Operations Research and Fuzziology 运筹与模糊学, 2023, 13(6), 6204-6210
Published Online December 2023 in Hans. https://www.hanspub.org/journal/orf
https://doi.org/10.12677/orf.2023.136614

Hans汉斯

基于沪深300指数的多因子选股模型研究

赵恒江

贵州大学经济学院，贵州 贵阳

收稿日期：2023年9月20日；录用日期：2023年11月29日；发布日期：2023年12月7日

摘　要

量化投资的方法在国外已被广泛使用，依数学模型投资的优势可以避免个人投资情绪影响，这使得量化投资在国内有良好的发展势头。本文以沪深300指数各成分股为研究对象，通过爬虫抓取研究对象的各金融指标，建立基于回归法的多因子量化选股模型，根据得到模型计算收益率挑选成分股，建立证券组合，最后根据投资组合的平均收益率评价量化投资选股模型。通过研究得出：基于多因子模型进行量化选股可取得超过基准市场收益，这说明量化投资在一定条件的有效性，可以给投资者提供更加有效的投资组合和建议。

关键词

沪深300，量化投资，多因子模型，主成分因子

Research on Multi-Factor Stock Selection Model Based on HS 300 Index

Hengjiang Zhao

School of Economics, Guizhou University, Guiyang Guizhou

Received: Sep. 20th, 2023; accepted: Nov. 29th, 2023; published: Dec. 7th, 2023

Abstract

Quantitative investment method has been widely used in foreign countries. The advantage of investment based on mathematical model can avoid the influence of individual investment emotions, which makes quantitative investment have a good momentum of development in China. This paper takes the constituent stocks of HS 300 index as the research object, constructs a multi-factor quantitative stock selection model based on regression method by crawling various financial indicators of the research object, and selects constituent stocks according to the obtained model to calculate the return rate, establishes a securities portfolio, and finally evaluates the quantitative

文章引用: 赵恒江. 基于沪深 300 指数的多因子选股模型研究[J]. 运筹与模糊学, 2023, 13(6): 6204-6210.
DOI: 10.12677/orf.2023.136614

investment stock selection model according to the average return rate of the portfolio. The result shows that quantitative stock selection based on multi-factor model can achieve more returns than the benchmark market, which shows the effectiveness of quantitative investment under certain conditions, and can provide investors with more effective investment portfolios and suggestions.

Keywords

HS 300 Index, Quantitative Investment, Multi-Factor Model, Principal Component Factor

1. 引言

量化投资在国外已经发展了较长时间，尤其是大型投资机构已将其作为进行投资决策的方法，量化投资逐渐趋于成熟，而在中国，量化投资是近年来才开始崭露头角的，并取得了较快发展。简单来说，量化投资使用现代数学和统计学的方法，从大量数据中找寻规律而获得超额收益率。量化投资的目标是通过对历史数据的研究、分析和模拟来预测未来的结果。这种策略严格按照计算机程序设定的数学模型进行投资，可有效避免人的主观情绪及其他因素的影响，并且计算机相对于人有更加强大的计算能力。随着大数据和人工智能不断深入社会生活的各个方面，国内量化投资者对于如何发挥机器学习在量化交易中的优势进行了较多研究。由于机构投资者雄厚的实力，目前主要是国内部分机构投资者开始运用量化投资开展资产管理业务，提高资产管理业绩。各类投资基金取得了较快发展，量化投资在实践运用得到了越来越多机构和投资者的青睐。

量化投资在国外从最初的技术分析方法，到后来发展成为以金融理论为支撑的金融工具，取得了巨大的成功。量化投资与定性投资的区别在于，量化投资是在定性的基础上建立模型，凭借计算机强大的处理数据能力，全面精确实现理性投资，克服资产管理人的心理等因素的干扰。金融量化投资利用金融产品(股票、债券、期权、期货等)从交易终端市场价格的异常波动中获利，实现跨市场、跨周期、甚至跨产品的套利交易。

和国外成熟市场相比，中国 A 股市场发展历史较短，上海证券交易所和深圳证券交易所成立仅仅才 30 多年，投资理念不够成熟，投资者以个人投资者为主，市场常常不是反应不足就是过度反应。随着 A 股发行股票数量不断增加，基金规模不断扩大，量化投资能有效避免非理性市场和不合理目标造成的负面影响，通过量化交易模型获得超额收益，这表明量化投资将会成为中国未来投资策略的一个发展新趋势。

2. 文献综述

国内外学者对因子的选择进行了深入的拓展，利用机器学习进行模型优化来选择股票或预测股价。Fama 和 French 在其三因素模型的基础上做了进一步研究，发现利润因素和增长因素可以用来解释股票的风险，并且比原来的三因素模型有更好的解释力，评价股票更加准确，并认为这两个因素解释了企业面临的部分风险[1]。Guerard 在前人的基础上进一步研究，从金融分析师预期数据、动量因子和基本面数据三个因子来构建模型，该模型具有很好的适用性，证明了该模型适用广泛[2]。Gu *et al.*解决了市场现有

因子数量繁多且部分因子间高度关联的问题，通过树的算法和神经网络算法改进风险溢价度量，简化的资产定价经济机制研究表明投资回报上机器学习模型优于传统多因子模型[3]。

相较而言，国内关于量化投资的研究起步较晚，初期只是简单将国外的研究方法运用于国内市场，主要是通过实证检验模型的有效性以及分析国内外差异的原因。方浩文讲述了量化投资的历史演变过程，接着分析其发展趋势，从总体上概括道来量化投资的机理与分类，为我国量化投资发展提供建议[4]。郭喜才认为量化投资在我国刚刚起步，还存在着不足，必须进行风险控制和加强监管促进其有效发展[5]。彭志认为量化投资和高频交易带来流动性，不仅如此，还进一步分析也带来了市场风险，这对于金融市场而言增加了系统性风险，需要加强监管[6]。林德发和杨潇宇基于多因子模型得到的投资组合的收益率表明组合收益率超过了沪深 300 指数，说明了模型是可行的[7]。王淑燕改进焦健的六因子模型得到八因子模型，八因子模型比六因子模型更具有准确地预测效果，用随机森林算法预测，说明了该模型良好的性能[8]。李文星和李俊琪刘宇轩等将中国金融状况指数(FCI)引入多因子选股模型之中，构建了基于金融周期的行业轮动多因子选股模型，引入该因子后提升了模型的表现[9]。

从国内外学者研究来看，国外关于量化投资的研究已经形成了较为成熟的体系，国内主要是将国外较为成熟的交易模型应用于国内市场，或者将机器学习与传统的多因子模型进行比较分析。最初研究重点关注因子的选取和模型的适用性，目前学者越来越重视模型的实用性。即使相同的因子在不同的时间和市场中表现不尽相同，因此对因子的选取及调整优化是很有必要的。本文以沪深 300 指数成分股为研究对象，结合当下热门的人工智能算法，对传统的多因子模型进行改进，建立基于回归法的多因子量化选股模型，根据模型筛选出来的证券投资组合的表现来衡量量化选股模型。

3. 研究设计

3.1. 数据选取

本文使用的成分股数据指标包括了已知的大部分指标类型，由每股收益、净资产收益率、总资产收益率等共计 52 个指标。通过对这些指标进行主成分分析，这些指标是根据爬虫得到地数据指标来的，由这些指标构建了 12 个主成分因子，本质上主成分因子的指标是相同的，只是贡献率有所不同，再对 12 个主成分因子回归，形成多因子回归量化模型。

3.2. 数据标准化处理

本文获取的数据包含多个范畴的指标，这些指标的取值范围和测量单位也不同。通过这种数据处理方法可以使沪深 300 指数成分股在各个候选指标上的数据具有相同的单位和参照标准，并能准确地反映各成分股数据在该指标数据总体中所处位置，有助于投资者充分了解沪深 300 指数成分股在各个指标上的差异情况。

4. 实证分析

4.1. 基本概念

由于连续复利收益率的概率分布接近于正态分布，对于实际验证的简化具有相当大的帮助，本文使用连续复利收益率(对数收益率)衡量股票投资收益率。此外，本文还引入超额收益率指标进适应性检验时，股票超额收益率是以沪深 300 指数年收益率为基准，在数值上等于个股年收益率减去基准年收益率，代表个股跑赢股票市场取得超过市场部分的收益率。

当增长潜力为正时，其数值越大，说明在新的年度报告公布之前，该只股票价格很可能进一步向上增长。通过对个股的潜在增长潜力进行排序，看看各支股票在整体中的表现，筛选出排名靠前的股票构

建投资组合，以这些股票组成投资组合可以帮助投资者进行理智而有效的投资。本文根据一般的研究经验，选择排名靠前的 10 只股票构建投资者的投资组合，并用该组合对模型进行检验。

4.2. 主成分分析

首先，在进行模型构建前，利用 python 软件抓取 2022 年指数成分股标准化后的各变量数据，对数据进行标准化处理后进行 KMO 和 Bartlett 检验，主要是看看因子之间的相关性，检验结果如表 1 所示。

Table 1. KMO and Bartlett test results
表 1. KMO 和 Bartlett 检验结果

KMO 取样适切性量数	0.78204	
巴特利特球形度检验	近似卡方	59979.39483
	显著性	0.000

该统计量为 0.782，如果 Bartlett 球度检验判断相关阵是单位阵，这不符合一般的判断条件，说明各变量独立，就无法进行因子分析，本文的检验方法就无效，这里对应的 P 值小于 0.05，

如果一个特征根的值明显低于前一个特征根的值，那么特征根变化显著，并且这个特征根很小，后面的特征根变化也很小，说明后面添加的因素具有很小的解释力，那么前几个特征根则是所需要提取的公共因数。从图 1 可以看出，后面因素添加对信息增加作用不大，从第 10 个因素开始，后面因素的影响变得较小，可进一步根据主成分分析确定因素，以达到合理模型需要的解释力。

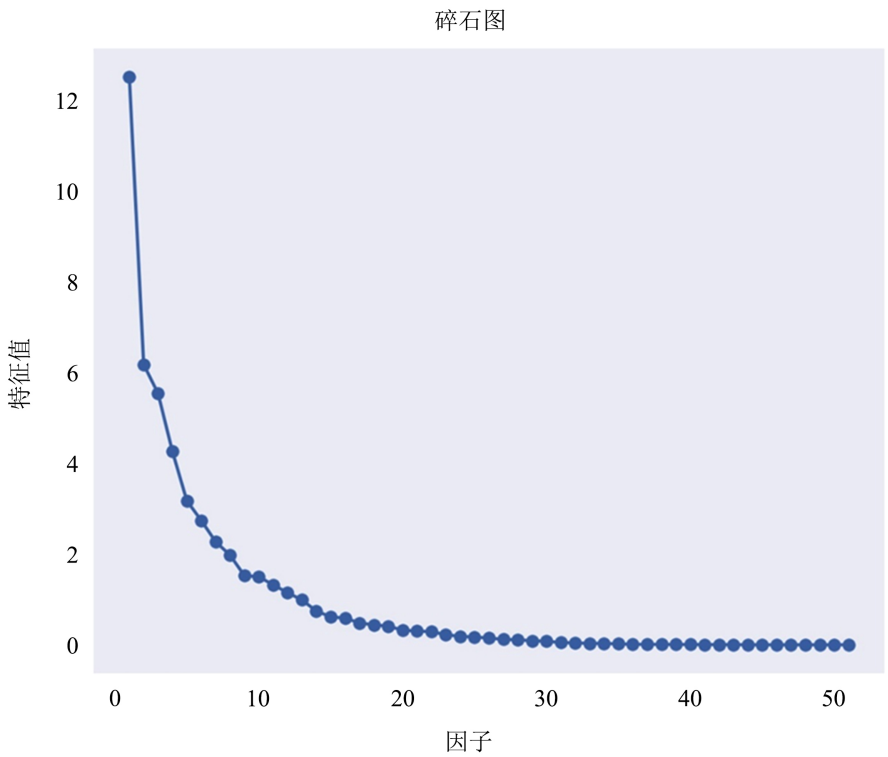

Figure 1. Scree plot
图 1. 碎石图

本文根据主成分分析选择特征根大于 1 的所有因数，然后统计这些计算出累积贡献率为 87.10%，根据经验分析达到 85%以上解释力就可以，这满足模型需要的解释力，共确定 12 个主成分因子，并得到各变量之间旋转成分矩阵。

Table 2. Rotating component matrix (part)
表 2. 旋转成分矩阵(部分)

指标名称	Factor1	Factor2	Factor3	Factor4	Factor5	Factor6	Factor7	Factor8	Factor9	Factor10	Factor11	Factor12
roe_yearly	0.8277	−0.0194	0.1548	−0.0246	0.1409	0.3644	−0.2765	−0.0439	0.0066	−0.0692	0.0144	−0.0095
roa_yearly	0.8973	−0.0255	−0.0473	−0.2350	−0.0654	0.2550	−0.0806	−0.0552	0.0165	−0.0403	0.0279	−0.0458
total_revenue	0.1176	0.6050	−0.4030	0.4028	0.1407	0.1218	−0.2114	0.0086	−0.1191	0.0968	0.0101	−0.0709
revenue_ps	0.1166	0.6005	−0.4051	0.4042	0.1429	0.1256	−0.2104	0.0088	−0.1120	0.0980	0.0124	−0.0721
undist_profit	0.2821	0.8858	−0.0701	0.1013	−0.1629	−0.1493	−0.0003	−0.0589	0.0196	−0.0153	−0.0247	0.0402
extra_item	−0.1093	0.1498	−0.0804	0.0856	0.0083	0.0637	−0.1178	0.1277	0.1878	0.0843	−0.8291	−0.0205
profit_dedt	−0.1278	0.3283	0.6135	0.2173	0.0296	0.4249	0.3863	−0.0120	−0.0001	−0.0240	−0.1283	0.0130
assets_turn	0.3064	−0.0491	−0.5897	0.1987	0.0860	0.4226	−0.1157	−0.1278	−0.1663	−0.0359	0.0659	−0.0509

由表 2 来看，主成分因子 F_1 与总资产收益率、净资产收益率有较强的相关关系，与其他指标间的关系较弱；各指标在其他主成分因子中的相关强度也不同，虽然是相同指标构成的因子，但是贡献率不一样，但是不同的主成分因子与同一指标的相关性也不同。

4.3. 量化模型构建

以净资产收益率(roe)为因变量，以主成分因子 F_1、F_2、F_3、F_4、F_5、F_6、F_7、F_8、F_9、F_{10}、F_{11} 和 F_{12} 为自变量，进行回归得到系数，得到基于回归法的多因子量化选股模型，回归结果见表 3。

Table 3. Regression results of multi-factor quantitative stock selection model in 2022
表 3. 2022 年多因子量化选股模型回归结果

变量名称	F_1	F_2	F_3	F_4	F_5	F_6	F_7
OLS	14.1662^{***}	1.3217^{***}	−0.0514	2.1897^{***}	3.0194^{***}	−0.0560	−0.0205
变量名称	F_8	F_9	F_{10}	F_{11}	F_{12}	const	R^2
OLS	0.6307^{***}	0.3925^{**}	1.5241^{***}	−0.1858	−0.0922	16.2836	0.946

由表 3 看，12 个主成分因子中 F_3、F_6、F_7、F_{11} 和 F_{12} 的回归系数均不显著，它们对因变量的解释力很小，可以不考虑这些因数，故将这 5 个因素除去，以净资产收益率为因变量，对剩余的 8 个主成分因子重新建立选股模型。模型的估计方程为：

$$Y = 16.2836 + 14.1662F_1 + 1.3217F_2 + 2.1897F_4 + 3.0194F_5 + 0.6307F_8 + 2.1897F_4 + 0.3925F_9 + 1.5241F_{10}$$

然后对每只股票的增长潜力进行排名，看看股票表现如何，选取排名靠前的前 10 只股票建立股票证券组合。检验该投资组合的收益，看看是否超过沪深 300 指数收益，从而检验模型的实际投资情况。

Table 4. Portfolio performance
表 4. 证券组合投资表现

股票名称	实际收益	股票名称	实际收益
江西铜业(600362)	0.0869	斯达半导(603290)	0.1294
贝泰妮(300957)	0.2899	合盛硅业(603260)	0.5510
同花顺(300033)	0.3267	卓胜微(300782)	0.4144
万泰生物(603392)	0.5759	圣邦股份(300661)	0.3587
新希望(000876)	−0.2614	海大集团(002311)	0.1123
平均复合收益率	0.2584	基准年收益率	0.0383
超额收益率	0.2201		

由表 4 看，模型所选择的投资组合获得的平均复合收益率为 25.84%，超额收益率为 22.01%，表明了多因子量化选股模型在实证检验中是可行的。此外该模型构建的投资组合整体投资效果更好，在这个投资组合中仅有一只股票是负收益的，其它的都是正收益的，说明选出的股票整体本来就不错，可以作为投资的借鉴。

由于可获得数据的限制，本文只研究了利用一年的多因子量化选股模型进行量化选股，然而可以对以往各年的数据进行相同建模，基于同样的方法构建投资组合，得到股票池在实际投资收益，并于指数相比较。然而，投资者的目标或许不仅仅是超过市场基准，超过市场基准也并不总是给投资者带来正的回报，虽然指数本身可能会负增长，但正的实际回报也可以说明量化选股模型在实际运用中的有效性。如果投资者想要获得比市场更高的超额收益，需要不断学习，做到与时俱进，才会在复杂的资本市场中才会如鱼得水，获得源源不断地回报。

5. 结论与建议

量化投资在未来成为投资的趋势，必将被推崇，对其研究将会很强的现实意义。本文以沪深 300 指数各成分股为研究对象，根据各年不同的指数行情，以主成分分析法构建因子，将这些因子建立基于回归法的多因子量化选股模型，然后将量化模型筛选出的证券组合与下一年的实际投资绩效相比较，检验量化选股模型，得出以下主要结论：第一，基于回归法的多因子量化选股模型比指数有更高的收益，表明了模型的有效性。基于主成分分析的多因子模型可以充分发掘出衡量股票的风险，更加全面准确评价股票，所形成的投资组合能获得更高的收益。第二，影响股票收益率的变量在各年不同，在主成分分析中，虽然构建的主成分因子指标类型一样，但主成分因子中各个指标占比不同，说明了在不同的年份中影响股票收益的变量不断变化，要根据实际情况不断调整，这正是量化投资的相对优势。

根据以上的研究分析，为提高投资者在股票市场中的收益，提出以下投资建议：第一，由主成分分析法可以选出表现更好的股票，它们是具有正收益的，其构成的投资组合也将会带来正的收益。第二，构建投资组合必须关注个股风险分析，个股风险变化如财务造假、资产重组、新业务开拓等将会改变对该股的评价，可能需要将其从投资组合中剔除，重新构建投资组合。第三，量化投资可以有效消除人的

主观认识的缺陷，客观全面评价股票，量化投资主要是用历史数据来预测股票未来收益变化，因此量化投资是一个动态的过程，需要在一定时刻重新建立投资组合，这是一个不断调整的过程。

参考文献

[1] French, K.R. and Fama, E.F. (2015) A Five-Factor Asset Pricing Model. *Journal of Financial Economics*, **116**, 1-22. https://doi.org/10.1016/j.jfineco.2014.10.010

[2] Guerard, J.B., Markowitz, H. and Xu, G.L. (2015) Earnings Forecasting in a Global Stock Selection Model and Efficient Portfolio Construction and Management. *International Journal of Forecasting*, **31**, 550-560. https://doi.org/10.1016/j.ijforecast.2014.10.003

[3] Gu, S.H., Kelly, B. and Xiu, D.C. (2018) Empirical Asset Pricing via Machine Learning. Social Science Electronic Publishing, Shanghai. https://doi.org/10.3386/w25398

[4] 方浩文. 量化投资发展趋势及其对中国的启示[J]. 管理现代化, 2012(5): 3-5.

[5] 郭喜才. 量化投资的发展及其监管[J]. 江西社会科学, 2014, 34(3): 58-62.

[6] 彭志. 量化投资和高频交易: 风险、挑战及监管[J]. 南方金融, 2016(10): 84-89.

[7] 林德发, 杨潇宇. 跑赢沪深 300 指数的成分股组合构建——基于多因素模型的实证分析[J]. 中国商贸, 2014(2): 83-84.

[8] 王淑燕, 曹正凤, 陈铭芷. 随机森林在量化选股中的应用研究[J]. 运筹与管理, 2016, 25(3): 163-168+177.

[9] 李文星, 李俊琪. 基于多因子选股的半监督核聚类算法改进研究[J]. 统计与信息论坛, 2018, 33(3): 30-36.