E-Commerce Letters 电子商务评论, 2024, 13(3), 8619-8628
Published Online August 2024 in Hans. https://www.hanspub.org/journal/ecl
https://doi.org/10.12677/ecl.2024.1331054

Hans汉斯

# 基于支持向量机模型的多因子量化选股策略

咪　纳，赵予宁，何秉坤

贵州大学经济学院，贵州　贵阳

收稿日期：2024年5月20日；录用日期：2024年6月6日；发布日期：2024年8月22日

## 摘　要

**随着金融计算机领域的迅猛发展，量化投资正在扮演越来越重要的角色。多因子选股模型作为量化投资领域的重要组成部分，是量化投资策略选取优质股票组合的有效工具。本文以沪深300成分股为研究对象，综合考虑了成长类、技术类、价值类、情绪类以及动量类对股价有影响的因子，运用因子暴露分析、相关系数以及因子IC确定能够显著影响股票收益率的因子，再建立具有最优参数的核函数支持向量机模型，选取2010年1月1日到2023年1月1日作为时间区间进行训练与回测。得出结论：沪深300成分股的有效因子为市值、股本、换手率、ROE和PE；由高斯核函数支持向量机模型选股策略构建的投资组合策略能够保持投资组合的多样性，具有较高的风险回报能力，且收益较为稳定，能够跑赢大盘，证明了模型的有效性。**

## 关键词

**多因子，支持向量机，量化选股**

# Multi Factor Quantitative Stock Selection Strategy Based on Support Vector Machine Model

**Na Mi, Yuning Zhao, Bingkun He**

School of Economics, Guizhou University, Guiyang Guizhou

Received: May 20th, 2024; accepted: Jun. 6th, 2024; published: Aug. 22nd, 2024

## Abstract

**With the rapid development of the financial computer field, quantitative investment is playing an increasingly important role. The multi factor stock selection model, as an important component of**

文章引用: 咪纳, 赵予宁, 何秉坤. 基于支持向量机模型的多因子量化选股策略[J]. 电子商务评论, 2024, 13(3): 8619-8628. DOI: 10.12677/ecl.2024.1331054

**quantitative investment, is an effective tool for selecting high-quality stock portfolios in quantitative investment strategies. This article takes the 300 constituent stocks of Shanghai and Shenzhen as the research object, comprehensively considers the factors that affect stock prices, including growth, technology, value, emotion, and momentum. Factor exposure analysis, correlation coefficients, and factor IC are used to determine the factors that can significantly affect stock returns. Then, a kernel function support vector machine model with the optimal parameters is established, selecting January 1, 2010 to January 1, 2023 as the time interval for training and backtesting. Conclusion: The effective factors of the 300 constituent stocks in Shanghai and Shenzhen are market value, share capital, turnover rate, ROE, and PE; The investment portfolio strategy constructed by Gaussian kernel support vector machine model stock selection strategy can maintain the diversity of the investment portfolio, have high risk return ability, and stable returns, which can outperform the market, proving the effectiveness of the model.**

## Keywords

**Multi Factor, Support Vector Machine, Quantitative Stock Selection**



## 1. 引言

### 1.1. 研究背景

互联网、金融科技、大数据计算、云空间以及人工智能的迅猛发展推动了量化金融进入飞速发展的时代，随着投资方式、金融市场和 IT 技术的不断演进，量化交易技术和策略持续迎来更新。在这个发展过程中，多因子选股模型通过综合考虑各种因子与股票收益之间的关系，成功建立适合的模型来挑选出优质的股票组合。由于其表现较为稳定，多因子选股模型成为当前量化投资领域的研究热点，同时也是量化金融领域中不可或缺的重要组成部分。

机器学习对量化投资领域的发展具有推动作用。支持向量机(SVM)是一种基于结构风险最小化和统计学习 VC 维理论的建模方法，以其出色的泛化能力和扎实的理论基础著称。SVM 在处理高维度、复杂且非线性的基期学习问题方面表现出色。投资领域对 SVM 的应用前景广阔。利用支持向量机机器学习方法，寻找 A 股市场中的有效因子，构建多因子选股模型成为研究重点。这种方法探索从众多影响股票价格的因子中筛选出优质股票的可能性因子，并进一步构建有效的投资策略。因此，将支持向量机用于多因子选股可谓是一个备受关注的研究方向。

### 1.2. 研究意义

本文旨在运用多因子选股模型选出 A 股市场中的有效因子，再通过支持向量机模型确定投资策略，其理论与现实意义在于：

1. 多因子选股模型的本质是通过综合分析众多影响股票价格的因素及其相关关系，评估投资策略的收益价值。支持向量机方法在非线性、高维度大数据处理和分类问题中表现出独特的优势。将该模型运用在多因子选股策略中，可以丰富其使用方式，同时为其他方法的改进与提升提供了很好的思路和借鉴。这一应用具有相当的理论价值。

2. 本文的研究可以验证多因子选股模型在 A 股市场中的有效性，同时也能够找出影响股市价格变化的因子组合，对市场监管者和参与者具有实质性的价值。通过提供因子分析，有助于他们更好地跟踪市场价格的动态变化，从而提高决策与投资的准确性和及时性。这对于有效监管市场、降低风险以及优化投资组合都具有积极的推动作用。

## 2. 文献综述

Fama (1992)将 CAPM 和 APT 相结合，在实证研究中发现，股票的收益受市场风险、股票估值和股票市值三个方面的影响。据此，二人提出了著名的 Fama-French 三因子模型，单因子模型被推广为三因子模型[1]。Carhart (1997)三因子模型的基础上引入了反映动量效应的动量因子，提出了四因子模型[2]。Fama 和 French (2015)提出五因子模型，新加入盈利能力因子和投资风格因子[3]。范龙振(2003)利用 Fama-Macbeth 两步回归方法，建立了一个新的多因素投资组合优化模型，研究结果表明账面市值比和市盈率、总市值，价格等指标显着解释股票平均回报率，在添加市盈率因子之后，传统的三因子模型能够更加合理的说明上述因子的影响效应[4]。高春亭(2016)运用两步回归法对账面市值比，利润、并对投资因素与个股收益率的关系做了回归分析检验，结果表明上述因素对股票收益率解释的显著程度依次递减，但五因子模型总体好于三因子模型[5]。Joseph D. Piotroski (2000)选取了 9 个关键基本面指标进行打分排序，选取综合得分最高的股票选择构建策略组合[6]。Quah，Srinivasan 则采用人工神经网络(ANN)对股票进行选择并构建投资组合，输入变量包括财务变量和技术变量，而输出变量则是股票表现[7]。高岩(2004)利用了层次分析法，综合考虑主观因素、客观因素和宏观因素对证券进行排序选择[8]。苏靖宇(2018)基于行情因子和财务因子，将多因子选股模型和模糊 C 均值聚类算法结合对沪深 300 成份股构建投资组合[9]。吕凯晨(2019)通过基于基本面的多因子打分模型构建出一个战胜市场指数的量化选股模型[10]。

## 3. 支持向量机有效因子的筛选及实证检验

保证支持向量机模型具有较好准确性的前提是挑选出对股价具有影响的候选因子，有效性分析对有效因子指标的选取非常重要。通过股票池股票因子指标有效性分析，筛选适合选股分析有效因子，为下文支持向量机进行选股分析打下基础。

### 3.1. 数据来源

本文选取了沪深 300 成分股。其中 2010 年 1 月 1 日~2017 年 12 月 31 日共 84 个月 25,200 组个股数据为训练集，用来选择与策略有关的要素；2018 年 1 月 1 日~2023 年 1 月 1 日共 60 个月 18,000 组个股数据为样本外测试集，用于之后回测及收益率的分析。文章选取了沪深上市公司收益率数据、财务数据和行情数据为样本数据，其中选择股票涨跌幅数据为股票收益率代表变量。

### 3.2. 数据预处理

常用的影响股票的候选因子包括营运能力、资本结构、成长能力、盈利能力、估值、现金流量、技术性等几大类指标数据。为了能够更好的体现出我国市场当今的实际情况，这里选取了沪深 300 成分股的交易数据为样本，时间跨度为 2010 年 1 月 1 日至 2017 年 12 月 31 日共 1,663,200 个因子数据，所使用的数据来自于聚宽数据库，包括了公司财务相关数据、停复牌信息和收益率数据等。

在因子分析中需要进行缺失值和异常值的处理、标准化和中性化处理，以确保数据的准确性与可靠性。在处理数据的过程中，排除了处于停牌状态的股票以及 ST 股票(特别处理股票)等不合适的数据。同时剔除次新股和错误股票数据，确保所使用的数据准确可靠。

### 3.3. 支持向量机有效因子分析

1. 因子暴露分析

因子暴露度通常指股票的一些特征或者属性，通过 Barra 的风险模型将每个公司的特征做 Z-score 后直接得到，通过截面回归的方式估计出因子溢价。

在这里，我们每周测量因子相对于整个市场的偏离程度，同时考虑可比性和统一的规模。所使用的数据是因子的每日排名。计算步骤如下：按降序对每日因子进行排序；提取属于某一指数的成分股，计算各因子的排名平均值；风险敞口 = (指数因子排名平均值 − 当天的中端市场排名)/当天的股票总数。

由计算可以得出：市值与股本因子的偏差是每年最高的，沪深 300 指数的偏差稳定在 40%。然而，净利润增长率因子每年都接近 0，这表明该因子对沪深 300 指数的有效性较低。

2. 相关性分析

对所选取的因子之间的相关关系绘制热力图，如图 1 所示。相关系数越大的因子其颜色显示越深，也是我们重点关注选取的对象。

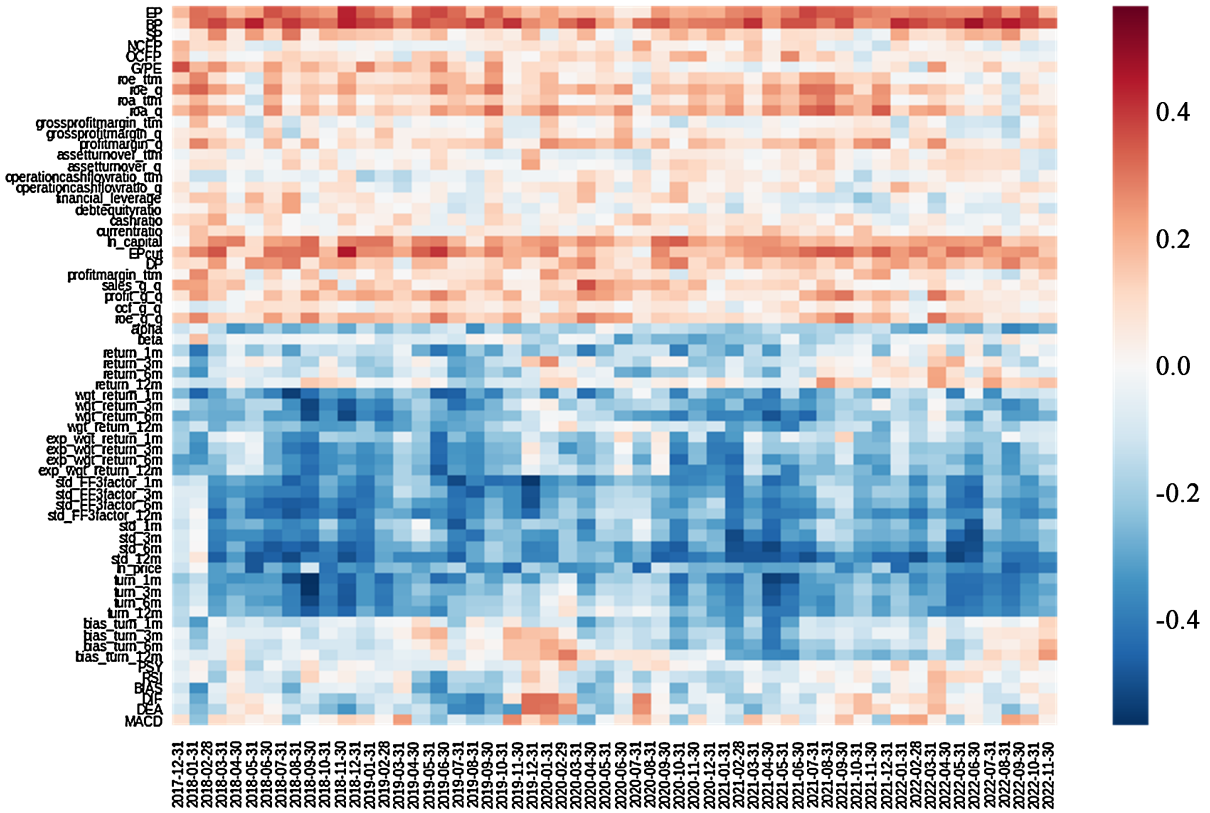


**Figure 1.** Heat map of factor correlation relationship
**图 1.** 因子相关关系热力图

(1) 相关性平均值

图 2 显示了沪深 300 指数中各种因子的相关性。结果显示，相关性最高的是股本和成交量，达到 0.88，其次是 EPS 和 ROE，达到 0.64。然而，换手率和市值之间的相关性降低了。以上是相关性的平均值。为了考虑相关性的稳定性，下表显示了沪深 300 指数中各因子相关性的标准差。标准偏差越小，相关性就越稳定。

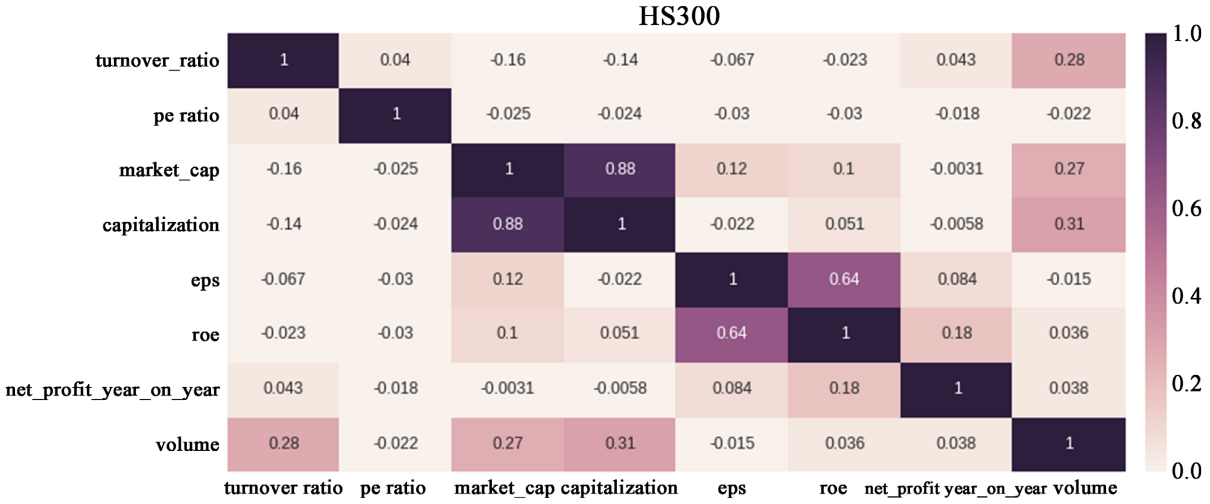


**Figure 2.** Average correlation value
**图 2.** 相关性平均值

(2) 相关性标准差

图 3 展示了各因子在沪深 300 指数中的相关性标准差，结果表明：波动性最高的是 PE 和净利润增长率，其次是换手率和成交量，但波动性降低。

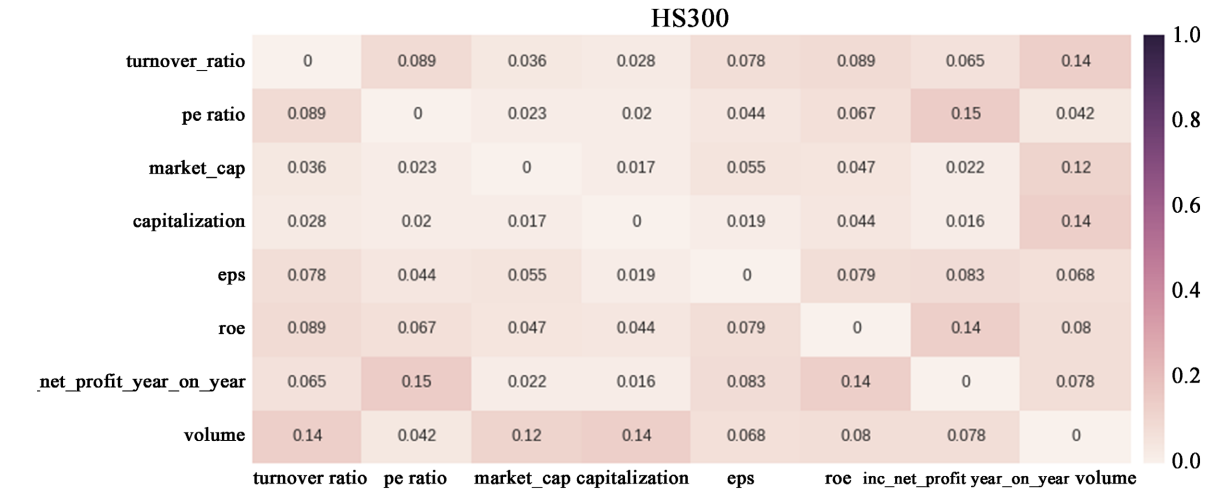


**Figure 3.** Correlation standard deviation
**图 3.** 相关性标准差

(3) 相关强度

从图 4 中可以得出以下结论：股本与市值是有明显稳定的相关性的，ROE 和 EPS 因子也有着较强的相关性，但是其他因子之间的相关强度的绝对值基本在 1~5 之间，相关强度最低的组合为 PE 和股本，换手率和市值，市值和换 PE，股本和换手率以及 EPS 和股本等。

3. 因子 IC 值分析

除了考虑因子的相关性，因子的选股能力是第三个评判标准。当 IC 绝对值大于 0.05 时，结论为该因子相对有效：相反，它是绝对无效的或者不稳定的。IC 值能够有效反映候选因子对于下期股票收益的预测能力，其正负代表它的作用方向。

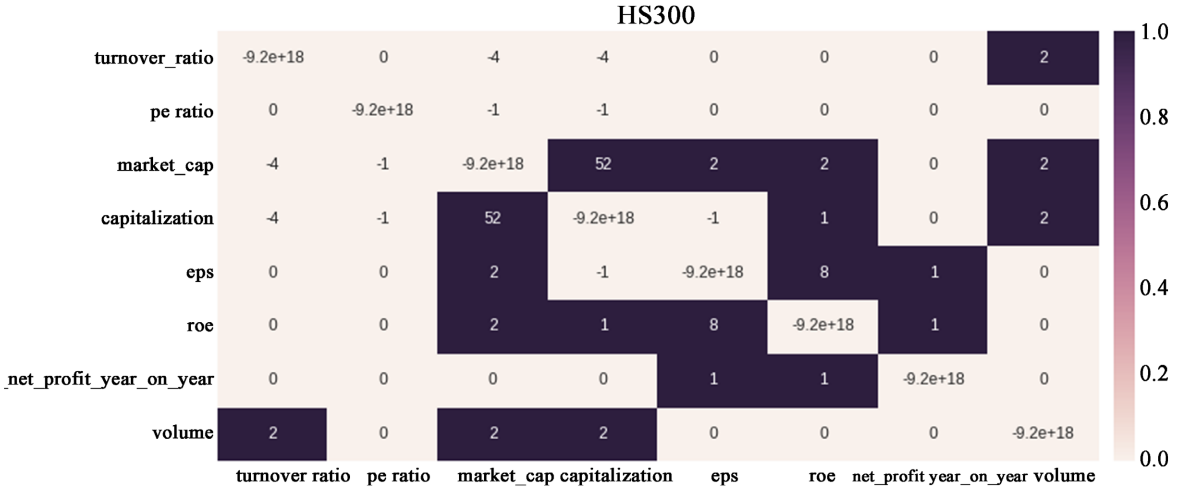


**Figure 4.** Correlation strength
**图 4.** 相关强度

(1) IC 均值分析

本研究通过因子的周 IC 值来选择解释力度较高的因子，结果如图 5：

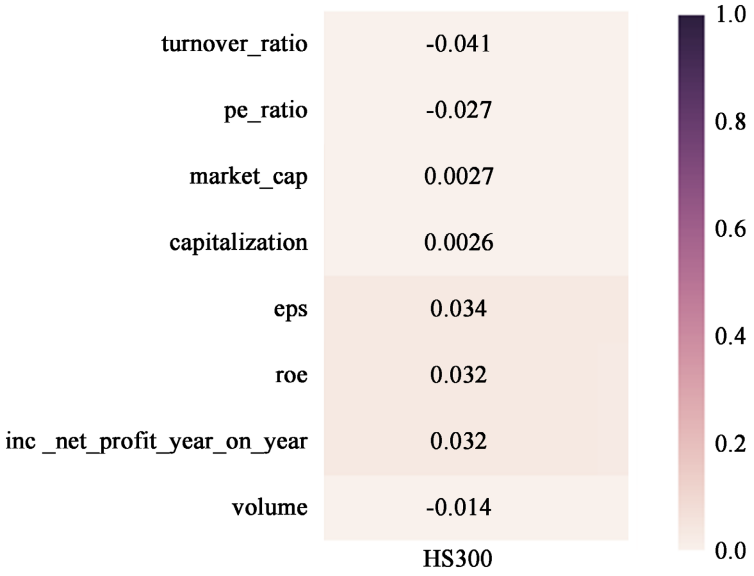


**Figure 5.** IC mean analysis
**图 5.** IC 均值分析

周平均 IC 值都在±1%以上，这表明因子的解释力度都很好。其中，换手率的选股能力最强，其次为 EPS 因子，而股本因子效果在所有因子中最差。为了进一步分析因子的选股能力，本文对 IC 的波动性进行分析，图 6 展示了年度的 IC 均值。

对历史 IC 序列的每个年度均值进行分析，可得：各个因子的波动性正常，换手率因子波动性最大，净利润增长率因子波动性最小。

(2) 绝对值均值分析

为了考察绝对选股能力，下表中展示了各因子 IC 的绝对值的平均值。

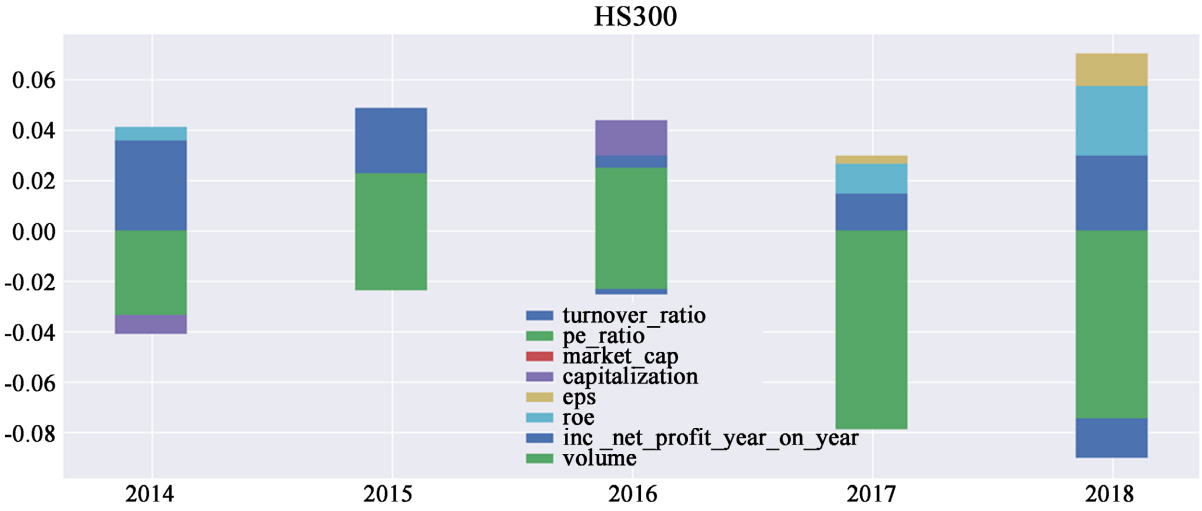


**Figure 6.** Annual IC Mean
**图 6.** 年度的 IC 均值

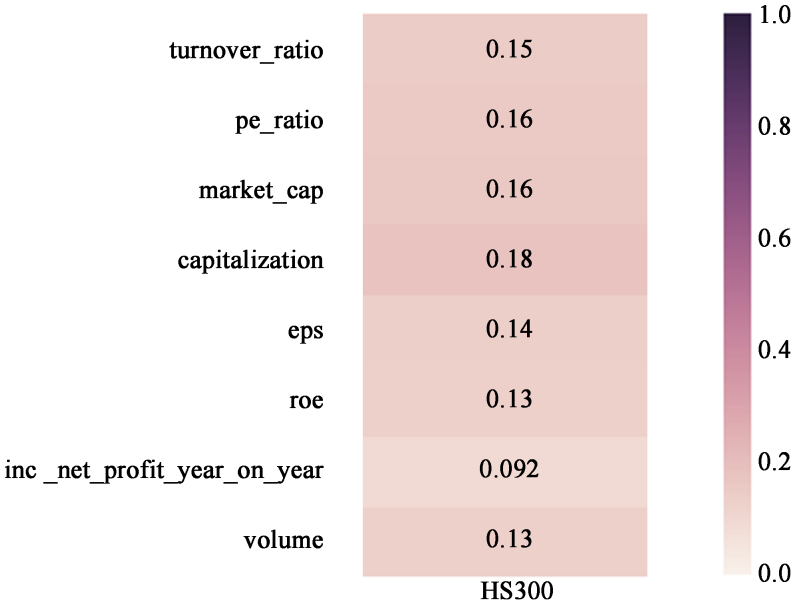


**Figure 7.** Absolute mean analysis
**图 7.** 绝对值均值分析

由图 7 可知，整体上可见 IC 的绝对值的平均值较高。股本因子绝对值平均值最大，可见股本因子的波动性最大，其次为市值因子和 PE，净利润增长率因子的绝对值平均值最小，该因子的波动性最小。由上可得到的结论与 IC 均值年度分析结果基本一致。因此，对于各因子 IC 的绝对值平均值来说，除了换手率因子较强、净利润增长率较弱外，其他因子的绝对选股能力差异不大。

综上，沪深 300 指数的市值和股本偏离程度最高，偏离程度中等的因子包括换手率、ROE、PE、EPS，偏离程度最低的因子是净利润增长率；相关性强度最低的组合是 PE 与股本、换手率与市值、市值与 PE、股本与换手率、EPS 与股本；除换手率因子较强、净利润增长率较弱外，其他因子的选股能力差异不大。基于上述结论，可以得出结论，在选择风险敞口高、相关强度低、选股能力强的因素时，市值、股本、换手率、ROE 和 PE 更适合作为因子组合。

## 4. 支持向量机模型的构建

本章在进行优质股的筛选过程中主要使用基于多因子的核函数支持向量机模型。我国股票市场总体结构是十分复杂，因此，不同因子对股票收益率的影响关系通常也具有非线性、复杂性等特点。本章基于上一章节有效性分析筛选出的有效因子，构建出一个具有优秀选股能力的支持向量机模型。

### 4.1. 基本流程

本研究以聚宽量化平台为基础，在现有研究的基础上，使用基于支持向量机的多因子选股模型。基本思路如下：先通过因子检验选出有效因子并将其作为输入特征集对模型进行分析。其次选取股票涨跌幅数据为代理变量计算股票收益率，根据收益率的正负将股票分成不同类别的标签。输入数据训练后确定支持向量机选股模型的最优超参数。再对样本外区间内的个股进行预测，选取上涨概率最大的个股构造投资组合，将所得的策略模拟回测，得出回测结果。

1. 数据获取。本文的研究对象为沪深 300 个股，样本区间为 2010 年 1 月 1 日至 2017 年 12 月 31 日，共 1,663,200 个因子数据，数据频率为月。

2. 特征和标签提取。以检验出来的有效因子的因子值作为样本的原始特征并进行排序。

3. 特征预处理。对有效因子在构建多因子选股模型之前需要对原始数据进行预处理。为了生成训练集，从样本内区间提取因子值，并对股票涨跌幅数据进行延迟一期处理。对于股票涨跌幅数据缺失的情况，删除这些股票的数据以获得有效的训练样本。

4. 样本内训练。选用高斯核函数建模。通过训练集上的样本训练后确定高斯核函数支持向量机模型的最优参数值。

5. 样本外测试：确定模型超参数后对模型进行训练，获得优化的支持向量机选股模型，再以全部股票有效因子值为模型输入变量，获取样本外区间股票收益率预测值以及对股票上升可能性的判断，选择上涨概率最大的某只个股构造投资组合策略。本文的测试区间为样本外区间：2018 年 1 月 1 日至 2023 年 1 月 1 日。

### 4.2. 模型参数设置

本文将验证集进行交叉验证实现参数设计，采用方法为 K 折交叉验证，将数据分成 K 组，随机抽取 1 组作为验证集，剩余 K-1 组训练集；训练集建立模型后，将验证集放到模型中，得到预测标签。

首先，确定核函数，考虑到高斯核函数能够将任意维数据映射到无穷维空间，使用高斯核函数建模更有意义。紧接着，选取合适的超参数是高斯核函数支持向量机建模的关键环节，超参数包括惩罚系数 C 值和核函数表达式中的参数 gamma 值。同时对 C 和 gamma 值进行遍历，找到全局最优解，参数寻优最常用的方法是网格搜索法。即对所有可能的参数组合进行穷举搜索，依次检验每一对参数的影响效果，以找到最优参数组合。

本文以 10 折交叉验证误差为目标，参数惩罚系数 C 取值范围为{0.01, 0.03, 0.1, 0.3, 1, 3, 10}，核函数参数 gamma 值取值范围为{$e^{-4}$，$3e^{-4}$，$e^{-3}$，$3e^{-3}$，0.01, 0.03, 0.1, 0.3, 1}。最终获得的最优超参数为 C 取 10，gamma 取 0.01。然后，对数据进行训练，在样本内进行交叉验证，根据计算结果可知，高斯核 SVM 模型样本内训练集和交叉验证集合正确率分别为 86.0%和 54.1%，AUC 分别为 0.859 和 0.541。

### 4.3. 模型运行

1. 投资策略

本策略运行于聚宽量化平台，并实现回测策略的模拟交易与预测。回测的初始资金是 100 万元，交

易手续费率为万分之三，回测频率是每天一次，对应于回测起止区间中交易日的时间。战略的业绩基准是上证指数。在此基础上构建了高斯核函数支持向量机调仓策略组合。调仓策略根据月度调整，也就是每个月的第 1 个交易日出售目前所持全部股份，然后选取本月首个交易日内可能性最大的 10 支个股买入，在购买时各股票购买权重按照各股票上涨概率的加权平均计算，完成当月建仓作业。然后将基准策略组合和前述调仓策略所得持仓记录进行性能比较，以观察基准策略及该策略之投资表现，验证所建立的支持向量机量化策略模型之有效性及可行性。

2. 支持向量机模型构建

经过 K 折交叉验证法及网格搜索，获得最优超参数构造出最优高斯核函数支持向量机模型。接下来对 2018 年 1 月 1 日至 2023 年 1 月 1 日的月数据进行股票收益率预测，以每月涨幅最大的 10 只个股构造投资组合，每只个股权重是每只个股涨幅的加权平均，模型运行结果如图 8 所示：


**Figure 8.** Running results of kernel function model
**图 8.** 核函数模型运行结果

综上，由核函数支持向量机模型选股策略构建的投资组合策略能够保持投资组合的多样性，具有较高的风险回报能力。说明本文通过因子效能分析选择有效因子作为输入特征变量、以收益率为因变量，通过选择高斯核函数支持向量机模型的最优参数，能够得到具有良好泛化能力的支持向量机模型，所得到的多因子选股模型具有较好的选股性能。

## 5. 总结与展望

### 5.1. 研究总结

本文考虑影响股票收益率的各个维度的因子，通过因子相关性分析筛选出有效因子，构建基于多因子的高斯核函数支持向量机选股策略，并验证模型策略在市场的表现。通过实证回测发现：

1. 本文对市场中所有股票候选因子进行有效性分析，以此筛选出有效因子为：市值、股本、换手率、ROE 和 PE。

2. 筛选出有效因子后，将其作为支持向量机选股模型的输入变量集，并将之前得到的股票涨跌幅数

据进行分类，其标签值作为因变量，利用网格搜索对所有数据进行遍历，通过 K 折交叉验证法对比各参数值的正确率，选取正确率最高的参数值作为最优参数值，建立一个最优高斯核函数支持向量机的训练模型。在聚宽量化平台中对样本外测试期的 A 股市场进行回测分析。最终模型的回测结果说明，基准业绩组合和高斯核函数支持向量机模型所选股票构建的投资组合策略多样性较高，风险偏小，并有着较高的风险回报能力，能够稳定获得超额收益，成功跑赢大盘，取得了不错的实盘效果，具有良好的推广能力。

### 5.2. 研究展望

随着我国投资市场的不断发展，可以预见到支持向量机等机器学习模型将在未来的投资决策中占据举足轻重的地位。在这个基础上，仍然可以进行更多的工作提高所选取因子的有效性：

1. 在静态训练期进行训练及测试中本文均采用固定长度进行选取，可以尝试进行超参数的动态寻优以及滚动预测进一步增加模型的真实性。

2. 有效因子的筛选是整个支持向量机建模的基础，但在因子选取的过程中，只是在常用的候选因子内进行选择，因此需要考虑更多常规因子外的财务因子、技术因子和宏观因子对股价造成的影响，进一步优化选股因子。

## 参考文献

[1] Fama, E.F. and French, K.R. (1996) Multifactor Explanations of Asset Pricing Anomalies. *The Journal of Finance*, **51**, 55-84. https://doi.org/10.1111/j.1540-6261.1996.tb05202.x

[2] Carhart, M.M. (1997) On Persistence in Mutual Fund Performance. *The Journal of Finance*, **52**, 57-82. https://doi.org/10.1111/j.1540-6261.1997.tb03808.x

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

[4] 范龙振, 王海涛. 上海股票市场股票收益率因素研究[J]. 管理科学学报, 2003(1): 60-67.

[5] 高春亭, 周孝华. 公司盈利、投资与资产定价: 基于中国股市的实证[J]. 管理工程学报, 2016, 30(4): 25-33.

[6] Piotroski, J.D. (2000) Value Investing: The Use of Historical Financial Statement Information to Separate Winners from Losers. *Journal of Accounting Research*, **38**, 1. https://doi.org/10.2307/2672906

[7] Quah, T. and Srinivasan, B. (1999) Improving Returns on Stock Investment through Neural Network Selection. *Expert Systems with Applications*, **17**, 295-301. https://doi.org/10.1016/s0957-4174(99)00041-x

[8] 高岩, 杨国孝. 基于层次分析法的选股决策[J]. 数学的实践与认识, 2004(10): 62-68.

[9] 苏靖宇, 方宏彬. 基于沪深 300 成份股的多因子量化选股策略研究[J]. 福建商学院学报, 2018(1): 21-28.

[10] 吕凯晨, 闫宏飞, 陈翀. 基于沪深 300 成分股的量化投资策略研究[J]. 广西师范大学学报(自然科学版), 2019, 37(1): 1-12.