北京大学

硕士研究生学位论文

题目：多因子模型投资绩效研究-以沪深 300 为股票池

姓　　名：	钟杰
学　　号：	1401210888
院　　系：	软件与微电子学院
专　　业：	计算机技术
研究方向：	电子商务与物流
导师姓名：	李杰教授

二０一七年七月

版权声明

摘要

股票市场投资可以分为主动投资和被动投资两种方式。被动投资的方法是尽可能跟踪某个股票指数，以指数的成分股和比例配置投资组合，目的是取得与大盘相近的收益。被动投资认为市场是有效的，因而价格反应了所有的信息，因而通过选股、择时等方法都无法战胜市场。主动投资认为市场不是有效的，因而有可能取得超越市场的收益率。目前，主动投资可分为以基本面和技术面为代表的定性投资和定量投资，也称量化投资。量化投资以数学、物理学、统计学等学科为工具，利用历史数据检验模型化的投资理念和投资策略，然后利用成功的策略和理念进行投资的定量投资方法。相对于其他量化投资策略，多因子模型已在国外的实践中取得了较大的成功。

多因子模型认为股票的超额收益率可以由共同因子和特异因子来解释，因而多因子模型的关键是挖掘出对股票收益率具有解释作用的共同因子。影响股票收益率的因子非常多，根据因子类型的不同，多因子模型可以分位宏观多因子模型、基本面多因子模型和统计多因子模型。据全球领先的风险模型机构 Barra 的研究，基本面多因子模型的表现优于宏观和统计多因子模型，因此本文运用基本面多因子模型。

本文的目的有三个。一是对多因子模型的每个环节进行细化，以便发展出一套具有较强可行性的模型构造方法；二是运用多因子模型对国内股票市场作实证分析，目的是为了验证多因子模型的在国内市场是否适用，以及中国股票市场是否是有效的；三是用多种不同的因子加权方式加总股票的总分，并与传统等权加权方式的投资绩效进行比较。本文以沪深 300 为股票池，选取了估值因子、成长因子、资本结构等 7 类共 29 个基本面因子，运用 2007 年至 2016 年股票行情数据与财务数据进行验证。在 2007 年至 2013 年的样本内检验期，在每一期的期初对股票进行排序分组，计算分组收益率与因子 IC，以 IC 考察因子的预测能力，以分组收益率考察因子的收益能力，综合二者表现情况筛选出有效的因子。在 2014 年至 2016 年的样本外检验期，以筛选出的有效因子作为选股指标，运用多种不同的因子加权方法加总股票总分，对总分排序构造多空股票组合，并利用年化超额收益率、胜率、最大回撤，以及 IR 等绩效评价指标对组合绩效进行比较。实证结果表明：一是多因子模型适合中国股票市场，中国股票市场是弱有效的；二是本文的因子加权方式构造的投资组合都能战胜市场；三是由于因子权重是基于历史数据的，因而并不存在最优的方法。

关键词：量化投资，多因子模型，权重

AN EMPIRICAL ANALYSIS OF THE MULTIFACTOR MODEL ON HS300 STOCKS

Jie Zhong (Computer Technology)
Directed by Jie Li

Abstract

In general, there are two kinds of investment styles in stock market, including passive strategy, also called index-tracking strategy, and active strategy. Passive strategy tries to track some indexes, making efforts to get the market return in the way of mimicking the stocks and weights in the in index. Compared to active strategy, passive strategy believes that Market is not efficient, so that stock price reflects all information and no one can get extra returns by selecting stocks, timing, or using other kinds of investment strategies. That disappoints active managers. In tradition, fundamental analysis and technical analysis are the mainstream active investment strategies. In the past few decades, quantitative strategy has rising up in western countries, which comprehensively uses several kinds of subjects, including mathematics, physics, statistics etc. Quantitative strategy models its investment philosophies and strategies and use history data to test whether the strategy can make extra returns.

In the contrast of other quantitative strategies, multifactor model has succeeded in overseas. Multifactor model owes the stock yields to several common factors, which represent system risks, and special factor, which is the risk that only affects one stock. According the factors, multifactor model can be divided into three kinds, including macroeconomics model, fundamental model and statistics model.

There are three objects in this article. First, to construct a detailed and systematic back test method so that others can refer to when they use the multifactor model to invest. Second, use the hs300 stocks and its data to test whether the model suits Chinese stock market, and whether the market is efficient. Third, try to use other weighting methods to compute stock score, which attempts to get more returns compared with equal weight. In this article, about 29 fundamental factors, deriving from financial reports, has been used to back test from January 2007 to December 2016. From January 2007 to December 2013, the object is to select efficient factors, and from January 2014 to December 2016, is to construct investment portfolios.

KEY WORDS: Quantitative investment, Multifactor model, Weight

目 录

第一章 绪论

1.1 研究背景

自资本市场诞生以来，无数的人们前赴后继，希望用自己的聪明才智分享资本市场的收益。根据投资者对市场是否有效的信念，可以将投资分为主动投资和被动投资。被动投资认为投资者无法战胜市场，因而尽可能跟踪某个股票指数或其他基准，使得投资组合的业绩与业绩基准偏离最小。通过复制基准指数的成分及权重配置组合，获取与大盘相近的收益。主动投资认为市场不是完全有效的，相信通过选股、择时等手段，投资组合的收益可以战胜某个股票指数或其他基准。主动投资分为基本面分析、技术分析及量化投资。

在金融市场上，基本面分析是最古老的方式。基本面分析是利用金融学理论，从宏观经济、行业以及公司财务报表等方面，寻找价格低于内在价值或具有高成长性的股票。同时，基本面分析也关注一些市场动态，如政府政策及行业变化等信息。1934 年价值投资之父本杰明.格雷厄姆和佛兰克.多德出版了被誉为“投资圣经”的《证券分析》，开创了价值投资之道。格雷厄姆的学生沃伦.巴菲特将价值投资发扬光大，其投资的可口可乐、迪士尼、华盛顿邮报及美国运通公司使伯克希尔.哈撒韦取得了巨大的成功。从 1965-2009 年，伯克希尔.哈撒韦的年投资回报率为 22%，明显高于同期标准普尔 500 指数 9.3%的收益率。2001 年 12 月 31 日，公司的股票以每股 75600 美元的价格进行交易，成为美国历史上价格最高的上市公司股票。然而，基本面分析的潜在假设是通过分析事件而推测其对股票或市场的影响，但在现实世界中，原因常常是模糊的，而且消息与价格，原因与效果之间的关系并不确定。

诞生于 20 世纪初的道氏理论是所有技术分析的鼻祖。技术分析是通过图表和指标对过往的历史价格信息进行研究以便挖掘出股价历史走势的规律，并依照此规律预测股价未来的走势。技术分析认为市场行为包容消化一切，股价波动可以定量分析和预测，常用的技术分析理论有道氏理论、波浪理论、江恩理论等。所有的技术分析都建立在市场行为包容消化一切、价格以趋势方式演变、历史会重演这三大假设之上。技术分析的优点在择时上，可以告诉投资者何时买卖股票。它的缺点是依赖于历史会重演，即要求市场是无效的，由于资本市场是不断变化演进的，因而技术分析的效果受到挑战。

随着资本市场的发展，市场上投资产品日益增加且市场有效性逐渐增强。近年来，互联网和移动互联网的普及更是让信息急剧膨胀，信息更加透明，传播的速度更快。传统的基本面分析按照宏观分析、行业分析和公司分析的模式极其考验基金经理和卖方研究员的选股能力和信息处理能力，他们无法覆盖数量如此巨大的股票，更无法有效

处理每天纷至沓来的巨量信息。因此，要在巨大的沙砾中淘出黄金，无异于大海捞针。

1971 年美国著名的巴克莱投资公司发行了第一支指数基金，标志着量化投资的开始。詹姆斯.西蒙斯是著名的数学家，也是全球最知名的对冲基金经理之一。由于利用各种数量化模型进行股票投资，西蒙斯被人称为“模型先生”。华尔街的传统是雇佣全球知名商学院的毕业生，而西蒙斯是一个“另类”。西蒙斯并不信任正统金融学理论，从来不雇佣金融专业的学生，文艺复兴科技的雇员都是顶级院校的数学、物理等“硬学科”的毕业生，他们建立的模型在资本市场上大放异彩。1987 至 2007 年，大奖章基金的平均年收益率高达 35%，若考虑高达 44%的收益提成，则实际基金的年收益率超过 60%，远超巴菲特同期的业绩。

量化投资(Quantitative Investment)在海外诞生了 40 多年，宽客(quant）从被嘲笑的群体逐渐被华尔街认可。由于中国资本市场发展时间短、数据少、做空机制不完善、人才培养及研究落后等原因，量化投资在国内还处于起步阶段。然而，量化投资在中国有巨大发展潜力。第一，随着中国经济的发展，国人收入水平快速提高，对投资的热情持续高涨。第二，量化投资在国外取得的惊人成果吸引了越来越多的人从事量化研究。第三，国内资本市场正在不断完善，融资融券、股指期货等做空工具的推出，衍生品的发展为量化投资策略提供了工具。第四，由于计算机技术的发展、互联网和移动互联网的普及以及跨学科交叉融合的优势，客观上要求我们借鉴先进的技术和各学科的知识，在投资者水平不断提高的情况下，利用量化投资获取超越市场的收益。

1.2 研究目的及意义

随着中国资本市场规模和交易制度的日趋完善，国内资产管理和研究行业面临机遇和挑战。近年来，国内券商、基金等各个机构都加大了对量化投资的研究，而且还引进了许多国外先进的投资策略和投资方法。同时，留学海外人员和华尔街从业人员回国创业或工作的人越来越多，量化投资正为越来越多的人所知。量化投资是继基本面分析和技术分析两种主流投资方法后，同时借鉴基本面分析和技术分析中优秀的投资理念和投资方法外，还融合了数学、统计学、计算机科学等各种不同学科的思想和技术，并以历史数据检验为基础，具有很强的科学性和发展潜力。

量化投资是一种主动管理方法，其存在的基础是市场的非有效和弱有效。以某个指数为业绩基准，若某种主动管理方法能够取得超过市场的收益率，则可以证明市场不是有效的，同时也表明该方法适合此种市场。多因子模型假定股票的收益率是由许多风险因素驱动的，因而不管市场如何变化，总是可以找到影响股票收益率的有效因子。从当前中国股票市场的现状来看，在众多的量化投资策略中，多因子模型是为数不多的可以运用的量化策略。

本文的研究目的有三个。一是对多因子模型的每个环节进行细化，以便发展出一套具有较强可行性的模型构造方法；二是运用多因子模型对国内股票市场作实证分析，目的是为了验证多因子模型的在国内市场是否适用，以及中国股票市场是否是有效的；三是用多种不同的因子加权方式加总股票的总分，并与传统等权加权方式的投资绩效进行比较。

对多因子模型的研究非常有意义。首先，由于量化研究有较高的准入门槛，国内对多因子选股的研究都集中在券商、基金等机构手中，他们的研究往往重实证分析而轻理论解释，普通投资者难以理解他们的研究报告。本文对多因子模型的研究同时兼顾理论基础和实证检验，将理论与实践相结合，做到每一步都有据可依，能给普通投资者了解多因子模型一个参考。同时，多因子模型是量化选股模型中一个重要模型，在国外已被许多良好业绩证实为一个有效的模型，因此，对它的深入研究无疑能促进多因子模型的推广和运用。

其次，多因子模型是基本也是最有效的选股模型，它是很多量化选股模型构建的基础。因此，研究其他复杂的量化投资模型离不开对多因子模型的研究。

最后，多因子模型结合基本面因子、技术面因子以及行为金融学的相关理论，对它的研究能更深入地了解驱动中国股票市场涨跌的因素，有利于增强基金经理选股的能力，提高基金的管理水平。

综上所述，对多因子量化选股模型的研究具有极其重要的理论意义和现实意义。

1.3 国内外研究现状

1.3.1 国外研究现状

金融资产定价一直是金融界的热点问题。Sharpe、Lintner 与 Mossin 于 1964 年提出了资本资产定价模型（Capital Asset Pricing Model, CAMP）表明资产的预期超额收益率只与其承担的系统性风险有关[1]。在早期的许多研究中，如 Black、Jensen 和 Scholes 支持 CMAP 理论。然而，80 年代后期，许多学者发现一些市场异象，如公司规模、账面市值比也会对收益率产生影响。Stattman 和 Rosenberg, Reid 及 Lanstein 发现公司账面市值比与公司股票收益率成正比[2]。Banz 发现了规模效应，即小公司的股票具有比 CAMP 可解释的更高的收益率，且扣除风险因素后依然成立，即股票的平均收益率与公司的规模成反比[3]。Fama 和 French 以 1989-1992 年间的股票市场为样本，提出了市场、规模和账面市值比的三因素模型，他们认为市值小的股票的年收益率大大高于市值大的股票，且账面市值比大的股票的收益率大大高于账面市值比小的股票[4]。随后，Fama 的博士生 Clifford Asness 发现了除市场、规模和估值以外的动量因子[6]。Fama 的另一博士生 carhart 后来写成四因子模型的论文[7]。后来，学界在四因子模型的基础上

又发现了质量和波动率因子，更加全面的解释了股票收益率的差异。Lev 和 Thiagarajan 从研究报告中挑选出分析师最看重的 12 个指标入手，建立了多元回归模型来研究超额收益[8]。Partha S. Mohanram 从盈利能力、增长稳定性、财务稳健性三个方面选取 9 个因子作为选股指标，在 PB 排名前 1/5 的股票中选择股票构建组合，取得了很好的市场表现[9]。 Fama, French(2015)在原三因子模型的基础上，加入了盈利能力和投资模式，表明五因子模型比三因子模型的解释作用更强，但加入两个因子后原来的价值因子变得多余[10]。

1.3.2 国内研究现状

对多因子模型的研究，不论 Fama-French(1993)三因子模型、carhart(1997)四因子模型、Fama-French(2013)五因子模型，还是其他多因子模型，国内主要利用国外成熟的方法对国内 A 股市场作实证研究。

靳乐云、刘霖（2001）利用不同流通市值股票的收益率之差作为另一因子，建立了两因子模型，发现不存在无风险收益率时的两因子模型比 CAMP 具有更好的解释能力[11]。范龙振、余世典（2002）的研究证实中国 A 股市场存在显著的市值效应、账面市值比效应、市盈率效应和价格效应[12]。邓长荣、马永开（2005）用深市股票数据验证了 Fama-French 三因子模型在中国股市的有效性[13]。毛小元、陈梦根、杨云红（2008）基于改进的三因子模型，研究了配股对股票收益的影响，表明配股后股票长期低绩效并不是配股公司特有的现象[14]。刘依明（2012）综合了多因子模型和波动率策略下的动量和反转策略，运用六年的历史数据构建投资组合获得了超额收益[15]。勾东宁、王维佳（2015）将沪深两市 16 家上市银行按规模和账面市值比为四组，运用 Fama-French 三因子模型作实证检验[16]。李倩、梅婷（2015）研究了 Fama-French 三因素模型在不同时期的适用性，发现模型在股市衰退时期的适用性最好[17]。章宏帆（2015）综合运用基本面选股策略和动量策略构建了获得超额收益的模型[18]。赵胜民、闫红蕾、张凯（2016）利用我国 A 股市场的数据，对比了 Fama-French 三因素模型和五因素模型在 A 股的适用性，结果表明三因素模型更适合我国股票市场[19]。吴敏华（2016）以深主板、中小板和创业板内的 A 股股票为样本，尝试用 Fama-French 五因子模型解释股票收益率，实证结果表明五因子模型对股票收益的解释强于三因子模型，且中国股市的规模效应和利润效应显著[20]。凌士勤、付力（2016）在多因子模型的基础上加入风格轮动模型，用历史数据检验，结果表明加入风格轮动的模型比多因子模型的表现更加突出[21]。黄若冰（2016）在传统多因子模型的基础上，结合 K-means 算法将股票池分成不同的类，再从中选出有价值的股票，模型获得了较高的收益并承担了较小的风险[22]。

1.4 本文所做的工作

在对多因子模型的实证过程中，本文广泛参考了许多国内外著作和各大研究机构的报告，力求做到对已有研究成果的充分了解。多因子模型逻辑简单易懂，但各个操作环节繁杂，在前人研究的基础上，本文做了以下几个工作：

第一，在理论方面，本文详细论述了多因子模型实证分析各个步骤，力求做到为以后的实证分析提供依据，以增强多因子模型的理论基础；

第二，相比于以往的分析框架，本文充分将 IC 用于单因子有效性检验和因子加权，而且利用了更加严格的因子有效性筛选标准，增强了因子的稳定性；

第三，以往的研究用的大多是因子的原始值，本文用回归的方法使因子行业和风格中性化，充分控制了风险暴露，以使组合的投资风格更加稳健；

第四，以传统等权为参照，尝试利用 IC、IC_IR、最大化复合因子单期 IC、最大化复合因子 IC_IR 四种因子加权方式优化因子权重。

1.5 文章结构与技术路线

本文共分六章，同时兼顾理论阐述和实证分析。下面对各章节的重点内容作一个简要介绍。

第一章为绪论。先介绍了选题的背景、目的和意义，然后对国内外多因子模型的相关研究做了回顾，最后简要介绍了本文研究的技术路线及所做的工作。

第二章是基础理论部分。尽管量化投资是一种主动管理方法，与传统金融理论相悖，但量化投资的许多思想源于传统金融理论和投资思想。因此，本章先对 20 世纪 50 年代以后的现代金融理论做了简要梳理，以便展示金融理论是如何逐渐数量化的。最后，介绍了量化投资和多因子选股的相关理论。

第三章是多因子模型的实证方法。本章介绍了多因子模型的实证方法，对每一步都做了细化，以便在第四章和第五章作实证分析。本文多因子模型的具体包括数据的获取与处理、单因子有效性检验、投资组合构建及组合业绩评价四个步骤。

第四章开始多因子模型在沪深 300 中的实证分析。本章的任务是在 2007 年 1 月至 2013 年 12 月的样本内检验期中，以因子的大小对股票排序，将股票分为 5 组，通过综合 IC 和收益率，从 29 个备选因子中挑选出有效并稳健的因子。

第五章是投资组合构建及组合业绩评价。在 2014 年 1 月至 2016 年 12 月的样本外检验期，利用挑选出的有效因子，以不同的因子加权方式获得股票在所有因子上的总分，按总分给股票排序，选取一定比例的股票构建多空组合，最后再根据年化收益率、信息比率、胜率等指标对组合的业绩进行评价。

第六章是结论和展望。先对前五章的内容作了回顾，在对前文分析的基础上得出本文的结论，然后分析了本文的不足之处及对后续研究的展望。

本文的研究技术路线：

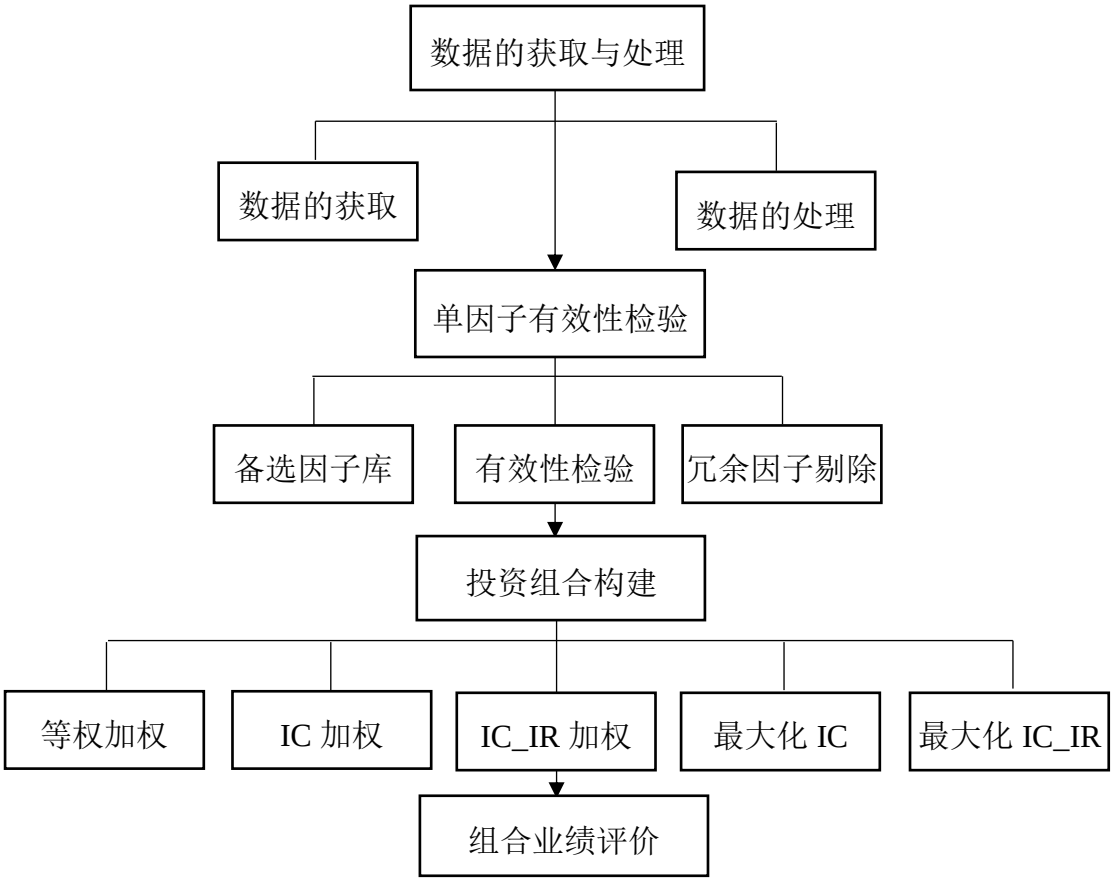


图 1.1 研究技术路线

第二章 基础理论

现代金融学认为市场是有效的，因而主动管理是没有必要的，而量化投资是一种主动管理，区别于传统基本面分析和技术分析等定性投资方法，量化投资充分运用数学、统计学、计算机科学的方法。虽然量化投资与传统金融理论相悖，但它借鉴了许多传统金融中理论和思想。因此，有必要对现代金融理论作一个简要的梳理，以加深对量化投资的理解。

2.1 现代金融理论的发展

正式的金融研究始于 20 世纪初，它的出现源于随机数学和统计学。1900 年，年轻的法国数学家巴舍利耶在其博士论文《投机的理论》中将概率论应用于股票市场的研究，提出了股票价格遵循“随机游走”的关键模型。巴舍利耶的博士论文被其同时代的人大大忽略，到了 1964 年才被译成英文并出版，现代经济学和金融学理论的一个庞大体系正是由那时开始发展出来。

1952 年 Markowitz 在其博士论文《portfolio selection》中提出了奠定现代投资组合理论基础的均值-方差分析模型和投资组合有效边界模型[23]。Markowitz 发现在整本《证券分析》中都找不到对于风险的数量化定义，于是他用期望收益率表示收益，用方差表示风险，同时提出了确定最优资产组合的基本模型和分散化的概念。均值-方差模型是一个最优化问题，它表明一个理性的投资者应该在收益一定的情况下最小化风险或是在风险一定的情况下最大化收益。此外，在一个投资组合中，系统性风险无法分散，而非系统风险可以通过增加股票数量的方式分散。

在 Markowitz 投资组合理论的基础上，Sharpe[1]、Lintner[24]与 Mossin 于 1964 年提出了资本资产定价模型（Capital Asset Pricing Model, CAMP）。根据 CAMP 理论，资产的预期超额收益率与其承担的系统性风险成正比。

CAMP 的数学表述为：

$$E(R_i)-r_f=\beta_i(E(r_m)-r_f)，\ 其中\beta_i=\frac{\mathrm{cov}(R_i,R_m)}{\sigma_m^2} \qquad 式(2.1)$$

其中：

R_i：股票或组合的收益率；

β_i：股票或组合的超额收益率与市场超额收益率之间的协方差除以市场组合超额

收益率的方差；

R_m：市场收益率；

r_f：无风险收益率；

Sharpe[25]在其博士论文《对于“资产组合”分析的简化模型》中提出了单因素模型。单因素模型把经济系统中的所有相关因素作为一个总的宏观经济指标，假设它对整个证券市场产生影响，并进一步假定其余的不确定性是公司所独有的。其一般形式为：

$$R_i = a_i + b_i F + \varepsilon_i \qquad \text{式(2.2)}$$

其中：

R_i：股票收益率；

F：影响股票收益率的系统性因素；

ε_i：影响股票收益率的非系统性即公司层面的因素；

上式假定：任意证券i的随机项ε_i与因子F不相关，任意证券i与j的随机项ε_i与ε_j不相关。

科恩和波格[26]在《商业学刊》上发表《对两种投资组合选择模型的经验评价》中首先提到了多因素模型，从而拓展了单因素模型。该模型表示如下：

$$R_i = a_i + b_{i1}F_1 + b_{i2}F_2 + ... + b_{ik}F_k + \varepsilon_i \qquad \text{式(2.3)}$$

其中：

R_i：股票收益率；

b_{ik}：股票 i 对因子 k 的暴露度，也称因子载荷。为了能在实践中使用，我们假设因子暴露都能在收益率考察期之前获知；

F_k：因子 k 的收益率。因子收益率或是在考察期末被归因到因子上，或是直接在考察期上被观测到；

ε_i：股票 i 的特异收益率，即不能被共同因子所解释的部分，也称奇异收益率。

假设每只股票的特异收益率与因子收益率不相关，且每只股票的特异收益率与其他股票的特异收益率之间也是不相关的。因素模型并没有告诉我们决定股票收益率的因素是哪些以及因素的个数是多少。根据选择的因子风险的不同，多因素模型可以分为统计因子模型、宏观因子模型、基本面因子模型以及混合模型。选择的因子风险是否合理将决定因子模型能否解释股票的超额收益率。

Ross[27]提出了套利定价理论（Arbitrage Pricing Model, APT）。套利定价理论指出任何股票的预期超额收益率都由该股票的因子暴露度以及每个因子的预测决定。APT 中预期超额收益率的表达式为：

$$E(r_i) = \sum_{k=1}^{K} X_{n,k} * m_k \qquad 式(2.4)$$

其中：

r_i：股票超额收益率；

$X_{n,k}$：股票 n 对因子 k 的暴露度，也称因子载荷；

m_k：因子 k 的预测收益率；

Samuelson 和 Fama[28]提出了有效市场假说（Efficient Market Hypothesis,EMH)。它有三个逐渐弱化的理论基础。一、假设投资者是理性的，因而能正确评估资产的价值。二、即使存在非理性的投资者，只要投资者的交易策略是不相关的，市场也可能是有效的。这是因为投资者的行为是随机的，大量的非理性交易者的行为就会相互抵消。三、即使非理性投资者的行为是相关的，由于套期保值者的存在，会抵消非理性投资者的作用。

Fama 依据价格中所反映的信息将市场有效性分为三个层次：弱式有效市场，半强式有效市场以及强式有效市场。在弱式（weak form）有效市场中，当前的市场价格反映了所有历史价格中所包含的信息，因而只使用历史价格和成交量数据不能战胜市场，此时用图表和指标对历史价格所进行的技术分析无效。在半强式(semi-strong form)有效市场中，当前的市场价格反映了过去价格的信息和所有公开的信息（如财务报表和新闻报道），因而只使用公开信息（历史价格、基本面信息、分析师的公开评级等）不能战胜市场，此时基本面分析无效。在强式(strong)有效市场中，当前市场价格反映了所有信息（包括历史价格中的信息，公开信息和私人信息），此时即使是内幕消息也无法发现价值被低估的股票，因而投资者无论怎样都无法战胜市场。

从 20 世纪 80 年代开始，作为对经典的有效市场假说的挑战，行为金融学迅速发展，有力地解释了许多经典理论所无法解释的市场异象，如日历效应、小盘股效应、动量及反转效应等。行为金融学是以市场主体非理性为基础的。行为金融学认为，资本市场是由人组成的市场，由于人类情绪以及认知偏差的影响，人们并非按照经典经济学的效用函数框架进行投资决策。

对于行为金融学，学术界至今没有一个公认的严格定义。Thaler(1993)将其称为“思路开放式金融研究”（open-minded finance），认为只要在关注现实世界时，考虑到人的并不完全理性，即可认为是开始行为金融学的研究了。Hsee(2000)认为, 行为金融学是将行为科学、心理学和认知科学上的成果运用到金融市场中产生的学科，它的主要研究方法是基于心理学实验结果提出决策时的心理特征假设来研究投资者的实际投资决策行为。

从金融理论的发展可以看出，资产定价一直是金融研究的热点问题。资产定价也是微观金融理论的核心问题，除了股票价格问题，还涉及债券，期权、期货等衍生品的定价问题，如在期权定价领域发展除了著名的布莱克-斯科尔斯-莫顿期权定价模型。这些理论和模型给量化投资提供了的参考和思路。

2.2 量化投资概述

量化投资（Quantitative Investment）是利用数学、统计学、物理学、计算机科学等学科通过数学建模，将投资理念和投资策略注入模型，并将历史数据代入模型以验证其有效性的一种科学投资方式。量化投资将投资者经常使用的基本面投资方法、技术分析、现代投资组合理论以及历史数据中反映出的统计规律建立数学模型，消除了投资决策中的情绪化因素，保持了投资的一致性。

2.2.1 传统投资与量化投资

量化投资与传统定性投资一样，都是基于市场非有效或弱有效，是一种主动型投资策略。但是，量化投资与定性投资又有许多方面的不同。

与传统的定性投资依赖于上市公司调研和基金经理个人经验及主观判断不同，量化投资是将定性思想和定量规律进行量化的过程。依靠计算机进行信息处理和投资决策，量化投资克服了人脑对信息处理能力的局限性，可以处理海量的数据，覆盖股票数量多，是一种“宽度”投资方法，而定性投资能对个股进行较为深入的研究。由于依赖模型与数据，因而其投资策略较为客观，克服了投资者的认知偏差以及人类恐惧、贪婪等弱点。

量化投资不同于技术分析。技术分析是对市场历史运行规律的一种总结和运用，但缺乏对这种规律的投资理念的认识。因此，技术分析无法估计该规律的可持续性和未来的变化。量化投资策略覆盖了投资的整个流程，从选股、择时、行业选择、资产配置、风险控制等方面都体现了定性的投资理念，具有良好的理论基础。

量化投资也不同于基本面分析。基本面分析从宏观、行业、公司等角度对公司所处环境及公司自身进行定性研究，不同的人、不同的角度对同一家公司的股票可能会得出完全不同的结论。而量化投资充分利用对市场的理解和数学、物理学、统计学、计算机科学等客观技术，是定性与定量的结合。

尽管量化投资与基本面分析有较大不同，但量化投资与传统基本面分析和技术分析并不是对立的关系，量化投资借鉴了基本面分析和技术分析中可以量化的投资理念和投资策略。

2.2.2 量化投资的分类

量化投资的主要内容有量化选股、量化择时、股指期货套利、商品期货套利、高频交易、期权/可转债和标的证券之间的套利、统计套利等。

量化选股是指用数量化的手段选出能战胜市场基准收益的股票组合。量化选股模型分为基本面选股和市场行为选股。基本面选股包括多因子模型、风格轮动模型和行业轮动模型。市场行为选股有资金流模型、动量反转模型、一致预期模型、趋势追踪模型和筹码选股模型。本文主要讨论多因子模型。

择时交易的核心是用某种方法判断大盘的走势是上涨、下跌还是震荡。若判断是上涨，则买入持有；若判断是下跌，则卖出清仓；若判断是震荡，则高抛低吸。这种方法能获得远高于简单买入持有策略的收益率。量化择时是指利用数量化的方法对大盘的走势进行判断。常用的量化择时方法有趋势择时、市场情绪择时、有效资金模型等。

股指期货套利是利用股指期货市场存在的不合理价格，同时参与股指期货市场与股票现货市场交易，或同时进行不同期限、不同（但相近）类别股票指数合约交易，以赚取差价的交易方法。股指期货分为期现套利、跨期套利、跨市套利以及跨品种套利。

商品期货套利指在买入或卖出某种期货合约的同时，卖出或买入相关的另一种合约，并在某个时间将两种合约平仓。商品期货套利主要分为期现套利、跨期套利、跨市场套利和跨品种套利。

高频交易是指利用高性能计算机迅速捕捉市场中转瞬即逝的价差以寻求获利的交易方式，如某种证券买卖价差的微小变化或不同交易所之间的微小价差。由于高频交易的复杂性，一旦计算错误将给市场造成巨大的冲击，因此，各国金融监管机构对高频交易都由不同的限制。

期权是一种以股票、期货等作为标的资产的衍生证券，具有收益无限而风险有限的优点。期权和股票之间的套利有两种方式。多头套利是指在做多股票的同时买入看跌期权，空头套利是指在卖空股票的同时买入看涨期权。此外，期权之间可以相互组合，根据期限及期权种类的不同，可以分为跨式套利、蝶式套利等。

统计套利是先找出相关性最好的若干对投资品种（股票或者期货等），再找出每一对投资品种的长期均衡关系（协整关系），当某一对品种的价差（协整方程的残差）偏离到一定程度时开始建仓，买进被低估的品种，卖空被高估的品种，等到价差回归均衡时可获利。统计套利包括股票配对交易、股指对冲、融券对冲和外汇交易对冲。

2.2.3 量化投资的优点和缺点

量化投资既借鉴了传统投资中的许多理念，又善于运用最新的科学方法，它具有许多其他投资方法所不具有的优点。

第一，量化投资不仅能克服人的主观情绪偏差和人性弱点，客观评价交易机会，而

且能充分利用科学技术的优势，更全面、系统、准确、及时地捕捉交易机会。

第二，通过对历史数据的深度挖掘，容易发现很多隐藏很深而又复杂的数据规律，可以快速地利用和发现许多其他市场参与者未察觉的交易机会。

第三，数量化和程序化的交易方法，可以快速捕获市场的交易机会，其交易迅速，效率高，运作成本低，能在一定程度上减少基金管理人的运营成本。

虽然量化投资具有许多极好的特性，但也有其自身致命性的缺陷。

第一，量化投资在极端的波动性和模型策略的重复性上有很大的风险。市场的波动性是量化投资获利的关键，但极端波动性可致量化投资于死地，即出现“黑天鹅事件”。由于量化投资是以大概率的条件下获利为基础，因而，在出现难以预测的极端小概率事件时，量化投资策略将遭受严重的打击。即便有两位诺贝尔经济学家坐镇，且交易基于成熟的计算机模型和风险管理策略，长期资本管理公司仍然在俄罗斯债券违约后亏损数十亿美元，最终难逃被收购的风险。

第二，从单一量化策略来看，某些策略是有效的。然而，当市场上太多的人使用相同的策略时，策略的有效性会降低，甚至引发灾难，而这种状况对于量化策略的开发者来说又是无法预知的。

第三，量化投资策略以数据和数学模型为核心，只能量化可以量化的因素和投资理念，而投资是一种艺术和科学的结合体，很多伟大的思想和理念是无法通过模型量化的，这使量化投资在很多情况下是有缺憾的投资方法。

第四，量化投资是基于历史可以重演这一假设的。以分析历史数据的投资方法，有可能造成模型对数据的过拟合，而且市场是千变万化的，过去有效的模式未来不一定有效。

投资是一门复杂的学问，任何投资派别都有其自身的缺陷。尽管量化投资的缺点也很多，但并不能因此而弃之不用。不论是巴菲特、查理.芒格、索罗斯，还是西蒙斯，这些伟大的投资者都用各自的方法取得了非凡的收益。在世界各地的宽客的不断努力下，量化投资必能不断取得进步！

2.3 多因子模型

现代金融理论认为，股票的预期收益是对股票持有者所承担风险的报酬。APT 模型指出套利是资本市场有效的基础，如果市场不是有效的，则会存在无风险套利机会，因而套利会使市场重新回到均衡。同时，APT 模型认为风险资产的预期收益率与一组风险因素相关，多因子模型即是在 APT 模型中发展出来的。

多因子模型(Multifactor Model)正是对风险-收益的刻画，它认为股票的收益率可以由一组代表系统性风险的共同因子和一个仅与该股票有关的特异因子解释。

多因子模型的一般形式为：

$$r_n(t)=\sum_{k=1}^{K} X_{n,t}(t)*b_k(t)+u_n(t) \qquad \text{式(2.5)}$$

其中，

$r_n(t)$：资产 n 在时期 t 的超额收益率，即收益率减去无风险收益率；

$X_{n,t}(t)$：资产 n 在时期 t 期初对因子 k 的暴露度，也叫因子载荷；

$b_k(t)$：因子 k 在时期 t 的收益率；

$u_n(t)$：资产 n 在时期 t 的特异收益率，即总收益率中不能被共同因子解释的部分。

在实际运用中，因子载荷需要在收益率考察期前获得，因子收益率可以在考察期末或是考察期上被观测到。

多因子模型是对风险与收益间关系的定量表达，不同的因子代表不同的风险类型。根据风险因子的不同，多因子模型有以下三种类型：

宏观因子模型。金融理论认为，股票市场与外部经济之间存在着某种可被证实的关联，因而诸如通货膨胀率、利率、汇率等宏观经济变量对股票市场有一定的影响。虽然宏观经济变量具有一定的解释力，但也存在某些缺陷，如基于历史数据所估计的响应系数可能会随着时间的变化而变化。

基本面与市场因子模型。基本面因子模型和市场因子模型基于股票自身的属性，基本面因子模型利用公司财务数据，如市盈率、总市值等，而市场因子模型则利用波动率、换手率、成交量等基于股票价格和成交量计算出的统计量。

统计因子模型。通过收集大量股票的收益率数据，统计因子模型借助一些统计工具，如主成分分析、最大似然估计等生成多种统计因子，如利用主成分分析从股票收益率样本协方差矩阵中提取出 k 个主成分，利用这些主成分来解释股票收益率。但统计工具要求每个资产对因子的暴露度在估计时段内是恒定的，且存在因子的直观含义难以理解，因子的估计过程容易受伪相关性的影响的不足，因而通常避免使用统计因子。

根据 BARRA 对三种多因子模型的研究，基本面因子的效果要好于其他两类模型。

多因子模型是套利定价理论在量化投资中的典型运用。模型通过历史数据找到影响收益率的有效因子，利用有效因子作为选股标准，从而构建能超越市场基准收益的投资组合。多因子模型的构建主要有打分法、排序法和回归法三种方法。

打分法最早由 Fama 和 French 在 FF 三因素模型中提出。打分法是挖掘能预测股票收益的因子，然后根据股票的每个因子值在截面上的相对位置给出股票在该因子上的得分，再按照一定的权重将股票的各个因子得分相加得到股票的最终得分，最后依照该得分对股票进行排序、筛选，以构造投资组合。打分法的优点是比较稳健，不容易受

极端值的影响。

排序法是按照多个因子的大小分别排序，选出排名靠前的股票构建投资组合。基于因子排序的选股模型需要明确哪些是重要因子，哪些是次要因子。在此基础上，股票按照重要因子排序，然后再按照次要因子排序，最终选出符合要求的组合。

回归法利用因子的风险暴露与股票下期收益率之间的线性关系，以最小二乘法拟合出因子收益率，然后将最新的风险暴露带入回归方程得到对股票下期收益率的预测，并以此作为选股依据。回归法的优点是能及时调整股票对各因子的敏感性，而且不同的股票对不同的因子的敏感性也可以不同，缺点是容易受极端值的影响，在股票对因子敏感性变化较大的市场情况下效果较差。

第三章　多因子模型的实证步骤

多因子模型的主要目的是要从众多的因子中找到能对股票收益率有解释作用的因子，简单的说就是能从因子的大小中区分出收益率不同的股票。多因子模型并未明确影响股票收益率的因素，也并未限制影响因素的个数，其实证过程有很大的灵活性。依据本文研究的目的，本章对多因子模型实证的各个环节给出自己的方法，大体分为数据的获取与处理、单因子有效性检验、组合构建及组合业绩评价四个步骤。

3.1 模型设定

实证分析前，需要对相关内容进行初步设定：

回测区间：回测区间的设定有两点需要注意，一是中国股票市场股权分置改革前后的变化，二是不同的市场阶段市场的走势不同。2005 年开始的股权分置改革，旨在消除非流通股和流通股之间的制度差异，适应资本市场改革开放和稳定发展的要求。股权分置改革前后，市场中的投资者、行业和规模上都有较大的变化，因而前后两个市场是不同的，一般从 2006 年后开始回测较好。回测区间的选择最好能覆盖不同的市场状况，如能覆盖牛市、熊市、震荡市，使因子能经受不同的市场状况的考验，增强模型的健壮性。由于因子在不同市场阶段的有效性不同，因此可以在不同市场阶段研究选股因子的效果，以挑选适合某一市场行情的有效因子，在对市场有过明确的预判后，可以有针对性地利用这些因子的选股。

股票池：不同的股票池构建的投资组合的效果不同。可以选择全股票市场，也可以根据不同的标准选择股票池，如根据不同的指数成分股，可以选择沪深 300 指数成分股、中证 300、中证 500 等；根据不同的行业，可以选择机械设备、建筑材料、有色金属等；根据不同风格，可以选择大盘股、中盘股和小盘股，周期和非周期行业。已有的研究表明，全股票市场一般比指数成分股、行业、风格效果更好。选择不同的股票池，可以对同一个因子的选股效果进行比较。

备选因子库：多因子模型是一个统计模型，它没有明确规定因子的种类和影响的大小，获取更多更有效的因子有助于最大限度地捕获股票的收益。对因子的要求只有一个，即要求因子载荷在在考察期初就可以确定。常见的因子有宏观经济因子、基本面因子和市场因子以及统计因子。宏观因子，来源于股票市场与外部经济力量之间的某种关联，如油价变动、利率、汇率等。宏观因子的解释能力可能非常强，但由于估计误差和数据获取等问题一般不使用。基本面类和市场类因子，基本面类市盈率、总市值等

可以从公司财务报表中计算而来；市场类属性包括过去某一时间段上的收益率、过去某一时间段上的波动率、期权的隐含波动率、换手率等。通过最大似然分析和预期最大化分析等得到统计因子。由于统计因子的直观含义难以理解，因子的估计过程容易受到伪相关性的影响，而且统计工具也不能捕获暴露度随时间变化的因子，因而我们避免使用统计因子。

调仓频率：调仓频率可以按周、月、季度、年等不同的时间跨度进行调仓，调仓频率的大小影响组合成本，调仓频率过高带来股票调入调出组合的手续费过高，很有可能会抹平组合的收益。另外，由于公司有财务报告期，因此调仓频率的设定决定了不一定能获取当天的财务数据，此时需要根据报告期和调仓频率计算获取财务数据的日期。

业绩基准：用于投资组合业绩评价。主动投资管理的目的是为了获得超过市场平均水平的 alpha 收益。市场收益包括所有可以投资的标的组成的市场，由于无法知道全部标的构成的市场收益，因此，在评价一个投资组合的业绩表现时，通常会选择一个市场业绩基准，常用的市场业绩基准一般为交易所发布的某个市场指数，如 HS300 指数，中证 500、中证 800 等

数据库：目前，有许多如 BLOOMBERGE、WIND、CEIC 等数据库，也有雅虎财经、新浪财经等门户网站可以下载数据。如果需要获取的数据量较大，可以通过 wind 提供的函数接口编程获取数据。

3.2 数据的获取与处理

3.2.1 数据的获取

实证检验中非常关键的一环是关于数据的获取。需要获取的数据有股票的财务数据、行情数据，业绩基准的行情数据等。

在设定好上一步的相关内容后，按照以下步骤取出数据：

第一，根据回测区间和调仓频率计算组合构建日期，组合构建日期必须在股票交易日。然后，计算获取行情数据和财务数据的日期。行情数据、估值因子和规模因子可以获取构建日当天的数据。需要注意的是，国内上市公司财务报告期为最迟每年的 4 月底出当年一季报和去年年报，8 月底出 9 中报，10 月底出三季报，因而一些财务因子，如盈利因子、成长因子等无法获取当天的数据，只能根据财务报表的报表日获取最近的数据；

第二，获取组合构建日成分股、股票所属的行业和需要剔除的股票；

第三，依据取因子日期和成分股获取股票的财务数据和行情数据，以及业绩基准的行情数据。

3.2.2 数据的处理

在选股日，首先应剔除三类股票：

一是 ST/*ST 股票。ST 股票是指上市公司在财务或其他方面出现异常而被限制单日涨跌幅在 5%以内的股票。这些股票在做策略的时候应该作为异常值剔除，但考虑到幸存者偏差的问题，在选股日是 ST/PT 的股票不一定代表在过去也一直是 ST/PT 的股票，一旦不再 ST/*ST，股票将重新纳入股票池。

二是上市不满一年的股票，考虑到新股的特殊性，上市不满 1 年的股票不纳入股票池。

三是无法交易的股票，包括停牌的股票和选股日无法交易的股票。

在余下的股票中，再按以下步骤处理因子值：

第一，空值和负值的处理。对于估值因子和规模因子这类反向因子，由于负值已经失去意义。为避免数据不一致的问题，负值和空值分别用除这两类值外的最大值和均值代替。

第二，去极值。因子常会出现个别极大或极小值，为了更好地计算因子收益，首先应该对因子去极值。去极值有许多方法，如固定比例法、“中位数法”、“3σ”等。由于固定比例的设定较为主观，需要对照数据的分布特点来设定，而模型涉及的数据量非常大，可行性较差，所以一般不用。下面介绍“中位数法”和“3σ法”。

“中位数法”去极值表示为：

$$\begin{cases} f_{i,upper} = f_m + N * f_{MAD}, f_i \geq f_m + N * f_{MAD} \\ f_{i,\text{lower}} = f_m - N * f_{MAD}, f_i \leq f_m - N * f_{MAD} \end{cases} \quad \text{式(3.1)}$$

其中，

f_i：第i个因子值；

f_m：所有因子值 f_i 的中位数；

f_{MAD}：所有因子值 f_i 与 f_m 距离的中位数；

N：偏离的倍数；

“3σ法”去极值表示为：

$$\begin{cases} f_{i,upper} = \mu + 3\sigma, f_i \geq \mu + 3\sigma \\ f_{i,lower} = \mu - 3\sigma, f_i \leq \mu - 3\sigma \end{cases} \quad \text{式(3.2)}$$

其中，

f_i：第i个因子值；

μ：所有因子值 f_i 的均值；

σ：所有因子值 f_i 的方差；

第三，行业与风格中性化。部分因子在不同行业和风格股票中相差非常大，按照排序的方法选出的股票可能会偏向某一个行业，为了使选出的股票不受行业和风格的影响，需要行业与市值的中性化。

在选股日用因子（如市盈率）对行业和市值作回归，取残差作为新的因子值。在t期，记 $\vec{f}_t=(f_{1,t},f_{2,t},...,f_{N,t})^T$ 为期初 N 只股票在某个因子上的取值；假设股票划分为 D 个一级行业，则 $I_t=(\vec{i}_{1,t},\vec{i}_{2,t},...,\vec{i}_{D,t})$ 为 $N\times D$ 行业矩阵，若第 n 只股票属于第 d 个行业取值为1，否则取值为0；记 $\vec{M}_t=(m_{1,t},m_{2,t},...m_{N,t})^T$ 为 n 只股票总市值取自然对数后在横截面上标准化后的值。在期初，以 $\vec{f}_t$ 为因变量，I_t 与 $\vec{M}_t$ 为自变量做回归，取残差 $\vec{\varepsilon}_t$ 为新的因子值，即

$$\vec{f}_t \sim I_t+\vec{M}_t+\vec{\varepsilon}_t \qquad \text{式(3.3)}$$

对行业的划分，国际上通用的是富时行业分类（ICB）和国际行业分类标准（GICS），国内有证监会、中证、申万等行业划分方法。国内普遍采用的是申万行业划分方法，共划分为28个一级行业：

采掘、化工、钢铁、有色金属、建筑材料、建筑装饰、电气设备、机械设备、国防军、汽车、家用电器、纺织服装、轻工制造、商业贸易、农林牧渔、食品饮料、休闲服务、医药生物、公用事业、交通运输、房地产、电子、计算机、传媒、通信、银行、非银金融、综合。

第四，标准化求 *zscore*

在t期期初，用均值-标准差法对横截面上所有股票的因子值标准化，表示为

$$zscore=\frac{\vec{f}_t-E(\vec{f}_t)}{std(\vec{f}_t)} \qquad \text{式(3.4)}$$

第五，$Gram-Schmidt$ 正交化

在使用最大化复合因子 IC_IR 加权时，需要使因子线性无关。因此，对第四步中的标准化残差做 $Gram-Schmi$ 正交化，以使因子线性无关。在t期期初，记 $(\vec{f}_1,\vec{f}_2,...,\vec{f}_K)$ 为横截面上股票 K 个因子的值，$(\vec{f}_1^{\,0},\vec{f}_2^{\,0},...,\vec{f}_K^{\,0})$ 为正交化后的因子值，则

$$\vec{f}_1^{\,0}=\vec{f}_1 \qquad \text{式(3.5)}$$

$$\vec{f}_2^{\,0}=\frac{1}{\sqrt{1-\rho_{21}^2}}(\vec{f}_2-\rho_{21}^2\vec{f}_1^{\,0}) \qquad \text{式(3.6)}$$

$$\overrightarrow{f_3^{\,0}} = \frac{1}{\sqrt{1-\rho_{32}^2-\rho_{31}^2}}(\overrightarrow{f_3} - \rho_{32}\overrightarrow{f_2^{\,0}} - \rho_{31}\overrightarrow{f_1^{\,0}}) \qquad \text{式(3.7)}$$

$$\overrightarrow{f_k^{\,0}} = \frac{1}{\sqrt{1-\rho_{k,1}^2-\rho_{k,2}^2-\ldots-\rho_{k,k-1}^2}}(\overrightarrow{f_k} - \rho_{k1}\overrightarrow{f_1^{\,0}} - \rho_{k2}\overrightarrow{f_2^{\,0}} - \ldots - \rho_{k,k-1}\overrightarrow{f_{k-1}^{\,0}}) \qquad \text{式(3.8)}$$

3.3 单因子有效性检验

3.3.1 *IC* 与检验步骤

将回测区间划分为样本内检验期与样本外检验期。在样本内检验期，目的是从备选因子库中挑选出对股票收益率有较好解释作用的有效因子，也就是检验因子的大小与股票未来的收益是否有相关性；在样本外检验期，用有效因子挑选股票构建适应市场变化并能大概率战胜市场的股票组合。

在样本内检验期，因子有效性检验包括因子预测能力和收益能力两个方面。用 IC 来检验因子预测能力，用分组收益率来检验因子的收益能力。

信息系数（Information Coefficient, 简称 IC），有两种不同的定义。

定义 1：因子对股票下期收益率的预测与股票实际收益率之间的相关系数，表示为：

$$IC_k(t) = corr(r^e(t), r(t)) \qquad \text{式(3.9)}$$

其中，

$IC_k(t)$：因子 k 的在 t 期的 IC 值

$r^e(t)$：因子 k 在 t 期期初对股票下期收益率的预测

$r(t)$：股票 t 期末的实际收益率

通过 t-1 期期末的股票实际收益率对 t-1 期期初的因子 k 作回归计算出因子收益率，再将 t 期期初的因子值代入得到 t 期期初对期末收益率的预测值 $r^e(t)$，进而求出 $IC_k(t)$。

定义 2：t 期期初的因子值与 t 期期末的股票实际收益率之间的相关系数

$$IC_k(t) = corr(f(t), r(t)) \qquad \text{式(3.10)}$$

其中，

$IC_k(t)$：因子 k 的在 t 期的 IC 值

$f(t)$：因子 k 在 t 期期初的因子值

$r(t)$：股票在 t 期期末的实际收益率

相关系数的计算有皮尔逊(person correlation)相关系数和斯皮尔曼(spearman correlation)相关系数，皮尔逊相关系数直接用原始值计算相关系数，而斯皮尔曼相关系数用原始值的位次代替原始值计算相关系数。当两个变量满足正态分布时，皮尔逊和斯皮尔曼相关系数很接近，但斯皮尔曼相关系数不依赖于变量的正态分布特性，因此计算 *IC* 时多采用斯皮尔曼相关系数。

信息比率（Information Ratio, IR），因子 *IC* 均值与标准差的比值，综合考虑了因子的预测能力和稳定性，计算方法如下：

$$IR_k = \frac{\frac{1}{T_1}\sum_{t=1}^{T_1} IC_{kt}}{\sqrt{\frac{1}{T_1 - 1}(IC_{kt} - \frac{1}{T_1}\sum_{t=1}^{T_1} IC_{kt})}} \qquad \text{式(3.11)}$$

其中，

IR_k：因子 k 的 IR；

T_1：样本内检验期的组合构建次数；

IC_{kt}：因子 k 在 t 期的 *IC* 值。

对有效因子的筛选，包括以下三个步骤：

第一， *IC* 和分组收益率的计算；

第二，有效性分析；

第三，冗余因子的剔除。

下面对每一步具体的实施进行细化。

3.3.2 第一步：IC 和分组收益率计算

此步骤有三个计算目标，分别是分组收益率、*IC* 和因子之间的相关系数。先计算，后续用到这些结果再说明。

在样本内检验期内，第 1 期开始，直至检验期末，滚动对股票排序并分组。具体来说，在每一期 t，按以下步骤进行计算：

第一，数据处理。3.1.3 所述第一至第四步处理股票和因子值；

第二，计算分组收益率。对每一因子，在期初将股票按因子值的大小从小到大排序，分成 n 组，持有至期末，以等权重分别计算各组的收益率；

第三，计算 IC 值。在期初，按前述方法，计算每个因子的 IC 值；

第四，计算因子值之间的相关系数矩阵 $\mathrm{corrM}atrix_t(1,2,...,k)$。在期初，运用各个因子序列计算因子之间的相关系数。

为方便后续阐述，将因子的大小与股票收益率负相关的因子称为负向因子，即因

子越小，股票的收益率越大，一般估值因子和规模因子属于此类；将因子的大小与股票收益率呈正相关的因子称为正向因子，即因子越小，股票的收益率越小，一般成长因子属于此类。另外将正向因子的第五组、负向因子的第一组称为优势组合，负向因子的第五组、正向因子的第一组称为劣势组合。

3.3.3 第二步：有效性分析

经过 3.2.2 的计算，得到了样本内检验期内每个因子的分组收益率时间序列和 *IC* 时间序列。前文所述，*IC* 用于评估因子的预测能力，分组收益率用于评估因子的盈利能力。

对于每一个因子的 IC 时间序列，计算 IC 的均值、标准差、IR、p 值、IC 值序列中大于 0（或小于 0）的比例，并画出 IC 的分布图。IC 均值用于判断因子预测能力的大小，标准差用于评价因子预测能力的稳定性，IR 综合了因子预测能力的有效性和稳定性，比例用于判断因子预测效果的一致性，p 值用于判断因子的预测能力是否显著。

对每个因子的分组收益率时间序列，计算优势组合与劣势组合的收益差时间序列，根据收益差计算几何平均收益率，年化收益率，大于 0 的期数与总期数的比例，收益差的 p 值，并画出各组月均收益率的柱状图。

判断方法：

首先，判断因子单调性。观察 IC 分布图，以考察 IC 在整个样本内检验期的正负值比例，某一方向越多越好，说明一致性较强；观察分组收益率图，考察因子与收益率是否有单调性，并有一定的区分度，从而初步筛选出因子。

其次，剔除 IC 均值、胜率过低，或 t 检验不显著的因子；通过优势组合与劣势组合的收益差的 p 值，考察优势组合是否能显著战胜劣势组合，再剔除收益能力较低的因子。

最后，经过综合比较，从备选因子库中初步筛选出有效因子。

3.3.4 第三步：冗余因子的剔除

在 3.3 节的组合构建时，需要综合利用有效因子的信息。通过有效性检验的因子可能会存在信息源的重叠，因子间也会存在较强的相关性，因而有必要进一步剔除信息重复的因子。

若在样本检验期内得到每一期的因子间相关系数矩阵 $\mathrm{corrM}atrix_t(1,2,...,k)$，对所有相关矩阵求平均值，

$$\overline{\mathrm{corrM}atrix}=\frac{1}{T_1}\sum_{t=1}^{T_1}\mathrm{corrM}atrix_t(1,2,...,k) \qquad \text{式(3.12)}$$

设置阈值为 0.5，剔除因子间相关系数超过 0.5 的两个因子中的一个，留下更有效的一个。

3.4 组合构建

经过 3.2 节有效因子的筛选，就可以根据筛选出的有效因子在样本外检验期构建投资组合了。

3.4.1 组合构建的步骤

设样本外检验期共有 T_2 个组合构建日，从样本外检验期开始直到结束，在每一期 t 构建组合的步骤为：

第一，数据处理。按 3.1.3 的方法处理股票因子和收益率数据；

第二，在期初，计算股票 *zscore* 总分。用某种因子加权方式计算股票的总得分。因子加权方式有等权重加权、复合因子加权、基于 IC 的加权等方式，详细说明参见 3.3.2 节；

第三，排序选股。在期初，对该选股日的所有股票按第三步求出的总分排序，买入总分靠前的一定数量股票，卖出总分最低的相同数量股票，持有至该期期末。组合中股票仓位的配置对组合的业绩有一定的影响，在股票仓位的配置上有以下几种方法：

1） 等权重，即组合中的每个股票具有同等的权重。这是最普遍的方法

2） 流通市值加权，即组合中的股票权重取决于股票的市值大小

3） 风格/行业中性加权。如按照大、中、小盘和周期、非周期行业的股票配置不同的仓位，以体现不同的投资风格。此外，根据不同行业，如金融、机械重工等不同行业的股票也可以配置不同的仓位。

第四，在期末，计算组合收益率和同期业绩基准收益率。

3.4.2 因子加权方法

根据多因子模型理论，股票的收益是由不同的因素共同驱动的，有效因子的挑选及因子赋权方式是多因子模型的关键。多因子模型假设每一类因子会单独产生收益，但因子收益是动态变化的过程。在不同的风格子集，不同的宏观环境及市场环境中，都会有不同的有效因子。因此，好的赋权方式能将有效的因子发挥更大的威力，而差的赋权方式则会将低估更有效的因子，因而造成组合回报大大低于预期。

在加总股票得分时涉及到因子权重的问题，由于因子的权重是因子对于模型的重要性，若某一因子有优于其他因子的表现，就要在其权重分配中体现出来。按检验的过程中，因子权重是否变化，可以分为静态加权和动态加权方式。静态加权方式即等权重

加权方式，动态加权方式有基于 IC 加权，聚类分析等方式。

3.4.2.1 等权重加权

等权重加权是典型的静态权重分配法，它认为每个因子有相同的有效性，所以无需差别对待，每一期模型都会为每个因子赋予同样的权重。在因子间相关性较低、有效性差异不明显的情况下会产生较好的效果。等权重加权时其他加权方法验证时对比的基准。等权重方法的优点在于权重的稳定性。由于是静态的权重，模型最终挑选股票组合的换手率低。但这种方法较为主观，没有考虑不同因子之间有效性的差异，受因子间线性相关性的影响较严重。设有 k 个因子，则因子权重为：

$$\vec{w} = (\frac{1}{k}, \frac{1}{k}, ..., \frac{1}{k})' \qquad \text{式(3.13)}$$

3.4.2.2 合成因子等权

为了避免因子共线性的影响，可以把相关性高的因子归为一类，通过某种加权方式合成新的因子，或用其中最有效的因子来代替同类因子。但这种方法的难点在于如何对因子合理分类。基于因子的经济含义，可以把相同逻辑的因子归为一类，但这种方法主观性较强，基本面如估值、成长、盈利等因子经济意义较为明确，而技术面因子的逻辑并不好区分。因此，对因子的分类不确定性较大，受主观因素影响较大。

3.4.2.3 *IC* 和 *IC _ IR* 加权

当因子间存在较强相关性，且有效性差异较大时，等权加权方式会损失强有效因子带来的收益。考虑到了因子有效性的差异，IC 和 IR 加权是把因子过去一段时间的 IC 均值和 IR 作为权重，对各个因子的 zscore 进行加权汇总。在某一期t，分别利用因子过去 S 期的 IC 计算 IC 均值、标准差及 IR，分别得到该期因子的权重 $\overline{IC_k}$ 与 IR_k。

$$\overrightarrow{w_{IC}} = (\overline{IC_1}, \overline{IC_2}, ..., \overline{IC_K})' \qquad \text{式(3.14)}$$

$$\overrightarrow{w_{IR}} = (IR_1, IR_2, ..., IR_K)' = (\frac{\overline{IC_1}}{std(IC_1)}, \frac{\overline{IC_2}}{std(IC_2)}, ..., \frac{\overline{IC_K}}{std(IC_K)})' \qquad \text{式(3.15)}$$

3.4.2.4 最大化复合因子单期 *IC* 加权

记有 N 只股票，K 个因子构成的 $N \times K$ 的因子矩阵 $F = (\vec{f_1}, \vec{f_2}, ..., \vec{f_K})$，因子 $\vec{f_i}$ 经过了标准化处理。因子权重为 $\vec{w} = (w_1, w_2, ..., w_K)'$，股票下期股票收益率为 $\vec{r}$。

记单个因子的 IC 为 $IC_i = corr(\vec{f_i}, \vec{r})$，IC 向量为 $IC = (IC_1, IC_2, ...IC_K)$ 两个因子间的

协方差$\sigma_{i,j}=\mathrm{cov}(\overrightarrow{f_i},\overrightarrow{f_j})$，则$\sigma_i$=1，协方差矩阵为$\Sigma=(\sigma_{i,j})_{i,j=1}^{K}$ 。

以权重 w 加总各个因子 zscore 得分得到复合因子 $\overrightarrow{f_c}$ ，

$$\overrightarrow{f_c}=F.w=\sum_{i=1}^{K}w_i.\overrightarrow{f_i} \qquad \text{式(3.16)}$$

$$std(\overrightarrow{f_c})=\sqrt{\overrightarrow{w}'\Sigma\overrightarrow{w}} \qquad \text{式(3.17)}$$

复合因子 f_c 的 IC 为，

$$\begin{aligned} IC_c &= corr(\overrightarrow{f_c},\vec{r}) \\ &= \frac{\mathrm{cov}(\overrightarrow{f_c},\vec{r})}{std(\overrightarrow{f_c}).std(\vec{r})} \\ &= \frac{\sum_{i=1}^{K}w_i\,\mathrm{cov}(\overrightarrow{f_i},\vec{r})}{std(\overrightarrow{f_c}).std(\vec{r})} \\ &= \frac{\sum_{i=1}^{K}w_i\,corr(\overrightarrow{f_i},\vec{r})std(\overrightarrow{f_i})}{std(\overrightarrow{f_c})} \\ &= \frac{\sum_{i=1}^{K}w_iIC_i\sigma_i}{\sqrt{\overrightarrow{w}'\Sigma\overrightarrow{w}}} \end{aligned} \qquad \text{式(3.18)}$$

由于因子经过了标准化处理，因此上式变为

$$IC_c==\frac{\sum_{i=1}^{K}w_iIC_i}{\sqrt{\overrightarrow{w}'\Sigma\overrightarrow{w}}} \qquad \text{式(3.19)}$$

为使上式最大化，对 w 求偏导，得

$$\overrightarrow{w}=\Sigma^{-1}IC \qquad \text{式(3.20)}$$

3.4.2.5 最大化复合因子 *IC_IR* 加权

综合考虑 IC 的大小及稳定性，Qian, Hua[31]在 QPEM 中提出了以最大化复合因子的 IR 值为目标的因子加权方法。沿用 3.3.2.5 的记号，下面推导其计算方法。

在样本外检验期的每个选股日,需要用到前T_0 期的 IC 时间序列，构成$T_0\times K$ 的 IC 矩阵，记

$\overline{IC}$：因子 IC 的均值向量；

$\sum_{IC}$：因子 IC 的方差-协方差矩阵；

$\vec{w}$：因子的权重向量；

$IC_{c,t}$：复合因子 f_c 在 t 期期初的 IC

表示为：

$$\overline{IC} = (\overline{IC_1}, \overline{IC_2}, ..., \overline{IC_K}) \quad \text{式(3.21)}$$

$$\sum_{IC} = (\rho_{ij})_{i,j=1}^{K} \quad \text{式(3.22)}$$

$$\vec{w} = (w_1, w_2, ..., w_K)' \quad \text{式(3.23)}$$

$$IC_{c,t} == \frac{\sum_{i=1}^{K} w_i IC_{i,t}}{\sqrt{\vec{w}' \sum \vec{w}}}, t = 1, 2, ...T_0 \quad \text{式(3.24)}$$

为复合因子 f_c 在 t 期期初的 IC，对因子作 Gram-Schmit 正交化后因子协方差成为单位矩阵，

$$IC_{c,t} = \frac{\sum_{i=1}^{K} w_i IC_{i,t}}{\sqrt{\vec{w}' \vec{w}}} = \frac{1}{\tau} \sum_{i=1}^{K} w_i IC_{i,t} \quad \text{式(3.25)}$$

则复合因子 f_c 的 IC 均值和标准差分别为，

$$\overrightarrow{IC_c} = \frac{1}{\tau} \sum_{i=1}^{K} w_i \overline{IC_i} = \frac{1}{\tau} \vec{w}' \overline{IC} \quad \text{式(3.26)}$$

$$std(IC_c) = \frac{1}{\tau} \sqrt{\sum_{i=1}^{K} \sum_{j=1}^{K} w_i w_j \rho_{ij} \sigma_{IC_i} \sigma_{IC_j}} = \frac{1}{\tau} \sqrt{\vec{w}' \Sigma_{IC} \vec{w}} \quad \text{式(3.27)}$$

复合因子的 IR 为，

$$IR = \frac{\overline{IC_c}}{std(IC_c)} = \frac{\vec{w}' . \overline{IC}}{\sqrt{\vec{w}' \sum_{IC} \vec{w}}} \quad \text{式(3.28)}$$

为了使复合因子的 IR 最大，求解目标函数：

$$\max_{w} IR_c = \frac{\vec{w}'.\overline{IC}}{\sqrt{\vec{w}'\sum_{IC}\vec{w}}} \qquad \text{式(3.29)}$$

求$\vec{w}_t$的偏导数，得

$$\frac{\partial(IR_c)}{\partial\vec{w}} = \frac{\overline{IC}}{\sqrt{\vec{w}'\sum_{IC}\vec{w}}} - \frac{(\vec{w}'.\overline{IC})\sum_{IC}\vec{w}}{\left(\vec{w}'\sum_{IC}\vec{w}\right)^{3/2}} \qquad \text{式(3.30)}$$

令上式为零，得

$$(\vec{w}'.\sum_{IC}.\vec{w})\overrightarrow{IC} = (\vec{w}'.\overrightarrow{IC})\sum_{IC}\vec{w} \qquad \text{式(3.31)}$$

得到因子的最优权重向量为：

$$\overrightarrow{w^*} = \lambda\sum_{IC}^{-1}\overrightarrow{IC} \qquad \text{式(3.32)}$$

其中λ是使权重之和为 1 的数。

3.4.3 样本协方差矩阵的压缩估计

在 3.3.2 节用最大化复合因子 *IR* 的方法求最优权重时需要估算因子 *IC* 的协方差矩阵。最常用的估计量是样本协方差矩阵$\hat{\sum}_{IC}$。样本协方差矩阵是总体协方差矩阵的无偏估计量，且在正态假设下还是极大似然估计。但样本协方差矩阵的估计误差非常大，当因子数量超过时间样本数量T_0 时矩阵不可逆；即使因子数量小于时间样本数量从而矩阵可逆，样本协方差矩阵的逆矩阵也不是协方差矩阵逆矩阵的无偏估计。因而样本协方差矩阵计算最优权重不可行。

Ledoit[32]提出了压缩估计量方法，用一个方差小但偏差大的高度结构化的估计量F作为压缩目标，和样本估计量S 做一个调和，以牺牲部分偏差来获得更稳健的估计量，即

$$\hat{\sum}_{shrink} = \delta^* F + (1-\delta^*)S, 0 \le \delta \le 1 \qquad \text{式(3.33)}$$

计算压缩协方差矩阵需要知道压缩目标F和最优压缩强度δ^*，压缩目标F有单参数形式、平均相关系数形式等三种不同的形式，本文选择固定相关系数形式。具体来说就是假定因子之间的 *IC* 相关系数是相等的，则固定相关系数就是所有因子间相关系数的平均值。

最优压缩强度的计算通过最小化压缩估计量与真实协方差的期望距离来计算。压缩估计量与真实协方差的期望距离为

$$L(\delta) = \| \delta F + (1-\delta)S - \sum \|^2 \qquad \text{式(3.34)}$$

则

$$\min_{\delta} E(L(\delta)) = E(\| \delta f + (1-\delta)S - \sum \|^2) \qquad \text{式(3.35)}$$

求解上式的公式推导见附录 A。

3.5 组合业绩评价

3.4 节构建了多空投资组合，得到了多空组合和业绩基准的收益率时间序列。组合的业绩到底好不好，有一些评价的指标。对于组合的业绩评价，必须同时兼顾组合的收益与风险。

设股票或组合 p 投资 T 个时期，记以下变量：

T ：股票或组合投资的期数

t ：第 t 个时期，$t = 1,2,3,...,T$

r_f ：无风险收益率

R_p：股票或组合的收益率

R_B ：业绩基准的收益率

r_p ：股票或组合的超额收益率

3.5.1 概念界定

由于在实际使用中，对一些概念存在模糊的地方，如 alpha，平时使用时指的是相对基准的超额收益，而理论中的 alpha 是指实际收益率超出预期收益率的部分，也就是不能由风险因子解释的部分。本文为了规范概念的使用，界定以下几个概念。

无风险收益率：投资于短期国库券、货币市场基金等无风险资产时所获得的利率

超额收益率：风险资产的实际收益率与实际无风险收益率的差值，即

$$r_p(t) = R_p(t) - r_f \qquad \text{式(3.36)}$$

阿尔法（alpha）：股票或组合的实际超额收益率与预期超额收益率的差值，即

$$alpha_p(t) = R_p(t) - r_f - \beta_p(R_B(t) - r_f) \qquad \text{式(3.37)}$$

主动收益率：风险资产的实际收益率超出业绩基准实际收益率的部分，即

$$\mathrm{Er}_{pB}(\mathrm{t}) = R_p(t) - R_B(t) \qquad \text{式(3.38)}$$

3.5.2 衡量收益

3.5.2.1 总收益率

总收益率是在一定期限内，股票或组合投资获得的回报率。组合 p 在整个时期 T 的复合总收益率为

$$R_p(1,T)=\prod_{t=1}^{T}(1+R_p(t))-1 \qquad \text{式(3.39)}$$

由于总收益率受投资期限的影响，一般而言，投资期限越长收益率越高。不同的组合有不同的投资期限，因此必须把收益率转化为同一时间段才能比较，如同一月份、季度、年等。

3.5.2.2 算术平均收益率与几何平均收益率

算术平均收益率以单利投资相同期限求得的收益率，用各期收益率的均值计算，表示为，

$$AR_p=\frac{1}{T}\sum_{t=1}^{T}R_p(t) \qquad \text{式(3.40)}$$

几何平均收益率是以复利方式投资相同期限得出的单期收益率，计算方法为

$$GR_p=\sqrt[T]{\prod_{t=1}^{T}(1+R_p(t))}-1 \qquad \text{式(3.41)}$$

可以证明，在任何情形下，总有 $Er_p \le Ar_p$，算术平均收益率是未来收益率的预测，而几何平均收益率能更加精确地度量投资业绩。

3.5.2.3 年化复合收益率

年化复合收益率把总收益率转化到“一年”这个维度上，计算方法为

$$AYR_p=\sqrt[\frac{T_0}{T}]{\prod_{t=1}^{T}(1+R_p(t))}-1 \qquad \text{式(3.42)}$$

其中，T_0 表示根据 T 判断的一年中对应的天数、月份数等，如有 T 个月，则 T_0 为 12，T 个季度，则 T_0 为 4，以此类推。

3.5.2.4 超额收益率

股票或组合年化超额收益率为年化复合收益率与无风险利率之差，表示为

$$AYER_{pf} = \sqrt[\frac{T_0}{T}]{\prod_{t=1}^{T}(1+R_p(t))} - AYr_f \qquad 式(3.43)$$

其中， AYr_f 为年化的无风险利率。

3.5.2.5 主动收益率

主动收益率用投资组合的年化复合收益率减去业绩基准组合的年化复合收益率。定义为：

$$AYER_{pB} = \sqrt[\frac{T_0}{T}]{\prod_{t=1}^{T}(1+R_p(t))} - \sqrt[\frac{T_0}{T}]{\prod_{t=1}^{T}(1+R_B(t))} \qquad 式(3.44)$$

T_0 的意义同上。

3.5.3 衡量风险

3.5.3.1 夏普比率（SP）

夏普比率（Sharp Ratio,SP）是超额收益率的均值与标准差的比值，记 u_p 为组合收益率的均值， σ_p 为组合收益率的方差，则

$$\mu_p = \frac{1}{T}\sum_{t=1}^{T} R_p(t) \qquad 式(3.45)$$

$$\sigma_p = \sqrt{\frac{1}{T-1}\sum_{t=1}^{T}(R_p(t)-\mu_p)^2} \qquad 式(3.46)$$

$$S_p = \frac{\mu_p - r_f}{\sigma_p} \qquad 式(3.47)$$

夏普比率反映的是单位投资组合平均收益率超过无风险收益率的程度。夏普比率越大，说明投资组合的单位风险能获得的风险收益越高。

3.5.3.2 信息比率 IR

信息比率(Information Ratio, IR）是评价一个多头策略时所用的衡量指标，衡量单位主动风险所带来的主动收益。定义为：

$$\mu_{Ep} = \frac{1}{T}\sum_{t=1}^{T}(R_p(t) - R_B(t)) \qquad \text{式(3.48)}$$

$$\sigma_{Ep} = \sqrt{\frac{1}{T-1}\sum_{t=1}^{T}(R_p(t) - R_B(t))^2} \qquad \text{式(3.49)}$$

$$IR = \frac{\mu_{Ep}}{\sigma_{Ep}} \qquad \text{式(3.50)}$$

其中 μ_{Ep} 是主动收益率的均值，σ_{Ep} 是主动收益率的标准差。信息比率越高，表示基金有超过市场基准的业绩表现。

3.5.3.3 最大回撤（MDD）

最大回撤(maximum drawdown，MDD)是任一时间段内基金的总资产走到最低点时收益率下降幅度的最大值，用来描述买入产品可能出现的最糟糕的情况。在任意时期t，计算t时期之前的基金最大净值，计算该t时期基金净值相对该最大净值的回撤，整个T时期的最大回撤即为所有回撤的最大值，表示如下：

$$MDD = \max\left(\frac{\max(N(1,t-1)) - N(t)}{\max(N(1,t-1))}\right) \qquad \text{式(3.51)}$$

其中 $N(t)$ 表示t时期的基金净值。最大回撤是基金运作历史上出现过的最大亏损，与成立时间长短有关，是一个随运作时间延长而变大的指标。

3.5.4 胜率(HR)

胜率(hit ratio, HR）是组合战胜业绩基准的概率，用战胜业绩基准的次数与总的交易次数之比来度量,表示如下：

$$HR = \frac{N((R_p(t) - R_B(t)) > 0)}{T} \qquad \text{式(3.52)}$$

如果组合战胜业绩基准的概率较大，则能获取超额收益的概率也较大，反之，如果超配组合战胜业绩基准的概率较小，则组合很可能无法获取超额收益。

第四章　实证分析 I:有效因子的筛选

在第二、三两章的理论基础上，本文开始多因子模型的实证分析。实证分析分为两部分，其中第四章是第一部分，讨论从备选因子库中筛选出最终的有效因子；第五章是第二部分，运用几种不同的加权方法求股票在所有因子上的总分，然后排序构建投资组合。

4.1 模型设定

在开始回测之前，先要对模型进行设定：

测试区间：2007 年-2016 年，总共 10 年。其中样本内回测区间为 2007 年-2013 年，样本外回测区间为 2014 年-2016 年；

股票池：HS300 指数成分股。因为本文的目的不是为了比较在不同指数成分股下因子的有效性，且 HS300 指数中成分股为沪深两市中规模大、流动性好的股票，比较好的反应了沪深股市的行情，所以本文以 HS300 指数成分股为股票池；

备选因子库：本文共选取 7 大类共 29 个财务因子，具体参见 4.2 节；

调仓频率：月度调仓，每月最后一个交易日选股持有至下月最后一个交易日；

股票加权方法：等权重配置

业绩基准：HS300 指数；

数据库：数据全部来源为 Wind；

行业划分方法：申万一级行业；

在取数据时，有些因子，如盈利因子、偿债能力的数据在组合构建日还未发布，只能取最近报告的数据替代，如下表所示。

表 4.1 组合构建日与财务因子报告期

组合构建期	部分财务因子取值日期
1 月、2 月、3 月	去年三季报
4 月、5 月、6 月、7 月	当年 1 季报、去年年报
8 月、9 月	当年中报
10 月、11 月、12 月	当年三季报

按照 3.1.2 节的步骤和 wind 提供的 Matlab 函数接口，编程获取数据，模型的其他

部分也使用 Matlab 编程实现。

4.2 构建备选因子库

目前，根据金融理论、投资理念及数据可获得性等角度考虑，可以利用的因子主要有三类，分别是宏观经济因子、与股票有关的财务指标、行情指标以及根据股票价量关系衍生出的技术指标。本文选取以下因子进入备选因子库：

首先考虑盈利因子。价值投资者根据购买股票价值被低估的股票，其关键是估算股票的内在价值。根据金融理论，股票的内在价值是股票未来收入现金流的贴现值。因而一家公司能否创造利润，是判断一家公司好坏的基本要素。常用的盈利因子主要有销售净利率、毛利率、净资产收益率、资产收益率、息税前利润与营业收入比等。

价值因子是很重要的一类因子。价值类因子表征股票价格的高低，也是价值投资者常用的选股指标。公司是贵还是便宜，可以通过判断公司创造每股收益需要付出的价格，即市盈率来判断。其他与股票估值相关的因子还有市净率、市销率、市现率，PEG、企业价值倍数等。

成长因子是与股票的业绩成长相关的因子。以成长风格著称的彼得.林奇通过购买那些具有较高成长性的股票而获得优秀的投资业绩。常见的成长因子有净利润增长率、营业利润增长率、营业收入增长率、每股收益增长率、净资产增长率、股东权益增长率、经营活动产生的现金流量金额增长率等。

质量因子是与股票的财务质量、资本结构相关的因子，如流动比率、速动比率、存货周转率、总资产周转率等。

规模因子是与股票规模相关的因子。按照已有的检验，股票有规模效应，即小市值的股票往往具有较大的上涨空间，也称小盘股效应。常见的规模因子有总资产、总市值、流通市值、自由流通市值、流通股本、总股本等。

本文共选取七类共 29 个因子，汇总如下表所示，其计算方法见附录 B。

表 4.2 备选因子库

因子类别	序号	指标
盈利因子	1	净资产收益率 ROE（平均）
	2	总资产净利率 ROA
	3	投入资本回报率 ROIC
	4	销售净利率
	5	销售毛利率

续表 4.2 备选因子库

估值因子	6	市盈率 TTM
	7	市盈率（TTM，扣除非经常性损益）
	8	市销率 TTM
	9	市现率（现金流量 TTM）
	10	市净率 LF
	11	企业价值倍数
成长因子	12	经营活动产生的现金流量净额（同比增长率）
	13	净利润（同比增长率）
	14	净资产（同比增长率）
	15	总资产（同比增长率）
	16	净资产收益率（摊薄）（同比增长率）
营运能力	17	固定资产周转率
	18	存货周转率
	19	总资产周转率
	20	流动资产周转率
	21	应付账款周转率
偿债能力	22	流动比率
	23	速动比率
	24	经营活动产生的现金流量净额除流动负债
规模因子	25	总市值
	26	自由流通市值
资本结构	27	流动负债权益比
	28	流动资产除总资产
	29	权益乘数

4.3 单因子有效性检验

在样本内检验期，按照 3.2 节的方法计算因子 *IC* 、分组收益率和相关系数，以做后续之用。对数据的处理及其他相关说明如下：

第一，用“中位数法”去极值；

第二，单因子检验的数据处理只需要到标准化；

第三，以处理后的因子值计算斯皮尔曼相关系数作为 *IC* 。

下面对每个因子的进行分析，以筛选出有效因子。

首先由 *IC* 分布图和分组收益率图粗略筛选出有一定单调性的因子，限于篇幅，正文部分仅展示市净率 LF 的 *IC* 和收益率图，其余因子见附录 C。

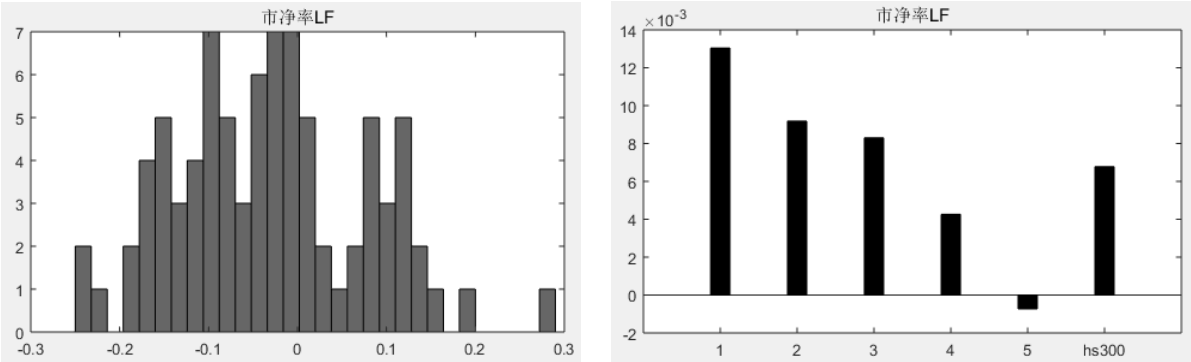


图 4.1 市净率 LF IC 分布图和分组收益率

从上图可以看到市净率 LF 的 *IC* 大部分都是负值，分组收益率单调递减，说明是负向因子，且分组收益率单调性和区分度非常好，可以入选。其他筛选出的因子如下表，其中负号表示负向因子，正号表示正向因子。

表 4.3 初步筛选出的因子

因子类别	序号	指标	方向
估值因子	1	市销率 TTM	-
	2	市净率 LF	-
	3	企业价值倍数	-
成长因子	4	经营活动产生的现金流量净额（同比增长率）	+
	5	净利润（同比增长率）	+
	6	总资产（同比增长率）	-
	7	净资产收益率（摊薄）（同比增长率）	+
营运能力	8	总资产周转率	+
	9	流动资产周转率	+
规模因子	10	总市值	-
	11	自由流通市值	-

筛选出 12 个因子，规模因子全部入选，估值因子入选 3 个，且都为负向因子，偿债因子全部剔除。

计算与 *IC* 相关的统计值，列表如下，可以看到在 10%的显著性水平下，因子全部通过了检验，说明 *IC* 全部异于 0，具有一定的预测能力；除总资产（同比增长率）胜率小于 0.5 外其他全部大于 0.5，说明预测方向较一致；经营活动产生的现金流量净额（同比增长率）、总资产（同比增长率）、流动资产周转率的 *IC* 绝对值小于 0.01 预测能力较弱。市净率 LF 的 *IC* 均值为-0.0314，小于 0 的比率 0.6548，*IR* 为-0.2941，表现除了很强的预测能力。总资产（同比增长率）的 *IC* 小于 0 的比例太低，预测方向与均值不一致，将其剔除；经营活动产生的现金流量净额同比增长率、流动资产周转率的 *IC* 均值小于 0.01，将其剔除，剩余 8 个因子。

表 4.4 因子 IC 相关统计量

因子名称	Mean	std	IR	ratio	ttest
市销率 TTM	-0.0221	0.0727	-0.2618	0.6071	1.0000
市净率 LF	-0.0314	0.0911	-0.2941	0.6548	1.0000
企业价值倍数	-0.0275	0.0606	-0.3958	0.6071	1.0000
净利润（同比增长率）	0.0323	0.0737	0.3883	0.6190	1.0000
净资产收益率（摊薄）（同比增长率）	0.0359	0.0733	0.4502	0.6786	1.0000
总资产周转率	0.0120	0.0534	0.1935	0.5833	1.0000
总市值	-0.0407	0.1675	-0.2028	0.5000	1.0000
自由流通市值	-0.0294	0.1494	-0.1621	0.5357	1.0000

计算与收益率相关的统计因子，考察因子的收益能力，列表如下。在 10%的显著性水平下，因子全部通过了检验，说明优势组合显著地战胜了劣势组合。入选的因子胜率全部大于 0.5，说明收益能力较强。估值因子、成长因子和规模因子的收益较其他因子更强势，大部分接近或超过 1%。总资产周转率的 IR 较低，将其剔除。

表 4.5 因子分组收益率相关统计量

因子名称	mean	std	IR	HR	ttest
市销率 TTM	0.0107	0.0279	0.3852	0.6071	1.0000
市净率 LF	0.0138	0.0361	0.3812	0.6548	1.0000
企业价值倍数	0.0092	0.0276	0.3332	0.6429	1.0000
净利润（同比增长率）	-0.0059	0.0265	-0.2205	0.6190	1.0000

续表 4.5 因子分组收益率相关统计量

净资产收益率（摊薄）（同比增长率）	-0.0093	0.0265	-0.3519	0.6786	1.0000
总资产周转率	-0.0023	0.0225	-0.1013	0.5714	1.0000
总市值	0.0176	0.0562	0.3128	0.5714	1.0000
自由流通市值	0.0131	0.0498	0.2639	0.5833	1.0000

综合考虑因子预测能力与收益能力，初步筛选出以下 7 个因子：

表 4.6 通过单因子有效性检验的因子

因子类别	序号	指标	方向
估值因子	1	市销率 TTM	-
	2	市净率 LF	-
	3	企业价值倍数	-
成长能力	4	净利润（同比增长率）	+
	5	净资产收益率（摊薄）（同比增长率）	+
规模因子	6	总市值	-
	7	自由流通市值	-

4.4 冗余因子的剔除

筛选出 7 个有效因子后，由于同类因子之间存在相关性，因而应进行相关性检验，如 3.2.4 所述，计算出的因子相关矩阵均值 $\overline{\text{corr}Matrix}$ 如下表：

表 4.7 因子相关性均值矩阵

因子	1	2	3	4	5	6	7
1	1.0000	0.4078	0.3843	0.0250	0.0082	0.0287	0.0065
2	0.4078	1.0000	0.3682	0.0329	0.0457	0.1852	0.1406
3	0.3843	0.3682	1.0000	0.0125	0.0031	0.0125	0.0026
4	0.0250	0.0329	0.0125	1.0000	**0.8487**	0.0183	0.0051
5	0.0082	0.0457	0.0031	0.8487	1.0000	0.0331	0.0179
6	0.0287	0.1852	0.0125	0.0183	0.0331	1.0000	**0.8418**
7	0.0065	0.1406	0.0026	0.0051	0.0179	0.8418	1.0000

由表中数据所示，净利润（同比增长率）和净资产收益率（摊薄）（同比增长率）的相关系数为 0.8487，总市值与自由流通市值的相关系数为 0.8418，分别应将其中一个剔除。

净资产收益率（摊薄）（同比增长率）在 IC 和收益率上都优于净利润（同比增长率），因此将净利润（同比增长率）剔除。

虽然总市值比自由流通市值在 IC 和收益率上优于自由流通市值，但自由流通市值能正确的反应股票的大小，且自由流通市值对股票的区分能力较强，因此应剔除总市值。

最终选出的 5 个有效因子：

表 4.8 有效非冗余的因子

因子类别	序号	指标	方向
估值因子	1	市销率 TTM	-
	2	市净率 LF	-
	3	企业价值倍数	-
成长因子	4	净资产收益率（摊薄）（同比增长率）	+
规模因子	5	自由流通市值	-

第五章　实证分析 II:组合构建与业绩评价

基于第四章筛选出的有效因子，本章运用先用等权重方法构建组合，并与业绩基准相比较，以验证多因子模型的有效性。接着，尝试利用 IC 、IC_IR、最大化复合因子单期 *IC* 、最大化复合因子 IC_IR 加权四种因子加权方式构建组合，并比较各组合的绩效。

在开始组合构建之前，先对模型相关要素进行设定：

样本外测试区间：2014 年 1 月-2016 年 12 月，共 36 个月

股票池：HS300 指数成分股

有效因子：共 5 个，见 4.4 节

IC 计算回溯期数：12 期

因子加权方式：等权重、*IC* 、IC_IR、最大化复合因子 *IC* 、最大化复合因子 IC_IR

调仓频率：月频，每月末最后一个交易日构建组合，持有至下月最后一个交易日

股票数：以总分排序后，选取前 10%与后 10%的股票构造多空组合

股票仓位配置：等权重配置股票

初始净值：1

交易成本：忽略

业绩基准：HS300 指数

无风险利率：2%

5.1 等权重加权

以等权重加权因子 zscore，即因子权重为，

$$w=(0.2,0.2,0.2,0.2,0.2)' \qquad \text{式}(5.1)$$

在等权重中，估值因子的比重为 0.6，成长因子为和规模因子都为 0.2，并不区分因子间有效性的差异，估值因子在总分中占的比重更大。

每个月最后一个交易日调整组合中的股票，下表给出 2014 年 1 月组合中的股票。左边为多头组合的股票，右边为空头组合的股票。

表 5.1 2014 年 1 月组合中的股票

000725.SZ	京东方 A	600887.SH	伊利股份
600100.SH	同方股份	002653.SZ	海思科
000876.SZ	新希望	000895.SZ	双汇发展
601168.SH	西部矿业	601001.SH	大同煤业
600688.SH	上海石化	002241.SZ	歌尔股份
601600.SH	中国铝业	600690.SH	青岛海尔
600219.SH	南山铝业	600705.SH	中航资本
601607.SH	上海医药	000598.SZ	兴蓉环境
600027.SH	华电国际	600783.SH	鲁信创投
601098.SH	中南传媒	002038.SZ	双鹭药业
600664.SH	哈药股份	002415.SZ	海康威视
600062.SH	华润双鹤	002310.SZ	东方园林
000768.SZ	中航飞机	601866.SH	中远海发
601618.SH	中国中冶	000581.SZ	威孚高科
600271.SH	航天信息	000061.SZ	农产品
600859.SH	王府井	600637.SH	东方明珠
600694.SH	大商股份	002024.SZ	苏宁云商
600827.SH	百联股份	600406.SH	国电南瑞
600588.SH	用友网络	600276.SH	恒瑞医药
600096.SH	云天化	002450.SZ	康得新
000933.SZ	*ST 神火	000831.SZ	*ST 五稀
600216.SH	浙江医药	002236.SZ	大华股份
601258.SH	庞大集团	002230.SZ	科大讯飞
600362.SH	江西铜业	600315.SH	上海家化
600267.SH	海正药业	600259.SH	广晟有色
600642.SH	申能股份	002353.SZ	杰瑞股份
600655.SH	豫园商城	600111.SH	北方稀土
002001.SZ	新和成	000826.SZ	启迪桑德
600266.SH	北京城建	600256.SH	广汇能源

组合的初始净值为 1，期末净值为 3.1989，总收益率为 219.89%；组合年化收益率为 47.34%，同期 HS300 年化收益率为 12.42%，主动收益率为 34.93%；胜率为 58.33%；夏普比率为 1.2889，同期 HS300 夏普比率为-0.2612，即单位风险无法带来收益；*IR* 为 0.702；最大回撤为 4.79%，同期 HS300 最大回撤为 40.56%，HS300 的损失风险极大。多因子模型的各项指标都优于同期 HS300 指数，显示了极强的选股能力。

表 5.2 多空组合与 HS300 业绩对比

衡量指标	多空组合	HS300
期末净值	3.1989	1.4206
总收益率	2.1989	0.4206
几何平均收益率	0.0328	0.0098
年化收益率	0.4735	0.1242
主动收益率	0.3493	
最大回撤	0.0479	0.4056
胜率	0.5833	
夏普比率	1.2889	-0.2612
信息比率	0.7102	

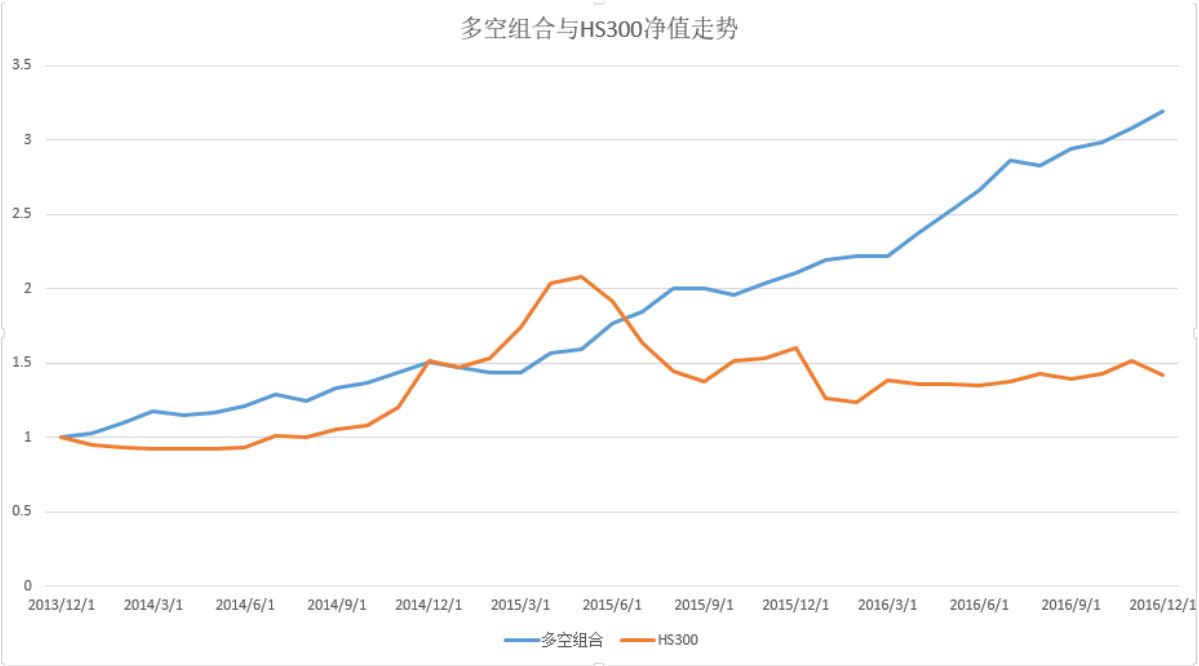


图 5.1 因子等权重下多空组合净值走势

5.2 其他方式加权

5.2.1 IC 加权

以样本内检验期因子 IC 的均值作为权重，突出有效因子的作用。因子 IC 加权的权重为

$$w_{ic} = (0.1509, 0.2147, 0.1881, 0.2007, 0.2456)' \quad \text{式(5.2)}$$

可以看到，市销率、企业价值倍数的权重都下降了，而市净率、自由流通市值和净资产收益率同比增长率的比例却变大了，其中净资产收益率同比增长率权重的增幅最大。

5.2.2 IC_IR 加权

IC 加权只考虑了预测能力的大小，没有考虑稳定性，因而尝试以每个因子的 IR 作为权重，综合考虑因子的预测能力。

$$w_{ir} = (0.1674, 0.188, 0.2531, 0.1036, 0.2879)' \quad \text{式(5.3)}$$

净资产收益率同比增长率的权重进一步加大，但自由流通市值的权重却变小了近一半，估值因子中的企业价值倍数在 IC_IR 加权中变大，说明它的 IC 标准差较小，预测能力较稳健。

5.2.3 最大化复合因子单期 IC 加权

最大化复合因子单期 IC 以因子样本期内的因子协方差矩阵的逆矩阵均值乘 IC 均值向量，得到因子权重向量为

$$w = (0.0369, 0.2361, 0.1341, 0.2378, 0.3551)' \quad \text{式(5.4)}$$

市销率的权重仅为 0.0369，减少近 80%，而净资产收益率同比增长率增加 77%，最大化复合因子单期 IC 突出了净资产收益率同比增长率的作用。

5.2.4 最大化复合因子 IC_IR 加权

最大化复合因子 IC_IR 加权以最大化复合因子的 IR 值为目标，由于 IC 的样本协方差矩阵可能不可逆，因而运用压缩估计的方法估计总体因子协方差矩阵，求出的压缩估计矩阵，最优压缩系数为 0.8140，说明样本协方差矩阵的估计误差极大，因而只给予其 0.1860 的权重。

求出因子权重向量为

$$w_{ir\_\max} = (0.188, 0.188, 0.199, 0.046, 0.4544)' \quad \text{式(5.5)}$$

净资产收益率同比增长率的权重增大至 0.4544，说明其对 IR 的影响极大；而自由流通市值的权重只有 0.046，估值因子的权重则全部下降。

5.3 不同加权方式下组合绩效比较

5.2 节算出了不同因子加权方式下的因子权重向量，构建组合后，得到评价组合的指标如下表

表 5.3 多种加权方式下投资组合绩效比较

业绩指标	Equal_w	IC	IC_IR	IC_MAX	IR_MAX	HS300
期末净值	3.1989	3.7087	3.1117	2.9132	3.1278	1.4206
总收益率	2.1989	2.7087	2.1117	1.9132	2.1278	0.4206
几何平均收益率	0.0328	0.0371	0.0320	0.0301	0.0322	0.0098
年化收益率	0.4734	0.5479	0.4599	0.4282	0.4624	0.1242
主动收益率	0.3493	0.4237	0.3358	0.3041	0.3383	
最大回撤	0.0479	0.0376	0.0496	0.0339	0.0441	0.4056
胜率	0.5833	0.5556	0.5556	0.6389	0.5556	
夏普比率	1.2889	1.5922	1.1110	1.1027	1.1340	-0.2612
信息比率	0.7102	0.9109	0.7119	0.6838	0.6901	

由表中可知，所有加权方式下组合的表现极大的超越了业绩基准。业绩基准的年化收益率只有 12.42%，但多因子模型所构建的投资组合最低的也有 42.82%；业绩基准的最大回撤为 40.56%，说明损失的风险极大，而本文构建的组合最大回撤全部在 5%以下，非常的稳健；多因子模型的夏普比率全部大于 1，而同期 HS300 的夏普比率为负，说明承担单位风险无法给指数带来收益。

主动收益率最大的是 IC 加权，高达 42.37%，但其最大回撤和胜率不是最优的；虽然最优化复合因子单期 IC 的主动收益率只有 30.41%，但其胜率和最大回撤的表现最好。IC_IR 加权在主动收益率、最大回撤方面劣于最大化复合因子 IR 加权，但其信息比率却更高，说明其承担主动风险带来的收益更高。

总体来说，IC 的表现最优，但其承担的风险也更大，在超过业绩基准时能极大的取得收益；等权方式依然是较好的加权方式，其不区分因子有效性的不同，由于因子与

收益率并非因果决定关系，反而有可能取得较好的结果。由于所有的因子优化方法都是基于历史数据的，其假设是将来能重复历史，但股票市场极其复杂，在模型的各种假设下成立的东西在现实中不一定成立，未来与过去总会有不同。

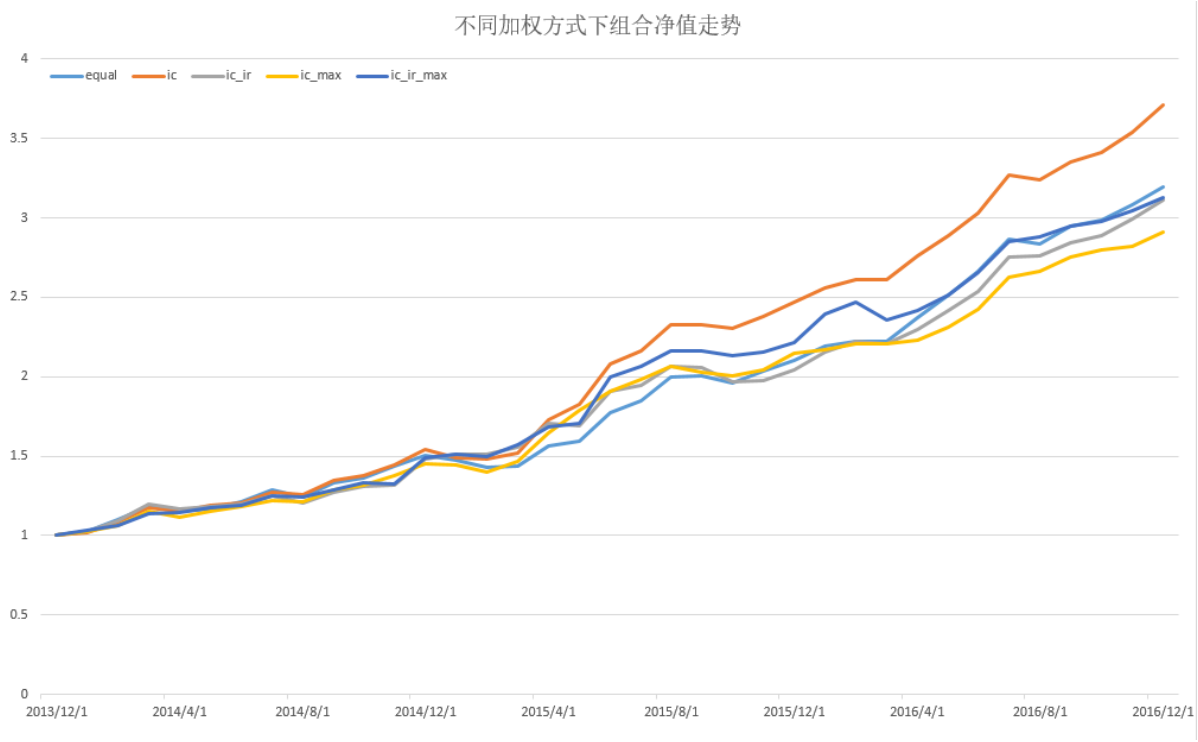


图 5.2 多种加权方式下组合累计净值曲线

第六章　总结与展望

6.1 结论

本文选取基本面盈利因子、估值因子、资本结构等 7 大类共 29 个因子，运用排序的方法，最后确定了可能存在相关性的 12 个有效因子，然后在样本外检验期将这 12 个有效因子用于组合构建。

本文实现了多因子模型从数据获取与处理、单因子有效性分析、到投资组合构建与绩效评价的全过程。从实践中，可以发现多因子模型具有较大的灵活性。从实证过程中，可以得到如下结论：

第一，多因子模型构建的投资组合战胜了业绩基准，这说明多因子模型是适用于中国股票市场的，且中国股票市场并不是完全有效的，主动管理方法大有可为。

第二，从本文的结果来看，所有因子加权方式下的投资组合都优于业绩基准，IC 加权方式总体较优，但各种加权方法各有千秋，并不存在最优的加权方法，采用何种加权方法取决于对收益和风险的权衡。由于对与权重的优化都是基于历史数据的，如 IC 加权是利用过去一段时间的 IC 均值作为权重加权，历史上某个因子的 IC 较高就会给将来的因子赋予更高的权重，但将来因子的效果未必如过去的 IC 所反应出来的那么好，因而对权重的优化可能存在历史最优解，但对将来未必也是最优的。

第三，从挑选出的因子看，估值、成长和规模因子表现了很好的选股能力，表明在中国股票市场低估值和高成长的公司是比较好的投资标的，价值和成长投资也适用于中国股票市场。

6.2 本文不足之处

本文在前人研究的基础上，对沪深 300 指数成分股进行了多因子选股实证分析。但是，由于研究重点及自身条件的限制，本文有许多不足之处：

首先，在因子的挑选上，未加入技术面因子。正因为量化投资结合了基本面与技术面的分析方法，量化投资才能取得非凡的业绩。本文只选取了 29 个财务指标，包括盈利因子、估值因子、成长因子、偿债能力、规模因子、资本结构、营运能力七大类因子，全部属于基本面因子。由于股票市场本质是由人组成的市场，人性的弱点和自身能力的限制也会反映在股市中，因而许多市场异像，如动量和反转效应、日历效应等，因此可以在模型中适当加入行为金融学已有的结论，更深层次地获得 alpha 收益。

其次，由于本文构建的是多空组合，而多空组合容易受大盘的影响。更合理的方式

是构建多空组合后，运用 HS300 股指期货等做空工具对冲市场风险。由于不同因子的有效性随时间及市场情况的不同而变化。在未运用对冲方法对冲贝塔风险时，组合的收益容易受大盘的影响。在组合有正收益时，也不容易区分是受益于市场大势还是由于主动投资经理进行主动管理后的结果。由于主动投资经理的目的是为了获取超过市场基准的超额收益，因而运用不同的对冲方法对冲市场风险对判断业绩归因非常重要。

6.3 未来研究方向

多因子模型逻辑简单，但非常灵活，其中还有许多可以研究的地方。

第一，本文只选取了基本面的财务因子。然而，由于影响股票涨跌的因素非常多，如技术因子、分析师预测因子、宏观因子等。因此，如何选择尽量多的有效因子以更深度挖掘股票 alpha 对提高多因子模型选股效果非常重要。

第二，由于因子在不同的市场行情，不同的风格（如大小盘、周期及非周期）、不同的行业（如机械、农林等）的有效性有较大差别。可供选择的股票池如全市场选股、沪深 300、中证 500、中证 800 等，可以研究同一因子在不同股票池的有效性。同一因子在牛市、熊市及震荡市中的表现可能不同。同一因子在不同的行业，如市盈率对金融行业与对钢铁行业的选股效果可能不同。因此，如何针对不同的市场行情、风格与行业选出有效的因子以运用这些有效因子进行特定股票池、特定行业选股也是一大热门研究方向。

第三，在时间窗口的设定上，本文选择一个月为调仓日。由于调仓的频率对股票交易成本有较大影响。如果调仓过频，可能导致过高的交易成本抹杀取得的收益，导致最终选股效果不好，而如果调仓频率过低，则无法将亏损的股票及时卖出，也无法将好的股票及时买入。因而在实际运用多因子模型时，可以根据需要研究不同时间窗口下选股效果的差异，以选取较适合的时间窗口进行调仓。

第四，在挑选出有效因子之后，本文用将股票按因子值的大小评分，然后计算股票各因子打分的相关性矩阵，通过相关性检验来消除多重共线性。除了本文的方法外，能否用其他方法来消除多重共线性，如运用计量经济学中的一元线性回归，即一个因子对另一个因子进行回归，然后运用统计学的方法对因子相关性作出判断，也是一个研究的方向。

第五，多因子模型目前常用的方法是打分法和回归法。根据已有的研究，多元线性回归的效果比运用打分法选股的效果差。多元线性回归假定股票的收益与各因素是线性关系，因此，可以对线性模型进行改进，研究非线性情况下多因子模型的选股效果。

第六，在因子赋权方法上，除了本文运用的方法外，还可运用其他学科的方法，如

数据挖掘中的聚类分析方法，层次分析法等多种方法对因子赋权，并比较不同赋权方式的优劣。

第七，本文主观上选取前 10%与后 10%的股票作为投资组合。然而，根据经典金融学的理论，组合中的股票数量有助于分散非系统性风险，但股票数量也不是越多越好。在股票数量增加到一定情况下，组合股票数量增加的边际非系统性风险减少量会降低。因此，如何确定组合中股票数量也非常重要。

第八，本文在挑选股票后构建多空组合，这种组合构造方式容易受大盘的影响。由于受中国资本市场发展的限制，目前可以用于对冲股票市场风险的做空工具不多。未来，随着更多衍生品等工具的推出，可以构造股票多空组合，并运用股指期货等工具，获取市场中性的超额收益。

最后，在组合股票仓位的配置上，除给予股票相同权重的等权重赋值方法外。可根据不同的投资，不同的经济环境状况提高某些股票的权重，以突出其优于其他股票的优势。

附录 A 样本协方差矩阵的压缩估计

如正文所述，需要求解最优化方程

$$\min_{\delta} E(L(\delta)) = E(\| \delta F + (1-\delta)S - \sum \|^2) \qquad \text{式(A.1)}$$

要求解上式，首先求压缩目标 F，再求压缩强度δ，记

K：因子数量，$1 \le i \le K$

T：样本期数，$1 \le t \le T$

IC_{it} ：因子 i 在 t 期的 IC 值

$\overline{IC_i}$ ：因子 i 在样本期内的 IC 均值

$$\overline{IC_i} = \frac{1}{T}\sum_{t=1}^{T} IC_{it} \qquad \text{式(A.2)}$$

将总体与样本的变量列于下表：

表 A.1 因子的总体与样本变量

变量	总体	样本
因子协方差矩阵	$\sum$	S
因子协方差	σ_{ij}	s_{ij}
因子间相关系数	$\rho_{ij}, \rho_{ij} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}}$	$r_{ij}, r_{ij} = \frac{s_{ij}}{\sqrt{s_{ii}s_{jj}}}$
因子间平均相关系数	$\bar{\rho}, \bar{\rho} = \frac{2}{(K-1)K}\sum_{i=1}^{K-1}\sum_{j=i+1}^{K}\rho_{ij}$	$\bar{r}, \bar{r} = \frac{2}{(K-1)K}\sum_{i=1}^{K-1}\sum_{j=i+1}^{K} r_{ij}$
固定相关系数矩阵		F
固定相关系数矩阵中的元素	$\varnothing_{ii} = \sigma_{ii}, \varnothing_{ij} = \bar{\rho}\sqrt{\sigma_{ii}\sigma_{jj}}$	$f_{ii} = s_{ii}, f_{ij} = \bar{r}\sqrt{s_{ii}s_{jj}}$

F 即为方程中的压缩目标，下面求解δ。

Frobenius 范式

$$\| z \|^2 = \sum_{i=1}^{N} \sum_{j=1}^{N} z_{ij}^2 \qquad \text{式(A. 3)}$$

根据已有的研究

$$\delta = \frac{\kappa}{T} \qquad \text{式(A. 4)}$$

$$\kappa = \frac{\pi - \eta}{\gamma} \qquad \text{式(A. 5)}$$

因而问题转化为求解π、η和γ三个变量

记

$$\hat{\pi}_{ij} = \frac{1}{T} \sum_{t=1}^{T} \{(IC_{it} - \overline{IC_i})(IC_{jt} - \overline{IC_j}) - s_{ij}\}^2 \qquad \text{式(A. 6)}$$

$$\hat{\theta}_{ii,ij} = \frac{1}{T} \sum_{t=1}^{T} \{(IC_{it} - \overline{IC_i}) - s_{ii}\}\{(IC_{it} - \overline{IC_i})(IC_{jt} - \overline{IC_j}) - s_{ij}\} \qquad \text{式(A. 7)}$$

$$\hat{\theta}_{jj,ij} = \frac{1}{T} \sum_{t=1}^{T} \{(IC_{jt} - \overline{IC_j}) - s_{jj}\}\{(IC_{it} - \overline{IC_i})(IC_{jt} - \overline{IC_j}) - s_{ij}\} \qquad \text{式(A. 8)}$$

则 Ledoit 和 Wolf 证明

$$\hat{\pi} = \sum_{i=1}^{K} \sum_{j=1}^{K} \hat{\pi}_{ij} \qquad \text{式(A. 9)}$$

$$\hat{\eta} = \sum_{i=1}^{K} \hat{\pi}_{ii} + \sum_{i=1}^{K} \sum_{j=1, j \neq i}^{K} \frac{\bar{r}}{2} \left(\sqrt{\frac{s_{jj}}{s_{ii}}} \hat{\theta}_{ii,ij} + \sqrt{\frac{s_{ii}}{s_{jj}}} \hat{\theta}_{jj,ij} \right) \qquad \text{式(A. 10)}$$

$$\hat{\gamma} = \sum_{i=1}^{K} \sum_{j=1}^{K} (f_{ij} - s_{ij})^2 \qquad \text{式(A. 11)}$$

求出最优压缩强度为

$$\hat{\delta}^* = \max\{0, \min\{\frac{\hat{\kappa}}{T}, 1\}\} \qquad \text{式(A. 12)}$$

附录 B 备选因子计算方法

表 A.1 备选因子的计算方法

盈利因子
1.净资产收益率 ROE（平均）= 最近 4 个季度归属母公司净利润/平均净资产
2 总资产净利率 ROA = 最近 4 个季度归属母公司净利润/平均总资产
3.投入资本回报率 ROIC = EBIT（1-税率）/（有息负债+权益）
4.销售净利率 = 净利润/营业收入
5.销售毛利率 = 毛利/营业收入
估值因子
6．市盈率 TTM = 市值/最近 4 个季度归属母公司净利润
7.市盈率（TTM，扣除非经常性损益） = 市值/最近 4 个季度扣除非经常收益净利润
8．市销率 TTM = 市值/最近 4 个季度营业收入
9.市现率（现金流量 TTM） = 市值/最近 4 个季度净现金流量
10.市净率 LF = 市值/当期净资产
11.企业价值倍数 = 剔除货币资金后的企业价值/最近 4 个季度税息折旧及摊销前利润
成长因子
12．经营活动产生的现金流量净额（同比增长率） =（经营活动产生的现金流量净额-上年同期经营活动产生的现金流量净额）/上年同期经营活动产生的现金流量净额
13.净利润（同比增长率） =（当期净利润-上年同期净利润）/上年同期净利润
14.净资产（同比增长率） = (当期净资产-上年同期净资产)/上年同期净资产
15.总资产（同比增长率） =（当期总资产-上年同期总资产）/上年同期总资产
16.净资产收益率（摊薄）（同比增长率） =（当期净资产收益率-上年同期净资产收益率）/上年同期净资产收益率
营运能力
17.固定资产周转率 = 营业收入/(期初固定资产净额+期末固定资产净额)/2
18.存货周转率 = 营业成本/（期初存货净额+期末存货净额）/2
19.总资产周转率 = 营业收入/(期初资产总额+期末资产总额)/2
20.流动资产周转率 = 营业收入/（期初流动资产+期末流动资产）/2
21.应付账款周转率 = 营业收入/（期初应付账款+期末应付账款）/2

续表 A.2 备选因子的计算方法

偿债能力
22.流动比率 = 流动资产总额/流动负债总额
23.速动比率 = （流动资产总额-存货）/流动负债总额
24.经营活动产生的现金流量净额除流动负债 = 经营活动产生的现金流量净额/流动负债
规模因子
25.总市值 = 总股本*股票收盘价
26.自由流通市值 = 自由流通股本*股票收盘价
资本结构
27.流动负债权益比 = 流动负债/所有者权益
28.流动资产除总资产 = 流动资产/总资产
29.权益乘数 = 资产总额/股东权益总额

附录 C IC 分布图和分组收益率图

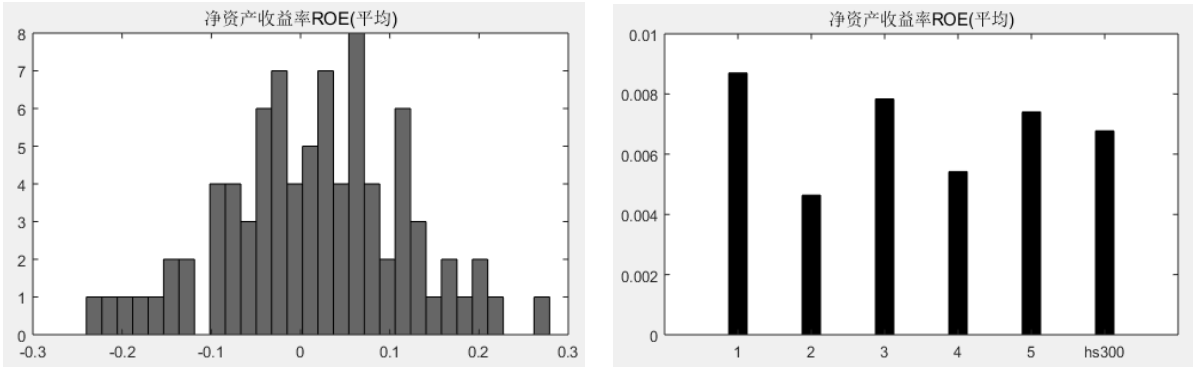


图 B.1 资产收益率 ROE(平均)IC 分布图和分组收益率

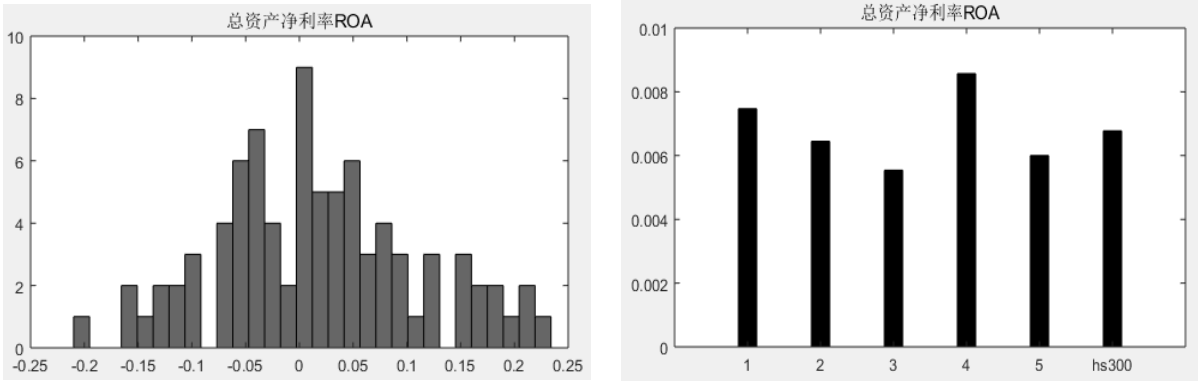


图 B.2 总资产净利率 ROA IC 分布图和分组收益率

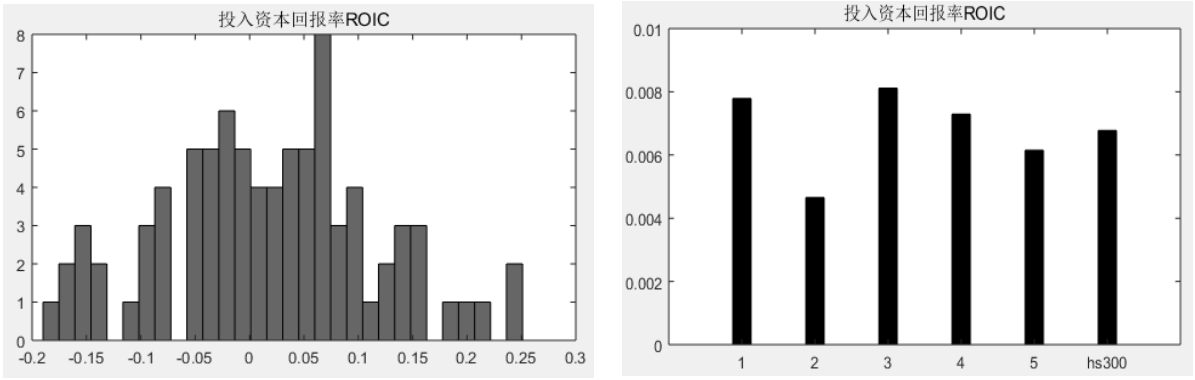


图 B.3 投入资本回报率 ROIC IC 分布图和分组收益率

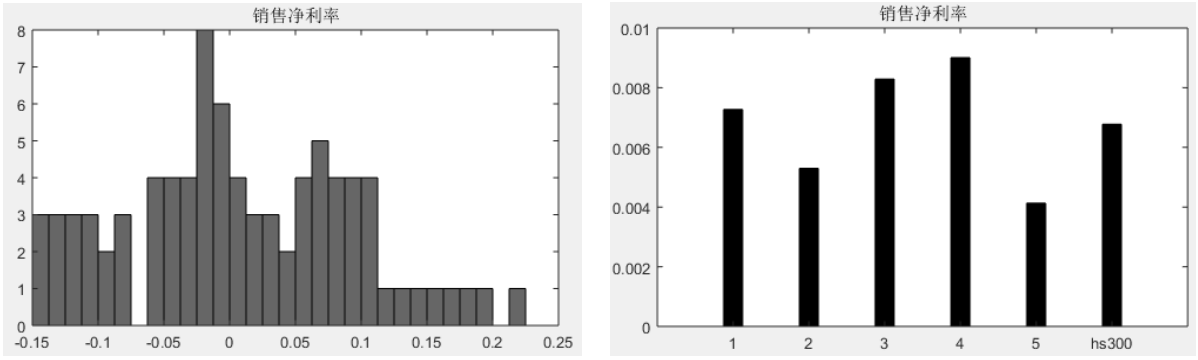


图 B.4 销售净利率 IC 分布图和分组收益率

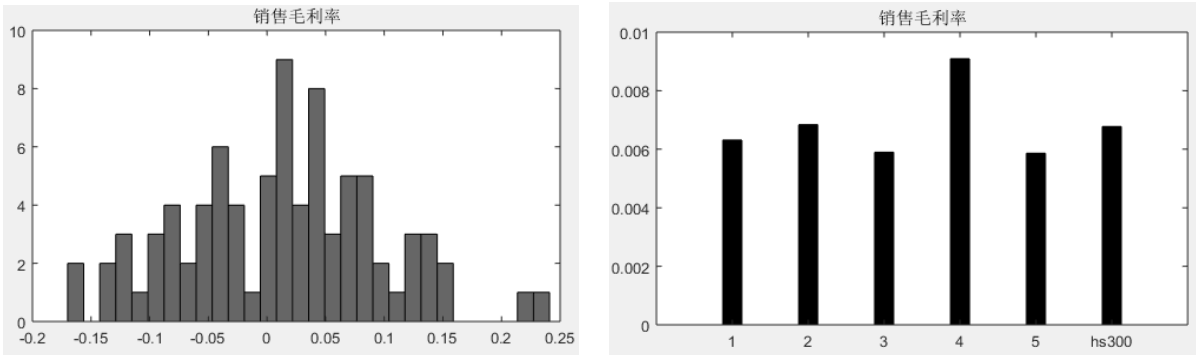


图 B.5 销售毛利率 IC 分布图和分组收益率

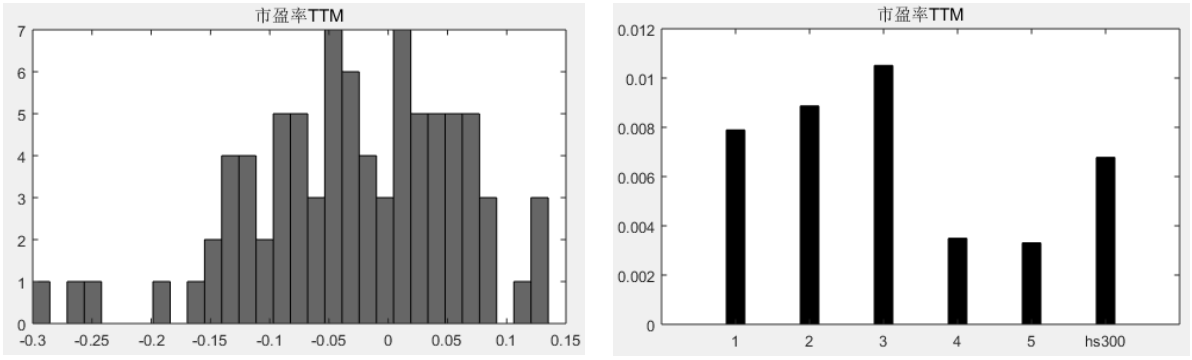


图 B.6 市盈率 TTM IC 分布图和分组收益率

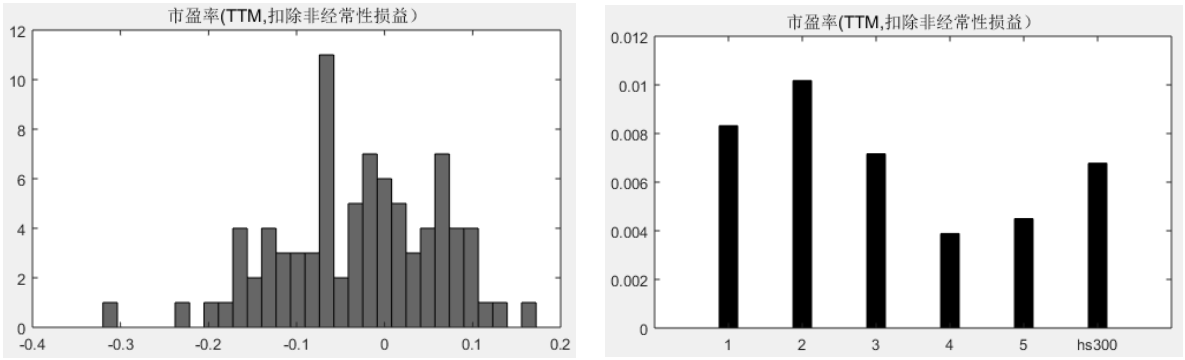


图 B.7 市盈率（TTM，扣除非经常性损益）IC 分布图和分组收益率

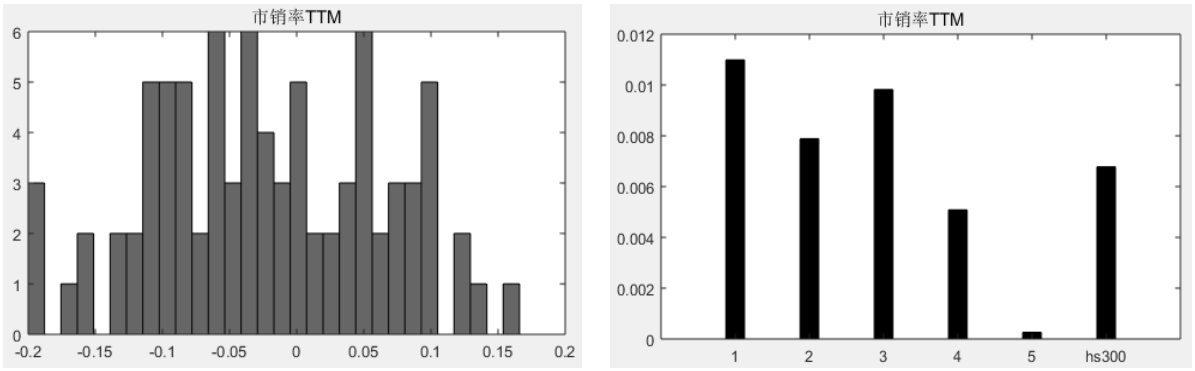


图 B.8 市盈率 TTM IC 分布图和分组收益率

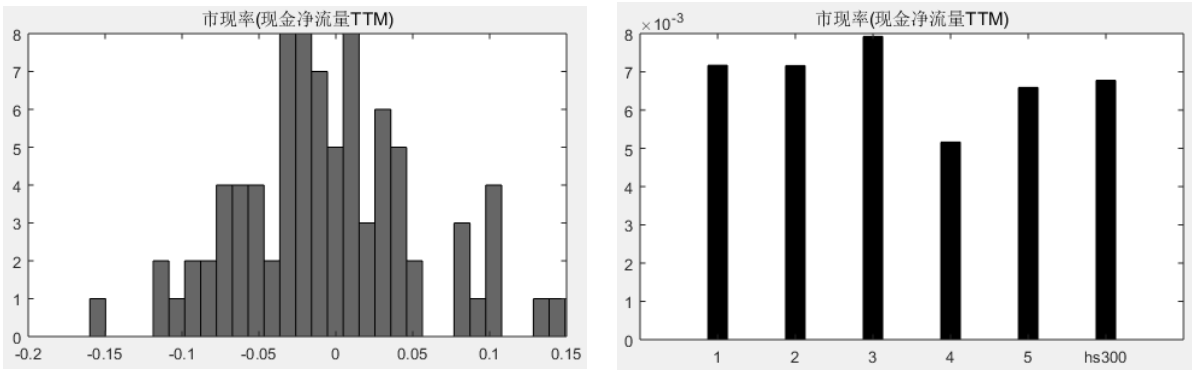
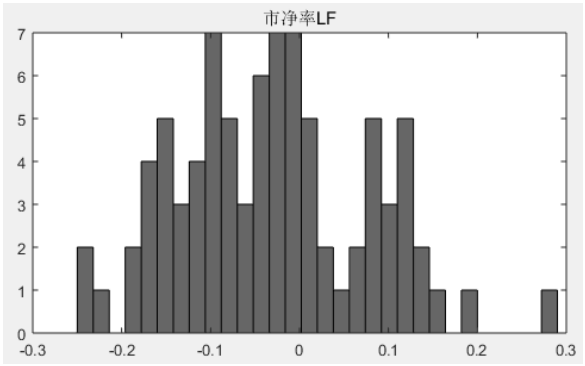


图 B.9 市现率（现金净流量 TTM）IC 分布图和分组收益率



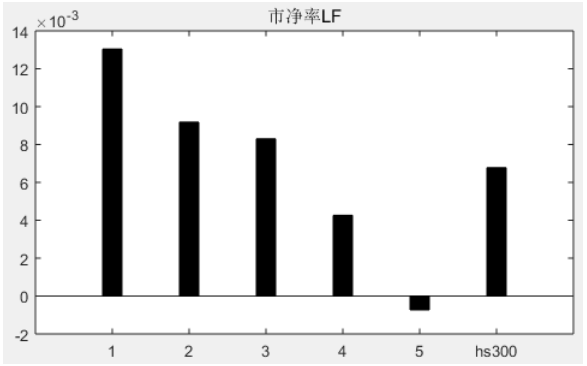
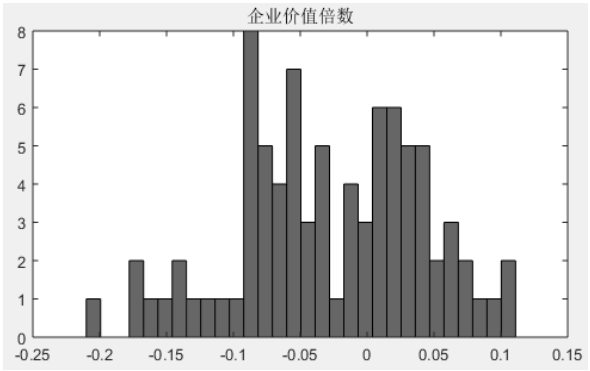


图 B.10 市净率 LF IC 分布图和分组收益率



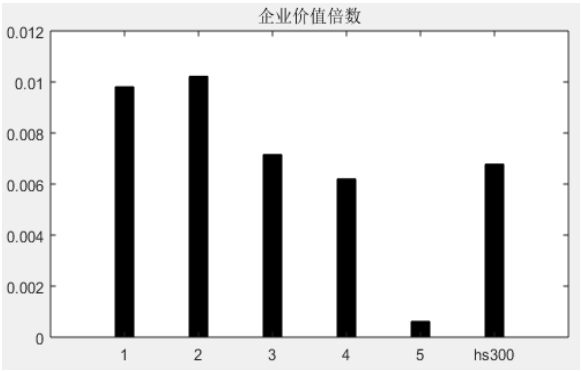
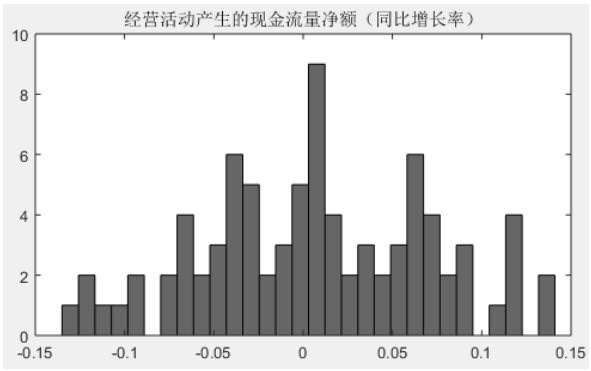


图 B.11 企业价值倍数 IC 分布图和分组收益率



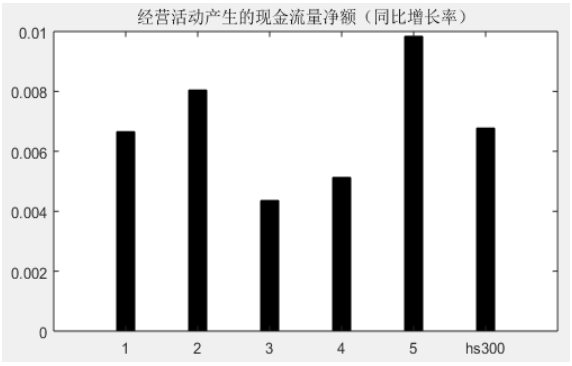


图 B.12 经营活动产生的现金流量净额（同比增长率）IC 分布图和分组收益率

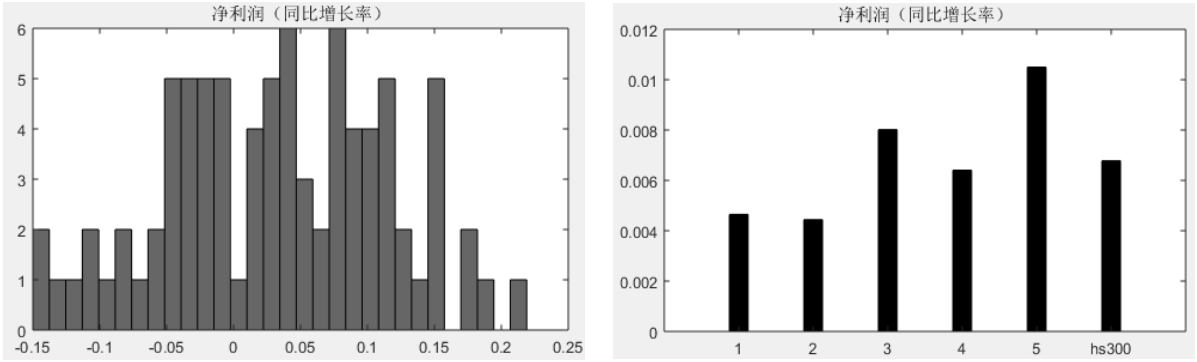


图 B.13 净利润（同比增长率）IC 分布图和分组收益率

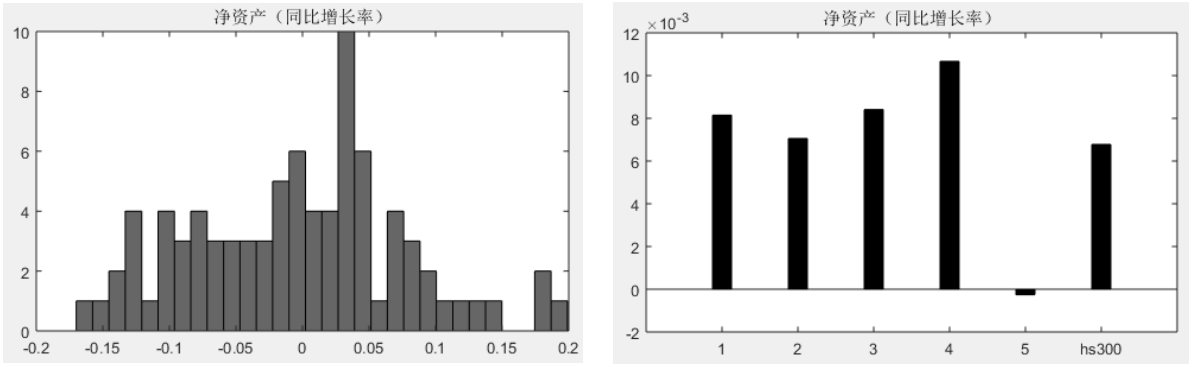


图 B.14 净资产（同比增长率）IC 分布图和分组收益率

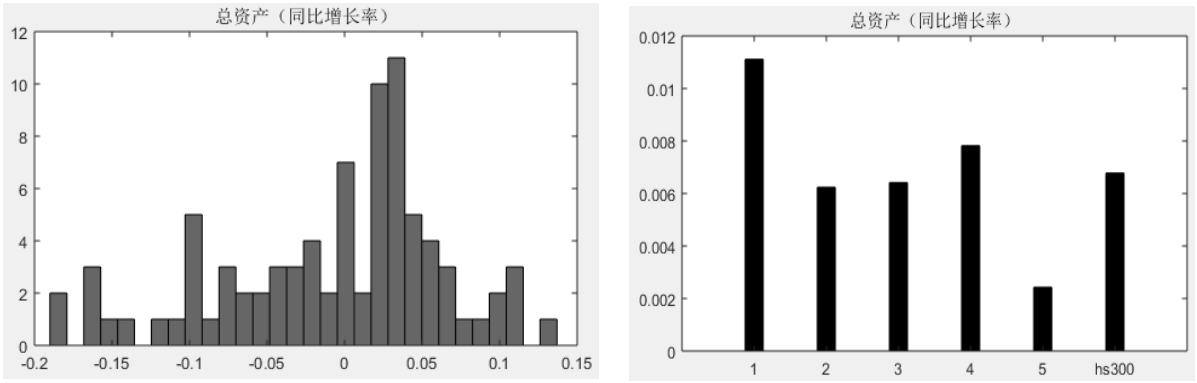


图 B.15 总资产（同比增长率）IC 分布图和分组收益率

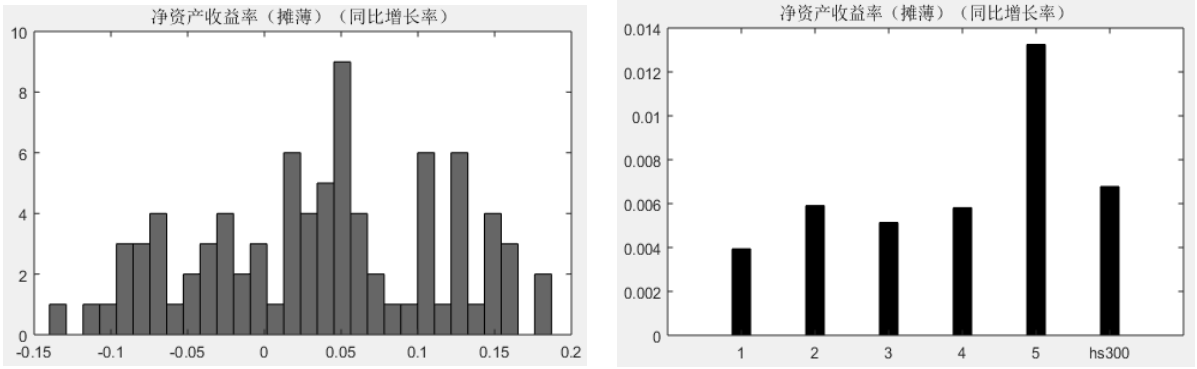


图 B.16 净资产收益率（摊薄）（同比增长率）IC 分布图和分组收益率

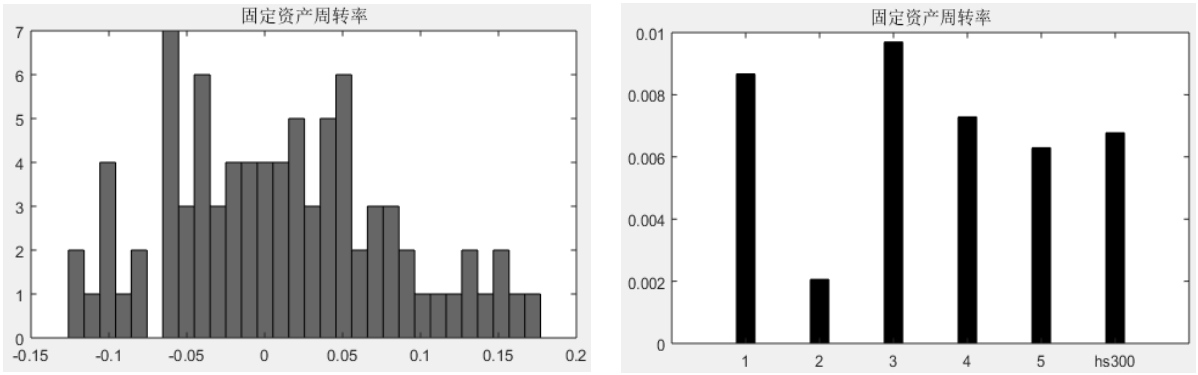


图 B.17 固定资产周转率 IC 分布图和分组收益率

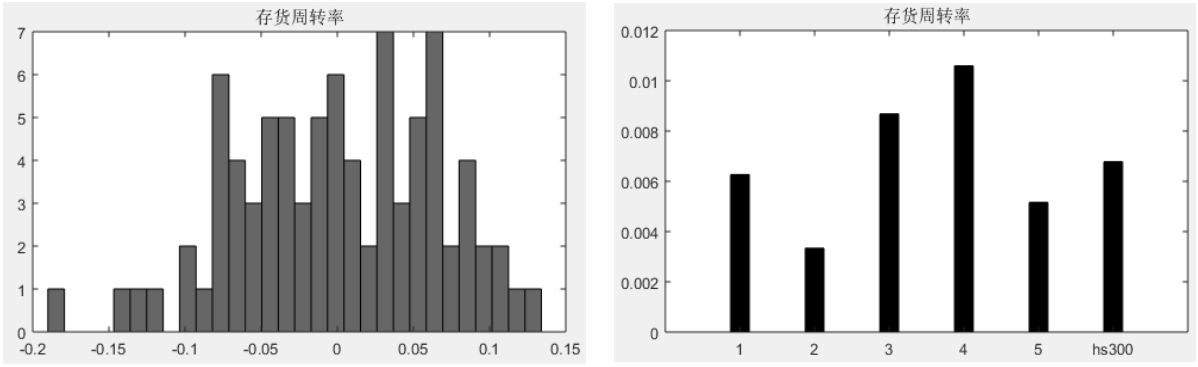


图 B.18 存货周转率 IC 分布图和分组收益率

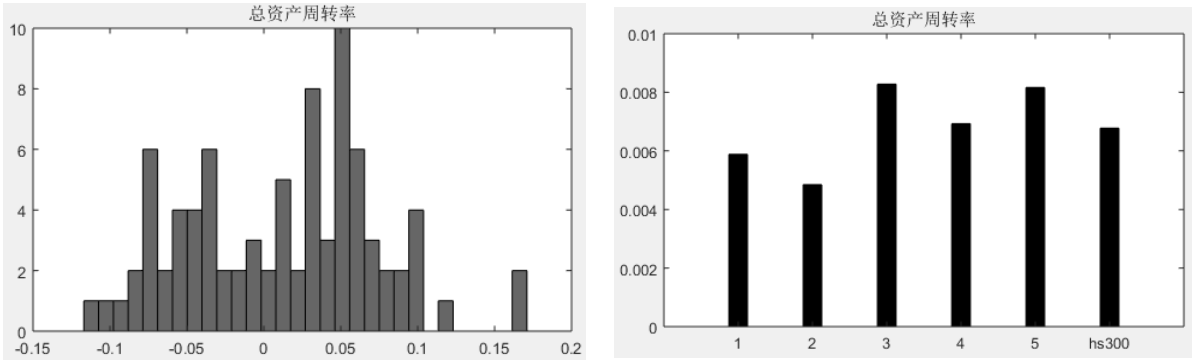


图 B.19 总资产周转率 IC 分布图和分组收益率

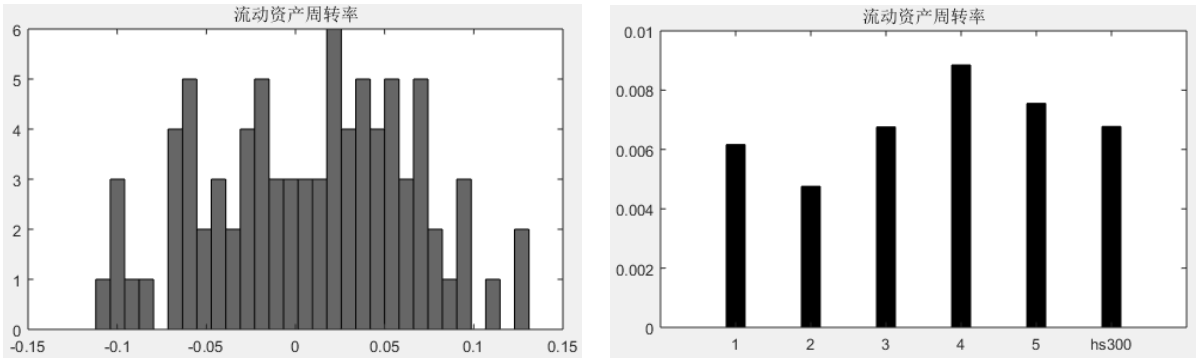


图 B.20 流动资产周转率 IC 分布图和分组收益率

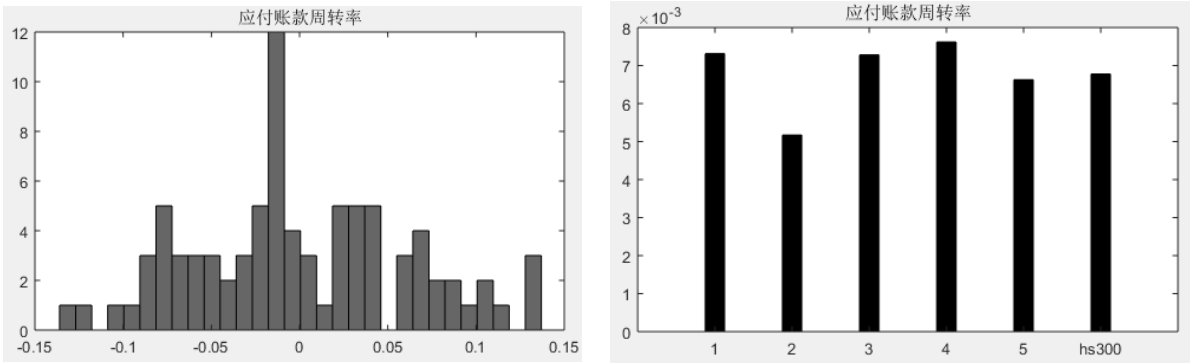


图 B.21 应付账款周转率 IC 分布图和分组收益率

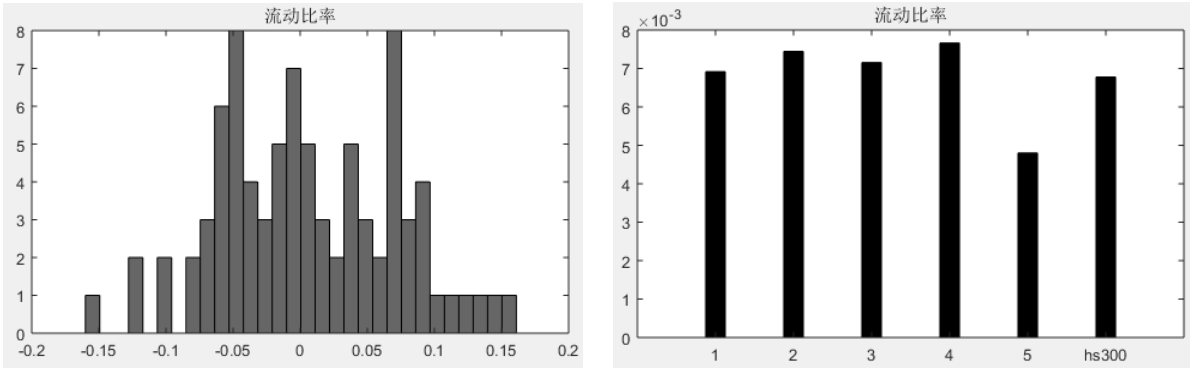


图 B.22 流动比率 IC 分布图和分组收益率

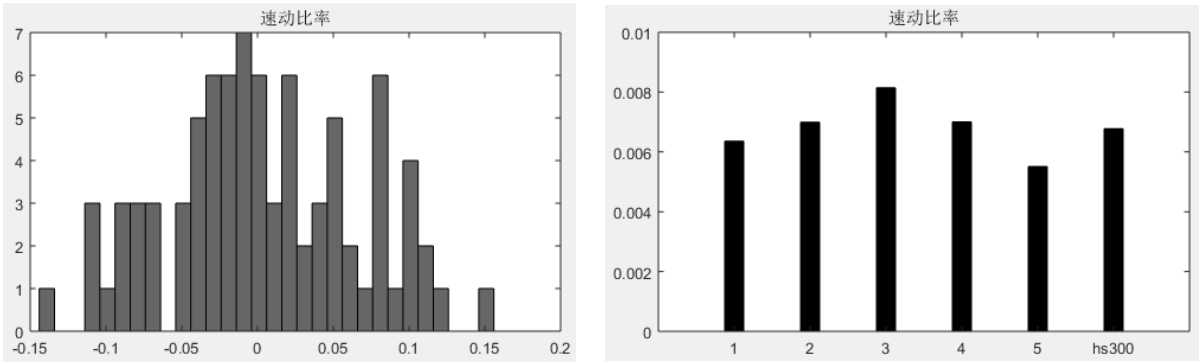


图 B.23 流动比率 IC 分布图和分组收益率

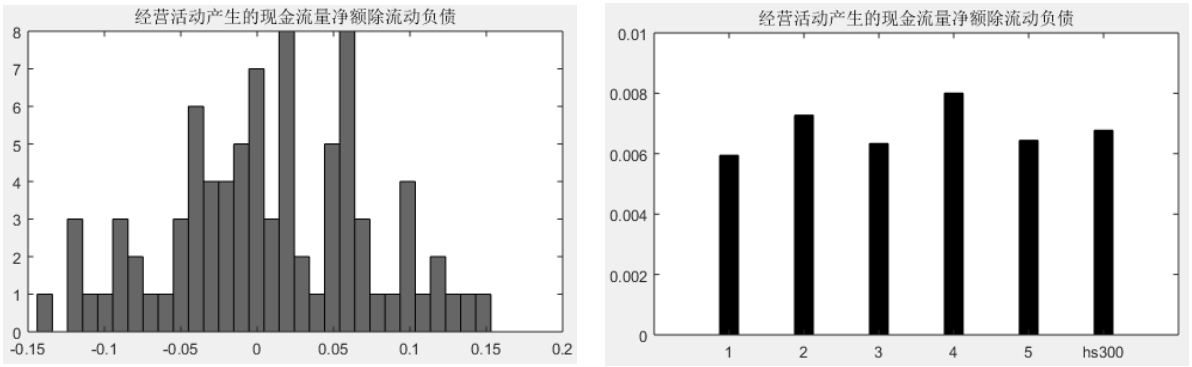


图 B.24 经营活动产生的现金流量净额除流动负债 IC 分布图和分组收益率

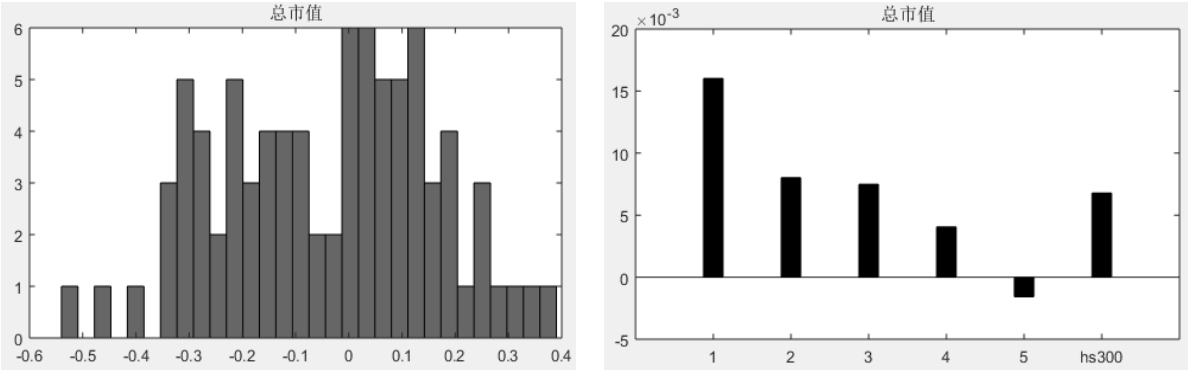


图 B.25 总市值 IC 分布图和分组收益率

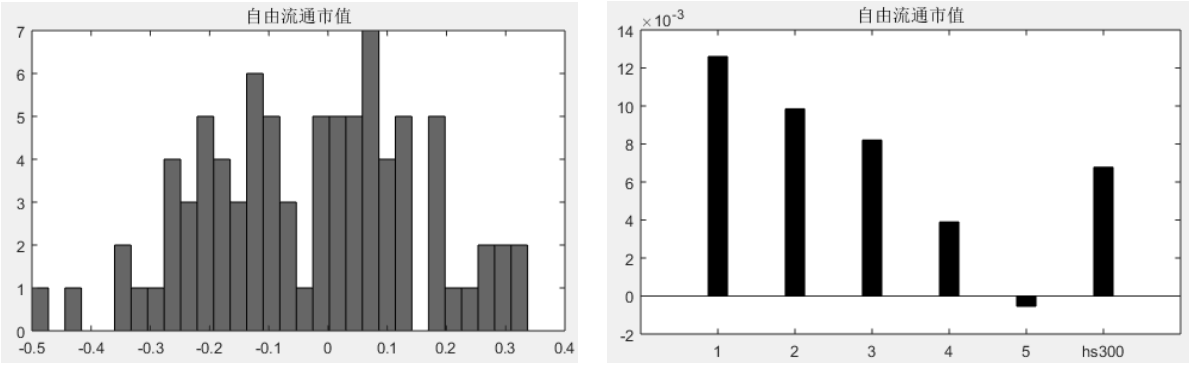


图 B.26 自由流通市值 IC 分布图和分组收益率

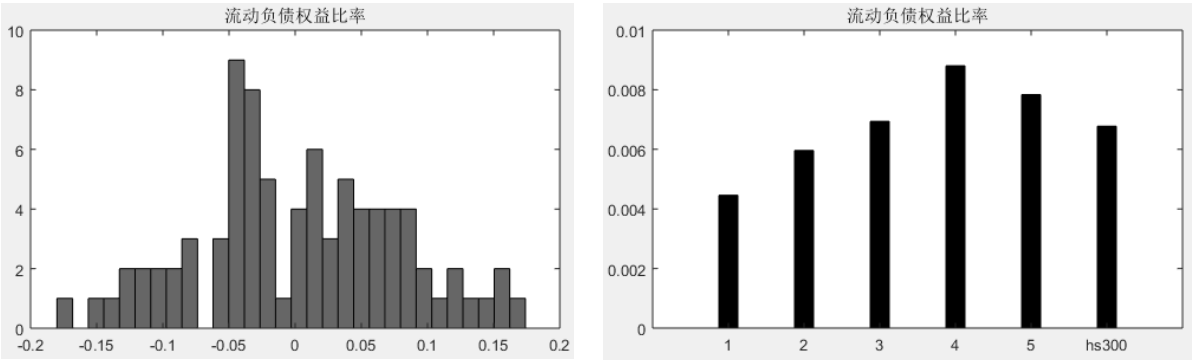


图 B.27 流动负债权益比率 IC 分布图和分组收益率

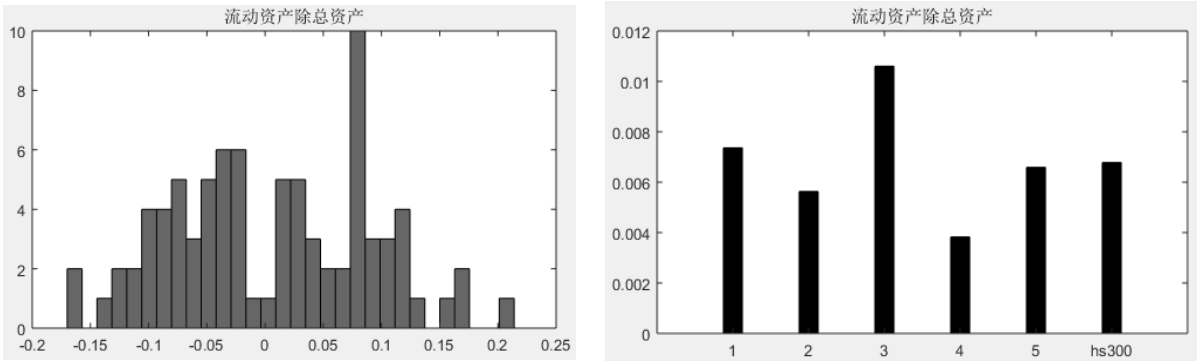


图 B.28 流动资产除总资产 IC 分布图和分组收益率

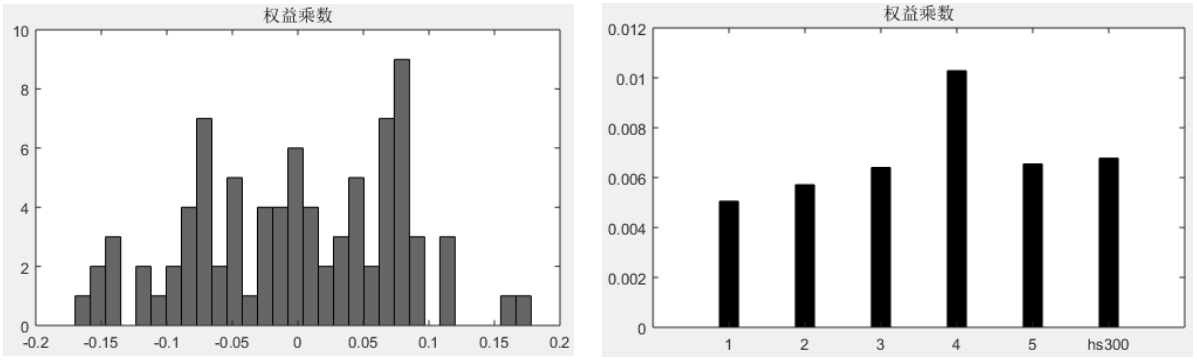


图 B.29 权益乘数 IC 分布图和分组收益率

参考文献

[1] Sharpe W F. Capital asset prices: A theory of market equilibrium under condition of risk [J]. Journal of Finance, 1964, 19(3):425-442.

[2] Stattman D. Book values and expected stock returns [J]. Practical Financial Modeling, 1980，259-261.

[3] Banz RW. The relationship between return and market value of common stocks [J]. Journal of Financial Economics, 1981, 9(1):3-18.

[4] Fama E F, French K R. The cross-section of expected stock returns [J]. Journal of Finance, 1992, 47(2): 427-465.

[5] Fama E F, French K R. Common risk factors in the returns on stocks and bonds [J]. Journal of Financial Economics, 1993, 33(1):3-56.

[6] Asness C S. The interaction of value and momentum strategies [J]. Financial Analysts Journal, 1997, (2):29-39.

[7] Carhart M M. On persistence in mutual fund performance [J]. The Journal of Finance, 1997, 52(1):57-82.

[8] Lev B, Thiagarajan SR. Fundamental information analysis [J]. Journal of Accounting Research, 1993, 31(2):190.

[9] Mohanram P S. Separating winners from losers among low book-to-market stocks using financial statement analysis [J]. Social Science Electronic publishing, 2004, 10(2-3):133-170.

[10] Fama E F, French K R. A five-factor asset pricing model [J]. Journal of Financial Economics, 2015, 116(1):1-22.

[11] 靳乐云，刘霖.中国股票市场的双因子定价模型[J]. 经济科学，2001, (5):92-99.

[12] 范振龙，余世典. 中国股票市场的三因子模型[J]. 系统工程学报，2002,17(6):537-546.

[13] 邓长荣，马永开. 三因素模型在中国证券市场的实证研究[J]. 管理学报，2005, 2(5):591.

[14] 毛小元，陈梦根，杨云红. 配对对股票长期收益的影响：基于改进三因子模型的研究[J]. 金融研究, 2008, (5):114-129.

[15] 刘依明. 基于 alpha 策略的量化投资研究-来自 A 股市场的经验证据[D]. 对外经济贸易大学, 2012.

[16] 勾东宁，王维佳. 基于 Fama-French 三因子模型对我国上市银行股的实证检验[J]. 统计与决策, 2016, (21):158-161.

[17] 李倩，梅婷. 三因素模型方法探析及适用性再检验：基于上证 A 股的经验数据[J]. 管理世界, 2015, (4):184-185.

[18] 章宏帆. 运用量化投资策略实现超额收益 Alpha 的理论与实践[D]. 浙江大学, 2015.

[19] 赵胜民，闫红蕾，张凯. Fama-French 五因子模型比三因子模型更胜一筹吗-来自中国 A 股市场的经验证据[J]. 南开经济研究, 2016, (2):41-59.

[20] 吴敏华. Fama-French 五因子模型在中国 A 股市场的实证研究[D]. 吉林大学, 2016.

[21] 凌士勤，付力. 基于风格轮动的多因子选股模型的实证研究[J]. 商情, 2016, 25.

[22] 黄若冰. 基于 K-Means 聚类的多因子选股模型[J]. 商情, 2016, 34.

[23] Markowitz H. Portfolio selection. Journal of finance [J], 1952, 7(1):77-91.

[24] Lintner J. The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets [J]. The review of economics and statistics, 1965, 13-37.

[25] Sharpe W F. A simplified model for portfolio analysis [J]. Management Science, 1963, 9(1):277-293.

[26] Cohen K J, Pogue J A. An empirical evaluation of alternative portfolio-selection models [J]. Journal of Finance, 1967, 40(2):166-193.

[27] Ross S A. The arbitrage theory of capital asset pricing [J]. Journal of Economic Theory, 1976, 7(2): 5-16.

[28] Fama E F. Efficient capital markets: a review of theory and empirical work [J]. Journal of Finance, 1970, 25(2):383-417.

[29] Fama E F. Market Efficiency, Long-term returns and behavioral finance [J]. Journal of Financial Economics, 1998, 49(98):283-306(24).

[30] Fama E F, MacBeth J. Risk, return, and equilibrium: empirical tests [J]. Journal of Political Economy 1973, 81.

[31] Qian E E, Hua R H, Sorensen E H. Quantitative equity portfolio management: modern techniques and applications [M]. Chapman & Hall/CRC, 2007.

[32] Ledoit O, Wolf M. Improved Estimation of The covariance matrix of stock returns with an application to portfolio selection [J]. Journal of Empirical Finance, 2002, 10(5):603-621.

[33] Ledoit O, Wolf M. Honey, I shrunk the sample covariance matrix [J]. Journal of Portfolio Management, 2003, 30(4):110-119.

[34] Kothari S P, Shanken J, Richard G S. Another look at the cross section of stock returns [J]. Journal of Finance, 1994, 49.

[35] Tortoriello R. Quantitative strategies for achieving alpha [M]. McGraw Hill, 2009.

[36] Grinold R C, Kahn R N. A Quantitative approach for providing superior returns and controlling risk (2th edition) [M]. McGraw Hill, 2000.

[37] Barra. United States equity, risk model handbook [M]. Barra Inc., 1998.

致谢

时光飞逝，转眼就要毕业离开校园了，非常不舍。2012 年 9 月末，我独自一人从广东来到北京，只为了心中伫立的那座博雅塔。在北大虽然只有短短的 3 年时间，却收获了非常多，遇到了许多志同道合的同学和朋友，更有幸能遇到我的导师李杰教授。

还记得当时听师兄说想请李老师当导师的同学非常多，心里还非常紧张，生怕选不上。于是靠着那份无来由的勇气，给老师发了封邮件，东拉西扯地胡乱介绍了一下自己。没想到老师很快就答应当自己的导师了，这正是李老师的可爱可敬之处，为人宽容又热诚。之后的第一堂电子商务课就被老师震撼了。李老师不但课讲的好，条理分明，逻辑清晰，而且给我们分享了许多人生经验，非常受用。与老师平时联络不多，但每次她都能给我很大指导。有一次老师和我们每个人谈心，我把自己积攒许久的问题都一股脑抛给了老师，现在想来有太多冒昧之处，真感觉老师太好了。

本文的写作也得到了老师的许多指导。李老师从确定论文题目、开题、论文格式、写作方法等许多方面都给了我很多的指导。老师从教几十年，学术功底深厚，在一些关键的问题上总能给出建设性的意见。

感谢李老师的宽容和睿智，她让我明白作为一个北大学生应该用自信、乐观、积极的心态来面对生活，在生活中与人为善、广结善缘。

感谢我的父母，他们用自己的辛劳和对生活百折不挠的抗争向我诠释了作为人生活在这个世界所应有的精神力量。感谢我的舅舅，从小对我的指导才让我有幸进入北大。他不但是我的舅舅，还是我的朋友、师兄、校友。感谢我的几位姨妈，在我念书的这几年多我经济上的支持及对我家里的照顾，正是由于你们，才让我没有后顾之忧，心无旁骛地体验在北大的这三年美好的时光。感谢多年来一直支持我的同学和朋友，你们的关心和鼓励让我感受到了友情的力量。感谢自己，多年来那颗不断进取的心！我知道，自己做得还远远不够，还有许多事没有完成，但我会一直努力。

感谢北大!校长林建华说过：“北大从来不只是一座校园，它是人们心中的图腾，是人们的精神圣地”。北大思想自由，兼容并包，在这里，每一位学子都能找到属于自己的地方，或学术、或社团、或发展自己的兴趣。感谢北大深厚的学术氛围和丰富多彩的校园生活，让我拓宽了视野，得到了精神的极大满足。感恩北大，希望自己在以后的人生道路上始终带着北大人那份特有的情怀，永远不忘北大精神。

北京大学学位论文原创性声明和使用授权说明

原创性声明

本人郑重声明：所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不含任何其他个人或集体已经发表或撰写过的作品或成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。本声明的法律结果由本人承担。

论文作者签名：日期：年月日

学位论文使用授权说明

（必须装订在提交学校图书馆的印刷本）

本人完全了解北京大学关于收集、保存、使用学位论文的规定，即：

- 按照学校要求提交学位论文的印刷本和电子版本；
- 学校有权保存学位论文的印刷本和电子版，并提供目录检索与阅览服务，在校园网上提供服务；
- 学校可以采用影印、缩印、数字化或其它复制手段保存论文；
- 因某种特殊原因需要延迟发布学位论文电子版，授权学校□一年/□两年/□三年以后，在校园网上全文发布。

（保密论文在解密后遵守此规定）

论文作者签名：导师签名：

日期：年月日