目录

[摘要 2](#_Toc3316014)

[Part 1 引言 2](#_Toc3316015)

[Part 2 文献综述 2](#_Toc3316016)

[Part 3 模型与算法选择 2](#_Toc3316017)

[3.1 普通最小二乘线性回归（OLS） 2](#_Toc3316018)

[3.2 弹性网络（Elastic Net） 2](#_Toc3316019)

[3.3 2](#_Toc3316020)

[3.4 Boosting 2](#_Toc3316021)

[3.5 Random Forest 2](#_Toc3316022)

[Part 4 输入特征选取 2](#_Toc3316023)

[4.1 个股因子选取 3](#_Toc3316024)

[4.2 宏观变量选取 4](#_Toc3316025)

[4.3 行业变量选取 4](#_Toc3316026)

[Part 5 实证结果与分析 5](#_Toc3316027)

[5.1 准备工作 5](#_Toc3316028)

[5.2 模型训练与验证 6](#_Toc3316029)

[Part 6 结论与展望 7](#_Toc3316030)

[致谢 7](#_Toc3316031)

## 摘要

本文试图将机器学习领域的经典算法，引入到量风险资产的超额收益这一资产定价领域的经典问题中来。具体来讲，本文主要做了以下三个方面的工作：

首先，本文汇集了对中国市场的A股超额收益有显著影响的三个方面因子，即个股因子，宏观因子和行业因子，以其作为输入值进行建模。此外，除了传统的OLS回归方法，本文构建了适用于评价机器学习模型表现的样本外R-square统计量，用以对比Huber损失函数下的OLS回归，弹性网络以及boosting和bagging等集成学习模型的效果，并发现

其次，本文试图深入到每一个模型的内核，去分析究竟是哪些输入变量对结果产生了显著影响。为此，本文针对每一种机器学习算法的特性，设计了评估自变量重要性的算法，从而得到最终的结果，从OLS模型来看。。。

最后，为了更进一步地探究各模型的稳健性，本文将问题从单个资产转向资产组合，效仿Fama-French（1993）和Fama-French（2015）的做法，将所有股票分别根据市值和净市率、市值和投资、市值和盈利能力、市值和动量3\*2分为6组，考察模型对组合表现的预测能力。结果显示，

## Part 1 引言

## Part 2 文献综述

## Part 3 模型选择与模型评价

现有的资产定价实证论文中，不论选用什么样的预测变量，其核心的预测算法几乎都是OLS线性回归算法。本文首要的创新之处就在于试图将机器学习领域的经典模型引入资产定价实证中来，以观察这种非线性的学习算法能否挖掘出更多的有用信息。这一部分将依次对于本文选用的模型的基本原理做一个介绍，以便对机器学习领域不熟悉的读者理解文章内核精神。具体来讲，针对每一个模型，本文都将从三个核心的方面来对其进行介绍：第一是该算法的优化目标，也就是损失（Loss）的计算方式；第二是优化过程中使用的算法；第三则是影响到算法效果的主要超参数，也就是CV验证集需要优化的对象。

### 3.1 普通最小二乘线性回归（OLS）

### 3.2 弹性网络（Elastic Net）

### 3.3 Boosting

### 3.4 Random Forest

### 3.5 模型评价标准

当使用机器学习模型来学习数据时，传统的可决系数评价标准将不再适用于我们的问题。所以，在这里本文定义了一种更适合机器学习模型的评价标准：样本外R-square，其计算方式如下：

其中，值的是测试集，也就是完全处于样本外的部分。该评价标准计算了模型对于所有样本外数据的平均预测误差，可以适用于所有的机器学习模型。

### 3.6 模型有效性来源的分析

## Part 4 输入特征选取

模型的输入变量分为三个部分，一部分是个股的因子值，二是因子值与宏观变量的交叉项，最后是行业虚拟变量。这一部分将对这三部分特征的选取做一个详细的介绍。

### 4.1 个股因子选取

个股的因子部分，Belly, Xiu（2017）等人的论文，汇总了95个在美国市场上被认为有效的异常因子，但大量的研究显示，在美国市场上显著的异常因子，并不一定适合中国A股市场。因此，需要考虑中国的情况选择合适的因子，这里我们同样参考了Liu（2018）中列举出的9大类14个在中国市场比较显著的因子，其取值均可由个股的市场数据和财务报表数据整理计算得到，这些数据将来自国泰安经济金融研究数据库。

具体的因子定义以及详细的计算方式列示如下：

1. Size: 个股的市值，由t时点的收盘价乘以发行在外总股本数得到。（总市值而非流通市值）
2. Value: 个股的价值，包括市盈率（Earnings-price ratio，EP)，市净率（Book-to-market ratio，BM)以及市现率（Cash-flow-to-price，CP)三个因子。
3. Profitability：个股的盈利能力，主要以权益收益率ROE来衡量。
4. Volatility：个股的波动率，包括以下两个部分
   1. 一月期滚动波动率（VOL）：以日度收益率计算得到的过去20个交易日的收益率标准差。
   2. MAX：过去20个交易日中的最高日度收益率
5. Investment(INVES)：公司的投资幅度，以年度的总资产增长率进行衡量。需要注意的是，总资产增长率以最近一份年报中的总资产，相对于其前一年年报中的总资产的增长率为标准。
6. 收入：个股的收入相关指标，包括以下两个因子
   1. 应计收入（Sloan（1996））：

其中，和分别代表流动资产和流动负债相对于上年的增长额， 代表现金相对上年的增长，代表流动负债中负债部分的增长，代表应纳税额增长，Dep则代表过去一年的折旧摊销额。

* 1. 净运营资产（Net-operating-assets, NOA）:

1. Illiquidity(Amihud(2002))：个股的流动性，以t时点个股收益率的绝对值，除以其同期交易量来衡量。
2. Turnover：个股的换手率，主要包括以下两个方面
   1. 12个月换手率(TO)：过去250个交易日的平均换手率（换手率 = 交易量/发行在外股份数）
   2. 1月期异常换手率(EXTRA\_TO)：过去20个交易日的平均换手率于过去250个交易日换平均手率的比值
3. Reversal：个股的反转效应，以过去20个交易日的累计收益率进行衡量

为了验证因子选取的有效性，在这一部分里，本文将沿用经典资产定价文献所使用的方式，从以下两个方面对所选的因子进行分析：首先，对于每个因子，本文使用上个月的因子值，将沪深两场所有股票分为5组，以市值加权和等权两种方式构建组合，按月调仓，最终得到5个组合的月度收益率，并对以做多第5组同时做空第1组的方式构建的high-minus-low组合的收益率进行t检验，从而验证因子在横截面上的有效性。其次，为了考察各个因子之间的相关性，以第一步中构建的各因子的high-minus-low组合的时间序列收益率，计算组合间的相关系数，从而反映因子之间的相关性。

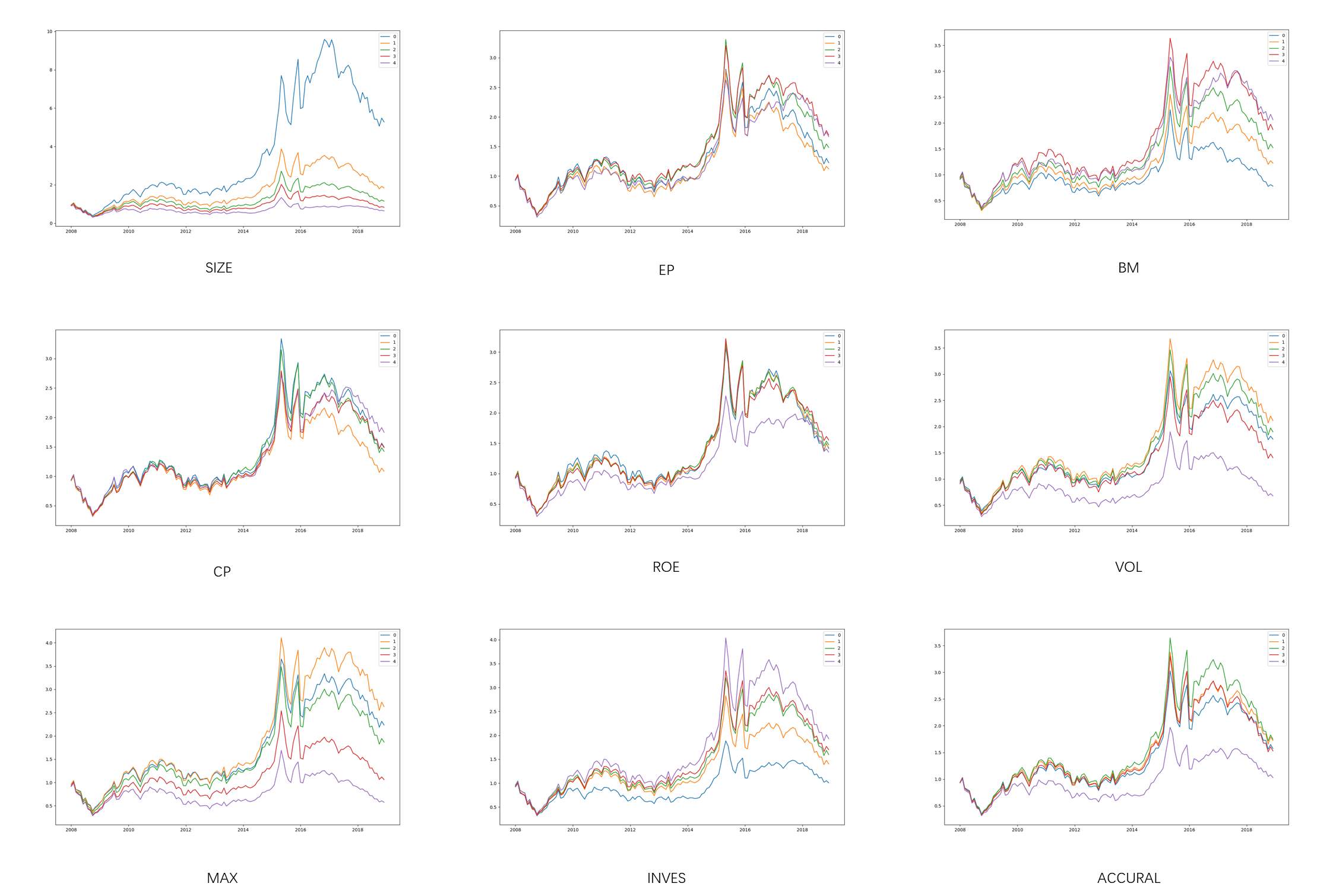
此外，考虑到2017年以来A股市场风格的转变，本文做了两个时间区间的测试，一个是2008年-2016年，另一个则是2008-2018年。在这两个时间区间下，分别得到的因子的t值列示在下表中:

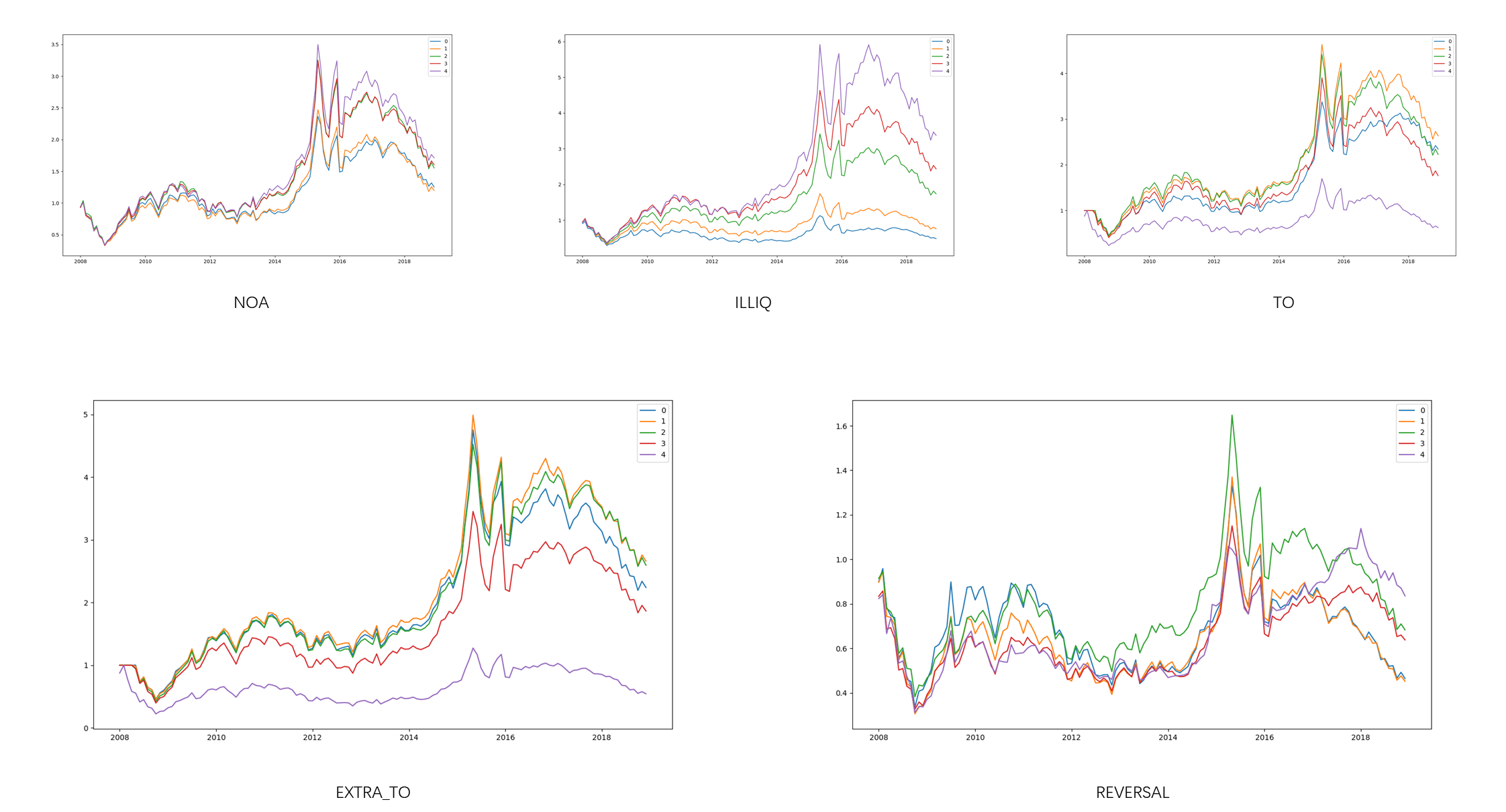
|  |  |  |
| --- | --- | --- |
|  | 2008-2018 | 2008-2016 |
| SIZE | -2.453 | -3.681 |
| **EP** | **0.593** | **-0.308** |
| BM | 1.252 | 0.484 |
| **CP** | **0.989** | **0.067** |
| ROE | -0.703 | -1.709 |
| VOL | -3.416 | -2.305 |
| MAX | -6.032 | -4.462 |
| INVES | 1.911 | 2.878 |
| ACCRUAL | -1.968 | -2.443 |
| NOA | 1.412 | 1.537 |
| ILLIQ | 3.955 | 4.113 |
| TO | -2.802 | -2.080 |
| EXTRA\_TO | -3.221 | -2.734 |
| REVERSAL | 1.343 | 0.889 |

从表中可以看出， EP因子和CP因子的显著性非常差，BM因子、ROE因子、NOA因子和REVERSAL因子的显著性也一般，需要进一步的考察。而其他因子，主要是流动性、换手率、投资三个方面的因子，在两个时间区间内都显示出了较强的显著性。

此外，由于本文与传统论文的根本区别点在于将焦点放在机器学习算法，和非线性关系的拟合之上，所以只考虑t检验的结果是不够的，下面两页给出的14张图分别绘制在2008-2018年这个区间下，每个因子的分组累积净值曲线，以供我们观察是否存在一些非线性的关系。从中可以看出，尽管BM因子的high-minus-low组合的表现一般，但各组之间的区分度是比较明显的，只是这种区分并非是完全的线性关系，所以不能用简单的high-minus-low组合的t统计量来衡量，所以，本文选择保留BM因子；NOA和REVESAL因子也存在类似的情况，所以将被保留。然而，EP和CP两个因子从图像上来看几乎没有区分度，因此，在此本文选择剔除这两个因子，保留剩余12个个股因子。

最后，下表给出了在2008-2018这个时间区间下，所有因子之间的相关系数矩阵，从中可以看出，应计收入因子和价值类因子之间的相关性较强，此外，投资因子和流动性因子的相关性较强，两个换手率类的因子之间，以及两个波动率类的因子之间都有比较强的相关性。如果用传统的OLS算法，将相关性较强的因子同时放在模型中难免会导致多重共线性的问题，但机器学习算法都有很好的变量动态选择和预防过拟合的机制，所以不需要剔除掉高度相关的因子。





|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | ACCRUAL | BM | CP | REV | EP | INVES | ILLIQ | MAX | NOA | ROE | SIZE | TO | EX\_TO | VOL |
| ACCRUAL | 1.00 | **0.84** | **0.89** | 0.12 | **0.83** | -0.73 | -0.72 | -0.14 | -0.77 | 0.42 | 0.72 | -0.28 | 0.14 | -0.28 |
| BM | **0.84** | 1.00 | **0.89** | 0.02 | 0.77 | -0.74 | -0.73 | -0.29 | -0.74 | 0.22 | 0.50 | -0.25 | 0.11 | -0.44 |
| CP | **0.89** | **0.89** | 1.00 | 0.13 | **0.87** | -0.75 | -0.71 | -0.32 | -0.75 | 0.45 | 0.67 | -0.38 | 0.00 | -0.42 |
| REVERSAL | 0.12 | 0.02 | 0.13 | 1.00 | 0.11 | 0.03 | -0.01 | 0.15 | -0.13 | 0.22 | 0.21 | -0.31 | -0.06 | 0.11 |
| EP | **0.83** | 0.77 | 0.87 | 0.11 | 1.00 | -0.73 | -0.61 | -0.28 | -0.70 | 0.72 | 0.75 | -0.39 | 0.05 | -0.39 |
| INVES | -0.73 | -0.74 | -0.75 | 0.03 | -0.73 | 1.00 | 0.71 | 0.16 | 0.64 | -0.40 | -0.63 | 0.22 | -0.09 | 0.26 |
| ILLIQ | -0.72 | -0.73 | -0.71 | -0.01 | -0.61 | 0.71 | 1.00 | -0.05 | 0.78 | -0.12 | -0.49 | 0.05 | -0.19 | 0.02 |
| MAX | -0.14 | -0.29 | -0.32 | 0.15 | -0.28 | 0.16 | -0.05 | 1.00 | 0.15 | -0.21 | -0.04 | 0.28 | 0.41 | 0.87 |
| NOA | -0.77 | -0.74 | -0.75 | -0.13 | -0.70 | 0.64 | 0.78 | 0.15 | 1.00 | -0.32 | -0.58 | 0.17 | -0.08 | 0.23 |
| ROE | 0.42 | 0.22 | 0.45 | 0.22 | 0.72 | -0.40 | -0.12 | -0.21 | -0.32 | 1.00 | 0.74 | -0.41 | -0.04 | -0.24 |
| SIZE | 0.72 | 0.50 | 0.67 | 0.21 | 0.75 | -0.63 | -0.49 | -0.04 | -0.58 | 0.74 | 1.00 | -0.42 | 0.00 | -0.14 |
| TO | -0.28 | -0.25 | -0.38 | -0.31 | -0.39 | 0.22 | 0.05 | 0.28 | 0.17 | -0.41 | -0.42 | 1.00 | 0.64 | 0.31 |
| EXTRA\_TO | 0.14 | 0.11 | 0.00 | -0.06 | 0.05 | -0.09 | -0.19 | 0.41 | -0.08 | -0.04 | 0.00 | 0.64 | 1.00 | 0.32 |
| VOL | -0.28 | -0.44 | -0.42 | 0.11 | -0.39 | 0.26 | 0.02 | **0.87** | 0.23 | -0.24 | -0.14 | 0.31 | 0.32 | 1.00 |

### 4.2 宏观变量选取

宏观变量部分需要的是对于整个A股市场有普遍影响的宏观指标，既包括宏观经济指标，也包括一些反应金融市场整体运行状况的指标。参考Welch and Goyal（2008）的做法，选出了8个宏观变量，分别的计算方式以及数据来源列示如下：

1. 市盈率（Earnings Price Ratio, EP）:来自wind数据库的申万A股平均市盈率数据

Earnings：滚动过去12个月的Earnings之和。

1. 市净率（Book-to-Market Ratio, BM）: 来自wind数据库的申万A股平均市净率数据
2. 权益净扩张（Net Equity Expansion, NTIS）:由wind数据提供的A股权益净增发数据和市值数据计算而得。

Net Issue指的是滚动的过去12个月的净权益增发额之和，每个月的Net Issue则可由以下公式估计而得：

其中，rt指的是整个市场过去一月的不带分红收益率。

1. 短期国债利率（Short Term Yield, STY）：1年期国债到期收益率，来自国泰安数据库。
2. 长期国债利率（Long Term Yield, LTY）：10年期国债到期收益率，来自国泰安数据库。
3. 期限利差（Term Spread, TMS）：长短期国债利率的差。
4. 信用利差（Default Spread, DFY）: 信用利差指的是信用债到期收益率和国债到期收益率之间的差异，在现有的数据库中，没有比较成熟的信用利差数据，因此，本文从彭博数据库取得了中债的信用债总指数和国债总指数的平均现金流法到期收益率数据，取其差值作为信用利差。
5. 滞后一月期通胀率（One-month-lagged Inflation, INFL）:来自国家统计局官网的月度CPI同比涨幅。这里之所以要滞后一月，是因为当月的CPI数据一般而言要在下月才公布。以中国为例，国家统计局一般会在本月的10号左右公布上月的CPI数据。

### 4.3 行业变量选取

对于中国股市的实证研究，常用的行业分类有申银万国行业分类/中证行业分类/证监会行业分类等几种。本文选用的是证监会的一级行业分类，该分类标准将整个市场上的股票划分为19个大行业，这部分数据来自国泰安经济金融研究数据库。

## Part 5 实证结果与分析

### 5.1 准备工作

#### 5.1.1 数据预处理

参考liu（2018）等人的做法，本文选取的研究标的为来自上/深两市的全部A股数据，以月度为研究频率，并对涵盖了沪深两市A股的整个股票池做如下两个方面的过滤：第一，因为新上市的股票往往会出现价格波动过大的情况，所以我们在每个时间点上，过滤掉上市不足六个月的股票；其次，由于使用的是月度收益率，为了保证样本的有效性，需要过滤掉在过去一个月中，由于停牌等因素导致的交易日不足15天的股票。

对于原本的季度财务数据，也需要做两个方面的处理。首先，CSMAR提供的财务数据所对应的时间是会计期间，而不是财务报表的发布时间，所以我们需要根据财报发布的时间，对原始的财务数据对应的时间区间进行调整，以防用到未来数据。其次，我们将季度的财务数据日度化用于因子值的计算，保证在每个时间点上，用到的财务数据均为截至当前时间点，最近一次发布的财务报表数据。

此外，考虑到财务数据的可得性，和中国股票市场的改革进程，选取的数据期间为2005-01-01 至 2018-12-31。

#### 5.1.2 输入变量和输出变量计算

在整个研究中，作为输出变量的个股超额收益，由个股的收益率与对应的无风险利率相减得到，这两部分数据均来自国泰安经济金融研究数据库（CSMAR），其中的无风险利率基准选用的是银行一年期固定利息率。

输入变量方面，主要是上文提到的三个方面，个股的因子方面，我们首先根据日度的市场交易数据和财务数据，得到日度的因子值，最后按月取均值，作为当月的因子值；宏观因子中，市盈率、市净率、国债利率数据是日度的，同样取月度均值，其他为月度因子，除CPI滞后一月外，不做其他处理；行业因子则要考虑个股在行业划分上的变动性，得到每个月每支股票的行业因子，再处理成0/1形式的独热矩阵，作为最后的输入值。最后，对12个个股因子，每个都将分别与8个宏观因子计算交叉项，加上其本身，最终得到12 \* 8 + 12 = 108维度的输入值，并进行去极值和标准化处理，加上19维的行业因子矩阵，最终的输入变量为127维。

由于许多公司的财务报表数据有一定的缺失，加上因子计算导致的时间区间上的损失，完成了数据的预处理和因子值计算之后，最后得到的数据期间为2008年1月到2018年12月。

最终得到的整个研究期间的有效样本数量为164981个，其中，每个年度的样本数量，以及包含在样本中的股票数量列示在下表中：

|  |  |  |
| --- | --- | --- |
| 年份 | 股票数量 | 数据量 |
| 2008 | 891 | 5652 |
| 2009 | 1071 | 11276 |
| 2010 | 1108 | 12176 |
| 2011 | 1234 | 12756 |
| 2012 | 1473 | 15303 |
| 2013 | 1638 | 17323 |
| 2014 | 1707 | 17267 |
| 2015 | 1723 | 15038 |
| 2016 | 1786 | 17069 |
| 2017 | 1853 | 18468 |
| 2018 | 2020 | 20097 |
| 合计 |  | 162425 |

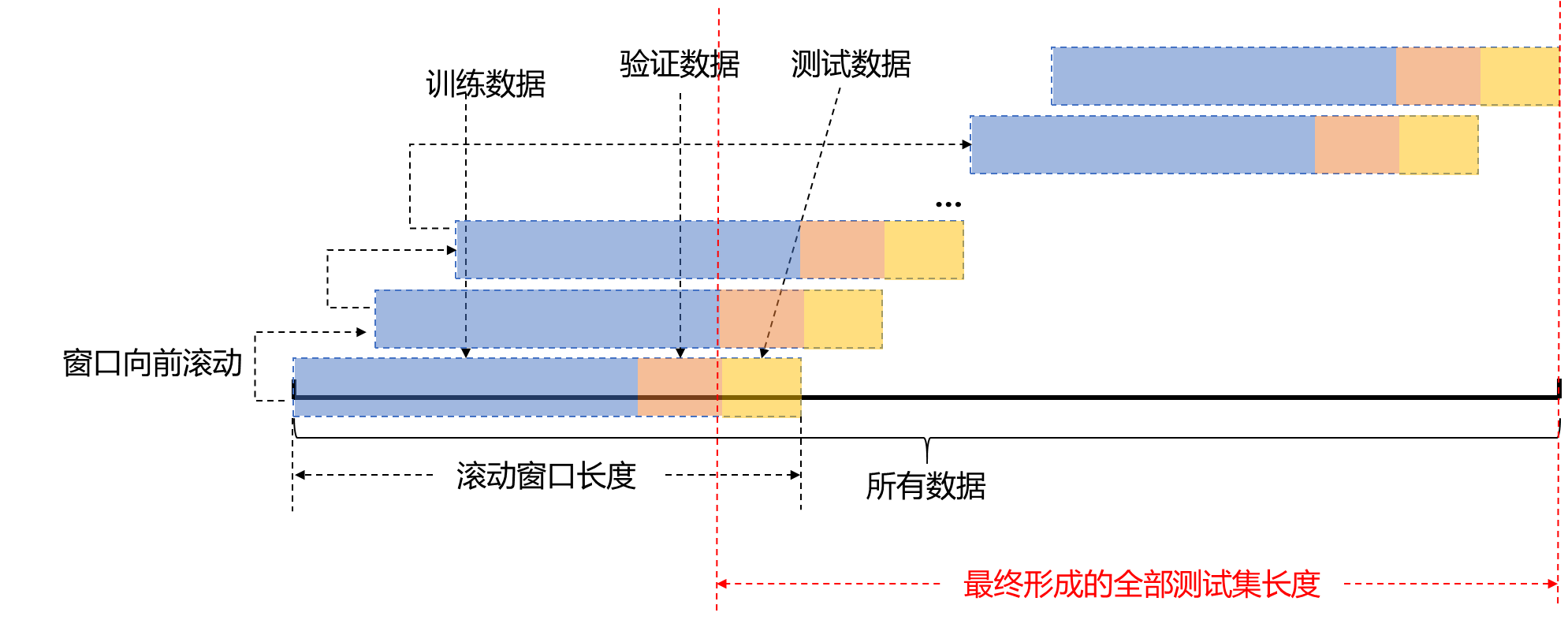
### 5.2 模型训练与验证

现有文献中，对于机器学习模型的训练过程，一般分为以下三种处理方式：

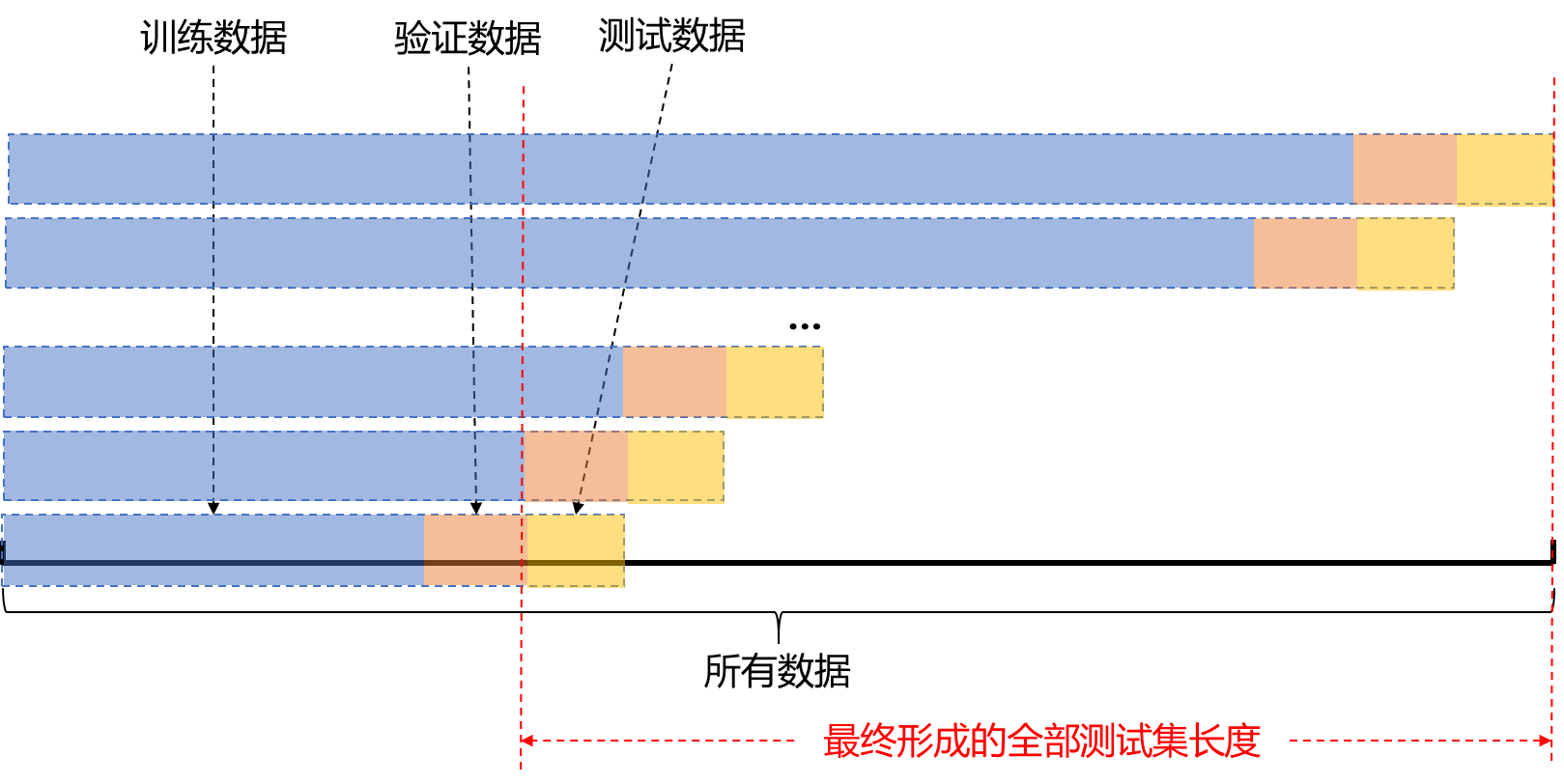
第一种是固定期间训练，将整个数据期间划分为训练集、交叉验证集、测试集三个部分，分别进行模型训练，超参数修正和最后的模型效果验证。这种方式可以直观地用下图表示：



第二种方式则是滚动期间训练，这种方式会预先确定一个固定长度的滚动窗口，在整个数据期间内，按这个窗口向前滚动训练，而窗口内部又会划分为训练集、验证集和测试集三个部分，随着整个滚动过程的推进，使用的训练集的长度是固定不变的。这一过程同样可以以下面的流程图表示：



第三种方式则是不变锚点滚动期间训练，这种方式和第二种方式一样，通过不断向前滚动而获得一个更长的训练集长度，区别在于它的滚动窗口并非固定长度的，而是用于以整个数据期间的起点，作为每次滚动窗口的起点向前滚动，这样就会导致训练集长度不断增加，好处是可以学习到所有的历史数据中的信息，坏处则是会加大整个模型训练过程中的计算量。以流程图表示则是下面这样一个过程：



在本文中，我们选择的是上述三种方式中的第三种，即不变锚点滚动训练期间的方式，主要是出于三个方面的原因：首先，在我们的研究期间中，中国的股票市场一直是处于一个不断完善发展的阶段的，经济大环境也在不断变化，所以对数据进行一次性的拟合，并不能很好地捕捉所有期间的市场内在逻辑，滚动拟合的方式更适合不断变化的经济金融环境；其次，我们的研究期间一共只有12年，如果采取一次性划分方式，测试集的长度是比较短的，而滚动拟合的方式可以得到一个更长的测试期间，更好地验证模型的稳健性，但是，滚动拟合又会同时导致训练数据的长度不足，因此，使用不变锚点滚动的方式，可以同时保证训练集和测试集的长度满足需求。

本文以5年为第一个滚动的窗口，之后不断累加。为了保持时间顺序，不进行交叉验证，而是以整个窗口的最后一年作为测试集，倒数第二年作为验证集，其余年份均作为训练集。因此，第一个测试集将从13年开始，并随着滚动过程不断增长，最终得到一个包含了13-18年6年数据的测试集结果，这一结果将是完全的样本外结果，用于计算我们在第三部分中定义的模型评价标准——样本外R-square（R2-oos）。

## Part 6 结论与展望

从本文的研究出发，后来者可以针对这个问题，在以下两个方面进行更深入的研究：

首先，在因子的选取问题上，本文主要是借鉴了目前已经被已有文献证实在A股市场存在显著超额收益的因子，这样做难免就会受到时滞问题的一些影响，尤其是2017年以来，市场风格的转变导致size等传统因子失效，以后的研究需要纳入更多更新，更有效可靠的因子进行分析。

第二，在算法的选取问题上，考虑到A股市场的历史较短，数据量少，还达不到所谓“大数据”的标准，因此本文选择的算法集中在传统的机器学习领域，而深度学习领域的神经网络等算法没有涉猎。随着中国股市的发展和可用数据的增多，或许可以尝试更多机制更复杂，更智能的算法。

## 致谢

郭老师、刘老师、梁老师

小美、若男