大数据、机器学习与资产定价

●潘水洋

摘要:文章采用机器学习领域中的记忆神经网络、支持向量机、随机森林捕获多个定价因子之间的非线性定价结构。基于中国 A 股市场数据将机器学习模型与多因子线性定价模型进行了全面比较。实证结果表明,机器学习非线性定价模型在样本外预测精度、分组多空策略业绩表现均优于传统线性定价模型。基于机器学习理论的非线性资产定价模型能够更加准确对中国股票投资组合未来收益率进行预测,可以指导投资者做出更合理的投资决策从而提升中国资本市场的定价效率。

关键词:大数据;机器学习;资产定价

一、引言

从计量经济学角度对传统线性资产定价模型进行实证时经常发现该模型的解释能力不足,或者回归截距项显著大于零,或者贝塔系数不显著等一系列问题。造成这些问题的根源可能是解释变量遗漏,可能是期望收益的时变性,也可能是贝塔系数的时变性,还可能是随机项方差的时变性(如条件异方差)等。因此对线性资产定价模型的修正便出现了三个不同的努力方向:(1) 将失败归因于解释变量的遗漏, 并据此提出了多因子资产定价模型(Fama & French,1992;Jegadeesh & Titman,1993;Fama & French;2015);(2)将失败归因于贝塔系数的时变性,并据此提出了变参数(贝塔)资产定价模型;(3)将失败归因于随机项方差的时变性,并据此提出了 GARCH 类资产定价模型等。然而上述三个方向给出的模型并没有从根本上解决资产定价模型解释能力不足的弱点,本课题认为单纯依靠增加解释变量提高模型解释能力是一个不能穷尽的过程,与此同时国内外许多文献发现金融市场存在长记忆、尖峰厚尾、分形等非线性特征(徐绪松,2001;李红权,2008),然而对线性资产定价模型进行修正的文献并没有认识到这些非线性现象,这可能是许多线性定价模型解释力不足的重要原因。为解决以上问题,本文创新性的引入机器学习领域中的经网络模型、支持向量机、随机森林模型捕获定价因子之间的非线性关系。新模型与线性定价模型在样本外拟合优度、多空策略业绩表现做了严格对比。本文将从以下几个方面构建模型:(1)收集中国 A 股数据,根据国内外文献构建出一系列具有预测功能的因子,包括以财务数据为基础的基本面因子、以交易数据为基础的技术指标因子、以互联网非结构化数据为基础的情绪因子,形成因子大数据集合;(2)分别采用机器学习理论中的神经网络、支持向量机、随机森林从大数据集中辨识出隐藏在数据背后的因子非线性定价结构;(3) 与线性资产定价模型进行比较,检验机器学习理论非线性资产定价模型解释力和样本外预测能力。本文的贡献主要体现在以下方面:(1)脱离了依靠增加解释变量个数提高线性资产定价模型解释力的传统实证资产定价的研究框架;(2) 将机器学习领域中的非线性模型运用在金融资产定价领域,实现定价因子之间的非线性关系捕获,丰富了资产定价领域的研究方法;(3)采用机器学习理论的非线性资产定价模型能够对中国股票投资组合未来收益率进行更加准确预测,对投资者构建有效投资策略具有很强的借鉴意义。

二、文献回顾

随着计算机技术的不断革新,证券资产交易的相关数据都可以被采集,并且正在以前所未有的速度增长。在现阶段人们不仅可以获取反映公司运营状况的财务数据,也可以获取上市公司逐笔高频交易数据。借助互联网技术,通过自然语言处理算法人们还可以自动获取各个上市公司的舆情信息、搜索热度等反映投资者对公司未来预期的情绪因子数据。以财务数据为基础的基本面因子数据、以交易数据为基础的技术指标因子数据、以互联网非结构化数据为基础的情绪因子数据共同成为股票收益率驱动因子大数据来源。与此形成鲜明对比的是,大数据背景下,学术界并未跳出资产定价模型传统研究框架:从 Sharpe (1964),Limner (1965) 提出的 CAPM 模型, 到 Fama 和 French (1992) 提出的三因子定价模型, 到 Jegadeesh 和 Titman(1993)加入动量效应的四因子定价模型再到 Fama 和 French(2015)增加了盈利因素 RMW 和投资因素 CMA 提出的五因子模型都没有脱离小样本数据背景下线性资产定价研究框架,这一研究框架导致了资产定价模型存在解释力不足。与此同时,以机器学习理论为基础的大数据分析技术正成为大数据分析与处理的有效工具 (余凯,2013)。机器学习理论的不断深入发展,导致了“大数据+深度模型”时代的来临。随着金融大数据不断累积,机器学习将在金融大数据分析中发挥关键作用。

国内外许多文献发现金融市场存在非线性特征,大量文献对此进行了实证检验和报道(Ding,1993;Panas,2001;徐绪松,2001;Davidson,2002;李红权,2008)。股票市场具有非线性特征意味着随机游走与有效市场假设将失效,以此为理论假设基础的现代资本市场理论以及其他依赖正态分布或有限方差性质的金融计量学模型都将面临严重的质疑。在国内,徐绪松(2001)对中国 A 股的实证发现中

国股票对数收益率偏离正态分布且不是独立同分布序列，存在非线性效应。李红权(2005)通过修正的 R/S 分析与 ARFIMA 模型对中国股市收益率及其波动性的长期相关性进行了实证研究。结果表明：中国股市具有显著的非线性特征，虽然收益率序列的自相关性较弱，但波动性序列却表现出显著的长期记忆效应。李红权(2008)采用非线性动力学分析方法，分析中国股市波动的本质特征与形成机制。结果表明中国股市波动具有显著的分形动力学非线性特征。郝清民(2007)采用 R/S 法和 ARFIMA 模型也发现我国股市存在长记忆非线性效应，并且深市比沪市的长记忆更强。苑莹、庄新田(2008)和余俊(2008)等也持有相同观点。在资产定价模型的修正过程中，股市具有非线性现象这个广泛存在的金融异像被选择性忽视，本文认为其主要原因在于刻画因子之间的非线性定价结构无法给出一个具体的解析式。为了解决这一问题，本文并不直接寻找非线性定价结构解析式，而是采用机器学习领域中的预测模型去自动捕获因子之间的非线性定价结构。只要因子数据集与股票收益率之间存在非线性定价结构，机器学习就能够从历史数据中辨识出隐藏在数据背后的定价结构。机器学习以数据驱动为核心来构建模型，并依靠交叉验证选择最优模型，是辨识因子之间的非线性定价结构的有力工具。

当前机器学习在经济学文献中应用还较为少见，但最近已经获得了高度的关注(Varian，2014；Mullainathan and Spiess，2017)。机器学习包含两大类别，分别为非监督式学习(Unsupervised Learning)和监督式学习(Supervised Learning)。非监督式学习主要关注自动寻找规律，通常应用在图像视频识别、文本数据挖掘等领域。监督式学习主要关注的是预测。给定一个训练数据集 x 和结果变量集 y，监督式机器学习通过最小化均方误差能够自动找到 x 与 y 之间的映射关系，同时 x 与 y 之间的映射关系能够在新数据集中对 y 进行预测。监督式学习这一类机器学习算法在计量经济学领域最有应用价值，许多学者对此进行了探索。Doudchenko 和 Imbens(2016)阐述了机器学习如何与双重差分 DID 结合进行因果识别。Varian(2016)对机器学习与断点回归结合的未来方向进行了深入探讨。Athey(2015，2017)介绍了监督机器学习在公共资源分配、因果推断领域的应用。Chalfin 和 Aaron(2016)的文章指出使用机器学习预测员工生产率可以获得巨大的社会福利收益。Hartford(2017)采用神经网络等非线性方法对工具变量选择进行第一阶段的估计。

在资产定价领域，Rapach(2013)采用机器学习领域的 LASSO 回归预测全球股票回报率。Hutchinson(1994)采用神经网络对衍生品价格进行预测。Khandani(2010)、Butaru(2016)采用回归树模型预测信用卡违约概率。Harvey 和 Liu(2016)采用机器学习领域中的自助法对多个资产定价因子的有效性进行了检验。Kelly(2015)采用机器学习中的因子降维方法和构建多个预测模型组合预测股票回报率，取得了较好的预测结果。从上述文献中可以看出，与常规线性模型相比，机器学习理论中的神经网络、支持向量机、随机森林模型有一些独特的优势。首先，机器学习模型能够自动识别金融数据背后的隐含特征，不需要人工干预。其次，传统的线性资产定价模型有效的前提条件是金融经济系统是线性的。然而国内外许多文献发现金融市场存在长记忆、尖峰厚尾、分形等非线性特征，这些特征预示着金融时间序列存在非线性动力学系统。传统线性模型对此无能为力，而机器学习模型却能识别出金融数据中的非线性规律。

三、定价因子大数据集构建

随着计算机技术的不断革新，金融数据越来越容易获取。从预测研究的国际趋势来看，使用大数据集进行预测越来越广泛，大数据技术逐渐成为很多国家进行金融经济预测的新方法和新工具。从已有的资产定价模型文献研究结果来看，影响资产定价的因子有很多，本文将影响资产定价的因子分为三大类，分别为基本面因子、技术指标因子、情绪因子。

1. 基本面因子由公司已发布的财务报表数据构建。根据已有的文献研究结果，本文从公司盈利能力、估值、现金流、成长性、负债等多维角度全方位构建基本面因子。本文采用的基本面因子如下：

EP 因子：净利润/总市值；
BP 因子：净资产/总市值；
SP 因子：营业收入/总市值；
CFP 因子：经营性现金流/总市值；
GPE 因子：净利润同比增速与市盈率的比值；
Sale_G_Q 因子：营业收入季度同比增速；
Profit_G_Q 因子：净利润季度同比增速；
OCF_G_Q 因子：经营性现金流季度同比增速；
ROE_G_Q 因子：ROE 季度同比增速；
ROE 因子：净资产收益率；
ROA 因子：资产收益率；
Grossprofitmargin 因子：毛利率；
Profitmargin 因子：净利率；
Assetturnover 因子：资产周转率；
Operationcashflowratio 因子：经营性现金流/净利润；
Debtequityratio 因子：长期债务/净资产；
Currentratio 因子：流动比率；
Cashratio 因子：现金比率。

2. 技术指标因子。大量从事技术分析的交易者战胜市场获取超额收益的事实表明市场并不是完全有效的，技术指标因子也能够从一定程度对资产未来收益率进行预测。本文基于中国 A 股日频交易数据分别构建趋势类、反转类、波动率、流动性、动量等技术指标因子。本文采用的技术因子如下：

Holder_avgpercent_GN 因子：户均持股比例相对于前 N 季度的变化率，N=1，2，3，4；

Holder_num_GN 因子：股东户数相对于前 N 季度的变化率，N=1，2，3，4；

Ln_size 因子:流通市值的对数;

Return_N 因子:过去 N 个月的涨跌幅,N=1,3,6,12;

Beta 因子:过去 1 年与上证综指回归估计的 Beta;

Std_N 因子:过去 N 个月的个股波动率,N=1,3,6,12;

Std_Res_N 因子:过去 N 个月与上证综指回归的残差波动率,N=1,3,6,12;

Turn_N 因子:过去 N 个月的个股日均换手率,N=1,3,6,12;

Bias_turn_N 因子:过去 N 个月与过去 24 个月的个股日均换手率比值,N=1,3,6,12。

3. 情绪因子。随着行为经济学的发展,有研究者发现大众情感能在一定程度上预测股票未来收益率。本文通过上市公司官方网站每天访问量构建出情绪因子。

以上由基本面因子、技术指标因子、情绪因子形成 50 个定价因子大数据集,在此基础上,本文采用机器学习理论对隐藏在因子大数据集背后的非线性结构进行辨识,找出因子非线性定价结构,并给出预测结果。

四、实证分析

N 个资产的线性资产定价模型可以用 K 个因子线性表示,线性资产定价模型的一个重要缺陷就是解释力不足。导致这一结果的可能原因是模型忽略了因子之间存在的非线性结构。对于如何辨识出资产定价模型中的因子非线性结构,国内外文献并未做相应的研究。本文尝试采用机器学习理论中的神经网络、支持向量机、随机森林对隐藏在因子大数据背后的非线性定价结构进行辨识。需要指出的是,机器学习模型并不对非线性定价结构做任何前提假设,而是寄希望通过对因子大数据进行分析,自动辨识出因子非线性定价结构,在这个基础上再以中国 A 股数据对此非线性定价结构进行实证分析。

基于机器学习理论的非线性资产定价模型能够辨识出股票收益率与定价因子大数据集之间的非线性映射。本文分别选取支持向量机、神经网络、随机森林实现非线性映射。为了检验机器学习非线性资产定价模型的有效性,我们采用样本外模型预测值的拟合优度以及构建多空策略业绩表现来检验模型的有效性,基准模型为线性多因子定价模型。

本文选取 1997 年 1 月至 2017 年 12 月沪深两市 A 股股票的月度收益率、收盘价、合并财务报表中财务数据,再剔除 *ST 股票、已经退市的股票、上市不足 1 年的股票构建因子大数据集,所有数据来源于 wind 资讯数据库。将 1997 年 1 月至 2007 年 12 月的数据作为模型训练数据,用于估计线性定价模型参数和机器学习非线性资产定价模型参数。2008 年 1 月至 2017 年 12 月的数据作为样本外验证数据,用于检验模型样本外的预测能力。估计模型参数之后,从 2008 年 1 月开始,在每个月月初,将股票最新的因子数据分别输入多因子线性定价预测模型、神经网络模型、支持向量机模型、随机森林模型分别得到每个模型对股票下个月收益率的预测值,再计算模型样本外的拟合优度。

各个预测模型的股票收益率样本外拟合优度如下:线性模型拟合优度 0.31;神经网络模型拟合优度 0.46;支持向量机模型拟合优度 0.37;随机森林模型拟合优度 0.41。从以上结果来看,采用机器学习模型股票收益率拟合优度都高于线性因子定价模型样本外拟合优度。

为了进一步检验机器学习非线性定价模型的有效性,本文基于线性模型、机器学习模型构造多空策略组合:多因子线性组合预测策略和机器学习因子非线性组合预测策略。

多因子线性组合预测策略构造过程如下:每个月初,线性定价模型基于最新因子截面数据预测股票下一期收益率,然后将股票收益率预测结果按降序排列等分为 10 组,做多第 1 组股票,做空第 10 组股票。

机器学习因子非线性组合预测策略构造过程如下:每个月初,各个机器学习模型基于最新因子截面数据预测股票下一期收益率,然后将股票收益率预测结果按降序排列等分为 10 组,做多第 1 组股票,做空第 10 组股票。各个模型策略表现如下:

线性模型多空策略。多空组合年化收益 17.12%;多空组合收益年化波动率 13.31;夏普比率 1.93;最大回撤 6.93%。

神经网络模型多空策略。多空组合年化收益 25.12%;多空组合收益年化波动率 10.28;夏普比率 2.97;最大回撤 5.61%。

支持向量机模型多空策略。多空组合年化收益 21.87%;多空组合收益年化波动率 12.47;夏普比率 2.53;最大回撤 7.39%。

随机森林模型多空策略。多空组合年化收益 22.52%;多空组合收益年化波动率 11.68;夏普比率 2.64;最大回撤 7.57%。

从以上结果可以看到,机器学习模型多空策略年化收益、年化波动率、夏普比率、最大回撤均要好于线性模型多空策略,进一步表明机器学习模型捕获到了因子之间的非线性关系,策略表现要优于线性模型。

参考文献:

[1] Athey S. Beyond Prediction:Using Big Data for Policy Problems [J]. Science,2017,355(6324):483-485.

[2] Athey S. Machine Learning and Causal Inference for Policy Evaluation [C]. Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,2015.

[3] Sharpe W F. Capital Asset Prices:A Theory of Market Equilibrium under Conditions of Risk [J]. Journal of Finance,1964,19(3):425-442.

[4] Limner J. The Valuation of Risk Assets and Selection of Risky (下转第 33 页)

1243-1272.
[5] Hooper, P. and Mann, C. L. . Exchange Rate Pass-Through in the 1980s: the Case of U. S. Imports of Manufactures[R]. Brookings Papers on Economic Activity,1989,(89):297-337.
[6] Hooper, P. and Marquez, J. , Exchange Rates, Prices, and External Adjustment in the United States and Japan[R]. International Finance Discussion Paper, No. 456, Board of Governors of the Federal Reserve System, Washington D. C. , Oct. ,1993.
[7] Hung, W. et al., Pricing Exports: a Cross Country Study [J]. Journal of International Money and Finance,1993,(12):3-28.
[8] Knetter, M. M.. Price Discrimination by U. S. and German Exporters[J].The American Economic Review,1989,(79):198-210.
[9] Knetter, M. M.. Exchange Rate and Corporate Pricing Strategies[R].NBER Working Paper No. 4151, Aug. ,1992.
[10] Knetter, M. M.. International Comparisons of Pricing-to-Market Behavior [J]. The American Economic Review,1993,83(3):473-486.
[11] Ohno, K. . Export Pricing Behavior of Manufacturing:A U. S.-Japan Comparison[R]. International Monetary Fund Staff Papers,1989,36(3):550-579.
[12] Park, T. A. and Pick, D. H. . Imperfect Competition and Exchange Rate Pass-through in U. S.Wheat Exports[M].Industrial Organization and Trade in the Food Industries. Westview Press, Oxford,1996.
[13] Patterson, P. M. et al., Price Discrimination by U. S. High-Value Food Product Exporters, Industrial Organization and Trade in the Food Industries[M]. Westview Press, Oxford,1996.
[14] Winkelmann, L. and Winkelmann, R. , The Costs of Non-Tariff Barriers to Trade: Evidence from New Zealand [J]. Review of World Economics,1997,(133):270-281.
[15] Yang, J. . Exchange Rate Pass-Through in the U. S. Market:A Cross-Country and Cross-Product Investigation[J]. International Review of Economics and Finance,1995,4(4):353-371.

基金项目:北京语言大学院级科研项目(中央高校基本科研业务专项资金资助)(项目号:18YJ040009)。
作者简介:冯耀鹏(1978-),女,汉族,河南省登封市人,北京语言大学商学院讲师, 对外经济贸易大学经济学博士,研究方向为国际贸易理论与政策。
收稿日期:2018-12-11。

(上接第 8 页)

Investments in Stock Portfolios and Capital Budgets[J]. Review of Economics and Statistics,1965,47(1):13-37.
[5] Rapach E. ,K. Strauss,and Z. Guofu. International Stock Return Predictability: What is the Role of the United States[J]. Journal of Finance,2013,68(4):1633-1662.
[6] Hutchinson M, L. Andrew, and P. Tomaso. A Nonparametric Approach to Pricing and Hedging Derivative Securities via Learning Networks [J]. Journal of Finance,1994,49(3):851-889.
[7] Khandani,E. , K. Adlar, and L. Andrew. Consumer Credit-Risk Models via Machine Learning Algorithms[J]. Journal of Banking & Finance,2010,34(11):2767-2787.
[8] Butaru, Florentin, and C. Qingqing. Risk and Risk Management in the Credit Card Industry[J]. Journal of Banking & Finance, 2016,72(3):218-239.
[9] Harvey, R. Campbell, and Y. Liu. and the Cross-Section of Expected Returns[J]. Review of Financial Studies,2016,29(1):5-68.
[10] Kelly, Bryan, and S. Pruitt. The Three-Pass Regression Filter:A New Approach to Forecasting Using Many Predictors[J].Journal of Econometrics,2015,186(2):294-316.
[11] 徐绪松,陈彦斌.深沪股票市场非线性实证研究[J]. 数量经济技术经济研究,2001,23(3):110-113.
[12] 李红权,马超群.金融市场的复杂性与风险管理[M]. 北京:经济科学出版社,2006.
[13] 李红权,马超群.股市收益率与波动性长期记忆效应的实证研究[J].财经研究,2005,12(8):29-37.
[14] 李红权,汪寿阳,马超群. 股价波动的本质特征是什么?——基于非线性动力学分析视角的研究 [J]. 中国管理科学,2008,21(7):1-8.
[15] 郝清民. 中国股市收益率长记忆性 R/S 非线性分析 [J]. 管理工程学报,2007,13(2):115-117.
[16] 余凯,贾磊,陈雨强.深度学习的昨天,今天和明天[J].计算机研究与发展,2013,50(9):1799-1804.

基金项目:国家社科基金青年项目“大数据背景下基于深度学习理论的非线性资产定价模型研究”(项目号:18CJY057)。
作者简介:潘水洋(1986-),男,汉族,湖南省岳阳市人,北京大学经济学院博士后,研究方向为金融大数据分析。
收稿日期:2018-12-16。