基于自编码机器学习的金融大数据资产定价研究

唐国豪 朱琳 廖存非 姜富伟
（湖南大学金融与统计学院，长沙 410006
中央财经大学金融学院，北京 100089）

摘要：本文在中国股票市场上，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。本文的研究发现，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，本文发现我国股票市场异象具有时变特征。此外，本文从宏观经济状态和经济政策两个角度进行分析，研究结果表明，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：自编码机器学习方法 金融大数据 Logistic 回归 横截面收益预测
中图分类号：F832.5 **文献识别码**：A

Abstract: Using the improved autoencoder machine learning approach, and financial big data which contains nearly a hundred firm characteristics, we predict and explain stock returns in Chinese market, and analyses the source of the predictability. This paper finds that the autoencoder-based characteristic extracts the composite signal of effective information from the selected characteristics, and obtains significant cross-section returns. The results on the importance of factors show that stock market anomalies are time-varying, and the weights of these anomalies vary in different periods. In addition, from the two perspectives of macroeconomics and economic policy, this paper makes economic analyses of the relationship between the autoencoder-based characteristic's return predictability and macroeconomics policy. The results show that the autoencoder-based predictor captures effective macroeconomic information and hedges market risks during huge market bubbles and excess speculation periods. The autoencoder-based characteristic responds well to fiscal and monetary policies, and captures the changes in economic environment which are influenced by economic policies.

Key words: Autoencoder, financial big data, Logistic regression, the cross-section of stock returns

作者信息：
唐国豪，湖南大学金融与统计学院副教授，湖南省长沙市岳麓区石佳冲路 109 号，邮编：410006，电子邮箱：ghtang@hnu.edu.cn，电话：13401190049；
朱琳，通讯作者，湖南大学金融与统计学院博士研究生，湖南省长沙市岳麓区石佳冲路 109 号，邮编：410006，电子邮箱：zhulin1@hnu.edu.cn，电话：13678876615；
廖存非，湖南大学金融与统计学院博士研究生，湖南省长沙市岳麓区石佳冲路 109 号，邮编：410006，电子邮箱：cfliao@hnu.edu.cn，电话：15526400291；
姜富伟，中央财经大学金融学院教授，北京市海淀区学院南路 39 号，邮编：100089，电子邮箱：fwjiang@qq.com，电话： 18511086494。

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。2019 年央行正式发布了《金融科技发展规划》，指出“到 2021 年要实现金融与科技的深度融合，要使金融科技成为推动金融转型升级的新引擎，金融服务实体经济的新途径，防范化解金融风险的新利器”。在此背景下，使用以机器学习为代表的人工智能技术和金融大数据提升金融市场的定价效率，具有不容忽视的社会经济价值。

股票市场的实证资产定价模型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市场异象指标（Hou 等[1]），即能够提供超额收益的独特因子。Cochrane[2]在美国金融学年会的主旨演讲中，将这些海量定价指标称为因子动物园（Factor Zoo）。这些定价指标中可能饱含着同一类定价信息，并且单个指标的预测能力随时间逐渐减弱。McLean 和 Pontiff[3]发现，揭示因子的以往论文在发表后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际需要。

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，Bali 等[4]指出，由于变量过多，若用传统的回归方法对数百个股票因子同时进行回归，将会带来“维数灾难”。同时，传统的组合分析（Portfolio Analysis）和 Fama-MacBeth 回归并未综合考虑因子间的交互作用；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，其主要目的在于找到不同风险溢价下对应的资产收益，而受制于市场时变性和信息获取难度等问题，资产风险往往难以用唯一的模型有效度量。Cochrane[2]认为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，使用综合方法进行信息提取。

为了解决上述问题所面临的挑战，本文尝试采用基于自编码机器学习方法从公司特征中提取共同信息，构建定价和预测模型来解释截面收益。对于提取因子中的共同信息，Kozak 等[5]和 Kelly 等[6]分别运用主成分分析（Principal Components Analysis，PCA）和增量主成分分析（Instrumented Principal Components Analysis, IPCA）方法来进行提取，发现基于只依赖于较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 只能通过对指标进行线性变化来获得静态线性因子，但越来越多的研究发现资产收益和指标之间还存在非线性关系（Campbell 和 Cochrane[7] ，Bansal 和 Yaron[8]以及 He 和 Krishnamurthy[9]等）。Pohl[10]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，我们需要构造非线性动态模型来进行非线性降维变换，从而更好地提取金融数据中的非线性信息。

本文使用的自动编码器是一类在无监督学习中使用的一种人工神经网络，也是目前机器学习领域主要的降维模型。其功能是通过将输入信息作为学习目标，对输入信息进行表征学习（Representation learning）。Gu 等[11]表明，自编码器可以被认为是一个与主成分分析（PCA）相对应的非线性神经网络。自编码器模型是比线性因子更一般的模型，因为其除线性变换外，它允许非线性降维变换，已在除金融以外的多个领域被证明是有价值的。例如，Hinton 和 Salakhutdinov[12]表明，在图像识别方面，深度非线性自编码器的性能要优于线性自动编码器。

在梳理和参考了国内外实证资产定价的文献基础上，本文利用金融数据，独立构建了

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，可以从不同视角探索金融市场包含的多维信息。此外，本文率先针对中国股票市场使用自编码（Autoencoder, AE）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。使用收益数据进行训练和预测，从本质上是一个监督学习的过程。所以本文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达到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，新构建的自编码模型在中国股票市场进行非线性信息特征的提取和分析，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。

本文的研究结果显示，基于自编码机器学习的资产定价模型能够较好地在中国市场解释和预测收益。第一，新模型收益预测能力强，按照预测值对横截面股票构建投资组合，本文发现新模型的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，新模型在经典的多因子模型（三因子、五因子模型）中超额收益仍均在显著性水平 5%上显著。第二，模型中训练集大小的设定不同和自编码结构的不同，预测结果也大有不同。为了降低信噪比，训练集不是越大越好，12 月的训练集大小和中国经济周期的变动规律相符，因此预测结果最好；另外，一个或两个隐藏层的自编码器所购在的投资组合的月平均收益率为佳，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，也验证了 Gu 等[11]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，本文采用逐月滚动构造模型，输出每个模型网络中输出层的各因子权重值。观察到的结果是，股市异象的确具有时变特征，不同时期算法对不同因子赋予的权重不同。具体来说，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，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，本文还探究了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，将根据自编码因子构造投资组合所获得的超额收益与市场状况和宏观经济政策相联系。我们选取一系列具有代表性的宏观指标进行分析并发现，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的变化。

本文的研究有一定的现实意义和理论贡献。近年来，金融学开始探讨机器学习对现有研究范式的价值和意义，美国市场的大量研究尝试使用包括神经网络机器学习方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。例如 Light 等[13]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；Gu 等[11]对比了包括主成分分析、偏最小二乘法、LASSO、随机森林和神经网络在内的各类机器学习模型在美国股票市场的应用，结果发现非线性的机器学习模型的预测表现可以有效地超越传统线性模型，同时浅层网络模型要优于深度学习模型；Feng 等[14]在进行股价预测时引入无套利约束并结合了深度学习网络模型。

不少研究也着眼于中国市场。洪永淼和汪寿阳[15]认为大数据和机器学习对经济学的研究范式和研究方法到来了机遇和挑战；苏治等[16]对近年来在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机器学习的应用给出了评述；陈卫华等[17]基于深度学习和股票论坛数据对股市波动率进行了预测，发现深度学习模型显著提升了预测

精度；李斌等[18]基于 96 个异象因子，运用 12 种机器学习算法，构建股票收益预测模型及投资组合，发现非线性方法能获得较高的投资绩效。本文紧跟前沿，实证上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，还尝试去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之间的关联。

本文余下部分的组织架构如下。第二部分主要介绍了理论模型和研究方法，包括自编码器和逻辑回归。第三部分主要描述了数据来源以及公司特征指标的定义。第四部分是主要的实证结果与分析。第五部分是自编码因子的经济学分析。第六部分是结论与建议。

1 理论模型与研究方法

1.1 标准自编码器

自编码器是一种特殊的神经网络，这个系统的输出变量试图去近似输入变量。输入变量通过中间层（压缩层）的神经元，对输入层的变量进行编码，得到输入到编码的映射，然后再将其解压缩并映射到输出层（解码），结构如下图所示。

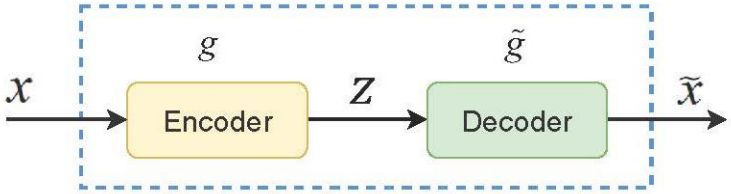

图 1 自编码模型示意图

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 $x=(x_1,x_2,\cdots,x_N)^T$ 。$\tilde{x}_k$ 是输出层 N 个神经元的输出，以及输出向量为 $\tilde{x}=(\tilde{x}_1,\tilde{x}_2,\cdots,\tilde{x}_N)^T$ 。

在压缩层的向量为 $z=(z_1,z_2,\cdots，z_K)^T$ ，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 $x=(x_1,x_1,\cdots，x_N)^T$ 。

自编码器的编码过程为：

$$z=g(b+Wx), \tag{1}$$

其中，W 是 $K\times N$ 维的权重矩阵，b 是一个被称作是偏差参数的 $K\times 1$ 维向量。本文使用线性整流函数（Rectified linear unit, ReLU）作为输入层的非线性激活函数。

自编码器的解码过程为：

$$\tilde{x}=\tilde{g}(W^Tz+b^T), \tag{2}$$

其中，解码层的激活函数 $\tilde{g}(\cdot)$ 可以和编码层的不同，这里使用 Sigmoid 函数。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。激活函数的选取与后续使用 logistic 回归模型有关。

对于自编码器，本文重点关注压缩层的编码，或者说是输入层到压缩层的映射。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。实际上，自主学习原始数据的特征表达是神经网络和深度学习的核心之一。本文就是将压缩层的信息作为后续机器学习算法的输入，采用 Logistic 回归模型。

后续采用 Logistic 回归模型主要有以下考虑。首先，Logistic 回归属于判别式模型，是一个二分类回归，同时伴有很多模型正则化的方法，不必像在用朴素贝叶斯那样担心特征是否相关。其次，与决策树、SVM 相比，模型的可解释性好，可以得到一个概率解释，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。

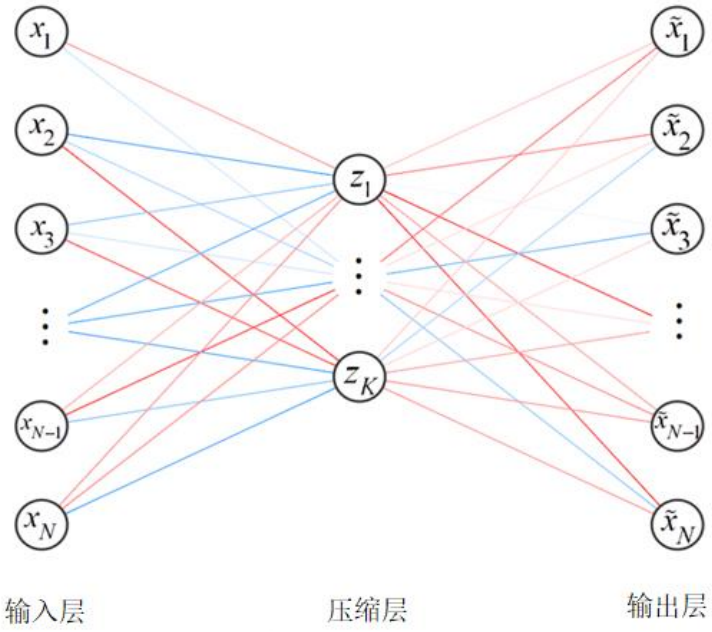

图 2 标准自编码器网络结构示意图

1.2 堆叠自编码器（Stacked Autoencoder）

前面的标准自编码器模型表示为三层的神经网络，因为训练的需要，本文将原始数据作为监督误差来训练整个网络。等训练结束后，将输出层去掉，只关心从输入层 X 到压缩（隐藏）层 h 的变换。若将 h 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，这就是 Bengio 等[19]提出的堆叠自编码器（Stacked Autoencoder, SAE）。需要注意的是，整个训练不是一蹴而就的，而是需要逐层进行。换句话说，n 个 AE 按顺序训练，第 1 个 AE 训练完成后，将其编码器的输出作为第 2 个 AE 的输入，以此类推。最后得到的特征作为分类器的输入，完成最终的分类训练，结构如下图所示。

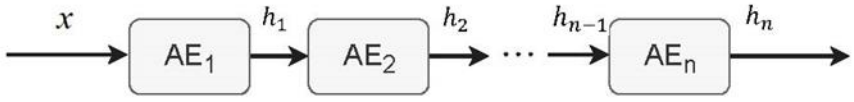

图 3 堆叠自编码模型示意图

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 $x^{(l)}=(x_1^{(l)},x_1^{(l)},\cdots,x_{K^{(l)}}^{(l)})^T$。在每一个隐藏层中，前一层的额输入在传入下一层前经过非线性激活函数 $g(\cdot)$。递归输出公式为

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

在每一隐藏层 l 中参数的数量为 $K^{(l)}(1+K^{(l-1)})$，$b^{(l-1)}$ 是一个 $K^{(l)}\times 1$ 维的列向量。自编码

器的最终的输出为

$$G(x,b,W)=b^{(L)}+W^{(L)}x^{(L)}, \tag{4}$$

利用与输入相同的维数来逼近 x 。

非线性 x 函数的递推公式为：

$$x_{i,t-1}^{(0)}=x_{i,t-1}, \tag{5}$$

$$x_{i,t-1}^{(l)}=g(b^{(l-1)}+W^{(l-1)}x_{i,t-1}^{(l-1)}),l=1,2,\cdots,L, \tag{6}$$

$$z_{i,t-1}=b^{(L)}+W^{L}x_{i,t-1}^{(L)}. \tag{7}$$

(5)式表达的是初始化网络模型的为第 i 家公司在 t-1 期的公司特征数据 $x_{i,t-1}$，(6)式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，(7)式描述了一组 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”方法减少模型的样本内性能，以期提高模型的样本外稳定性。惩罚项的设定减少模型的拟合噪声，防止过拟合。

设 $L(W;\cdot)$ 为优化自编码器的损失函数，用 W 来概括因子网络模型(5)式到(7)式到中的权重参数。将评估目标定义为：

$$L(W;\cdot)=\frac{1}{NT}\sum_{t=1}^{T}\sum_{i=1}^{N}\left\|x_{i,t}-\tilde{x}_{i,t}\right\|^{2}+\phi(W;\cdot). \tag{8}$$

其中，$\phi(W;\cdot)$ 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，它采取的形式为

$$\phi(W;\lambda)=\lambda\sum_{j}\left|W_j\right|. \tag{9}$$

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，用哑变量的方法对响应变量进行编码。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 假公司的收益率进行排序，每个月收益排名前 30%标记为上涨类别 1，后 30%的股票标记为下跌类别 0，剩余样本不参与训练。具体如下所示：

$$y_{i,t}=\begin{cases}1, & \text{上涨}\\ 0, & \text{下跌}\end{cases}. \tag{10}$$

Logistic 回归对 y 属于某一类的概率而建模而不直接对响应变量 y 建模。对本文而言，Logistic 回归建立股票上涨概率模型

$$P_{i,t} \equiv P(y_{i,t}=1|z_{i,t-1}) = \frac{1}{1+e^{-z^T(x_{i,t-1})\beta_t}}, \tag{11}$$

$$z^T(x_{i,t-1}) = x_{i,t-1}^T \Gamma. \tag{12}$$

其中 $x_{i,t-1}$ 是原始横截面公司特征，$z_{i,t-1}$ 是 $x_{i,t-1}$ 作为输入变量经过自编码器学习的压缩层的变量。

为评估分类器的概率输出，本文在概率估计上定义对数似然损失(Log-likelihood Loss)函数，也称逻辑回归损失(Logistic Loss)或交叉熵损失(Cross-entropy Loss)。对数损失通过惩罚错误的分类，实现对分类器的准确度(Accuracy)的量化。最小化对数损失基本等价于最大化分类器的准确度。对于 t 时刻的模型，对数损失函数的计算公式如下:

$$L(y_{i,t}, P_{i,t}) = -\frac{1}{N}\sum_{i=1}^{N}(y_{i,t}\log p_{i,t} + (1-y_{i,t})\log(1-p_{i,t})). \tag{13}$$

其中，$y_{i,t}$ 是所属的每个类别的概率值，$z_{i,t}$ 是逻辑回归模型的输入变量，N 为输入样本量，$p_{i,t}$ 为预测输入实例 $z_{i,t-1}$ 属于类别 1 的概率。对所有样本的对数损失表示对每个样本的对数损失的平均值，对于完美的分类器，对数损失为 0。

当得到数据模型后，在评价和比较分类模型的优劣时，关注的是其泛化能力，不能只关注模型在某个验证集上的表现。因此，一般的做法是借助统计学的方法，评估模型在某个验证集分类表现的显著性。本文针对包含二元分类问题的自编码模型，输出分类准确率分数和 AUC 作为评估模型的依据，依据自编码器和 Logistic 回归的损失函数的收敛性调整模型迭代次数。

2 数据来源

2.1 数据来源

本文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。本文选取 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，诸如资本资产定价模型（Capital Asset Pricing Model, CAPM）、Fama-French 因子模型数据（三因子、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）股票数据、企业产权分类数据等也来自国泰安数据库。在人工智能因子的经济学分析中，本文从 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，本文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（Fama 和 French[20]）。首先，本文将财务数据的样本时间区间起始点选在 2000 年后。这是由于中国股票市场从 90 年代初期建立的头十年里，市场机制不健全、上市公司数量较少、公司的财务造假和内幕交易现象较严重，因此难以得到令人信服的实证资产定价研究结论。我国在 2000 年左右加入了世界贸易组织（WTO），以此为契机，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。因此，根据目前研究中国股票市场的研究 Carpenter 等[21]和 Allen 等[22]的经验，本文将分析的起始点选在 2000 年后。

数据库中存在着一定比例的缺失值，其处理过程分为两步：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 398919 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 3406 条，如图 4 所示。

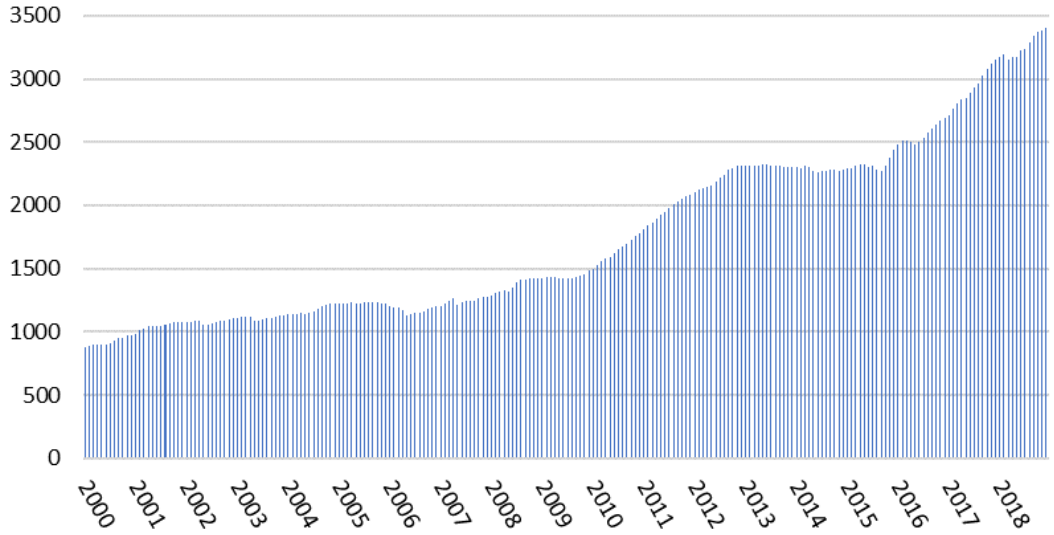

图 4 2000 年 1 月至 2018 年 12 月月度有效样本数量

因子的描述性统计显示，不同因子的取值在数量级及分布上存在显著差异，可能导致预测偏差，比如：①量级较大的特征在预测时占据主导地位；②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，本文将公司特征数据进行归一化处理，映射到 0 到 1 之间。

由于机器学习算法需要一段样本期作为训练集，本文设定不同长度的滚动窗口，采用逐月滚动的训练方法。除此之外，选取 10%的交叉验证（Cross-validation）样本进行检验。最终进行横截面收益预测的区间为 2007 年 7 月至 2018 年 12 月。

2.2 公司特征变量

对于金融大数据的构建，Mclean 和 Pontiff[3]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。本文全面梳理了实证资产定价的文献，整理了与公司股票预期收益相关的 89 个公司特征变量。公司特征指标名称、定义及文献出处详见附录 1。上述指标通过国泰安数据库中财务报表数据和股票交易数据计算而得。综合考虑指标的重要性和数据的可获得性，本文使用的公司变量包含了 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。

3 实证结果与分析

这一部分将从横截面股票收益预测的角度出发，研究不同自编码器结构是否有明显的差别，进行合理的比较与分析，并进一步考察各预测因子在模型中的重要程度和其时变性。

3.1 自编码器网络结构

本文在实证中，共测试 9 种不同网络架构的自编码器（Autoencoder），其简写与结构

为下表所示：

表 1 自编码器网络结构

序号	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

其中，AE(32)、AE(64)和 AE(128)为一层自编码结构，即为前文的标准自编码器，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，均属于前文中的堆叠式自编码器。

3.2 投资组合的构建

投资组合分析法（Portfolio Analysis）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，该方法在横截面股票收益异同的研究中最常见。投资组合分析法的实质是按不同指标对股票进行排序和分组，构建投资组合并持有一定的时期（年、季、月）。然后，在计算投资组合收益的基础上，比较不同投资组合的收益及超额收益是否有显著差异，从而检验不同指标对于横截面股票收益是否产生影响。相较于线性回归，投资组合分析法能够较少地受到线性假设的限制，同时对于个别的极端值也不敏感，因此在实证资产定价领域得到了广泛应用。该方法相当于一种利用每一段等长的过去数据对一定时间内的未来结果进行反复预测的方法。

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。与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 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。

3.3 横截面收益预测结果

我们使用投资组合法检验自编码器结构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上涨概率的大小构建的各投资组合的月平均收益、夏普比率（年化后）、CAPM-α、FF3-α、FF5-α 以及各组 t 值（方括号中）。其中，月平均收益和月度的超额收益（α）在表格中以百分比形式汇报。

表 2　AE(128,64)自编码器结构的横截面股票收益

投资组合	月平均收益	夏普比率	CAPM-α	FF3-α	FF5-α
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]

分析表 2 的结果可知，各投资组合的月平均收益随着上市公司的预期上涨概率的增加而增加。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。这表明通过投资于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 以上。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。多空对冲组合的月平均 CAPM-α、FF3-α、FF5-α 分别为 1.84 （t=3.64）、1.26 （t=3.63）、1.12%（t=2.29），均在 5%水平以上显著。这说明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（滚动窗口为 12 个月）

自编码器结构	AE(32)			AE(64)			AE(128)		
投资组合	Low	High	H-L	Low	High	H-L	Low	High	H-L
月平均收益	-0.27	1.41	1.67	-0.26	1.42	1.68	-0.31	1.34	1.65
t值	[-0.32]	[1.81]	[3.36]	[-0.31]	[1.84]	[3.36]	[-0.37]	[1.71]	[3.32]
夏普比率	-0.2	0.41	0.99	-0.2	0.42	0.99	-0.22	0.39	0.98
FF5-α	-0.81	0.23	1.04	-0.81	0.20	1.01	-0.84	0.10	0.94
t 值	[-2.62]	[0.78]	[2.16]	[-2.71]	[0.67]	[2.13]	[-2.89]	[0.33]	[2.00]

自编码器结构	AE(32,16)			AE(64,32)			AE(128,64)		
投资组合	Low	High	H-L	Low	High	H-L	Low	High	H-L
月平均收益	-0.21	1.33	1.54	-0.17	1.29	1.46	-0.28	1.54	1.82
*t*值	[-0.24]	[1.74]	[3.06]	[-0.20]	[1.68]	[2.87]	[-0.34]	[1.97]	[3.54]
夏普比率	-0.18	0.39	0.90	-0.17	0.37	0.85	-0.21	0.46	1.05
FF5-α	-0.75	0.16	0.91	-0.7	0.06	0.77	-0.79	0.32	1.12
t 值	[-2.54]	[0.55]	[1.90]	[-2.35]	[0.23]	[1.60]	[-2.63]	[1.09]	[2.29]
自编码器结构	AE(32,16,8)			AE(64,32,16)			AE(128,64,32)		
投资组合	Low	High	H-L	Low	High	H-L	Low	High	H-L
月平均收益	-0.09	1.22	1.32	-0.05	1.41	1.47	-0.22	1.36	1.58
*t*值	[-0.11]	[1.61]	[2.69]	[-0.06]	[1.8]	[2.96]	[-0.26]	[1.75]	[3.09]
夏普比率	-0.14	0.35	0.79	-0.13	0.41	0.87	-0.19	0.40	0.91
FF5-α	-0.66	0.13	0.79	-0.60	0.14	0.74	-0.72	0.14	0.86
t 值	[-2.34]	[0.46]	[1.78]	[-2.09]	[0.48]	[1.58]	[-2.43]	[0.47]	[1.80]

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，输入变量所包含的这些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。

由于现有文献中有关机器学习算法中对于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 和图 5 所示。

表 4 不同滚动窗口的自编码器的多空组合（H-L）在横截面的股票收益

自编码器序号	1	2	3	4	5	6	7	8	9
6 个月滚动窗口									
月平均收益	1.13	1.08	1.08	0.93	1.11	1.12	0.6	1.03	1.18
*t*值	[2.23]	[2.10]	[2.08]	[1.82]	[2.11]	[2.09]	[1.1]	[1.95]	[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									

月平均收益	1.25	1.30	1.43	1.25	1.35	1.33	1.47	1.29	1.30
*t*值	[2.74]	[2.86]	[3.11]	[2.81]	[2.93]	[2.86]	[3.49]	[2.89]	[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									
月平均收益	0.96	1.06	1.11	1.15	1.08	1.02	0.91	0.95	0.94
*t*值	[2.09]	[2.34]	[2.45]	[2.63]	[2.28]	[2.20]	[2.14]	[2.10]	[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									
月平均收益	0.95	1.07	1.12	0.86	1.00	1.16	1.13	0.66	1.26
*t*值	[2.08]	[2.41]	[2.48]	[1.89]	[2.09]	[2.52]	[2.47]	[1.37]	[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									
月平均收益	0.98	1.14	1.17	0.86	1.01	1.13	0.67	0.80	1.01
*t*值	[2.04]	[2.49]	[2.53]	[1.77]	[2.12]	[2.41]	[1.45]	[1.64]	[2.16]
夏普比率	0.6	0.74	0.74	0.52	0.63	0.71	0.43	0.48	0.64

注：为简明起见，自编码结构用序号代替，自编码器序号所对应的自编码结构见表 1；12 个月滚动窗口的横截面收益如表 3 所示。

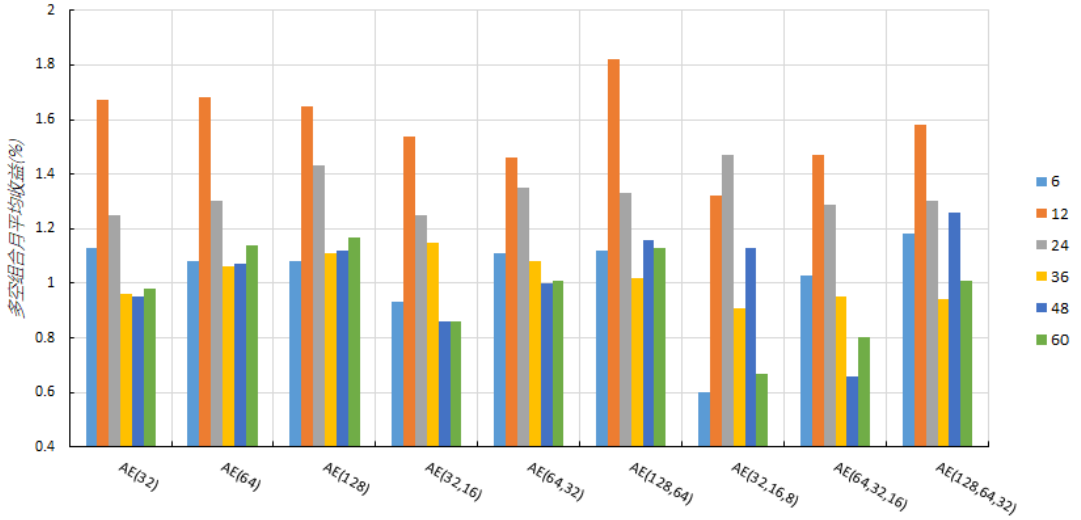

图 5 不同滚动窗口的自编码器的多空组合在横截面的股票收益

结合表 2 的结果可以发现，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 1.81%，除此之外的其他结构也在 1.36%到 1.68%之间的范围，明显高于其他的滚动窗口。此结果从图 5 展示的更加清晰，12 个月为训练样本的结果（图中为橙色）的多空组合月平均收益最好，24 个月（灰色）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[19]在运用机器学习算法时最优的滚动窗口长度一致。前文提到，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。训练集的长度设置的过大，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。在前者的研究中普遍认为，在多种宏观经济周期中，和股市密切相关的是基钦周期，又称库存周期，以企业库存变化为主要驱动因素和划分依据，股票行情的强弱与库存周期的强弱成正相关关系。每个库存周期平均长度 3 年左右，其中，上升期 1-1.5 年，下行期 1-1.5 年，所

以训练集不宜过大，否则会使训练信息中包含更多噪音，进而增加了降低信噪比，这为 12 个月训练集的结果最优提供了理论基础。

3.4 自编码因子与其他因子指标的双变量排序

本小节，我们进一步考察基于自编码预测值和其他指标的双变量排序的投资组合的表现，因子指标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）、资产收益率（ROA、非流动性（ILL）和换手率（TURN）。

根据 Hou 等[1]的研究，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；按照因子指标的数值从小到大进行排序，按 30 和 70 百分位点将全体公司分为三组投资组合，我们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% 分位数点公司的称为“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（size）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，我们还选取了衡量投资水平的指标净运营资本（NOA）、盈利类指标资产收益率（ROA）以及非流动性（ILL）和换手率（TURN）四个指标，使研究更加全面。

表 5 基于自编码预测值和其他因子指标双变量排序（Double Sorts）的投资组合表现

投资组合	公司规模（SIZE）			账面市值比(BM)			净运营资本（NOA）		
	微型	小型	大型	成长	中	价值	低	中	高
Low	1.43	0.42	-0.36	-0.04	-0.38	-0.47	-0.08	-0.24	-0.44
	[1.35]	[0.43]	[-0.42]	[-0.05]	[-0.43]	[-0.49]	[-0.09]	[-0.28]	[-0.53]
High	2.35	1.24	0.91	1.19	1.37	1.79	1.17	1.39	1.73
	[2.48]	[1.45]	[1.26]	[1.40]	[1.64]	[2.16]	[1.51]	[1.79]	[1.97]
H-L	0.91	0.82	1.26	1.23	1.75	2.26	1.25	1.63	2.17
	[1.96]	[1.91]	[2.35]	[2.37]	[3.41]	[3.21]	[2.32]	[2.91]	[4.01]
投资组合	资产收益率(ROA)			非流动性（ILL）			换手率（TURN）		
	低	中	高	低	中	高	低	中	高
Low	-0.53	-0.46	0.02	-0.51	0.83	1.00	0.37	-0.21	-0.58
	[-0.58]	[-0.53]	[0.02]	[-0.59]	[0.94]	[1.08]	[0.42]	[-0.25]	[-0.67]
High	1.90	1.41	1.13	1.03	1.22	2.07	1.59	1.55	1.28
	[2.33]	[1.71]	[1.41]	[1.37]	[1.52]	[2.34]	[2.10]	[1.75]	[1.45]
H-L	2.43	1.87	1.11	1.54	0.39	1.17	1.22	1.76	1.84
	[3.67]	[3.37]	[2.26]	[2.59]	[0.87]	[2.47]	[1.84]	[3.44]	[3.12]

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，分别为 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%，总的来说，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%到高净运营资本 2.17%（t=4.01）；而自编码投资在所有盈利能力的公司中均有超额收益，

但在盈利能力低的公司中对收益率的区分更好，从 Low 组的-0.53%到 High 组 1.90%，多空组合获得 2.43%月平均收益。除此之外，对于非流动性和换手率两个因子的双变量排序可以看出，自编码多空投资组合在高流动性和高换手率的公司种能获得更高的回报。

3.5 因子的重要度与时变性

本文进一步考察各类指标在不同模型中的重要程度以及其时变性。现学术界在考察预测因子的重要程度中普遍借鉴 Kelly 等[23]的做法，即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，比较微观因子的相对重要性。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，并不能得到因子的重要程度随时间的变化。

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，随着市场结构性改革和流动性的增加，单个异象的预测能力在随时间逐渐减弱。在美国，大多数异象在 2001 年以后开始逐步减弱，由 12 个异常组成的交易策略的平均回报和夏普比率减少了一半以上。Harvey 等[24]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。在我国市场，尹力博等[25]研究表明，在样本期内，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，而规模、特质波动率、换手率以及市场风险异象正逐渐显现并仍有增强的趋势。

因此对因子在时间上的有效性的验证的价值尤为重要。本文输出标准自编码器的输出层各因子权重，由于本文所建立的模型是逐月滚动的，所以本文通过输出的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。

本文将模型输出层的权重值取绝对值，然后每 6 个月进行加总平均，最后按顺序输出排在前十的因子，得到表 6。通过表 6 可以看出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。具体而言，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 07 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 17 年以后无形资产类因子的权重在逐步加大。

表 6 自编码器输出层因子权重排序

year	1	2	3	4	5	6	7	8	9	10
2001	MOM1M	PRCDEL	TAN	GM	CHMOM	PRC	ROA	FOE	ROE	MAXRET
	MOM1M	TAN	PRCDEL	ROE	FOE	EPS	MOM6M	CHMOM	PRC	MAXRET
2002	ROE	EPS	PRCDEL	FOE	MOM6M	TAN	MOM1M	CHMOM	AGE	CHTX
	INDMOM	FOE	EPS	MOM6M	CHMOM	AGE	PRCDEL	TAN	ROE	MAXRET
2003	INDMOM	EPS	TAN	FOE	MAXRET	CHMOM	MOM6M	IVOL	AGE	PRCDEL
	INDMOM	EPS	TAN	PRCDEL	CHMOM	FOE	MAXRET	MOM6M	AGE	MOM1M
2004	INDMOM	EPS	PRCDEL	CHMOM	MOM6M	AGE	FOE	TAN	MOM1M	ROE
	INDMOM	CHMOM	MOM6M	ROE	PRCDEL	AGE	MOM1M	EPS	FOE	FM
2005	INDMOM	ROE	CHMOM	AGE	MOM1M	PRCDEL	FOE	dPIA	MOM6M	GM
	GM	PRCDEL	DOLVOL	FOE	CHMOM	AGE	INDMOM	TBI	CFOA	MOM1M
2006	PRCDEL	DOLVOL	GM	FOE	CFOA	AGE	TAN	CHMOM	IVC	BETASQ
	PRCDEL	TAN	GM	AGE	DOLVOL	CFOA	CHMOM	FOE	CFP	EP
2007	PRCDEL	GM	AGE	TAN	CHMOM	EPS	CFOA	CFD	INDMOM	EY
	EPS	CHMOM	TAN	GM	EY	PRCDEL	ROE	TBI	IP	AGE
2008	EPS	ROE	EY	PRCDEL	EP	GM	TAN	IVOL	CHMOM	RNA
	PRCDEL	ROE	EPS	IVOL	EY	TAN	FOE	DOLVOL	CFOA	ROA
2009	ROA	TAN	PRCDEL	EY	EPS	IVOL	CFOA	DOLVOL	ROE	CFD
	ROA	TAN	EY	CFOA	PRCDEL	EPS	GM	DOLVOL	ROE	IVOL
2010	ROA	PRCDEL	GM	EPS	EY	CFOA	TAN	ROE	dPIA	DOLVOL
	PRCDEL	GM	EPS	dPIA	CFOA	ROA	TAN	EY	FM	FOE
2011	CFOA	GM	TAN	PRCDEL	dPIA	EPS	FM	ROA	EY	CHMOM
	CFOA	TAN	GM	FOE	dPIA	EY	PRCDEL	FM	ROA	DOLVOL
2012	CFOA	GM	TAN	PRCDEL	FM	FOE	EY	IVC	DOLVOL	dPIA
	FM	CFOA	PRCDEL	TAN	GM	dPIA	IVC	DOLVOL	OCFP	CAC
2013	FM	PRCDEL	TAN	GP	GM	dPIA	DOLVOL	IVC	OCFP	CAC
	FM	TAN	GP	PRCDEL	IVC	GM	DOLVOL	PRC	ROIC	CFP
2014	FM	TG	TAN	IVC	GM	DOLVOL	CFP	EY	EP	EPS
	TAN	GM	DOLVOL	CFP	CP	PRCDEL	EY	TG	INDMOM	FM
2015	PRCDEL	TAN	CFP	DOLVOL	GM	CP	TBI	OCFP	CHMOM	INDMOM
	PRCDEL	DOLVOL	TAN	CFP	CINVEST	TBI	IP	OCFP	IVC	CHMOM
2016	PRCDEL	DOLVOL	TAN	BETASQ	GP	CFP	IP	EY	OCFP	PRC
	TAN	PRCDEL	TG	DOLVOL	GP	CFP	EPS	OCFP	EP	BETASQ
2017	TAN	EY	DOLVOL	PRCDEL	EPS	OCFP	CFP	PRC	CHMOM	EP
	TAN	EY	PRCDEL	SG	IVC	GM	DOLVOL	EPS	CHTX	CFP
2018	EY	PRCDEL	TAN	TG	IVC	CFD	GM	EP	CHTX	EPS
	EY	CFD	PRCDEL	TAN	SIZE	GM	TG	EP	IVC	CHTX

估值与成长类
投资类
盈利类
惯性类
交易摩擦类
无形资产类

注：表中的英文缩写所代表的公司特征指标详见论文附录 1 公司特征指标名称、定义以及文献出处。

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。

对于动量效应的研究，国内外学者主要从传统风险收益模型和行为金融，但包括 Fama-French 因子模型在内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，这和表中股市早期市场因子的权重也比较大的结果相吻合。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 07 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 和 Robert[26]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，传统的回归分析和投资组合分类可能不足以处理大量的预测变量。机器学习方法提供了一种自然的方法来适应高维预测集和灵活的函数形式，并采用“正则化”方法来选择模型和减轻过度拟合偏差（Gu 等[11]）。然而，自编码器模型本质上是不透明的，通常被称为“黑匣子”。在这一部分中，将超额收益与市场状况的变化联系起来，探讨与资产价格息息相关的上市公司特征受到哪些宏观经济变量的影响，重点放在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。

本文研究投资组合的收益是否随市场状态而变化，具体受到哪些因素的影响，本文参考 Avramov 等[27]在探讨预测收益是否随时间变化时采用的方法，采用以下月度时间序列回归方程进行分析：

$$HML_t = \alpha + \varphi Dummy_{t-1}^{high} + \beta' F_t + e_t \tag{14}$$

其中，HML_t 是指在t月的所有股票使用自编码模型构造的多空对冲（Long-short）组按市值加权的超额收益。β 是5×1维列向量，F_t 是指 Fama 和 French（2013）中的五个常见风险因子叠加，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，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，因此本文参考 Cooper 等[28]以及 Han 等[29]的在探讨资产收益与经济周期关联时采用的方法进行分析。同时，本文报告 Newey - West 调整后的t值。

4.1 与宏观经济状态的联系

首先，本文研究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。宏观经济是个体经济的总和，企业的投资价值会在宏观经济的总体中综合反映出来。因此，宏观经济变量会影响整个股票市场的投资价值。Charles等[30]研究了股票市场发展的制度和宏观经济决定因素，表明宏观经济因素是新兴市场国家股市发展的重要决定因素。因此，本文参考Goyal和Welch[31]和Avramov等[27]，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，该值等于每股价格与每股收益的比，市盈率指标不仅综合考虑了金融资产的价格和收益，同时也考虑了盈利能力、成长能力、风险等因素。Lipe 和 Kormendi[32]的 NPVGO 模型说明市盈率能反映市场风险与收益，陈占锋[33]指出市盈率可以度量股票市场的泡沫大小。何诚颖[34]说明以市场平均市盈率指标作为股票市场定价基准应是有效的。由此看出，市场市盈率是衡量股市热度和健康状况的重要指标。陈国进等[35]发现宏观风险的长期性和宏观经济的不确定性对于中国股票市场的股权溢价之谜有很好解释作用。

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 7 采取了（5.1）式中的回归方程，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 7 自编码因子与宏观经济状况

宏观经济变量	国内生产总值(GDP)	采购经理人指数(PMI)	PMI:生产	PMI:新订单	市场市盈率
α	0.28	0.33	0.41	0.28	0.44
	[0.51]	[0.45]	[0.86]	[0.37]	[0.72]
φ	1.72	1.52	1.38	1.62	1.37
	[2.16]	[1.63]	[1.67]	[1.77]	[1.65]
MKT	0.65	0.70	0.70	0.70	0.65
	[3.67]	[4.26]	[4.44]	[3.88]	[3.32]
SMB	0.62	0.72	0.70	0.70	0.70
	[2.44]	[3.90]	[3.09]	[3.23]	[3.52]
HML	-0.18	-0.17	-0.17	-0.17	-0.18
	[-2.20]	[-2.37]	[-2.28]	[-2.64]	[-2.00]
RMW	-0.13	-0.14	-0.11	-0.11	-0.18
	[-0.42]	[-0.47]	[-0.38]	[-0.39]	[-0.62]
CMA	-0.12	-0.14	-0.12	-0.12	-0.12
	[-0.47]	[-0.66]	[-0.47]	[-0.51]	[-0.44]

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)，这说明自编码方法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这些因素的影响。

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。庄芳等[36]研究表明,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。黄赜琳和朱保华[37]构建了财政税收的实际经济周期模型，实证分析财政政策调整对中国实体经济波动的影响，以及以税收政策变动为代表的供给政策在调节中国宏观经济波动过程中的效果显著。王少林等[38]表明我国货币政策与股票市场之间存在非对称的互动关系。Schnabl 等[39]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[40]和贾盾等[41]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[42]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[43]表明，我国的货币政策能有效影响社会融资规模，社会融资规模也对经济增长、物价水平、投资消

费等实体经济指标产生较大影响，社会融资规模是反映金融与经济关系的良好指标。盛松成和谢洁玉[44]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指标。

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，Ang等[45]研究了股票收益横截面中总波动风险的定价。发现对总波动率高度敏感的股票平均回报率较低。除此之外Brunnermeier和Pedersen[46]和Adrian 和 Shin[47]等理论研究表明，由于资金紧张和风险偏好降低，波动性加大会降低做市商的流动性供应能力。因此，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 8 自编码因子与宏观经济政策

宏观经济变量	税收收入	人民币存款准备金率	社会融资规模	股票市场波动性
α	1.77	1.89	0.28	2.17
	[3.87]	[3.91]	[0.44]	[4.73]
φ	-1.29	-1.58	1.57	-2.12
	[-1.71]	[-2.11]	[1.84]	[-3.00]
MKT	0.68	0.68	0.75	0.69
	[4.11]	[3.89]	[3.88]	[4.14]
SMB	0.69	0.72	0.71	0.68
	[3.50]	[3.67]	[3.14]	[3.37]
HML	-0.17	-0.18	-0.17	-0.18
	[-2.26]	[-2.30]	[-1.97]	[-2.25]
RMW	-0.07	-0.15	-0.06	-0.16
	[-0.22]	[-0.53]	[-0.19]	[-0.46]
CMA	-0.09	-0.15	-0.14	-0.19
	[-0.32]	[-0.51]	[-0.54]	[-0.82]

表 8 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率。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，月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。

5 结论与建议

本文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。本文构建了包含 89 个与公司特征密切相关的金融大数据指标集，使用自编码模型对上市公司股票的上涨概率进行预测，并进行投资组合分析。此外，本文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。

本文的而研究发现，自编码方法能够提取包含公司特征的金融大数据中的信息，得到包含有效信息的复合信号，能够在横截面股票收益预测中获得显著的超额收益。对因子重要度的研究中表明，股票市场异象具有时变特征，不同时期算法对不同因子赋予的权重不同。在探究自编码方法提取的复合信号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。

本文的研究进一步推动了基于中国市场的机器学习资产定价研究，同时为探讨机器学习方法的预测经济来源进行了分析。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独特特征。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长足发展。

参考文献：

[1] Hou, K., Xue, C., and Zhang, L. Replicating anomalies [J]. Review of Financial Studies, 2020,33(5), 2019 – 2133.

[2] Cochrane, J.H.Presidential Address:Discount Rates [J].The Journal of Finance, 2011, 66 (4) :1047-1108.

[3] Mclean R D, Pontiff J. Does Academic Research Destroy Stock Return Predictability? [J]. Journal of Finance, 2016, 71(1):5-32.

[4] Bali, T.G., R.F.Engle, and S.Murray.Empirical Asset Pricing:The Cross Section of Stock Returns [M].Hoboken, New Jersey:Wiley, 2016.

[5] Kozak, S., S.Nagel, and S.Santosh. Shrinking the Cross Section [J].Journal of Financial Economics,2020, Vol. 135, Issue 2, 271-292.

[6] Kelly, B. T., Pruitt, S., and Su Y.. Characteristics are covariances: A unified model of risk and return[J]. Journal of Financial Economics, 2019,134(3): 501-524.

[7] 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.

[8] Bansal R , Yaron A . Risks for the Long Run: A Potential Resolution of Asset Pricing Puzzles[J]. Journal of Finance, 2010, 59.

[9] Zhiguo He, Arvind Krishnamurthy. Intermediary Asset Pricing[J]. American Economic Review,2013,103(2): 732-70

[10] Pohl W , Schmedders K , Wilms O . Higher Order Effects in Asset Pricing Models with Long-Run Risks[J]. The Journal of Finance, 2018, 73(3).

[11] Gu, S., Kelly, B. T., Xiu, D.. Empirical asset pricing via machine learning[J]. Review of Financial Studies, 2020,Vol. 33, Issue 5, 2223-2273.

[12] Hinton, Geoffrey E, and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks[J], science. 2006,313, 504-507.

[13] Light, N., D.Maslov, and O.Rytchkov. Aggregation of Information about the Cross Section of Stock Returns:A Latent Variable Approach [J].The Review of Financial Studies, 2017, 30 (4) :1339-1381.

[14] Feng G , He J , Polson N G . Deep Learning for Predicting Asset Returns[J]. Papers, 2018.

[15] 洪永淼,汪寿阳.大数据革命和经济学研究范式与研究方法[J].财经智库,2021,6(01):5-37+142-143.

[16] 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［J］. 金融研究, 2017, (05) :111-126.

[17] 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 ［J］. 管理世界, 2018, (1) :180-181.

[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(08):61-79.

[19] Bengio, Y. , Lamblin, Pascal , Popovici, D. , Larochelle Hugo and Montreal, U.. Greedy layer-wise training of deep networks[J]. Advances in Neural Information Processing Systems 19 (NIPS'06).2007,153–160

[20] Fama, E. F., and French, K. R..Common risk factors in the returns on stocks and bonds[J], Journal of Financial Economics, 1993,Vol.33(1),pp.3~56.

[21] Carpenter, Jennifer N., Fangzhou Lu, and Robert F. Whitelaw. The real value of China's stock market[J], National Bureau of Economic Research, 2015, No. w20957.

[22] Allen, F., Qian, J., Shan, C., and Zhu, J. 2018. Dissecting the long-term performance of the Chinese stock market. SSRN working paper.

[23] Kelly B T , Pruitt S , Su Y . Some Characteristics are Risk Exposures, and the Rest are Irrelevant[J]. SSRN Electronic Journal, 2017.

[24] Harvey, C.R., Y.Liu, H.Zhu.…and the Cross-Section of Expected Returns [J].The Review of Financial Studies, 2016, 29 (1) :5-68.

[25] 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(08) :14-25.

[26] Hall and Robert E.. The stock market and capital accumulation[J]. American Economic Review, 2001, vol.91, pp.1185–1202.

[27] Avramov D , Cheng S , Metzker L . Machine Learning versus Economic Restrictions:

Evidence from Stock Return Predictability [J]. 2019. SSRN Electronic Journal. 10.2139.
[28] Cooper, M. J., Gutierrez R. C. and Hameed A., 2004, Market states and momentum. Journal of Finance 59, 1345-1365.
[29] Han, Y., Zhou G. and Zhu Y.. A trend factor: Any economic gains from using information over investment horizons?[J].Journal of Financial Economics ,2016, 122(2), 352-375.
[30] Charles, Amo, Yartey. The institutional and macroeconomic determinants of stock market development in emerging economies [J]. Applied Financial Economics. 2010, 20: 1615–1625.
[31] Welch I , Goyal A . A Comprehensive Look at The Empirical Performance of Equity Premium Prediction[J]. Social ence Electronic Publishing, 2008, 21(4):1455-1508.
[32] Lipe, R． Kormendi. Mean Reversion in Annual Earning and Its Implications for Security Valuation[J]． Review of Quantitative Finance and Accounting,1994,4: 205-242．
[33] 陈占锋.上海股票市场A股泡沫问题:市盈率测量与综合解释 [J].世界经济,2002(07):63-70.
[34] 何诚颖. 中国股市市盈率分布特征及国际比较研究 [J]. 经济研究, 2003(09):74-81+95.
[35] 陈国进, 黄伟斌, Tribhuvan Puri. 宏观长期风险与资产价格:国际比较与中国经验 [J]. 世界经济, 2014, 37(06):51-72.
[36] 庄芳, 庄佳强, 朱迎. 我国财政政策和货币政策协调配合的定量效应——基于协整向量自回归的分析[J]. 金融研究, 2014(12):71-85.
[37] 黄赜琳, 朱保华. 中国的实际经济周期与税收政策效应[J]. 经济研究, 2015, 50(03):4-17+114.
[38] 王少林, 林建浩, 杨燊荣. 中国货币政策与股票市场互动关系的测算——基于FAVAR-BL方法的分析[J]. 国际金融研究, 2015(05):15-25.
[39] Drechsler I , Savov A , Schnabl P . A Model of Monetary Policy and Risk Premia[J]. Journal of Finance, 2018, 73(1):317-373.
[40] 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 [C]. 《国际货币评论》2018年第二季度合辑. :中国人民大学国际货币研究所, 2018:184-205.
[41] 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(07):76-95.
[42] 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, 053(008):68-83.
[43] 盛松成. 社会融资规模与货币政策传导 [J]. 金融研究, 2012(10):1-14.
[44] 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.
[45] Ang, A., Hodrick, R.J., Xing, Y. and Zhang, X.. The cross-section of volatility and expected returns[J], The Journal of Finance,2006,Vol.61(1),pp.259-299.
[46] Brunnermeier, M. K., and L. H. Pedersen. Market liquidity and funding liquidity[J]. Review of Financial Studies, 2009, 22:2201–2238.
[47] Adrian, T., and H. S. Shin. 2010. Liquidity and leverage. Journal of Financial Intermediation 19:418–437.

附录 1

公司特征指标名称、定义及文献出处

缩写	名称	文献出处
估值类指标（20 个）		
AM	资产市值比（Assets-to-market）	Fama and French (1992)
BM	账面市值比（Book-to-market equity）	Rosenberg, Reid, and Lanstein (1985)
CFP	现金流股价比（Cash flow-to-price）	Lakonishok, Shleifer, and Vishny (1994)
DER	债务股本比（Debt-to-equity ratio）	Bhandari (1988)
DP	红利价格比（Dividend-to-price ratio）	Litzenberger and Ramaswamy (1982)
EP	市盈率（Earnings-to-price）	Basu (1983)
LG	债务增长（Liability growth）	Richardson, Sloan, Soliman, and Tuna (2005)
OCFP	营运现金流价格比（Operating cash flow-to-price）	Desai, Rajgopal, and Venkatachalam (2004)
PY	股息股价比（Payout yield）	Boudoukh, Michaely, Richardson, and Roberts (2007)
Rev1	反转（Reversal）	De Bondt and Thaler (1985)
SG	可持续增长率（Sustainable growth）	Lockwood and Prombutr (2010)
SMI	销量增长与存货增长差（Sales growth minus inventory growth）	Abarbanell and Bushee (1998)
SP	销量价格比（Sales-to-price）	Barbee, Mukherji, and Raines (1996)
TG	纳税增长（Tax growth）	Thomas and Zhang (2011)
DA	财务杠杆长期负债/长期负债与股东权益之和	苏冬蔚(2005)
DPR	股利支付率	蒋志强,田婧雯,周炜星(2019)
NCF	现金流量价格比	赵春光（2004）
FM	基于现金流构建的 BM 因子	干伟明，张涤新 （2018）
FOE	基于现金流构建的 ROE 因子	干伟明，张涤新 （2018）
NAPS	每股净资产(Net Assets Per Share)	陈信元,陈冬华,朱红军（2002）
投资类指标（12 个）		
ACC	应计收入（Accruals）	Sloan (1996)
PACC	百分比应计收入（Percent accruals）	Hafzalla, Lundhom, and Van Winkle (2011)

CAPXG	资本开销增长率（Capital expenditure growth）	Mcconnell and Muscarella (1985)
dBe	股东权益变化（Change in shareholders' equity）	Richardson, Sloan, Soliman, and Tuna (2005)
dPIA	固定资产与存货变化率（Changes in PPE and inventory-to-assets）	Lyandres, Sun, and Zhang (2008)
IA	投资资产比（Investment-to-assets）	Cooper, Gulen, and Schill (2008)
IVC	存货变化率（Inventory change）	Thomas and Zhang (2002)
IVG	存货增长率（Inventory growth）	Belo and Lin (2011)
NOA	净运营资本（Net operating assets）	Hirshleifer, Hou, Teoh, and Zhang (2004)
CINVEST	公司投资（Corporate investment）	Titman, Wei, and Xie (2004)
IC	投资资本比	王茵田,朱英姿（2011）
GS	销售增长率（Growth-of-sale）	肖军,徐信忠(2004)
盈利类指标（23 个）		
ATO	资产换手率（Asset turnover）	Soliman (2008)
CFOA	现金流资产比（Cash flow over assets）	Asness, Frazzini, and Pedersen (2017)
CP	现金生产率（Cash productivity）	Chandrashekai and Rao (2009)
CTA	现金资产比（Cash-to-assets）	Palazzo (2012)
CTO	资本换手率（Capital turnover）	Haugen and Baker (1996)
EBIT	息税前收益（Earnings before interests and taxes）	Greenblatt (2006)
EY	企业收益率（Earnings yield）	Greenblatt (2006)
GM	边际毛利（Gross margins）	Novy-Marx (2013)
GP	毛利率（Gross profitability）	Novy-Marx (2013); Jiang, Qi, and Tang (2018)
NPOP	净利润（Net payout over profits）	Asness, Frazzini, and Pedersen (2017)
RNA	净运营资产收益（Return on net operating assets）	Soliman (2008)
ROA	资产收益率（Return on assets）	Balakrishnan, Bartov, and Faurel (2010); Jiang, Qi, and Tang (2018)
ROE	股权收益率（Return on equity）	Hou, Xue, and Zhang (2015); Jiang, Qi, and Tang (2018)
ROIC	投资型资产收益率（Return on invested capital）	Greenblatt (2006)

TBI	应税所得与账面资产比（Taxable income-to-book income）	Green, Hand, and Zhang (2013)
Z	Z 评分（Z-score）	Dichev (1998)
SALEREV	销售收入比（Sales to receivables）	Ou and Penman (1989)
CHTX	纳税变化额（Change in tax expense）	Thomas and Zhang (2011)
EPS	每股盈余	赵宇龙（1998）
CEPS	永久经济盈余 CEPS	赵宇龙,易琮（1999）
TA	有形资产与营业收入比	罗婷,朱青,李丹（2009）
IP	投资收益占比	翟进步,罗玫（2014）
TMT	主营业务利润率环比增长率	干伟明,张涤新（2018）
趋势类指标（7 个）		
CHMOM	6 个月惯性变化（Change in 6-month momentum）	Gettleman and Marks (2006)
VOLT	交易量趋势（Volume trend）	Haugen and Baker (1996)
INDMOM	行业惯性（Industry momentum）	Moskowitz and Grinblatt (1999)
MOM1M	1 个月惯性（1-month momentum）	Jegadeesh and Titman (1993)
MOM6M	6 个月惯性（6-month momentum）	Jegadeesh and Titman (1993)
MOM12M	12 个月惯性（12-month momentum）	Jegadeesh and Titman (1993)
MOM36M	36 个月惯性（36-month momentum）	Jegadeesh and Titman (1993)
交易摩擦类指标（15 个）		
BETA	系统性风险系数（Market beta）	Fama and MacBeth (1973)
BETASQ	系统性风险系数方差（Beta squared）	Fama and MacBeth (1973)
IVOL	异质性收益波动率（Idiosyncratic return volatility）	Ali, Hwang, and Trombley (2003)
ILL	非流动性（Illiquidity）	Amihud (2002)
MAXRET	最大日收益（Maximum daily returns）	Bali, Cakici, and Whitelaw (2011)
PRC	股价（Price）	Blume and Husic (1973)
PRCDEL	延迟股价（Price delay）	Hou and Moskowitz (2005)
DOLVOL	人民币交易量（RMB trading volume）	Chordia, Subrahmanyam, and Anshuman (2001)
SIZE	公司规模（Frim size）	Banz (1981)

STD_DOLVOL	人民币交易量波动率（Volatility of RMB trading volume）	Chordia, Subrahmanyam, and Anshuman (2001)
STD_TURN	换手率波动率（Volatility of turnover）	Chordia, Subrahmanyam, and Anshuman (2001)
RETVOL	收益率波动率（Return volatility）	Ang, Hodrick, Xing, and Zhang (2006)
TURN	股票换手率（Share turnover）	Datar, Naik, and Radcliffe (1998)
ZEROTRADE	零交易天数（Zero trading days）	Liu (2006)
TOR	销售收入(Total Operaing Revenue)	陈信元,陈冬华,朱红军(2002)
无形资产类指标（12 个）		
AGE	公司年龄（Firm age）	Jiang, Lee, and Zhang (2005)
CFD	现金流负债比（Cash flow-to-debt）	Ou and Penman (1989)
CR	流动比率（Current ratio）	Ou and Penman (1989)
CRG	流动比率增长（Current ratio growth）	Ou and Penman (1989)
QR	速动比率（Quick ratio）	Ou and Penman (1989)
QRG	速动比率增长（Quick ratio growth）	Ou and Penman (1989)
SC	销量现金比（Sales-to-cash）	Ou and Penman (1989)
SI	销量存货比（Sales-to-inventory）	Ou and Penman (1989)
OL	经营杠杆（Operating leverage）	Novy-Marx (2011)
TAN	资产有形度（Tangibility）	Hahn and Lee (2009)
IBV	无形资产/股本(Intangible Book Value)	薛云奎,王志台（2001）
CAC	现金及其等价物对固定资产净值比(Cash asset to Capital)	王茵田,朱英姿（2011）