第 22 卷第 3 期　　经 济 学（季 刊）　　Vol. 22, No. 3
2022 年 5 月　　China Economic Quarterly　　May, 2022

深度学习与中国股票市场因子投资

——基于生成式对抗网络方法

马　甜　姜富伟　唐国豪*

摘　要　本文运用深度学习模型研究中国股票市场的收益预测与因子投资。我们使用 148 个微观企业特征变量构建因子大数据集，并采用生成式对抗网络（GAN）方法构建深度学习模型。研究发现，相较于线性模型，深度学习模型在收益预测精度和因子投资绩效上均有很大提升。本文还分析了不同类型因子在中国股市的重要性，探索了金融深度学习预测的经济理论机制解释。本文对中国金融市场高质量发展和金融科技应用探索均有重要意义。

关键词　深度学习，资产定价，因子投资
DOI：10.13821/j.cnki.ceq.2022.03.05

一、引　　言

资本资产定价模型（CAPM）和 Fama-French 三因子模型等金融市场资产定价模型的关键在于探究金融市场预期投资收益的决定因素。随着大数据时代的到来，越来越多的股票市场异象和定价因子被发现，以大数据和机器学习为代表的人工智能技术正在逐渐融入并改变金融研究范式。其中，对高维因子大数据的有效信息提取和利用近年来成为资产定价理论与实证研究的焦点。比如，围绕因子大数据信息有效性和冗余性等问题，金融学术界使用机器学习方法，从高维因子大数据中识别出可以持续稳定获得超额收益的少数核心定价因子（Harvey *et al.*, 2016；Hou *et al.*, 2019），并开展股票收益预测和投资策略构建（Feng *et al.*, 2018a；Jiang *et al.*, 2018；姜富伟等，2019；Gu *et al.*, 2020）。

随着互联网云计算等技术的发展，深度学习神经网络模型已广泛应用在各行各业，包括以人脸识别和文本识别为代表的计算机视觉领域和以自然语

* 马甜，中央民族大学经济学院；姜富伟，中央财经大学金融学院；唐国豪，湖南大学金融与统计学院。通信作者及地址：姜富伟，北京市海淀区学院南路 39 号，100098；电话：18511086494；E-mail：fwjiang@qq.com。本文得到国家自然科学基金项目（72072193，71872195，72003062）资助。感谢《经济学》（季刊）编辑部、三位匿名审稿人以及李斌、黄卓、余乐安、杨晓光、邹恒甫、李建军、周国富等学者对本文的宝贵意见，当然文责自负。

言处理（natural language processing，NLP）为代表的语音识别领域。由于接近于人脑思维的结构特性和现实应用中的优异表现，深度学习正逐渐成为人工智能的代名词。在金融领域，深度学习也受到金融学术界和业界的广泛关注，但相关应用和研究处于起步摸索阶段。本文创新运用深度学习神经网络模型研究了中国股票市场的收益预测与因子投资问题。我们认为，深度学习一方面可以通过数据降维和特征提取解决高维陷阱问题，有效挖掘股票特征因子大数据中的显性（observable）和隐性（latent）信息；另一方面，不同于主成分分析等经典线性机器学习方法，深度学习考虑了数据间的非线性相依关系，可以有效提取股票因子大数据中的非线性信息，并带来资产定价研究的非线性范式革命（Chen *et al.*，2019）。

作为无监督深度学习的代表，生成式对抗网络（generative adversarial networks，GAN）模型最初由 Goodfellow（2014）提出，诞生至今经历了急速发展，包括 Google、OpenAI、Facebook、Apple 等互联网巨头均投入了大量研究资源深化和扩展 GAN 的应用（Salimans *et al.*，2016；Zhao *et al.*，2016），并成功应用在图片生成（Gadelha *et al.*，2017）和视频生成（Vondrick *et al.*，2016）等领域。受博弈论中二元零和博弈的启发，GAN 的框架中包含一对相互对抗的模型：生成器和判别器。不同于传统依靠反向传播进行参数更新的神经网络模型，GAN 由于引入了对抗网络而具备了“进化”的特性：生成器利用特征因子大数据集对股票市场收益进行预测，判别器通过正误判断及比较真实实现收益和预测收益，指导生成器完成参数更新。在迭代过程中，二者不断提高各自的生成能力和判别能力来寻找二者间的纳什均衡，从而最终的预测结果相比单一神经网络模型更接近真实收益分布。

国内金融领域尚无使用 GAN 模型进行研究的文献，本文首次将 GAN 应用于我国股票市场预测与投资，并强调深度学习预测背后的经济理论思考和探索。相比国外成熟市场，中国股票市场散户参与度较高，市场波动性更大。因此，为准确把握我国股票市场的动态变化规律，模型需要具有很敏锐的适应性和动态学习能力。而 GAN 模型特有的判别器系统带来更为动态的模型适应性，在面对新一期的样本数据时，不仅通过生成器中的记忆单元保留了时序数据的趋势项，还通过判别器进一步过滤噪声。

本文基于上述理论框架开展实证研究。我们首先构建了包括 74 个企业特征因子和 74 个行业特征因子共计 148 个指标的中国股票市场特征因子大数据集。我们接着围绕横截面和时间序列两个不同的维度开展预测建模，其中横截面主要强调个股之间的差异，而时间序列更关注市场的动态波动，从不同视角探索金融市场对于高维因子信息的反应。实证研究发现，相比长短期记忆模型，GAN 模型在不同样本期均具有更灵活、更精准的预测能力。本文在深度学习模型构建和因子大数据分析方面的特色创新如下：

第一，更有效的特征提取和对非线性信息的利用。传统线性回归模型忽

视了金融大数据内在的潜在信息因子、稀疏性和非线性等数据性质。本文在构建股票特征大数据的基础上，使用了 GAN 模型对中国股票市场进行非线性信息特征的提取和分析。实证结果相比线性模型有了显著的提升，表明了我国股票市场中上市公司数据的非线性特征包含有重要的预测信息。

第二，对于时序数据的有效处理。本文在构建 GAN 模型时，使用了更适合时序数据处理的长短期记忆网络（long short term memory，LSTM）模型，LSTM 模型作为循环神经网络（recurrent neural network，RNN）的改进，弥补了 RNN 模型在处理数据时“短时记忆”的问题，在对时序数据的处理中有着天然的优势。金融数据长期存在自相关特性，资产定价领域中动量效应更是作为经典的异象因子被广泛使用，LSTM 通过记忆单元保留有效信息，并通过遗忘单元过滤掉“噪声”信息，针对不同的资产类型匹配不同的记忆长度，而普通 RNN 模型无法保留长期的数据记忆。针对中国股票市场数据，本文在上述基础上进行模型框架优化并给出相关参数设置，对后续 LSTM 模型在金融市场的应用研究提供了参考。

第三，更“智能”的预测模型。不同于传统神经网络模型优化过程中单纯地使用梯度下降的方式，生成式对抗网络引入了“博弈”的过程，生成模块得到预测数据后，判别器将其与真实数据进行比对分类，评估并否定其预测结果，而最终的优化结果即要求生成器生成的数据骗过判别器达到以假乱真的程度。经济学中完全竞争市场具有最优效率，而通过引入判别器这一“竞争”者，生成式对抗模型在结构上优于单一的预测模型。

本文实证研究发现，相较于经典线性模型，GAN 深度学习模型在股票收益预测精度和因子投资绩效上均有非常显著的提高。在股票收益时间序列预测方面，GAN 深度学习样本外预测 R^2 最高达到 0.89%，显著好于线性模型，模型预测精度提升效果在 5%水平内显著。在股票横截面因子投资策略方面，使用预测值排序法构建投资组合，发现基于 GAN 深度学习模型构建的多空对冲因子投资组合的月平均收益率和夏普比率分别为 1.13%和 0.71，显著好于线性模型，且其 FF3 和 FF5 模型超额收益在 5%水平内显著。

本文还深入探究深度学习预测背后的经济理论机制。投资组合构成分析，发现科技类股票相比传统行业贡献更高收益。因子重要度分析，发现最重要的特征因子有三类：价格及交易量趋势类指标、流动性类指标、基本面类指标，其中前十大特征因子的贡献度占到了所有 148 个因子的 40%左右。错误定价理论分析，发现深度学习模型对于低金融摩擦、低波动性以及高流动性类股票更为有效。宏观经济状况分析，围绕宏观经济活跃度、经济政策与金融市场不确定性、投资者情绪等多个角度展开论证，发现深度学习模型可以有效捕捉我国宏观经济或金融市场中潜在的风险因素（Huang *et al.*，2015）。微观企业状况分析，发现深度学习模型能有效预测企业在中短期（未来一年内）的盈利、收入和现金流等基本面状况信息。

本文余下内容安排如下：第二部分为文献综述，第三部分为模型和数据描述，第四部分为实证结果研究，第五部分和第六部分为微观和宏观视角的经济学分析，最后一部分为本文结论。

二、文 献 回 顾

深度学习模型是指隐藏层大于两层的神经网络模型，其已广泛应用在各行各业，但在金融领域的应用还刚刚起步，仍处于探索阶段：Heaton *et al*.（2016）开发了一种深度学习模型来智能选择股票并构建投资组合；Gu *et al*.（2020）使用包括神经网络模型在内的多类机器学习算法对美国股市进行了收益预测，发现神经网络预测准确度更高，构建的投资组合夏普比率更高；Feng *et al*.（2018a）同样构建神经网络模型来预测股价；Feng *et al*.（2018b）引入了无套利约束结合深度学习来估计风险溢价。金融界对生成式对抗网络（GAN）模型的应用成果很少，只有 Chen *et al*.（2019）使用 GAN 构建了一种非线性资产定价模型，其中生成器使用了长短期记忆（LSTM）网络和前馈网络模型，判别器使用了 RNN 模型，使用的数据包括了 46 个美国市场股票特征以及 178 个宏观经济变量，发现其投资组合收益和夏普比率显著高于 Fama-French 三因子模型。

国内研究中，苏治等（2017）对金融实证里应用的深度学习模型进行了总结归纳，并对未来中国金融市场中深度学习的应用给出了评述。曾志平等（2017）对 A 股市场股票日度收益进行了图像分类，对上升、下降、波动趋势的三类股价使用深度网络模型进行训练，并按照预测信号进行量化交易，其准确度达到了 90.54%；陈卫华和徐国祥（2018）基于长短期记忆网络模型和股票论坛数据对股市波动率进行了预测，发现模型显著提升了预测精度。李斌等（2019）收集 1997—2018 年 A 股市场的 96 项异象因子数据，构建了 12 种机器学习算法驱动的股票收益预测模型及相应的基本面量化投资模型，系统性地对比了机器学习驱动的模型在中国市场取得的实证绩效；Jiang *et al*.（2018）、姜富伟等（2011，2019，2021）系统研究了中国股票市场收益可预测性，并针对中国股票市场构建了 70 多个企业特征因子大数据库，系统比较了多类机器学习模型的预测能力，发现神经网络模型等非线性模型表现最好。

三、模型及数据描述

（一）研究模型

1. 模型总体设计

本文总体研究框架如图 1 所示：

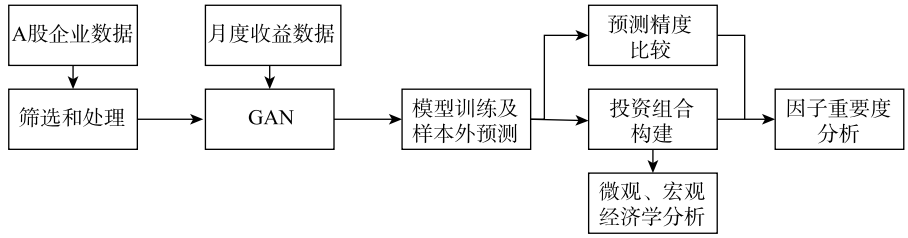

图 1　总体研究框架

本文的核心是使用 GAN 深度学习模型进行股票收益预测，构建多因子资产定价模型：

$$r_{i,t}=f_t(z_{i,t-1}:\theta)+\mu_{i,t}, \tag{1}$$

其中 $r_{i,t}$ 为 t 期股票 i 收益数据，$z_{i,t-1}$ 为 t−1 期股票 i 企业特征向量，函数 f_t 代表了 t 期考虑了变量间非线性关系的 GAN 深度学习模型。而在拟合计算 t 期函数的结构时，本文采用扩展窗口（expanding windows）法设定训练集和验证集。具体来说，样本初始训练集为 2003 年 1 月至 2006 年 12 月，初始验证集为 2007 年 1 月至 2008 年 12 月，利用得到的预测模型估计样本期为 2009 年 1 月至 2009 年 12 月的股票收益；之后每年年初保持验证集和测试集长度不变，训练集长度增加一年，最终得到的样本外预测集为 2009 年 1 月至 2017 年 12 月共 108 个月收益预测数据，如图 2 所示。

图 2　本文所使用的训练集、验证集与预测集结构

2. 生成式对抗网络模型

生成式对抗网络（GAN）作为无监督深度学习的一种，近年来广泛应用于人工智能的各个领域。典型的 GAN 包括两个模块，生成器（generative model）和判别器（discriminative model）。初始 GAN 模型受制于数据结构和模型复杂度，当判别器过于精准时会导致生成器参数更新时无法收敛；后续学者改进初始模型，提出了 Wasserstein GAN（WGAN）具有良好的收敛特性，模型优化过程也更加稳定。本文构建用于金融资产定价的 GAN 模型时，同样使用 WGAN，并设定生成器为 LSTM，以便利于具有时序特征的数据生成预测收益；判别器使用卷积神经网络（convolutional neural networks，CNN），用于区分预测值和真实值的差异。

深度学习算法结构如图 3。具体而言，在生成器中引入企业特征因子

$z_{i,t-1}$ 来获得初步估计的月度收益 $\hat{r}_{i,t,G}$；第二步再将真实收益 $r_{i,t}$ 与估计收益 $\hat{r}_{i,t,G}$ 导入判别器中，其中 $r_{i,t}$ 归为“1”类而 $\hat{r}_{i,t,G}$ 归为“0”类，并训练判别器；第三步固定判别器模型并重新迭代生成器，使得到的 $\hat{r}_{i,t,G}$ 尽可能在判别器中被归为“1”类。

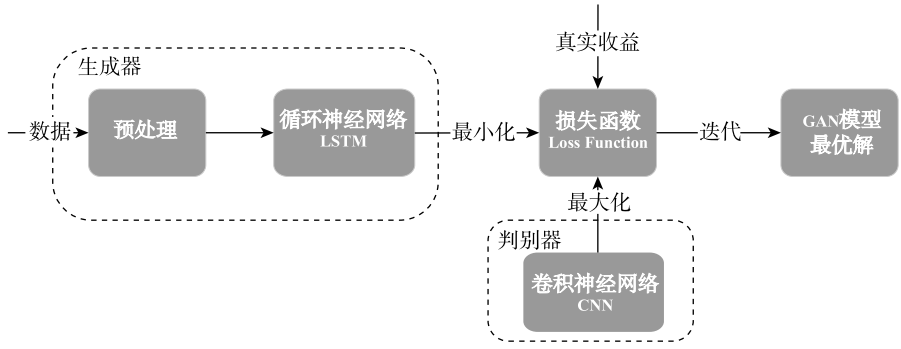

图 3 深度学习的算法结构

GAN 模型在训练过程中判别器的设定会较大程度影响到最终的预测精度，考虑到判别器为一个二分类模型，最初的 GAN 使用交叉熵（cross entropy）表示其目标函数：

$$J(D)=E_{x\sim p_{\text{data}}(r_{i,t})}\left[\log D(r_{i,t})\right]+E_{z\sim p_z(z_{i,t-1})}\left[\log\left(1-D(G(z_{i,t-1}))\right)\right], \tag{2}$$

其中，E 代表期望，D 和 G 分别为判别器和生成器，作为 Shannon（1948）信息论中一个重要概念，交叉熵主要用于度量两个概率分布间的差异性信息，公式（2）中对于真实收益 $r_{i,t}$ 和预测收益 $G(z_{i,t-1})$，最优的判别器 D 为使得 $D(r_{i,t})$ 为 1 而 $D\left(G(z_{i,t-1})\right)$ 为 0，即最大化 $J(D)$，而由于 G 和 D 是二元零和博弈，即生成器最优结果为使得 $D\left(G(z_{i,t-1})\right)$ 为 1，因此生成器的目标函数 $J(G)=-J(D)$，最终 GAN 的迭代问题可表示为在判别器最大化信息差异即交叉熵的前提下生成器通过迭代更新参数来最小化预测值和真实值的差别：

$$\begin{aligned}\min_G \max_D V(D,G)=&E_{x\sim p_{data}(r_{i,t})}\left[\log D(r_{i,t})\right]\\&+E_{z\sim p_z(z_{i,t-1})}\left[\log\left(1-D(G(z_{i,t-1}))\right)\right].\end{aligned} \tag{3}$$

考虑到在优化上述函数过程中会出现梯度消失和模型崩溃问题，我们使用改良的 Wasserstein 距离来替换原目标函数即公式（2）：

$$L=E_{x\sim p_{data}(r_{i,t})}\left[f_w(r_{i,t})\right]-E_{z\sim p_z(z_{i,t-1})}\left[f_w\left(G(z_{i,t-1})\right)\right], \tag{4}$$

其中，f_w 为新的判别器，即本文采用的 GAN 算法中使用的 CNN 模型，L 代表了估计值与真实值之间的差异，即生成器模型优化使得 L 达到最小的估计

收益 $\hat{r}_{i,t,G}$ 为最终模型输出值。优化目标函数时一般使用梯度下降（gradient descent）的算法，对于给定的函数 $L(\theta)$ 算法沿函数一阶导，即梯度 $\nabla_\theta L(\theta)$ 的反方向更新 θ 来最小化 $L(\theta)$，即 $\theta=\theta-\eta\nabla_\theta L(\theta)$，其中 η 为设定的迭代步长，本文使用随机梯度下降法（stochastic gradient descent，SGD）优化 GAN 模型。

我们同时构建了样本外 R^2 指标来表征模型预测能力：

$$R^2=1-\frac{\sum_{t=1}^{T}\sum_{i=1}^{N_t}(r_{i,t}-\hat{r}_{i,t})^2}{\sum_{t=1}^{T}\sum_{i=1}^{N_t}r_{i,t}^{\ 2}},\tag{5}$$

其中，$r_{i,t}$ 和 $\hat{r}_{i,t}$ 分别为各期股票和组合实际收益和预测收益值，R^2 取值范围为 $(-\infty,1]$，其值越高表明模型预测能力越好，当模型完整预测各期股票收益时 $R^2=1$。在模型迭代过程中，当训练集的 R^2 会持续上升，而验证集的 R^2 在上升到一定水平时会出现下降时，此时本文认为模型出现了过拟合并停止训练。

综上使用 GAN 模型进行预测主要分为以下几步（见图 4）：（1）将各股企业特征数据导入生成器（LSTM）进行训练，并得到初始预测值；（2）利用判别器（CNN）进行分类，将预测值归为“0”，真实股票收益归为“1”；（3）完成一次训练后，固定判别器模型参数，同时多次迭代生成器使得生成的新预测值尽可能被归入“1”类；（4）为了防止过拟合，引入验证集，并在验证集损失函数不再继续下降时停止训练，并单独取出生成器模型进行 $t+1$ 期收益预测；（5）滚动完成所有样本外时间预测。

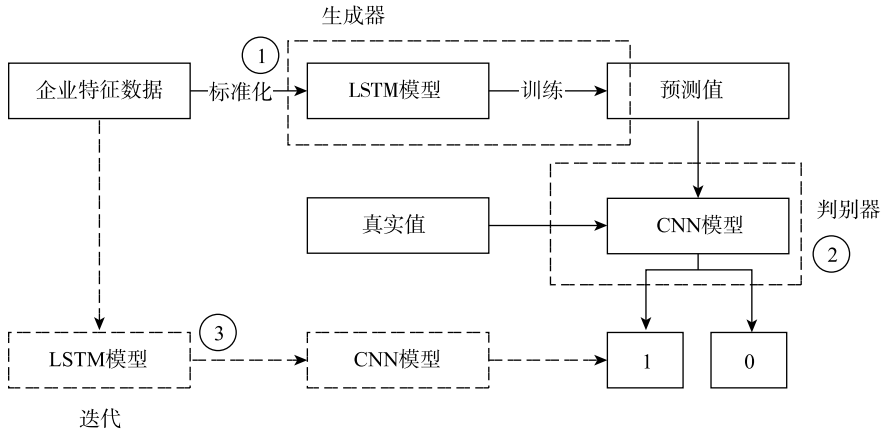

图 4　使用 GAN 模型进行预测的步骤

GAN 模型中使用的基础模型均基于神经网络，本文在拟合神经网络模型时设定损失函数（loss function）为预测值与实际值间的差异：

$$Loss=\frac{1}{N}\sum_{i=1}^{N}\left(\sum_{j=1}^{M}\frac{T_{j,i}}{T_j}(y_{j,i}-\hat{y}_{j,i})^2\right), \tag{6}$$

其中，$\hat{y}_{j,i}$ 为股票 j 在 i 时刻预测收益，$y_{j,i}$ 为实际收益，T_j 为股票 j 样本期月度累计上市时间，$T_{j,i}$ 为第 i 期时累计上市时间，整体上函数提高了近期预测误差的权重并减少远端预测结果的影响，同时考虑了新上市股票波动较大的情况增加了其误差权重。

此外，为了使结果更加稳健，本文在计算时采用集成训练（ensemble）的方法：在每次计算中构建不同的随机种子（random seeds）初始化模型，最终模型预测结果为单次计算结果的平均。

（二）数据介绍

本文选取 2003 年 1 月至 2017 年 12 月中国 A 股市场所有股票收益数据，计算股票超额收益使用的无风险收益采用月度的一年期国债收益率，市场组合收益为 A 股市场所有股票收益流通市值加权平均得到，数据来源为国泰安数据库。图 5 给出了本文使用的月度股票样本数。可以看出，我国上市公司数量随着年份呈增长趋势，整体均值为 1 520 家。

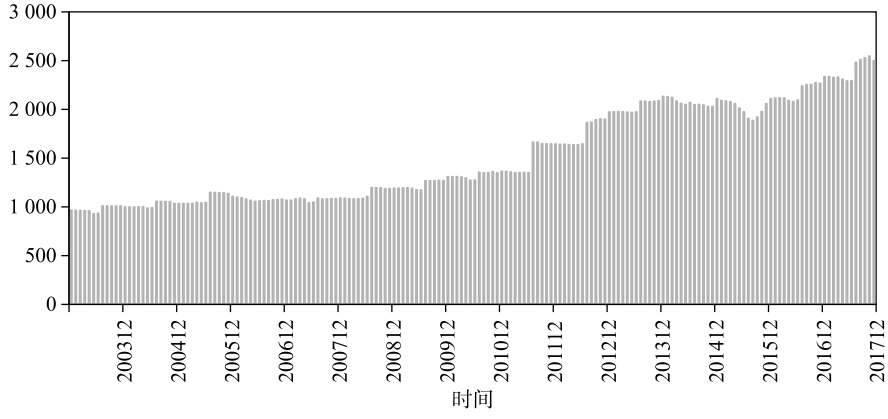

图 5 样本集月度有效企业数

特征因子大数据库方面，本文依据 Jiang *et al*.（2018）构建了 6 个大类 74 个企业特征因子指标：估值与成长类、投资类、盈利类、惯性类、交易摩擦类和无形资产类。[1] 本文还计算了每一期的 74 个企业特征因子指标，最后总计获得 148 个特征因子指标。考虑到不同变量间的数量级存在差异，本文对各变量月度横截面数据进行了标准化处理，即 $z_{scale}=\frac{z-\bar{z}}{\sigma_z}$，其中 $\bar{z}$ 和 σ_z 为各月度数据的均值和方差。

1 限于篇幅，企业特征因子指标的详细介绍未列示，感兴趣的读者可向作者索取。

四、实证研究

（一）股票收益预测精度

我们首先研究了 GAN 深度学习模型对我国 A 股收益的预测能力。表 1 展示了 GAN 模型以及线性回归、弹性网络以及长短期记忆网络模型三类基准模型的样本外预测 R^2（见公式（5））。研究结果发现，传统的线性回归模型 LR 表现最差，R^2 为－6.45%。弹性网络模型 EN 考虑了多重共线性和过拟合等问题，模型表现好于 LR，为 0.15%。长短期记忆网络模型 LSTM 考虑了非线性，并通过记忆单元保存前期有效时序信息，相比线性模型有更好的表现，预测能力达到 0.45%。但是，GAN 深度学习引入动态博弈，相比单一 LSTM 模型表现上有了进一步的提升，其预测表现最好，R^2 达到 0.89%[2]，显著优于线性模型 LR、EN 和非线性模型 LSTM。

表 1 同时给出了各模型均方预测误差（mean squared forecasting error，MSFE）的计算结果，以及与基准 LR 模型的数值比：

$$MSFE=\frac{1}{T}\sum_{t=1}^{T}\frac{1}{N_t}\sum_{i=1}^{N_t}(r_{i,t}-\hat{r}_{i,t})^2. \tag{7}$$

表 1 显示，各类模型中，LR 预测误差最大，达到了 2.09%；相比 LR，其他三类模型 *MSFE* 均小于 2%，其中 GAN 模型误差值最小，为 1.93%，相比 LR 减少了约 8%。

表 1　模型样本外预测 R^2 及 MSFE

(%)	LR	EN	LSTM	GAN
R^2	－6.45	0.15	0.45	0.89
MSFE	2.09	1.96	1.95	1.93
比值	—	93.77	93.30	92.34

图 6 展示了各类模型在 2009—2017 年各年度下的预测精度 R^2，相比 EN 和 LSTM，GAN 模型在不同年度的表现更为稳定，在三类模型表现均出现下降的 2009—2012 年度，弹性网络 R^2 最低为－8.36%，区间波动幅度达到了 15.7%。而 GAN 模型区间波动幅度仅为 6.8%，且后续年度中 GAN 模型在预测稳定性和预测精度上要高于对照模型。

[2] Gu *et al*.（2020）文中针对美国市场计算了各机器学习模型的 R^2，其中最高的为三层神经网络的 0.40%，低于本文使用的 GAN 模型。

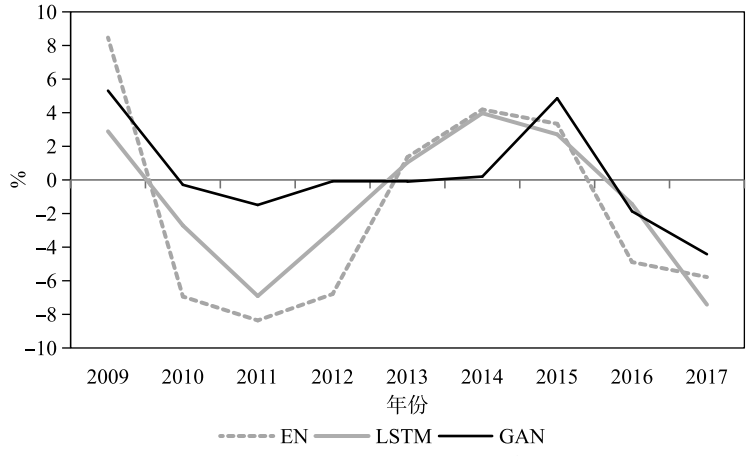

图 6 各模型年度与预测 R^2

图 7 展示了 GAN 模型对各行业股票的预测精度（2009—2017 年度）[3]，预测精度较高的行业包括住宿和餐饮（H）、科技（M）、卫生（Q）和综合（S），而能源（D）、租赁服务（L）和文化娱乐（R）的行业 R^2 为负值。另外占比最高的制造业（C）模型预测精度为 0.35%。

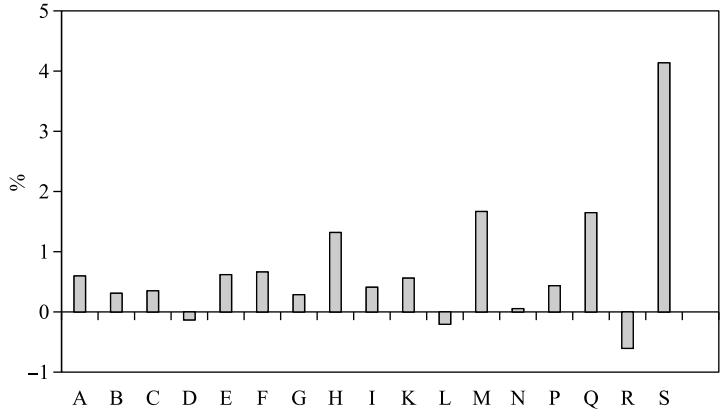

图 7 GAN 模型行业预测 R^2

参考 Gu *et al*.（2020），我们使用 Diebold-Mariano 检验各模型间的预测差异，检验公式如下

$$d_{12,t}=\frac{1}{n_3}\sum_{i=1}^{n_3}\left((\hat{e}_{i,t}^{(1)})^2-(\hat{e}_{i,t}^{(2)})^2\right), \tag{8}$$

其中 $\hat{e}_{i,t+1}^{(1)}$ 和 $\hat{e}_{i,t+1}^{(2)}$ 表示两类模型在 t 期对于股票 i 的预测误差，n_3 为当期的个股数，定义 $DM_{12}=\frac{\overline{d}_{12}}{\hat{\sigma}_{d12}}$，$\overline{d}_{12}$ 和 $\hat{\sigma}_{\bar{d}_{12}}$ 分别为 $d_{12,t}$ 的均值和标准差。表 2 显示，各

[3] 行业分类依据中国证监会《上市公司行业分类指引》中的一级行业分类代码。

模型中表现最好的是 GAN 模型，表现最差的是 LR 模型，LSTM 模型表现不如 GAN，但显著高于其他模型。

表 2　模型间预测误差比较

	EN	LSTM	GAN
LR	0.76	**4.14**	**4.13**
EN		**1.66**	**1.82**
LSTM3			**1.76**

注：表中的数为正数表明该列包含的模型的预测能力优于与此相对应的行所指的模型，标粗字体为比较结果在 5%显著性水平内显著。

（二）因子投资组合表现

我们接着根据 GAN 深度学习模型预测构建因子投资策略并研究其绩效。我们在月度频率上依据不同模型对个股的预测收益大小构建投资组合：在每月的第一个交易日，依据模型预测的收益结果对样本股票进行升序排序，等分成 10 个投资组合。其中第一组表示预测收益最小的前 10%的股票，标记为“L”组；第十组表示预测收益最高的前 10%的股票，标记为“H”组。其他组按预测收益的高低分别记为 2 组至 9 组。本文通过做多“High”组，卖空“Low”组构建多空对冲组合，即“H-L”组，并将投资组合持有一个月，在下个月初重复上述过程构建和持有新的投资组合直到样本期结束。组合收益的权重采用流通市值加权。

表 3 汇报了不同模型下各因子投资组合的收益（Ret）、方差（Std）和年化的夏普比率（SR）。研究结果发现，基于神经网络的两类模型多空组合 H-L 收益及夏普比率高于线性模型，其中 GAN 模型 H-L 收益为 1.13%（年化为 13.56%），相比 LSTM（0.93%）提升 24%，夏普比率同样达到了最高的 0.71。

图 8 展示了各模型的多空对冲组合 H-L 的累计收益与同期市场收益对比。发现，GAN 模型整体表现最好，各个时期中回撤均较小，累计最高收益在 2015 年 5 月达到了 124%。

表 3　各模型投资组合分析

	LR			EN		
	Ret	Std	SR	Ret	Std	SR
L	1.15	8.56	0.46	1.09	8.74	0.43
2	1.21	8.54	0.49	1.03	8.74	0.41

（续表）

	LR			EN		
	Ret	Std	SR	Ret	Std	SR
3	1.07	8.70	0.43	0.94	8.51	0.38
4	1.22	8.66	0.49	1.24	8.36	0.51
5	1.61	8.81	0.63	1.65	8.50	0.67
6	1.34	8.68	0.53	1.47	8.47	0.60
7	1.19	8.13	0.51	1.52	8.16	0.64
8	1.55	8.05	0.66	1.42	8.12	0.60
9	1.38	8.25	0.58	1.37	8.52	0.56
H	0.56	8.14	0.24	0.56	8.10	0.24
H-L	−0.59	5.22	−0.38	−0.52	5.92	−0.30
	LSTM			GAN		
	Ret	Std	SR	Ret	Std	SR
L	0.07	9.76	0.02	0.32	7.50	0.15
2	0.25	9.22	0.09	1.22	8.19	0.51
3	0.78	9.45	0.29	0.91	8.44	0.37
4	0.49	10.02	0.17	1.21	8.58	0.49
5	0.66	9.54	0.24	1.12	8.30	0.47
6	0.73	9.58	0.26	1.20	8.52	0.49
7	0.71	9.56	0.26	1.28	8.66	0.51
8	0.90	9.89	0.31	1.54	8.72	0.61
9	1.20	9.66	0.43	1.25	8.95	0.48
H	1.00	9.54	0.36	1.45	8.75	0.57
H-L	0.93	6.54	0.49	1.13	5.50	0.71

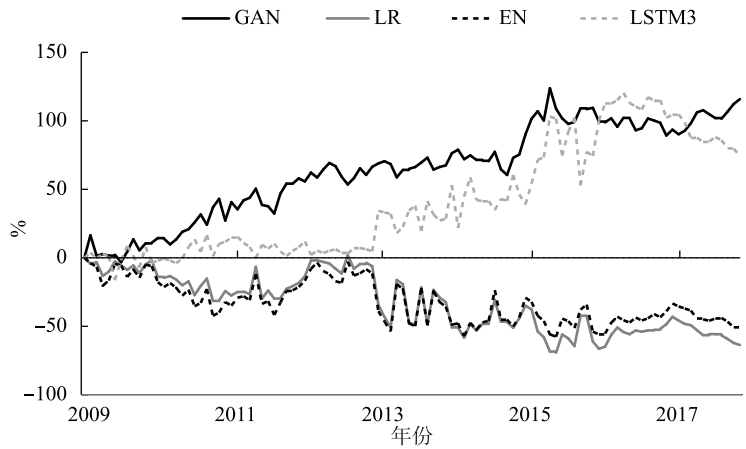

图 8 多空策略组合累计收益

我们使用 Fama-French 三因子和五因子模型[4]对模型 H-L 组合收益进行拟合，并给出超额收益 α 和拟合优度 R^2。发现，线性模型 LR、EN 组合的 FF3-α 和 FF5-α 不显著，其收益可由多因子模型解释。但是，GAN 模型的 FF3-α、FF5-α 分别达到了 1.13%和 1.01%，年化为 13.56%和 12.12%，并在 5%水平显著，不能被现有的 Fama-French 多因子模型解释。

表 4　模型在 FF3 和 FF5 中的超额收益及拟合优度

	LR	EN	LSTM	GAN
FF3-α	0.20	0.36	0.69	**1.13**
R^2	43.4	44.6	2.00	2.00
FF5-α	0.00	0.15	1.18	**1.01**
R^2	45.1	45.5	2.90	2.90

注：标粗字体为在 5%显著性水平内显著。

（三）因子重要度分析

本文以单独剔除各因子后模型 R^2 下降程度代表该因子的重要度，比较各特征因子的相对重要性。图 9 为排序最高的 10 个企业特征因子在 GAN 模型中的重要度。

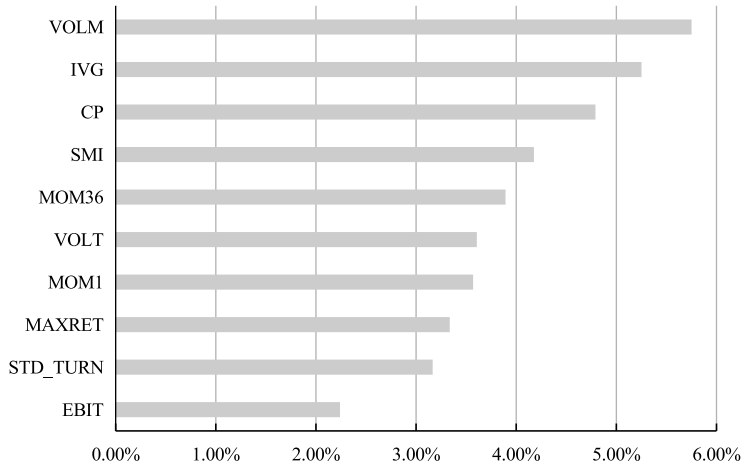

图 9　企业特征因子重要度

[4] 三因子和五因子数据来源中国资产管理研究中心，见 http://sf.cufe.edu.cn/kxyj/kyjg/zgzcglyjzx/index.htm，访问时间：2020 年 3 月 20 日。

图 9 显示，对各因子重要度进行归一化处理后，最重要特征因子可以分为三类：第一类为价格及交易量趋势类，包括 1 个月动量（MOM1）、36 个月动量（MOM36）、历史最高收益（MAXRET）、交易量趋势（VOLT）和交易量动量（VOLM）等；第二类为流动性指标类，包括换手率（TURN）和换手率方差（STD_TURN）等，换手率通常用来反映市场流动性，换手率越高表示股票交投活跃，流动性好，变现能力强，投资者更倾向于购买这类股票，但换手率标准差的突然放大往往伴随着较大的波动风险和投机风险；第三类为基本面指标，包括存货增长率（IVG）、销量增长与存货增长差（SMI），现金生产率（CP）和息税前利润（EBIT）等。销售和库存的变动代表了公司的周转能力，行业内优良的企业具有将产品快速变现的能力，一定程度上增强了企业抵御风险的能力，降低了企业风险。同时，无论是存货增长率还是销售与库存增长额差值均反映了企业的运营状态和其基本面价值，而后者是股票价值最重要的驱动因素。整体上前十大特征因子贡献度占到了所有因子的 40％左右。

五、基于错误定价理论的收益预测来源分析

我们接着从金融摩擦、波动、流动性等角度研究错误定价理论对 GAN 深度学习模型预测能力的解释能力。我们参照 Fama and French（2015），使用条件双排序组合法，在每年 7 月底按照企业特征在 30％和 70％分位点将股票分为低（L）、中（M）、高（H）三组，这一分类持续到下一年 6 月，并在下一年度 7 月重复上述步骤，在上述各组中每月按照预测收益在 30％和 70％分位点再进行分组，最终得到的每月投资组合数为 9 个（3×3），全样本区间为 2009 年 7 月至 2017 年 12 月。按照流通市值加权的方式计算各组合月度收益，并检验在企业特征分类下的多空股票组合（H-L）收益的显著性。

（一）金融摩擦

表 5 为使用市值和账面市值比因子与 GAN 模型预测收益进行双排序的组合构建结果，相比中小市值股票，GAN 模型对于高市值股票（Size 中的 B 列）的预测能力更高，多空组合月度收益为 0.71％高于中小市值股票样本，相比大市值股票，小市值股票交易摩擦更高，因此 GAN 模型对金融交易摩擦低的股票定价更为有效；而账面市值比组合中，GAN 对中性股票（BM 中的 M 列）的预测为显著的，多空组合月度收益为 0.88％（$t=2.29$），FF3 检验的 α 为 1.05％（$t=2.27$），成长型股票（BM 中的 L 列）和价值型股票（BM 中的 H 列）的多空组合收益并不显著。

表 5　双排序组合结果（交易摩擦）

	Size			BM		
	S	M	B	L	M	H
L	1.36	0.79	0.04	0.39	−0.07	0.44
M	1.51	1.05	0.47	0.52	0.70	0.69
H	1.68	0.97	0.75	0.69	0.81	1.11
H-L	0.32	0.18	0.71	0.30	**0.88**	0.67
t 值	1.17	0.65	1.73	0.94	2.29	1.52
FF3-α	0.31	0.11	0.70	0.32	**1.05**	0.54
t 值	1.05	0.39	1.47	0.92	2.27	1.01

注：标粗字体为在 5%显著性水平内显著。

GAN 模型对于大市值股票的预测能力更好，主要原因在于相比小市值企业，大市值企业的股票金融交易摩擦更低，股价更多地反映了基本面的信息，因此基于基本面大数据的预测模型表现更好；而国内的小市值企业受“借壳”等因素影响（Liu *et al*.，2019），股票往往偏离核心价值呈现更大的差异化表现，此外小市值股票交易摩擦高且股价波动大，易受到散户以及游资这类“非理性”群体的关注，因此其股价更多受到市场情绪的影响。

（二）波动及不确定性

表 6 给出使用贝塔和特质波动率因子与 GAN 模型预测收益进行双排序的组合构建结果，其中贝塔（BETA）和特质波动率（IVOL）指标参考 Fama and MacBeth（1973）和 Ali *et al*.（2003）构建而成，其中贝塔使用各股周收益率与市场组合收益滚动回归得到，滚动窗口为 3 年，计算公式为：

$$R_{i,t}=a_{i,t}+\beta_{i,t}R_{m,t}+\mu_{i,t}, \tag{9}$$

其中，R_i 为各股收益，$R_{m,t}$ 为市场组合收益。特质波动率同样使用 3 年窗口期的各股周收益和周市场组合收益回归的残差取标准差计算得到。相比高波动股票，GAN 模型对于低贝塔股票（Beta 中的 L 列）的预测能力更高，其多空组合月度收益为 0.74%；GAN 模型对高特质波动率股票（IVOL 中的 H 列）的预测更为显著，多空组合月度收益为 0.79%（$t=2.24$）。

表 6　双排序组合结果（波动及不确定性）

	BETA			IVOL		
	L	M	H	L	M	H
L	0.22	0.46	0.27	0.44	0.49	−0.34
M	0.81	0.64	0.48	0.91	0.62	0.31

（续表）

	BETA			IVOL		
	L	M	H	L	M	H
H	0.96	0.52	0.50	1.25	0.59	0.45
H-L	0.74	0.06	0.23	0.81	0.10	**0.79**
t 值	1.76	0.22	0.77	1.82	0.28	2.24
FF3-α	0.78	−0.02	0.25	0.71	0.20	0.71
t 值	1.64	−0.06	0.74	1.44	0.55	1.71

注：标粗字体为在 5%显著性水平内显著。

GAN 模型对于低波动性股票的预测效果更好，其原因同样在于高波动性股票更多包含了市场情绪等信息，而 GAN 模型主要反映了基本面信息，且 GAN 模型更新频率较低（按年更新）影响了模型对短期情绪变动的响应能力；再者，姜富伟等（2021）指出中国股市存在明显的风险定价偏误，高波动（贝塔）股票的风险与收益并不对等，因此模型在估计这类股票收益时会存在较大误差。

（三）换手率和流动性

表 7 给出使用换手率（TURN）和流动性（ILLIQ）因子与 GAN 模型预测收益进行双排序的组合构建结果，其中换手率指标和流动性指标参考 Datar *et al*.（1998）和 Amihud（2002）构建而成。其中换手率的计算公式为：

$$TURN = VOL_t / LNS_t, \tag{10}$$

其中，VOL_t 为 t 期的总交易，LNS_t 为 t 期的总流通股票量，本文使用的 TURN 指标为最近 3 个月的均值。ILLIQ 指标定义为绝对收益除以交易量的最近 12 个月均值。可以看到，相比高换手率股票，GAN 模型对于低换手股票（TURN 中的 L 列）的预测能力更高，其多空组合月度收益为 0.79%；GAN 模型对低流动性股票（ILLIQ 中的 L 列）的预测为显著的，多空组合月度收益为 0.75%（$t = 1.95$），而中高流动性股票的多空组合收益并不显著。

表 7 双排序组合结果（流动性）

	ILLIQ			TURN		
	L	M	H	L	M	H
L	0.00	0.75	1.33	0.14	0.25	0.22
M	0.44	0.88	1.47	0.89	0.45	0.52

（续表）

	ILLIQ			TURN		
	L	M	H	L	M	H
H	0.75	0.85	1.33	0.93	0.79	0.72
H-L	**0.75**	0.10	0.00	0.79	0.54	0.50
t 值	1.95	0.34	0.00	1.81	1.76	1.51
FF3-α	0.70	0.02	0.04	0.80	0.53	0.51
t 值	1.51	0.05	0.14	1.60	1.54	1.33

注：标粗字体为在5%显著性水平内显著。

总之，上述三类错误定价理论探索表明，GAN深度学习模型整体上对大市值、低波动、高流动性的股票预测精度更高。大市值、低波动、高流动性的股票错误定价程度更低，其股价更多反映出基于基本面的价值信息，这说明GAN模型预测能力是基本面驱动而不是投机情绪驱动。

六、深度学习因子的经济机制分析

本部分主要从我国的宏观经济活跃度、经济政策与金融市场不确定性、投资者情绪等视角，研究了基于深度学习的资产定价模型与宏观经济和金融市场风险之间的关联。使用的回归方程如下：

$$HML_t = \alpha + \beta_{MKT} r_{MKT,t} + \beta_{SMB} r_{SMB,t} + \beta_{HML} r_{HML,t} + \beta_{RMW} r_{RMW,t} + \beta_{CMA} r_{CMA,t} + \varphi Dummy_t^{high} + \varepsilon_t, \tag{11}$$

其中，被解释变量 HML_t 代表第 t 期的由GAN方法构造的多空对冲组合（H-L组）的超额收益。式（11）中，本文使用Fama-French五因子模型中的市场因子、规模因子、价值因子、盈利因子和投资因子，分别用 $r_{MKT,t}$、$r_{SMB,t}$、$r_{HML,t}$、$r_{RMW,t}$、$r_{CMA,t}$ 表示，作为解释多空对冲组合收益的控制变量。对于每一个宏观经济变量，本文在样本期内按照二分法分成高、低两个时期，并构造虚拟变量 $Dummy^{high}$（指标处于取值较高的时期）。

（一）与深度学习因子负向关联的宏观经济变量

本文基于深度学习方法对股票收益预测的高低构造了多空对冲组合，该组合反映了深度学习方法捕捉到的基于公司特征的横截面收益变动规律，因此本文将其称为深度学习因子。本节总结了与深度学习因子存在负向关联的经济变量，见表8（方括号内汇报了 t 值），即当经济处于这些变量的高值区间时，深度学习因子捕捉到的正向预期收益关系将减弱。

表 8 表明，当我国的宏观经济处于扩张状态时，深度学习因子的预测效果有所降低。其中，当企业的新增固定资产投资处于高值时，深度学习因子所带来的月度超额收益会降低 1.62%（$t=-1.69$）。另外，当社会消费品零售总额上升时，深度学习因子所带来的超额收益会下降 2.28（$t=-2.07$）。根据基于消费的资产定价模型，当投资者更加倾向于将资产进行消费时，其金融资产的收益率将降低。当社会融资规模较大或 IPO 首日收益率较高时，深度学习因子的月度收益率会分别降低 2.03%（$t=-1.93$）或 1.90%（$t=-2.15$）。社会融资规模较大或 IPO 首日收益率较高往往意味着投资者情绪处于一个较高的状态，因此股票市场价格易于处在偏离均衡价格的水平。最后，本文发现当再贴现利率处于较高状态时，深度学习因子的月度超额收益将降低 1.66%（$t=-2.39$）。此外值得注意的是，在表 8 的结果中，原本深度学习因子能够产生显著的超额收益，而一旦上述的经济变量处于高值区间时，这些显著的超额收益会降低不少。因此，在使用深度学习方法进行投资实践中，应特别注意上述的宏观经济以及市场风险。

表 8 负向影响深度学习因子的经济变量

宏观经济变量	新增固定资产投资	社会消费品零售总额	社会融资规模	IPO 首日收益率	再贴现利率
α	2.18	2.93	2.60	1.70	1.13
	[2.36]	[2.58]	[3.04]	[2.75]	[2.20]
φ	−1.62	−2.28	−2.03	−1.90	−1.66
	[−1.69]	[−2.07]	[−1.93]	[−2.15]	[−2.39]
MKT	0.34	0.32	0.36	0.35	0.33
	[1.40]	[1.56]	[1.61]	[2.04]	[1.57]
SMB	−0.07	−0.09	−0.05	−0.13	−0.09
	[−0.32]	[−0.30]	[−0.17]	[−0.40]	[−0.30]
HML	0.22	0.21	0.19	0.20	0.21
	[1.82]	[1.77]	[1.65]	[1.77]	[1.84]
RMW	0.53	0.51	0.49	0.49	0.47
	[1.48]	[1.44]	[1.36]	[1.42]	[1.51]
CMA	−0.34	−0.37	−0.37	−0.30	−0.40
	[−1.15]	[−1.21]	[−1.44]	[−0.99]	[−1.29]

（二）与深度学习因子正向关联的经济变量

本节探讨了与深度学习因子存在正向关联的经济变量，即当经济处于这些变量的高值区间时，深度学习因子捕捉到的正向预期收益关系将增强，结果见表 9。

对表 9 的结果进行分析，当我国金融市场较活跃（波动率大）、外部市场不确定性较大、外贸较活跃、物价指数较高、消费者满意度较高时，深度学习因子的超额收益率显著增加。具体而言，当我国的市场波动率处在较高区间时，此时的深度学习因子月度收益率会增加 2.35%（$t=2.49$）。这说明市场较为活跃、不确定性较高时，深度学习方法的有效程度更高。而当美国的贸易政策不确定性增强时，此时深度学习因子的月度收益率会提高 1.92%（$t=2.48$）。当我国的外贸货物量较高时，深度学习因子的月度超额收益会增加 2.20%（$t=1.70$）。这说明繁荣的对外贸易能够有助于我国企业更好地进行生产，进而获取更高的利润。CPI 环比处于较高水平时，深度学习因子的月度收益将增加 2.01%（$t=2.49$）。此时的物价指数较高，短期内企业通过售卖产品获得的现金流较高，因此收益上升。当消费者满意度较高时，此时深度学习因子的月度收益上升 2.19%（$t=1.85$）。这说明消费者满意度较高时，消费者对企业产品甚至是企业资产的青睐程度较高，此时企业基本面特征得到进一步改善的可能较大，而深度学习因子正好能捕捉到这部分信息。同样值得关注的是，在表 9 的结果中，在控制了经济变量后，深度学习因子本不能产生显著的超额收益，只有当上述的经济变量处于高值区间时，才能获得显著的超额收益。因此，在使用深度学习方法进行投资实践中，应特别关注上述的经济变量反映的市场状态，把握较好的投资机会。

表 9　正向影响深度学习因子的经济变量

宏观经济变量	市场波动率	美国贸易政策不确定性	外贸货物量	CPI 环比	消费者满意度
α	−0.12	−0.39	−1.04	−0.07	0.54
	[−0.26]	[−0.60]	[−0.84]	[−0.14]	[0.84]
φ	2.35	1.92	2.20	2.01	2.19
	[2.49]	[2.48]	[1.70]	[2.49]	[1.85]
MKT	0.30	0.37	0.29	0.32	0.40
	[1.37]	[1.62]	[1.34]	[1.49]	[1.40]
SMB	−0.07	0.01	−0.10	−0.04	0.03
	[−0.32]	[0.02]	[−0.35]	[−0.15]	[0.10]

（续表）

宏观经济变量	市场波动率	美国贸易政策不确定性	外贸货物量	CPI 环比	消费者满意度
HML	0.20	0.22	0.23	0.23	0.20
	[1.89]	[2.63]	[2.03]	[2.38]	[1.77]
RMW	0.39	0.41	0.37	0.36	0.39
	[1.48]	[1.23]	[1.18]	[1.25]	[1.11]
CMA	−0.36	−0.48	−0.43	−0.52	−0.40
	[−1.10]	[−1.56]	[−1.57]	[−2.10]	[−1.36]

（三）深度学习与公司基本面状况预测

上述实证结果已显示 GAN 模型提取的信息能够正向预测个股层面的未来收益，本节试图探讨这种收益预测的正向相关是否也和公司基本面状况信息预测相关。我们参考 Fama and MacBeth（1973），使用 Fama-MacBeth 回归的方法进行研究，见公式（12）。

$$FP_{i(t+j)} = c_0 + c_1 GAN_{it} + c_2 SIZE_{it} + c_3 BM_{it} + c_4 MOM_{it} + c_5 BETA_{it} + \varepsilon_{it}, \quad (12)$$

其中，$FP_{i(t+j)}$ 代表第 i 家公司在未来 j 期的公司基本面特征；GAN_{it} 代表第 i 家公司在 t 期的收益预测信息；$SIZE_{it}$、BM_{it}、MOM_{it}、$BETA_{it}$ 是回归的控制变量，分别为第 i 家公司在 t 期的规模、价值、惯性、系统性风险敏感程度的大小。本文使用基于季报数据的销售收入增长率、净利润增长率、毛利润率、资产收益率、现金流资产比、现金流价格比来衡量公司的基本面特征，见表 10。

表 10 表明，深度学习模型能够有效捕捉到公司在未来 24 个月内的基本面状况信息。具体而言，对销售收入增长率、净利润增长率、毛利润率、资产收益率等销售与盈利指标的预测效果在 6 个月左右最显著，此后随着预测时期长度的增加，其预测效果逐渐降低。当预测长度达到 36 个月时，深度学习指标的预测性基本不显著。而对现金流资产比、现金流价格比等现金流指标，深度学习模型的预测效果在 12 个月预测期限下最显著，t 值分别达到 2.45 和 2.59，此后逐渐降低，到 36 个月时基本不显著。GAN 模型在提取公司特征因子信息时包含了多种类型的指标，比如包括市场交易类指标、趋势类指标、流动性指标等中短期交易指标，这些市场交易类因子指标可能更多地反映了公司近期的运营状况，以及投资者对于公司近期表现的看法。总之，表 10 表明，GAN 深度学习模型可以有效捕捉公司在中短期（未来一年以内）的销售、盈利、现金流等基本面状况信息。

表 10　深度学习指标与公司基本面特征的多期预测

公司基本面特征	$j=6$	$j=12$	$j=24$	$j=36$
销售收入增长率	0.52 [2.03]	0.43 [1.89]	0.38 [1.63]	0.15 [0.96]
净利润增长率	0.57 [2.11]	0.39 [1.90]	0.41 [1.93]	0.18 [1.10]
毛利润率	0.35 [2.06]	0.21 [1.66]	0.29 [1.97]	0.17 [1.54]
资产收益率	0.14 [2.26]	0.16 [2.21]	0.09 [1.78]	0.03 [0.74]
现金流资产比	0.26 [2.21]	0.27 [2.45]	0.22 [1.82]	0.18 [1.50]
现金流价格比	0.82 [2.34]	0.95 [2.59]	0.63 [1.73]	0.32 [1.30]

七、结论与启示

本文使用 2003—2017 年中国股票市场数据构建了企业特征因子大数据库，并利用基于生成式对抗网络（GAN）的深度学习模型，研究了中国股票市场的可预测性和因子投资策略，得出以下主要结论。首先，相比线性模型，GAN 深度学习预测精度显著提升，其最高预测 R^2 为 0.89。其次，依据模型预测构建股票因子投资组合，发现多空策略下投资收益为 1.13%（年化为 13.56%），最高夏普比率为 0.71，累计投资收益显著高出同时期的其他模型收益。本文还将各因子重要度进行排序，并给出理论分析。最后，经济理论机制分析，我们探索了错误定价理论的解释能力，还找到了能够负向或正向影响深度学习预测效果的经济变量，发现深度学习模型能有效挖掘股票市场与宏观经济风险、投资者情绪之间的关系，还能有效预测微观企业在未来一年内的盈利、收入和现金流等基本面状况。

本文的社会经济意义在于，更深刻地理解深度学习等人工智能技术对于金融市场高质量发展的价值。利用金融大数据和人工智能技术，投资者能更高效地开展投资管理和风险管理；金融监管者也能更好地监控金融市场动态和管控系统性风险，完善金融市场的制度，提升金融市场运行质量和效率，提高金融服务实体经济高质量发展的能力。本文的学术意义在于，从大数据分析和人工智能视角来理解中国股票市场的特征规律，深入开展经济机制分析，探讨人工智能技术与经济金融理论之间的逻辑关联。展望未来，人工智能将持续融入并改进金融理论，更深刻地揭示金融市场的复杂运行规律；同时，金融与人工智能交叉融合也要更重视理论创新和探索，密切联系宏观经济运行、投资者非理性行为、政府政策制定和文化制度要素，为模型调参和数据特征工程等提供经济理论指导，提升人工智能技术在金融领域的适用性。

参 考 文 献

[1] Ali, A., L. Hwang, and M. Trombley, "Arbitrage Risk and the Book-to-Market Anomaly", *Journal of Financial Economics*, 2003, 69 (2), 355-373.

[2] Amihud, Y., "Illiquidity and Stock Returns: Cross Section and Time-series Effects", *Journal of Financial Markets*, 2002, 5 (1), 31-56.

[3] Chen, L., M. Pelger, and J. Zhu, "Deep Learning in Asset Pricing", Working Paper, 2019.

[4] 陈卫华、徐国祥，"基于深度学习和股票论坛数据的股市波动率预测精度研究"，《管理世界》，2018 年第 1 期，第 180—181 页。

[5] Datar, V., N. Naik, and R. Radcliffe, "Liquidity and Stock Returns: An Alternative Test", *Journal of Financial Markets*, 1998 (1), 203-219.

[6] Fama, E., and K. French, "A Five-factor Asset Pricing Model", *Journal of Financial Economics*, 2015, 116 (1), 1-22.

[7] Fama, E., and J. Macbeth, "Risk, Return, and Equilibrium: Empirical Tests", *Journal of Political Economy*, 1973, 81 (3), 607-636.

[8] Feng, G., N. Polson, and J. Xu, "Deep Learning in Asset Pricing", Working Paper, 2018a.

[9] Feng, G., J. He, and N. Polson, "Deep Learning for Predicting Asset Returns", Working Paper, 2018b.

[10] Gadelha, M., S. Maji, and R. Wang, "3D Shape Induction from 2D Views of Multiple Objects", 2017 International Conference on 3D Vision, 2017, 402-411.

[11] Goodfellow, I., J. Pouget-Abadie, and M. Mirza, "Generative Adversarial Nets", International Conference on Neural Information Processing Systems, 2014, 2672-2680.

[12] Gu, S., B. Kelly, and D. Xiu, "Empirical Asset Pricing via Machine Learning", *Review of Financial Studies*, 2020, 33 (5), 2223-2273.

[13] Harvey, R., Y. Liu, and H. Zhu, "…and the Cross-Section of Expected Returns", *Review of Financial Studies*, 2016, 29, 5-68.

[14] Heaton, J., N. Polson, and J. Witte, "Deep Learning in Finance", 2016, arXiv: 1602.06561.

[15] Hou, K., C. Xue, and L. Zhang, "Which Factors?", *Review of Finance*, 2019, 23 (1), 1-35.

[16] Huang, D., F. Jiang, J. Tu, and G. Zhou, "Investor Sentiment Aligned: A Powerful Predictor of Stock Returns", *Review of Financial Studies*, 2015, 28 (3), 791-837.

[17] Jiang, F., G. Tang, and G. Zhou, "Firm Characteristics and Chinese Stocks", *Journal of Management Science and Engineering*, 2018, 4 (3), 259-283.

[18] 姜富伟、涂俊、D. Rapach、J. Strauss、周国富，"中国股票市场可预测性的实证研究"，《金融研究》，2011 年第 9 期，第 107—121 页。

[19] 姜富伟、唐国豪、衣英男、周国富，"金融大数据资产价格预测：人工智能视角"，工作论文，2019 年。

[20] 姜富伟、马甜、张宏伟，"高风险低收益？基于大数据和机器学习的动态 CAPM 模型的解释"，《管理科学学报》，2021 年第 1 期，第 109—126 页。

[21] 李斌、邵新月、李玥阳，"机器学习驱动的基本面量化投资研究"，《中国工业经济》，2019 年第 8

期，第 61—79 页。

[22] Liu，J.，R. Stambaugh，and Y. Yuan，“Size and Value in China”，*Journal of Financial Economics*，2019，134，48-69.

[23] Salimans，T.，I. Goodfellow，W. Zaremba，et al.，“Improved Techniques for Training Gans”，Advances in Neural Information Processing Systems，2016，29.

[24] 苏治、卢曼、李德轩，“深度学习的金融实证应用：动态、贡献与展望”，《金融研究》，2017 年第 5 期，第 111—126 页。

[25] Vondrick，C.，H. Pirsiavash，and A. Torralba，“Generating Videos with Scene Dynamics”，Conference on Neural Information Processing Systems，2016，613-621.

[26] 曾志平、萧海东、张新鹏，“基于 DBN 的金融时序数据建模与决策”，《计算机技术与发展》，2017 年第 4 期，第 1—5 页。

[27] Zhao，J.，M. Mathieu，and Y. Lecun，“Energy-based Generative Adversarial Network”，2016，arXiv1609.03126.

Deep Learning and Factor Investing in Chinese Stock Market
—Based on Generative Adversarial Networks

TIAN MA
(*Minzu University of China*)
FUWEI JIANG*
(*Central University of Finance and Economics*)
GUOHAO TANG
(*Hunan University*)

Abstract We estimate a non-linear asset pricing model with deep neural networks which are applied to China's A-share stock market with a large set of firm characteristics-based factors. Empirically, the deep learning model outperforms other linear benchmarks out-of-sample both in stock and portfolio level. We also analyze the relative importance of firm characteristics factors and explore economic mechanism explanations. This study has important implications for market efficiency and asset pricing of Chinese financial market in the big data age.

Keywords deep learning, asset pricing, factor investing

JEL Classification C45, G11, G12

* Corresponding Author: Fuwei Jiang, School of Finance, Central University of Finance and Economics, No. 39 South Xueyuan Road, Haidian District, Beijing 100089, China; Tel: 86-18511086494; E-mail: fwjiang@qq.com.