doi:10.19920/j.cnki.jmsc.2024.09.006

基于自编码机器学习的资产定价研究①
——中国股票市场的金融大数据分析视角

唐国豪[1], 朱 琳[2*], 廖存非[3], 姜富伟[4,5]
(1. 湖南大学金融与统计学院, 长沙 410006; 2. 广东财经大学金融学院, 广州 510320;
3. 南京理工大学经济管理学院, 南京 210094; 4. 厦门大学经济学院, 厦门 361005;
5. 厦门大学王亚南经济研究院, 厦门 361005)

摘要: 本研究在中国股票市场上,使用自编码机器学习方法和包含近百个公司特征变量的金融大数据,对资产价格进行解释和预测,并对自编码因子进行全面的宏观经济分析. 研究发现,自编码因子能够从包含公司特征的大量信息中提取到有效的收益预测信号,并在横截面上获得显著的超额收益. 在对因子重要度的研究中,研究发现我国股票市场异象具有时变特征. 此外,研究从宏观经济状态和经济政策两个角度分析表明,基于自编码的投资模型的有效性与宏观经济息息相关,它能够在市场泡沫成分较大和投机气氛较浓的情况下成功对冲市场风险,且能捕捉到由财政政策和货币政策所导致的市场环境的变化.

关键词: 自编码器; 机器学习; 金融大数据; Logistic 回归; 横截面收益预测
中图分类号: F832.5 **文献标识码**: A **文章编号**: 1007-9807(2024)05-0082-16

0 引 言

随着金融科技的飞速发展,以大数据、人工智能、云计算、区块链为代表的金融科技新技术不断涌现,新的金融科技革命正在悄然发生. 中国人民银行印发的《金融科技发展规划(2022 年—2025 年)》指出"要以深化金融数据要素应用为基础,以支撑金融供给侧结构性改革为目标,以加快推进金融机构数字化转型为主线:力争到 2025 年,我国金融科技整体水平与核心竞争力实现跨越式提升,数据要素价值充分释放、数字化转型高质量推进". 股票市场作为我国金融市场的重要组成部分,拥有上市公司海量的交易数据与各种类型的财务数据,因此具备率先应用以机器学习人工智能技术为代表的金融科技的天然优势. 如何使用金融大数据提升金融市场的定价效率,通过定价效率的提升使得资金高效流入最具有投资价值的公司,进而助力企业供给侧改革与数字化转型的高质量发展,具有较高的社会经济价值.

专注于解释股票市场资产收益的资产定价理论与实践在近几十年来已取得了长足的发展,国内外学者到目前为止已发现了四百多个能预测股票收益的异象指标[1],即能够产生超额收益的独特因子. 但这些定价指标中可能包含着同一类定价信息,并且单个指标的预测能力往往在被发现后,其预测能力随时间发展而逐渐减弱. McLean 和 Pontiff[2] 发现,被论文所揭示的因子往往在发表后其超额收益相对之前下降了三分之二. 因此,只依靠发现单个定价指标来进行资产价格预测和解释难以满足目前研究和实践需要.

另一方面,依赖传统方法进行多指标金融大数据的资产定价研究也遇到了不小的困难. 第一,由于变量过多,若用传统的回归方法对数百个股

① 收稿日期: 2020-08-31; 修订日期: 2022-12-17.
基金项目: 国家自然科学基金资助项目(71872195; 72072193; 72003062; 72342019); 国家社会科学基金资助重大项目(22&ZD063).
通讯作者: 朱 琳(1995—), 女, 山东青岛人, 博士, 讲师. Email: zhulin@gdufe.edu.cn

票因子同时进行回归，将不可避免地带来“维数灾难”. 同时，传统的组合分析（portfolio analysis）和 Fama-MacBeth 回归并未考虑因子间的交互作用，具有一定的局限性. 第二，无论是传统资本资产定价模型（Capital Asset Pricing Model，CAPM）还是以 Fama-French 模型为代表的多因子模型，由于受制于市场时变性和单个因子信息获取完整度等问题，使用唯一的模型往往难以有效度量资产风险. Cochrane[3] 在美国金融学年会的主旨演讲中指出，在存在大量噪声和高度相关的预测因子的金融市场，研究人员应该使用新颖的计量经济学方法综合地进行信息提取.

为了解决上述问题所面临的挑战，学者们开始探讨使用机器学习的方法对因子信息进行降维并提取有价值的信息. Kozak 等[4] 和 Kelly 等[5] 分别运用主成分分析（Principal Components Analysis，PCA）和增量主成分分析（instrumented PCA，IPCA）方法，发现基于只依赖于少数主成分的模型就可以较好地解释和预测截面收益. 然而，PCA 仅是通过对指标进行线性变化来获得静态线性因子，但越来越多的研究指出资产收益和指标之间还存在非线性关系[6,7]. Pohl 等[8] 等表明，当存在非线性关系时，线性近似会导致预测估计的巨大误差. 因此，在提取多指标的公共信息时，需要应用非线性模型进行降维变换，从而更好地提取金融数据中的非线性信息. 因此，本研究尝试采用基于自编码（Autoencoder，AE）机器学习方法从公司特征中提取非线性信息，作为无监督学习的代表，Gu 等[9] 表明自编码器是比线性模型更一般的允许非线性降维变换的神经网络模型. 研究发现，自编码机器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，另一方面，可以有效提取数据中的非线性信息从而带来资产定价研究的非线性范式革命.

在梳理和参考了国内外实证资产定价的文献基础上，研究构建了针对中国股票市场的金融因子大数据库，该数据库包含 89 个企业特征因子，可以从不同种类的公司特征中探索金融市场包含的多维信息. 此外，研究率先针对中国股票市场使用自编码模型提取来自公司特征的信息，将人工智能技术、金融大数据与股票市场资产定价相结合，并探究机器学习对股票市场收益的预测和解释能力. 由于使用收益数据作为目标进行训练和预测，所以该研究在自编码器网络的基础上后续增加逻辑回归（logistic regression），综合使用机器学习方法自动地寻找数据中的复杂结构和模式达到辅助预测的目的. 与传统模型相比，自编码模型在中国股票市场中通过非线性变换能从众多微弱的单个预测信号中提取出预测能力强且有意义的复合信号.

研究结果显示，基于自编码机器学习的资产定价模型能够在中国市场上较好地解释和预测收益. 第一，自编码模型收益预测能力强，依据模型预测值构造股票投资策略，发现多空对冲组合的月平均收益率、夏普比率最高分别为 1.82%（t = 3.54）和 1.05，投资组合在经典的多因子模型（三因子模型、五因子模型）中超额收益仍均在 1% 的显著性水平上显著. 第二，模型预测结果的差异显著受到训练集大小的设定及自编码结构差异的影响. 由于金融数据中的信噪比较低，训练集不是越大越好，12 个月的训练集大小和中国经济周期的变动规律大致相符，预测结果最好；另外，使用一个或两个隐藏层的自编码器模型得到的投资组合月平均收益率为佳，这表明在使用神经网络法进行金融大数据预测时，并不是模型越复杂越好，同时验证了 Gu 等[10] 的“浅层学习的表现优于深层学习”发现. 第三，为了探究异象因子的重要度和时变性，本研究逐月滚动地输出每个模型网络中输出层的各因子权重值，观察到股市异象的确具有时变特征，不同时期模型对因子赋予的权重不同. 具体而言，2006 年之前，我国市场的动量类因子权重占主要地位，其次是包含市场因子在内的交易摩擦类因子，而 2007 年以后权重以盈利类因子为主；直到 2012 年，转变为估值成长类因子权重增加，17 年以后无形资产类因子的权重在逐步加大.

为了打破机器学习的“黑箱”问题，本研究将自编码因子的投资组合收益与市场状况和宏观经济政策相联系，探究自编码方法提取的复合信号在收益可预测性方面的经济解释. 通过选取一系列具有代表性的宏观指标进行分析发现，基于自编码机器学习的投资模型可以成功捕捉有效的宏观经济信息，能够在市场泡沫成分较大和投机气氛较浓的情况下成功对冲市场风险，对财政政策

和货币政策有很好的反应,能敏锐察觉到经济政策变化所带来的市场经济环境变化.

本研究有一定的现实意义和理论贡献. 美国市场的大量研究尝试使用包括神经网络机器学习算法在内的大数据方法构建预测和定价模型,来解决因子大量涌现而给传统研究方法带来的挑战. 本研究紧跟前沿,但不同于 Light 等[11]采用“偏最小二乘”(Partial Least Squares ,PLS)来检验公司特征对预期收益的预测能力,本研究着眼于公司特征的非线性关系的提取. 不同于 Dong 等[12]和 Gu 等[10]在美国市场对比了的一系列机器学习方法的预测效果,本研究着重于对自编码特征的信息提取与收益来源的深度解释. 由于金融市场的特征存在一定的关联性和信噪比低等问题,本研究发现所使用的方法其实践投资效果可以有效超越传统线性模型,但同时也发现在金融预测中浅层算法要优于深层复杂算法.

随着中国金融市场重要性的不断提升,本研究进一步推进了中国股票市场科学定价问题的前沿探索. 研究着力于应对大数据时代背景下机器学习对经济学的研究范式和研究方法带来的机遇[13-17];也是对未来中国金融市场中机器学习的应用前景进行实践探索. 近期不少研究也关注了机器学习方法在中国金融市场信息提取的应用[18, 19],本研究从动态非线性的视角丰富了中国市场股票截面收益影响因素的研究,丰富了资产定价实证研究的机器学习工具,通过动态非线性机器学习算法探索横截面股票收益的决定因素,并尝试探究自编码算法在收益可预测性方面的经济解释以及与宏观经济状态之间的关联.

1 理论模型与研究方法

1.1 标准自编码器

自编码器模型是目前机器学习领域主要的非线性动态降维模型,其功能是通过将输入信息作为学习目标进行表征学习(representation learning). 中间层(压缩层)的神经元对输入层的变量进行编码,得到从输入层到编码层的映射,然后再将其解压并映射到输出层(解码),结构如图 1 所示.

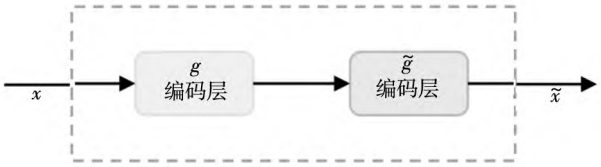

图 1 自编码模型示意图

Fig. 1 Standard autoencoder model

标准自编码器可以用数学表达式表示. N 是输入和输出层的神经元的个数,x_k 是输入层 N 个神经元的输入,以及输入向量为 $\boldsymbol{x} = (x_1, x_2, \cdots, x_N)'$. $\tilde{x}_k$ 是输出层 N 个神经元的输出,以及输出向量为 $\tilde{\boldsymbol{x}} = (\tilde{x}_1, \tilde{x}_2, \cdots, \tilde{x}_N)'$.

在压缩层的向量为 $\boldsymbol{z} = (z_1, z_2, \cdots, z_K)'$,$K$ 是在压缩层的神经元的个数,前一层的输入向量经过一个非线性的激活函数 $g(\cdot)$ 传递到输出层. 初始化网络,输入层使用公司特征的横截面数据为原始输入 $\boldsymbol{x} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_N)'$.

自编码器的编码过程为

$$\boldsymbol{z} = g(\boldsymbol{b} + \boldsymbol{W}\boldsymbol{x}) \tag{1}$$

其中 $\boldsymbol{W}$ 是 $\boldsymbol{K} \times \boldsymbol{N}$ 维的权重矩阵,$\boldsymbol{b}$ 是一个被称作是偏差参数(bias parameters)的 $K \times 1$ 维向量. 研究使用线性整流函数(Rectified Linear Unit, ReLU)作为输入层的非线性激活函数,表达式为 $g(m) = \max(m, 0)$.

自编码器的解码过程为

$$\tilde{\boldsymbol{x}} = \tilde{g}(\boldsymbol{W}'\boldsymbol{z} + \boldsymbol{b}') \tag{2}$$

其中解码层的激活函数 $\tilde{g}(\cdot)$ 使用 Sigmoid 函数,表达式为 $\tilde{g}(m) = \dfrac{1}{1 + e^{-m}}$. Sigmoid 函数是一种常见的 S 型函数,可以把变量映射到 [0, 1] 之间.

对于自编码器,本研究重点关注输入层到压缩层的编码过程. 一般而言,压缩层神经元的个数 K 小于输入变量神经元的个数 N,表示网络自主学习到输入样本的特征,试图以更小的维度去描述原始数据而尽量不损失数据信息,以进行压缩表示. 因为模型训练的需要,需将原始数据作为目标来训练整个网络,而等训练结束后,则将输出层去掉,只关心从输入层 X 到压缩(隐藏)层的变

换. 本研究将压缩层的信息作为后续机器学习算法的输入，后续机器学习模型采用的是 Logistic 回归模型.

选择 Logistic 回归模型作为研究工具主要基于以下几点考量. 首先，Logistic 回归作为一种判别式模型，特别适用于处理二分类问题，包含了多种模型正则化手段，有效解决了使用朴素贝叶斯模型时可能遇到的特征独立性要求问题. 其次，相较于决策树和支持向量机(SVM)，Logistic 回归模型提供更佳的可解释性，并能输出概率解释，对于预测金融市场中股票涨跌的概率问题具有天然的优势. 除此之外，模型内存资源占用小，运算速度较快，计算量仅和特征的数目相关，适合于处理大规模数据集.

1.2　堆叠自编码器(stacked autoencoder)

如果只关心从输入层 X 到压缩(隐藏)层的变换，前文的标准自编码器模型可表示单层神经网络结构，在此基础上，Larochelle 等[20]提出堆叠自编码器(Stacked Autoencoder, SAE)，将隐藏层再作为原始信息，训练一个新的自编码器得到新的特征表达. 第 1 个 AE 训练完成后，将其输出作为第 2 个 AE 的输入，以此类推，l 个 AE 按顺序训练，结构如图 2 所示.

L 个隐藏层的自编码器可以用递推公式表示. 设 $K^{(l)}$ 是第 l 层神经元的数量，$l = 1,2,\cdots,L$，$x_k^{(l)}$ 是 l 层 k 个神经元的输出，则这一层所有的输出变量组成的矩阵为 $\boldsymbol{x}^{(l)} = (\boldsymbol{x}_1^{(l)}, \boldsymbol{x}_2^{(l)}, \cdots, \boldsymbol{x}_{K^{(l)}}^{(l)})$. 在每一个隐藏层中，前一层的输入在传入下一层前经过非线性激活函数 $g(\cdot)$. 第 $l(l>0)$ 层的神经网络的递归输出公式为

$$\boldsymbol{x}^{(l)} = g(\boldsymbol{b}^{(l-1)} + \boldsymbol{W}^{(l-1)}\boldsymbol{x}^{(l-1)}) \tag{3}$$

其中 $\boldsymbol{W}^{(l-1)}$ 是 $K^{(l)} \times K^{(l-1)}$ 维的权重参数矩阵，$\boldsymbol{b}^{(l-1)}$ 是 $K^{(l)} \times 1$ 维向量，在每一隐藏层 l 中参数的数量为 $K^{(l)}(1 + K^{(l-1)})$.

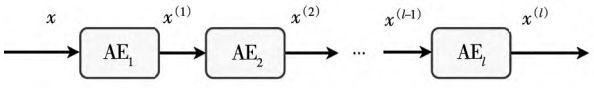

图 2　堆叠自编码模型示意图

Fig. 2 Stacked autoencoder model

非线性 x 函数的递归过程用公式表达为

$$x_{i,t-1}^{(0)} = x_{i,t-1} \tag{4}$$

$$x_{i,t-1}^{(l)} = \boldsymbol{g}(\boldsymbol{b}^{(l-1)} + \boldsymbol{W}^{(l-1)} x_{i,t-1}^{(l-1)}) \quad l = 1,2,\cdots,L \tag{5}$$

$$\boldsymbol{z}_{i,t-1} = \boldsymbol{b}^{(L)} + \boldsymbol{W}^{(L)} x_{i,t-1}^{(L)} \tag{6}$$

其中横截面上的公司股票用 i 标记，$i = 1,2,\cdots,N$，月份用 t 标记，$t = 1,2,\cdots,T$，这里假设股票为平衡面板数据，而关于数据缺失值的处理将在 2.1 节讨论. 式(4)表达的是初始化网络模型的第 i 只股票在 $t-1$ 期的公司特征数据 $x_{i,t-1}$，式(5)的方程描述了特征数据通过隐藏层神经元传播时的非线性(和交互的)转换，式(6)描述了一组 K 维因子如何从终端输出层产生.

为了防止过拟合，最常见的机器学习方法是在目标函数上附加一个惩罚项，以得到更简洁的设定. 这种“正则化”方法通过减少模型的样本内的过度拟合问题，以期提高模型的样本外预测效果[10]. 惩罚项的设定减少模型的拟合噪声，防止过拟合.

设 $L(\boldsymbol{W};\cdot)$ 为优化自编码器的损失函数，用 W 来概括因子网络模型式(4)到式(6)到中的权重参数. 将评估目标定义为

$$L(\boldsymbol{W};\cdot) = \frac{1}{NT}\sum_{t=1}^{T}\sum_{i=1}^{N} \|x_{i,t} - \tilde{x}_{i,t}\|^2 + \varphi(\boldsymbol{W};\cdot) \tag{7}$$

其中 $\phi(\boldsymbol{W};\cdot)$ 是对模型进行正则化的惩罚函数，称为正则化项. 有很多惩罚函数 $\phi(\cdot)$ 可供选择，它们虽然使用了不同的正则化项，但最终都实现了约束参数从而防止过拟合的效果. 这里用 LASSO 或“l_1”惩罚，采取的形式为

$$\phi(\boldsymbol{W};\lambda) = \lambda\sum_j |W_j| \tag{8}$$

这里的参数 λ 就是正则化系数，可以取大于 0 的任意值. λ 的值越大，对模型中参数的惩罚力度越大，就会有更多的参数被训练为 0.

1.3　Logistic 回归

若在回归问题中因变量 Y 为二元定性变量，用哑变量的方法对因变量进行编码. 假设第 i 家公司在 t 时刻的收益为 $r_{i,t}$，在 t 时刻对市场上所有的 N 只股票的收益率进行排序，每个月收益排名前 30%标记为相对高收益类别 1，认为其是在 t 时刻市场上表现优异(上涨)的股票；后 30%的股票标记为相对低收益类别 0，认为其是在 t 时刻市

场上表现较差的股票;剩余样本不参与训练. 具体如下所示

$$y_{i,t}=\begin{cases}1, & \text{相对高收益}\\ 0, & \text{相对低收益}\end{cases} \tag{9}$$

Logistic 回归对 y 属于某一类的概率建模而不直接对因变量 y 建模. 对本研究而言,Logistic 回归建立股票上涨概率模型

$$p_{i,t}\equiv P(y_{i,t}=1\mid z_{i,t-1})=\frac{1}{1+\mathrm{e}^{-z'(x_{i,t-1})\beta_t}} \tag{10}$$

其中 $\boldsymbol{z}(x_{i,t-1})$ 是以 $x_{i,t-1}$ 作为输入变量经过自编码器网络式(4)~式(6)的非线性变换得到的压缩层的变量,$x_{i,t-1}$ 是第 i 家公司在 $t-1$ 期的原始横截面公司特征. $p_{i,t}$ 为在输入 $\boldsymbol{z}_{i,t-1}$ 的条件下 Logistic 回归属于类别 1 的概率.

为评估分类器的概率输出,在概率估计上定义对数似然损失(Log-likelihood Loss)函数. 对数损失通过惩罚错误的分类,实现对分类器准确度(accuracy)的量化. 对数损失函数的计算公式如下

$$L(\boldsymbol{p};\cdot)=\frac{1}{NT}\sum_{t=1}^{T}\sum_{i=1}^{N}(y_{i,t}\log p_{i,t}+(1-y_{i,t})\log(1-p_{i,t})) \tag{11}$$

其中 $y_{i,t}$ 是所属的每个类别的概率值,$p_{i,t}$ 为预测输入变量 $z_{i,t-1}$ 属于类别 1 的概率. 依据自编码器和 Logistic 回归的损失函数的收敛情况调整模型迭代次数.

2 数据

2.1 公司特征因子指标

关于金融大数据的构建,McLean 和 Pontiff[2] 研究了上百个能显著预测股票收益的指标,Hou 等[1]将目前可以解释股票市场横截面收益的因子分为惯性、价值与增长、投资、盈利、无形资产与交易摩擦六类. 研究全面梳理了实证资产定价的近百篇文献,综合考虑数据的可获得性,通过国泰安数据库中财务报表数据和股票交易数据,整理计算出与公司股票预期收益相关的 89 个公司特征因子指标②,其中包含 20 个估值类指标、12 个投资类指标、23 个盈利类指标、7 个趋势类指标、15 个市场类指标以及 12 个无形资产类指标.

2.2 数据来源与处理

本研究的研究对象是中国股票市场的上市公司,主要包括在上海证券交易所、深圳证券交易所中上市的所有的 A 股公司以及创业板公司. 研究选取 2000 年 1 月至 2018 年 12 月中国 A 股市场所有上市公司为研究样本,数据为月度频率. 其中,上市公司的股票收益以及财务状况的数据主要从国泰安数据库(China Stock Market and Accounting Research, CSMAR)获得. 例如,月度频率的股票收益数据,上市公司年度、季度财务报表、资产负债表、现金流量表等数据. 此外,诸如 CAPM 模型、Fama-French 因子模型数据(三因子模型、五因子模型)、停复牌公司信息数据、特别处理(Special Treatment, ST)与特别转让(Particular Transfer, PT)等股票数据也来自国泰安数据库. 在自编码因子的经济学分析中,研究从 Wind 数据库、锐思数据库中获得宏观经济变量的数据.

为了保证实证分析结果的严谨性和准确性,研究根据资产定价领域的主流研究范式对所选取的数据进行样本筛选和修正. 首先,根据目前 Carpenter 等[21]研究中国股票市场的经验,将财务数据的样本时间区间起始点选在 2000 年. 这是由于 2000 年后我国的市场经济的发展程度更加完善,股票市场机制更加成熟,上市公司的财务披露质量和监管力度显著提高. 其次,数据库中的原始数据存在着一定比例的缺失值,研究采用了 Gu 等[10]和李斌等[18]对指标缺失值的处理方法,过程分为两步:1)若某只股票在第 t 月收益率数据缺失(通常由股票连续停牌造成),则剔除该股票在月份 t 上的所有数据;2)若某只股票的因子值缺失,则以该公司所属行业的横截面均值进行填充.

在剔除 ST、PT 股票,并处理完缺失值后,2000 年1 月至 2018 年 12 月的有效样本总计 398 919 条. 总体来看,月度样本量随年份呈上升趋势,由 2000 年 1 月的 882 条有效样本增加至 2018 年12 月的 3 406 条,如图 3 所示.

② 由于篇幅限制,公司特征因子指标名称、定义及文献出处,留存备索.

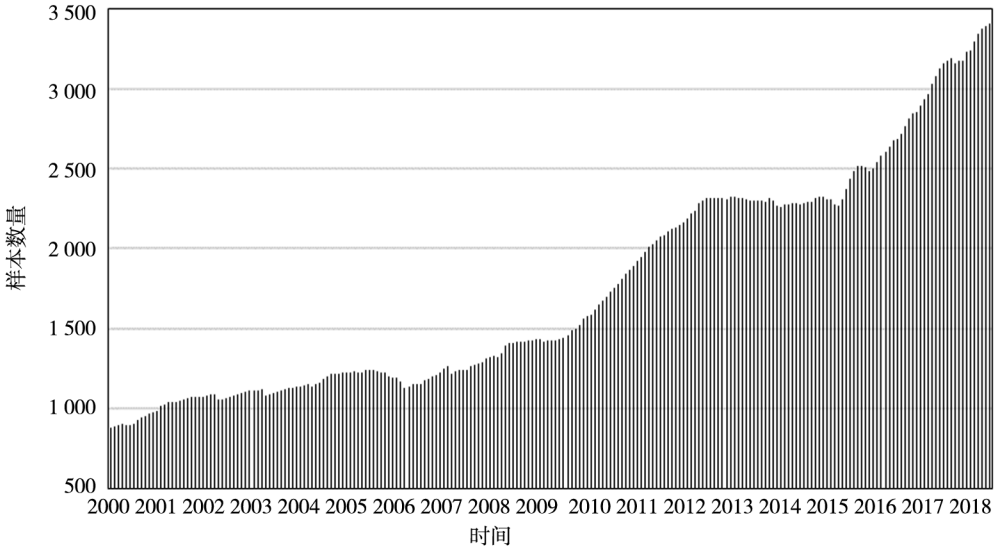

图 3　2000 年 1 月至 2018 年 12 月月度有效样本数量

Fig. 3 Effective samples number from Jan. 2000 to Dec. 2018

公司特征的原始数据值在数量级及分布上存在显著差异，可能导致预测偏差：量级较大的特征在预测时占据主导地位；数量级的差异会引起部分机器学习算法迭代收敛速度减慢. 由此，研究需将公司特征数据进行归一化处理，映射到 0 到 1 之间. 标准化方式为

$$x_{norm} = \frac{x - \min_x}{\max_x - \min_x} \tag{12}$$

其中 $\min_x$ 和 $\max_x$ 分别为各月度公司特征数据的最小值和最大值.

鉴于机器学习算法依赖一段样本期作为训练集，研究设定不同长度的滚动窗口，采用逐月滚动的训练方法. 此外，为了进一步验证模型的准确性和稳健性，本研究还选取 10% 的样本作为交叉验证(cross-validation)样本. 研究最终进行横截面收益预测的样本区间为 2007 年 7 月至 2018 年 12 月.

3　实证结果与分析

本节将从横截面股票收益预测的角度出发，采用投资组合分析法，根据自编码因子构建投资组合并研究其绩效，比较与分析不同自编码器结构是否有明显的差别，并进一步考察各公司特征因子在预测模型中的重要程度和时变特征.

3.1　自编码器网络结构

实证研究中共测试 9 种不同网络结构的自编码器，所有结构编号和简写如表 1 所示. 其中，AE(32)、AE(64)和 AE(128)为标准自编码器，AE(32,16)、AE(64,32) 和 AE(128,64)为二层自编码结构，AE(32,16,8)、AE(64,32,16)和 AE(128,64,32)为三层自编码结构，均属于堆叠式自编码器.

3.2　投资组合的构建

投资组合分析法最早由 Eugene F. Fama 提出，如今已成为实证资产定价领域被广泛使用的分析方法，相较于线性回归，该方法较少地受到线性假设的限制，对于个别的极端值也不敏感，能够有效检验不同指标对横截面股票收益的影响.

本文使用的投资组合法是在月度频率上根据自编码器对每家公司的股票预测相对高收益(上涨)的概率构建投资组合：在每月的第一个交易日，分别根据上个月公司最新的特征变量进行预测，按照预测的数值大小从小到大对样本中全体公司进行排序，等分成 10 组投资组合，并用编号对每组进行标记；第一组表示预测上涨概率最小的前 10% 的股票，标记为“Low”组，第十组表示预测上涨概率最高的前 10% 的股票，标记为“High”组，其他组按概率预测的高低分别记为 2 组至 9 组；同时，做多“High”组，卖空“Low”组，并将该投资组合记为多空对冲组合，即“High-Low”组；这些投资组合的持有周期为一个月，到了下个月，重复上述过程构建和持有新的投资组合. 在本文

中,根据上市公司上个月的市值大小分配每一只股票在投资组合中所占权重,即构造市值加权的投资组合. 一般而言,如果最终“High-Low”(下文简写为“H-L”)组能够产生显著收益,则证明使用不同的方法进行排序的预测收益指标能够显著预测横截面股票收益的差别.

表 1 自编码器网络结构

Table 1 Network structure of autoencoder model

编号	简写	输入层节点数	压缩层节点数	输出层节点数				
①	AE(32)	89	32	89				
②	AE(64)	89	64	89				
③	AE(128)	89	128	89				
		输入层节点数	隐藏层节点数	压缩层节点数	隐藏层节点数	输出层节点数		
④	AE(32,16)	89	32	16	32	89		
⑤	AE(64,32)	89	64	32	64	89		
⑥	AE(128,64)	89	128	64	128	89		
		输入层节点数	隐藏层节点数	隐藏层节点数	压缩层节点数	隐藏层节点数	隐藏层节点数	输出层节点数
⑦	AE(32,16,8)	89	32	16	8	16	32	89
⑧	AE(64,32,16)	89	64	32	16	32	64	89
⑨	AE(128,64,32)	89	128	64	32	64	128	89

3.3 横截面收益预测结果

本研究使用投资组合法检验自编码因子对横截面股票收益的预测能力. 表 2 列出了依据自编码器得到的股票的相对高收益(上涨)概率的大小构建的各投资组合的月度原始平均收益、夏普比率(年化后)、$CAPM-\alpha$、$FF3-\alpha$、$FF5-\alpha$ 以及各组 t 值(方括号中). 其中,月度平均原始收益(Ret)和月度的超额收益(α)在表格中以百分比形式汇报.

表 2 自编码因子的横截面股票收益

Table 2 Autoencoder predictors sorted portfolios risk-adjusted results

投资组合	*Ret*	夏普比率	$CAPM-\alpha$	$FF3-\alpha$	$FF5-\alpha$
Low	-0.28[-0.34]	-0.21	-0.70[-2.28]	-0.80[-2.73]	-0.79[-2.63]
2	0.08[0.10]	-0.09	-0.33[-1.37]	-0.60[-2.73]	-0.61[-2.62]
3	0.38[0.49]	0.03	-0.02[-0.09]	-0.26[-1.13]	-0.31[-1.29]
4	0.59[0.75]	0.10	0.19[0.74]	-0.16[-0.74]	-0.21[-0.93]
5	0.69[0.88]	0.14	0.28[1.16]	-0.10[-0.52]	-0.15[-0.75]
6	0.86[1.11]	0.21	0.46[1.74]	-0.05[-0.26]	-0.13[-0.64]
7	0.77[0.97]	0.17	0.36[1.32]	-0.17[-0.92]	-0.29[-1.43]
8	0.89[1.11]	0.21	0.48[1.45]	-0.16[-0.74]	-0.25[-1.09]
9	0.97[1.25]	0.25	0.57[1.77]	-0.08[-0.37]	-0.23[-1.05]
High	1.54[1.97]	0.46	1.14[2.99]	0.47[1.64]	0.32[1.09]
H-L	1.82[3.54]	1.05	1.84[3.64]	1.26[3.63]	1.12[2.29]

注:该表汇报自编码因子构建的投资组合的月度平均原始收益(*Ret*,%),夏普比率(年化后),投资组合的超额收益($CAPM-\alpha$、$FF3-\alpha$ 和 $FF5-\alpha$,%)及其 t 统计量([·]). “H-L”指多空对冲投资组合. 其中,该结果的自编码模型为 AE(128, 64),滚动窗口为 12 个月.

分析表 2 的结果可知,各投资组合的月平均收益随着自编码机器学习对上市公司预期的增加而增加. 具体而言,月度平均收益从预期收益最低组的 -0.28%(t = -0.34)增加至预期收益最高组的 1.54%(t = 1.97),预期收益最高与最低组的月度平均收益之差为 1.82%, t 值为 3.54,在 1%水平上显著. 这表明通过投资按预期收益大小所构建的多空对冲组合,平均每年

将获 21.84%的收益. 同时,夏普比率也存在着从低到高的单调递增规律,多空对冲组合的夏普比率更是高达 1.05. 此外,用 CAPM 模型、Fama-French 三因子模型、Fama-French 五因子模型衡量的超额收益都随着各组预期收益的增加而增加. 多空对冲组合的 CAPM $-\alpha$、FF3 $-\alpha$、FF5 $-\alpha$ 分别为 1.84%($t=3.64$)、1.26%($t=3.63$)、1.12%($t=2.29$),均在 1%水平以上显著,自编码方法能够显著预测中国股票市场横截面收益.

表 3　不同编码器结构的横截面股票收益

Table 3 Performance of the autoencoder predictors constructed with different structures

自编码器编号		①	②	③	④	⑤
Ret	Low	-0.27 [-0.32]	-0.26 [-0.31]	-0.31 [-0.37]	-0.21 [-0.24]	-0.17 [-0.20]
	High	1.41 [1.81]	1.42 [1.84]	1.34 [1.71]	1.33 [1.74]	1.29 [1.68]
	H-L	1.67 [3.36]	1.68 [3.36]	1.65 [3.32]	1.54 [3.06]	1.46 [2.87]
夏普比率	H-L	0.99	0.99	0.98	0.90	0.85
FF5-α	H-L	1.04 [2.16]	1.01 [2.13]	0.94 [2.00]	0.91 [1.90]	0.77 [1.60]
自编码器编号		⑥	⑦	⑧	⑨	
Ret	Low	-0.28 [-0.34]	-0.09 [-0.11]	-0.05 [-0.06]	-0.22 [-0.26]	
	High	1.54 [1.97]	1.22 [1.61]	1.41 [1.80]	1.36 [1.75]	
	H-L	1.82 [3.54]	1.32 [2.69]	1.47 [2.96]	1.58 [3.09]	
夏普比率	H-L	1.05	0.79	0.87	0.91	
FF5-α	H-L	1.12 [2.29]	0.79 [1.78]	0.74 [1.58]	0.86 [1.80]	

注：该表汇报不同自编码器结构得到的因子值构建的投资组合的月度平均原始收益（*Ret*，%），夏普比率（年化后），投资组合的超额收益（*FF5*$-\alpha$，%）及其 t 统计量（[·]）. 为简明起见,自编码结构用表 1 中的编号代替,所有模型的滚动窗口为 12 个月.

接下来探讨基于 12 个月的滚动窗口,不同自编码器结构的横截面股票收益情况. 根据表 3 的结果显示,包含一个隐藏层和两个隐藏层的自编码器能够获得较好的横截面收益预测结果,其多空对冲组合月收益最高达 1.82%($t=3.54$),夏普比率为 1.05,多种模型衡量的超额收益也在 5%水平显著,特别的,标准(一层)自编码器随着压缩层节点的变化横截面收益结果显著且保持相对稳定,分别为 1.67%($t=3.36$)、1.68%($t=3.36$)和 1.65%($t=3.32$). 但是,随着神经网络法中隐藏层不断增加,模型复杂度逐渐上升,不仅没有提高模型的预测能力,反而多空对冲组合的月平均收益率有所下降. 其中,三层自编码器月平均收益率分别为 1.32%($t=2.69$)、1.47%($t=2.96$)和 1.58%($t=3.09$). 这说明在使用神经网络法进行金融大数据预测时,并不是模型越复杂越好. 与深度学习在如计算机视觉领域的蓬勃发展不同,这些领域往往拥有更海量的数据集. 但金融市场中包含着各类公司信息和交易信号,这些信息中难免有许多与未来资产价格无关的噪音. 当隐含层层数越多时,输入变量所包含的噪音一定程度上会影响最终输出结果的预测精度. 未来的研究若能在神经网络中加入经济约束(依据不同的定价模型),更精确地衡量输入层变量之间的相互关系,或许能改善深层学习的预测结果.

鉴于现有文献有关机器学习算法中训练集大

小的设定不尽相同,因此本文设定不同的滚动窗口,分别为 6 个月、12 个月、24 个月、36 个月、48 个月和 60 个月,探究不同训练集大小对横截面收益的影响,结果如表 4 所示.

结合表 2 发现,以 12 个月为训练样本得到的多空组合的月平均收益最高达到 1.82%,除此之外的其他结构也在 1.32%到 1.68%之间的范围,明显高于其他的滚动窗口. 总体而言,12 个月为训练样本的结果的多空组合月平均收益最好,24 个月次之,而更短或更长时间的训练样本都没有达到更好的效果,这与李斌等[18]在运用机器学习算法时最优的滚动窗口长度一致. 首先,金融市场中包含着各类公司信息和交易信号,这些信息中难免有许多与未来资产价格无关的噪音. 训练集的长度设置过大,反而会引入更多的噪音,从而影响预测结果. 其次,宏观经济趋势影响股票市场价格,经济周期决定股票市场的价格区间. 股票行情的强弱与宏观经济周期的强弱成正相关关系,所以最佳的训练集大小应与经济周期相一致.

表 4 不同滚动窗口的自编码因子的多空组合在横截面的股票收益

Table 4 Performance of the autoencoder predictors constructed with different rolling windows

自编码器编号	①	②	③	④	⑤	⑥	⑦	⑧	⑨
6 个月滚动窗口									
Ret	1.13 [2.23]	1.08 [2.10]	1.08 [2.08]	0.93 [1.82]	1.11 [2.11]	1.12 [2.09]	0.60 [1.10]	1.03 [1.95]	1.18 [2.20]
夏普比率	0.66	0.62	0.61	0.54	0.62	0.62	0.33	0.57	0.65
24 个月滚动窗口									
Ret	1.25 [2.74]	1.30 [2.86]	1.43 [3.11]	1.25 [2.81]	1.35 [2.93]	1.33 [2.86]	1.47 [3.49]	1.29 [2.89]	1.30 [2.80]
夏普比率	0.81	0.84	0.92	0.83	0.86	0.84	1.03	0.85	0.83
36 个月滚动窗口									
Ret	0.96 [2.09]	1.06 [2.34]	1.11 [2.45]	1.15 [2.63]	1.08 [2.28]	1.02 [2.20]	0.91 [2.14]	0.95 [2.10]	0.94 [2.09]
夏普比率	0.62	0.69	0.72	0.78	0.67	0.65	0.63	0.62	0.62
48 个月滚动窗口									
Ret	0.95 [2.08]	1.07 [2.41]	1.12 [2.48]	0.86 [1.89]	1.00 [2.09]	1.16 [2.52]	1.13 [2.47]	0.66 [1.37]	1.26 [2.75]
夏普比率	0.61	0.71	0.73	0.56	0.62	0.74	0.73	0.41	0.81
60 个月滚动窗口									
Ret	0.98 [2.04]	1.14 [2.49]	1.17 [2.53]	0.86 [1.77]	1.01 [2.12]	1.13 [2.41]	0.67 [1.45]	0.80 [1.64]	1.01 [2.16]
夏普比率	0.60	0.74	0.74	0.52	0.63	0.71	0.43	0.48	0.64

注:该表汇报不同滚动窗口得到的自编码因子值构建的多空对冲组合(H-L)的月度平均原始收益(*Ret*,%),夏普比率(年化后)及其 t 统计量([·]). 为简明起见,自编码结构用表 1 中的编号代替;12 个月滚动窗口的横截面收益如表 3 所示.

3.4 控制其他公司特征的双变量排序

本小节进一步考察基于自编码因子和其他指标的双变量排序的投资组合的表现,选取了在股票市场上具有代表性的因子指标,分别为公司规模(SIZE)、账面市值比(BM)、净运营资本(NOA)和换手率(TURN).

参考 Hou 等[1]的分析方法,每月对自编码预测值和其他因子值进行独立的双变量排序,并在月度频率上进行再平衡. 具体而言,每月初按照自编码预测值从小到大对样本中全体公司进行排序,等分成 10 组投资组合;按照其他因子指标的数值从小到大进行排序,按 30 百分位点和 70 百分位点将全体公司分为三组投资组合,持有这些组合一个月并计算市值加权的投资组合月度收益.

对于账面市值比(BM),低于 30% 分位数点的公司称为“成长型(growth)”公司,30%到 70%之间的公司称为“中性”(neutral),高于 70%的公司称为“价值型(value)”公司. 根据公司规模的三个分位点可分为“微型(micro)”公司、“小型(small)”公司和“大型(large)”公司. 如此之外,还选取了衡量投资水平的净运营资本(NOA)和衡量流动性水平的换手率(TURN)两个指标,使研究更加全面.

表 5 展示了自编码预测值和其他因子指标的双变量排序(double sorts)的投资组合表现. 结果显示,自编码预测值的溢价在不同水平上有不同的表现,具体来看,在大型公司中多空投资组合的月平均收益为 1.26%(t = 2.35),高于微型公司和小型公司的收益率,分别是 0.91%和 0.82%;在价值型公司中多空投资组合的月平均收益为 2.26%(t = 3.21),高于微型公司和小型公司的 1.23%和 1.75%,总体而言,基于自编码的多空组合收益主要集中在大型公司和价值型公司. 另外,投资组合随投资水平的提高单调增加,范围从低净运营资本的 1.25%(t = 2.32)到高净运营资本 2.17%(t = 4.01);对于流动性水平而言,换手率高的公司的投资组合月平均收益为 1.84%(t = 3.12)高于换手率低的公司. 换言之,自编码多空投资组合在高投资水平和高流动性的公司能获得更高的回报.

表 5　基于自编码因子和其他公司特征双变量排序(double sorts)的投资组合表现

Table 5 Returns of the portfolios double sorted on predictors andother firm characteristics

投资组合	公司规模(SIZE)			账面市值比(BM)		
	微型	小型	大型	成长	中	价值
Low	1.43	0.42	-0.36	-0.04	-0.38	-0.47
	[1.35]	[0.43]	[-0.42]	[-0.05]	[-0.43]	[-0.49]
High	2.35	1.24	0.91	1.19	1.37	1.79
	[2.48]	[1.45]	[1.26]	[1.40]	[1.64]	[2.16]
H-L	0.91	0.82	1.26	1.23	1.75	2.26
	[1.96]	[1.91]	[2.35]	[2.37]	[3.41]	[3.21]
投资组合	净运营资本(NOA)			换手率(TURN)		
	低	中	高	低	中	高
Low	-0.08	-0.24	-0.44	0.37	-0.21	-0.58
	[-0.09]	[-0.28]	[-0.53]	[0.42]	[-0.25]	[-0.67]
High	1.17	1.39	1.73	1.59	1.55	1.28
	[1.51]	[1.79]	[1.97]	[2.10]	[1.75]	[1.45]
H-L	1.25	1.63	2.17	1.22	1.76	1.84
	[2.32]	[2.91]	[4.01]	[1.84]	[3.44]	[3.12]

3.5　因子的重要度与时变性

国内外学者到目前为止已发现了四百多个能预测股票收益的市场异象指标,然而 Harvey 等[22]发现随着市场结构性变化和流动性的增加,单个异象的预测能力随时间逐渐减弱; McLean 和 Pontiff[2]分析了这些指标在被学者提出前后的区别,发现不少收益预测指标在得到了学界以及市场的广泛关注后往往其预测能力会消失. 在中国市场,姜富伟等[23]基于大数据和机器学习的智能动态 CAPM 模型,检验了时变系统性风险对我国股市收益解释能力; 尹力博等[24]研究表明中国股票市场经历了一次风格转换,以账面市值比异象、市盈率异象为代表的价值型异象正在逐渐减弱甚至消失. 因此,在时间维度上验证因子有效性就尤为重要.

本小节进一步考察各类指标在不同模型中的重要程度及其时变性. 目前学术界在考察预测因子的重要程度的研究中普遍参考 Gu 等[10],以单独剔除各因子后模型样本外 R^2 下降程度代表该因子的重要度,比较因子的相对重要性. 但这种方法得到的因子的重要性是对于因子在样本期内的平均而言,并不能得到重要程度随时间的变化.

本研究逐月滚动地输出标准自编码器的输出层各因子权重,不仅能得到因子的重要程度,更能得到其每个月权重的变化,进一步验证因子的时变性特征. 图 4 展示了各时间点自编码器输出层中重要度排序最高的十个公司特征因子③. 具体的做法是,逐月输出自编码器输出层的因子权重值,取绝对值后在每 6 个月进行加总平均,最后按顺序展示重要度排在前十的因子.

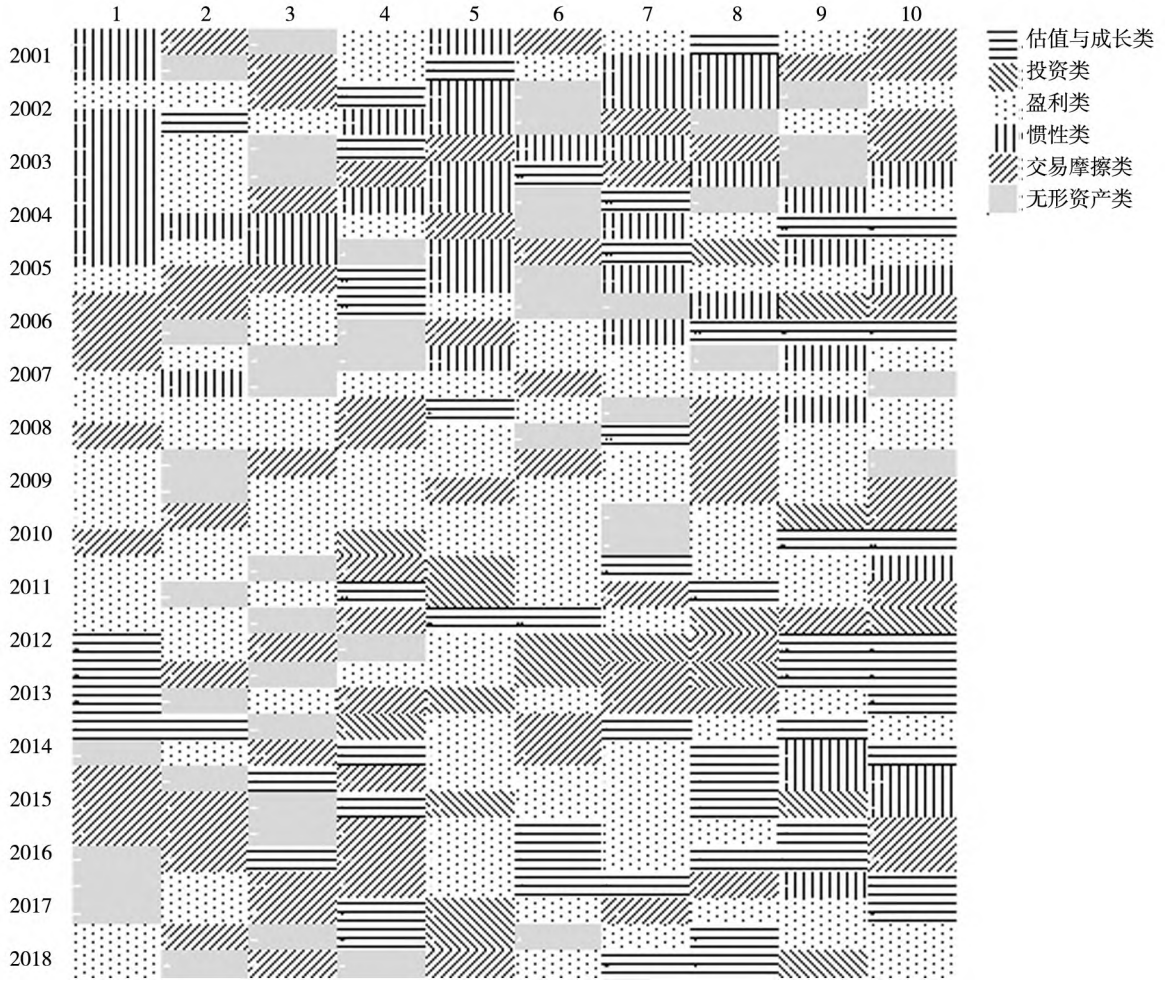

图 4 自编码器输出层企业特征因子重要度

Fig. 4 Weight of firm factors in output layer of autoencoders

图 4 结果表明,因子的重要度随时间有规律地变化,与宏观经济有密切的关联性. 2006 年之前,动量类因子权重占主要地位,其次是包含市场因子在内的交易摩擦类因子,而 2007 年以后权重以盈利类因子为主;直到 2012 年,转变为估值成长类因子权重增加,而 2017 年以后无形资产类因子的权重在逐步加大. 因子预测的重要性变化也反映了不同阶段能为投资者带来高收益的上市公司特征是时变的:从早期能引起投资者关注的动量类特征到目前反映公司软实力的无形资产特征,随着我国经济发展的阶段不同,金融市场优质公司的定义也在不断变化. 引起第一个时间节点变化的原因可能是 2005 年股改前后的市场表现区别较大. 股权分置改革主要通过完善股票市场的运行机制对股票市场的有效性产生积极影响. 从整体趋势上来说,股改后中国股票市场趋于有效和成熟.

关于动量效应的研究,国内外学者主要从传

③ 由于篇幅限制,图 4 的结果并没有展示因子的名称,仅以因子类别作为区分,更详尽的结果留存备索。

统风险收益模型和行为金融，但至今没有形成统一的解释. 对于中国股票市场而言，短期投资者占比高与中国市场短期的动量效应和中长期的反转效应有密切关系. 中国股市早期受政策影响较大，同时宏观经济波动也导致了所有股票同时涨跌的情况出现，单只股票的走势往往与市场联动性较强，这和表中股市早期动量类因子的权重较大的结果相吻合. 而对于盈利类因子，企业的盈利能力是决定公司股价走向的根本因素，因为资产的盈利能力决定着其未来产生现金流的贴现值从而影响资产的定价，因此在 2007 年以后盈利类因子在权重中占主导地位.

值得注意的是，自编码器对无形资产因子赋予一定权重，尤其是 2017 年以后无形资产因子（研发、品牌价值、人力资本等）的权重在加大，这与最近学术界对无形资产因子的研究相符. Hall[25] 指出，无形资产是股票市场价值重要组成部分，它的重要性正在显著增加. 投资者只有正确衡量企业的无形资产才能更准确地进行投资. 同时，随着我国企业逐步重视对无形资产的投入，以及有关无形资产的测算方法的完善，无形资产对于企业估值的作用越发重要.

4　自编码因子的经济学理论

Cochrane[3] 指出，传统的回归方法不足以处理大量的预测变量，机器学习方法提供了适应高维预测集和灵活的函数形式，然而，包括自编码器模型在内的机器学习方法因为其复杂性和多维性导致决策过程的不透明，通常被称为“黑箱”. 本节尝试打破机器学习的“黑箱”问题，探究自编码方法提取的复合信号在收益可预测性方面的经济解释，将收益与市场状态的变化联系起来，探讨依据自编码构造的投资组合收益具体受到哪些宏观经济因素的影响.

本研究在探讨预测收益是否随时间变化时采用以下月度时间序列回归方程进行分析，

$$RET_t^{AE} = \alpha + \varphi\, Dummy_{t-1}^{high} + \boldsymbol{\beta}' F_t + e_t \quad (13)$$

其中 RET_t^{AE} 是指在 t 月依据自编码模型构造的多空对冲（High-Low）投资组合收益，$\boldsymbol{\beta}$ 是 5 × 1 维列向量，F_t 是指 *Fama-French* 因子模型中的五个常见风险因子，包括市场因子（*MKT*）、规模因子（*SMB*）、账面市值因子（*HML*）、盈利能力因子（*RMW*）和投资因子（*CMA*）. $Dummy_{t-1}^{high}$ 是一个虚拟变量（dummy variable），对于每一个宏观经济变量，在整个样本期内按照二分法分成两个时期，变量大于样本中值取 1，否则取 0. 如果某个经济变量能对多空对冲组合收益产生影响，则一般有较显著的估计系数 φ. 相较于月度重构的投资组合，许多与宏观经济相关的变量是季度甚至更低频的数据，因此本研究参考 *Cooper* 等[26] 和 *Han* 等[27] 探讨资产收益与经济周期关联时采用的方法进行分析.

4.1　与宏观经济状况的联系

首先，本研究探讨自编码因子与宏观经济之间的关系. 股市涨跌变化与国民经济整体环境及结构变动密切相关. 企业的投资价值会反映在宏观经济的总体中，宏观经济变量会影响整个股票市场的投资价值. 参考 Welch 和 Goyal[28] 的研究，选择国内生产总值（GDP）、采购经理人指数（PMI）、市场市盈率衡量宏观经济运行状况.

GDP 是我国新国民经济核算体系中的核心指标，它反映了一国（或地区）的经济实力和市场规模. 我国制造业采购经理人指数（PMI）的指标体系包括库存、就业、产量、进口和订单量等，相较于 GDP 等宏观经济变量是季度甚至更低频的数据，我国 PMI 指标体系每月月初更新，更具时效性，能够及时预测经济增长. 市盈率是价格—盈利乘数，该指标不仅综合考虑了金融资产的价格和收益，同时也是衡量股市热度和健康状况的重要指标.

本节从宏观经济状况角度探讨前文提取的自编码因子是否受这些宏观经济因素的影响. 表 6 采取了式（13）中的回归方程，展示了根据前文的自编码因子所构造的投资组合收益与宏观经济状态之间的关系.

表 6 自编码因子与宏观经济状况④

Table 6 Autoencoder factor and macroeconomic situation

变量	α	φ
国内生产总值(GDP)	0.28 [0.51]	1.72 [2.16]
采购经理人指数(PMI)	0.33 [0.45]	1.52 [1.63]
PMI:生产	0.41 [0.86]	1.38 [1.67]
PMI:新订单	0.28 [0.37]	1.62 [1.77]
市场市盈率	0.44 [0.72]	1.37 [1.65]

注:该表汇报模型(13)中常数项(α)和虚拟变量 $Dummy_{t-1}^{high}$ 的回归系数 φ 及其 Newey-West 调整后的 t 值([·])。

从上表的结果可以看出,当 GDP 和 PMI 处于更高水平的时期,即我国所处的经济形势和经济环境较好,投资组合收益能增加 1.72%(t = 2.16)和 1.52%(t = 1.63),尤其在反应制造业企业的生产情况和内需情况的生产指数和新订单指数处于高水平时,对组合收益的影响更加显著,分别增加 1.38%(t = 1.67)和 1.62%(t = 1.77),说明自编码因子成功捕捉到了有效的宏观经济信息,以此获得超额收益。除此之外,市盈率作为衡量市场泡沫大小和市场危险性的指标,在较高的市盈率水平不仅能对投资组合产生正的显著影响,即组合收益能增加 1.37%(t = 1.65),还能获得超额收益 0.44%,这说明运用自编码方法能够在市场泡沫成分较大和投机气氛较浓的情况下成功对冲市场风险,获得超额收益。

4.2 与宏观经济政策的联系

本节从宏观经济政策的角度探讨前文提取的自编码因子是否受财政政策和货币政策、社会融资规模以及市场波动性等因素的影响。

一般而言,财政政策和货币政策是国家宏观调控经济运行的两大基本政策。财政政策和货币政策协调配合下对我国宏观经济波动产生的冲击效应非常显著。Drechsler 等[29]指出准备金的流动性溢价在未预期的货币政策对股票市场收益率的影响关系中具有中介效应。财政政策冲击能更多地解释我国产出的中期波动,而货币政策冲击则能更好地分析短期经济周期波动。此外,朱小能和周磊[30]和贾盾等[31]都指出预期外的货币政策或经济政策的不确定性都会对资本市场收益产生影响;邓可斌等[32]证明宽松货币、财政政策能有效地降低我国股市的系统性风险。

社会融资规模是指一定时期内实体经济从金融体系获得的资金总额,是全面反映政府和社会金融对实体经济的资金支持以及金融与经济关系的总量指标。盛松成和谢洁玉[33]证实社会融资规模(增量与存量)是能够反映金融体系与实体经济的联系、综合反映全社会融资总量和结构的良好指标。而市场波动性则体现了可能由宏观经济政策所造成的不确定性,在市场动荡时期,由于流动性枯竭,股票市场的收益可能会降低。

表 7 结果显示,我国货币、财政等经济政策能够显著影响自编码因子所构造的投资组合收益。税收对国民收入是一种收缩性力量,税收收入的增加,说明政府抑制总需求从而减少国民收入,国家财政投资规模缩小。当税收收入处于高位时,组合投资收益会降低 -1.29%(t = -1.71)。当人民币存款准备金率提高,表明央行实行紧缩的货币政策,收紧流动性,收缩货币量和信贷量,对投资组合收益产生显著负的影响,导致投资收益降低 -1.58%(t = -2.11)。较高的社会融资规模平均提升 1.57%(t = 1.84)的月平均收益,社会融资规模出于高水平,说明各类融资支持实体经济的状况较好,央行货币政策有利于实体经济的发展,对投资组合收益率的影响有促进作用。除此之外,当市场波动性增强时,此时的市场不确定性增加,投资组合月平均收益会降低 -2.12%(t = -3.00),但同样能获得超额收益,由此看出根据自编码因子构造的多空对冲组合并没有对冲市场波动性的风险,这可能和中国股票市场散户居多的特殊性有关。

综上所述,根据自编码因子构造的投资组合能够对财政政策和货币政策有很好的反应,成功捕捉到经济政策变化所带来的经济环境的变化,平均收益与经济政策息息相关。由此也说明,宏观经济政策的制定对股市的影响显著,如果政策制

④ 限于篇幅,正文未汇报 MKT、SMB、HML、RMW 和 CMA 的回归系数及其 t 值,完整的结果留存备索。

定者能够合理准确适时的颁布经济政策，有助于信息到资产价格的有效传导，势必会对资本市场乃至整个社会经济繁荣发展发挥重大作用.

表 7　自编码因子与宏观经济政策

Table 7 Autoencoder factor and macroeconomic policy

变量	α	φ
税收收入	1.77 [3.87]	-1.29 [-1.71]
人民币存款准备金率	1.89 [3.91]	-1.58 [-2.11]
社会融资规模	0.28 [0.44]	1.57 [1.84]
股票市场波动性	2.17 [4.73]	-2.12 [-3.00]

注：该表汇报模型(13)中常数项(α)和虚拟变量 $Dummy_{t-1}^{high}$ 的回归系数 φ 及其 Newey-West 调整后的 t 值([·]).

5　结束语

本研究将自编码机器学习方法应用到中国市场的横截面实证资产定价研究中. 在梳理近百篇与定价指标相关文献的基础上，本研究使用中国股票市场上市公司交易与财务数据构建了包含 89 个公司特征因子的金融大数据集，采用自编码模型预测上市公司股票的上涨概率，并进行投资组合分析以评估投资策略的绩效. 此外，本研究从宏观经济状态和经济政策关联两个角度，对自编码因子的收益预测来源进行了经济理论分析.

研究发现，自编码方法能够在金融大数据中提取到具有预测能力的复合信号，依据自编码因子构建的投资组合在横截面上获得显著的超额收益. 对因子重要度的研究中表明，股票市场中的公司特征因子具有时变特征，不同时期算法对不同因子赋予的权重不同. 在探究自编码因子在收益可预测性的经济解释方面，基于自编码的投资模型可以捕捉有效的宏观经济信息，能够在市场泡沫成分较大和投机气氛较浓的情况下成功对冲市场风险，对财政政策和货币政策有很好的反应，能敏锐察觉到经济政策变化所带来的市场经济环境的变化.

本研究进一步丰富了基于中国市场的机器学习资产定价研究，在对机器学习收益预测效果进行经济机制探索的同时，进一步研究了中国股票市场在人工智能技术应用层面的独特特征. 随着我国资本市场的制度不断健全、发展日趋完善，未来利用人工智能方法进行投资策略的构建和交易会逐渐成为主流. 通过对金融大数据的使用和分析，监管者也能更好地了解市场动态. 研究者、实践者因此也更需要跟上时代发展的潮流，从大数据分析和人工智能视角来理解市场的运行规律，探讨中国股票市场的时变性、独特性. 机器学习方法已被证明能够有效提取金融数据种的有价值预测信息，未来的研究还需要通过向计算机再学习的过程，更进一步联系经济理论，从而更深刻地揭示资本市场的变动规律，服务于我国金融市场的效率提升和与数字化转型的高质量发展.

参 考 文 献：

[1] Hou K, Xue C, Zhang L. Replicating anomalies[J]. Review of Financial Studies, 2020, 33(5): 2019 - 2133.
[2] McLean R D, Pontiff J. Does academic research destroy stock return predictability? [J]. Journal of Finance, 2016, 71(1): 5 - 32.
[3] Cochrane J H. Presidential address: Discount rates[J]. Journal of Finance, 2011, 66(4): 1047 - 1108.
[4] Kozak S, Nagel S, Santosh S. Shrinking the cross-section[J]. Journal of Financial Economics, 2020, 135(2): 271 - 292.
[5] Kelly B T, Pruitt S, Su Y A. Characteristics are covariances: A unified model of risk and return[J]. Journal of Financial Economics, 2019, 134(3): 501 - 524.
[6] Campbell J Y, Cochrane J H. By force of habit: A consumption-based explanation of aggregate stock market behavior[J]. Journal of Political Economy, 1999, 107(2): 205 - 251.
[7] Bansal R, Yaron A. Risks for the long run: A potential resolution of asset pricing puzzles[J]. Journal of Finance, 2004, 59(4): 1481 - 1509.
[8] Pohl W, Schmedders K, Wilms O. Higher order effects in asset pricing models with long-run risks[J]. Journal of Finance, 2018, 73(3): 1061 - 1111.

[9]Gu S, Kelly B, Xiu D. Autoencoder asset pricing models[J]. Journal of Econometrics, 2021, 222(1): 429 – 450.
[10]Gu S, Kelly B, Xiu D. Empirical asset pricing via machine learning[J]. Review of Financial Studies, 2020, 33(5): 2223 – 2273.
[11]Light N, Maslov D, Rytchkov O. Aggregation of information about the cross section of stock returns: A latent variable approach[J]. Review of Financial Studies, 2017, 30(4): 1339 – 1381.
[12]Dong X, Li Y, Rapach D E, et al. Anomalies and the expected market return[J]. The Journal of Finance, 2022, 77(1): 639 – 681.
[13]洪永淼, 汪寿阳. 大数据如何改变经济学研究范式? [J]. 管理世界, 2021, 37(10): 40 – 55 + 72 + 56.
Hong Yongmiao, Wang Shouyang. How is big data changing economic research paradigms? [J]. Journal of Management World, 2021, 37(10): 40 – 55 + 72 + 56. (in Chinese)
[14]崔毅安, 熊 熊, 韦立坚, 等. 金融科技视角下的计算实验金融建模[J]. 系统工程理论与实践, 2020, 40(2): 373 – 381.
Cui Yi'an, Xiong Xiong, Wei Lijian, et al. Agent-based modeling from the perspectives of FinTech[J]. Systems Engineering: Theory & Practic, 2020, 40(2): 373 – 381. (in Chinese)
[15]马 慧, 靳庆鲁, 王 欣. 大数据与会计功能——新的分析框架和思考方向[J]. 管理科学学报, 2021, 24(9): 1 – 17.
Ma Hui, Jin Qinglu, Wang Xin. Big data and accounting function: A new analysis direction and framework[J]. Journal of Management Sciences in China, 2021, 24(9): 1 – 17. (in Chinese)
[16]伍之昂, 赵新元, 黄 宾, 等. 基于文献计量的大数据管理决策研究热点分析[J]. 管理科学学报, 2021, 24(6): 117 – 126.
Wu Zhi'ang, Zhao Xinyuan, Huang Bin, et al. Status and trends in big-data-driven managerial decision-making on bibliometric[J]. Journal of Management Sciences in China, 2021, 24(6): 117 – 126. (in Chinese)
[17]姜富伟, 林奕皓, 马 甜. "去刚兑"背景下的企业债券违约风险: 机器学习预警和经济机制探究[J]. 金融研究, 2023, (10): 85 – 103.
Jiang Fuwei, Lin Yihao, Ma Tian. Research on an early warning model of corporate bond default and its economic mechanism based on machine learning[J]. Journal of Financial Research, 2023, (10): 85 – 103. (in Chinese)
[18]李 斌, 邵新月, 李玥阳. 机器学习驱动的基本面量化投资研究[J]. 中国工业经济, 2019, (8): 61 – 79.
Li Bin, Shao Xinyue, Li Yueyang. Research on machine learning driven quantamental investing[J]. China Industrial Economics, 2019, (8): 61 – 79. (in Chinese)
[19]马 甜, 姜富伟, 唐国豪. 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[J]. 经济学(季刊), 2022, 22(3): 819 – 842.
Ma Tian, Jiang Fuwei, Tang Guohao. Deep learning and factor investing in Chinese stock market: Based on generative adversarial networks[J]. China Economic Quarterly, 2022, 22(3): 819 – 842. (in Chinese)
[20]Larochelle H, Bengio Y, Louradour J, et al. Exploring strategies for training deep neural networks[J]. Journal of Machine Learning Research, 2009, 10: 1 – 40.
[21]Carpenter J N, Lu F Z, Whitelaw R F. The real value of China's stock market[J]. Journal of Financial Economics, 2021, 139(3): 679 – 696.
[22]Harvey C R, Liu Y, Zhu H Q. ... and the cross-section of expected returns[J]. Review of Financial Studies, 2016, 29(1): 5 – 68.
[23]姜富伟, 马 甜, 张宏伟. 高风险低收益? 基于机器学习的动态 CAPM 模型解释[J]. 管理科学学报, 2021, 24(1): 109 – 126.
Jiang Fuwei, Ma Tian, Zhang Hongwei. High risk low return? Explanation from machine learning based conditional CAPM model[J]. Journal of Management Sciences in China, 2021, 24(1): 109 – 126. (in Chinese)
[24]尹力博, 韦 亚, 韩复龄. 中国股市异象的时变特征及影响因素研究[J]. 中国管理科学, 2019, 27(8): 14 – 25.
Yin Libo, Wei Ya, Han Fuling. Study on charcteristics and influence factors of time-varying anomalies in China's stock market[J]. Chinese Journal of Management Science, 2019, 27(8): 14 – 25. (in Chinese)
[25]Hall R E. The stock market and capital accumulation[J]. American Economic Review, 2001, 91(5): 1185 – 1202.
[26]Cooper M J, Gutierrez R C, Hameed A. Market states and momentum[J]. Journal of Finance, 2004, 59(3): 1345 – 1365.

[27]Han Y F, Zhou G F, Zhu Y Z. A trend factor: Any economic gains from using information over investment horizons? [J]. Journal of Financial Economics, 2016, 122(2): 352 - 375.

[28]Welch I, Goyal A. A comprehensive look at the empirical performance of equity premium prediction[J]. Review of Financial Studies, 2008, 21(4): 1455 - 1508.

[29]Drechsler I, Savov A, Schnabl P. A model of monetary policy and risk premia[J]. Journal of Finance, 2018, 73(1): 317 - 373.

[30]朱小能, 周　磊. 未预期货币政策与股票市场——基于媒体数据的实证研究[J]. 金融研究, 2018, (1): 102 - 120.

Zhu Xiaoneng, Zhou Lei. Monetary policy surprises and stock returns: Evidence from media forecasts[J]. Journal of Financial Research, 2018, (1): 102 - 120. (in Chinese)

[31]贾　盾, 孙　溪, 郭　瑞. 货币政策公告、政策不确定性及股票市场的预公告溢价效应——来自中国市场的证据[J]. 金融研究, 2019, (7): 76 - 95.

Jia Dun, Sun Xi, Guo Rui. Monetary announcements, policy uncertainty, and equity premium in China[J]. Journal of Financial Research, 2019, (7): 76 - 95. (in Chinese)

[32]邓可斌, 关子桓, 陈　彬. 宏观经济政策与股市系统性风险——宏微观混合 β 估测方法的提出与检验[J]. 经济研究, 2018, 53(8): 68 - 83.

Deng Kebin, Guan Zihuan, Chen Bin. Macroeconomic policies and systemic risk in China's stock market: An approach based on integrated hybrid betas[J]. Journal of Financial Research, 2018, 53(8): 68 - 83. (in Chinese)

[33]盛松成, 谢洁玉. 社会融资规模与货币政策传导——基于信用渠道的中介目标选择[J]. 中国社会科学, 2016, (12): 60 - 82 + 205 - 206.

Sheng Songcheng, Xie Jieyu. The scale of social financing and the transmission of monetary policy: Based on choice of intermediate credit channel targets[J]. Social Science in China, 2016, (12): 60 - 82 + 205 - 206. (in Chinese)

Asset pricing research based on autoencoder machine learning: A financial big data analysis perspective of the Chinese stock market

TANG Guo-hao[1], *ZHU Lin*[2*], *LIAO Cun-fei*[3], *JIANG Fu-wei*[4, 5]

1. College of Finance and Statistics, Hunan University, Changsha 410006, China;
2. School of Finance, Guangdong University of Finance & Economics, Guangzhou 510320, China;
3. School of Economics and Management, Nanjing University of Science and Technology, Nanjing 210094, China;
4. School of Economics, Xiamen University, Xiamen 361005, China;
5. The Wang Yanan Institute for Studies in Economics, Xiamen University, Xiamen 361005, China

Abstract: This paper predicts stock returns in the Chinese market by using an improved autoencoder machine learning approach and financial big data encompassing approximately one hundred firm characteristics. The findings demonstrate that the autoencoder factor can extract predictors from a large amount of information containing company characteristics to forecast returns and can achieve significant excess returns in the cross sections. Additionally, the analysis on the significance of factors reveals that the anomalies are time-varying in the Chinese stock market. Additionally, the predictive efficiency of the autoencoder method correlates with macroeconomic conditions and economic policies. The autoencoder-based long-short portfolios can effectively mitigate market risk especially during substantial market bubbles and heightened speculative periods, demonstrating well resilience to the shifts of fiscal and monetary policy-induced economic conditions.

Key words: autoencoder; machine learning; financial big data; Logistic regression; the cross-section of stock returns