摘要

随着信息时代的进一步发展，各种数据被广泛地应用于各行各业中。本文研究的是高维时间序列数据，高维数据的主要特点是数据维度较大，在模型识别和参数估计方面存在困难。为了更好地对高维时间序列数据进行建模估计，文章将机器学习模型引入潜在因子模型中，并利用股票日收益率等数据对国证 300 指数相关市场进行分析。

Chang 等（2015）提出了一种具有潜在因素、内生性和非线性的高维随机回归模型，该模型将可观测因素引入模型中，且未对回归项和潜在因素回归过程施加平稳条件，但模型的不足在于使用线性估计拟合可观测因子部分。传统的线性估计对非线性数据和其他具有更复杂内在关系的数据拟合效果不佳，导致参数估计效果受到影响。近年来随着机器学习的发展和应用，各种机器学习模型在非线性数据建模方面具取得了一系列的成功，因此本文考虑在 Chang 等（2015）的基础上，对可观测因子部分采用机器学习方法进行估计，得到估计后的残差部分，然后基于此残差部分对潜在因子过程进行因子估计，并据此构造了基于机器学习方法的潜在因子模型。模型在可观测因子部分引入的机器学习方法能够显著地提升其对于非线性数据的拟合能力，进而提升对残差部分的预测准确性，从而提升后续的参数估计准确性。

文章对上述基于机器学习的潜在因子模型以潜在因子个数 r 的识别准确率和未知因子载荷矩阵**A**的估计效果为评价标准进行数值模拟实验，设计了一个线性数值模拟部分和两个非线性数值模拟部分。结果发现，基于机器学习的潜在因子模型在这三个数值模拟部分相较于基于线性估计和多项式估计的潜在因子模型表现更好，特别是在非线性数值模拟部分，参数估计更加准确，且更加稳定。通过对 2015 年 1 月至 2022 年 12 月的国证 300 指数相关行业股票日收益率数据进行实际数据分析发现，基于机器学习的潜在因子模型能很好地对参数进行估计，而且在预测效果对比部分，无论是整体预测还是分行

业预测，其效果都更好。

关键词：高维时间序列数据；潜在因子模型；机器学习；股票日收益率预测

Abstract

With the further development of the information age, various data are widely used in all walks of life. This paper studies high-dimensional time series data. The main feature of high-dimensional data is that the data dimension is large, and there are difficulties in model identification and parameter estimation. In order to better model and estimate the high-dimensional time series data, this paper introduces the machine learning model into the potential factor model, and uses data such as daily stock returns to analyze the relevant markets of the National Securities 300 Index.

Chang et al.(2015) proposed a high-dimensional random regression model with potential factors, endogenous and nonlinear. The model introduced observable factors into the model, and did not impose stationary conditions on the regression term and potential factor regression process. However, the disadvantage of the model is to use linear estimation to fit the observable factors. The traditional linear estimation has poor fitting effect on nonlinear data and other data with more complex internal relations, resulting in the impact of parameter estimation. In recent years, with the development and application of machine learning, various machine learning models have achieved a series of successes in nonlinear data modeling. Therefore, based on Chang et al.(2015)'s research, this paper estimates the observable factors using machine learning method to obtain the estimated residual part, and then estimates the potential factor process based on this residual part, Based on this, a potential factor model based on machine learning method is constructed. The machine learning method introduced in the observable factor part of the model can significantly improve its fitting ability for nonlinear data, thus improving the prediction accuracy of the residual part, and thus improving the subsequent parameter estimation accuracy.

This paper carries out numerical simulation experiments on the potential factor model based on machine learning with the identification accuracy of the number of potential factors r and the estimation effect of the unknown factor load

matrix as the evaluation criteria, and designs a linear numerical simulation part and two nonlinear numerical simulation parts. The results show that the potential factor model based on machine learning performs better in these three numerical simulation parts than the potential factor model based on linear estimation and polynomial estimation, especially in the nonlinear numerical simulation part, the parameter estimation is more accurate and more stable. Through the analysis of the actual data of the daily return rate of the stocks of the relevant industries of the National Securities 300 Index from January 2015 to December 2022, it is found that the potential factor model based on machine learning can well estimate the parameters, and in the comparison part of the prediction effect, both the overall prediction and the prediction by industry are better.

Key words: high-dimensional actual sequence data; Potential factor model; Machine learning; Forecast of daily stock yield

目录

1.绪论

1.1 研究背景与意义

1.1.1 研究背景

随着科学技术水平的发展，我们已经处于信息时代的快车道上，数据在人类社会中扮演着越来越重要的角色，各种类型的数据被广泛运用于各行各业中，本文所要研究的方法就是针对于现实生活中常见的时间序列数据。

时间序列数据是指按照一定的时间间隔排列的一组数据，其时间间隔可以取不同的单位，例如周、天、小时、分、秒。而时间序列分析就是为了从这些时序数据中提取有价值的统计信息，从而达到总结历史规律和预测未来发展趋势等目的。时间序列数据的应用场景非常广泛，在经济学、医学、天文学、海洋学等方面都有很高的使用频率。

在当今大数据的背景之下，随着储存能力和计算能力的提升，大量或巨量的高维数据出现在生产生活之中，高维数据的主要特点是样本量通常较小，而其数据维度十分巨大，这就导致在模型识别和参数估计等方面的困难。另一方面，经济、社会、自然现象的面板数据研究、金融市场分析、通信工程等各种实际问题都对高维时间序列的建模和预测提出了需求。当时间序列数据具有中高维度时，如何对其进行建模分析始终是一个挑战。

今年年初，中国人民银行、市场监管总局、银保监会、证监会四部门印发《金融标准化“十四五”发展规划》。规划提出，到 2025 年，与现代金融体系相适应的标准体系基本建成，金融标注化的经济效益、社会效益、质量效益、和生态效益充分显现，标准化支撑金融业高质量发展的地位和作用更加凸显。报告中的一个重点便是科技赋能及数字化转型，提议银行业金融机构应紧跟最新科技动态，持续探索科技赋能，有序推进在产品与服务、合规与

风控、管理与运营等方面创新应用的落地，在降本增效的同时提升核心竞争力。

股市作为金融行业中的重要一环，对股市的数据进行建模分析具有较大现实意义和应用前景，本文所使用的股市数据就属于高维时间序列数据，本文所采用的模型可以视为是 Chang 等（2015）的发展，其在将高维向量时间序列表示为低维潜在因子的过程中，引入三个新的特征，第一个特征是给因子模型增加了一个回归项，第二是未对回归量和潜在因子过程施加平稳条件，最后重点研究了带非线性回归的因子模型。本文在该研究的基础上，加入机器学习相关方法，形成基于机器学习方法的潜在因子模型。

机器学习是人工智能的一个分支，它的基本思想是从样本数据中学习得到知识，然后用于实际推断和决策，它和普通程序的一个显著区别是需要样本数据，是一种数据驱动的方法。机器学习目前被广泛应用于模式识别、计算机视觉、语音识别、自然语言处理等领域，其主要包含决策树、随机森林、支持向量机、卷积神经网络、递归神经网络等模型，它们被用于不同的任务。机器学习能够适用于多种数据类型，能够拟合更加复杂的内在关系。

将机器学习模型引入因子模型增强了模型的拟合能力和泛化能力，也增加了模型的应用场景。

1.1.2 研究意义

首先，从研究方法来看，使用改进后的因子模型对以下两方面的问题起到了针对性的作用：（1）在降低数据维度、参数估计方面，因子模型具有较大的优势，对于高维的时间序列数据，因子模型将高维的时间序列过程用潜在的因子模型来表征，这样就达到了降低维度的目的。同时由于高维数据在参数估计过程中往往会出现维度灾难的问题，即随着维度的增加，估计参数所需要的样本呈指数级增加，从而使得参数估计难以进行。在这种背景下，改进后的因子模型巧妙地将高维时间序列过程表示为一些低维的潜在因子过程，这样就大大减少了需要估计的参数数量，使得参数估计得以进行。（2）本文对所使用的因子模型进一步拓展，将机器学习方法引入了因子模型，从而更好地拟合改进后因子模型中的可观测部分。拓宽了模型的适用范围，使得模

型能在现实中有更多的应用场景。

其次，本文所使用的模型符合股票收益率数据的特性。一方面，由于股票的收益率随着时间的变化而变化，它是一个时间序列数据，同时市场上的股票的数量众多，这多只股票所形成的数据又符合了高维性的特点，其构成的数据就是高维时间序列数据。另一方面，改进后的因子模型包含了可观测过程和潜在因子过程，在现实中，股票收益率受到许多因素的影响，例如企业经营成绩、政策环境、宏观经济运行情况等等，这些因素可以作为可观测过程进入模型，然后本文通过增加机器学习算法对可观测过程部分的计算进行改进，增加了模型的稳健性。

最后，加入机器学习模型具有实际意义。随着大数据时代的到来，我们能够获取到的数据越来越多，数据所呈现出的关系也越来越复杂，简单的线性回归已经很难精确地刻画数据背后所隐藏的规律。基于机器学习的潜在因子模型受约束少，对数据的分布一般不做任何要求，且其适应能力强，稳健性高。

1.2 研究内容与框架

1.2.1 研究内容

股票的收益率作为衡量股票业绩表现的重要指标之一，受到许多因素的影响，且股票收益率数据属于高维的时间序列数据，高维时间序列数据的分析和建模充满了挑战和机遇。本文以 Chang 等(2015)建立的具有潜在因子、内生性、非线性的高维随机回归模型为基础，该模型主要包含三个部分，第一部分是可观测的线性回归部分，第二部分是潜在因子过程，第三部分为随机扰动项，本文在该模型的基础上进行了深入的研究，主要的研究在于给该模型第一部分可观测模块引入机器学习方法，这样就提高了模型的适用范围以及稳健性。本文的具体研究内容就包含对引入机器学习的因子模型进行数值模拟，通过模拟研究模型效果，然后将模型运用于股票收益率的预测当中，对模型的适用能力进行评估。

1.2.2 研究框架

本文的研究基于 2015 年 1 月 5 日至 2022 年 12 月 1 日的 165 支股票日收益率数据，以及国证行业指数数据。由于以往因子模型的可观测部分主要采用线性回归模型，本文将机器学习方法引入模型中，探究其模型表现。主要研究内容分为以下几个部分：

第一章为绪论部分，介绍了本文的研究背景和研究意义，说明股票日收益率预测对现实生活的意义，总结研究的内容和框架。

第二章为文献综述部分，主要介绍三个方面的研究情况，首先介绍因子模型研究概况，然后介绍机器学习模型研究现状，最后介绍股票日收益率相关研究进展

第三章为模型构建。在对基于线性估计的潜在因子模型进行分析后，建立引入机器学习方法的潜在因子模型。

第四章为数值模拟。建立一个线性模拟部分和两个非线性模拟部分，针对这三个部分对比基于线性估计、基于多项式估计以及基于机器学习的潜在因子模型，分别得到参数估计值并评估模型实际表现。

第五章为实际数据分析。基于股票日收益率数据以及国证行业指数数据，对数据进行描述性分析，然后使用基于线性估计、基于多项式估计以及基于机器学习的潜在因子模型对该数据参数进行估计，对比模型效果，探究行业差异。

第六章为总结与展望。总结本文在因子模型的进一步改进和股票收益率预测研究中取得的结果，同时分析研究中的不足之处，为后续的研究提供思路。

本文的主要组织结构如图 1-1 所示。

1.3 创新点

本文研究的是基于机器学习方法的潜在因子模型，其主要创新点有以下三点：

（1）本文对具有潜在因子、内生性、非线性的高维随机回归模型进行了进一步的改进，针对其在可观测因子部分对非线性数据估计效果不佳的问题，

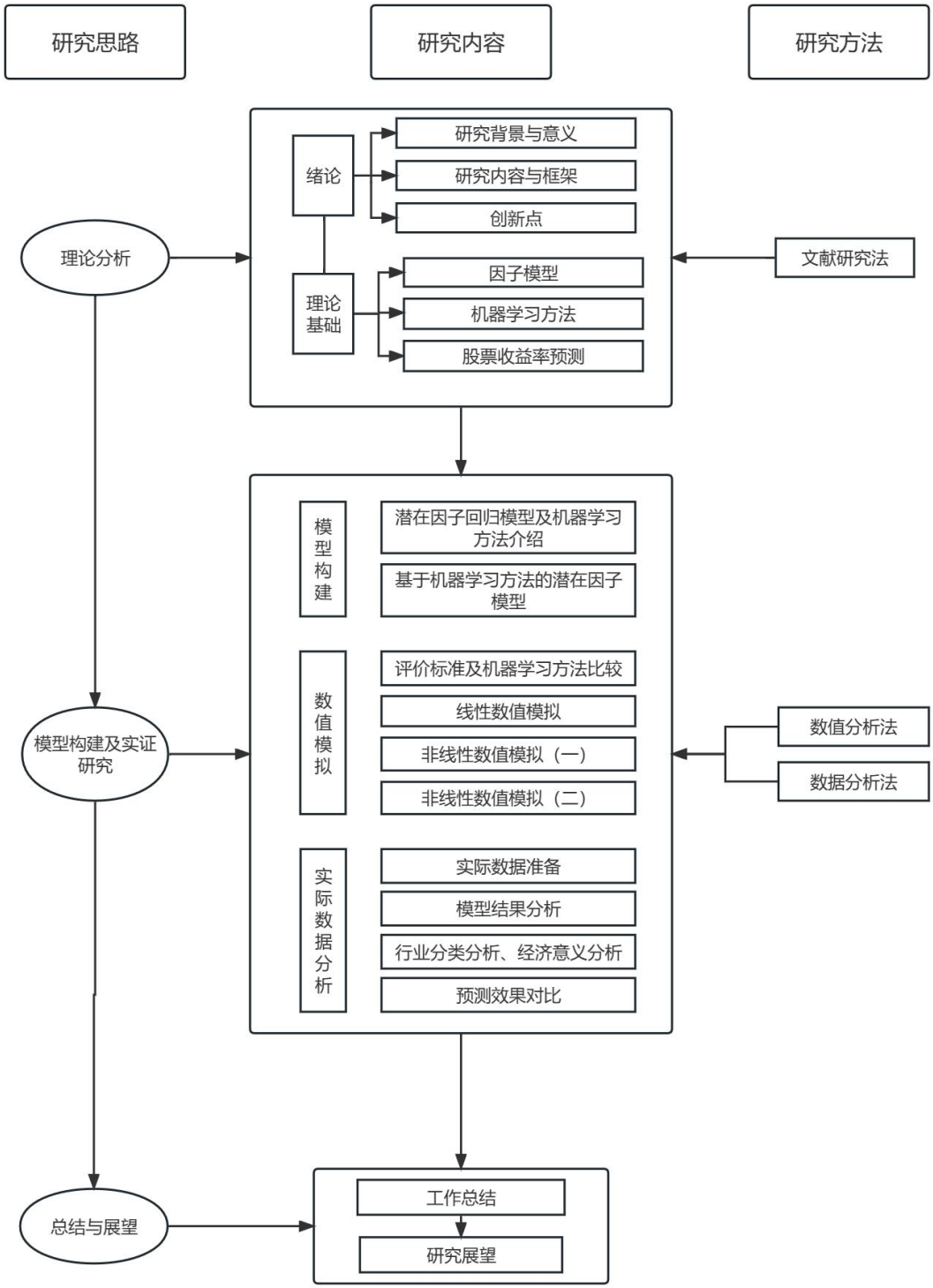
研究思路
研究内容
研究方法
绪论
研究背景与意义
研究内容与框架
创新点
理论基础
因子模型
机器学习方法
股票收益率预测
理论分析
文献研究法
模型构建
潜在因子回归模型及机器学习方法介绍
基于机器学习方法的潜在因子模型
数值模拟
评价标准及机器学习方法比较
线性数值模拟
非线性数值模拟（一）
非线性数值模拟（二）
实际数据分析
实际数据准备
模型结果分析
行业分类分析、经济意义分析
预测效果对比
模型构建及实证研究
数值分析法
数据分析法
总结与展望
工作总结
研究展望

图 1-1：论文组织结构

将机器学习方法引入模型之中，利用机器学习方法估计可观测因子部分，得到残差估计后对潜在因子部分进行因子估计，既提升了参数估计的准确性，同时也提高了模型的拟合能力和泛化能力，使得模型的应用场景也有所拓宽。

（2）在数值模拟部分，本文设计了一个线性模拟部分和两个非线性数值模拟部分。总体看来，与基于线性估计和多项式估计的潜在因子模型相比，基于机器学习的潜在因子模型在对潜在因子个数 r 和因子载荷矩阵 $\mathbf{A}$ 二者的预测上都取得了更好的效果，能够对非线性特征的数据进行更好的拟合。

（3）在实际数据分析部分，本文将改进后的模型运用于股票收益率数据上，然后对模型的预测准确率进行探究，发现在整体数据集上，基于机器学习方法的潜在因子模型预测的 RMSE 最小，预测效果最好，对于分行业的数据集，11 个行业中的 9 个基于机器学习方法的潜在因子模型预测效果最佳。这也体现了基于机器学习方法的潜在因子模型在实际数据中相较于已有方法的提升。

2.文献综述

2.1 因子模型

目前，具有中大维度的时间序列数据的建模仍然是一个挑战。高维时间序列数据的降维成为了研究的重点之一，降维的主要目的是使高维时间序列模型的参数估计得以进行，同时提取时间序列数据的重要信息。传统的时间序列建模方法主要分为：分段线性近似、符号表示法、离散小波变换等。

分段线性近似(Piecewise Linear Representation, PLR)是数据表征的方法之一，这种方法已被许多研究者用于时间序列数据的聚类、分类、索引和关联规则的挖掘。廖俊等（2010）总结认为分段线性近似的思想是将原有的时间序列数据用多条首尾相邻的直线段代替表示，如何确定分段点是该算法的核心。Keogh 等（2002）提出 PLR 的主要分支算法又包含三种：滑动窗口算法、自上向下算法、自下向上算法。滑动窗口算法的工作原理是，在时间序列的第一个数据点上锚定一个潜在段的左侧点，然后尝试用越来越长的段向右侧近似该数据。自上向下算法的工作原理是考虑时间序列的每一种可能的划分，并在最佳位置进行分割。自下向上算法是自上向下算法的自然补充。该算法首先创建时间序列的最佳近似，因此使用$n/2$ 段来近似n长度的时间序列。接下来，计算每对相邻段合并的成本，算法开始迭代合并成本最低的对，直到满足停止条件。谢婷玉等（2018）在分段线性近似方法的基础上，提出了基于重要点双重评价因子的时序趋势提取算法．该算法首先定义重要点作为时间序列分段点的备选集，并给出两种重要点评价因子的定量计算方式，用距离因子度量其相对差异程度，用趋势因子在全局上度量其对整体趋势的影响程度，综合评价每个重要点对整体趋势的重要程度来选取分段点。该方法有效削弱了噪声干扰，且提高了提取的精度。

Lin 等（2007）最早提出符号表示法(SAX)，SAX 算法的主要做法就是将时间序列进行符号化表示。在保持了原方法的低复杂度的情况下，在范围查询中提供了令人满意的灵敏度和可选择性。SAX 算法是时间序列的一种新的符号表示。这种新的表示方法使得维数降低，此外，还允许用符号方法来定义距离度量，后一个特性让某些数据挖掘算法可以在符号表示的基础上运行，例如聚类、查询、异常检测等数据挖掘任务。Zhong 等（2008）为了改善 SAX 算法对时序信息描述不完整的缺陷，提出了基于统计特征的时序数据符号化算法。相比于 SAX 算法而言，改进后的算法将时序符号看作矢量，而各时序子段的均值和方差则分别作为描述其平均值及发散程度的分量。该算法相较于 SAX 算法而言得到的结果更为准确。黄俊杰等（2022）提出了一种融合形态趋势信息的时间序列符号聚合近似方法。这个方法的主要改进在于通过一种新的趋势指标来对子序列段的趋势进行拟合，并用改进后的指标的符号限量对时间序列进行表征。该方法在某些数据集上进行了实验，实验结果表面改进后的方法相比 SAX 算法有更好的分类预测效果，提升了 SAX 算法对时间序列数据局部趋势的信息提取能力。

离散小波变换(DWT)是对基本小波的的尺度和平移进行离散化。Chan&Fu（1999）深入研究了 DWT 的应用，提出了使用 Haar 小波变换进行时间序列索引，且通过实验证明了 Harr 变换的性能优于离散傅里叶变换。

本文使用的模型是由 Sims（1980）提出的向量自回归模型(VAR)不断发展而来的。向量自回归模型是计量经济学中的一个常用模型，Lam 等（2011）研究的基于公因式的高维时间序列降维问题。该研究从降维的角度出发，采用了一种不同的统计方法，将p维时间序列分解为两个部分：低维因素驱动的动态部分和作为向量白噪声的静态部分。随后 Lam 等（2012）又进一步研究了该模型的理论性质。Chang 等（2015）在此模型的基础上，提出了一种具有潜在因素、内生性和非线性的高维随机回归模型。该模型的主要贡献有以下三个方面：首先是给模型增加了一个回归项，该回归项可以将可观测因素引入模型中；然后该模型未对回归项和潜在因素回归过程施加平稳条件，这大大扩展了模型的应用场景，因为在实际问题中的许多重要因素不是固定的；最后重点研究了带非线性回归的因子模型，通过将非线性回归函数表示为若干基函数的线性组合，将问题转化为具有大量线性回归量的模型。本文就是

在 Chang 等（2015）的模型基础上进一步进行研究，将非参数回归和深度学习引入模型的回归项之中。

2.2 机器学习方法

机器学习方法一般包括监督学习、无监督学习、强化学习等。监督学习是指从标注数据中学习预测模型的机器学习问题，标注数据表示输入输出的对应关系，预测模型对给定的输入产生相应的输出，监督学习的本质是学习输入到输出的映射的统计规律。无监督学习是指从无标注数据中学习预测模型的机器学习问题，无标注数据是自然得到的数据，预测模型表示数据的类别、转换或概率。无监督学习的本质是学习数据中的统计规律或潜在结构。强化学习是指智能系统在环境的连续互动中学习最优行为策略的机器学习问题。目前主流的机器学习模型包括决策树、随机森林、梯度提升树（GBDT）、支持向量机、KNN 算法等等，由于本文主要使用了决策树、随机森林、GBDT 等模型，故主要对这几个模型的研究现状进行梳理。

决策树算法是一个预测模型，他代表的是对象属性与对象值之间的一种映射关系。决策树算法被广泛应用于各个领域，王琪（2010）结合供应链金融的特点，建立基于决策树的供应链金融模式信用风险评估体系，使得商业银行可以更好地从供应链角度评估小企业信用。魏信等（2010）利用中尺度分辨率的 ETM+TM5 影像,使用决策树分类技术对北京市城区土地利用分类进行研究,得到北京市中心城区土地利用分布影像。一些研究提升了决策树模型的性能，Narish 等（2011）提出了一种学习倾斜决策树的新算法，该算法使用了一种评估超平面的策略，在决策树的每个节点，找到两个类的聚类超平面，并使用其角平分线作为该节点的分割规则，提升了小型决策树的性能。胡宸等（2015）提出基于决策树数据挖掘技术的设计空间优化方法,对原始设计域进行划分,缩小优化算法对初始设计点的起始范围,提高了优化算法的设计效率。

随机森林模型采用集成学习思想，将许多棵决策树整合成随机森林，通过平均得到最终结果。随机森林相较于决策树在预测准确性和防止过拟合方面有一定的提升。于晓红等（2016）根据已有的非均衡少量样本,分别采用随机

森林分类和回归算法进行建模,模型对各级风险样本的识别正确率均达到了100%,具有很好的实用价值和预测能力。袁敏等（2009）基于蛋白质的氨基酸序列,将组合离散增量和伪氨基酸组分信息共同作为预测参数,采用随机森林分类器,对 8 类膜蛋白进行了预测。王爱平等（2011）提出了一种增量式极端随机森林分类器(incremental extremely random forest,简称 IERF),用于处理数据流,特别是小样本数据流的在线学习问题.

GBDT 模型全称为 Gradient Boosting Decision Tree，GBDT 是是使用了向前分布算法的加法模型。组成 GBDT 的弱学习器被限定为 CART 回归树。对于提升树而言，其核心思想为残差拟合，GBDT 梯度提升树使用损失函数的负梯度作为提升树算法中残差的近似值。张元平等（2013）在基于 HMM 的 Trainable TTS 的语音生成后端,引入 GBDT 算法,分别应用在频谱、基频、时长三个维度的语音参数上，发现应用在 LSF 频谱模型聚类上,系统的主观倾向性得分提高了 15.6%。Wang 等（2016）提出了一种基于 LR 算法和 GBDT 算法的融合模型，用于在移动环境中推荐垂直行业商品，其 F1-score 与基线模型相比提高了 2%-36%。Yang 等（2017）提出了一种集成学习方法——梯度增强决策树（GBDT），用于基于环路检测器收集的高速公路交通量数据进行短期交通预测，其预测精度一般高于 SVM 和 BPNN。马文臻等（2022）基于梯度提升决策树原理构建卫星工程参数异常智能检测方法,利用量子科学实验卫星任务的工程数据开展应用验证与分析,与原采用的"阈值+规则表达式"异常检测方法相比,将平均准确率提升了约两个百分点,达到 98%以上,可有效减少漏报和错报,并将检测速度提升了大约 6 倍。

2.3 股票收益率预测

随着时代的发展，我国的资本市场也愈发的成熟，资本市场为我国的经济发展也作出了很大的贡献。对于资本市场来说，股票收益率是衡量一家公司或一个行业经营状况或发展水平的重要指标，股票收益的率的计算公式比较简单，其值等于收益额和原始投资额的比值。但由于股票市场的复杂性，股票收益率这一指标受到许多因素的影响，基于此，探索稳定高效的预测模型就具有很大的理论意义和实践价值。目前，对于股票收益率的预测主要有两

个方向，第一个方向是通过传统的计量经济学模型对收益率进行预测，第二个方向是将机器学习算法引入到股票收益率的预测之中。

计量模型在股票预测建模方面通常假定金融时间序列满足某些前提条件，在这些条件满足的情况下，模型的可靠性才得到相应的保证。Min（2005）选取产品的市场占有率作为预测对象，建立马尔可夫模型，并进行了实例分析。Shi（2005）采用自回归求积移动平均法对《上海市统计年鉴 2002》提供的固定资历产投资额资料进行了分析。结果显示 ARIAM 模型提供较准确的预测效果可用于未来的预测并为上海市全社会固定资产投资提供可靠依据。徐珺（2017）构建了基于 ARMA 模型的黄金期货价格预测方法。在对黄金期货价格时间序列作平稳化的基础上，检验并探究该时间序列适合的模型，然后进行了相关检验。结果表明 MA 序列在该模型上表现更佳，静态预测方法表明该模型相对平均误差为 1.69%，模型具有一定的现实价值。王莉（2020）选取 1991 年 4 月 3 日至 2020 年 1 月 10 日期间的深证综合指数收盘价为样本数据，对样本数据进行对数处理，然后进行一阶差分处理以保证其序列平稳性。发现处理后的序列具有尖峰厚尾性、异方差性以及杠杆效应等特点。最后基于这些特点建立了 ARMA－GARCH 和 ARMA－EGARCH 两种模型分别对序列进行拟合、描述及分析。比较之后确定 ARMA(1,1)－GARCH(1,1)模型拟合效果更优。Zhang 等（2018）运用 MF-X-DMA 方法对“一带一路”倡议中三个股指在全球和局部时期多重分形特征进行了研究，并对该特征进行了进一步的分析。研究表明,中国股市与其他三个股市之间的交叉相关性是多重分形的，相较于其他的股市，中国股市存在比短期行为更多的长期交易行为。且在“一带一路”倡议提出之后，多重分形的特征与之前的特征相比持久性得到了加强。

股票收益率预测的第二个方向是采用机器学习方法。Wang 等（2020）认为人类在成长、生活中积累很多历史经验“数据”，利用这些数据，人类总结出生活的“规律”，也就是得到了经验，当人类遇到一些未知的问题时，就可以通过这些经验对未来进行推测，从而指导自己的生活和工作。机器学习就类似于人类思考的过程，机器学习的核心思想在于通过建立模型，提取数据中潜在信息，然后使用模型进行预测。而模型“训练”与“预测”就可以对应到人类生活中的总结和推测过程。机器学习应用的场景十分广泛，诸

如模式识别、统计学习、数据挖掘、计算机视觉、语音识别、自然语言处理等领域。近年来，股票收益率预测也使用到了很多机器学习和深度学习的模型，与传统的计量经济学模型相比，机器学习方法在假设条件这一方面的限制更少，且具有很强的处理非线性和非平稳特征问题的能力。

例如，Jigar Patel 等（2015）研究比较了四种模型在预测印度股票市场的走势和股价指数的表现。四种预测模型分别为神经网络、支持向量机、随机森林、朴素贝叶斯，在数据输入方法的选择上，使用了两种不同的方式。第一种方式是使用原始股票交易数据计算十个技术参数，第二种方式侧重于将这些技术参数表示为趋势确定性数据。然后分别比较这两种数据输入方式在不同模型上的表现。实验结果表明，对于第一种数据输入方式，随机森林模型表现最优。对于第二种数据输入方式，各模型表现不佳。张鹏等（2022）提出了一种基于机器学习方法的两步骤多元化投资组合优化模型。该模型首先通过 XGboost、支持向量回归、K 近邻算法筛选得到预测收益率较高的股票，并评估相应的模型。然后进行投资组合优化，采用均值-下半方差模型、均值-方差模型和等比例模型确定所选股票的投资比例。实证结果表明，XGboost+MSV 的模型表现最佳，其收益率和风险控制能力均为最优。Chandar（2021）利用神经网络模型提高股市预测精度，基于反向传播神经网络(BPNN)、径向基函数神经网络(RBFNN)、时滞神经网络(TDNN) 3 种神经网络以及遗传算法(GA)、粒子群算法(PSO)、人工蜂群算法(ABC)等自然启发算法的潜力,提出了 9 种新的日内股票价格预测集成模型。分别命名为 GA-BPNN、PSO- BPNN、ABC- BPNN、GA-RBFNN、PSO-RBFNN、ABC-RBFNN、GA-TDNN、PSO-TDNN、ABC-TDNN。采用自然启发算法优化 ANNs 参数。由历史数据计算出的技术指标作为模型的输入。提出的混合模型在四个数据集上得到验证。采用均方根误差(RMSE)、击中率(HR)、错误率(ER)和预测精度 4 个统计指标来衡量模型的性能。结果表明，PSO - BPNN 模型对日内股票价格的预测精度最高。

基于现有的研究，本文将机器学习方法引入到因子回归模型中，在潜在因子模型的基础上，通过机器学习方法的预测能力减小预测残差，提高模型的拟合能力和准确度。

3.模型构建

3.1 背景介绍

股票的收益率受到许多因素的影响，这些因素可以分为可观测因子和不可观测因子。可观测因子指那些我们在实际生活中可以观测到的因子，例如大盘指数等，而不可观测因子指那些我们无法量化的影响因子，它们可能是一些政策因素的影响或一些细微的难以统计的影响因子。本章节将首先对潜在因子回归模型以及机器学习方法进行介绍，然后对本文建立的基于机器学习方法的潜在因子模型进行进一步阐释。

3.2 潜在因子回归模型

3.2.1 模型设定

本文使用的模型是潜在因子模型的进一步发展，而潜在因子模型是在 Lam 和 Yao（2011）的模型基础上发展得到的。Lam 和 Yao（2011）将高维向量时间序列表示为低维潜在因子过程和向量白噪声过程的线性相加。Chang 等（2015）提出的潜在因子模型就是在该模型的基础上，进一步的发展。潜在因子回归模型的主要改进有三个方面。

第一个方面是在原模型的基础上增加了一个回归项。这个回归项所代表的就是可观测的因子的影响，这是一个有力的补充，例如现实问题中，PM2.5 浓度受到各种污染物和气象条件的影响，居民消费价格指数受到粮食、能源供给等影响,在预测时就可以通过可观测因子进入模型中。第二个方面是未对可观测因子和潜在因子过程施加平稳条件，这有效地扩大了应用范围，因为在现实中会出现许多非平稳的情况。第三个方面是研究了带非线性回归项的

模型。

潜在因子回归模型的基本形式如下：

$$\mathbf{y}_t = \mathbf{D}\mathbf{z}_t + \mathbf{A}\mathbf{x}_t + \boldsymbol{\varepsilon}_t,$$

其中$\mathbf{y}_t$和$\mathbf{z}_t$是可观测的$p \times 1$维和$m \times 1$维时间序列，$\mathbf{x}_t$是$r \times 1$维的潜在因子过程，$\boldsymbol{\varepsilon}_t$是白噪声，$\boldsymbol{\varepsilon}_t \sim \mathrm{WN}(0, \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}})$，且$\boldsymbol{\varepsilon}_t$与$(\mathbf{z}_t, \mathbf{x}_t)$无关，$\mathbf{D}$是未知的系数矩阵，$\mathbf{A}$是未知的因子载荷矩阵。潜在因子的个数$r$是一个未知的固定常数。当有观测值$\{(\mathbf{y}_t, \mathbf{z}_t): t = 1, \ldots, T\}$时，目标就是估计$\mathbf{D}$、$\mathbf{A}$和$r$，此模型一般适用于$p$大于$T$的高维模型。

3.2.2 参数估计

从形式上看，对$\mathbf{D}$的参数估计可以看作是一个标准的最小二乘问题,

$$\mathbf{y}_t = \mathbf{D}\mathbf{z}_t + \boldsymbol{\eta}_t, \boldsymbol{\eta}_t = \mathbf{A}\mathbf{x}_t + \boldsymbol{\varepsilon}_t.$$

这里 $\mathrm{cov}(\boldsymbol{z}_t, \boldsymbol{\eta}_t) = 0$，将$\mathbf{D}$表示为$\mathbf{D} = (\mathbf{d}_1, \ldots, \mathbf{d}_p)^{\mathrm{T}}$，那么$\mathbf{D}$的最小二乘估计可以表示为以下形式：

$$\mathbf{D} = (\widehat{\mathbf{d}}_1, \ldots, \widehat{\mathbf{d}}_p)^{\mathrm{T}},$$

$$\widehat{\mathbf{d}}_i = \left(\frac{1}{T}\sum_{t=1}^{T} \mathbf{z}_t \mathbf{z}_t{}^{\mathrm{T}}\right)^{-1} \left(\frac{1}{T}\sum_{t=1}^{T} y_{i,t}\, \mathbf{z}_t\right),$$

其中，$y_{i,t}$是$\boldsymbol{y}_t$的第i个元素。

与 Lam 等（2011）的估计方法类似，对因子载荷空间$\boldsymbol{\mathcal{M}}(\mathbf{A})$的估计基于残差的估计$\widehat{\boldsymbol{\eta}}_t$，$\widehat{\boldsymbol{\eta}}_t = \boldsymbol{y}_t - \widehat{\mathbf{D}}\boldsymbol{z}_t$。这里为了进一步的说明估计的步骤，引入一些公式，

$$\boldsymbol{\Sigma}_x(k) = \frac{1}{T-k}\sum_{t=1}^{T-k} \mathrm{cov}(\mathbf{x}_{t+k}, \mathbf{x}_t),$$

$$\boldsymbol{\Sigma}_{x\varepsilon}(k) = \frac{1}{T-k}\sum_{t=1}^{T-k} \mathrm{cov}(\mathbf{x}_{t+k}, \boldsymbol{\varepsilon}_t),$$

$$\boldsymbol{\Sigma}_{\eta}(k) = \frac{1}{T-k}\sum_{t=1}^{T-k} \mathrm{cov}(\boldsymbol{\eta}_{t+k}, \boldsymbol{\eta}_t).$$

例如，当$\mathbf{x}_t$是平稳序列时，$\boldsymbol{\Sigma}_x(k)$就是$\mathbf{x}_t$滞后k阶的自协方差矩阵，对于任

意的$k \neq 0$，计算可得，

$$\boldsymbol{\Sigma}_{\eta}(k) = \mathbf{A}\boldsymbol{\Sigma}_{x}(k)\mathbf{A}^{\mathrm{T}} + \mathbf{A}\boldsymbol{\Sigma}_{x\varepsilon}(k).$$

对于一个指定的固定正整数$\bar{k}$，定义，

$$\mathbf{M} = \sum_{k=1}^{\bar{k}} \boldsymbol{\Sigma}_{\eta}(k)\,\boldsymbol{\Sigma}_{\eta}(k)^{\mathrm{T}},$$

假设$\mathbf{M}$的秩为r，这是因为潜在因子过程$\mathbf{x}_t$是r维的。对于任何向量$\mathbf{b}$，若$\mathbf{b}$与因子载荷空间$\boldsymbol{\mathcal{M}}(\mathbf{A})$正交，即$\mathbf{b} \perp \boldsymbol{\mathcal{M}}(\mathbf{A})$，都有，

$$\mathbf{M}\mathbf{b} = 0,$$

因为在$\boldsymbol{\mathcal{M}}(\mathbf{A})$不变的情况下，$\mathbf{A}$的选择几乎是任意的，故我们可以取$\mathbf{M}$非零特征值对应的标准正交特征向量作为$\mathbf{A}$的列，设$\mathbf{A} = (\mathbf{a}_1, \ldots, \mathbf{a}_r)$，$\mathbf{a}_1, \ldots, \mathbf{a}_r$为前$r$大个特征值满足$\lambda_1 \geq \cdots \geq \lambda_r > 0$对应的标准正交特征向量。由于$\mathbf{A}$是由各标准正交特征向量构成的矩阵，$\mathbf{A}$是满足$\mathbf{A}\mathbf{A}^{\mathrm{T}} = \mathrm{I}_r$半正定的矩阵。当$\mathbf{M}$的$r$个非零特征值不同时，若我们忽略$\mathbf{a}_j$和$-\mathbf{a}_j$的替换，$\mathbf{A}$是唯一的。我们可以通过如下的方式估计$\mathbf{A}$，

$$\widehat{\boldsymbol{\eta}}_t = \mathbf{y}_t - \widehat{\mathbf{D}}\mathbf{z}_t,$$

$$\widehat{\boldsymbol{\Sigma}}_{\eta}(k) = \frac{1}{T-k}\sum_{t=1}^{T-k} (\widehat{\boldsymbol{\eta}}_{t+k} - \overline{\boldsymbol{\eta}})(\widehat{\boldsymbol{\eta}}_t - \overline{\boldsymbol{\eta}})^{\mathrm{T}},\ \overline{\boldsymbol{\eta}} = \frac{1}{T}\sum_{t=1}^{T} \widehat{\boldsymbol{\eta}}_t\,.$$

以上的估计式自然而然地形成了对$\mathbf{A}$的估计，记$\mathbf{A}$的估计式$\widehat{\mathbf{A}} \equiv (\hat{\mathbf{a}}_1, \ldots, \hat{\mathbf{a}}_r)$。这里$\hat{\mathbf{a}}_1, \ldots, \hat{\mathbf{a}}_r$就是$\mathbf{M}$的估计$\widehat{\mathbf{M}}$的前$r$大个特征值满足$\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_r > 0$对应的标准正交特征向量。我们所定义$\widehat{\mathbf{M}}$的计算式如下，

$$\widehat{\mathbf{M}} = \sum_{k=1}^{\bar{k}} \widehat{\boldsymbol{\Sigma}}_{\eta}(k)\,\widehat{\boldsymbol{\Sigma}}_{\eta}(k)^{\mathrm{T}}.$$

以上的这些估计都是基于r已知的条件，在现实情况中r往往是未知的，故我们还需要对r进行估计，这是关键的一步。本文中采用 Chang 等（2015）一样的方法，通过比率估计器来对潜在因子个数r进行估计，其计算式如下，

$$\hat{r} = \operatorname{argmin}\left\{\frac{\hat{\lambda}_{j+1}}{\hat{\lambda}_j} : 1 \leq j \leq R\right\},\ \text{其中}\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_p,$$

在估计中一般取$R = p/2$。

3.2.3 模型不足

潜在因子回归模型存在的主要问题是模型形式较为单一，采用多元线性回归的方式对参数进行估计，对数据的拟合能力存在偏弱，若数据内在的关系不符合多元线性回归的方式，其预测效果将受到一定的影响。例如在现实中，在拟合可观测因子和应变量之间的数量关系时，由于现实世界的复杂性，其内在关系往往会较为复杂，简单的多元线性回归效果不佳，这又进一步的导致了残差 η_t 估计的精度下降。而估计得到的残差是我们对潜在因子回归个数r和因子载荷矩阵$\mathbf{A}$进行估计的基础，这就会对潜在因子回归个数r和因子载荷矩阵$\mathbf{A}$的估计造成影响，使得参数估计的准确度下降。

3.3 机器学习方法介绍

3.3.1 决策树模型

为了解释 GBDT(Gradient Boosting Decision Tree)梯度提升树模型的原理，就必须先介绍决策树的概念。GBDT 是一种基于决策树的集成算法。决策树是一种基本的分类和回归的方法，决策树模型正如其名字所言，是一种树形结构的模型，其中每个内部节点表示一个属性上的测试，每个分支代表一个测试输出，每个叶子节点代表一种类别。

决策树是用一系列分支语句表示的模型，决策树按任务可以分为分类树与回归树，按特征选择方法主要有三种，ID3 决策树算法、C4.5 决策树算法、CART 决策树算法。CART 决策树（Classification And Regression Tree）既可以用于分类也可以用于回归，由于本文的模型目的是回归预测，故主要介绍适用于回归任务 CART 决策树算法，CART 决策树的基本构造方法如下：

1.初始化特征集合和数据集合；

2.计算数据集合信息熵和所有特征的条件熵，选择信息增益最大的特征作为当前决策节点；

3.更新数据集合和特征集合（删除上一步使用的特征，并按照特征值来划分不同分支的数据集合）；

4.重复 2，3 两步，若子集值包含单一特征，则为分支叶子节点。

CART 决策树算法的划分标准是 GINI 指数，GINI 指数可以衡量数据划分的不纯度，它的值在 0 到 1 之间，GINI 指数的值越小，表明样本集合的纯净度越高，而 GINI 指数越大表明样本集合越杂乱。对于一个数据集T，将某一个特征作为分支标准，其第i个取值的 GINI 指数的计算公式如下：

$$gini(T_i) = 1 - \sum_{j=1}^{n} p_j^2 ,$$

其中n表示数据的类别数，p_j表示样本属于第 j 个类别的概率。

在计算出某个样本所有取值的 GINI 指数后，就可以得到 Gini Split Info:

$$Gini_{split}(T) = \sum_{i=1}^{2} \frac{N_i}{N} gini(T_i) ,$$

其中N表示样本总数，N_i表示属于特征第i个属性值的样本数。此时选择 GINI 指数最小的特征作为划分标准。

CART 决策树中回归树的具体构造采用自顶向下的贪婪式递归二分的方式。这里的贪婪，是指每一次划分只考虑当前最优，即每一次划分是在之前的基础上将某个区域一分为二。从数学上定义，设预估结果$y \in R$，特征向量为$X = [x_1, x_2, \ldots, x_p]$，回归树的两个步骤为：

1.把整个特征空间X切分为 2 个没有重叠的区域R_1, R_2，其中j是选择的第j个特征，s是划分标准，$x^{(j)}$是每个样本中第j个变量的取值，其计算公式如下，

$$R_1(j,s) = \{x|x^{(j)} \le s\}, \quad R_2(j,s) = \{x|x^{(j)} \ge s\},$$

2.计算这两个区域的的误差，并求解下式，

$$\min_{j,s}\left[\min_{c_1}\sum_{x_i \in R_1(j,s)}(y_i - c_1)^2 + \min_{c_2}\sum_{x_i \in R_2(j,s)}(y_i - c_2)^2\right],$$

其中，c_1、c_2是子集y_i的平均数，通过不断迭代上式，取不同的特征j和其对应的s，得到误差最小的值，作为划分的标准。

3.3.2 随机森林模型

随机森林模型是一种有监督学习算法，它利用多个决策树对样本进行训练、分类并预测。在对数据进行分类的同时，还可以给出各个变量的重要性评分，

评估各个变量在分类中所起的作用。随机森林模型中的“随机”主要包含两方面，第一是指随机森林在原始训练数据中有放回的选取等量的数据作为训练样本，二是在生成各个决策树的时候，从所有特征中随机抽取一部分。这两种随机进一步减小了决策树之间的相关性，提高了模型的准确性。

随机森林模型的具体实现过程如下，

1.从原始数据集中随机有放回地抽取 n 个样本生成 m 个训练集，然后分别在这 m 个训练集上生成 m 个决策树模型。

2.每棵树在生成的时候，从 K 个特征中随机选取 k 个特征，构成这棵树的特征集，再从该特征集中选取最优特征，构成树的子节点。

3.把生成的多棵决策树组成随机森林，对于分类问题，按照多棵分类树投票决定最终分类的结果；对于回归问题，由多棵树预测值的均值决定最终预测的结果。

3.3.3 GBDT 梯度提升树模型

（1）Boosting 算法

集成学习方法一般分为两种，Bagging 算法和 Boosting 算法。集成学习的优势就在于它可以和各种各样的其他分类、回归算法进行结合，达到提高模型准确率和稳定性的效果，同时，也可以降低方差，有效避免过拟合的发生。

Bagging 算法最早由 Leo Breiman（1994）提出，是一种在每个自助样本集上建立基分类器，最后通过投票指派得到测试样本最终类别的方法。其过程是从 m 个样本训练集中随机采样（有放回的重复随机抽取），然后得到 T 个采样集，在这 T 个采样集上分别训练弱学习器，最后通过一定的策略将这 T 个弱学习器集合在一起，形成强学习器。

Boosting 算法和 Bagging 算法的主要区别就在于多个学习器之间的生成规则。Bagging 算法的若学习器是相互独立的，弱学习器之间不存在什么联系，属于并行式关系，而 Boosting 算法的若学习器之间存在依赖关系，必须按照顺序迭代生成。Boosting 算法的过程为：

1.给训练集中每个样本建立权值w_i，表示对每个样本的权重，这个权重会

随着迭代的进行更新，对于被错误分类的样本，其权重在下一次迭代中将增大。

2.与此同时，对于那些分类准确率较高的弱分类器，增加其权重；对于分类准确率较低的弱分类器，减小其权重，使其在表决中所占比例降低。每迭代一次，都将得到一个弱分类器。

3.最后通过弱分类器权值将其结合起来，得到最终的模型。

本文使用的 GBDT 模型就采用了 Boosting 算法。

（2）GBDT 模型

GBDT 模型是以决策树为基函数的模型并以此成为提升树。提升树的提升方法为基函数的线性组合与向前分布算法。对于向前分布算法，模型见下式:

$$f(x) = \sum_{m=1}^{M} \beta_m b(x;\gamma_m),$$

其中，$b(x;\gamma_m)$为基函数，γ_m为基函数的参数，β_m为基函数的系数，该式为加法模型。

实际上，当损失函数$L(y,f(x))$和训练数据给定的条件下，该加法模型转化为损失函数最小化的问题:

$$\min_{\beta_m,\gamma_m} \sum_{i=1}^{N} L\left((y_i, \sum_{m=1}^{M} \beta_m b(x_i;\gamma_m))\right).$$

一般要求解该优化问题比较困难，为了求解这一问题，向前分布算法的思想是利用加法模型的特点，从前往后，每一步只单独学习一个基函数和它对应的系数，用局部最优去逼近全局最优，减小优化问题的复杂度，每一步需要优化的损失函数如下：

$$\min_{\beta,\gamma} \sum_{i=1}^{N} L(y_i, \beta b(x_i;\gamma)).$$

GBDT 模型由 Freidman〔2001〕年提出，全称为 Gradient Boosting Decision Tree，GBDT 是使用了向前分布算法的加法模型。组成 GBDT 的弱学习器被限定为 CART 回归树。对于提升树而言，其核心思想为残差拟合，GBDT 梯度提

升树使用损失函数的负梯度作为提升树算法中残差的近似值。GBDT 回归算法的基本流程如下：

1.初始化学习器,c的均值可以设置为样本y的均值。对于训练集 D = $\{(x_i, y_i), i = 1,2,\dots,m\}$，$x_i \in R^d$，最大迭代次数为$T$，损失函数为$L$，

$$f_0(x) = argmin_c \sum_{i=1}^{n} L(y_i, c),$$

2.对迭代次数$t = 1,2,\dots,T$。首先对样本$i = 1,2,\dots,m$计算负梯度，

$$r_{it} = -\left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right]_{f(x)=f_{(t-1)}(x)},$$

然后利用$\{(x_i, r_{it}), i = 1,2,\dots,m\}$拟合得到第$t$棵 CART 回归树，其对应的叶子节点区域为$\{R_{tj}，\ j = 1,2,\dots,J\}$，其中$J$为回归树$t$的叶子节点个数。

对于叶子区域$j = 1,2,\dots,J$，计算最佳拟合值：

$$c_{tj} = argmin_c \sum_{x_i \in R_{tj}} L(y_i, f_{t-1}(x) + c),$$

再更新强学习器，

$$f_t(x) = f_{t-1}(x) + \sum_{j=1}^{J} c_{tj}\, I(x \in R_{tj}),$$

3.最后得到强学习器$f(x)$的表达式如下：

$$f(x) = f_T(x) = f_0(x) + \sum_{t=1}^{T}\sum_{j=1}^{J} c_{tj}\, I(x \in R_{tj}).$$

3.4 基于机器学习方法的潜在因子模型

3.4.1 模型设定

由于基于机器学习的潜在因子模型在模型设定和参数估计方法上类似，我们这里不对每个机器学习方法单独介绍，而是采用统一的符号代表机器学习模型进行阐释。

本文对现有的潜在因子模型进行改进，将机器学习方法引入到模型中，主

要是为了利用机器学习模型来拟合因变量$\mathbf{y}_t$和可观测因子$\mathbf{z}_t$之间的关系，从而进一步对模型中的其他参数进行估计。由于$\mathbf{y}_t$的维度非常高，为了提高估计的准确性，我们这里提出一种预测模型，首先将$\mathbf{y}_t$（$t = 1,2,\ldots T$）分成v个维度的数据，$v = 1,2,\ldots p$，然后将每个$\mathbf{y}_{vt}$作为因变量，将可观测因子$\mathbf{z}_t$形成的$\mathbf{z}_t^{'}$作为自变量，公式如下，

$$\mathbf{y}_{vt} = \begin{bmatrix} \mathrm{y}_{v1} \\ \vdots \\ \mathrm{y}_{vt} \end{bmatrix}_{T\times 1}, \mathbf{z}_t^{'} = \begin{bmatrix} \mathrm{z}_{1,1} & \cdots & \mathrm{z}_{1,m} \\ \vdots & \ddots & \vdots \\ \mathrm{z}_{T,1} & \cdots & \mathrm{z}_{T,m} \end{bmatrix}_{T\times m},$$

其中$\mathbf{y}_{vt}$指的便是$\mathbf{y}_t$在t时刻第v个维度对于的数值。以股票市场的收益率数据为例，假如$\mathbf{y}_t$是一个 100×1 的矩阵，那么高维时间序列$\mathbf{y}_t$（$t = 1,2,\ldots T$）就包含了 100 只股票在每个t时刻的的收益率数据，那么对于某个维度v，$\mathbf{y}_{vt}$其实就是指第v只股票的时间序列数据。而对于这里的可观测因子$\mathbf{z}_t$，它是$m \times 1$ 维的时间序列，形成的$\mathbf{z}_t^{'}$包含了所有时间的可观测因子信息。

我们这里的具体做法是对于p个维度的训练集$\{(\mathbf{z}_t^{'}, \mathbf{y}_{vt}), v = 1,2,\ldots p\}$分别建立共计$p$个机器学习模型。对于某一维度$v$，$\mathbf{z}_t^{'}$的行向量作为机器学习模型的自变量，$\mathbf{y}_{vt}$对应的行元素作为因变量，这样就构成了用于机器学习模型拟合的一条记录，由于时间序列长度为T，我们这里共计可以得到T条记录用于某个维度p的拟合。

对于某一维度v，其训练集 $\mathrm{D} = \{(\mathbf{z}_t^{'}, \mathbf{y}_{vt}), v = 1,2,\ldots p\}$，训练后得到强学习器$f(\mathbf{z})_v$，

$$f(\mathbf{z})_v = f_0(\mathbf{z})_v + \sum_{t_{iterate}=1}^{T_{iterate}} \sum_{j=1}^{J} c_{t_{iterate}j}\, I(z \in R_{t_{iterate}j}).$$

我们共计可以得到p个强学习器$\{f(\mathbf{z})_v, v = 1,2,\ldots,p\}$,利用$f(\mathbf{z})_v$，可以预测得到第$v$个维度$\mathbf{y}_{vt}$的预测值$\hat{\mathbf{y}}_{vt}$，进而得到所有维度的预测值$\{\hat{\mathbf{y}}_{vt}, v = 1,2,\ldots,p\}$。这也是我们将这$p$个强学习器结合在一起的方法，最终得到一个总学习器$f(\mathbf{z}_t)$，方法示意图如下，

$$f(\mathbf{z}_t) \xleftarrow{\text{结合}p\text{个强学习器}} \begin{cases} f(\mathbf{z})_1 \\ \vdots \\ f(\mathbf{z})_p \end{cases},$$

基于机器学习的潜在因子模型如下：

$$\mathbf{y}_t = f(\mathbf{z}_t) + \mathbf{A}\mathbf{x}_t + \boldsymbol{\varepsilon}_t,$$

其中，与前面的因子模型类似，$\mathbf{y}_t$和$\mathbf{z}_t$分别是可观测的$p \times 1$维和$m \times 1$维时间序列，$\mathbf{x}_t$是$r \times 1$维的潜在因子过程，$\boldsymbol{\varepsilon}_t$是白噪声，$\boldsymbol{\varepsilon}_t \sim \mathrm{WN}(0, \boldsymbol{\Sigma}_\varepsilon)$，且$\boldsymbol{\varepsilon}_t$与$(\mathbf{z}_t, \mathbf{x}_t)$无关，$\mathbf{A}$是未知的因子载荷矩阵。潜在因子的个数$r$是一个未知的固定常数。与前面的潜在因子模型不同，由于我们这里使用的是机器学习模型替换多元线性回归模型，我们不需要再估计$\mathbf{z}_t$的未知参数矩阵$\mathbf{D}$。

3.4.2 参数估计

基于机器学习的潜在因子回归模型的参数估计过程与潜在因子回归模型的参数估计方式相似，主要的不同在于对残差$\boldsymbol{\eta}_t$的估计之上。我们需要估计的参数有三个，残差$\boldsymbol{\eta}_t$、未知的因子载荷矩阵$\mathbf{A}$以及潜在因子的个数r。

对于残差$\boldsymbol{\eta}_t$的估计，我们首先将模型分为两部分，

$$\mathbf{y}_t = f(\mathbf{z}_t) + \boldsymbol{\eta}_t, \ \boldsymbol{\eta}_t = \mathbf{A}\mathbf{x}_t + \boldsymbol{\varepsilon}_t,$$

其中$\mathrm{cov}(\mathbf{z}_t, \boldsymbol{\eta}_t) = 0$，首先通过结合后的机器学习模型对$\{(\mathbf{z}_t, \mathbf{y}_t), t = 1,2,\dots,T\}$进行拟合，然后通过结合后的机器学习模型预测得到$\mathbf{y}_t$的估计值$\hat{\mathbf{y}}_t$，$\boldsymbol{\eta}_t$的估计值$\hat{\boldsymbol{\eta}}_t$的计算式如下，

$$\hat{\boldsymbol{\eta}}_t = \mathbf{y}_t - \hat{\mathbf{y}}_t.$$

在完成了对$\hat{\boldsymbol{\eta}}_t$的估计之后，未知的因子载荷矩阵$\mathbf{A}$以及潜在因子的个数$r$的估计与前面潜在因子回归模型对其估计类似，这里不再赘述，只列出关键的估计部分，

$$\hat{\boldsymbol{\Sigma}}_\eta(k) = \frac{1}{T-k}\sum_{t=1}^{T-k}(\hat{\boldsymbol{\eta}}_{t+k} - \overline{\boldsymbol{\eta}})(\hat{\boldsymbol{\eta}}_t - \overline{\boldsymbol{\eta}})^{\mathrm{T}}, \ \overline{\boldsymbol{\eta}} = \frac{1}{T}\sum_{t=1}^{T}\hat{\boldsymbol{\eta}}_t .$$

类似地，记$\mathbf{A}$的估计式$\hat{\mathbf{A}} \equiv (\hat{\mathbf{a}}_1, \dots, \hat{\mathbf{a}}_r)$。这里$\hat{\mathbf{a}}_1, \dots, \hat{\mathbf{a}}_r$就是$\mathbf{M}$的估计$\widehat{\mathbf{M}}$的前$r$大个特征值满足$\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_r > 0$对应的标准正交特征向量。$\widehat{\mathbf{M}}$的计算式如下，

$$\widehat{\mathbf{M}} = \sum_{k=1}^{\bar{k}} \hat{\boldsymbol{\Sigma}}_\eta(k)\, \hat{\boldsymbol{\Sigma}}_\eta(k)^{\mathrm{T}}.$$

潜在因子个数r的估计式如下，

$$\hat{r} = argmin\left\{\frac{\hat{\lambda}_{j+1}}{\hat{\lambda}_j}: 1 \le j \le R\right\}, \text{ 其中} \hat{\lambda}_1 \ge \cdots \ge \hat{\lambda}_p.$$

在估计中一般取$R = 2/p$。