揭开量化交易神秘面纱
运用前沿统计与机器学习模型研究中国期货市场

附赠
完整的数据集以及
分析用的
R和C++代码！

中国期货市场量化交易

（R与C++版）

李尉◎著

由浅入深
从简单线性模型
到复杂度决策树与
深度学习模型

系统全面
从分笔数据处理
到最终动态投资组合管理

贴近实践
本书策略用于实盘交易
多年表现稳定

看得懂 | 学得会 | 用得上

清华大学出版社

中国期货市场量化交易

（R与C++版）

李尉◎著

清华大学出版社
北京

内 容 简 介

本书主要介绍如何运用统计分析和机器学习等方法对中国期货市场量化交易进行建模分析。不仅覆盖了最基础的数据获取、数据清理、因子提取、模型构造以及最后的动态投资组合优化、C++ 编程实现等方面，而且有丰富的代码方便读者临摹学习和修改提升。本书中的数据首先是交易所最原始的期货分笔数据，在此基础上整合成 5 分钟 K 线，然后再计算预测因子，最后套入统计预测模型。在交易层面，采用严谨的滚动优化方式，充分考虑了滑点和手续费，严格测试。另外本书还覆盖了中低频的趋势策略以及高频的短趋势策略，最后也详细介绍了跨期套利策略，以及对读者择业就业的建议。

本书内容的广度和深度都是国内市场上少见的，适合相关专业人士和感兴趣的投资爱好者阅读，如高校数理类和经管类师生及证券、期货、私募证券、公募基金等量化交易相关从业人员，以及对机器学习在金融方面运用的相关人士和对量化交易感兴趣的各行各业人士。

图书在版编目(CIP)数据

中国期货市场量化交易 : R与C++版 / 李尉著. — 北京：清华大学出版社，2018
ISBN 978-7-302-50322-4

Ⅰ. ①中… Ⅱ. ①李… Ⅲ. ①期货市场－研究－中国 Ⅳ. ①F832.5

中国版本图书馆 CIP 数据核字（2018）第 114984 号

责任编辑： 刘志彬
封面设计： 汉风唐韵
版式设计： 方加青
责任校对： 宋玉莲
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：010-62770175　　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： 三河市国英印务有限公司
经　　销： 全国新华书店
开　　本： 170mm×240mm　　**印　　张：** 19.75　　**字　　数：** 319 千字
版　　次： 2018 年 11 月第 1 版　　**印　　次：** 2018 年 11 月第 1 次印刷
定　　价： 89.00 元

产品编号：078896-01

前言

期货市场是一个有着悠久历史的金融市场，早在几百年前，芝加哥一带的农民聚集在一起，商量一个有中央结算性质的交易场所，最终成立了芝加哥期货交易所。后来随着电子计算机技术的发展，期货交易所日益电子化，交易更为便捷，交易大厅的交易员也逐渐演变成计算机前的量化交易员和程序员。

相比期货交易，量化交易是一个更新鲜的概念。传统的期货交易有很多技术分析的书籍，如经典的《日本蜡烛图技术》《期货市场技术分析》等，一般更着重于使用技术图表分析 K 线形态，从而给交易员提升买卖点位，辅助交易员主观交易。

然而，量化交易不大一样，或者说是期货交易的升华。量化交易更多是运用现代统计学模型，包括机器学习、深度学习等模型来预测市场的价格变化，从而编写计算机程序，实现自动交易。更广泛地说，投资的整个过程，包括品种的选择、价格变化的预测、投资组合权重的分配、最小化交易成本地下单等，都可以使用相应的量化模型来分析，并且提供一套系统性、科学性的测试方法。因此，量化交易跟传统意义上辅助交易员下单的技术分析还是很不一样的。

对于一些基本面信息，本质上也可以融入量化模型中，因此，量化分析和基本面分析并不矛盾。并且现在市场上也有很多期货和股票方面的基本面量化的书籍供读者参考翻阅。

目前，国内期货市场蓬勃发展，量化交易方兴未艾。然而，目前国内很多私募量化基金交易的期货策略都是传统的程序化交易方法，与国外基于统计分析、机器学习模型的方法相比存在较大差距。然而国内却没有有关方面的书籍，即使有也是在股票投资方面，期货方面仍属空白。因考虑到广大理工科学生和科研人员对金融量化交易有着极大的热情，且本人有国内外期货量化交易多年

的经验，于是写作了本书。

本书特色

1. 国内率先系统性运用统计和机器学习模型研究中国期货市场的书

国外用机器学习模型研究股票与期货市场的书确实存在，但比较新，如《Machine Trading: Deploying Computer Algorithms to Conquer the Markets》，主要运用Matlab，分析的主要是美国股票市场日线数据。本书分析的是国内期货市场，使用的是分笔数据和5分钟K线数据，频率上要比市场上同类书籍高出不少。另外本书的模型都是本人实战多年的成果，有着良好的实盘交易记录，并且还给出了研究用的R代码和实盘用的C++代码，方便读者学习。能做到这点的，市面上无论中国还是美国，以本人的经验看，尚不存在。

2. 理论结合实际，由浅入深，娓娓道来

本书从最基本的分笔数据出发，如从如何获得数据、如何合成K线等，到最后的C++实盘交易程序，应该说量化交易的内容都有所涵盖。从最基本的基于买卖规则的策略，到最后基于深度学习、增强学习的策略都有所涉及，而且有详细的R和C++代码，方便大家自主学习。本人也有着丰富的国内外量化交易经验，不仅在美国对冲基金公司全职工作过，而且也在国内多家期货公司和私募基金工作过。本书里面的代码经历过多年实盘交易的检验，另外也会穿插介绍本人的职场经历，可以供各位参考。

3. 覆盖高频与中低频交易

绝大多数的量化交易书籍都不会涉及高频交易，本书却给出了研究高频交易模型的框架，同时检验了多种经典的机器学习预测模型。一般来说，相对于中低频交易，高频交易数据量更大，就可以训练更复杂的模型，因此本书也探讨了很多非线性的模型。但对于中低频交易的训练，还是以传统线性模型为主。

本书内容及体系结构

第1章　期货基本策略概要。简单介绍了目前国内流行的股票对冲、商品CTA、高频交易等策略，以及常见的程序化交易平台，对比了R、Python、Matlab等常见的分析语言，并且对全书进行了概括性的介绍，结合本人的经历发表了对国内量化交易市场的看法。

第 2 章　数据处理。详细介绍了国内商品期货分笔数据的数据结构、获取的方式、处理的方法等，以及如何从分笔数据合成 K 线数据，如何提取主力合约，如何编写更高效率的处理程序等。其中包括 R 与 C++ 相结合的编程方式，如何在 R 里面编译及调用 C++ 程序，如何使用多核并行计算等，而且有详细的 R 与 C++ 代码，为以后的建模做准备。

第 3 章　预测因子。任何模型本质上都是因子的组合方式。机器学习模型又被称为统计预测模型，因此里面用到的因子自然被称为预测因子，当然也有人称为特征因子。本章介绍了构造因子的方法，给出了一些常用的因子，并且还给出了测试因子的基本方法。这里的因子既有基于 K 线信息的因子，也有基于分笔数据的高频因子，方便各种策略使用。

第 4 章　基础统计模型。本章在第 3 章的基础上，运用一些经典的统计模型进行预测分析，并且使用了训练集、验证集和测试集的概念，严谨建模。本章使用的模型以线性模型为主，因为对于绝大多数情况，采用线性模型已经足够了。在模型测评方面，采用样本外的 R^2 作为主要依据，这种方法跟样本内的 R^2 和调整后的平方都不一样。

第 5 章　复杂统计模型与机器学习。本章讨论了更为复杂的统计模型，一般也被称为机器学习模型，包括决策树、随机森林、神经网络、深度学习等，并且对比了不同模型之间的表现。由于金融数据的高噪声、高维度特征，因此复杂的模型很多时候未必会比简单的模型更好。在中低频交易中，如果条件允许，花更多精力收集信息或许更为有效。

第 6 章　从预测到交易。有了预测模型之后，还要落实到交易才有意义。把预测结果转成交易信号有很多方法，本章会进行比较。当然还要结合品种的买卖价差和手续费，以及交易的频率等。对于股票配置型的策略和期货择时型的策略，会有不一样的处理方法。

第 7 章　策略模型深化。本章是对前面几章的总结和提炼，主要是在结果一致的情况下探讨一些提高计算速度的方法，从而提高研究效率。很多时候，量化研究过程需要很多次的搜索、迭代等运算，这会消耗大量时间。如果能提高计算速度，那么就可以大大提高研究的效率。事实上，最近神经网络、深度学习等方法重新流行起来，更多是依靠 GPU 等计算技术的发展。

第 8 章　投资组合优化。有了交易信号和资金曲线之后，下一步就是对各

个策略、各个品种的投资组合进行优化工作。事实上，这部分工作也可以用量化模型来完成。与之前的统计预测、机器学习不同，这部分更多是传统的运筹优化方面的模型，如均值－方差模型、Black-Litterman模型等。本章对比不同的投资组合优化的方法，并给出测评的结果及相关的程序。

第9章　投资组合优化深入研究。本章主要介绍了风险评价策略和增强学习（近似动态规划）等在投资组合里面的应用。其中近似动态规划属于动态投资组合优化的内容。另外，本章也介绍了策略的滚动优化和动态调整，对比了滚动优化和全局最优化的结果。事实上，如果处理得当，滚动优化可以取得比全局最优更好的效果。

第10章　C++实现策略。本章主要介绍了如何把R语言转换成C++，从而实现自动交易。本章还介绍了CTP接口的基本原理，以及转换策略的基本步骤，包括处理行情数据、K线数据、计算指标、计算仓位、合并策略等。本章主要是基于Linux的C++编程，系统默认是Ubuntu 16.04 LTS，读者掌握后就能自主编写全自动交易程序了。

第11章　实盘交易管理。上一章介绍了用C++实现实盘交易的程序。事实上交易过程中其实会遇到各种各样的问题，特别是自己用C++写程序，各类错误都要自己调试改正。本章系统介绍实盘交易会遇到的各种问题，并给出相应的解决方案。

第12章　套利交易。前面介绍的都是关于投机型趋势策略。因为没有做空的限制，所以商品期货从本质上都可以交易。本章就专门讨论套利类的策略，先从最基本的跨期套利开始，然后再简单介绍一些跨品种套利。

第13章　求职与工作。前面章节讲的都是量化交易研究与技术方面的问题。现实中，我们技术人员还要去工作，无论是担任期货的资管还是从事私募证券投资基金。因此，在学习了前面的知识之后，本章有关求职与工作相关的话题，作者有些经验可以和读者们分享。

本书读者对象

- 数学、统计、信息与计算科学、计算机、金融工程等专业本科生、研究生
- 高校数理类和经管类教师和科研人员

- 证券、期货、私募证券、公募基金等量化交易相关从业人员
- 对量化交易感兴趣的各行各业人士
- 对机器学习在金融方面运用的相关人士
- 人工智能方面想从事量化交易的相关人士
- 学习R语言或C++的相关人士
- 其他对中国期货市场量化交易感兴趣的人

感　谢

本书写作过程中，本人的妻子张妮洁女士正处于怀孕中，还要不断给本人写作提供各种各样的帮助。在此，本人表示对妻子衷心的祝愿。希望我的妻子能一切顺利，同时希望我们的宝宝能平安出生，成为新一代的“baby quant”。对于本书的更新和未来进展，以及相关程序、数据资料的获取，本人会在知乎“baby quant 谈量化金融”专栏里发布，敬请各位读者留意。

2018 年 1 月

baby quant

目录

第一章

期货基本策略概要

本章主要介绍期货量化交易策略的基本类型及发展的历程。目前国内的期货品种已经有数十个，其中金融期货包括股指期货和国债期货。商品期货则覆盖农产品、能源化工、有色金属、黑色金属、贵金属等各个板块。曾经股指期货的交易额占全部期货品种的90%，但2015年股指交易受限制之后，股指交易量下降了99%。目前的期货市场由商品主导，特别是螺纹钢、铁矿石等黑色系商品。下面分别介绍各品种对应的策略。

1.1 股指日内策略

在 2012-2015 年，最受欢迎的期货量化策略是股指日内策略。国内有很多期货程序化交易平台，如 TB 开拓者、金字塔交易系统等，很多本土的量化交易团队都会使用这些平台。这些平台的优点主要是回测、优化、模拟、交易都是同一个程序，不需要修改，而且平台会自动维护数据库，另外只需要支付很低的费用（如金字塔当时是 1800 元 / 年）就可以实现自动交易。以至于很多初创公司都会使用这类平台。

1.1.1 日内策略简单介绍

这类平台的期货历史数据主要是连续合约和指数合约。所谓连续合约，比如螺纹钢现在成交量最大的是 rb1305，那么连续合约对应的数据就是 rb1305 的数据，如果过了一段时间螺纹钢成交量切换到 rb1310，那么连续合约的数据就是 rb1310。所谓指数合约，指的是该品种上市的所有合约的加权合约。因此，实际上并不存在指数合约这个交易标的。之所以使用指数合约，是因为商品策略一般是隔夜策略，而连续合约在换月上会有较大跳空，回测的时候不准确，因此使用指数合约可以缓解跳空带来的影响。

我们可以看一个连续合约跳空的例子，如图 1-1 所示。

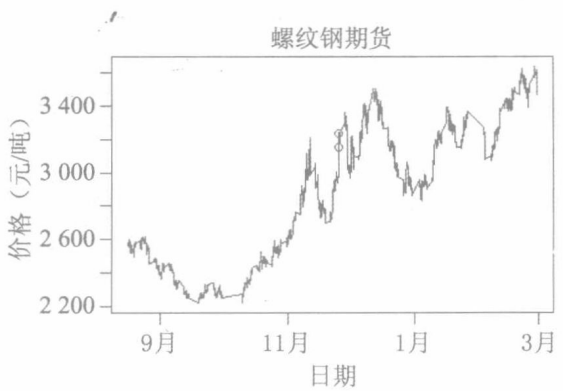

图 1-1 螺纹钢连续合约

图 1-1 中价格曲线的两个圆圈表示合约换月的日子，可以看到合约价格有较大的跳空。如果回测的时候没有注意到，就很可能捕捉到这个虚假盈利。如果要更精确地进行回测，则需要在换月之前平仓。第三方平台要将此写入程序里，不是那么的方便，如果用 R 语言等更通用的统计分析语言则方便得多。

有些人可能会问，现在很多微信号和淘宝店铺都有销售基于第三方平台的程序化交易策略，售价几十元到几百元不等，资金曲线也挺好看，既然如此，为何还要使用这么复杂的基于机器学习的量化交易呢？

事实上，很多上述策略设置了极低的手续费，比如万分之一，而且没有设置滑点，只能用最新价交易，这对商品来说每交易一次其实还能获得不少便宜。另外它们使用的是指数合约，对跳空没做专门处理，很可能策略捕捉的是不存在的行情。

正是因为使用第三方平台写隔夜策略有这些麻烦，所以很多人喜欢做日内策略。日内策略由于强制性日内平仓，因此就不会有赚取换月虚假跳空获利的情况。然而，国内上市期货品种虽然大约有 50 个，但绝大多数不适合日内策略。日内平仓可以看作一种风险控制手段，规避了隔夜的风险，放弃了潜在的利润，也避免了潜在的损失。然而，要规避这种风险是需要付出代价的。每天的平仓操作就是代价：一来需要支付手续费；二来需要牺牲滑点。因为平仓操作是必须成交的，所以一般会主动成交。

说到成交方式，一般有两种：一种称为被动成交；另一种称为主动成交。一般来说，如果需要立即成交，则会采取主动成交的方式，比如市价成交，或者加入很大滑点的限价成交。对于被动成交，则一般都是限价成交，比如说要买入，一般是当前的买 1 价，也可以是比买 1 价低的价。对于流动性比较差的品种，买卖价差非常大，被动成交可能是买 1 与卖 1 之间的某个价位挂单。

如果是第三方平台的程序化交易，一般都是趋势型策略，不会涉及频繁的挂撤单操作，因此基本可以假设是市价成交。对于上海期货交易所这种没有市价指令的交易所，一般是加了 3 个价位的限价指令，它的目的是立即成交。

好了，下面我们更仔细地介绍股指日内策略。

1.1.2 曾经辉煌

前面说过第三方平台非常适合股指日内策略。由于程序化交易、私募基金等是在2012年开始大规模兴起的，所以当时规模都还很小，几百万到几千万的水平，而股指日内策略的容量一般也是这么大。一个比较好的股指日内策略其实交易频率并不高，一般1周1次，即5天1次，如果一个团队储备了50个比较好的策略，那么平均下来每天有10个策略触发，可以做10手。当时股指合约的价值为50～100万元，因此，10个策略对应的容量是几千万元。日内策略的收益回撤比较隔夜策略好，如果每个策略的收益/最大回撤有1～2倍，50个策略合起来应该可以达到3～5倍，这已经是非常好的策略组合了。

如图1-2所示，这是一个股指日内策略组合的效果图。

图1-2 股指日内策略组合

这是5个策略的组合，每个策略配50万元资金，初始资金250万元，总收益66.83%，最大回撤5.38%，收益回撤比超过10倍。由此可以看出，策略在2013年8—12月持续振荡，因为在经历大跌之后的那个时期股指每天均窄

幅波动，真正获利多的是2014年年底开始的大牛市，一直持续了很长时间。一般股指日内策略回测标准是最新价加合约价值万分之三的成本。

2015年8月开始股指实盘，其中2015年10月之后持续稳定，实盘结果如图1-3所示。

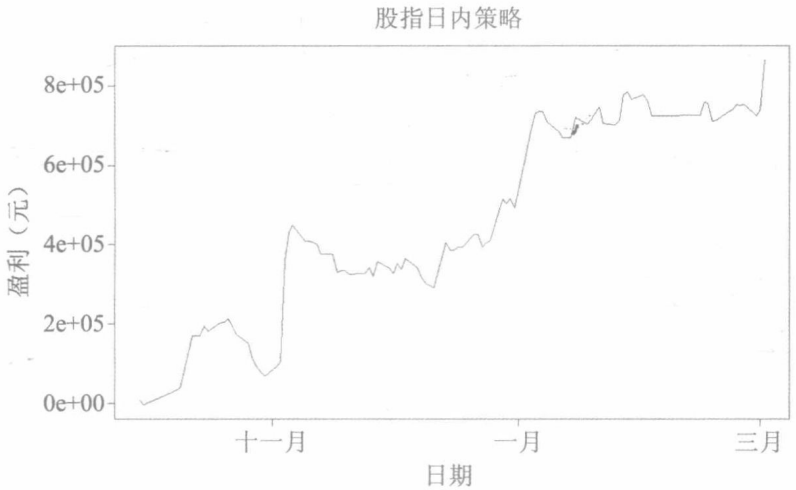

图1-3 股指日内策略实盘交易

2015年10月—2016年3月：夏普比3倍，累计收益/最大回撤5.68倍。应该说这段时间的股指日内表现都还是很不错，主要原因是在股指受到限制后很多厉害的量化团队放弃了这块业务，因此市场有效性降低，盈利变得更加容易。

1.1.3 近期发展

然而，2015年8月，证监会已经对股指进行了开仓限制，每天只能开仓10手，之后买卖价差变大，流动性变低，使股指日内策略的盈利难度变大。其实2015年8月—2016年3月还是可以的，毕竟当时的波动还很大，基本上还可以每个月都能盈利，但2016年1月经历了熔断，之后波动开始下降，2016年3月之后股指日内盈利就比较困难了，于是很多人采取了折中的办法，比如日内收盘不再硬性平仓，毕竟日内的波动已经无法覆盖平仓的成本了。

2015年8月之后日内平仓手续费增加了100倍，要规避这个限制，只能采取锁仓的方式。国内的交易方式跟国外不大一样，国内有“开仓”“平仓”

的概念。比如现在没有仓位的话，买入操作其实是买入开仓，给交易所增加新的持仓量；如果现在已经有了仓位，比如现在是空仓，那么可以选择买入平仓，相当于为交易所减少当前的持仓量。然而，即使现在已经有了空仓，也可以选择买入开仓，这样就规避了日内平仓高额的手续费，只需要在第二天再买入平仓和卖出平仓，即可平掉仓位，实现日内交易。

一开始交易所对保证金的措施是“单向较大金额”，即买入开仓和卖出开仓都要占用保证金，但只收取更大的那一边。后来交易所为了抑制日内交易，取消了这一规定，而且把保证金调整至40%，双向就是80%，几乎没有杠杆了，这进一步降低了日内交易的吸引力。

2017年开始了大盘股的牛市，俗称“漂亮50”和“悲惨3000”，或者说一九行情。即大盘股开启了持续上涨的趋势行情，这也给股指日内交易提供了机会。很多在2016年长期不盈利的策略在2017年也开始盈利。随着股指限制从10手到20手，保证金比例也相应下调，相信未来会越来越好。

1.2 商品趋势策略

商品趋势策略是最传统的量化交易策略，一般人们称为程序化交易策略。下面简单介绍一下。

1.2.1 程序化交易的挚爱

传统的程序化交易者都是交易商品起家的，因为股指期货2010年才上市。例如，TB开拓者陈剑灵就是做商品趋势策略起家，然后反过来收购了TB开拓者，这种商品趋势策略又统称为CTA策略。

一开始的CTA策略一般基于日线指标，因为商品波动太低，持仓时间需要比较长，另外第三方平台处理数据、优化策略的速度实在太慢，很难处理更高频率的数据，因为有数十个品种，如果每个品种每个策略都优化一下，再放到几十个品种重复做一遍，计算量会非常大。因此，CTA策略一般以日线为主，每天收盘前下单。

其实现在很多提及的机器学习、现代统计模型，实际上他们的建模过程比传统的基于灵感和规则的程序化交易模型死板很多，也正因为建模分析过程很死板，因此容易规模化，也更方便严谨测试。实际上，很多灵活处理的交易策略未必能很好地转成机器学习的模型。但从大规模生产和高速计算的角度来看，机器学习优势明显，此内容后面会提到。

下面看一下商品日线策略的例子，如图 1-4 所示。

名称	利润率	年回报	胜率	交易数	成功率	最大回撤...	MAR比率
全部综合	433.31%	22.31%	48.28%	116	13.79%	11.22%	1.99
S 螺纹钢指数	433.31%	22.31%	48.28%	116	13.79%	11.22%	1.99

图 1-4　螺纹钢日线策略

从 2009 年 3 月 27 日螺纹钢期货上市开始，至 2017 年 7 月 18 日，单手螺纹钢初始资金 3 万，由此可见年回报达到 22.31%，累计回报 433.31%，最大回撤 11.22%，交易次数只有 116 次，相当于每年 15 次左右。交易成本是交易所手续费加两个滑点，但由于交易次数非常少，因此加多一些滑点也不会差很多。

一般商品的买卖价差比较大，比如螺纹钢，买卖价差一般占合约价值的 0.3% ～ 0.6%，而手续费只有万分之一左右，因此对于商品低频交易，成本更多指的是买卖价差，而不是手续费。

还有一项叫作 MAR 比率，它实际上是年回报与最大回撤的比值。由于最大回撤一直增加，而年回报一般来说每年会有波动，但整体来说不会随着时间的增加而增加，因此 MAR 比率对时间长的资金曲线不利。比如一些国际知名的 CTA 产品，它的 MAR 或许只有 1 ： 1，因为它成立了 20 多年，总有回撤