equity markets. Working paper.

这篇文章。虽然市场对于A股壳价值的认识始终存在，但Lee et al. 研究了中国股市的壳价值和监管风险，并指出冗长的 IPO 发审制度造就了中国股市中独有的壳价值，而壳价值可以造成其他主流因子解释不了的预期收益率截面差异。此外，对于壳价值的研究还可以解释中国股市的一些现象 —— 比如，当考虑了壳价值因子后，小市值因子几乎就消失了，因此可以说小市值股票有效的内在原因是潜在的壳价值。

# 第六章 期货市场量化交易常见策略

## 1. 期货交易与股票交易的区别

## 2. 商品期货趋势型策略

# 第七章 其他衍生品定价模型

## 1. 蒙特卡洛模拟基础

## 2. 常见随机过程离散化

## 3. European Option（欧式期权）蒙特卡洛模拟定价

## 4. Exotic option（奇异期权定价）

## 5. Least-square monte-carlo for American option pricing

## (最小二乘蒙特卡罗对美式期权定价)

## 6. 期权套利原理

# 第八章 Pair trading （配对交易）

## 1. 套利原理示意图

## 2. 商品板块

## 3. 产业链

## 4. 跨品种价差

## 5. 时机交易

# 第九章 风险管理

## 1. 指标、模型与策略

## 2. 模块划分

## 3. 行情与数据

### 4. 日志框架

## 5. 风险控制与头寸管理

# 第十章 从理论到上线：策略的提出与实现

## 1. 指标、模型与策略

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.

## 2. 模块划分

## 3. 选择样本与数据

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。

## 4. 日志框架

## 5. 风险控制与头寸管理

## 6. 模型动力学

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。