量化投资学导论

ICSI

施想
俄亥俄州立大学

2023

课 纲

邮箱: orashshi@gmail.com 电话：(971) 8258828

时间: 周日 19:30 - 21:30 教室: Hitchcock Hall 031

课程简介

本课程旨在为社员提供量化投资的基础理论、工具和策略，培养社员具备独立设计、测试和执行量化投资和设计交易策略的能力。

目标学生

金融学、数学、统计学、计算机专业学生、金融工程师、量化研究员、及对量化交易感兴趣的投资者。

课程网站

GitHub 链接: https://github.com/orashshi/quantitativeinvestment

先修课程

微积分，Python。

课程目标

1. 了解量化投资的基本理论工具.
2. 了解多因子模型的运行原理与构建.
3. 股票分类模型的搭建.
4. 量化价值投资的原理.
5. 使用机器学习初探量化价值投资.
6. 建立股票初筛模型.
7. 入门量化投资，为潜在的实习机会做准备.
8. 作为量化交易课的前置课程.

课程结构

在本量化投资课程中，我们将深入探讨各种投资算法的核心逻辑。课程的目标不仅是让学生理解这些统计学方法或算法的工作原理，更重要的是，希望学生能够将理论知识转化为实践能力。

因此，每一堂课后，社员都将提交一份文件至 Github。学生将尝试对课上所讲解的统计学方法或算法进行代码复现。这是极其重要的，不仅能够巩固学生的理论知识，更能锻炼他们的编程技能和在未来职业中的实际应用能力。我们鼓励学生积极参与，通过实践来加深对量化投资的理解。

时间表

时间表以及每周课内主题是暂定的，可能会随实际教学情况发生变化。

Week 01, 2023.9.17(课) - 2023.9.23: 量化投资概论

- 什么是量化投资？
- 统计套利与有效市场假说.
- 数学，统计学，计算机要求解读.
- 数据，分析，模型，优化与算法简介.
- 量化投资的应用，以基金公司为例.

Week 02, 2023.9.24(课) - 2023.9.30: 常用统计学理论与应用

- 平均数，中位数，众数.
- 分位数，标准差.
- 夏普比率，离散度的度量.
- 收益率分布的对称性和偏度.
- 收益率分布的峰度.
- 实例分析.

Week 03, 2023.10.1(课) - 2023.10.7: 常用概率论知识与应用（上）

- 概率，期望，方差.
- 贝叶斯公式，计数原理.
- 离散型随机变量.
- 连续型随机变量.
- 蒙特卡洛模拟.
- 实例分析.

Week 04, 2023.10.8(课) - 2023.10.14: 常用概率论知识与应用（下）

- 抽样方法.
- 样本均值分布（中央极限定理）.
- 抽样偏差.
- 假设检验入门.
- 实例分析.

Week 05, 2023.10.15(课) - 2023.10.21: 休息

Week 06, 2023.10.22(课) - 2023.10.28: 休息

Week 07, 2023.10.29(课) - 2023.11.4: 回归分析

- 相关性分析.
- 线性回归与数据处理.
- 多元线性回归.
- 建模分析，问题与解决.

Week 08, 2023.11.5(课) - 2023.11.11: 休息

Week 09, 2023.11.12(课) - 2023.11.18: 时间序列分析简介

- 趋势模型.
- 自回归时间时间序列模型.
- 随机游走和单位根.
- 时间序列模型的季节性.
- 自回归条件异方差模型.
- 实例分析.

Week 10, 2023.11.19(课) - 2023.11.25: 多因子模型简介

- 多因子模型.
- 统计套利定价理论.
- 基本面多因子模型，价值投资学初探.
- 实践与应用，以港股为例.

Week 11, 2024.1.14(课) - 2024.1.20: 多因子模型的应用, 以股票分类模型为例

- 分类标准，成长股，周期股.
- A 股市场分析.
- 指标与有效因子挖掘.
- 数据处理.
- 因子灵敏度测试.
- A 股股票分类模型.

Week 12, 2024.1.21(课) - 2024.1.27: 量化价值投资基础

- 量化价值投资是如何防止人为的错误的?
- 格林布拉特的神奇公式.
- 神奇公式测试.
- 理论模型构造，量化价值投资思维实验.

Week 13, 2024.1.28(课) - 2024.2.3: 安全边际，以机器学习算法为例

- 机器学习算法初探.
- 利用 PROBM 模型预测.
- 度量财务风险.
- 算法筛选潜在的欺诈者，造假者.
- 安全边际的测量，时间序列分析模型的测量.
- 安全边际的数学模型构造与优化.

Week 14, 2024.2.4(课) - 2024.2.10: 护城河，公司质量量化研究

- 财务指标横比.
- 皮托尔斯基分值 (皮氏 F 分值)
- 改进后的财务实力分值 (FS 分值).
- 公司管理稳定性系数.
- 公司市场位置研究.
- 风险预测.

Week 15, 2024.2.11(课) - 2024.2.17: 股票初筛模型，寻找低价的优质股票

- 机器学习理论再叙.
- 输入，输出与神经网络.
- 因子相关性测试.
- 股价时间序列评估.
- 回测与算法迭代.

Week 16, 2024.2.18(课) - 2024.2.24: 量化价值投资的缺陷与总结

- 市场结构性问题，以 A 股市场为例.
- 风险预测的不可控性.
- 有效市场假说与统计套利再探.
- 稳健性因子.
- 量化交易初探.

目 录

1 量化投资概论

1.1 量化投资的定义

定义 1.1 量化投资策略，就是采用数量化手段构建而成并进行决策的策略。

具体解释起来，该定义包括两层含义。

- **数量化**的手段应该占主要成分。包括对整个投资决策流程和投资目标的数量刻画、数学模型的构建、对量化目标的最优化、对策略结果的数量化评价。这一部分是相对**主观的**。
- 策略构建完毕后，进行投资决策时必须具有明确的数量化规则，完全**不存在主观判断**的成分。这一特性使得整个策略可以在完全量化的设置下进行历史数据下的回溯测试，以及准确无误地指导投资操作。

那么我们来看这个例子，

例子 1.1 “移动平均线看起来很好时买入，看起来不好时卖出。”

这是一个量化投资策略吗？很明显，这不是。因为它的投资决策过程中，有主观判断的成分。问题有二：

- 该交易策略在表述上较为模糊，不是一个具有明确数量化规则的决策手段。交易员还是需要在交易过程中通过主观判断来完成买卖行为。
- 交易员在形成这样的交易规则时很难定量化地描述整个交易策略和交易过程，也就难以使用最优化之类的数量方法。在多数情况下，交易员可能更倚重于复盘等人工形式来完成这一类交易策略的构建。

例子 1.2 “价格线从下向上穿过移动平均线时买入，从上向下穿过移动平均线时卖出”

这是一个量化投资策略吗？是，因为这个非常的客观。由量化指标构建，也存在明确的量化交易规则。

指标交易的两个例子：

- Richard Donchian 所开发的通道规则（过去特定天数内的最高价和最低价为边界形成一个通道，当目前价格超出通道范围时，形成买卖决策）。

- Richard Dennis 的“海龟交易法则”，是在 Donchian 通道指标的基础上构建而成的，除了通道突破的买卖规则外，海龟交易法则还包括仓位大小的选择、随时间的调整、止损等多个组成部分，更接近于一个构架完整的交易策略。

1.2 量化投资的对象

量化投资的对象主要包括**股票**、**期货**、期权、债券、外汇、商品、房地产等金融资产。

1.3 研究量化投资需要哪些前置知识？

- 数学：概率论，统计学，微积分，线性代数，偏微分方程，随机过程，时间序列分析。
- 经济学：宏观经济学，微观经济学，投资学，资产定价理论，投资组合理论。
- 计算机：数据结构，算法，数据库，深度学习。

1.4 量化投资常见策略

- 基本面驱动策略 (FMQ)：当证券价格低于其内在价值时买入，高于其内在价值时卖出。由于证券的基本面价值是由按季度发布的财务报表和预测报告决定的，因此这类策略的时间跨度是一个季度。在进行基本价值分析时，常用的一些指标是公司质量、公司价值以及投资者情绪。
- 全球宏观策略：利用对市场事件和趋势的宏观分析来识别投资机会。也可以分为酌情策略和系统性策略。酌情策略是指投资者根据对市场的宏观分析。系统性策略是指投资者根据对市场的宏观分析，让模型和软件代替人工来进行决策。
- 收敛策略、相对价值策略和其他统计套利策略：收敛性策略是指同时投资于一组相似的资产，且这些资产的价格存在收敛的趋势。相对价值策略是指投资者同时买入一个资产，卖出另一个相关资产，从而获得两个资产之间的价差。其他统计套利策略是指利用统计学的方法，通过对历史数据的分析，发现市场上的定价错误，从而获得超额收益的一种投资策略。

- 高频交易策略 (HFT)，以西蒙斯为首，时间跨度为毫秒量级，且高频交易资产持有时间小于 1 秒。

1.5 量化投资的历史发展

- 20 世纪 50 年代，Harry Markowitz 提出了最优投资组合理论，使用均值，方差描述投资组合上的收益和风险。
- 20 世纪 60 年代，William Sharpe 提出了资本资产定价模型 (CAPM)，将股票在无风险收益之上的超额收益分解为两个部分，即市场部分和残余部分。股票的风险也相应地分为两个部分，对应起来分别是系统风险和非系统风险。
- 20 世纪 60 年代，Stephen Ross 基于 CAPM 从另外一些假设条件出发，得出了套利定价理论。理论模型还能够包含其他一些存在风险补偿的风险因子。
- 20 世纪 70 年代，Eugene Fama 提出了**有效市场假说**，认为市场上的股票价格已经包含了所有的信息，因此股票价格的变动是随机的，不可能通过分析预测股票价格的变动。
- 20 世纪 70 年代，Fischer Black、Myron Scholes 等人提出了期权定价模型 (BS 模型)，为期权量化投资提供了工具。

但说到量化的开山鼻祖，那必然是西蒙斯。西蒙斯是投资学的神话，主打高频交易，他的量化投资策略在 1988 年到 2008 年期间，年化收益率达到了 66%，而且波动率只有 16%，夏普比率达到了 4.1。这个收益率是相当惊人的，而且他的策略在 2008 年金融危机期间，也没有出现亏损。西蒙斯的量化投资策略，主要是基于统计套利的策略，他的策略是基于股票的配对交易，即同时买入一个股票，卖出另一个相关股票，从而获得两个股票之间的价差。

投资人	关键基金	年份	年化复合收益率*
詹姆斯·西蒙斯	大奖章基金	1988—2018	39.1
乔治·索罗斯	量子基金	1969—2000	32†
史蒂文·科恩	SAC 资本	1992—2003	30
彼得·林奇	麦哲伦基金	1977—1990	29
沃伦·巴菲特	伯克希尔·哈撒韦	1965—2018	20.5‡
瑞·达利欧	Pure Alpha 基金	1991—2018	12

图 1: 各大基金的年化复合收益率对比

推荐大家看一本书《征服市场的人》，安匀译。

1.6 有效市场假说

有效市场假说（Efficient Market Hypothesis，EMH）是由美国经济学家尤金·法玛于 1965 年提出的。在这个市场中，存在着大量理性的追求利益最大化的投资者，他们积极参与竞争，每一个人都试图预测单个股票未来的市场价格，每一个人都能轻易获得当前的重要信息。在一个有效市场里，众多精明投资者之间的竞争，会导致这样一种状况：在任何时候，单个股票的市场价格都反映了已经发生的和尚未发生、但市场预期会发生的事情，所以大家都无法获得超额收益。

该假说认为，市场上的股票价格已经包含了所有的信息，因此股票价格的变动是随机的，不可能通过分析预测股票价格的变动。

有效市场假说分为三种形式：

- 弱有效市场假说：市场上的股票价格已经包含了所有的历史价格信息，因此不能通过分析历史价格来预测股票价格的变动。
- 半强有效市场假说：市场上的股票价格已经包含了所有的公开信息，因此不能通过分析公开信息来预测股票价格的变动。
- 强有效市场假说：市场上的股票价格已经包含了所有的信息，包括历史价格信息和公开信息，因此不能通过分析任何信息来预测股票价格的变动。

总结来看，有如下的特点：

	技术分析（看图表）	基本面信息	私人信息（内幕）
弱有效市场	x	✓	✓
半强有效市场	x	x	✓
强有效市场	x	x	x

1.7 统计套利

统计套利是什么？统计套利是指利用统计学的方法，通过对历史数据的分析，发现市场上的定价错误，从而获得超额收益的一种投资策略。

- 基于模型的投资过程：期权/期货等资产的定价，均值回复
- 基于历史数据的投资过程：认为过去的资产间稳定关系在未来依旧存在
- 运用数量手段构建投资组合：估计相关指标 (价差等) 的概率分布

- 对常规风险因子进行规避：构建多空组合
- 无风险套利与统计套利：承担模型风险与资产异质量风险

具体到量化投资，统计套利的策略主要包括：

- 基于统计学模型的跨期套利：是一种基于历史价差水平的统计方式来挖掘价差稳定性以及变量间的长期均衡关系，从而制定相对客观的跨期套利策略。其无需对行情进行预期和估计，而且能够挖掘最大化的套利机会。
- 跨市场套利，大小盘轮动：商品期货跨市场套利，即根据同一品种或者具有相关性的商品期货合约之间的联动性，在某个交易所买入 (或卖出）某一交割月份的某种商品期货合约的同时，在另一个交易所上卖出（或买入) 同种或者具有较强关联性商品相对应的合约，利用两个交易所由于交易市场与原产地距离、供需关系等原因造成的价差变动来赚取利润。

1.7.1 跨期套利

我们来尝试看下这个例子：

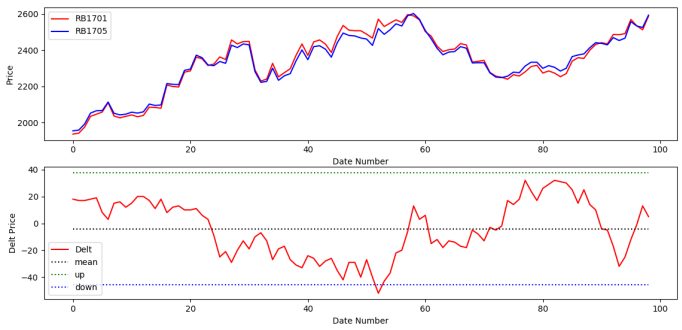

图 2: 两个时间序列

可以看出，两只股票具有同涨同跌的规律，长期以来两只股票的价差比较平稳，当价差变大或变小时，会有某种“力量”使它回归均值，我们的策略也正基于此。但我们是仅凭肉眼观测的，肯定是不严谨的不够量化的，这就需要数学工具协整性检验。如果两个股票具有强协整性，那么无论它们中途怎么走的，它们前进的方向总是一样的。

提到协整性，就不得不提平稳性。在数学中，平稳随机过程（Stationary random process）或者严平稳随机过程（Strictly-sense stationary random

process），又称狭义平稳过程，是在固定时间和位置的概率分布与所有时间和位置的概率分布相同的随机过程：即随机过程的统计特性不随时间的推移而变化。这样，数学期望和方差这些参数也不随时间和位置变化。平稳在理论上有严平稳和宽平稳两种，在实际应用上宽平稳使用较多。宽平稳的数学定义为：对于时间序列 Y_t，若对任意的 t, k, m，满足：

$$E(Y_t) = E(Y_{t+m}) \tag{1}$$

$$cov(Y_t, Y_{t+k}) = cov(Y_{t+k}, Y_{t+k+m}) \tag{2}$$

则称时间序列 Y_t 是宽平稳的。

平稳性（stationarity）是一个序列在时间推移中保持稳定不变的性质。“价差平稳”似乎不过有说服力，我们将价差做了一阶差分，可以看出，它始终是围绕着一个长期均值 0 在波动。

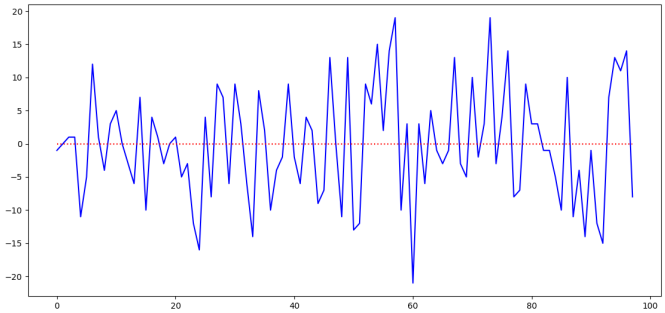

图 3: 一阶差分

定义整性为：如果某一非平稳序列 X_t 能够经过 d 次差分后变成平稳序列，就称该序列为 d 阶整性，也称为 d 阶单整，记为 $X(t) \sim I(d)$。

协整性是指若两个或多个非平稳的变量序列，其某个线性组合后的序列呈平稳性。假设两个非平稳时间序列 Y_t，X_t ，且有 $Y(t) \sim I(d)$ 和 $X(t) \sim I(b)$。如果存在某一参数向量 $(1-\beta)$，使得：

$$[Y_t - \beta X_t] \sim I(d-b) \tag{3}$$

其中，b 为正整数，$(1-\beta)$ 为协整向量，β 为协整系数。那么 Y_t，X_t 之间存在协整，记为：

$$Y_t, X_t \sim CI(d-b) \tag{4}$$

如果 $d-b=0$，那么：

$$Y_t, X_t \sim CI(d,d) \tag{5}$$

有：

$$\mu(Y_t - \beta X_t) \sim I(0) \tag{6}$$

意味着回归方程：

$$Y_t = \beta_t + \mu_t \tag{7}$$

是有意义的。协整概念是一个强有力的概念。因为协整允许我们刻画两个或多个序列之间的平衡或平稳关系。对于每一个序列单独来说可能是非平稳的，这些序列的矩，如均值、方差或协方差随时间而变化，而这些时间序列的线性组合序列却可能有不随时间变化的性质。

那么如何检验协整性呢？我们采用“Engle-Granger 两步协整检验法”，Engle-Granger 两步协整检验法用普通最小二乘法估计这些变量之间的平稳关系系数，然后用单位根检验来检验残差，如果序列是平稳的，则不存在单位根，否则就会存在单位根。

我们以两个序列 Y_t 和 X_t 为例，在检验协整性之前，首先要对序列的单整性进行检验，只有当两个序列单整阶数相同时，才有可能存在协整关系。

在 X_t 和 Y_t 具有相同单整阶数，通过单整性检验之后，我们用最小二乘法估计模型：

$$Y_t = \alpha + \beta X_t + \varepsilon_t \tag{8}$$

并计算相应的残差序列：

$$e_t = Y_t - \hat{\beta}_0 - \hat{\beta}_1 X_t \tag{9}$$

然后，检验残差序列的平稳性：

$$De_t = \delta e_{t-1} + \delta_0 + \theta_t + \sum_{i=1}^{k} \delta_i \times De_{t-i} + \varepsilon_t \tag{10}$$

利用 ADF 检验法，检验在上述估计下得到的回归方程的残差 e_t 是否平稳（如果 X_t 和 Y_t 不是协整的，则他们的任意组合都是非平稳的，因此残差 e_t 将是非平稳的）。也就是说，我们检验残差 e_t 的非平稳的假设，就是检验 X_t 和 Y_t 不是协整的假设。

那么在量化操作中，我们该如何实现以上的策略？我们可以采用如下的步骤：

- 首先获取套利（RB1801，RB1805）标的价格序列。
- 根据 EG 两步法 (1、序列同阶单整 2、OLS 残差平稳) 判断序列具有协整关系之后 (若无协整关系则全平仓位不进行操作)。
- 通过计算两个真实价格序列回归残差的 1.5 个标准差上下轨，并在价差突破上轨的时候做空价差，价差突破下轨的时候做多价差，在回归至标准差水平内的时候平仓。

1.8 量化投资的评估，以量化基金公司为例

量化基金是指利用模型和计算机代替人为决策来进行投资的基金。投资决策过程，一般包括三个关键组成部分：

- 输入模块：包括规则设定、数据获取、数据清洗、数据预处理等。
- 预测模块：用于估计未来证券的价格、证券的收益和评估风险的模型的参数。
- 构建投资组合模块：利用优化算法，根据预测模块的输出，构建投资组合。

对于基金经理或投资组合经理的业界的评估通常是用业绩来衡量的。假设市场中存在一个无风险的资产，其利率为 r_f。存在一个风险资产，其收益率为 r，且 r 的均值为 μ，方差为 σ^2。根据资本资产定价模型 (CAPM) 可知风险资产的 Sharpe 比率 (SR) 为：

$$SR = \frac{(\mu - r_f)}{\sigma} \tag{11}$$

一般来说，Sharpe 比率是越高越好的。因为 Sharpe 比率就是指每多承受一份风险所换来的预期收益，就是一个边际增量，换来的预期收益越大那么 Sharpe 比率也就越高，所以夏普比率是越高就越好的。对于一个量化基金来说，我们需要最大化夏普比率，我们可以采用凸优化的方式。

另一个常用的评估指标是 Treynor 比率 (TR)，定义为：

$$TR = \frac{(\mu - r_f)}{\beta} \tag{12}$$

其中，$\beta = Cov(r, r_M)/\sigma_M^2$, r_M 和 σ_M^2 分别是市场组合 M 的收益率和方差。

一般来说，Treynor 比率越高越好单位风险溢价越高。因为 Treynor 比率是基金的收益率超越无风险利率的值与系统性风险的比值，即每单位系统风险资产获得的超额报酬 (超过无风险利率 R_f 的部分)，适用于评价非系统风险完全分散的基金。

1.9 这节课的目标

量化的目标是什么？通俗来说，就是用计算机代替人的决策过程，用数学模型来描述投资过程，用统计学理论来描述投资目标，用机器来计算投资组合，用指标来描述投资结果，从而**征服市场**，这就是量化投资的目标。

1.9.1 量化投资和价值投资

参考文献

[1]《征服市场的人》，安匀译

[2]《量化交易》，Xin Guo