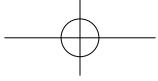




AI量化投资

李必文 著

清华大学出版社
北 京



内 容 简 介

本书旨在探索 AI 技术与投资策略的跨界融合。

全书分为上下两篇，共 10 章。上篇由量化思想、量化实践、量化方法、量化策略、风险控制绕不开凯利公式、交易信息系统外接共 6 章组成；下篇由遗传算法在黄金投资中的应用、大规模神经网络及股票非量价复合策略、小波分析及金融工程多维度应用和前沿研究与探索共 4 章组成。上篇主要阐述当前已有的量化知识并用独特鲜明的风格呈现出来，侧重计算机动态仿真技术；下篇聚焦探索未知的领域。全书注重金融实证、工程数学、计算机编程三者之间的跨界融合。

本书可作为量化基金从业人员和证券分析师的参考用书，也可作为金融专业、人工智能专业的高年级本科生、硕士和博士研究生的参考书（含毕业论文参考用书），还可以作为具备理工科背景且未来有志于从事 AI 量化投资人士的自学书籍。

本书封面贴有清华大学出版社防伪标签，无标签者不得销售。

版权所有，侵权必究。举报：010-62782989，beiqinquan@tup.tsinghua.edu.cn。

图书在版编目 (CIP) 数据

AI 量化投资 / 李必文著. —北京：清华大学出版社，2022.11

(深度投资分析丛书)

ISBN 978-7-302-62029-7

I. ① A… II. ①李… III. ①人工智能—应用—投资分析 IV. ① F830.593-39

中国版本图书馆 CIP 数据核字 (2022) 第 189447 号

责任编辑：施 猛

封面设计：常雪影

版式设计：方加青

责任校对：成凤进

责任印制：丛怀宇

出版发行：清华大学出版社

网 址：<http://www.tup.com.cn>，<http://www.wqbook.com>

地 址：北京清华大学学研大厦 A 座 邮 编：100084

社 总 机：010-83470000 邮 购：010-62786544

投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn

质 量 反 馈：010-62772015，zhiliang@tup.tsinghua.edu.cn

印 装 者：北京博海升彩色印刷有限公司

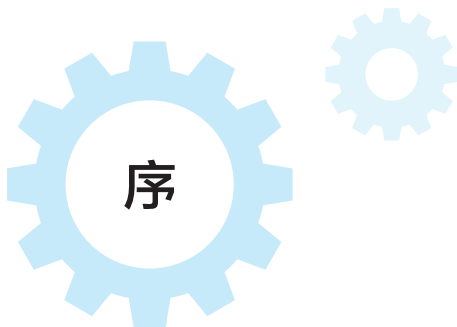
经 销：全国新华书店

开 本：180mm×250mm 印 张：15.25 字 数：281 千字

版 次：2022 年 12 月第 1 版 印 次：2022 年 12 月第 1 次印刷

定 价：188.00 元

产品编号：096929-01



从春风桃李
到秋雨梧桐
鲜衣走马
风霜不暮少年心
那些四季轮动的日子里
纵有寂寥惆怅
无悔匍匐前行
纵有伤痕累累
依旧孤帆济沧海
踉踉跄跄，懵懵懂懂
算而今
进入 AI 领域逾十载
辗转持牌私公募基金行业七余载

循规蹈矩是容易的
传授或组合信息茧房已有的知识也是容易的
教育学家斯普朗格 (E. Spranger) 曾说，
(教育) 不是传授已有的东西
而是要把人的创造力量诱导出来
将生命感、价值感唤醒



唤醒，是种教育手段

事物在交叉边缘往往更有生命力
恩格斯在《自然辩证法》里曾说，
在分子科学和原子科学的接触点上
物理学家和化学家都承认自己无能为力
然而应当在这点上期待最大的成果
那么
AI 与投资融合会发酵出什么呢？

形式大于实质毫无意义
只有创造价值才有价值
无论是逆流而上的价值投资
抑或顺势而为的 AI 量化投资
——金融时序数据的惯性

我有一个梦想
AI 在金融领域还在蹒跚探索
为推动全球 AI 量化投资进步一小步
渺小人物也想贡献一份绵薄之力
没有什么比儿时改变世界的梦想
更加让人心潮澎湃
但三十多年的道行
面对这份执著
还是会像狂士在佛像前失去些许自信

我有一个梦想
读者投资时未必照搬书本
读者在与作者思想进行碰撞时
能唤醒其内心的创造力
使读者迸发出灵感
思路喷涌而出

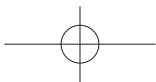
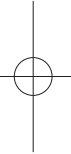
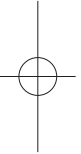


我还有一个梦想
这本书在出版十年之后
依旧有沉甸甸的营养
依旧有可借鉴的价值
所以
对于书中点滴内容都曾反复推敲

一场缘分
我有幸认识了清华大学出版社施猛老师
熬过无数个凌晨
增删改数十回
终于完成这部书稿
感谢施老师一路的相伴和指导
亦感谢清华社排版、设计、发行的各位老师
还要感谢我家人的支持
才能腾出大块的时间笔耕不辍

是日
山色空蒙裹细雨
记为序

作者
壬寅虎年荷月
于湖州·井空里





这是一本主要讲 AI 量化投资的书，与常规量化投资书籍有本质不同。

怀着满满的诚意，AI 技术与金融投资相互契合是本书希望表现的特色。比如第 8 章浓墨重彩地描述了作者在 AI 技术理论上的突破，技术深度和难度极大，且多数是独创技术理论；第 9 章主要介绍小波在金融领域的应用，尚属探索阶段，这章所列举的三个金融案例，具有启发性和探索性的双重价值。本书有以下几个特点。

(1) 阅读体验。按照数学建模的思路和步骤，本书对每个符号的讲解、每个公式的推导尽量深入浅出，推导方法力求别具一格。作者还做了一张表格，对每一章内容的技术难度做了难度系数的评级，用★表示，★数量越多表示对应内容的技术难度越大。所以，读者朋友们可以按照本书的章节顺序依次阅读，还可以根据每一章的难度系数跳跃式阅读，按由浅及深的方式阅读本书，也是没有太大问题的。作者在创作本书的过程中，已经考虑到读者的阅读体验。

(2) 体系化。量化思想、量化实践、量化方法、量化策略、[风险控制](#)、[信息交易系统接入](#)共同构成了量化体系，再由量化投资体系过渡到 AI 量化投资体系，循序渐进，娓娓道来。

(3) 创造性。用鲨鱼猎腥的方法类比人工神经网络的训练过程，用小波分析的方法去识别资本证券炒作的市场特征，而且不管炒作程度多高，都能基于小波多尺度分辨的技术把市场炒作信号筛出来。此外，本书更有大量延伸数学模型的创造。

(4) 稳定性。虽然资本市场是不断演变的，具有很强的时效性，但作者对 AI 算法研究了十几年，发现算法底层理论进步空间极小；有所不同的是，新的应用场景和大算力平台层出不穷，所以本书遴选的大部分实证案例，尽可能规避时效性的限制，刻意摒弃了时效性强的案例，侧重底层逻辑的阐述，致力于在“AI 量化投资”



领域对**不_点变量**的挖掘，而不是追求对新技术、新概念、新场景的探索。

(5) 原创深度。除了极其少数的内容参考了少量文献，本书绝大多数内容皆是原创，让读者少了一份似曾相识或雷同的感受，比如第 10 章里面的“LZ 证券能量守恒定理”，市面上其他著作鲜有与之相似或相同的内容，作者原创性地给出该定理的金融公式，并用实证案例进行了检验。本书几乎没有基础性的量化、AI 知识，作者努力追求全书内容都是硬核干货。此外，作者在写作本书的过程中，在保证描述精准的前提下，努力摒弃那些刻板、索然无味的文字，字里行间透射出灵动鲜活的魅力，让文字活泼而有张力。

(6) 可视化。本书大量采用了计算机绘制的图形图像，比如基于云滴智能技术观测主动管理型基金的风格漂移，凯利公式风控模型的计算机模拟，等等。

(7) 交叉融合。作者推崇不同学科、不同领域知识与实践的融会贯通，即便是表面上看起来风马牛不相及的领域，比如本书里面的“7.5.2 遗传算法与弗洛伊德梦的解析法”“10.3 最速降线：股票利空出尽的快速途径”，等等。严谨的物理学中有很多伟大的发现，前期都是靠猜的，那么，证券投资领域为何就不能“天马行空”地举一反三呢？因为两者的内在机理确实是一致的。

AI 量化投资技术高深，但不是说，金融或者 AI 博士学位是投资盈利的必要条件。事实上，投资圈存在大量高学历但是投资业绩很一般的基金经理或无指引价值的研究员。古人云：三人行必有我师；又云：世事洞明皆学问，人情练达即文章。缺乏投资和 AI 知识的普通投资者，或许亦能敏锐地捕捉到更加有效的“情绪因子”。知识储备程度、时效性程度、灵活应用程度三者基于乘法原则，共同决定了投资回报率，所以纵观本书的叙述结构，是相当开放的，体现了与各位读者探讨的诚意——没有绝对的专业，亦没有绝对的权威。

限于篇幅，本书内容不一定面面俱到；限于作者水平，虽竭尽全力，但是 AI 量化投资需要的知识实在是太浩瀚了，书中难免存在错误和纰漏，敬请读者朋友们不吝指正、勘误。反馈邮箱：wkservice@vip.163.com。

作者

2022 年 7 月于上海



在读者开卷之前，本书作者**着重强调**四点：

- 本书所涉及的任何数据不代表未来收益承诺
- 未来或许能再现，但无法预测
- 市场有风险，谨慎参与，自担投资风险
- 不盲目相信 AI 技术的力量，需辩证与理性看待



本书阅读及学习导引

全书分成上、下两篇。本书涉及金融投资、工程数学、计算机程序开发三个维度的知识。

为了使读者阅读本书时能够获得更好的体验，笔者对本书 10 章内容的难度进行了评级，★越多表示内容难度相对更高。

章号	章标题	难度系数
第 1 章	量化思想	★
第 2 章	量化实践	★★★★
第 3 章	量化方法	★★★★★★★
第 4 章	量化策略	★★★★★
第 5 章	风险控制绕不开凯利公式	★★★★★
第 6 章	交易信息系统外接	★★
第 7 章	遗传算法在黄金投资中的应用	★★★★★★
第 8 章	大规模神经网络及股票非量价复合策略	★★★★★★★★★★
第 9 章	小波分析及金融工程多维度应用	★★★★★★★★★★
第 10 章	前沿研究与探索	★★★★★

本书重心在于跨界融合——AI 技术与投资策略两者的深度融合。这是一本主要讲述 AI 量化投资的书，第 7 章、第 8 章、第 9 章三章内容占据了本书 50% 左右的篇幅，AI 量化投资与量化投资在侧重点上有本质区别。



上篇 量化体系

第 1 章 量化思想

3

1.1	超额 α 实证案例	3
1.1.1	定量构建三级基金池	4
1.1.2	通过“AOA”分析法进行大类资产配置	5
1.1.3	检验正交(独立)的多条投资回报流	6
1.1.4	组建风险收益模型	7
1.1.5	模型延伸：通过 JDBC 驱动连接 MySQL 接口	10
1.2	万物皆是算法	11
1.2.1	生物学的算法属性	11
1.2.2	机器学习算法与数学机械化概述	13
1.3	什么是量化投资	13
1.3.1	量化战胜市场	13
1.3.2	主观投资与量化投资	14
1.3.3	全球证券投资的上升策略	16
1.3.4	经典多因子量化三要素	17
1.3.5	多因子投资拓扑结构	17



1.3.6	算力与 tick 颗粒度	19
1.3.7	算法暴力会人为造成服务器“燃料”短缺	20
1.3.8	量化策略能盈利的底层逻辑	21
1.4	什么是 AI 量化投资	22
1.4.1	AI 量化投资与量化投资的本质区别	22
1.4.2	收益与高流动性显著相关	22
1.5	AI 用于投资策略的三项前提条件	24

第 2 章 量化实践 26

2.1	定量设计现金管理方案	26
2.2	基金科学定投	28
2.2.1	蒙特卡罗模拟的原理	28
2.2.2	误差分析	29
2.2.3	定投钝化	31
2.3	时间分辨率为何深度影响投资收益率	32
2.3.1	时间分辨率	32
2.3.2	投资时间颗粒度与复利理论最大值	33

第 3 章 量化方法 35

3.1	线性相关度	35
3.1.1	计算方法	35
3.1.2	用于宏观经济因子	36
3.1.3	构建不相关投资回报流	38
3.1.4	线性相关度延伸：Spearman 相关度	38
3.2	灰色相关矩阵	40
3.2.1	股票走势的灰色性	40
3.2.2	灰色关联度计算方法	40
3.2.3	用于宏观经济因子	42
3.3	多因子投资	44
3.3.1	因子暴露、因子溢价、因子模型、资产向量图	44
3.3.2	构建多因子投资向量通式	46
3.3.3	单因子溢价	47
3.3.4	双因子溢价	49



3.3.5	因子检验	50
3.3.6	多因子分析的主观性	57
3.4	云模型，不确定性的 AI	59
3.4.1	“云滴”的随机性与模糊性	59
3.4.2	云滴模拟基金的风格漂移	60
3.5	贝叶斯判别法，先验概率的 AI	62
3.5.1	后验概率	62
3.5.2	判别股票的动量反转	63
3.6	K-means 聚类法	67
3.6.1	数学原理	67
3.6.2	中证 1000 成分股量价因子聚类	68
3.6.3	A 股全指个股量价因子聚类	69
3.7	行为金融交易法	70
3.7.1	前景理论	70
3.7.2	面向未来	72
3.7.3	信息交易者与噪声交易者	73
第 4 章 量化策略		75
4.1	量化投资流程与策略分类	75
4.1.1	策略实现流程	75
4.1.2	量化策略分类	76
4.2	指数增强策略	76
4.3	对冲中性策略	78
4.3.1	策略原理	78
4.3.2	期货升贴水	79
4.3.3	ETF 对冲概述	80
4.4	网格交易策略	81
4.4.1	网格交易原理	81
4.4.2	沪深 300 实证案例	83
4.4.3	策略缺点与优化	84
4.5	风险平价策略	86
4.5.1	原始模型	86
4.5.2	常规模型	87



4.5.3 模型求解	90
4.5.4 增强模型	91
4.5.5 杠杆模型	93
4.6 CTA 策略	94
4.7 量化策略的趋势	96

第 5 章 风险控制绕不开凯利公式 97

5.1 普适的凯利公式	97
5.2 胜率和赔率决定冒险投资的仓位	98
5.3 公式推导	99
5.4 计算机动态模拟	100
5.5 凯利投资准则	104

第 6 章 交易信息系统外接 106

6.1 接入门槛	106
6.2 系统外接 IT 架构	107

下篇 AI 方法及投资策略

第 7 章 遗传算法在黄金投资中的应用 111

7.1 AI 算法思想	111
7.1.1 边缘活跃效应	111
7.1.2 AI 算法时代	112
7.1.3 量化投资在我国的发展阶段	113
7.2 遗传算法概述	113
7.2.1 重度参与飞机发动机设计	113
7.2.2 生物学基础	114
7.3 遗传算法关键数学原理	115
7.3.1 编解码	115
7.3.2 复制运算	117



7.4	遗传算法寻优举例	117
7.4.1	编码	118
7.4.2	评价个体适应度	119
7.4.3	新种群复制	119
7.4.4	交配点发生交配	120
7.4.5	基因突变	120
7.5	遗传算法拓展	122
7.5.1	协力进化	122
7.5.2	遗传算法与弗洛伊德梦的解析法	123
7.6	计算机程序设计框架	124
7.6.1	伪代码设计	124
7.6.2	参数设计准则	124
7.6.3	适应度函数调优	126
7.7	黄金现货对冲套利	127
7.7.1	黄金投资重要属性	127
7.7.2	定价要素	128
7.7.3	组建模型	129
7.7.4	求解非线性多约束双目标函数	131
7.7.5	套利策略设计思路	132

第 8 章 大规模神经网络及股票非量价复合策略 134

8.1	人工神经网络概述	134
8.1.1	网络结构	134
8.1.2	激励函数	136
8.2	经典神经网络模型	137
8.2.1	因式分解网络的学习步骤	137
8.2.2	基于视网膜感受视野设计 RBF 神经网络	139
8.3	设计 BP 神经网络	140
8.3.1	鲨鱼嗅闻血腥味原理	140
8.3.2	网络学习案例化解析	142
8.3.3	网络训练多尺度分析	151
8.4	设计 RBF 神经网络	152
8.4.1	梯度训练法	152



8.4.2 性能要素	152
8.5 评价神经网络的泛化能力	153
8.6 神经网络核心技术与策略	154
8.6.1 震荡的成因	154
8.6.2 演绎推理法与绝妙的排列组合算法	155
8.7 设计大规模径向基小波神经网络的深度理论	160
8.7.1 五层网络拓扑结构	160
8.7.2 符号约定	160
8.7.3 组建神经网络数学模型	161
8.7.4 调整伸缩因子 a	162
8.7.5 调整尺度因子 b	164
8.7.6 调整连接权值 w	165
8.7.7 a 、 b 、 w 动态调整量	166
8.7.8 添加网络动量	167
8.7.9 非常规定律	168
8.8 中证 1000 成分股非量价复合策略	169
8.8.1 股价预测的合理性与不合理性	169
8.8.2 多策略原理阐述	170
8.8.3 因子群挖掘	172
8.8.4 股票数据 ETL	174
8.8.5 股票数据归一化和添加白噪声	176
8.8.6 启动训练	176
8.8.7 股价预测准确率	178
8.8.8 交易指令设计	180
8.8.9 为什么有些机器学习量化策略没有成功	180
8.8.10 神经网络在股票优选策略中的独特优势	182

第 9 章 小波分析及金融工程多维度应用 183

9.1 小波分析框架理论	183
9.1.1 数学变换的含义	183
9.1.2 傅里叶变换的局限性在哪里	183
9.1.3 傅里叶变换和小波变换图谱对比	186
9.1.4 小波的定义及特征	187



目录

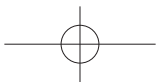
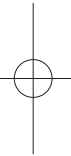
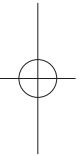
9.2	小波伸缩平移和变换	188
9.2.1	母小波伸缩平移	188
9.2.2	连续小波变换	189
9.3	小波分解与重构	189
9.3.1	平均和细节	189
9.3.2	多尺度分析	191
9.4	市场熵投资策略	195
9.4.1	海森堡测不准定律	195
9.4.2	小波窗函数影响市场动量因子	197
9.4.3	统一论的量化投资体系	199
9.5	抽取震荡中的道琼斯工业指数行情趋势	202
9.6	识别可转债的资本市场炒作行为	203
9.6.1	进可攻，退可守的可转债	203
9.6.2	数据洗涤	204
9.6.3	基于 db6 小波基对炒作信号进行变换	205
9.6.4	通过噪声分贝识别可转债市场炒作行为	207
9.6.5	模型延伸：提升时间分辨率	209

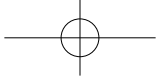
第 10 章 前沿研究与探索 211

10.1	斐波那契投资模型	211
10.1.1	通项式	211
10.1.2	简单是股票策略有效的必要条件	212
10.2	股票能量守恒定律	214
10.2.1	物理方程式	214
10.2.2	实证案例	215
10.3	最速降线：股票利空出尽的快速途径	217
10.3.1	参数方程组	217
10.3.2	延伸理解	218
10.3.3	利空出尽最速衰减规律	219

后记 221

参考文献 223

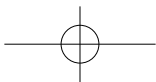
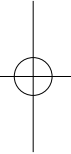
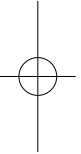
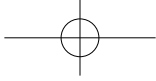


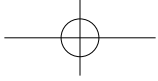


上篇 量化体系

上篇系统介绍了量化体系。笔者概括了上篇部分核心内容和特色内容。

章号	核心计算机技术与数学方程式	量化实证案例独特内容
第 1 章 量化思想	① JDBC 驱动连接 MySQL 接口程序 ② JMatLink 程序包调用混编	① 超额 α 收益案例的因式分解 ② 因子与因子群结构示意图 ③ AI 量化投资与量化投资的本质区别
第 2 章 量化实践	蒙特卡罗计算机模拟求解 π 值	高频 T0 交易与时间极限微分收益率
第 3 章 量化方法	① 本章原创数学方程式： $\cos\theta_i = \beta_i = r_i$ $BCTR = \lambda\beta_{BAPM} + (1-\lambda)\beta_{CAPM}$ ② Spearman 相关度与 Pearson 相关度的共生关系推导	① 灰色关联度矩阵用于宏观经济因子 ② 资产向量图 ③ 云滴模拟基金的风格漂移 ④ 因子的检验：有效性与稳定性
第 4 章 量化策略	从风险结构的视角切入（其他同类著作一般从风险平价角度切入），推导出风险平价的原始模型、常规模型、增强模型和杠杆模型四类进阶数学模型，逻辑层层递进	① 以沪深 300 指数为例，讲解了网络交易的原理 ② 对资产组合的收益进行因式分解
第 5 章 风险控制 绕不开的 凯利公式	凯利公式的计算机动态模拟	揭示现代金融投资的风控理论体系，旨在解决“在胜率和赔率已定的情况下，该用多少资金去冒险”这个“道”层面的问题，非“术”层面的问题
第 6 章 交易信息 系统外接	股票交易的完整 IT 架构，这是量化投资行业能进行自动化交易的基础	基于接入券商系统的视角进行介绍





第1章 量化思想

1.1 超额 α ^① 实证案例

让专业投资人士学习并掌握 AI(artificial intelligence, 人工智能) 技术, 而让 AI 算法工程师学会投资策略, 让具备理工学科背景并想在 AI 量化投资领域发展的读者能踏进 AI 量化投资的快速通道, 这是作者写这本书的愿景。

本节基于公募基金组合, 讲解一个 2020 年的量化实证案例。

目前, A 股市场有超过 5000 只股票, 而基金数量远远超过股票数量, 仅公募基金就有 10000 多只, 加上私募基金就更多了。从这个意义上说, 基金投资的本质是对 5000 多只个股进行筛选和组合, 而在这方面, AI 机器人比基金经理更擅长精密计算和判断, 且更加具备投资自律性。

2019 年 1 月 2 日至 2020 年 2 月 14 日, 某公募指数型基金组合的收益是同期上证指数的 2.6 倍左右, 年化收益率为 40%, 如图 1-1 所示。

以上为非回溯的数据, 而是实证案例。波动风险、回撤均较小, 夏普比率高, 因此, 投资者的持有体验更好。如此高收益是怎么实现的呢? 下面详细加以说明。

① 超额 α 指的是超越大盘指数的收益, 是不受大盘涨跌影响的收益。举例来说, 假设以沪深 300 指数为业绩比较基准, 那么该投资标的超过沪深 300 指数收益的部分就是超额 α 。与此对应的还有 β 收益, 这是一种受到大盘涨跌影响的相对收益, 也叫市场收益。

成立时间: 2019-01-02 高风险

累计净值

收益走势

● 本组合: 1.4470 ● 上证指数(470007): 1.1741



注: 投资有风险, 历史收益不代表未来收益。

图 1-1 量化策略实证案例的业绩走势



1.1.1 定量构建三级基金池

构建公募基金初选池、优选池、核心池三层基金池体系 (见表 1-1)，这跟信息技术中敏捷开发的底层逻辑是一样的，定期动态维护好各个基金池子，以备使用。

表 1-1 三层基金池

初选池	基金分类	股票基金		混合基金		债券基金	
		主动股基	指数股基	偏股混合	偏债混合	纯债债基	混合债基
	对应数量						
优选池	基金分类	主动股基	指数股基	偏股混合	偏债混合	纯债债基	混合债基
	成立年限	≥ 1 年	≥ 1 年	≥ 1 年	≥ 1 年	≥ 1 年	≥ 1 年
	基金规模	≥ 1 亿元	≥ 1 亿元	≥ 4 亿元	≥ 1 亿元	≥ 1 亿元	≥ 1 亿元
	区间 回报	3 月	≥中位数	≥中位数	≥中位数	≥中位数	≥中位数
		12 月	≥中位数	≥中位数	≥中位数	≥中位数	≥中位数
		3 年	≥中位数	≥中位数	≥中位数	≥中位数	≥中位数
	对应数量						
核心池	基金分类	主动股基	指数股基	偏股混合	偏债混合	纯债债基	混合债基
	收益标准差	20%		20%			
	最大回撤	15%		15%	40%	80%	40%
	下行概率	20%		20%	20%		20%
	詹森指数 (Jensen)	25%		25%		20%	
	夏普指数 (Sharpe)	20%		20%			
	跟踪误差		80%				
	信息比率 (IR)		20%				
	卡玛比率 (Calmar)				20%		20%
	索提诺比率 (Sortino)				20%		20%
	复权排序取中位值之上的基金						
	对应数量						

1.1.2 通过“AOA”分析法进行大类资产配置

资产配置遵循“倒三角”的原则，首先确定大类资产方向，其次确定行业龙头个股，或者具体的单只基金。

如果从基础研究起步，预判未来中短期的主力资产配置方向，那将是非常庞大、繁复、系统的金融工程，需要极高的专业度以及大量的财力支撑。

笔者使用一种叫做“AOA 分析法”^①的轻模式，并且结论不会差。

具体原理就是，使用文本挖掘技术对排名前十位券商的分析报告进行深度挖掘与分析，然后对文本挖掘的资产标签簇进行 *K-means* 聚类（后面章节会讲到该技术），取大类资产配置的最大公约数。

这是一种平台思维，在金融领域已经被较为广泛地运用，比如“FOF 基金”，也就是基金中的基金，俗称“母基金”，母基金侧重投资业绩优秀或者策略互补的基金，其本身并不直接对底层标的进行投资。再如“MOM 基金”，也就是基金管理人的管理人，以“人”为中心，侧重选择优秀的基金经理。

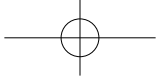
优选各个券商知名分析师的资产配置观点，不仅是为了集合投资观点，还能提高资产配置方向的容错率和鲁棒性。尽管这种方法与“投资成效来自非大众共识”形成一定程度的悖论，但重要的是，知名分析师的一致观点能间接塑造市场走向的预期。金融理论研究表明，市场行为存在随机漫步，凝聚共识能引导市场形成一致性的预期，所以在因子投资的方法里，有专门的“一致预期因子”。

笔者对券商 2019 年底分析报告进行文本挖掘，发现大类资产配置要点如下：

- 三四线城市地产不具有投资价值。
- 房地产销售面临下行压力。
- 买入人口净流入城市的房子。
- 未来 5 年最好的投资机会在中国。
- 美股、外汇、大宗商品存在投资机会。
- 消费 + 投资 + 政府支出扩大，美国经济向好。
- 大数据、云计算、人工智能是高景气度赛道。
-

实际上，几乎所有常规资产配置都是由 A 股（或对应基金）必选消费、A 股医疗、A 股银行、A 股地产、全球黄金、美股、美债、美国地产、全球能源这 9 个类别构成的。

^① 券商每年都撰写大量分析报告，我们用文本挖掘的计算机程序在报告内容里提炼投资观点，笔者把这种定量与定性相结合的基本面研究方法称为“analyst of analyst”（简称“AOA 分析法”）。



不同的时期，9 类资产配置的仓位大相径庭，但是这 9 类资产基本都在配置之列——对于未来行情预期较差的资产品种，会配置极少的仓位，比如 2% 或 1%，然后再逐步调整仓位；而对于中短期具有成长性的资产，则增加仓位。

军工板块也能形成独树一帜的投资回报流（包括稀土题材基金），但是军工行情波段短，股息率和净资产收益率（ROE）的价值评估体系有所不同，因为军工行业不以营利为目的，所以行情节奏难以把握。当一个投资标的的逻辑难以定性或者定量分析的时候，笔者宁愿舍弃这个资产标的，否则量化模型很难保证业绩一致性。

依据上述文本挖掘的结果，并结合基金池选择大类资产配置的行业指数基金，包括上证消费 80、深证红利、标普 500、医疗保健、黄金、美国房地产指数、上证 50、信用债。具体配置逻辑是：

- 之所以选择指数基金是因为费率比较低。
- 最好的投资机会是中国上证 50 指数基金。
- 看好美国经济，美股和美国地产进入大类资产范畴。
- 加入信用债的目的，不是基于业绩回报，跟业绩一点关系都没有——对绝对收益的增长贡献几乎毫无用处。加入信用债，可以在权益类资产行情剧烈波动的时候，平滑波动。
- 行业整体下行，此次不将国内房地产纳入大类资产配置的范围。

大类资产配置的核心是量化择时，量化择时或是量化中难度最大的策略。量化投资经常用到高频交易，不过高频策略通常只负责增厚收益，难以产生跳跃式的投资回报，要赚取丰厚的收益还得依赖择时。

资产组合投资中，对于基金经理或个股的尽职调查是不可或缺的。尽职调查内容主要包括投资策略、投资风格、未来的投资方向等，面谈或者线上沟通都是可以的，但需要有统一的模板。

1.1.3 检验正交（独立）的多条投资回报流

如果大类资产之间的相关度过高，势必会导致资产配置的冗余度过高，比如中证 800 指数与中证 500 指数、沪深 300 指数之间，无论什么时候，相关度大概率都在 95% 以上，这样的大类配置不利于风险控制和收益管理。但是，对于一个景气度高的行业，通过多只基金或者多只股票进行配置是可以的。

2014—2018 年大类资产相关度如表 1-2 所示。为什么要选择 2014 年到 2018 年这 5 年时间作为统计分析的刻度呢？因为一般认为金融周期是 5 年左右。从表 1-2 可以看出，

以上各项大类资产几乎没有相关度，除了美股与石油之间。统计学上规定，线性相关度的绝对值大于 0.7 为强相关，低于 0.3 为弱相关，介于两者之间为存在相关关系。

表 1-2 2014—2018 年大类资产相关度

相关系数	美国地产	上证 50	国内债券	黄金	标普 500	石油
美国地产	1.00					
上证 50	0.12	1.00				
国内债券	0.02	0.04	1.00			
黄金	0.11	-0.01	0.19	1.00		
标普 500	0.59	0.26	-0.04	-0.19	1.00	
石油	0.45	0.25	0.04	-0.06	0.67	1.00

1.1.4 组建风险收益模型

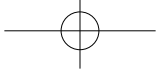
对每个行业指数基金进行仓位上下限的约束，尽量对各个投资回报流进行风险均衡（后面有专门的章节详细介绍），这些会演变成一个非线性多约束的目标函数最大化问题。目标函数就是最大化未来一年的投资风险收益比（定期迭代）。对于模型求解的方式，可以使用人工智能的遗传算法，笔者使用的是 MATLAB 规划问题的工具箱函数，当然还可以使用 Python、Lingo 或者其他计算机语言来求解。这些都是“术”的差别，只是工具不同而已。

如何进行计算机编程并不是本书的重点，在此一笔略过，本书重点落在 AI 量化投资的策略开发上面。本章为了展示实证案例的完整性和系统性给出了执行程序。

我们约定，多约束风险收益比的矩阵变量 $[w_1, w_2, w_3, \dots, w_8]$ 是向量， w_1 向量里面的变量 $A \sim H$ 是当前待求解的仓位数值， I 是常数变量，如表 1-3 所示。看不懂 1.1.4 节内容没关系，不影响本书的后续阅读。

表 1-3 资产配置待求解仓位矩阵

w_1	w_2	w_3	w_4	w_5	w_6	w_7	w_8
A	36.9550	-2.0884	28.5472	-15.4136	20.8561	1.1615	9.3378
B	53.7984	-14.6977	33.9559	-4.2677	27.8571	6.3203	22.6654
C	39.4396	-6.0911	29.5931	-20.4755	23.2514	-0.0415	6.6512



(续表)

W1	W2	W3	W4	W5	W6	W7	W8
D	24.9232	-4.1173	47.3720	-22.5686	24.0211	7.4256	16.1593
E	8.4371	18.1999	12.4197	-0.5497	12.9183	11.9656	9.2314
F	-5.2204	18.5183	3.1396	3.7450	10.3743	5.0798	1.0606
G	7.1466	13.3048	-2.6112	-0.7235	14.0661	9.0367	6.7412
H	0.0000	0.0000	3.1419	6.6689	1.9361	4.3347	4.8999
I	20.5967	-10.3091	7.4844	-23.98878	25.3865	-8.6025	-2.5377

约束条件如下:

- (1) $w1 \cdot w2 \geq I(1)$
- (2) $w1 \cdot w3 \geq I(2)$
- (3) $w1 \cdot w4 \geq I(3)$
- (4) $w1 \cdot w5 \geq I(4)$
- (5) $w1 \cdot w8 \geq 0.1$
- (6) $w1(5) + w1(7) \leq 0.4$
- (7) $w1(1) + w1(2) + \dots + w1(8) = 1$
- (8) $w1(1), w1(2) \cdots w1(8) \geq 0.05$
- (9) $w1(1), w2(2) \cdots w1(8) \leq 0.3$

目标函数为:

$$\text{MAX} [(W1 \cdot W7) / (W1 \cdot W6)]$$

附上 MATLAB 主程序:

```
clear all;
clc;
f=@(x)((-1)*((1.1615*x(1)+6.3203*x(2)+0.0415*x(3)+7.4256*x(4)+11.9
656*x(5)+5.0798*x(6)+9.0367*x(7)+4.3347*x(8))./(20.8561*x(1)+27.85
71*x(2)+23.2514*x(3)+24.0211*x(4)+12.9183*x(5)+10.3743*x(6)+14.0661*x(7)+1.93
61*x(8))));
X0=[0.1;0.1;0.1;0.1;0.1;0.1;0.1;0.1];
A=-1*[36.9550 53.7984 39.4396 24.9232 8.4371 -5.2204 7.1466
0.0000;
-2.0884 -14.6977 -6.0911 -4.1173 18.1999 18.5183 13.3048 0.0000;
28.5472 33.9559 29.5931 47.3720 12.4197 3.1396 -2.6112 3.1419;
-15.4136 -4.2677 -20.4755 -22.5686 -0.5497 3.7450 -0.7235 6.6689;
9.3378 22.6654 6.6512 16.1593 9.2314 1.0606 6.7412
4.8999];
B=1*[20.5967;-10.3091;7.4844;-23.98878;10];
```

```

Aeq=[1 1 1 1 1 1 1 1
0 0 0 0 1 0 1 0];
beq=[1;0.40];
Lb=[0.05;0.05;0.05;0.05;0.05;0.05;0.05;0.05];
Ub=[0.30;0.30;0.30;0.30;0.30;0.30;0.30;0.30];
x= fmincon(f,X0,A,B,Aeq,beq,Lb,Ub)
-1*f(x)

```

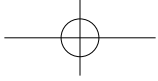
程序算出来标普 500、美国地产和 A 股医药总仓位超过 60%。在 2019 年 1 月 2 日至 2020 年 2 月 14 日的投资收益统计区间内，**近一年的时间内获得 44.7% 的收益，约是同期上证指数涨幅的 2.6 倍。**这种证券投资组合的风险收益比模型之所以在投资上存在较大的胜率，根本原因在于风险和收益存在阶段性错位。

通过这个量化策略的实际案例，我们认识到此量化策略也存在不足：**量化的指标在时间轴上必须具有惯性，否则遇到市场风格突变的行情，回撤通常很大。**在本投资组合中，A 股医药的仓位较重。但是我们发现，2021 年下半年和 2022 年上半年医药股票和相关基金累计跌幅逾 30%，这种市场突变对于量化组合策略冲击较大，模型就会失去原有效力。所以，没有任何一种投资策略是放之四海而皆准的，需要不断迭代。

不得不承认，投资实践中其实还有一些非理性但重要的因素存在，即便基于量化投资模型进行投资也不能忽略这些因素。比如资金量大小不同，会导致投资心态完全不同，又决定了投资风格与投资纪律都会不同。再如，基金投资一般都要经历“募、投、管、退”四个阶段，募集来的投资资金是有固定封闭期的（开放式公募除外），若遇到投资机遇极佳但是濒临临时开放申赎或者封闭期结束，也有可能错失投资机会。凡此种种，都会影响投资业绩。

一般而言，使用对冲手段可增强抵御风险的能力。以房贷为例，如果一个人每个月还款金额都非常大，且目前又实行贷款市场报价利率（LPR），若同时购买国债期货，理论上，能对冲 LPR 大幅上涨带来的还贷成本增加。不过，风险对冲是需要成本的，目前我国资本市场还没有丰富的做空工具。

此外，量化策略相比传统的价值投资理念，顶层设计思路也是不一样的。价值投资践行者芒格一旦买入股票就近乎将其忘记，直至上涨了很多倍才卖出。量化投资更偏向利用数学和计算机进行投资决策，比如在投资回报区间内，传统价值投资策略可能期望至少上涨 30%，但是量化策略满足于在一只基金或者个股上赚 10% 的利润，再在另外一只基金或股票上赚 10% 的利润，然后又在其他基金或股票上再赚 10% 的利润，这样就能达到 30% 的收益目标，集腋成裘。这两者的投资理念有根本的不同。



1.1.5 模型延伸：通过 JDBC 驱动连接 MySQL 接口

本节内容为选修。

公开市场上存在庞大的交易数据、财务数据和市场数据。MATLAB 软件主要用于矩阵计算、电路设计、计算机仿真等方面，还可以用来辅助设计半导体芯片，然而它并不太擅长数据分析和挖掘——MATLAB 是基于内存进行数据交互的，处理数据能力有限。由于 MATLAB 拥有强大的计算引擎，目前一部分在线量化平台使用 MATLAB 语言。

为了解决这个问题，笔者探索使用 JDBC 驱动技术，实现 MATLAB 与数据库系统之间实时通信，这需要编写接口驱动程序并调试。笔者以连接 MySQL 数据库为例，将接口驱动程序（已经调通）分享给各位读者。

```
% 在Matlab用JDBC驱动连接MySQL，可调用现有的connector驱动jar包学习方法
% 与连接其他主流数据库的驱动程序相比，仅连接符url和driver不同
% 1.下载mysql的jdbc驱动mysql-connector-java-5.1.7-bin.jar
% 2.在command windows中输入: edit classpath.txt 打开路径配置文件，在文件最后一行添加mysql-connector-java-5.1.7-bin.jar的完整路径
% 如：复制路径 E:\MySQL_Test\mysql-connector-java-5.1.7-bin.jar 到 classpath.txt文件的最后一行
% 3.在MySQL中新建一个登录名root，并设置密码为870807
% 4.采用如下代码在Matlab中访问MySQL
clear;close all;
databaseurl = 'jdbc:mysql://localhost:3306/experience'
driver = 'com.mysql.jdbc.Driver';
username = 'root';           % 登录名
password = '870807';        % 密码
databasename = 'experience'; % 数据库名称
conn = database(databasename,username,password,driver,databaseurl)
tic
% 读取数据库
curs=exec(conn,'select * from `order`')
% 将SQL执行结果返回到MATLAB工作空间
curs=fetch(curs);
% 得到SQL执行结果中的数据
data=curs.Data;
toc
% 关闭游标对象
close(curs);
% 关闭数据库连接对象
close(conn);
```

此外，如果 AI 量化计算平台的底层语言使用 Java 编写，再用先进的方法调用 MATLAB 的算力，可以通过 MATLAB engine 的方式实现。传统 COM 组件方案的

缺陷在于 MATLAB 中的很多复杂函数都不能正常使用。

由于 Java 移植性好, 适合搭建量化计算集群的平台, 而 MATLAB 计算引擎强大, 那么要实现 Java 和 MATLAB 的混编, 在 Java 中调用 JMatLink 包就可以使用了, 具体标准流程可以登录 JMatLink 相关网站进行学习, 其功能性远高于传统 COM 组件, 但是必须装上跟 Java 版本兼容的 MATLAB 软件。

1.2 万物皆是算法

1.2.1 生物学的算法属性

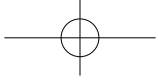
第二次世界大战期间, 意大利舰队驻扎在塔兰托海港, 控制着地中海的北面, 影响着北非的战局。在英国皇家海军偷袭塔兰托之前, 英国人通过情报人员找到了一张如图 1-2 所示的塔兰托港的平面黑白地图照片。



图 1-2 塔兰托港的平面地图照片——奇怪的“白斑”

照片上, 各种防御措施清晰可见, 唯有图 1-2 中连续的“白斑”实在令人费解。英国人起初猜想这些“白斑”是冲洗黑白照片时留下来的, 但是仔细辨别, 发现这些“白斑”排列规则, 显然是人为的。

英国人根据“白斑”进行计算和分析, 判定这些“白斑”是拦截偷袭攻击机的钢丝气球。顶端拴着悬浮的大气球, 一端用坚韧的钢丝连接地面, 这对于急速飞行的战斗机来说就是梦魇, 它们可以将正在飞行的战斗机撕成两半。英国人经过精密计算得



出，相邻两个气球之间的距离为 270 米，而英国“剑鱼”式攻击机可以在这个间距中安全通过。

1940 年 11 月 11 日，英国军队从光辉号航空母舰上分两个批次共出动 21 架“剑鱼”攻击机偷袭塔兰托港。意大利受到重创，同时改变了地中海地区海军力量的格局，优势朝英国倾斜。在这次战役中，对“白斑”的辨识和分析起到了至关重要的作用，其准确度的高低决定了偷袭是否可行。

这是一个典型的信息分析案例。

大家都知道一个常识，数据 (data) 是简单的、纯粹的事实和数字。广义的数据既包括纯粹的数据，也包括文本信息，其自身的使用价值较小。比如，某公司有 20% 的产品销往上海市。要想使这个数据更有价值，就必须将其与其他数据横向或纵向联系起来——如“这批产品被 16 ~ 25 岁的青年男女买走”，将其处理为信息，这条信息对于客户画像是有帮助的。

所谓信息 (information) 就是具有特定用途和使用价值的信息。知识 (knowledge) 是将信息与已知信息进行比较后得出的结论。为什么现代人崇尚知识，而不一定崇尚传统的经验呢？经验也可以成为有用的知识啊。因为经验一般都是感性的总结，不方便准确传承，更难以更新完善，所以经验常常过时。黑格尔曾说过，人类从历史中吸取的唯一教训，就是人类不会从历史中吸取教训。

无论是信息还是知识，都是从大量数据中发掘出的价值。**数据尤其是海量数据之所以存在巨大的商业价值，是因为在数据中蕴含着丰富的信息。**比如有的线上零售平台，它能依据算法捕捉女性的生理周期，再根据生理期阶段的不同，推测女性的不同心情和不同需求，有针对性地推送客户可能感兴趣的物品。高阶的机器学习算法亦是如此。

在零售领域，客户流失率曲线会严格遵循“幂律”定理，残差比率不超过 0.03%。从社会现象映射到数学定理，这当然也是一种算法。

孔子是我国文化圣人，代表着我国源远流长的文化“符号”。把一个自然人抽象成一个符号、一种精神、一个信仰，这种映射亦是一种算法——算法不仅限于狭义的加减乘除、开方、对数等。三国时期，诸葛亮依据天象推算出“连窥天河，有云如蛇，主大雾”，也是一种算法，从现代视角来看，这是一种非结构化信息的算法。甚至，少数人彼此之间的算计都是一种算法。所以一切皆是算法。

人类群体行为也是算法，最底层的是动物原始意识的算法，再往上是物种自然选择的算法，最后是孟德尔遗传定律的基因层面的算法，多层次的算法构成了群体遗传学。

1.2.2 机器学习算法与数学机械化概述

AI 智能算法是人工智能技术的冠上明珠，它深度依赖机器学习。

人类会自发使用高效的机器学习算法代替人类脑力劳动，重新组织生产关系。人类习性总是朝着“变懒”的方向进化。同样功能的互联网产品，若其中一个互联网产品比另外一个产品需要多戳一下电子屏幕，这样的互联网产品必将被市场淘汰。

从各类社交媒体、购物网站和地图软件等所形成的海量数据中能萃取很多信息和知识。数据规模增大之后，出现了几百、几千 TOPS 算力 (tera operations per second, 1TOPS 代表处理器每秒进行一万亿次运算) 的 AI 芯片，这些 AI 芯片都有支持机器深度学习的能力。

不论是粒子群算法还是蚁群算法，不论是遗传算法还是免疫算法，不论是神经网络算法还是模拟退火算法，本质上都是机器学习。机器学习具有一个显著特征：机械性。我国吴文俊院士在《数学机械化》(科学出版社，2003 年 3 月)一书中提到：“数学机械化是指数学在运算和证明过程中，只要前进了一步，都会有确定的、必然的下一步，直到到达结论，整个过程按照既定的刻板规律进行。”

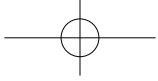
计算机将解析几何转化成向量进行证明，这是吴文俊院士于 20 世纪 70—80 年代首创的，这是数学机械化的一个成功应用。再如，线性代数里维度很大的矩阵求逆，虽然没有多大技术含量，但是人工求解着实烦琐，大算力的处理器能够按照数学机械化的指令快速求解。

可见，智能算法、机器学习和数学机械化大抵相同。本书讲解投资中的量化交易和 AI 机器学习算法是如何高胜率战胜非理性市场的。A 股市场上超过 70% 的资金来自个人投资者，机器人是严格自律的，而个人投资者是非理性的，无法戒贪、戒嗔、戒痴。

1.3 什么是量化投资

1.3.1 量化战胜市场

真正意义上的量化投资，笔者觉得应该是 1952 年由哈里·马科维茨 (Harry M. Markowitz) 提出来的，他在学术论文《资产选择：有效的多样化》中，首次应用资产组合报酬的均值和方差这两个数学概念，创立了“均值—方差模型”，从数学的角度明确地定义了投资者偏好，才使金融研究逐步摆脱了早期较为模糊的“文学性”特征。



为此，哈里·马科维茨、威廉·夏普、默顿·米勒在 1990 年共同获得了诺贝尔经济学奖，获奖理由是：对现代金融经济学理论的开拓性研究，为投资者、股东及金融专家们提供了衡量不同金融资产投资的风险和收益的工具，以估计、预测股票、债券等证券的价格。

关于量化投资，在金融圈还有一个令人津津乐道的故事。

1958 年，在拉斯维加斯赌场，一个叫做爱德华·索普的年轻人安静地坐在 21 点牌桌前，按照自己的节奏下注，如同一台机器。

很快，他就用几千美元赢了几万美元。这时候，赌场经理走过来，身边带了几个保镖，说：“你不是在赌博，你是在算牌，我们要把你赶出赌场！”这是索普第三次被驱逐出赌场。到了 1964 年，爱德华·索普已经成为所有赌场最不受欢迎的人。

索普又把算牌的思路用于购买股票。他运用大数定律进行投资。一只股票在明天能涨还是能跌，是说不好的，但是他可以推算股票涨跌一定幅度的概率，购买多只股票、多次购买股票就会收获超过 50% 的胜率。也就是说，如果股票的波动率是随机的，那么它就是可以量化的。这个发现，奠定了之后量化投资的基础。

索普的兴趣在于研究数学中的漏洞，然后将其写成书籍，他更像一位教授。1961 年，一位记者采访了他，发表了文章《数学家破解 21 点密码》，使索普一夜成名。1967 年，他和希恩·T. 卡索夫合写书籍《战胜一切市场的人》，再次名声大噪。1969 年 11 月 3 日，索普与杰伊·里根合伙创立了一家基金公司——可转换对冲合伙基金，1975 年该公司改名为普林斯顿—纽波特合伙公司。里根在普林斯顿管理基金，索普继续在纽波特的加州大学当数学教授，负责策略开发。

从 1971 到 1974 年，索普的对冲基金分别取得 26.7%、12%、6.5% 和 9% 的正收益。更让人惊讶的是，在最初的 6 年时间，索普的对冲基金只有 1 个月出现了亏损，其他时间都是盈利的。其中，1973 年开始的美国“漂亮 50”^① 泡沫崩溃导致美股平均下跌 48.2%。但是，索普基金采用中性套利策略，市场下跌和中性策略关联度较低，对所有的头寸都会进行对冲，剥离了市场 β 风险。这种惊人的稳定性促使中性策略开始发扬光大，也奠定了爱德华·索普在量化投资领域的行业影响力。

1.3.2 主观投资与量化投资

如果说爱德华·索普是量化行业的鼻祖，那么詹姆斯·哈里斯·西蒙斯就是华尔街的“异类”。

^① “漂亮 50” (nifty fifty) 是美国股票投资史上特定阶段出现的一个非正式术语，用来指 20 世纪 60 年代和 20 世纪 70 年代在纽约证券交易所交易的 50 只备受追捧的大盘股。

西蒙斯也是数学家，并在 1968 年开始领导纽约州立石溪大学的数学系。

1974 年，他与陈省身联合发表了论文《典型群和几何不变式》，创立了著名的 Chern-Simons 理论，这个理论被广泛应用到从超引力到黑洞的各个领域，并因此摘得数学界的皇冠——全美维布伦 (Veblen) 奖。就在其数学研究成就斐然的时候，1978 年，西蒙斯离开石溪大学，全职创立了一家外汇交易公司，1982 年该公司更名为“文艺复兴科技公司”。

大奖章基金成立于 1988 年 3 月，是文艺复兴公司的第一只基金，基金经理是两位著名的数学家：西蒙斯和埃克斯。他们都获得过 Veblen 奖，这便是大奖章基金名称的由来。

大奖章基金投资标的必须符合以下三个条件。

- 能在公开市场进行交易。
- 足够的流动性。
- 适合用数学模型和程序自动交易。

形成大奖章基金交易策略分以下三个步骤。

- 从历史价格数据中找到异常的信号。
- 用统计学计算和再现异常信号有无显著性。
- 观测价格变化是否符合异常信号统计学特征。

西蒙斯的精力集中在短线交易上 (笔者注：金融数据是时序数据，短线在动量上有连续性)，他几乎不雇用华尔街的分析师，仅通过数据搭建数学模型，寻找市场存在的微利机会并快速加杠杆交易获利。一般由交易员通过计算机自动交易，减少人为干预；倘若遇到风格突变的重大行情，亦会转到人工交易流程。西蒙斯创立了较为完整的量化策略框架体系。

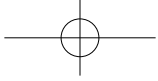
2005 年，西蒙斯成为全球收入最高的对冲基金经理，净赚 15 亿美元，差不多是索罗斯同期收益的 2 倍。从 1988 年开始，西蒙斯所掌管的大奖章基金年均回报率高达 35%，此后 15 年资产从未减少过。大奖章基金的管理费率为 5%，高出同业水平 2%。

2009 年 10 月 10 日，西蒙斯宣布他于 2010 年 1 月 1 日退休，保留公司荣誉主席职位。2020 年 4 月 6 日，西蒙斯以 1400 亿元人民币财富名列《胡润全球百强企业家》第 45 位。

站在巨人的肩膀上，笔者多维度比较了主观投资与量化投资的区别，如表 1-4 所示。

表 1-4 主观投资与量化投资比较

比较要素	主观投资	量化投资
代表人物	沃伦·巴菲特	詹姆斯·哈里斯·西蒙斯
身份	投资家	投资家、数学家



(续表)

比较要素	主观投资	量化投资
投资机构	伯克希尔·哈撒维	文艺复兴科技公司
2008 年次贷危机投资回报率	-15%	80%
过去 20 年年化收益率	20%	35%
投资风格	深度研究、集中持仓、长期持有	挖掘因子、分散投资、全频段收益

从长期来看,“小概率小回报”(散户)比不上“大概率小回报”(量化投资)的积少成多。如果量化策略每天的收益率达到 0.1%,则一年的收益率可以达到 28%,这便是量化策略众毛攒裘的道理。当然,“大概率大回报”(巴菲特)另当别论。量化投资未来还有很长的红利期,以策略清晰一致、行情涨跌都有盈利机会等优势,正强势崛起,人才需求缺口也会越来越大。

1.3.3 全球证券投资的上升策略

海量数据叠加 AI 算法,我国投资领域正在经历一次深层次的转变。

很多证券类的私募基金都在大力升级硬件设备,比如 AI 算法所用的服务器,这既是一种符合趋势的前瞻性战略举措,也是一种被动的防御策略。投资领域的“跑马圈地”,会加速淘汰没有 AI 运算能力的中小投资机构。

一方面,中小量化投资机构在资金募集方面毫无优势,资产管理规模越小,越募集不到充足的资金。另一方面,既然一些机构可以通过量化去做策略,其他机构也可以用量化去收割,就像一个加密一个解密。一旦在投资领域建立某种规则,就会出现另一种规则努力去打破或者跳过当前约束——另类的“勒夏特列原理”^①,大思路是一致的。

2018 年以后,全球基金规模排在前十位的对冲基金里,量化基金占 8 家左右。从 2017 年到 2020 年,我国量化基金尤其是私募基金快速增长,资产管理规模的年复合增长率超过 60%。其中,2019 年资产管理规模增至 2820 亿元,同比增长 88%;2020 年资产管理规模猛增至 8670 亿元,同比增长 204%。

量化策略已经成为全球证券投资的主流策略之一。私募基金比公募基金发展更快。对于公募股票型基金,我国证监会要求股票仓位不能低于 80%,无论是在牛市还是在熊市。但是,私募基金可以选择在二级市场行情不好的时候空仓,灵活很多。

^① 勒夏特列原理(Le Chatelier's principle),又名化学平衡移动原理,由法国化学家勒夏特列于 1888 年发现,其具体内容为:如果改变可逆反应的条件(如浓度、压强、温度等),化学平衡就会被破坏,并向减弱这种改变的方向移动。

1.3.4 经典多因子量化三要素

因子投资通常是量化投资机构的主力策略。

各家量化投资机构的策略大同小异，区别在于算力和因子不同。算力不同会导致交易程序的低延时程度不同；风险因子和因子群（因子组合）不同决定了策略的收益不同。

多因子投资由算力、算法和因子群三个要素构成，如图 1-3 所示。

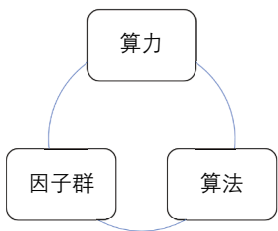


图 1-3 量化投资三要素

1.3.5 多因子投资拓扑结构

量化投资利用数学和计算机进行决策和交易。量化选股是量化策略里经常使用的策略，该策略通过算法不断挖掘因子，然后将因子组合形成“因子群”（也叫“因子组合”），并能按次序输出股票清单，满足这些因子群的股票按照权重逐个自动被程序买入，原有不满足“因子群”的股票则被系统自动卖出。

所谓因子，就是能够对不同资产的收益率进行解释的特征和特征值，比如高换手率对于流动性的影响；再如每周四 A 股下跌概率更高一些；又如个股利空出来之后，大概率不会只下跌一个交易日，会有多个交易日的下跌才可能利空出尽……这些结构性或者非结构性的信息都可以通过大数据洗涤，转化成“因子”。计算机利用因子组合形成投资策略，再以不同权重叠加（信号融合）的方式，使多个策略共同构成一只基金的量化策略。

常见因子分类和特征如表 1-5 所示，建议读者认真阅读，以理解运用。

表 1-5 常见分类因子和特征

序号	因子分类	特征描述
1	情绪因子	聊天笑脸及哭脸数量、VIX 恐慌指数、热搜、全民上网总时长
2	价值因子	市盈率（动态、静态）、市净率、现金流、应收应付账款等财务指标