

基于 LSTM 优化的金融市场风险评估

摘要

金融风险是金融市场不可忽视的概念。2008年金融危机之后，这一领域的研究更是到了顶点。近些年随着贸易保护主义与全球化的趋势不断冲突，市场对金融风险的重视程度日益提高。因此，是很有必要对于金融市场进行风险评估的。股票市场是经济的晴雨表，因此本文选择对于股票市场的指数做风险评估。首先对指数（沪深300）的对数化收益，进行计算，作为标签。随后，提取和设计了523组因子，在因子进行有效性分析，共线性处理后，分布形态上进行删选。将得到的294个因子进行归一化处理，最后用LSTM模型进行预测。同时对预测结果，进行优化，优化后的结果还与未进行删选的预测结果、非对数化标签的预测结果、多元非线性回归的预测结果进行比较，且使用了自行设计的评估函数进行分析。

结论表明，在金融市场，对数化标签要比非对数化的标签更加适合预测，优化后的LSTM预测性能比多元非线性回归要好、设计的评估函数对于预测需求有着良好的评估、论证了对于因子数据进行共线性处理的必要性。因子形态测试与归一化处理，更加适合本文预测模型。

关键词：金融风险，Beta 预测，对数正态分布，LSTM

LSTM-optimized risk assessment of financial markets

ABSTRACT

Financial risk is a concept that financial markets cannot ignore. After the 2008 financial crisis, research in this field reached its peak. In recent years, with the trend of trade protectionism and globalization, the market attaches great importance to financial risks. Therefore, it is necessary to carry out risk assessment of financial markets. The stock market is a barometer of the economy, so this paper chooses to make a risk assessment for the index of the stock market. First of all, the index (CSI 300) of the reciprocal income, as a label. Then, 523 sets of factors were extracted and designed, and after the factor effectiveness analysis, the distribution pattern was selected after the collinear processing. The 294 factors obtained were normalized and finally predicted by the LSTM model. At the same time, the prediction results are optimized, and the results after optimization are compared with the prediction results of undelated, the prediction results of non-log labels, and the prediction results of multi-nonlinear regression, and the evaluation functions designed by themselves are used for analysis.

The conclusion shows that in the financial market, logarithm label is more suitable for prediction than non-log label, the optimized LSTM prediction performance is better than multi-nonlinear regression, the evaluation function of design has a good evaluation of the forecast demand, and the necessity of collinear processing of factor data is demonstrated. The treatment of normalized, in this paper is more suitable for the prediction model.

KEY WORDS: Financial Risk; Beta Prediction; Positive Normal Distribution; LSTM

目 录

摘要	
ABSTRACT	II
目录	III
图、表目录	V
1 绪论	1
1.1 背景	1
1.2 研究内容与主要工作	2
2 文献综述	2
2.1 系统性风险与非系统性风险	2
2.1.1 定义与特点	2
2.1.2 衡量指标	4
2.2 金融市场的价格风险	8
2.2.1 企业自身因素导致的价格风险	8
2.2.2 其他投资者（竞争者）的行为导致的价格风险	9
2.2.3 投资者的交易习惯和行为	10
2.2.4 整体市场因素导致的价格风险	10
2.3 多元非线性拟合模型	11
2.4 复杂网络	12
2.4.1 聚类与分类	13
2.4.2 投资组合优化与投资者行为	14
2.4.3 市场结构特点与结构性风险评估	15
2.4.4 复杂网络模型归纳	16
2.5 长短期记忆模型	17
2.5.1 遗忘门	18
2.5.2 传入门	19
2.5.3 输出门	19
3 因子库与标签的建立	20
3.1 整体概述	20
3.2 数据来源与预处理	23
3.2.1 原始数据来源	23
3.2.2 复权日检测与复权化处理	23
3.2.3 额外信息说明	25
3.3 标签的生成与处理	25
3.3.1 标签的定义	25
3.3.2 标签的计算方法	26

3.3.3 标签计算方法的优化.....	27
3.3.4 标签索引值的确定.....	29
3.3.5 标签的进一步思考.....	29
3.3.6 标签最终结果.....	32
3.4 因子逻辑建立	33
3.4.1 因子逻辑论证.....	33
3.4.2 因子逻辑设计.....	34
3.4.3 因子生成.....	39
3.5 因子有效性检测	39
3.5.1 因子有效性检测概述与方法.....	39
3.5.2 因子有效性检测.....	40
3.5.3 因子有效性分析.....	54
3.6 因子共线性检测	54
3.6.1 因子共线性概述.....	54
3.6.2 共线性检测.....	55
3.6.3 因子共线性处理.....	58
3.7 因子分布形态测试	60
3.8 因子优化	67
3.9 因子删选结果	69
4 LSTM 模型建立	70
4.1 预测逻辑概述	70
4.2 建立 LSTM 模型	72
4.2.1 数据匹配与切分.....	72
4.2.2 删除恒定值与缺失数值处理.....	72
4.2.3 数据的特征缩放.....	74
4.2.4 确定模型与定义网络架构.....	75
4.2.5 模型初步结果与评估函数建立.....	77
4.3 模型分析与优化	79
4.3.1 训练比率.....	79
4.3.2 学习率.....	80
4.3.4 迭代次数.....	82
4.3.5 初始结果与优化结果对比.....	83
4.3.6 全因子预测与因子删选预测结果对比.....	85
4.3.7 对数化收益和非对数化收益的对比.....	87
4.3.8 对比多元非线性回归.....	89
4.3.9 思考与不足.....	91
4.4 模型测试结论	92
5 结论	93
参考文献	96
致 谢	错误!未定义书签。

图、表目录

图 2.1 LSTM 结构	18
图 2.2 遗忘门.....	18
图 2.3 传入门.....	19
图 2.4 输出门.....	19
图 3.1 沪深 300 指数 K 线图节选.....	26
图 3.2 对数化收益与非对数化收益累计收益对比图.....	28
图 3.3 对数化收益与非对数化收益收益增量对比图.....	28
图 3.4 非对数化收益分布图.....	30
图 3.5 因子一 xy 散点分析图.....	42
图 3.6 因子二 xy 散点分析图.....	42
图 3.7 因子三 xy 散点分析图.....	44
图 3.8 因子四 xy 散点分析图	44
图 3.9 因子五 xy 散点分析图.....	45
图 3.10 因子六 xy 散点分析图.....	45
图 3.11 因子七 xy 散点分析图.....	46
图 3.12 因子八 xy 散点分析图.....	46
图 3.13 因子九 xy 散点分析图.....	47
图 3.14 因子十 xy 散点分析图.....	47
图 3.15 因子十一 xy 散点分析图.....	48
图 3.16 因子十二 xy 散点分析图.....	48
图 3.17 因子十三 xy 散点分析图.....	50
图 3.18 因子十四 xy 散点分析图.....	50

图 3.19 因子十五 xy 散点分析图.....	51
图 3.19 因子十六 xy 散点分析图.....	51
图 3.19 因子十七 xy 散点分析图.....	53
图 3.19 因子十八 xy 散点分析.....	53
图 3.19 因子十九 xy 散点分析图.....	53
图 3.24 因子二十 xy 散点分析图.....	53
图 3.25 因子一分布图.....	61
图 3.26 因子二分布图.....	61
图 3.27 因子三分布图.....	62
图 3.28 因子四分布图.....	62
图 3.29 因子五分布图.....	64
图 3.30 因子六分布图.....	64
图 3.31 因子七分布图.....	65
图 3.32 因子八分布图.....	65
图 3.33 因子九分布图.....	66
图 3.34 因子十分布图.....	66
图 3.35 未优化因子分布图.....	68
图 4.1 归一化与标准化比较.....	75
图 4.2 初始预测结果对比图.....	77
图 3.30 因子六分布图.....	64
图 3.31 因子七分布图.....	65
图 3.32 因子八分布图.....	65
图 3.33 因子九分布图.....	66
图 3.34 因子十分布图.....	66

图 3.35 未优化因子分布图.....	68
图 3.36 优化后因子分布图.....	68
图 4.1 归一化与标准化比较.....	75
图 4.2 初始预测结果对比图.....	77
图 4.3 训练比例测试图.....	80
图 4.4 学习率测试图.....	81
图 4.5 迭代次数测试图.....	82
图 4.6 初始预测结果测试图.....	83
图 4.7 优化预测结果测试图.....	84
图 4.8 删选后因子预测结果测试图.....	85
图 4.9 全因子预测结果测试图.....	86
图 4.10 优化后对数化收益预测结果.....	88
图 4.11 优化后非对数化收益预测结果.....	88
图 4.12 多元非线性回归预测结果.....	90
图 4.13 最优 LSTM 预测结果.....	90
 表 3.1 索引未处理标签数据表（节选）.....	32
表 3.2 索引处理标签数据表（节选）.....	32
表 3.3 初始因子名称表（节选）.....	39
表 3.4 有效因子名称表（节选）.....	54
表 3.5 最终入选因子表（节选）.....	69
表 4.1 预测结果表.....	78
表 4.2 初始预测结果评估表.....	79
表 4.3 初始与优化使用参数表.....	83

表 4.4 初始与优化预测结果评估表.....	83
表 4.5 全因子与删选后因子参数表.....	85
表 4.6 全因子与删选后因子预测结果评估表.....	85
表 4.7 对数化与非对数化标签参数表.....	87
表 4.8 对数化与非对数化预测结果评估表.....	87
表 4.9 LSTM 模型与多元非线性回归预测结果评估表.....	89

1 緒論

1.1 背景

在全球的许多国家都经历过多次的由严重的银行业务崩溃或者汇率问题导致的金融危机。实际上，国际货币基金组织 180 多名成员中的 130 多个国家在 1980 年至 1995 年之间，都有经历过金融危机的经历。银行系统和货币系统带来的金融危机不仅在全球范围内频繁发生，而且从实际产出的下降和纳税人向银行存款人和其他金融索赔人的转移支付（财富转移）的增加来看，危机的代价往往是极其高昂的，因为这些人的资金已得到明示或暗示的保险或按面值得到担保^[1]。如何应对和防范金融危机，本质应该是一个政府决策和监管课题。自 2008 年美国的次贷危机导致的全球金融危机后，对于金融风险的测度以及防范一直是研究热点。

实际上 2008 年之后，学术界和实务界都意识到了要求各个银行机构均以自身的稳健经营为目标的微观监管手段和资本约束的机制规定，已经无法有效的维护金融市场的稳定。并且其的弊端在金融危机之后被多方所关注。2010 年巴塞尔协议 III 的实施，将中国银行业资本监管由关注微观风险推向了宏观与微观审慎相结合的实施阶段^[2]。对于系统性风险更加准确的衡量，信用融资水平和金融累积风险的识别等都是理论和学术界迫切需要解决的问题。

金融风险不仅可以应用于企业，政府实体，整体金融市场也可以应用于个人。这种风险是股东，投资者或者其他财务利益相关者潜在会蒙受损失的危险和不确定性。根据金融风险的不同的风险源和风险成因，金融风险的种类可以分为以下几种：1. 系统性风险，非系统性风险；2. 流动性、信用、股权等风险，这其中的分类有的是依据财务属性，有的则是宏观微观的区分。

金融风险过高毫无疑问是会带来风险损失的。这一特点在金融市场中，尤其是在股票市场中，是十分明显的。因此，本文将金融市场的价格风险定义为由于金融风险过高导致的金融市场的价格崩溃的风险。

对于这一价格风险的研究不仅可以帮助投资者、企业进行合理的决策，也有助于政府机构等宏观调控。

1.2 研究内容与主要工作

本文首先对价格风险进行了初步的定义，选择了合适的样本数据。从多个维度提取了因子，运用了LSTM模型进行预测，在进行了分析和优化后，将结果与多元非线形拟合和单因子预测进行了对比分析。

2 文献综述

2.1 系统性风险与非系统性风险

2.1.1 定义与特点

若将金融风险区分为系统性风险和非系统性风险。则系统性风险通常是指的是整体金融系统或者金融市场崩溃的风险。更具体而言，系统性风险是指整个系统受到单部分倒闭的影响而面临冲击的风险和概率^[3]。系统性风险也被认为是对实体经济产生重大影响的金融机构集体行为的结果^[4]。系统性风险管理中，传染效应一个重要研究对象，研究其可以帮助学界深入了解风险的传递机制以及防治，监管。典型的，学者包全永（2005）通过构建经济模型，研究了一个封闭银行系统以及银行间市场的银行系统性风险的传染机理^[5]。根据其研究可得，中国金融市场上的，银行间的系统性风险具有自放大性。这极有可能导致整体银行业的崩溃以及给金融市场造成极大的冲击。而学者张晓朴（2010）将系统性风险定义为：整个金融体系崩溃或丧失功能的必然性，并且对其的特点进行总结，得到系统性风险具有复杂性、突发性、传染快、波及广、危害大五个基本特征^[6]。当然，对系统性风险的定义，也存在一定的争议，学者Edward J. Kane（2010）提出系统性风险的官方定义忽略了政府官员在产生系统性风险中的作用。政策制定者在危机情况下吸收损失的倾向性，鼓励机会主义公司，都会导致其在监管部门内促进和利用激励性冲突^[7]。

非系统性风险一般指的是某些无法预测的特殊事件（通常这些特殊事件与个股本身有关）的爆发，导致的个股本身收益的不确定性和波动性。非系统性

风险是现代公司金融的一个重要研究方向。学者徐廷玮（2015）运用企业财务理论与社会网络理论，分别从企业非系统性风险的成因（企业非效率投资）和治理对策（独立董事网络治理）两个维度展开研究，得到了企业非效率投资以及过度投资与企业非系统性风险呈显著正相关的结论以及独立董事网络中心度与企业非系统性风险呈显著负相关的结论^[8]。这就意味着企业进行非效率投资的程度越高，非系统性风险就越大，并且独立董事的治理行为一定程度上取决于独立董事治理在董事网络中的地位和影响力。

总的来说，系统性风险和非系统性风险的研究方向是有所区别的。对于系统性风险，学术界主要是围绕着如何衡量和如何避免这一风险，以及其衣橱效应导致的损失展开的。而对于非系统性风险，本身由于其发生通常是受到某一个或者一些特殊因素的刺激，通常具有不可预测性，但是对于投资组合来说，这种风险是可以被持有并优化所持有投资组合的数量以及权重所分散的。因此，优化的投资组合是该领域的研究重点，并且更多的研究是针对非效率投资以及董事网络结构的角度进行分析。

值得补充的是，从历史上看，大多数国家的保险，银行和其他金融市场之间存在明显的分离，因此一个领域的事件通常对另一个领域几乎没有影响。但是，近年来，保险公司，银行和其他金融公司之间的壁垒已被部分拆除，从而导致它们之间的联系更加紧密，其活动之间存在更多的联系和重叠。因此，在衡量系统性风险过程中，通常需要分为两个阶段，研究的早期阶段主要是单个部门（比如：保险、银行）风险独立情况下的各自的系统性风险的测度，以及对总体风险的贡献的衡量，而目前的研究则有侧重不同部门之间风险传染的研究。

而对于这几个部门的评估，学者 Oscar Bernal (2014) 通过评估主要不同金融部门（银行，保险和其他金融服务行业）内的困境在多大程度上导致系统性风险，从而识别这些部分对于系统性风险的做出贡献。其得到了与保险业相比，银行业对系统性风险的贡献似乎更大的结论。在其研究的同一时期对比下，发现保险业是美国系统性风险最高的金融部门，而银行业对该领域的系统性风险的影响最小^[9]。

关于银行业，根据斯坦福大学等学者 Elyas Elyasian (2019) 测试了，四个在金融危机时期，不同规模的银行的行为差异的假设：（1）非传统银行活动假设，（2）组织结构假设，（3）会计风险假设，以及（4）会计保守主义假设。结果表明，大型银行和小型银行之间的行为差异可能是，至少部分与前三个假设有关^[10]。与非传统活动假设相关的因素适合绩效数据（ROA 和 EBIT / TA）。组织复杂性因素是与银行系统风险（SRISK）相关的最相关的变量集。由此，监管机构应关注与这两个假设相关的因素。

其次，对于保险业，根据学者 Faisal Baluch (2011) 的研究，金融危机确实在很大程度上影响了保险市场，而保险市场中的事件反过来又影响了危机的发展方式。但是，危机对保险公司的影响是非常不均衡的：在保险市场的某些部分，这种影响相对平缓，在其他部门则严重^[11]。

2.1.2 衡量指标

对系统性风险，学术界较为公认的衡量方式有以下几种：（1）beta 系数，（2）GARCH 模型，（3）在险价值，（4）预期资本不足，（5）条件风险价值，（6）边际预期不足等方法。

1) β 系数

市场风险的测度最早源自于 Markowitz 的投资组合理论的均值方差模型。也就是使用收益率的方差（也就是使用投资组合的收益的波动率）来衡量金融资产的不确定的风险。在资本资产定价模型（CAPM）中，Beta 是通过衡量两者之间的相关性来衡量单个股票相对于市场的移动方式的理论方法。

学者施东晖就以资本资产定价模型（CAPM）为基础，对中国股票市场进行羊群行为检验，得出了我国股市存在一定程度的羊群行为的结论^[12]，并且根据其研究，中国股市存在的这一羊群行为，会导致系统性风险在总体金融机构的风险中占有较高的比重。

根据理论，任意的投资组合的风险（可以用收益率的方差来计算）可以被分为系统性风险和非系统性风险。其中，非系统性风险可以被投资组合的优化所降低甚至消除，但是系统性风险无法分散消除。

2) 在险价值

在险价值，指在一定的置信水平下，持有的资产组合，由于价格变动所导致的潜在预期的最大损失。在险价值是金融风险管理中最为主流的风险定量分析方法之一。它估计在给定的时间内，一组投资可能（有给定的概率）损失多少。用公式表达即为：

$$VaR = -q_a \quad 2.1$$

$$\Pr\{R \leq q_a\} = a \quad 2.2$$

其中，持有期为 holding period，置信区间 $a = 5\%$ ，其中， q_a 是机构收益率的分位数。

VaR 在金融领域，被广泛应用于：风险管理，财务控制，财务报告和监管资本等四个重要领域，目前，在险价值已成为风险管理的行业标准^[13]。

该方法的也存在一定的局限性，学者江婕(2015)提出，VaR 是一个非一致风险测度，不满足次可加性的要求^[14]。实际上，相关风险度量应该为一个单调性、可加性的函数。但是，在险价值通常不是一种连贯的风险衡量指标，因为它不尊重子可加性。故此，在险价值可能会阻碍多元化^[15]。

更重要的是，VaR 无法反映处于分为数下方的尾部损失。实际上，对于固定的置信区间，VaR 并不会评估 VaR 违反时的损失幅度，因此作为风险管理的一个指标，VaR 有的时候并不是最合适。

3) 预期资本不足 (ES)

预期资本不足，是另一种在金融风险度量领域中用来评估投资组合的市场风险或信用风险的概念。预期缺口是最糟糕的情况下，资产组合的预期收益。ES 方法，准确的说，是一种在险价值的替代方法，这一方法对于收益损失分布的尾部形态更为敏感。

ES 的模型如下：

$$ES_a = \{R | R \leq q_a\} \quad 2.3$$

ES 方法以比较保守的方式估计风险，主要适用于利润较低的情况，对于高利润的 q 的情况，方法本身过于偏重了严重的损失，另一方面，与折现的最大损失对比，即使是对于较低的 q 预期的短缺，该方法也并不是仅仅考虑了单个个体最坏的结果。

但是，大部分情况下，ES 方法内认为是优于 VaR 的。阿特滋纳等学者（2005）证明了 ES 方法是满足连续函数条件下的次可加性，从而满足了风险管理的需求。同时学者 acharya（2010）指出 ES 可以捕捉更加极端情况下的负收益，因此，ES 方法是一种更加有效的风险测度的方法。此外，在发生极端事件时，证券化的银行的平均预期损失会更高^[16]。因此，通常认为 ES 方法是一种比在险价值（VaR）方法相对更有效的金融风险评估的方法。

4) 条件在险价值

条件风险价值（CoVaR）是在险价值的扩展风险度量方法，用于量化在指定时间段内超出置信水平的不太可能发生的情况下平均损失。条件在险价值目前是业内采用最为广泛的风险管理方法。例如：巴塞尔银行监管委员会（BCBS）在《交易簿基本审查》（FRTB）中，有条件在险价值替代 VaR，用于计算市场风险资本。

条件在险价值（CoVaR）方法注重了金融机构之间的风险联系。准确的说，条件在险价值的方法是指金融资产 i 损失位于 VaR_a^i 的水平时，金融资产 j 所处的 VaR 风险水平。其中， $1-a$ 为置信水平。

$$\Pr\{X^j \leq CoVaR^{j/i} | X^i = VaR_a^i\} = a \quad 2.4$$

而对于溢出风险价值定义为 CoVaR 与无条件的 VaR 的差值：

$$\Delta CoVaR_a^{i/j} = CoVaR_a^{i/j} - VaR_a^i \quad 2.5$$

条件在险价值是目前学术界较为主流的风险管理的工具。学者高国华等（2011）以测量金融机构溢出风险的条件 CoVaR 模型为基础，通过对 14 家上市商业银行的个股数据进行测算分析，得到了银行系统性风险贡献度与其自身 VaR 之间并无显著线性关系，对我国银行体系而言，系统重要性银行主要是四大

国有银行，尤其以建设银行、中国银行和工商银行的系统性影响最为显著，其他股份制银行的风险溢出和传染效应远小于这 3 家银行的结论^[17]。这为研究中国银行业风险管理政策监管提供了一定的借鉴意义。即：应该注重建设银行、中国银行和工商银行的风险管理，以避免其显著的风险溢出效应和传染效应带来的整体风险水平的偏好的现象。

而学者 Boucher C (2013) 针对条件风险价值进行了改进，提出了修正条件风险价值模型。Co-CoVaR 是 CoVaR 量度的模型风险校正版本（请参见上面的定义）。CoVaR 严重依赖于分位数估计，它对极端分位数的度量的误差非常敏感，并且在很大程度上受到对金融公司极端风险的轻度错误度量的影响^[18]。

5) 边际预期不足

边际预期缺口衡量的是当市场在给定的时间范围内跌破某一阈值时，公司的预期股本损失。MES 计算简单，因此监管者容易考虑^[12]。它可以计算为一家公司在市场最糟糕的 x% 时期的平均回报率。

$$SES_a = \sum_i y_i E\{r_i | R \leq q_a\} \quad 2.6$$

其中， y_i 代表第 i 个机构在资产组合中的所占权重，即：第 i 个机构的真实收益率。而组合中第 i 个机构对于 SES 的贡献程度则可以用下式表达：

$$MES_a^i = \frac{\Delta SES_a}{\Delta y_i} = E\{r_i \leq q_a\} \quad 2.7$$

6) Garch 模型

广义自回归条件异方差模型 (Garch 模型)，源自于 arch 模型。Arch 模型对时间序列有着几个基础的理论假设。Garch 模型针对 Arch 模型对时间序列变量的第二个假设（方差恒定），进行了改进和优化。因此，Garch 模型被广泛应用于金融风险的定量化衡量。

最近，Engle 和 Sriwardane (2014) 提出了一种新的波动率模型，其中财务杠杆通过他们所谓的“杠杆乘数”来放大股权波动率。该模型的确切规范是受信贷的标准结构模型驱动的；它的参数化足够灵活，准确，足以捕获公司资产波动性随机，资产收益可能跳跃，资产冲击不正常的环境。直接应用程序封装在新的系统风险度量 (SRISK) 中，其中的核心模型是结构 GARCH (SGARCH) 模型^[19]。

也有学者 Engle R (2014) 在结合了 Brownlees 和 Engle (2010) 开发的“边际预期短缺”的基础上，开发了动态 beta 模型，该模型有解释金融公司收益动态和时区异步的几个因素。且该模型结合了 DCC 模型和 β 参数的动态变化，单变量 GARCH 模型估计了误差项的波动性，动态 t-copula 估计了创新之间依赖关系的动态^[20]。

而对非系统性风险，更多的学术研究聚焦于如何能有效的分散系统性风险。学者陈奕延，李晔等（2018）等就利用了基于单调非增次模集函数，证明了在这一函数下，有多个资产的投资组合的非系统性风险比投资多个资产的非系统性风险的组合更低^[21]。

2.2 金融市场的价格风险

在金融市场中，对于金融市场的价格风险，是一个十分重要的概念。本文将价格风险分为以下的几种类型和影响的因素。

2.2.1 企业自身因素导致的价格风险

股票价格可以很好的反应企业自身的实力。而倘若企业发生了财务造假、偷税、漏税等行为，自然会造成对于金融市场上的股票价格造成冲击。典型的，有学者使用了 1995 年至 2008 年的美股数据作为样本，进行了实证分析。其得到了公司规避税款与公司的股票价格崩盘风险呈正相关^[22]。这也就意味着，长时间的积累和隐瞒坏消息（也就是避税和逃税行为），会在一定程度释放出来，从而导致股票市场价格崩溃。无独有偶的，另外一部分学者在研究企业社会责任和企业股票价格崩溃的风险关系时，得到企业的社会责任绩效（也就是 CSR），和企业股票价格崩溃的风险是呈现出负相关的^[23]。

值得注意的是，这些企业自身因素带来的价格风险通常是不预期的。正如上述学者所做的研究成果，一些企业的会主动隐瞒自身的负面消息，而当投资者发现这一消息时，通常已经无法挽回。因此，这类因企业自身因素导致的价格风险，通常可以被视为黑天鹅事件。股票的价格风险另一方也与公司的经理

人有着密切的关系。有学者实证得到了结论：条件保守主义限制了经理人的动力和夸大业绩并向投资者隐藏坏消息的能力，从而降低了股价崩盘的风险^[24]。

同样的，根据部分学者的研究，过度自信的经理会高估了其投资项目的收益，并将负的净现值（NPV）项目误认为是创造价值。他们也倾向于忽略或解释私人观察到的负面反馈。结果，负面的项目会保留时间太长，并且累积的不良业务，会导致股价暴跌。这也就是说，拥有过度自信首席执行官的公司比没有高度自信首席执行官的公司具有更高的股价崩盘风险^[25]。但是值得一提的是，这种CEO的过度自信所带来的风险，会收到公司的会计稳健性的约束。对于公司会计稳健性与公司股价的影响关系，在另一部分的学者研究中，发现了高素质的审计师可以通过降低尾部风险直接使投资者受益^[26]。准确的说，行业专家审核员（高素质的审计师）可减轻不透明性的影响，从而降低价格风险。

更加有意思的是，另外一部分学者，通过研究得到了，年轻CEO的公司更容易遭受股价崩盘的冲击，包括因负面新闻的揭露而造成的崩盘^[27]。

当然也部分的情况是，可以被预先所察觉的。例如公司的宗教程度。在部分学者的研究中发现，在总部设立在宗教文化水平较高的公司，会在表现出较低的可能性出现股票价格暴跌的风险^[28]。对于一个公司的总部所处地区的宗教文化水平高低，这些因素又是可以被投资者提前察觉的。类似的对于董事和高级职员保险的购买情况，也是可以被察觉的。部分学者，发现董事和高级职员保险与与股价崩盘风险负相关^[29]。这就意味着一些董事的行为信息，其实也和股票价格变化是密切相关的。

因此，这个角度思考，一些因为企业自身因素导致的价格崩溃的风险不能被视为黑天鹅事件。

2.2.2 其他投资者（竞争者）的行为导致的价格风险

由于股票市场是一个零和博弈的竞争市场，因此，竞争对手的行为，往往会影响着投资者的投资行为。比如，股价同步性和崩盘风险均与专门机构投资者对公司所有权的负面消息有密切关系。并且，由于机构投资者持有大量股份和投资期限长，因此有强烈的动机进行监督。相反，对于短暂的机构投资者来

说，这种关系变得积极，因为他们倾向于交易而不是监督。这一特征表现，机构监督限制了经理对公司现金流量的提取，从而降低了经理吸收的公司特定风险^[30]。而另外部分学者在研究流动性与股票价格风险的关系时，就得到了一些相似的结论。其得到了股票流动性增加了股票价格崩溃的风险的结论^[31]。总的来说，都可以归纳为不同投资周期的投资者的行为会对彼此产生影响。

2. 2. 3 投资者的交易习惯和行为

不同的投资者有着不同的交易习惯和行为。一些量化的投资者有着自己的交易逻辑。但是无论是交易习惯、交易行为还是交易逻辑，都是存在一定的程度的失效的可能性的。也就是说，投资者的交易逻辑本身有着一定的风险。

2. 2. 4 整体市场因素导致的价格风险

在金融市场中，有一些因素可以在某些时候重大的影响股票的价格。通常来说，这些因素会影响较多的个股，具有集体性的特点。因此，此时的价格风险也可以被认为是系统性风险。

根据部分学者的研究，过量振作消费对中国国有企业崩溃风险的影响^[32]。因为了高管有享受额外的待遇的机会，国有企业的高管便有动机隐瞒坏消息，从而导致未来股价崩盘风险增加了。这一过量振作消费的政府措施，就是一个典型的整体市场的影响因素。提及到整体市场的影响因素，最具有代表性的是利率。有学者就发现有证据表明空头利益与未来的崩溃风险正相关^[33]。

当然，整体市场因素，也包括有一些情绪影响。例如，整体市场的情绪是否乐观也会与市场的价格造成影响。以分析师与公司股票价格风险为例，部分学者研究得出以下结论。以中国市场为例，公司分析师覆盖面的增加会导致股票价格崩盘风险的增加，更加乐观的分析师意味着更高的崩溃风险^[34]。

与市场情绪密切相关的另一个因素是，社会信任。社会信任通常用以衡量一个社会的社会成员之间的相互信任程度。部分学者，证明了总部位于社会信任度高的地区的公司发生崩溃的风险较小^[35]。

指数一般来说，是反应一个市场大众情绪的有力代表，也是反应一个市场的行情变化的一个有力的参考指标。股票市场的指数变化通常与这个市场的系统性风险有着一定的相关性。这其实也就为后续本文建立模型的基础，奠定了基础。

本文在后续中描述风险，主要为市场整体的系统性风险带来的股票市场的价格奔溃的风险。具体定义为：由于市场的系统性风险过高，导致的股票市场的价格崩溃（指数的暴跌）。

2.3 多元非线性拟合模型

一般来说，拟合指的是，对于一些数据进行方程化的处理，使得其误差在某种度量意义下最小。而对方程化的方式，并没有限定。拟合算是个十分广泛的概念。拟合的曲线一般可以用函数表示，根据这个函数的不同有不同的拟合名字。

一般的线形模型是以参数 b 为系数的广义多项式，即：

$$y = b_1 g_1(x) + b_2 g_2(x) + \cdots + b_n g_n(x) \quad 2.8$$

式中， g_i 为已知的 n 个线性无关的连续函数，称为基函数。

在最小二乘意义下用线性模型拟合离散点组，参数 b 可通过解方程组来确定。在实际应用中，线性数据拟合通常是采用基函数来不断逼近实验的数据的。而由于实际数据的复杂性，很多时候会采用非线性拟合模型。

非线性拟合问题，一般有两种主流的处理方式，一种是利用变量代换求解，这类方法一般用于可线性化的问题。另外一种则是，不可线性化的问题，则通常采用，最优化理论中的无约束问题的解法进行求解。

目前有许多研究工作，应用了多元非线性拟合的方法，对实际预测问题进行了求解。大部分均有着不错的效果。

部分学者就利用多元局部多项式回归的概念来表征和预测深圳股市的股指序列。具体来说，就是在基于股指时间序列的非线性和混沌，考虑了多变量局部多项式预测方法和单变量局部多项式预测方法。根据其深圳成分指数的数值结果表明，多元预测变量的预测均方误差远小于单变量预测因子，且优于现有的三种方法^[36]。另外一大部分学者，对非线性的灰色伯努利模型(NGBM(1, 1))的应用产生了十分大的研究兴趣。根据部分学者的研究，这主要是因为非线性的灰色伯努利模型在时间序列预测中，要比经典的ARIMA模型更加具有优势^[37]。因此，选择多元非线性拟合模型，来作为本文的对比基准，是非常不错的选择。虽然非线性拟合的方法多种多样，但是刻意追求函数方法，在本文中，似乎并不明智。这主要是本文中，多元非线性拟合主要是作为对比的标的。而一部分的学者，在比较了自回归模型(autoregressive models)，纠错模型(error-correcting models)和多元最近邻回归模型(multivariate nearest-neighbor regression models)分别应用于最佳投资组合的预测之后，发现，没有任何一种预测技术能胜过任何其他方法^[38]。因此，本文在非线性拟合模型的方法选择上，使用较为基础的模型。然而，尽管非参数方法本身并不能显着提高预测准确性，但是，当非线性预测得到最佳投资组合时，非线性的方法可以获得更准确的预测。

因此，采用多元非线性拟合的方法作为本文的对比基础模型，是十分不错的选择。

2.4 复杂网络

近年来，随着人们对大型或大型数据集的兴趣日益浓厚，将大型图形和网络应用于金融大数据的趋势正在上升。许多研究基于股票价格之间的关系，着重于股票市场网络结构的构建和分析^[39]。大多数由大数据所构建的网络都显示出实质性的非平凡拓扑的特征。且其中的元素之间的关联方式，既不是纯规则的也不是纯随机的，而是介于二者中间。这些典型的特征包括有分布中的粗尾，高聚类系数，社区结构和层次结构等。对于定向网络模型，还有一些特定的结构。然而，过去研究的许多网络数学模型，例如网格图和随机图，都没有

具体的显示这些特征。实际上，目前可以通过具有中等数量交互作用的网络来实现相对的复杂的结构。这对应的理论背景是：针对一般概率，获得了最大信息内容（熵）。复杂网络的两个著名的和许多研究类是无标度网络和小世界网络，它的发现和定义都在外地典型案例研究。两者均具有特定的结构特征，其中前者的特征是，幂律度分布和短路径长度，后者的特征是高聚类。

2. 4. 1 聚类与分类

构造复杂网络，进行网络结构分析是一种逐渐受到社会科学领域重视的实用分析的工具。实际上，早在 2005 年，就有学者 Vladimir Boginski 根据一定时期内的开盘价格数据计算成对股票之间的相互关系而构建了股票市场网络结构图。并且在研究了市场网络结构图的结构性质随时间的演变，并在对所得结果的解释的基础上得出有关股市发展动态的结论^[40]。

根据其研究，在所考虑的时间间隔内，市场网络结构图的幂律结构是相当稳定的，因此可以说，其研究的自组织网络的概念适用于金融领域，从这个意义上说，股票市场可以被视为一个“自组织”系统。另一个重要的结果是，市场网络结构图的边缘密度以及最大集团规模在过去几年中稳步增加，这支持了最近被广泛讨论的关于经济全球化的众所周知的想法。同时作者还指出了通过计算市场图的集团划分，将金融工具集划分为相似对象组（聚类）的自然方法。这种方法可以通过考虑分区中的拟团来扩展，从而减少得到的簇的数量。

类似的，学者 R. N. Mantegna (1999) 使用股票价格的取对数后的每日时间序列数据，使用拓扑空间用以描述股票市场的结构。该拓扑图实质是从考虑投资组合的所有股票对之间的相关系数矩阵的角度，涵盖了每日股票价格对数的差的同步时间演变^[41]。该图相关的次要超空间的层次树叶提供了较为有意义的经济分类（聚类方法）。

而对于网络结构分析中，学者 Vladimir Boginski (2004) 还已生成的图中检测出集团和独立集合等特殊形式作出了具有明确的实际解释。最终说明该技

术基于股票价格数据对金融工具进行分类，从而可以更深入地了解股票市场的内部结构^[42]。

部分国内的学者也发现了股票市场海量的数据，可以表示为具有与顶点和边相关联的网络结构。学者黄伟强采用阈值法构建中国股票相关网络，然后研究该网络的结构特性和拓扑稳定性，并表明它遵循幂律模型^[43]。在其研究中，发现所构建的网络结构是较为脆弱的。倘若针对较为脆弱的网络结构特点联系金融风险管理的方法，也许可以进行深入的了解。

随着研究的深入，学者马骏也发现了整体中国股票市场网络的度分布是重尾的，但是当中国股票市场经历熊市时，它不能遵循幂律。而且在空头时期边缘密度非常大^[44]。这意味着，当市场呈现出不同的市场风格（也就是牛市和熊市两种市场风格）的时候，市场结构和投资者的行为会发生一些变化，一些投资策略也应该及时的调整和优化。

2. 4. 2 投资组合优化与投资者行为

学者 JP Onnela (2003) 在网络分析的过程中发现在崩溃期间，由于市场的强全局相关性，树在拓扑上收缩，这表现为平均占用层的低值。投资组合优化的多样化会导致了经典 Markowitz 投资组合的资产总是位于树的外叶上^[45]。而笔者猜想，倘若使用网络结构建模的过程中，将树的外的个股形成投资组合，是否会使投资组合的风险降为相对较低的水平。这是值得探索的。

另外一个较为有趣的研究是，学者 Kydros Dimitrios (2015) 在分析了希腊市场发现，根据股票的相关程度创建不同的网络。这些网络使用社交网络分析方法进行可视化和评估后，可以得到希腊市场是一个容易受到一些大投资者或经济环境影响的市场^[46]。这是对于希腊市场的投资者行为的一个较好的解释。同时也说明了，建立网络结构后，可以通过分析其，对金融市场的投资者的行为偏好进行分析。

根据学者马骏的另一份研究，在中国股市上，处于股市网络中心位置的股票可以从利好政策中获取更大的超额收益^[47]。这就意味着构建股票市场的网络结构图，是可以有利于投资组合的优化。

2. 4. 3 市场结构特点与结构性风险评估

而在当前有学者提出过建立结构性网络，进而可以分析整体市场的系统性风险和非系统性风险，为投资组合的建立提供建议。学者 hi K (2010) 构建复杂的网络来研究某固定时间段内所有美国股票的收盘价之间的相关性。这些网络节点是股票，它们之间的联系是由股票价格、价格回报率和交易量在一定时期内的变化的相互关系决定的，且结果显示，所有基于高度相关股票价格、价格回报率和交易量的网络，都显示出一种规模-程度分布^[48]。这表明，股票价格的变化受到相对较少的股票的强烈影响。由此，部分学者提出了一种新的方法来选择股票列入股票指数，选择那些较为有影响力的个股作为新的股票成分股。

对于市场结构和投资者消费偏好，学者 Huajiao li (2016) 构建了一个由 112 个国家（和地区）组成的衍生投资网络。通过计算不同的拓扑特征，定量分析了国家对外能源投资和对内能源投资的多样性，能源投资偏好，国家（地区）和最强大的能源投资国集团之间的双边能源投资关系的强度^[49]。实际上，这有利于能源市场的投资者进行更加理智的投资和风险警示。

同时，学 Monica Billio (2016) 者已经用不同的熵测度分析了欧洲系统性风险的时间演变，并为银行业危机构建了新的预警指标，并且实证分析得到了熵测度在预测银行危机中的预测能力是有效的^[50]。类似的，Joseph Andreas C 分析了 2002 年至 2012 年的跨境股权和长期债务证券投资网络，确定了两个针对金融危机的预警指标 34。并且在运用金融危机爆发的前期的数据，进行验证，验证了其预警指标的有效性。

而学者 Thong Nguyen-Huy (2019)，应用澳大利亚的数据进行分析的过程中发现：通过多维 Kruskal - Shephard 缩放方法的聚类分析所建立的时空多样化策略可能会切实降低澳大利亚的系统性天气风险。特别是整个区域内整个保险区域的平均降低风险的效果^[51]。

同样的，马君潞等学者在 2004 年等较早年份的分析中，发现在多银行构成诱导因素组的传染中，其程度取决于诱导因素组的银行构成，中国银行是决定传染程度的关键^[3]。此外，对于传染性渠道和相关性渠道代表了引发系统风险

的两种不同机制,这两种渠道对系统风险的影响又取决于金融网络的结构特征^[52]。因此,选择合适的相关联结数据,建立复杂网络结构是至关重要。

2.3.4 复杂网络模型归纳

而笔者也发现是近年来对于复杂网络建模的方法大多使用的最简单的时序相关性。并且数据只是包含日线级别的收盘价和开盘价,交易量三维度的数据。仅从这个角度分析,数据维度是不足的,并且存在可以丰富的空间,比如市值、所属指数等

对于金融市场的网络结构建模,大致优点可以归纳为:(1)分析市场的结构化特征,比如幂结构,为新的行业分类或者指数构成提供参考。(2)规避资产组合的风险,分析市场投资者的一些投资偏好,为市场监管者理解投资者行为,更好的监管提供参考。(3)进行风险评估和警示。

首先针对日线数据来说,就成熟市场而言,个股的交易量大多出现在开盘后和收盘前,而日内的其余时间都是交易量较低。这一特征会使得在建立相关性联系的时候,过高的评估个股于个股的相关性。因此,为了提高结果的准确性,可以从数据的切分角度,进一步的提高数据的分辨率,采用更小的级别数据。针对不同数据的维度,统一使用归一化处理的方法,使得数据在同一纬度进行分析。

对于个体节点之间的关联规则和关系,在构建网络结构中是比较重要的。A.Namaki,(2011)已在其研究结果中证明了,相关系数分布对所构建的网络结构技术具有重要影响^[53]。

有学者描述了一种用于构建多元非线性时间序列网络的统计学方法。其提供了一种通用的方法检验两个测得的时间序列之间的相互作用是否具有统计学意义^[54]。这意味着数据之间本身没有明显的定性相似性,也可以应用该方法。这一方法较为的适用于中国的金融市场。

实际上,股票市场最基本的元素应该是:开盘价,最高价,最低价,收盘价,交易量。因此,从建立基础的相关性联系上,可以涵盖以上的基本元素。

在实际操作中，股票市场的数据是时序数据，也就是说当个股与个股的相关性并不是长期一致的。因此，在建立相关性联系时，应该对于所选出的数据有所要求。传统的计量经济学方法对于类似的样本的数据的通常的处理方法是选取相同时间长度的样本数据。并且随着时间的更新，而舍弃最早的，也就是距离目前最远的部分数据，从而保持数据的长度相等并且提高模型的泛化能力。考虑到在风险分析的时候，上次更新的数据会对后续的风险产生一定的影响，也就是存在“继承”的关系。因此，该模型对于数据的要求应该是，数据长度要一致，并且对于超出时间长度的数据有所“遗忘”，形成数据上“有进有出”。而 `lstm` 模型，本身具有“有进有出”的特点，比较适合基于构建的网络结构数据的金融风险预测。

2.5 长短期记忆模型

长短期记忆（英文：Long Short-Term Memory，简称 LSTM）是神经网络模型的一种。它本质是一种非线性的模型。而在实际中，非线性模型又有着较好的适用性。杨子晖等学者在其构建的“全球金融市场与经济政策不确定性”的非线性关联网络，对全球 19 个主要国家（地区）的经济政策不确定性（EPU）与系统性金融风险传染关系展开研究中，得到了一系列的结论。其中股票市场是风险的主要输出方，而外汇市场则是风险的主要接受者，两者之间存在非对称传染效应。分样本研究发现，危机期间风险传染更加明显，且 EPU 在风险传染中发挥着重要作用；同时，股票市场是风险的源头，对外汇和 EPU 具有较强的溢出效应^[55]

在 `lstm` 模型中含有四个 S 函数单元，最左边函数依情况可能成为区块的 `input`，右边三个会经过 `gate` 决定 `input` 是否能传入区块，左边第二个为 `input gate`，如果这里产出近似于零，将把这里的值挡住，不会进到下一层。左边第三个是 `forget gate`，当这产生值近似于零，将把区块里记住的值忘掉。第四个也就是最右边的 `input` 为 `output gate`，他可以决定在区块记忆中的 `input` 是否能输出。

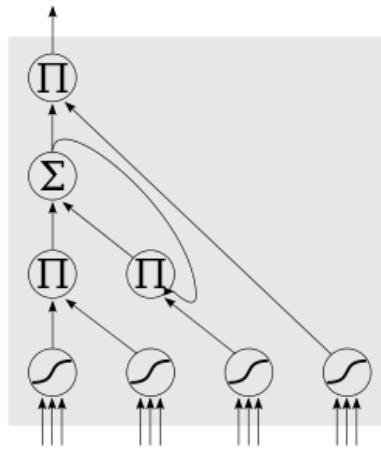


图 2.1 LSTM 结构图

将 LSTM 模型逐步分解，可以得到以下几个主要的内容。

2.5.1 遗忘门

在 LSTM 模型中，遗忘门的设置主要是需要实现，到底要哪一些的信息传入到模型，也等效于过滤掉哪一部分的信息。这一过程可以用以下的函数公式表示：

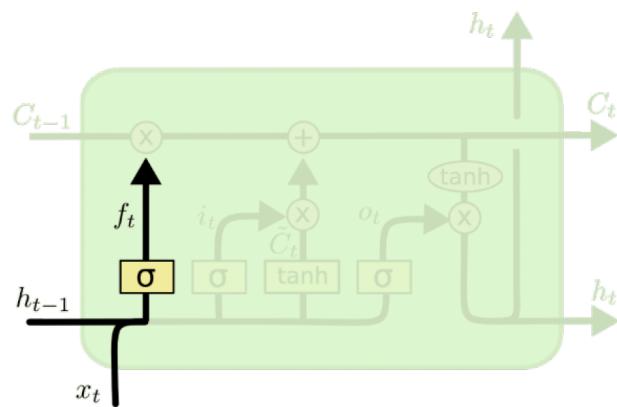


图 2.2 遗忘门

$$f_t = \text{sigmod}(W_f * [h_{t-1}, x_t] + b_f) \quad 2.9$$

2.5.2 传入门

传入门主要实现的功能是，决定让多少新的信息加入，并且更新 cell 状态。具体可以用如下的函数表达：

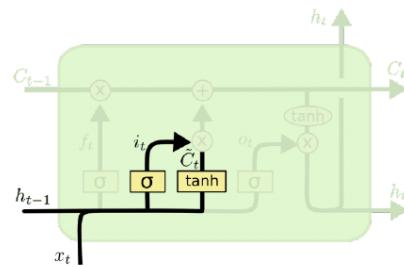


图 2.3 传入门

$$i_t = \text{sigmod}(W_i * [h_{t-1}, x_t] + b_i) \quad 2.10$$

$$\tilde{C}_t = \tanh(W_C * [h_{t-1}, x_t] + b_C) \quad 2.11$$

2.5.3 输出门

输出门，主要用以决定输出结果。这个输出结果主要是依赖于 cell 的状态。但是还经过一个过滤的处理。可以用以下的函数表达：

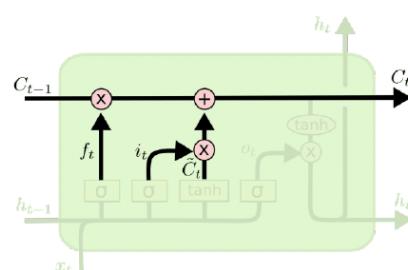


图 2.4 输出门

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \quad 2.12$$

实际上，由于独特的设计结构，即：遗忘门的存在，LSTM 适合于处理时间序列中间隔相对较长的数据。学者姚小强针对传统方法对多噪声、非线性的时间序列无法进行有效预测的问题，以多尺度特征融合为切入点，提出并验证了基于树结构长短期记忆（LSTM）神经网络的预测方法^[56]。并且将该方法应用于黄金期货的交易，实现了良好的回测效果。此外，学者孙瑞奇针对股票价格数据影响因素多的问题，选用收盘价、开盘价、最高价、最低价以及成交量这几个最关键的影响价格因素，作为神经网络每一层神经元的自变量，使用中美股市代表性的上证综指以及标准普尔 500 指数和道琼斯工业指数来进行研究^[56]。并且，根据其结果，该预测结果较为准确。

3 因子库与标签的建立

3.1 整体概述

本文采用的预测和评估方法是利用因子来预测成分股，再利用成分股个股的预测结果合成指数的预测结果。因此，因子在这里起到了十分重要的作用。在本章，先给出了，生成个股因子数据的数据来源的解释和说明。并且对这些原始的数据进行了符合文章数据要求的处理。这里，处理指的是对于原始股票数据前复权和后复权的数据处理。因为，未经复权的数据，经常性的会造成股票数据在价格上的不连续。这无论是对机器学习的预测还是对于生成原始的因子数据都是不利的。因此，首先需要进行统一复权的处理。

其次，对于设计因子而言，需要给出设计因子的逻辑。但是在给出逻辑之前，首先需要明确因子用于预测的对象。这里本文最终预测的是沪深 300 的指数数据。因此，预测的对象应该是指数。但是由于本文是通过预测个股合成指数，再来预测指数的。因此，个股也是本文的预测对象。但是，对于机器学习的模型而言，标签应该是只有一个的，那就是指数的收益。

对于标签——指数收益，应该需要考虑以下几个方面，使用传统的当天收盘价（close）到当天开盘价（open）的计算方式是存在一定问题的。因为对于指数而言存在跳空等现象，倘若使用传统的计算方法，容易出现收益率不连续的

情况。这一弊端的影响，类似个股使用未复权数据进行预测的情况。因此，为了将跳空情况考虑在内，本文使用计算指数收益率的时间范围为，当天指数的收盘价（close）到次日的收盘价（close）的指数收益变化的时间。在给定计算指数收益的之间之后，同样需要给予指数收益的计算方法给予定义，这里采用两种计算方法，一种为对数化的计算收益方式，另一种为非对数化的计算收益，具体定义在下文的公式中给出。同时，在计算标签的时候，还需要给予标签一个索引数值，这是为了将标签和因子数据最终匹配起来。这里采用较为方便且好分辨的时间作为索引数值。值得注意的是，根据本文定义的指数收益率计算时间范围中，索引数值应该被定义为次日的收盘价的时间，也就是次日的日期为这一收益率的日期索引数值。但是本文涉及的是一个预测模型，因此，模型的因子数据与标签数据的时间上，应该被要求错开。也就是：标签数据为 T1 日，而因子数据为 T0 日。这样子才可以避免未来函数的影响，提高模型泛化能力。在实际操作中考虑到因子数据的庞大，而标签的数量的较少的情况，且为了方便后续代码程序中的操作，故此，本文将标签的索引值集体向前移动一天。这也就意味着，标签原先的日期索引值为次日的日期，现在变为了当日的日期。这样子，在后续的预测模型的因子数据和标签的匹配中，只需要将具有相同的日期索引值的数据匹配在一起就可以了，这既避免了模型的未来函数的出现，提高了泛化能力，也方面了程序实现的代码工作量被大大减小了，并且代码错误风险降低了。

在规范定义且计算了标签之后，将进行因子设计。在设计因子之前，需要解释和定义一些因子设计的方向。在本文将从以下几个方面进行因子设计。第一类：单一指数类，单一指数类，指的是本文预测模型的预测对象为沪深 300 的指数，因此，沪深 300 指数的一些基本形态特征，本身对于预测模型就是有着积极的作用的。因此从单一指数的方向提取因子是可行的。第二类，跨指数的对比，跨指数的对比具体而言就是沪深 300 与中证 500 的对比。在实际交易中沪深 300 指数往往代表着市场中大市值的股票，而中证 500 指数往往代表着市场中小市值的股票，这主要是因为指数的定义不同而导致的。两类指数的对比（对比的强度等），通常会反应市场的一些特征，这是对于预测指数是起到积极的作用的。第三类，期现对比，也就是指数与其对应的期货的对比，期货作为现货的衍生品，本就是围绕着现货的价格上下波动的。因此，期货与现货的对比，也可以对预测现货价

格起到积极的作用。第四类，期货的远近合约对比，远近期货合约对比主要是依靠价差来反映一个基本的市场情绪，从而对预测现货的价格起到积极的作用。第五类，高波动率的个股统计，高波动率的个股反映的是市场中情绪最为激烈的票，因此，对于预测指数会有着积极的作用。第六类，特殊个股统计，也就是 st、近期复牌、与沪深 300 指数低相关的个股。这些个股一旦集体出现某种特征，往往代表着市场中出现了某种相对固定的风格。因此，对于预测指数有着积极的作用。第七类，国际市场特殊品种，国际市场的特殊品种有包括黄金、白银、原油这三个主要品种，选择这一分类的原因很简单，这些品种的交易，和股票市场是互补市场，因此，这类品种交易会影响沪深 300 指数的走势，对交易起到积极的作用。第八类，国际指数，国际指数对于沪深 300 的指数有着较大的影响，比如日经指数、韩国综合价格指数等，因为时区原因比国内早开盘半小时到一个小时，因此，这些指数在这段时间内的表现很有可能会影响国内沪深 300 指数的走势。其次，美国的纳斯达克指数、标普 500 这些具国际影响力的指数，本身就是可以反映全球股票的一种基本市场情绪的。因此，这也是对预测模型起到积极的作用的。

在上述因子设计的逻辑下，进行因子设计后，可以得到一些列的初始因子。但是这些因子并不是都可以使用的。因此，要进行因子有效性检验。也就是这一因子，究竟对于预测模型的标签是否有着积极的作用，值得检测。因为上述的因子的逻辑，都是在笔者的猜测下进行，是否真实有着积极的作用，仍然需要检测。因此，在下文中，会采用三种计算方法进行验证，分别为 xy 图，相关性检测，线性拟合的 K 数值。通过这三种方法得出初步的可用因子的结论。

而接下来，需要进行再一步的共线性检验。排除一些类具有高度线性相关的因子，减少代码的工作量以及模型的运算负荷。这里需要注意的是，通常模型要是仅仅用于预测的话，一般可以不处理多重共线性问题，因为这并不会极大的影响预测结果。但是考虑到本文的因子数量较多，如果不做处理，算力负荷过大。因此还是需要进行处理。造成共线性的原因可能会有以下的几种基础情况，第一类，因子都享有共同的时间趋势，第二类，因子之间存在相互的时间滞后关系。第三类，一些因子数据之间本身就是存在类似线性的映射关系。而对这些因子的检验，也不能仅仅依靠相关性检测，因为，在共线性检验中，相关系数高的，表明是存在多重共线性，但是相关系数低的，并不代表着不存在共线性，因此还

需要进行 vif 检验和条件系数检验，同时辅助以正则化的方法进行处理。

值得申明的是，需要先做的是因子有效性检测，在进行的是因子共线性检测，因为倘若先进性因子共线性，存在将一些有效因子错误排除的情况。具体情况会在下文中有详细的说明。

最后需要对一些因子的分布形态做进行的删选，将一些因子分布形态上并不接近对数正态分布的因子进行排除。原因是股票市场的回报率基本符合对数正态分布，因此采用相近的形态的因子进行预测，或许在准确性上有所提升。同时，对于接近对数正态分布的定义为：正态分布、卡方分布等，因为这些因子分布基本可以通过模型的迭代得到函数变化，转为对数正态分布。而对于一些被排除的因子，可以进行简单的优化得到对数正态分布，这在下文中也有处理流程。

3. 2 数据来源与预处理

3. 2. 1 原始数据来源

首先，本文的所采用的全部数据，来自同花顺 IFIND 金融数据库。本文选择使用三大指数之一的(沪深 300) 的表现来衡量股票市场价格风险。但是在价格风险测度过程中，价格风险每时每刻都是在处于波动中的。根据假设中的简化原则，本文中将价格风险的最短时间间隔定义为每一天，也就是每天更新一次。

3. 2. 2 复权日检测与复权化处理

对于股票的原始数据，需要复权处理的。准确的说，股票在除权和除息之后，股票的价格会发生一定的变化。但是实际上，持有股票的投资者的持有成本并没有发生变化。这一特点，会在预测模型中产生巨大的偏差。倘若不进行处理，很容易造成模型的精确率很低的情况，甚至在一些深度学习的模型中，极有可能造成模型不收敛的情况。具体来说，由于股票在除权和除息之后，股票的价格会有所下跌，而这个下跌幅度和正常的股票的价格变动叠加在一起，会使得在计算股票的收益率的时候，除权和派息的这一天的股票，价格波动要比正常情况更大(也

有可能是更小)。这就会导致在全局上(整个数据集的时间范围上),个股的收益率呈现出不连续的情况,从而对预测模型造成偏差,且对深度学习模型的收敛情况造成干扰。

对于数据的未复权的情况,由于同花顺数据库,提供了股票市场的复权化处理。但是出于谨慎行的考虑,本文进行第二次检测(考虑到数据库数据可能潜在出现问题)。本文采取的措施是,先进行检测,在进行复权化处理。首先,检测的数据是全体股票市场的数据。通常,股票市场上公司的股票的涨跌幅是有限制的。涨跌幅限制是国内股票市场的一种制度。主要目的是维持股票市场的稳定。也可以理解为抑制过度的投资行为,防止股票市场出现大量的暴涨暴跌现象,维持金融体制稳定的一种措施。国内的涨跌幅限制发展历史比较特殊。在历史上,实行过无涨跌幅限制,但是一度出现个股的暴涨暴跌现象,比如延中实业(现方正科技,600601)曾出现过一日涨幅超过100%的现象。这对于维持金融市场稳定是不利的。因此,在经过多次探索,国内股票市场实习,当日涨跌幅不超过10%的涨跌幅限制。而对于一些ST的个股,涨跌幅限制为5%。

根据这一特点,通过搜索涨跌幅超过20%的日期和普通个股寻找出来,并且将涨跌幅超过10%的ST个股寻找出来。在进行,注意确认。需要确认的是有以下的一些基本情况,涨跌幅限制是出现变化的。第一类,新股上市的首日,第二类,股改股票完成股改的首日,增发股票上市的当天,第三天退市股票恢复上市的当天。在排除这一些情况后,在结合时间进行,确认,是否存在遗漏复权的情况。倘若没有遗漏,则进行下一步,若有遗漏,怎进行复权化的处理。复权化的处理,有两种一种是向前复权,一种是向后复权。向前复权,就是保持现有价位不变,将以前的价格缩减,将除权前的K线向下平移,使图形吻合,保持股价走势的连续性。向后复权,就是保持先前的价格不变,而将以后的价格增加。上面的例子采用的就是向后复权。两者最明显的区别在于向前复权的当前周期报价和K线显示价格完全一致,而向后复权的报价大多高于K线显示价格。二者并没有好坏区分,只是计算口径不一样而言。在本文采用后复权处理。

后复权: 复权后价格=复权前价格×(1+流通股份变动比例)-配(新)股价格×流通股份变动比例+现金红利

3. 2. 3 额外信息说明

最后在这里进行解释，为什么只是利用涨跌幅进行基本的检测，而不是引入第三方提供的复权日信息，进行检索和确认。首先，同花顺金融数据库是提供了后复权的处理的。但是针对其可能潜在存在数据遗漏的情况，本文才进行检测的。倘若引入第三方的复权日信息，首先需要确认，第三方的数据是否真实，否则会造成由于为了避免一个小误差，而造成更大的误差的情况。其次，上述复权日检测是通过检测异常的涨跌幅日期，在通过信息检索，来确认是否有遗漏复权日的情况。在大多数情况下，这一方法是成立且有效的。但是也存在一种极端情况，就是除权派息，通常会导致股票价格下跌，而倘若股票价格下跌（除权派息带来的）与次日的股票正常的涨跌幅相抵消，就会使得股票的价格落在了正常的范围内。这就会造成误差。

对于这种情况，通常的除权派息日，股票一般波动较小。因此，概率比较小。但是出于谨慎原则，还是需要进行额外的处理。

由此，本文对于复权日的检测做了进一步的优化，具体如下：涨跌幅的限制是从股票的前一日收盘价开始计算的。且股票当天的涨跌幅的计算为当天的开盘价到当天的收盘价。除权派息对于股票的价格影响，主要是在前一日的收盘价到当天的开盘价的这一段时间。这一段时间，也反映了股票的跳空等现象的收益情况。因此，优化的复权日检测过程如下：

首先，计算股票前一日的收盘价到当天的开盘价的收益。倘若这一收益超过涨跌幅的比率，就会存在潜在的遗漏复权日的情况。

其次，在上述检测情况之后，在引入巨潮网的公示信息，再做逐一检测。确保原始数据，不存在问题。

3. 3 标签的生成与处理

3. 3. 1 标签的定义

标签用通俗的定义来说，就是结论，也就是因变量。对于标签的处理，相当

的重要。如果标签是分类的标签，那么本文的预测模型，就是分类模型。而倘若标签是回归数值，那么本文的预测模型就是回归模型。从深度学习的角度，标签的特征一般确定了学习模型的基本雏型。而在通常的预测模型中，标签也可以理解成因变量。在本文中，将标签定义指数收益率。下面给出标签的具体计算方法。

3.3.2 标签的计算方法

首先，在计算标签之前，明确标签的需求。根据实际模型，本文需要一个用以匹配因子数据的索引值和标签的实际数值。先准备标签的实际数值。这里由于在 3.2.1 的内容中，定义本文所确认的风险是以日为最小变化时间单位的。因此，标签的最小变动时间单位也应该是日。对于沪深 300 指数的日回报率的计算定义为：

$$\text{return}_{\text{no log}} = \frac{(close_t - close_{t-1})}{close_{t-1}} * 100\% \quad t \geq 2 \quad 3.1$$

对于这一计算方法，是交易所所使用的基本收益率计算方法。但是这一方法在预测模型中，可能存在一定问题。以下做具体说：

这一计算方法所确定计算收益率的时间段是合理。相较于计算当天的开盘价 (open) 到当天的收盘价 (close) 的时间段的收益而言，从前一天的收盘价到当天的收盘价的时间段的收益，包含了指数隔夜的收益变化情况。这是真实收益的重要组成部分。就以下图为例，说明隔夜变化的重要性：



图 3.1 沪深 300 指数 K 线图节选

在下图中，2020年3月25日的沪深300指数的开盘价为3711.48，而收盘价为3722.52，前一日，也就是2020年3月24日的收盘价为3625.11。由此，可以计算出，倘若按照当天的开盘价到收盘价的时间段的收益来作为当天的收益，则2020年3月25日的收益为0.3%，而倘若使用了前一日的收盘价到当天的收盘价的时间段的收益作为当天的收益的话，结果为2.69%。显然，隔夜部分的收益带来了巨大的影响。因此需要将其考虑进标签的计算。

其次，对于取得百分比的收益方式是存在不合理的地方的。虽然这类计算方式是很容易被大众所掌握，但是对于预测模型和金融分析，这是不利的。具体来说，从金融分析的角度，假设当股价从1元上涨为2元，则投资收益增长了100%，但是当股价又从2元跌回至1元的时候，收益却亏损了50%，但是实际价格已经回到了以前的正常水平。这样子显示收益，很容易误导投资者。从预测模型的角度而言，这样子计算得到的收益，通常会出现不连续的现象，导致在模型求解的时候出现不收敛的状况。

3.3.3 标签计算方法的优化

根据上文分析，需要对标签的计算方法进行优化。具体优化过程，则是，引入对数收益率的概念。

$$\text{return}_{log} = \log\left(\frac{\text{close}_t}{\text{close}_{t-1}}\right) * 100\% \quad t \geq 2 \quad 3.2$$

对于采用对数收益率的实际好处，会在下文的lstm模型分析地方有具体的对比。在这里，就先以，两种计算方法得到的数据的平滑程度（导数）做初步的判断。导数是很好的衡量曲线是否平滑的工具。

首先计算出，两种计算收益方式下的沪深300指数累计收益对比图。结果如下：

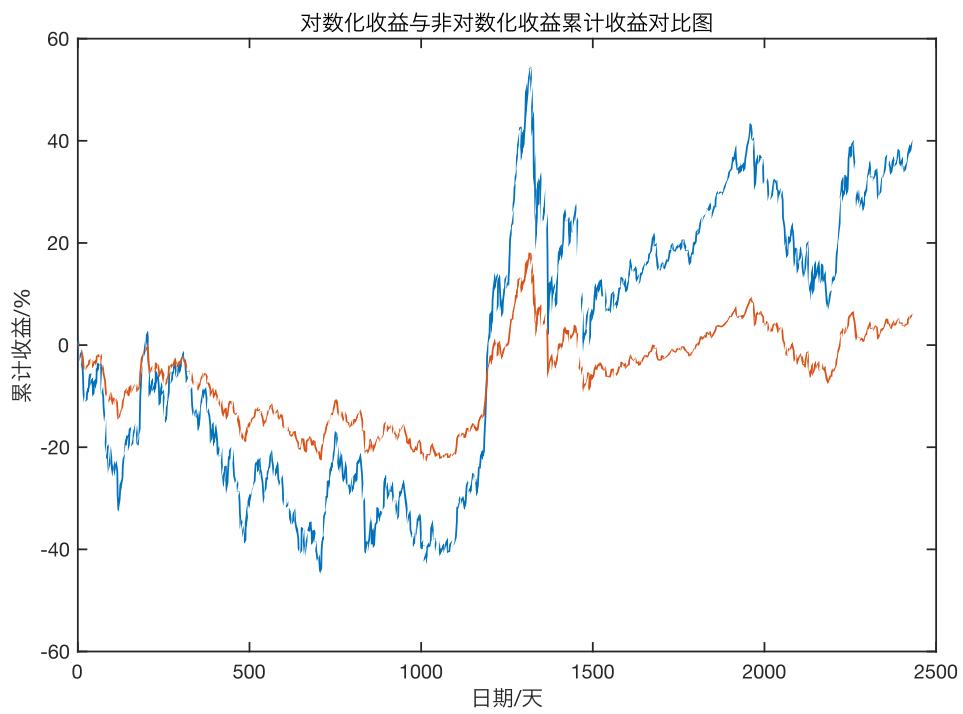


图 3.2 对数化收益与非对数化收益累计收益对比图

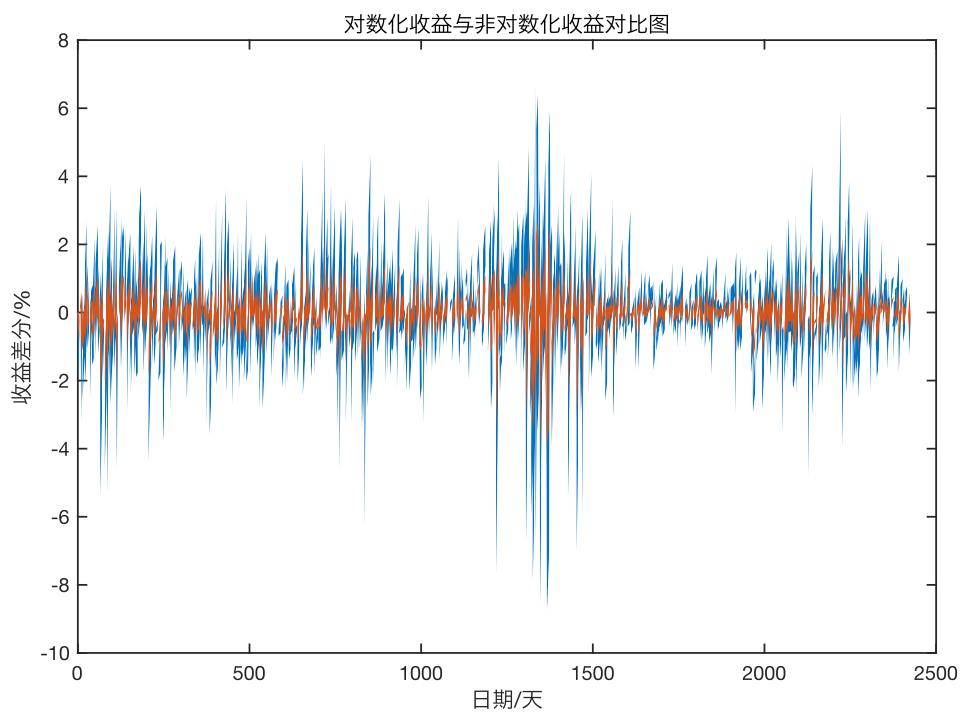


图 3.3 对数化收益与非对数化收益收益增量对比图

初步观察两条累计收益曲线，显然红色的曲线，也就是对数化收益的累计收益曲线，会更加平滑。具体结果，再对这两天曲线每一天的导数（做差），求得上述的第二张图。

很明显，对数化的收益的差值更加小，相对更加平滑了，基本满足了上文中提出的，标签相对平滑的要求。

3.3.4 标签索引值的确定

对于预测模型而言，需要给标签一个确定的索引值。这是为了将标签与因子数据相匹配起来。按照逻辑，需要确认的是， T_0 日的因子数据，应该是要匹配 T_1 日的标签数据的。这是为了明确避免未来函数的出现，导致模型的泛化能力不高问题的出现。在金融市场的预测模型中，本文就使用最常见的时间日期索引。也就是，在计算前一日的收盘价到次日的收盘价的过程中，应该将时间索引确定为次日的收盘价的时间。这是根据最晚的得到的必要计算信息是在次日的时候得到的所确定的。但是在实际代码和程序执行的过程中，发现这类错开一日的，索引匹配是十分麻烦的。因此，从简化程序设计的角度而言，这里将所有的标签的时期索引提前一天。这就使得，在后续的时间索引中，可以直接检索相同的时间索引，来完成匹配工作。这可以极大的提高代码的执行效率。

由此，标签的时间索引值确认为在计算前一日的收盘价到次日的收盘价的过程中的，前一日的日期。

3.3.5 标签的进一步思考

至此，对数化的收益标签，基本满足平滑的标签、反映指数的真实情况的要求。但是这一切的前提是直接采用沪深 300 的指数收益率作为标签。而根据这一前提，后续的 1stm 模型的雏型也就会被限定在回归模型上。这里需要论证的是，为什么不采用分类模型？现在给出一些论证：

倘若，本文采用分类模型，则，在标签就应该变为分类标签。也就是说，需要在计算出 2430 个收益率之后，进行分类，将其分为不同层次的标签。而对于

标签分类的基础，一定是根据收益率的排序。也就是说，应该在对 2430 个交易日计算所得的收益率进行排序，假设分为 10 类，则，按照排序结果十等分。每一类给予一定的标记（假设以 1-10 标记，收益率最高的日期到收益率最低的日期）。

这一方法在理论上，似乎是可行的。但是在实际操作中发现，沪深 300 的指数收益率的分布，并不是均衡的。

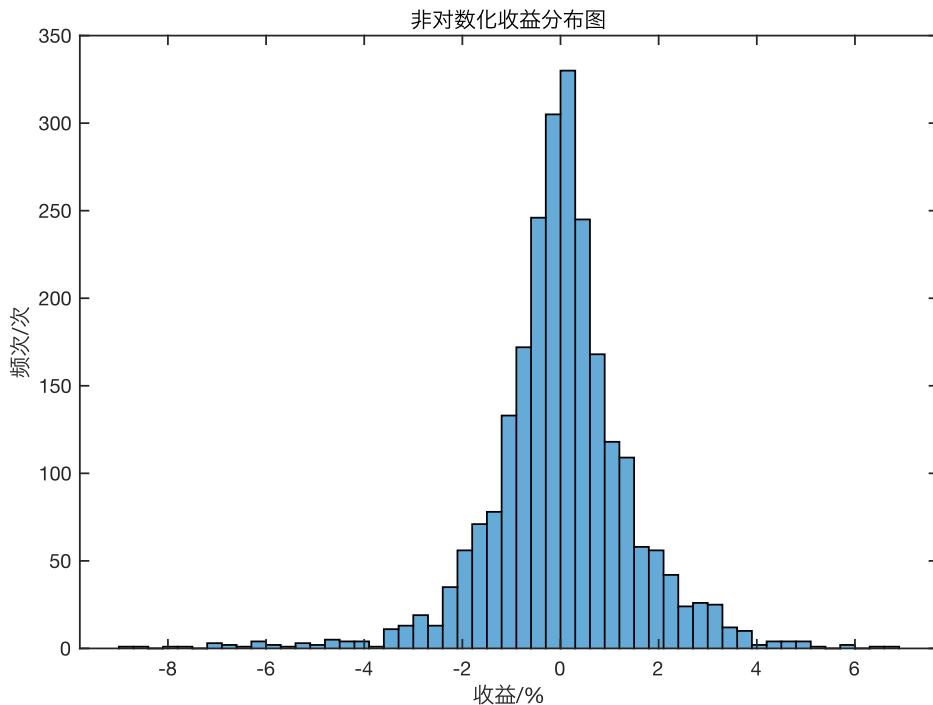


图 3.4 非对数化收益分布图

可以观察到，在-1%到 1%的指数收益的情况下，几乎占据了大部分的样本案例。这就导致了三个问题的发生。第一个：由于极端亏损和极端大涨的节点较少，标签数量太少的情况下（假设十等分），代表最差收益的那个标签所对应的实际日期中，实际上不仅仅是最差收益的日期之和，也包含了一些中等亏损的日期。对于极端大涨的对应标签，所包含的日期，也不仅仅是最好收益的日期之和，也包含了一些中等好的日期。这也就是分类标签数量难以确定，即使确定了，也不一定能反应真实的金融情况的问题。这一问题会导致模型可能预测结果是正确的，但是实际应用却造成了损失。因此是一个相对严重的问题。第二个：在两个相邻

的分类标签上，比如标签 1 与标签 2，由于标签的划分是人为给定，很有可能会造成 1.1% 的收益被归入标签 1，而 1.2% 的收益却被归入标签 2。这一问题会随着标签数量的增加而变得更加严重。第三个，由于收益率的分布，集中在 -1% 到 1%，若不采用排序收益率后，数量等分的方式，而是采用收益率分类的当时，很有可能在模型分类的过程中，无论样本数据是什么，都将其判别为标签 1，而就算判别错误，受到了惩罚，也没关系。因为这一类的标签的数量是最多的。这一问题，相对来说，可以通过模型的参数进行化解，但是还是会造造成不小的麻烦。

综上所述，本文先不考虑分类模型和分类标签。原因在于，标签类别的给予，若采用排序收益率后的数量等分的方式，过少的标签类别容易造成问题 1，过多的标签类别容易造成问题 2。采用收益率分类的方式，则容易造成问题 3。

当然上述问题也不是没有解决的方式，对于问题 1 和问题 2 可以引入聚类的概念，对收益率进行分类，按照类间差异最大化，类内差异最小化的原则，进行分类，从而在确认在采用排序收益率后的数量等分的方式下，确认最优的分类标签数量。对于问题 3，可以通过设计惩罚函数，实现对不同数量的标签的判别错误后，给予的错误惩罚是不一样的。对于数量较少的标签，一旦判别错误，给予更多的错误惩罚，来实现在判别效率上的一致性。最终解决问题 3 的现象。

但是上述方法，本文不给采用的原因在于，代码实现难度较大，同时也涉及一些其他的问题。比如聚类中的判别方法等问题。因此，本文的标签，还是遵循直接采用沪深 300 的指数收益率作为标签的原则。

3.3.6 标签最终结果

根据上述论证和处理，得到了如下的标签数据：

表 3.1 索引未处理标签数据表（节选）

日期索引	对数化标签	非对数化标签
20100106	-0.272711943363993	-0.625975017115407
20100107	-0.870328468999918	-1.98405863801024
20100108	0.108330183754440	0.249750825301172
20100111	0.0239535727715408	0.0551703528316492
20100112	0.654458655943917	1.51835843827630
20100113	-1.42087428138289	-3.21874328131896
20100114	0.603971268511792	1.40041038951929
20100115	0.171049446053878	0.394632536285139
20100118	0.223135946133856	0.515111664953458
20100119	0.0842789948783637	0.194247974679210
20100120	-1.42283462434131	-3.22311174974626
20100121	0.180536022853253	0.416564784072741
20100122	-0.543230164696848	-1.24304327034505

上述结果为 Matlab 计算所得，最后将上述结果在进行索引值的改进，得到了下表：

表 3.2 索引处理标签数据表（节选）

日期索引	对数化标签	非对数化标签
20100105	-0.272711943363993	-0.625975017115407
20100106	-0.870328468999918	-1.98405863801024
20100107	0.108330183754440	0.249750825301172
20100108	0.0239535727715408	0.0551703528316492
20100111	0.654458655943917	1.51835843827630
20100112	-1.42087428138289	-3.21874328131896
20100113	0.603971268511792	1.40041038951929
20100114	0.171049446053878	0.394632536285139
20100115	0.223135946133856	0.515111664953458
20100118	0.0842789948783637	0.194247974679210
20100119	-1.42283462434131	-3.22311174974626
20100120	0.180536022853253	0.416564784072741
20100121	-0.543230164696848	-1.24304327034505

3.4 因子逻辑建立

3.4.1 因子逻辑论证

在进行设计因子之前，需要先对因子设计的逻辑进行定义与论证。预测模型的本质是利用 T0 日期的因子数据，预测 T1 日期的标签。那么，在理想条件下，因子数据是和标签高度时序相关的。那么进一步放松条件，就可以得到 t0 日期的因子数据，应该也是和同期的标签高度相关的这一逻辑结论。而同期的标签数据已经是可以得到的，不存在未来函数。由此，因子挖掘可以从同期的标签得来。这是本文的第一个因子挖掘的逻辑。

其次，因子数据挖掘还可以遵守一个简单的贝叶斯逻辑。理论化的描述为：当 T1 日期的标签呈现出某一种特征的时候，统计全局样本中，出现这种特征的次数，记做 a，再统计标签出现这种特征的日期中，出现因子特征的日期次数，记做 b。求得 $e = b/a$ 的比值。接着寻找统计全局样本中，出现因子特征的日期次数，记做 c，在统计出现因子的特征的日期的次一日的出现标签特征的日期次数，记做 d。求得比值 $f = d/c$ 。

倘若 e 、 f 二者的比值越接近于 1，则该因子的有效性越高，成立的概率越大。等价的，定义 $q = e*f$ ，若 q 越接近于 1，则该因子有效的概率越大。

具体内容如下：假设：在当天指数出现了大跌的背景下，发现前一日的高波动个股市值变小。则进行因子验证。先统计所有的市场指数大跌的日期，并且进行标记 a。其次，将这些日期的前一日的高波动个股的市值情况进行分析，是否有所变小的趋势，将变小的日期的数量，标记为 b。如果 b/a 的比值非常高，则进行第二阶段的分析，寻找出历史上所有的高波动个股的市值情况进行分析，变小的趋势的日期数量，标记为 c。接着，寻找次日的指数表现，是否有大跌的现象。如果有，标记为 d。如果 d/c 的比例，也很高，则说明前一日的高波动个股市值变小与次日的指数大跌有着密切的联系。由此，前一日的高波动个股市值变小可以做一个有效的因子。

上述两种因子挖掘的方法，是本文出现两种主要的因子挖掘的方法。下文将阐述基于这些方法设计的因子逻辑以及对遵守因子逻辑而设计的因子，进行有效性检验。

3.4.2 因子逻辑设计

所有的因子设计逻辑都是围绕着预测标签展开的。在遵从上述两种基本的因子挖掘方法，以及综合了平时观察市场所得。本文提出了以下的几种因子逻辑，用以设计因子。

1) 单一指数类

这一因子具体解释为，本文的预测对象和标签为沪深 300 的指数收益。因此，沪深 300 指数本身就是很好的因子设计材料。沪深 300 指数的 K 线图的形态、交易量、开高低收的变化等，本身都可以用以设计因子。

当然在这里也不应该忽视的是，指数的 ETF。交易型开放式指数基金，通常又被称为交易所交易基金（Exchange Traded Fund，简称 ETF），是一种在交易所上市交易的、基金份额可变的一种开放式基金。沪深 300 的 ETF 也是一种重要的因子设计的素材，也不应该被忽视。

这里先列出一系列的因子猜想：

- (1) 沪深 300 指数，最高点出现日内的第几个小时，因子数值范围为 1, 2, 3, 4;
- (2) 沪深 300 指数的 body 最大的时候，出现在第几个小时，因子数值范围为 1, 2, 3, 4;
- (3) 沪深 300 指数一天中出现 tr 技术指标出现最高价的小时是第几个小时，因子数值范围为 1, 2, 3, 4;
- (4) 沪深 300 指数是否破前日高点，突破因子数值为 1，否则为 0;
- (5) 沪深 300 指数是否破前日低点，突破因子数值为 1，否则为 0;
- (6) 沪深 300 指数的收盘价是否大于 vwap 数值，突破因子数值为 1，否则为 0;
- (7) 沪深 300 指数成交量大于前日，突破因子数值为 1，否则为 0;
- (8) 沪深 300 指数成交量在最近 20 个交易日排在前 5，是则，因子数值为 1，否则因子数值为 0;
- (9) 沪深 300 指数，最近一次 body>atr 阳线，距离今日是交易日距离为多少，因子数值为交易日距离;
- (10) 沪深 300 指数，5min 下，连续 12 个 bar，令 $a = \text{最大最高价} - \text{最小最低价}$, 全日当中， $a > 0$ 的最大值出现在第几个小时;
- (11) 同上， $a < 0$ 的最小值出现在第几个小时。注释：a 理解为一波连续的上涨或下跌，本因子目的是找到最大的上升浪或下跌浪的近似值;
- (12) 沪深 300 指数是否出现快速反弹，出现因子数值为 1，否则为 0;
- (13) 沪深 300 指数是否出现快速反杀，出现因子数值为 1，否则为 0;
- (15) 沪深 300 指数是否出现小时级别，昨日的最高价>今日的最高价。并且昨日的最低价<今日的

最低价，如果没出现=0，出现，则等于出现的小时序号 1~4，注释：这一因子可以用于描述，一天中四个小时同期的变化。若因子数值为 1，则表示指数第一个小时的价格波动在变小。

2) 跨指数对比

跨指数对比，通常对比对象为沪深 300 指数与中证 500 指数，有时也可以选择为沪深 300 指数和上证 50 指数，沪深 300 指数和创业板指。这些对比是十分有意义的。首先，股市指数是由证券交易所或金融服务机构编制的表明股票行市变动的一种供参考的指示数字。因此，股票市场本身就是一种反映市场大众的投资者的一种指示。同时根据不同的编制方法，入围的股票数量和种类是有所不同的。由此，带来的指数也有所不同。这就会导致不同指数反应的市场特征是不一样的。例如，由于编织方法的不同，沪深 300 指数，强调规模和流动性作为标准，并赋予流动性更大的权重。由此，选出来的成分股通常具有流动性高，规模较大的特点。通俗的来说，就是市场中的市值较大的股票居多的一种指数。主要可以用来反应市场投资的需求。而作为重要的对比对象，中证 500 指数，全部 A 股中剔除沪深 300 指数成份股及总市值排名前 300 名的股票后，总市值排名靠前的 500 只股票组成，综合反映中国 A 股市场中一批中小市值公司的股票价格表现。而做为投资者，虽然在股票市场上，不是简单看作零和博弈的过程，但是两种指数的对比，基本可以反映市场上的资金流向的信息。这就使得这种对比产生了价值，是值得挖掘因子的。更加具体而言，这类对比所挖掘的因子，可以通过反应市场上的资金流向，而来预测本文的标签。

基于此，本文在给出一些的因子猜想：

(1) 沪深 300 指数的成交量增量是否大于中证 500 指数的成交量增量，是则，因子数值为 1，否则因子数值为 0；(2) 令 a 为指数的收盘价与成交量，取过去 20 日的数据，计算相关性的结果，沪深 300 指数的 a 是否大于中证 500 指数的 a ，是则，因子数值为 1，否则因子数值为 0；(5) 令 ret 等于指数的收益回报率， $reta=$ 收益回报率与 20 日均值的比值，沪深 300 指数的 $reta$ 是否大于中证 500 指数的 $reta$ ，是则，因子数值为 1，否则因子数值为 0；(6) $reta$ 在最近 20 个交易日的排名，沪深 300 指数的是否大于中证 500 指数的，是则，因子数

值为 1，否则因子数值为 0；(7) 令 $a=\text{沪深 } 300 \text{ 指数的连续 } 20 \text{ 日 ret}$ ； $b=\text{中证 } 500 \text{ 的连续 } 20 \text{ 日 ret}$ ； c 为二者的相关性，因子数值为相关性数值)， c 的数值作为因子；(8) 今日 $dhigh$ 是否昨日的 $dhigh$ ，是则，因子数值为 1，否则因子数值为 0，其中 d 意味着做差分。这一因子意味着今日的指数差在扩大；(9) 今日 $dlow$ 是否昨日 $dlow1$ ，是则，因子数值为 1，否则因子数值为 0；(10) 今日的 $(dclose-dopen)$ 是否大于昨日的 $(dclose1-dopen1)$ ，这意味着指数差的差距在缩小，是则，因子数值为 1，否则因子数值为 0。

3) 期现对比

期现对比，也就是指数与其对应的期货的对比，期货作为现货的衍生品，本就是围绕着现货的价格上下波动的。因此，期货与现货的对比，也可以对预测现货价格起到积极的作用。期货本身具有发现价格的作用，由于期货交易是公开进行的对远期交割商品的一种合约交易，在这个市场中集中了大量的市场供求信息，不同的人、从不同的地点，对各种信息的不同理解，通过公开竞价形式产生对远期价格的不同看法。期货交易过程实际上就是综合反映供求双方对未来某个时间供求关系变化和价格走势的预期。因此，期货与现货的对比就是一个很好的因子挖掘的方向。期限对比的典型的一些特征是，期货与现货之间，有价差，价差会随着时间，逐渐减少，最终实现，现货价格与期货价格最终趋于统一。

基于此，提出以下的部分因子 (1) 昨日 IF 波动率是否大于沪深 300 指数的波动率，是则，因子数值为 1，否则因子数值为 0；(2) IF 主力合约的昨日 $dhigh$ 是否今日 $dhigh$ ，是则，因子数值为 1，否则因子数值为 0；(3) IF 主力合约的昨日 $dlow$ 是否今日 $dlow$ ，是则，因子数值为 1，否则因子数值为 0；(4) IF 主力合约的昨日 $(dclose-dopen)$ 是否今日 $(dclose1-dopen1)$ ，是则，因子数值为 1，否则因子数值为 0。

4) 期货远近合约对比

期货的远近合约，主要是通过大众对近期合约的看多和看空程度，会在远期合约中体现。而期货远近月合约价差主要是指该期货近月合约价格与远月合约价格的差值，它主要反映了持仓成本（主要包含储存费用、利息、保险费等）。理

论上讲，期货远月合约价格通常高于近月合约价格，而近月合约价格又高于现货价格，随着交割月份的临近，持仓成本不断降低，期差逐渐减少，现货价格与期货价格最终趋于统一。

基于此，提出以下的因子猜想：

(1) 昨日近月涨幅是否更大，是则，因子数值为 1，否则因子数值为 0。收益计算方式为 $ret = (close - open) / close$; (2) 上周近月涨幅是否更大，是则，因子数值为 1，否则因子数值为 0; (3) 上个月近月涨幅是否更大，是则，因子数值为 1，否则因子数值为 0; (4) 上周近月波动率是否更大，是则，因子数值为 1，否则因子数值为 0。

4) 高波动率个股统计

高波动率的个股，一般指的是市场中最为活跃的个股，这一些个股通常交易量大，受到投资者的关注，通常也可以反应整体市场中的大部分或者头部投资者的投资态度和情绪。因此对于高波动个股的统计，可以反应市场一部分特征。

基于此，提出以下的几种因子猜想：

(1) 昨日涨停股，今日上涨的数量是否大于今日下跌的数量，用 $close - open$ 衡量涨跌，是则，因子数值为 1，否则因子数值为 0; (2) 昨日跌停股，今日下跌的数量是否大于今日上涨的数量，用 $close - open$ 衡量涨跌，是则，因子数值为 1，否则因子数值为 0; (3) 涨停封板后破板的股票数量是否在上升，一类是 $high = \text{涨停}$; 另一类是 $high = close = \text{涨停板}$ ，但是中间破板过。是则，因子数值为 1，否则因子数值为 0。

5) 特殊个股统计

特殊个股通常表现为 ST、近期复牌、与沪深 300 指数低相关的个股。这些个股一旦集体出现某种特征。一般代表着市场某种特殊的特征。典型的倘若 st 个股大量的涨停或者冲高，其实是反映了市场投机情绪过热。

基于此，提出以下的因子猜想：

(1) ST 个股的涨停数量，因子数值为涨停数量; (2) ST 涨停数量与昨日相比，是否增多，是则，因子数值为 1，否则因子数值为 0。

6) 国际商品品种交易情况

国际市场的特殊品种有包括黄金、白银、原油这三个主要品种，选择这一分类的原因很简单，这些品种的交易，有一些是和股票市场是互补市场，另一些则是和股票市场高度相关的，因此，这类品种交易会影响沪深 300 指数的走势，对交易起到积极的作用。

基于此，提出以下的因子猜想：

- (1) 黄金的收益率；(2) 黄金与白银的比价；(3) 石油的收益率

7) 国际指数情况

国际指数对于沪深 300 的指数有着较大的影响，比如日经指数、韩国综合价格指数等，因为时区原因比国内早开盘半小时到一个小时，因此，这些指数在这段时间内的表现很有可能会影响国内沪深 300 指数的走势。典型的，倘若日经指数、韩国综合价格指数都在第一个半小时内出现大跌的现象，对于沪深 300 指数，开盘也很有可能出现大跌现象。

其次，美国的纳斯达克指数、标普 500 这些具国际影响力的指数，本身就是可以反映全球股票的一种基本市场特征的，其收益率本身也就可以视为一种特征。因此，这也是对预测模型起到积极的作用的。

3.4.3 因子生成

通过上述逻辑，本文一共提出了 523 组因子猜想，由于因子名称较为混乱，根据其属性，进行二次整理。得到了以下的名称表格：

表 3.3 初始因子名称表（节选）

序号	名称	序号	名称
1	Alpha1	20	Shape_KLINE_C_7_5_60
2	Alpha2	21	Shape_KLINE_C_8_5_60
3	Alpha3	22	Shape_KLINE_C_9_5_60
4	Alpha4	23	Shape_KLINE_C_10_5_60
5	Alpha5	24	Shape_KLINE_C_7_15_60
6	Alpha6	25	Shape_KLINE_C_7_10_60
7	Alpha7	26	Shape_KLINE_C_7_15_20
8	Alpha8	27	Shape_KLINE_C_7_15_30
9	Alpha9	28	Shape_KLINE_C_7_15_40
10	Alpha10	29	Shape_KLINE_C_7_15_60
11	Alpha11	30	Shape_KLINE_C_17_5_000905_SKIN
12	Alpha12	31	Shape_KLINE_C_16_5_000905_SKIN
13	Alpha13	32	THS_CHARACTERISTIC_BDJZ2
14	Alpha14	33	THS_CHARACTERISTIC_JFZNZX1_30
15	Alpha15	34	Micro_KLINE_C_3_000905_SKIN
16	Alpha16	35	Coupling_KLINE_C_5_1
17	Alpha17	36	Coupling_KLINE_C_5_1
18	Alpha18	37	Coupling_KLINE_C_5_1
19	Alpha19	38	Coupling_KLINE_C_5_1

3.5 因子有效性检测

3.5.1 因子有效性检测概述与方法

在提出上述因子之后，这些因子是因子猜想而已，并没有进行有效性的验证。这就意味着并非有所得因子都可以在模型中发挥积极的作用。倘若不对这些无效因子进行剔除容易产生两种问题。首先，无效的因子，类似干扰信息，会对预测模型的预测结果产生干扰。其次，无效的因子，在后续的共线性检测中，可能会潜在的因子无效因子的存在而排除来某类有效的因子。具体来说，当无效因子和一类有效因子存在共线性关系之后，在进行共线性处理时，势必需要舍弃其中一

个因子，倘若从主观判断，容易造成误判。且，由于因子数量众多，在取得最优因子数量的过程中，势必需要考虑整体因子的相关性，而在计算整体的因子相关性的时候，就容易存在无效因子与整体因子的总体相关性较低，而有效因子相关性较高的情况，从而排除了有效因子，造成模型的误差偏大。

对于模型的有效性检测，本文主要从三个方面进行，第一，XY 散点图分析。这类分析，根据笔者的实践经验，通常在散点图的分布，会在中心区域密切的集中，而在四周出现一定的特征。因此，在分析这类图的时候，应该考虑放弃判断中间区域，因为这一区域的信息是无效混乱的。而在四周边缘区域，是有预测分析的价值。对于这一分析的逻辑也是可以得到论证的。首先，实际上可以假设深度学习的预测模型，存在一种过滤机制，将中心密集的区域的信息进行过滤，则可以提取有用的信息用以预测。其次，本文的预测目标为金融市场的风险评估，那么对于指数收益的极端情况，大涨或者大跌情况的预测，本文更加侧重预测指数的大跌，而无论是大涨还是大跌情况，都是属于收益的极端分布，理论上应该会分布于散点图的四周，而不是中心密集区域。因此，直接忽略密集区域的分析做法是可行的。

第二，根据上文所提到的因子挖掘的逻辑的论证部分，可以得到，因子数据和指数收益的应该是存在高度的时序相关的。因此，可以进行简单的时序相关的计算，判断因子是否有效。

第三，最后一种方法是，在上述结论暂不成立的情况下，或则上述结论呈现出线性相关性的情况下，使用线性回归 K 值判断，K 值数据的大小，可以用以判断因子的有效性，只要 K 值，不是接近于 0，则代表着因子的有效。倘若因子的 K 值距离 0 的偏离程度越大，则因子有效性越大。同时数值的正负可以反应这一因子数据对于指数收益而言是正向的还是负向的。

3.5.2 因子有效性检测

在这一部分，开始进行因子有效性检测与分析。由于因子数量庞大，这里先取得前 20 个因子对其展开具体详细的分析论述。而对于剩下的上百个因子的有效性检测，不进行具体的展开，而是在因子有效性分析环节，对于因子检测后的

结论进行分析和讨论，而不再具体探讨其具体的检测过程和方法。

在检测之前，首先需要说明的是，本文的预测逻辑是通过成分股的预测，合成指数预测。因此，在因子有效性分析环节，应该将指数的成份股的因子数据取得，进行分析。但是潜在的问题就出现了，以本文的分析对象沪深 300 指数为例，有 300 个成份股的因子数据，对应一个沪深 300 的收益（标签）。那么对于单一因子，则应该是由一个 300×1 的矩阵，对应一个标签，这在 xy 散点图上是无法体现的。因此，本文做了如下处理：首先，由于成份股合成的指数的方法是固定的。也就是说沪深 300 指数的编制，是通过入选的成分股乘以相应的权重编织而成的。因此，本文将因子数据乘以当时的指数编制的权重，就可以大致合成了一个因子数据，这个因子数据是反馈有指数信息的。这一背后的逻辑论证如下：沪深 300 的指数的成分股权重，其实反映了这个个股在指数中的影响。有些权重较低的个股，对指数价格影响较小，有些权重较高的个股，对指数价格影响较大。而沪深 300 指数的成分股的因子数据表现并不是始终一致的，肯定是有所不一样的。那么对于指数的影响也是不一样的。因此，需要去衡量这些同一个因子，在不同的成分股个股上，对于指数的影响程度。显然，使用编织沪深 300 时公布的指数成分股权重，是很合理的。

1) 因子一，名称 alpha 1

显然，通过这一散点图，可以得到一些结论。根据上文所提及的分析散点图的逻辑。应该暂时忽略中心密集的区域。由此，基本可以发现，该因子对于指数的影响，更加类似一个二次曲线的关系。本文再来计算二者的相关系数，二者的相关系数为 0.0061。这是一个很低的数值，基本是不符合本文的要求。因此，对于这一因子，考虑舍弃，并且视其为无效因子。

2) 因子二，名称 alpha 2

这一因子的散点图，呈现出的特点是，负相关（在排除中心密集区域的情况下）。而二者的相关系数，也是验证了上述结论，相关系数为 -0.0211。考虑到中心密集区域，是主管抛弃的。但是在计算相关系数的时候，还是考虑了这些密集区域。因此，相关系数会呈现出十分低的特点。

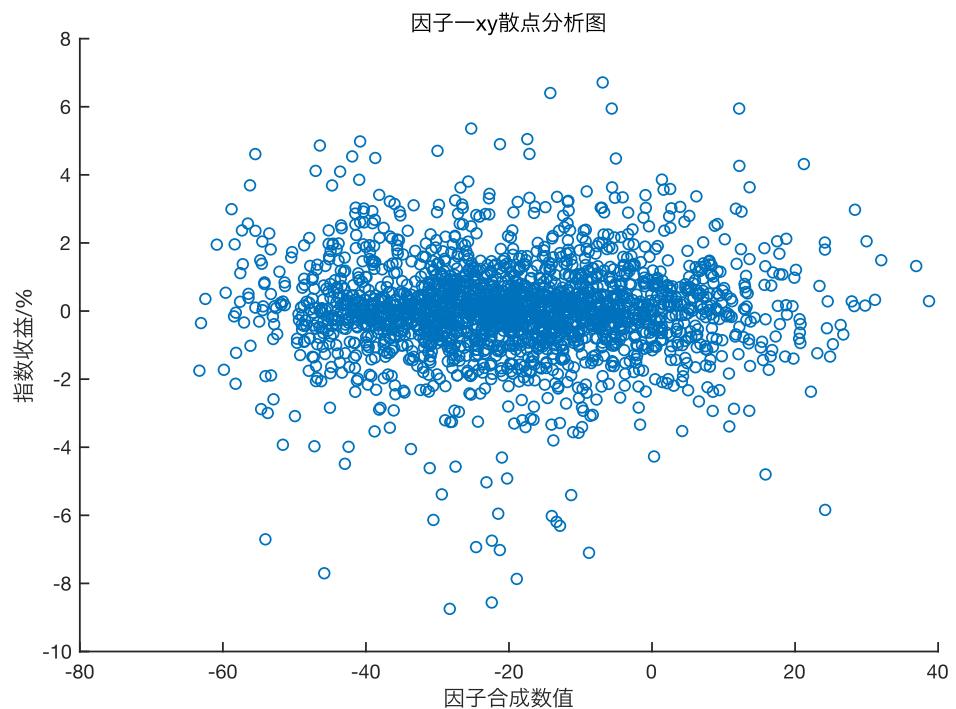


图 3.5 因子一 xy 散点分析图

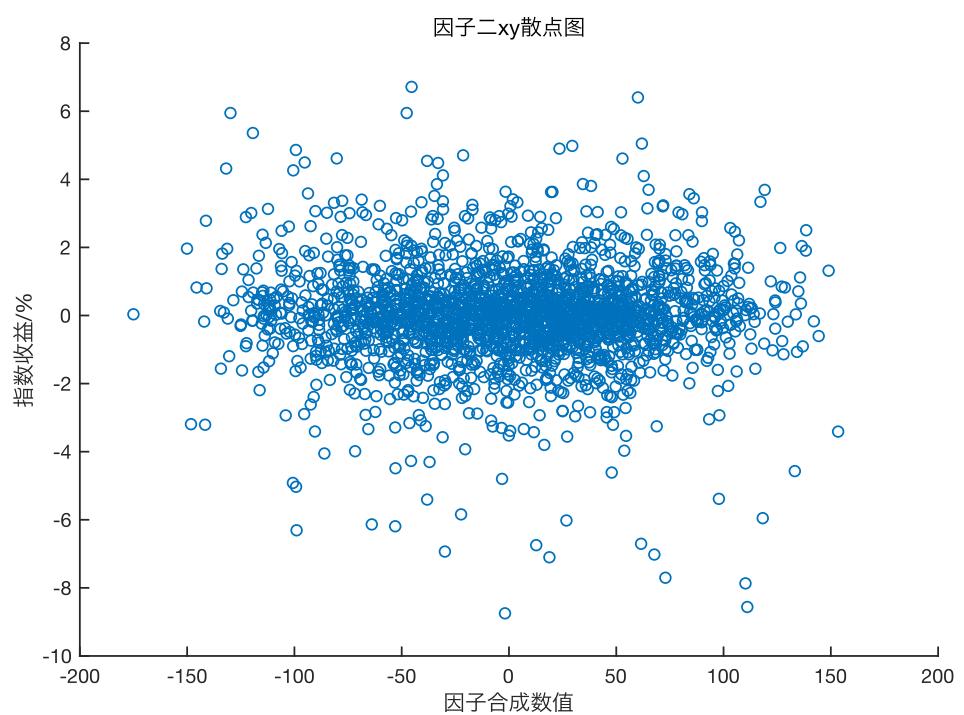


图 3.6 因子二 xy 散点分析图

3) 因子三，名称 alpha 3

相关系数为-0.0436，该因子的相关性特征，在对比之下。适合后续使用。

4) 因子四，名称 alpha 4

相关性系数为-0.0210，这一因子的特点较为明显，呈现出类似头重脚轻的分布，相关性特征在边缘区域，较为明显。是有效因子。

5) 因子五，名称 alpha 5

这一因子的分布，可以视为一个关于指数收益的二次函数。故此，相关性系数为0.024，偏低。但是特征分布较为明显，可以视为有效因子。

6) 因子六，名称 alpha 6

这一因子类似因子，在边缘处，呈现出二次函数的特点，但是总体相关性系数并不高，为0.0068。但是考虑其边缘分布形态较好，还是考虑其为有效

7) 因子七，名称 alpha 7

这一因子呈现出线性分布特征，其相关性系数为0.043。较为明显。因此，考虑将其纳入有效因子范围。

8) 因子八，名称 alpha 8

这一因子分布较为散漫，特征不明显，并且相关系数并不高为0.0192，因此考虑将将其视为无效因子。

9) 因子九，名称 alpha 9

这一因子的边缘分布形态较为有特色。呈现出负相关的特点。并且相关系数为-0.0223，应该存在一定的价值，考虑使其为有效因子。

10) 因子十，名称 alpha 10

这一因子在边缘形态的分布上较为有特点。集中在负半轴。相关系数为-0.0415，应该考虑其为有效因子。

11) 因子十一，名称 alpha 11

这因子的分布形态较为混乱，但是相关性系数较高，为0.0453。因此，考虑其为有效因子。

12) 因子十二，名称 alpha 12

这一因子的集中度较高，而边缘分布比较为集中，呈现负相关性的特点，且相关系数为-0.0630。应该是其为有效因子。

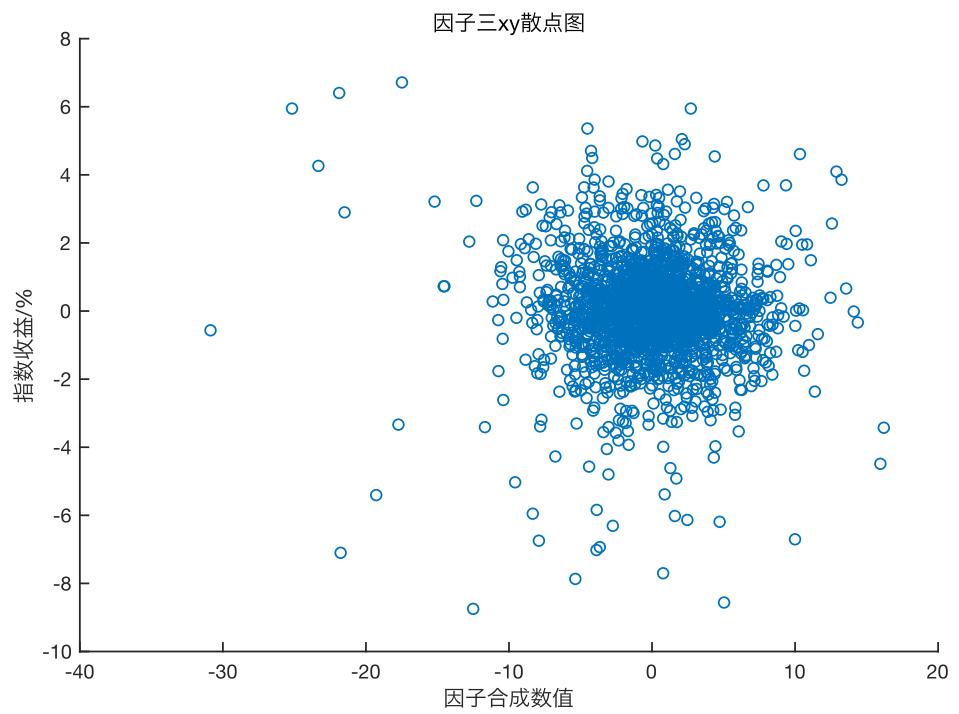


图 3.7 因子三 xy 散点分析图

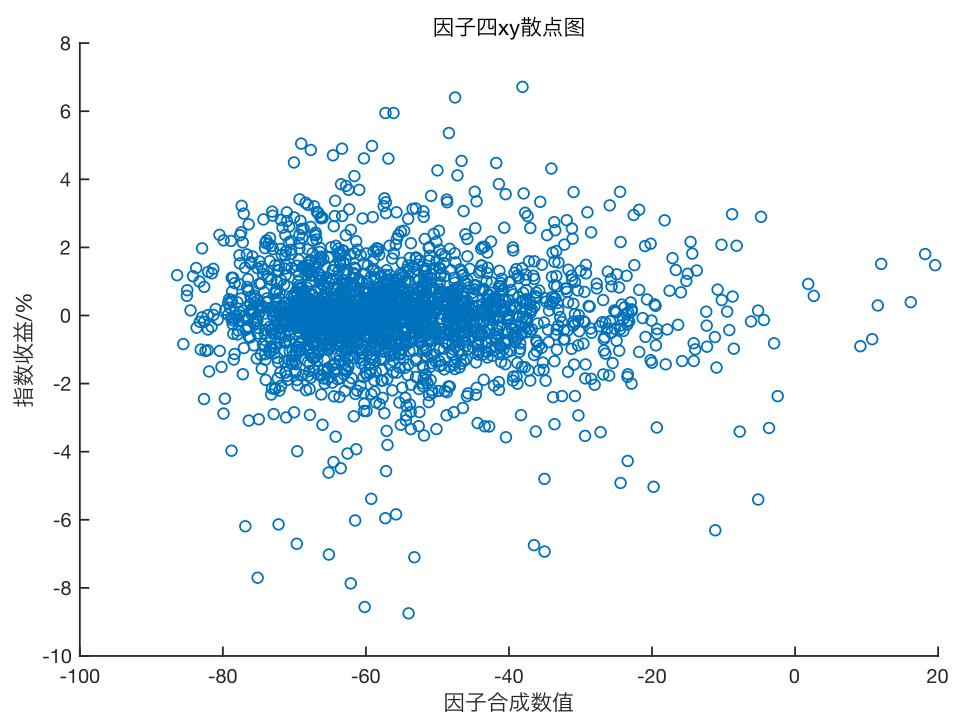


图 3.8 因子四 xy 散点分析图

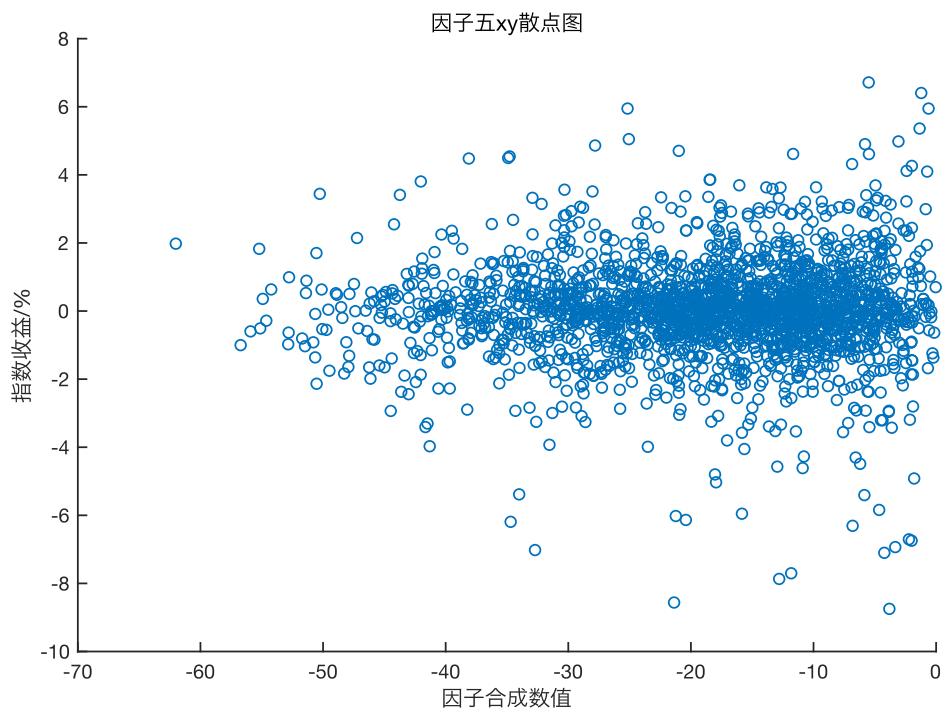


图 3.9 因子五 xy 散点分析图

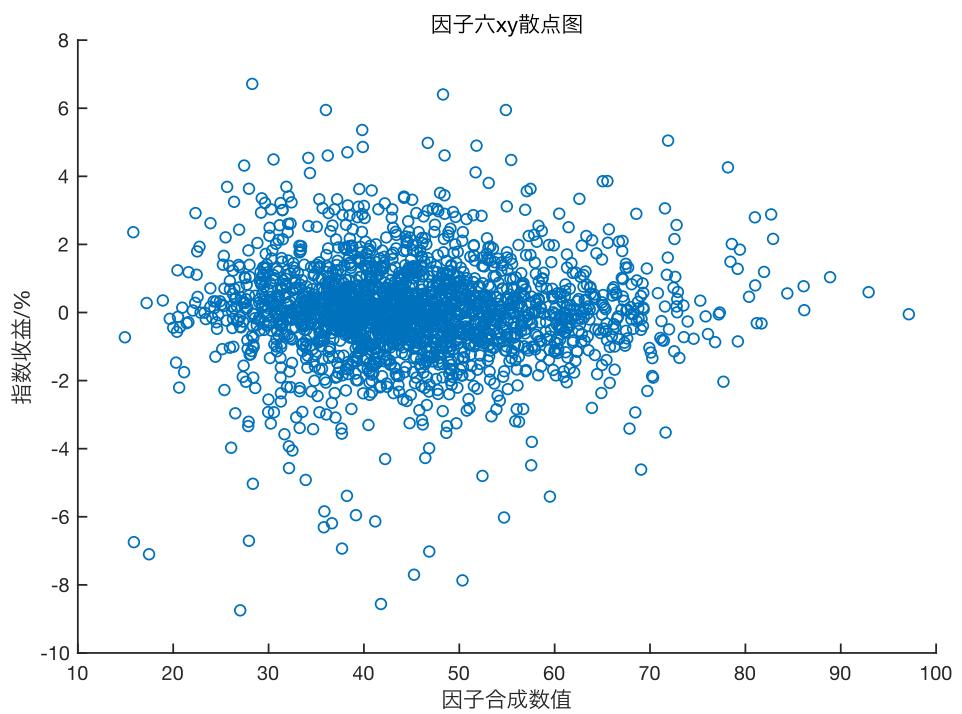


图 3.10 因子六 xy 散点分析图

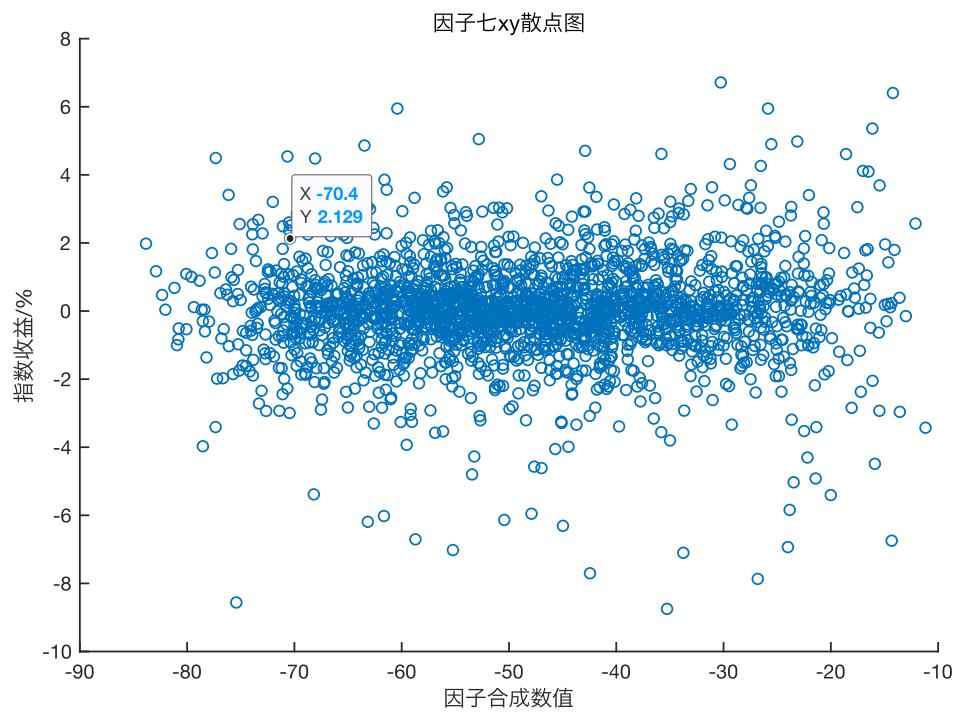


图 3.11 因子七 xy 散点分析图

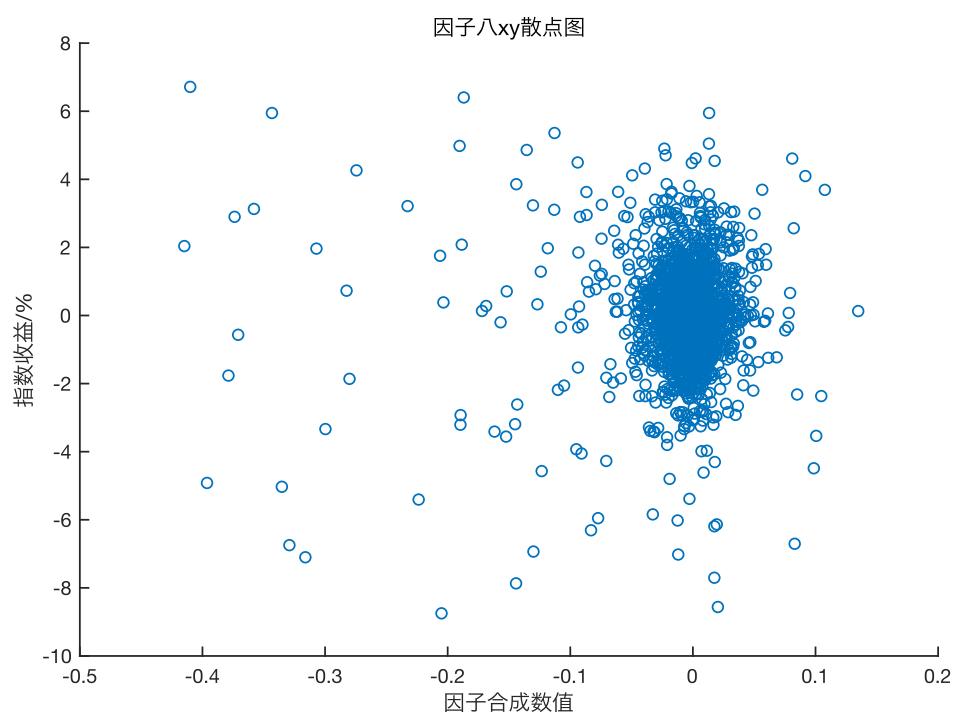


图 3.12 因子八 xy 散点分析图

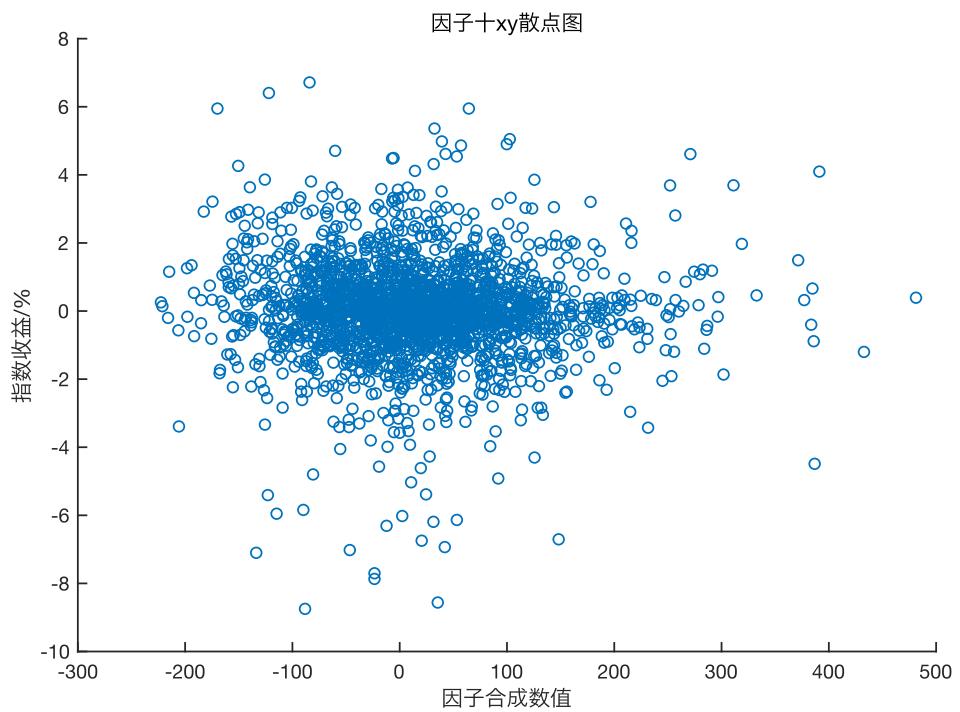


图 3.13 因子九 xy 散点分析图

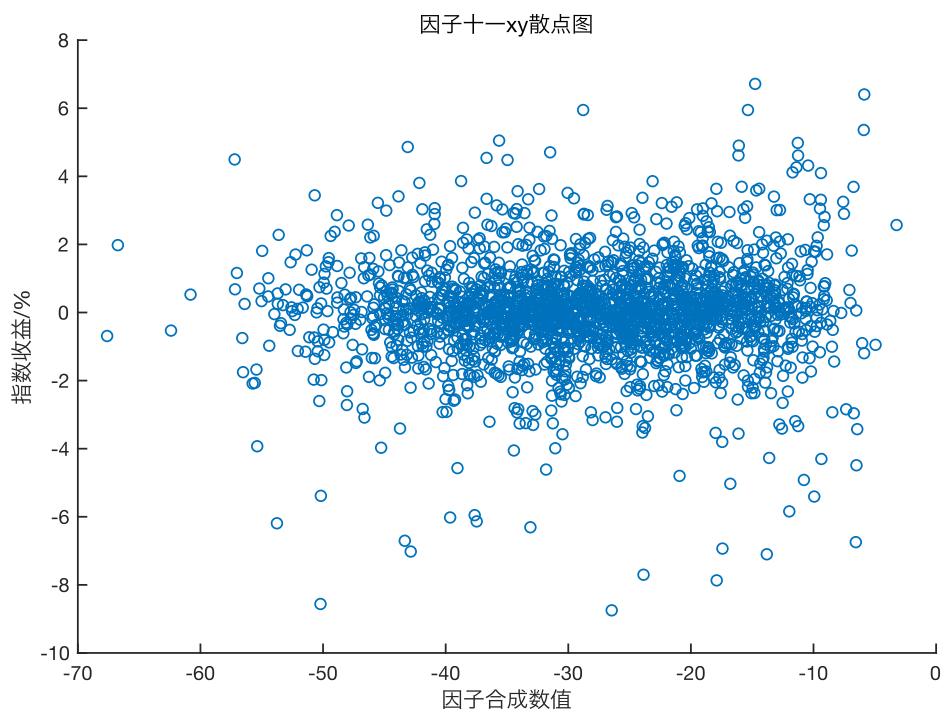


图 3.14 因子十 xy 散点分析图

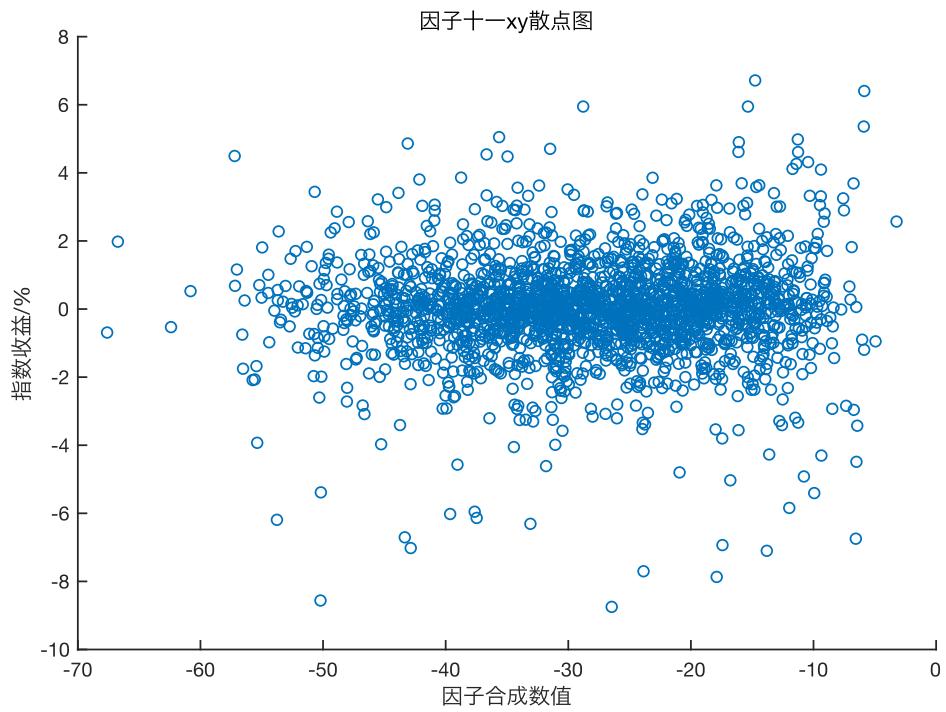


图 3.15 因子十一 xy 散点分析图

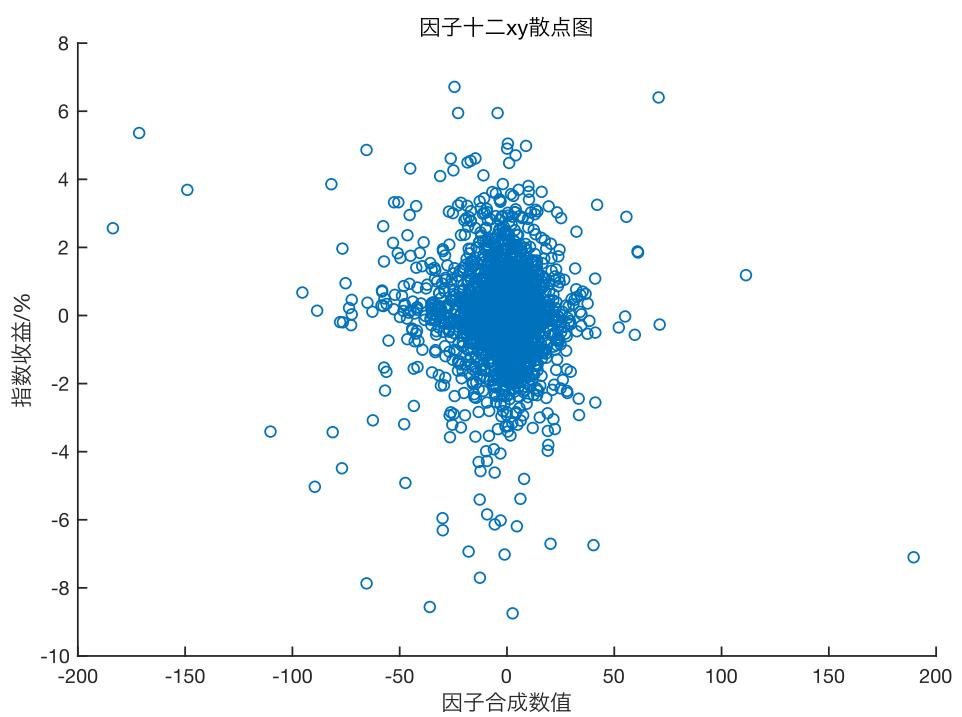


图 3.16 因子十二 xy 散点分析图

13) 因子十三，名称 alpha 13

这因子边缘分布较为混乱，相关性系数也较低，为 0.0134。因此，不考虑为有效因子。

14) 因子十四，名称 alpha 14

这因子边缘分布较为明显，但是相关性系数过于低，为 -0.0066。因此，不考虑为有效因子。

15) 因子十五，名称 alpha 15

该因子分布呈现出，线性分布的特点，相关系数较低，为 0.0128。因此考虑其成为有效因子。

16) 因子十六，名称 alpha 16

该因子的边缘分布较为混乱，相关系数也比较低，为 -0.0171。因此不考虑有效因子。

17) 因子十七，名称 alpha 17

该因子的分布较为胡乱，相关系数较为低，为 -0.0047。因此，不考虑为有效因子。

18) 因子十八，名称 alpha 18

这一因子，在边缘分布形态上，出现正相关的表现。相关系数为 0.0126，应该考虑其为有效因子。

19) 因子十九，名称 alpha 19

这一因子分布形态，呈现二次函数的特点，相关性系数不高，为 -0.0180。但是仍然考虑其为有效因子。

20) 因子二十，名称 alpha 20

这一因子呈现二次函数分布的特点，相关性系数为 0.0183，较为有价值，应该考虑其视为有效因子。

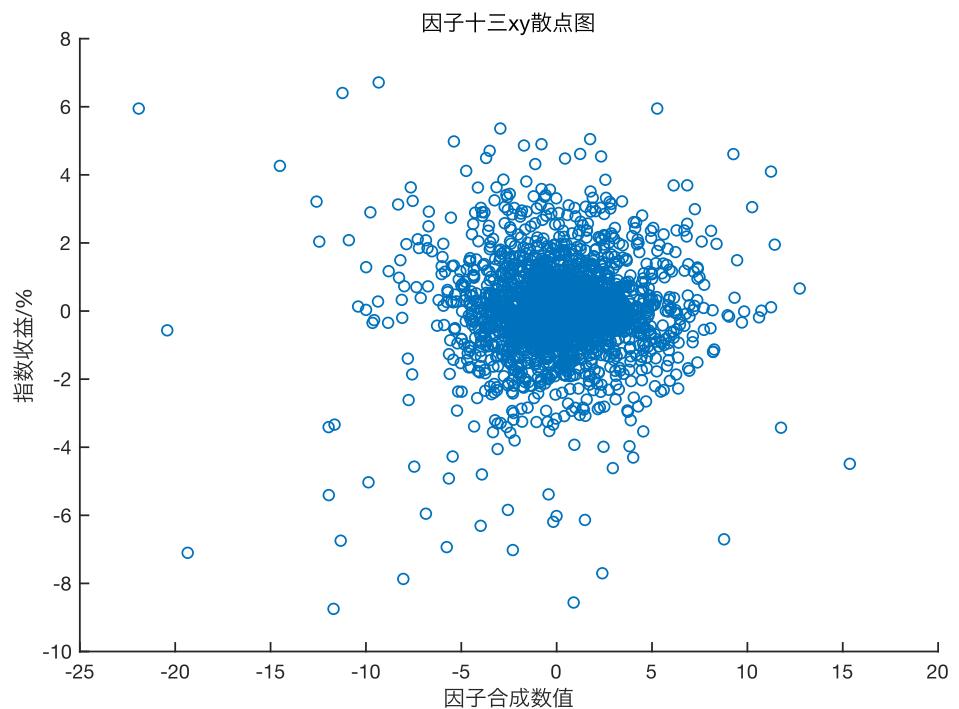


图 3.17 因子十三 xy 散点分析图

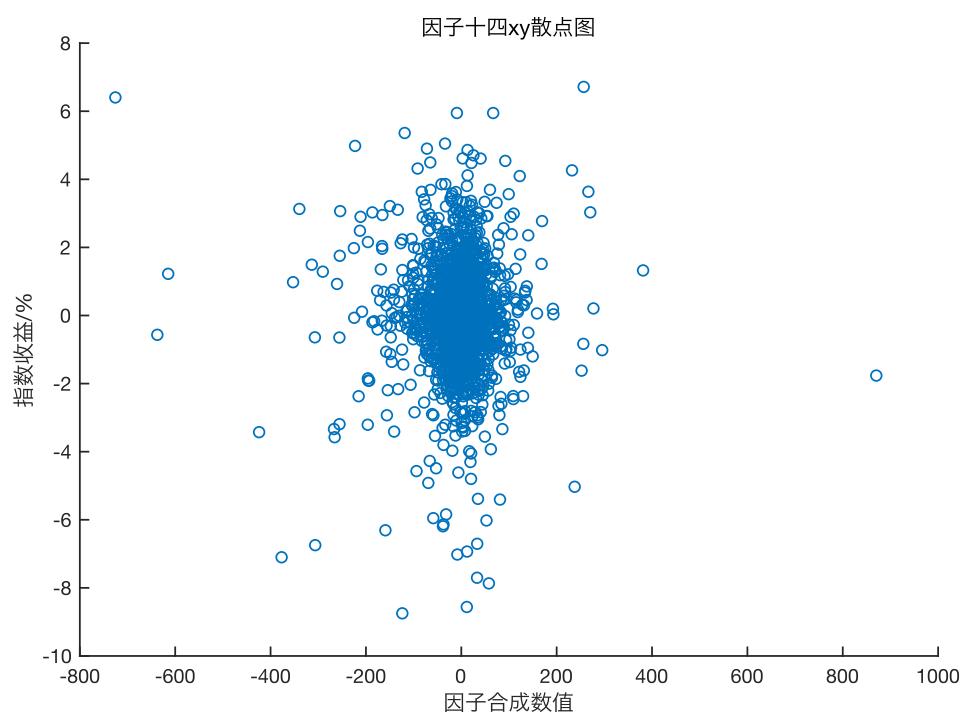


图 3.18 因子十四 xy 散点分析图

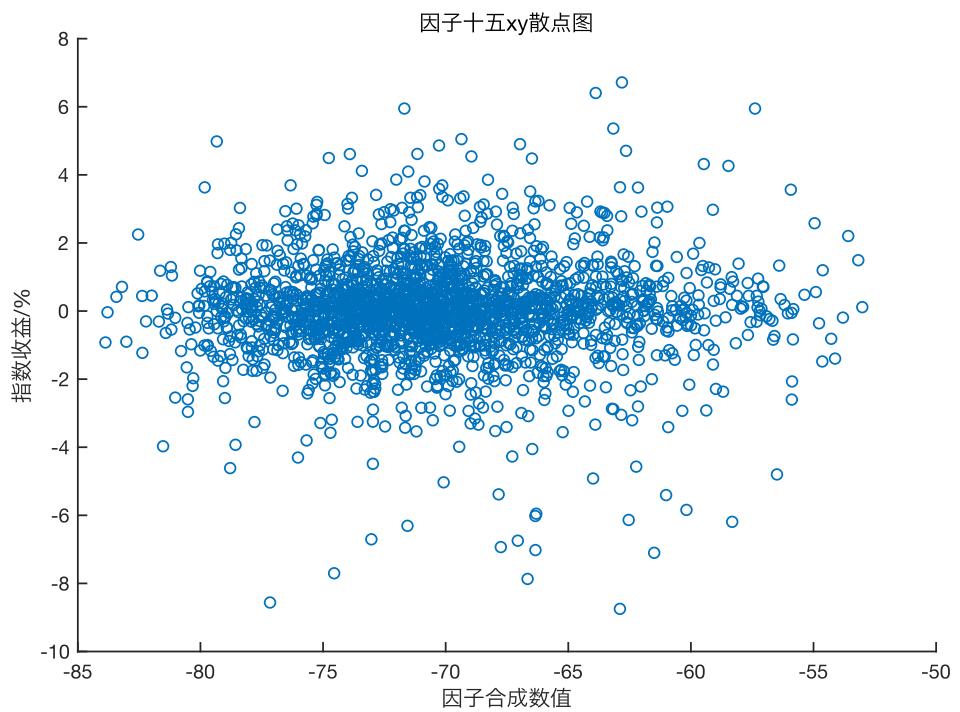


图 3.19 因子十五 xy 散点分析图

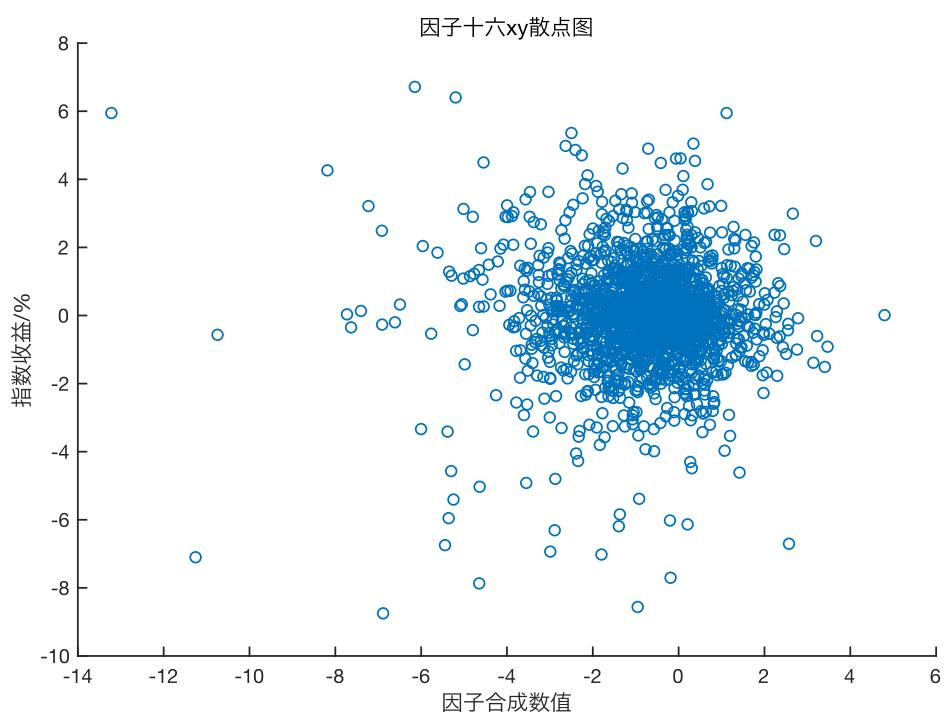


图 3.20 因子十六 xy 散点分析图

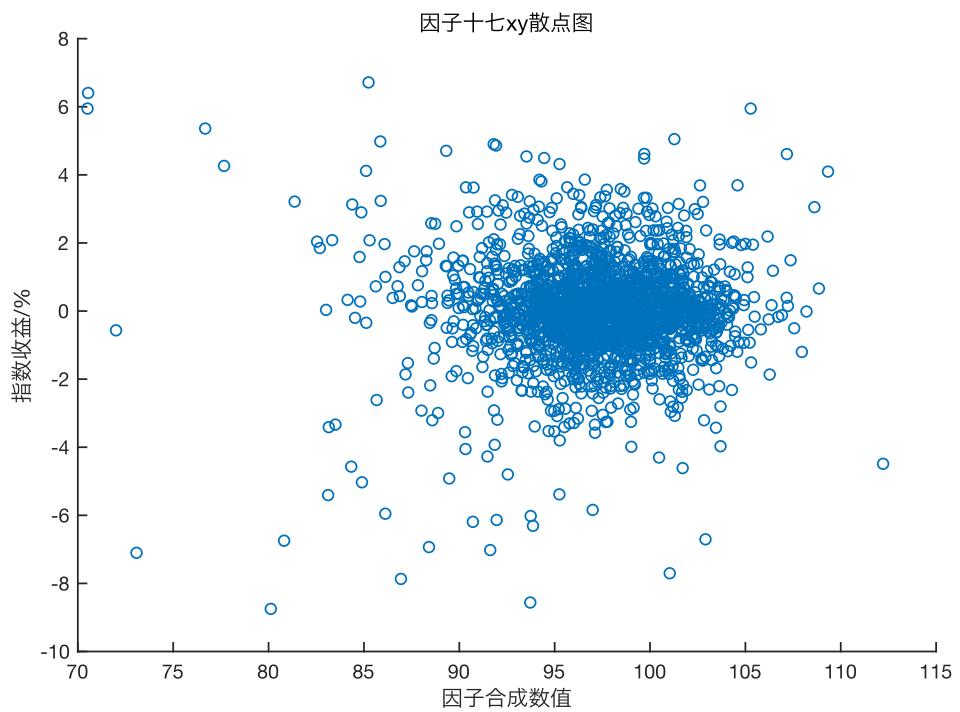


图 3.21 因子十七 xy 散点分析图

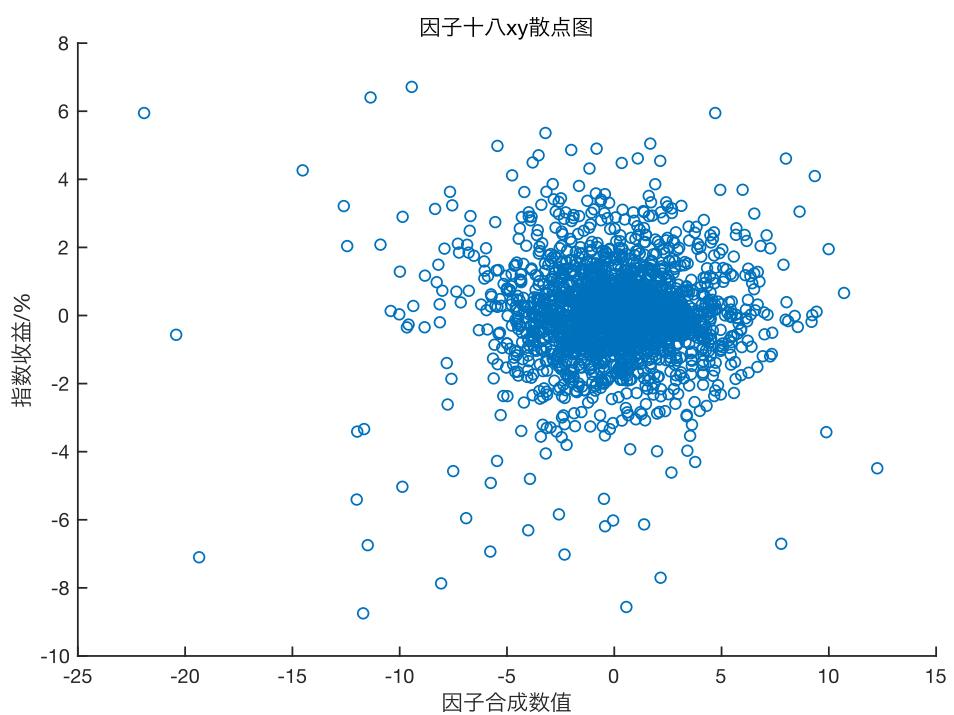


图 3.22 因子十八 xy 散点分析图

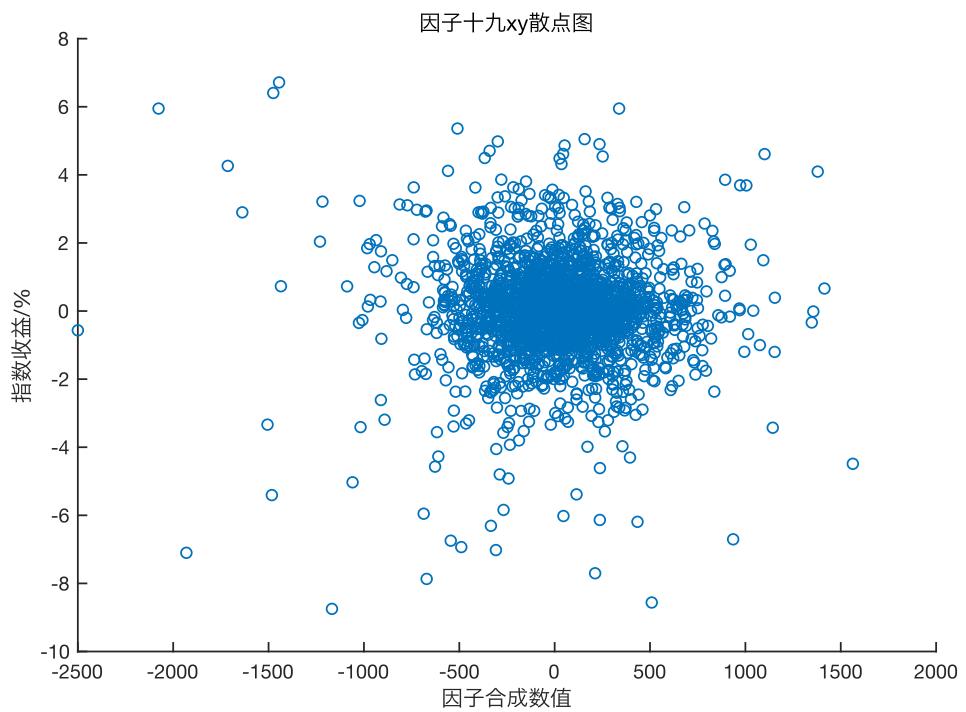


图 3.23 因子十九 xy 散点分析图

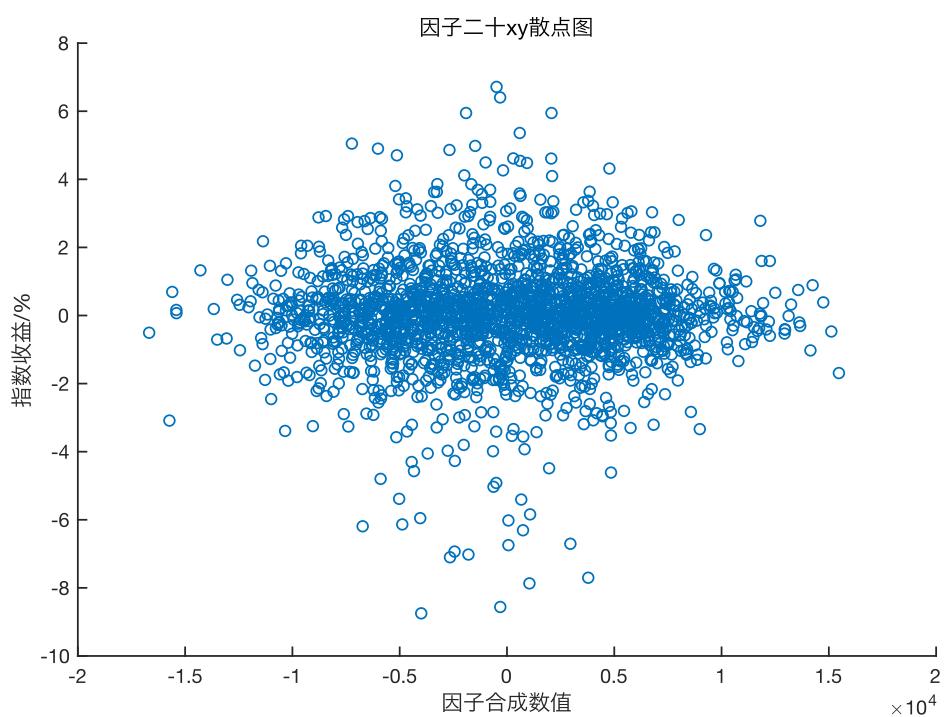


图 3.24 因子二十 xy 散点分析图

3.5.3 因子有效性分析

在经过检测分析之后，将入选的因子列出，得到了下表：

表 3.4 有效因子名称表（节选）

序号	名称	序号	名称
2	Alpha2	45	Micro_KLINE_C_11
5	Alpha5	48	Micro_KLINE_R_3
7	Alpha7	49	Micro_KLINE_C_27_1_399300_SKIN
8	Alpha8	50	Micro_KLINE_C_7_5_000905_SKIN
15	Alpha15	51	Micro_KLINE_C_27_3_000905_SKIN
19	Alpha19	66	Micro_KLINE_C_26_4_000905_SKIN
22	Micro_KLINE_C_5	69	Micro_KLINE_C_27_4_000905_SKIN
31	Micro_KLINE_C_6	79	Coupling_ENHANCE_R_2_3
32	Micro_KLINE_C_7	84	Label_TIME_R_1_10
33	Micro_KLINE_C_8	85	Label_TIME_R_5_3

3.6 因子共线性检测

3.6.1 因子共线性概述

多重共线性是出现在使用线性回归等模型算法的时候的一种问题。在其他算法和模型中，也是存在的问题的。但是对于另外一些模型和算法，比如决策树算法和贝叶斯模型，决策树的建模过程是逐步递进，每次拆分只有一个变量参与，这种机制本身就是含有抗多重共线性干扰的能力；而在贝叶斯模型中，前提假设就是变量之间是相互独立的，因此从理论层面上看，模型实际运行的是，应该是不存在多重共线性的问题，但是就算在贝叶斯模型中，虽然前提假设为变量之间相互独立，但是实际操作中，在使用该模型之前，还是需要进行简单的变量分析。更重要的是，对于回归类的算法，不论是一般线性回归，逻辑回归，都是需要同时考虑多个预测因子之间的关系的，因此多重共线性是不可避免的。很多时候，多重共线性是一个普遍的现象。对于共线性的理论背景解释可以有以下的内容说：假设有 k 个自变量的多元线性回归模型：

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \epsilon = \mathbf{BETA} * \mathbf{X} + \epsilon \quad 3.3$$

其中误差项是一个期望值为 0 且服从正态分布的随机变量：

$$\varepsilon \sim N(0, \sigma^2)$$

3. 4

则利用最小二乘法可得参数的估计值为:

$$\text{BETA} = X^{-1}y = (X^T X)^{-1} X^T y \quad 3.5$$

该求解公式唯一的条件是矩阵 X 是列满秩的, 不然会有无穷多解:

$$\text{BETA} = X^{-1}y + (1 - X^{-1}X)\zeta \quad 3.6$$

当各变量之间存在共线性问题, 即各变量之间存在部分线性相关时, 例如:

$$x_3 = x_2 + x_1 \quad 3.7$$

易知此时 X 近乎是不满秩的 (实际情况很难完全共线性), $X^T X$ 近乎是奇异的, X 的最小奇异值会非常小。他的影响将在下文得到说明。

3. 6. 2 共线性检测

首先, 进行扰动分析: 对于一个方程或者系统而言, 当输入有一个非常微小的扰动时, 我们希望方程或系统的输出变化也非常微小, 如果输出的变化非常大, 且不能被控制, 那这个系统的预测就无效了, 蝴蝶效应讲的就是这个。在矩阵计算中, 这叫做扰动分析。

设非奇异方阵 A 满足方程:

$$Ax = y \quad 3.8$$

它的精确解为:

$$x = A^{-1}y \quad 3.9$$

当 A 存在一个小扰动时, 得到以下方程:

$$(A + \delta A)\tilde{x} = y \quad 3.10$$

可以证明这一方程的解的扰动满足:

$$\frac{\|\delta x\|}{\|x\|} \leq \kappa(A) \frac{\|\delta A\|}{\|A\|} \quad 3.11$$

其中:

$$\kappa(A) = \|A^{-1}\| \cdot \|A\| \quad 3.12$$

是非奇异方阵的条件数, 且此时矩阵范数等价于矩阵最大的奇异值, 即矩阵的条件数等价于最大奇异值/最小奇异值

可以看到矩阵的条件数越大，扰动就越大，即 x 的求解值会变得非常不准确。回到上面讲的线性回归问题，容易证明最小二乘法的解满足下面的正定方程：

$$X^T X \widehat{BETA} = X^T y \quad 3.13$$

此时：

$$\kappa(X^T X) = \frac{\lambda_{\max} X^T X}{\lambda_{\min} X^T X} \quad 3.14$$

当方程有共线性问题时， X 的最小特征值非常小，相应的，上述的条件数会非常大。也就是说机器学习中的共线性问题实际上就是矩阵计算中的条件数问题。从实际应用的角度，一般若 $K < 100$ ，则认为多重共线性的程度很小，若是 $100 \leq K \leq 1000$ ，则认为存在一般程度上的多重共线性，若是 $K > 1000$ ，则就认为存在严重的多重共线性。

其次，进行方差分析。再从统计学的角度来看共线性。可以证明参数 $BETA$ 的协方差矩阵为：

$$\text{Var}(\widehat{BETA}) = \text{Var}(\widehat{BETA} - BETA) = \text{Var}[(X^T X)^{-1} X^T \varepsilon] \quad 3.15$$

又对任意的常数矩阵 A 和随机变量 x 有：

$$\text{Var}(Ax) = A \cdot \text{Var}(x) \cdot A^T \quad 3.16$$

代入上式即可得：

$$\text{Var}(\widehat{BETA}) = \sigma^2 X^T X^{-1} \quad 3.17$$

具体到每个参数，有：

$$\text{Var}(\widehat{BETA}_j) = \frac{\sigma^2}{(n-1) \text{Var}(x_j)} \cdot \frac{1}{1 - R_i^2} \quad 3.19$$

令：

$$VIF_i = \frac{1}{1 - R_i^2}$$

为方差膨胀因子 (variance inflation factor, VIF)。当

$$R_i^2 \rightarrow 1 \quad 3.20$$

时，即当第 i 个变量和其他变量之间存在线性关系时，VIF 趋于无穷大。所以 VIF 的大小反应了变量的共线性程度。一般地，当 VIF 大于 5 或 10 时，认为模型存在严重的共线性问题。

同时考虑参数显著性检验的 t 统计量：

$$t = \frac{\widehat{BETA}_l}{std(\widehat{BETA}_l)} \sim t(n - k - 1)$$

当存在共线性时，参数的标准差偏大，相应的 t 统计量会偏小，这样容易淘汰一些不应淘汰的解释变量，使统计检验的结果失去可靠性。

另外考虑线性回归的残差：

$$\hat{\varepsilon} = y - X\widehat{BETA} = M\varepsilon \quad 3.22$$

其中 M 是一个投影矩阵，且满足：

$$M = I - X(X^T X)^{-1}X^T \quad 3.23$$

易证明：

$$\|\hat{\varepsilon}\|_2^2 = \varepsilon^T M \varepsilon \leq \|M\|_F^2 \cdot \|\varepsilon\|_2^2 = (n - k) \|\varepsilon\|_2^2 \quad 3.24$$

而矩阵 M 的范数与 X 的条件数毫无关系，于是可以得出共线性并不影响模型的训练精度。但是对于泛化精度，由于参数的估计已经不准确啦，所以泛化误差肯定要差些，具体差多少，我还很难用公式表示出来。

总结一下，共线性问题对线性回归模型有如下影响：

- 1) 参数的方差增大；
- 2) 难以区分每个解释变量的单独影响；
- 3) 变量的显著性检验失去意义；
- 4) 回归模型缺乏稳定性。样本的微小扰动都可能带来参数很大的变化；
- 5) 影响模型的泛化误差。

其次，对于共线性问题的解决方法。根据上述描述，共线性问题有如下几种检验方法：

- 1) 相关性分析。检验变量之间的相关系数；
- 2) 方差膨胀因子 VIF。当 VIF 大于 5 或 10 时，代表模型存在严重的共线性问题；
- 3) 条件数检验。当条件数大于 100、1000 时，代表模型存在严重的共线性问题。

当变量数不多，样本数不是很大时，上述的方法是没问题的，检验某个变量有共线性问题时，可以结合实际业务考虑直接剔除该变量。但是有的时候变量数大到有上千个，VIF 的计算需要建立上千个回归模型（条件数仅能判定是否存在共线性，但不能找到对应的变量），这将耗费很长时间。事实上还可以从模型角度来直接规避共线性问题。

3.6.3 因子共线性处理

1) PCA 等降维法

主成分分析法作为多元统计分析的一种常用方法在处理多变量问题时具有其一定的优越性，其降维的优势是明显的，主成分回归方法对于一般的多重共线性问题还是适用的，尤其是对共线性较强的变量之间。当采取主成分提取了新的变量后，往往这些变量间的组内差异小而组间差异大，起到了消除共线性的问题。

2) 逐步回归法

逐步回归 (Stepwise Regression) 是一种常用的消除多重共线性、选取“最优”回归方程的方法。其做法是将逐个引入自变量，引入的条件是该自变量经 F 检验是显著的，每引入一个自变量后，对已选入的变量进行逐个检验，如果原来引入的变量由于后面变量的引入而变得不再显著，那么就将其剔除。引入一个变量或从回归方程中剔除一个变量，为逐步回归的一步，每一步都要进行 F 检验，以确保每次引入新变量之前回归方程中只包含显著的变量。这个过程反复进行，直到既没有不显著的自变量选入回归方程，也没有显著自变量从回归方程中剔除为止。

第一：建立全部变量的回归方程

第二：分别建立单独的回归方程，依照 t 检验和拟合度依次加入各变量来构建回归方程

第三：判断新引入的变量，对于之前的系数影响是否显著，是否符合实际以及对于拟合度的变量，来选择是否将变量引入模型中。

3) 岭回归、L2 正则化 (ridge regression)

岭回归是一种可用于共线性数据分析的有偏估计回归方法，它是一种改良的最小二乘估计法，通过放弃最小二乘法的无偏性，以损失部分信息、降低精度为代价获得回归系数更为符合实际、更可靠的回归方法，对条件数很大（病态数据）的拟合要强于最小二乘法。

在线性回归问题中，最小二乘法实际上是最小化问题：

$$\widehat{BETA} = arg_{BETA} \min ||X \cdot BETA - y||_2^2 \quad 3.25$$

而岭回归则是加入了 L2 惩罚项:

$$\widehat{BETA} = \arg_{BETA} \min \|X \cdot BETA - y\|_2^2 + C \|BETA\|_2^2 \quad 3.26$$

这样参数的方差不会过大，且随着惩罚项系数 C 的增大，共线性的影响将越来越小。在这个过程中，可以记录岭迹(ridge)的变化情况，通过对岭迹的波动来判断我们是否要剔除该变量。

岭回归解决共线性问题的处理，可以从矩阵计算的角度来看，L2 正则化下方程的解为：

$$\widehat{BETA} = (X^T X + CI)^{-1} X^T y \quad 3.27$$

在上一节我们讲到共线性代表正定矩阵 $X^T X$ 的条件数很大：

$$\kappa(X^T X) = \frac{\lambda_{\max}(X^T X)}{\lambda_{\min}(X^T X)} \quad 3.28$$

而当条件数很大时，矩阵的逆的数值计算也是非常不准确的，但是当我们给矩阵加上一个单位矩阵时，奇异性（不可逆）问题就被解决了。

进一步考虑对惩罚项对奇异值的影响，假设 X 的奇异值 (SVD) 分解为：

$$X = U \sum V^T \quad 3.29$$

则容易证明

$$\widehat{BETA} = (X^T X + CI)^{-1} X^T y = V D U^T y \quad 3.30$$

其中 D 是对角矩阵，且满足

$$D_{ii} = \frac{\sigma_i}{\sigma_i^2 + C} \quad 3.31$$

其反应了惩罚项是如何影响到条件数的。

4) LASSO 回归

LASSO 回归和岭回归类似，只不过将惩罚项由 L2 范数改为了 L1 范数

$$\widehat{BETA} = \arg_{BETA} \min \|X \cdot BETA - y\|_2^2 + C \|BETA\|_1 \quad 3.32$$

L1 范数没有 L2 范数那么圆润，毕竟存在不可导点，而且在 L1 范数下 LASSO 回归也给出了解析，但是相对于岭回归，LASSO 估计的参数能更容易收敛到 0。

5) ElasticNet 回归等

ElasticNet 回归同时兼顾了 L1 和 L2 惩罚项:

$$\widehat{BETA} = \arg_{BETA} \min \left\| X \cdot BETA - y \right\|_2^2 + C_1 \left\| BETA \right\|_1^2 + C_2 \left\| BETA \right\|_2^2 \quad 3.33$$

当许多变量是相关的时候, Elastic-net 是有效用的。Lasso 一般会随机选择其中一个, 而 Elastic-net 则会选在两个。

除此之外, 还有 L0 范数 (非零元的个数)、L1/2 范数等。

本文的采用的共线性处理方法为 L2 正则化, 这会直接体现在后续的模型端, 因此, 不再这里具体描述了。

3.7 因子分布形态测试

在上述处理和检测, 进行过之后, 在本节的因子分布形态测试主要是通过因子分布形态的角度, 对因子数据进行优化。因此, 只需要进行的是简单分析, 排除不适合的因子即可。对于不适合的因子的定义为: 因子分布形态上不接近对数正态分布。原因是股票市场的回报率基本符合对数正态分布, 因此采用相近的形态的因子进行预测, 或许在准确性上有所提升。同时, 对于接近对数正态分布的定义为: 正态分布、卡方分布等, 因为这些因子分布基本可以通过模型的迭代得到函数变化, 转为对数正态分布。

由于因子数量庞大, 在本文中先列举了前 10 个因子分布的情况。

1) 因子一

可以初步得到, 这一因子值的分布的形态上是接近卡方分布的。并且在 2011-2019 的九年內, 基本没有太大的变化。因此, 该因子无需二次优化。

2) 因子二

根据分布图, 这一因子值的分布形态上, 0 的数据较为突出, 根据校对, 是因为在产生因子数值的时候, 缺失的数值默认为 0。因此在后续的优化中, 将缺失的数据进行剔除。

3) 因子三

该因子的分布形态和因子有相同的问题, 缺失数值应该予以剔除。同时关注到 204 年因子分布出现了一定的偏离。但是均值还是在 0 左右, 只是极端样本更加偏向于左侧。

gene 1因子分布图

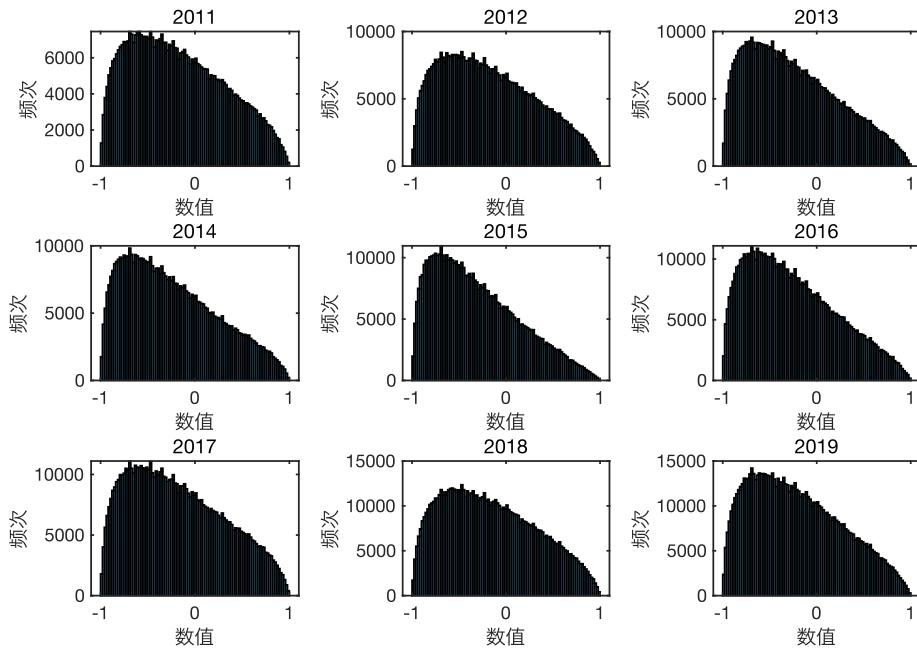


图 3.25 因子一分布图

gene 2因子分布图

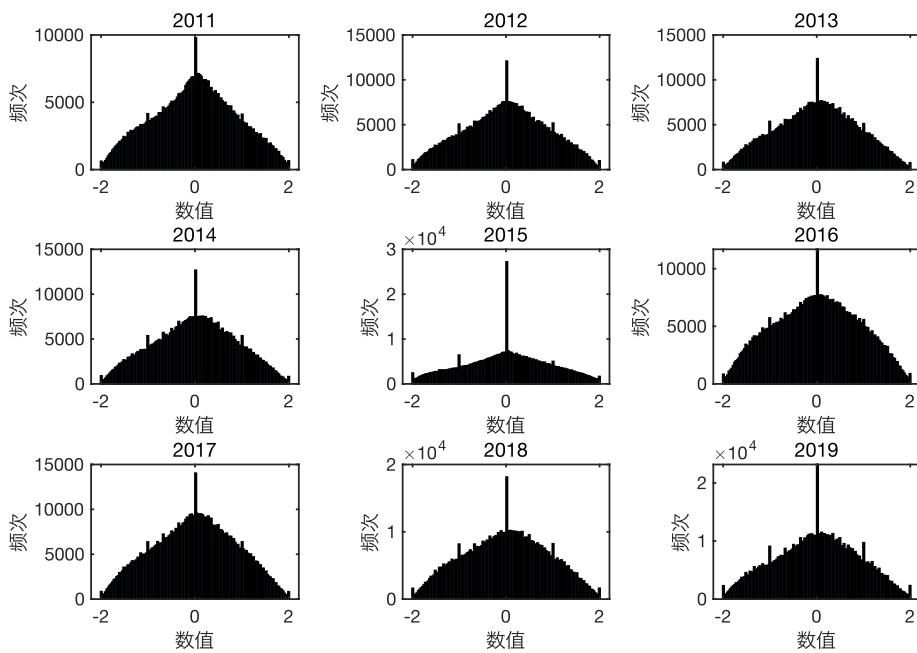


图 3.26 因子二分布图

gene 3因子分布图

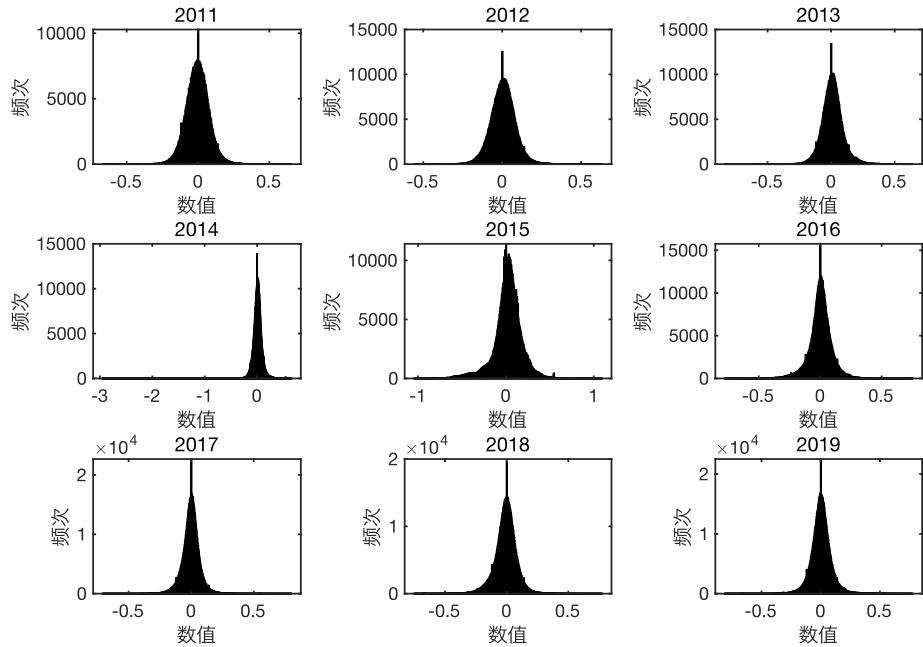


图 3.27 因子三分布图

gene 4因子分布图

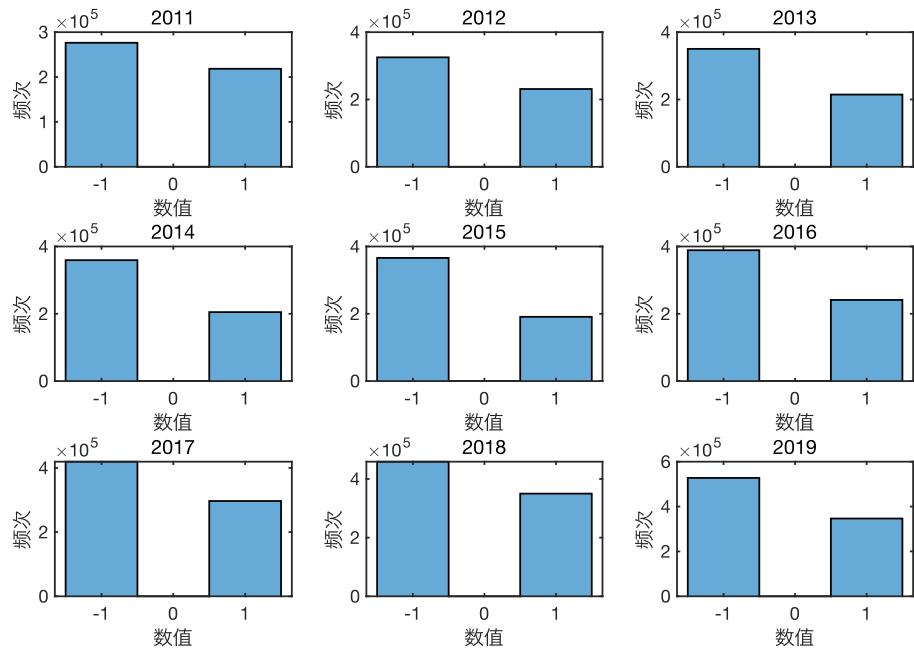


图 3.28 因子四分布图

4) 因子四

显然，这一个分类因子。暂不做处理。在优化部分会有对于这类因子进行额外的处理。

5) 因子五

这一因子呈现单侧递减的分布，但是在 0 出仍然出现较多的样本干扰。

6) 因子六

这一因子显然是没有意义的，因此在后续的模型中，将舍弃这一因子。原因是这一因子数值，大量分布在 0。

7) 因子七

这一因子分布，也是呈现单侧递减状态，但是相较于因子五，更加类似卡方分布。因子暂不做二次处理。

8) 因子八

这一因子分布比较均匀的。并且这一因子并不是分类因子。但是却是排序因子，因此，可以不做二次处理。

9) 因子九

这一因子类似于因子六，不做进一步讨论。

10) 因子十

这一因子出现的情况类似因子八，暂时不做考虑。

gene 5因子分布图

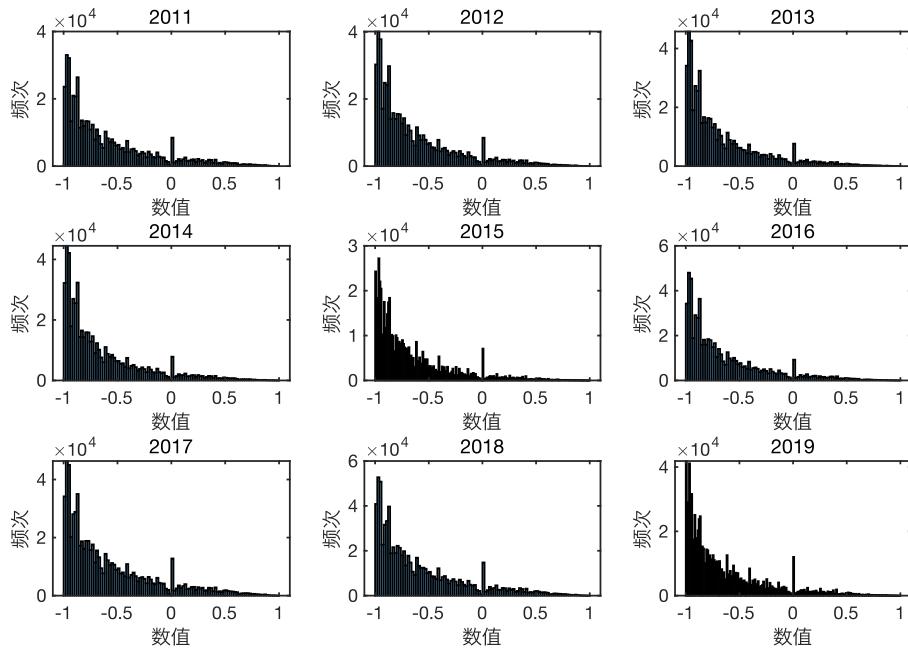


图 3.29 因子五分布图

gene 6因子分布图

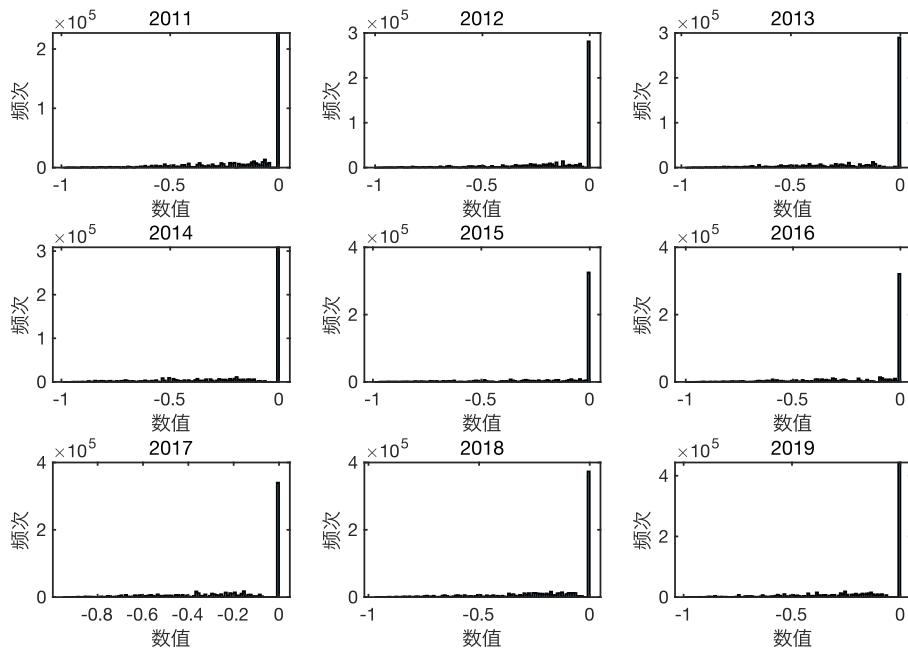


图 3.30 因子六分布图

gene 7因子分布图

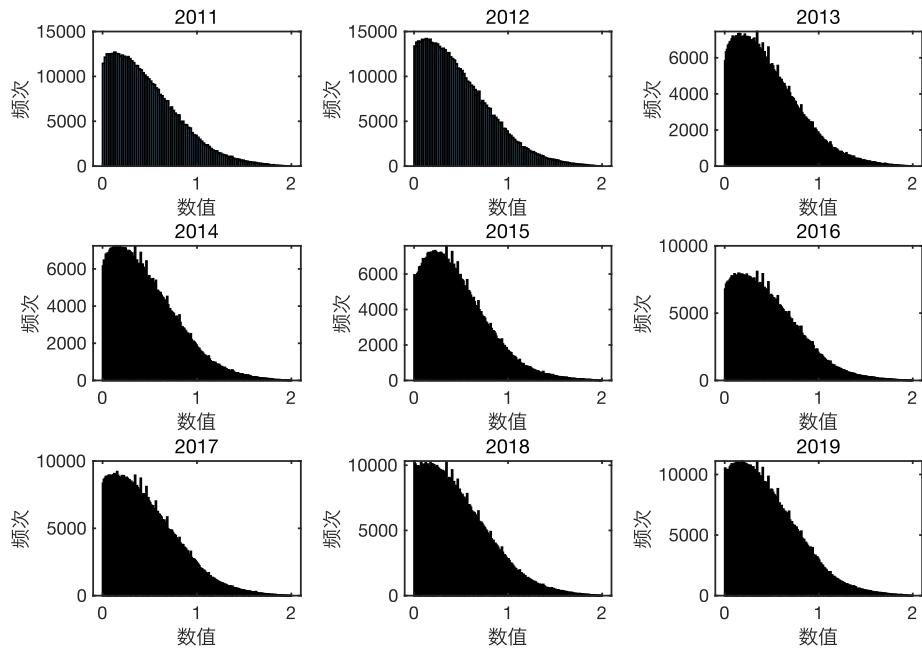


图 3.31 因子七分布图

gene 8因子分布图

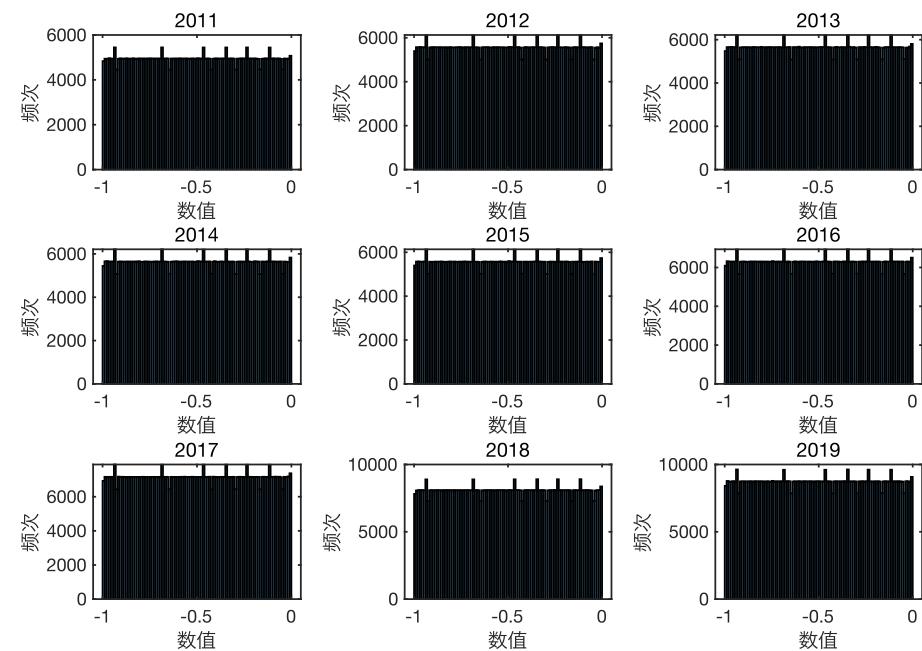


图 3.32 因子八分布图

gene 9因子分布图

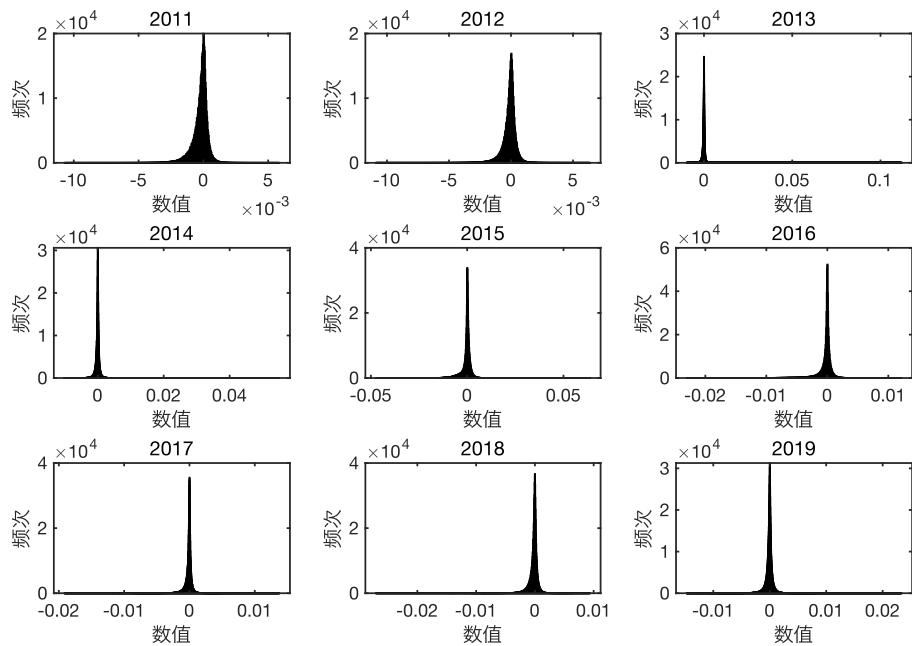


图 3.33 因子九分布图

gene 10因子分布图

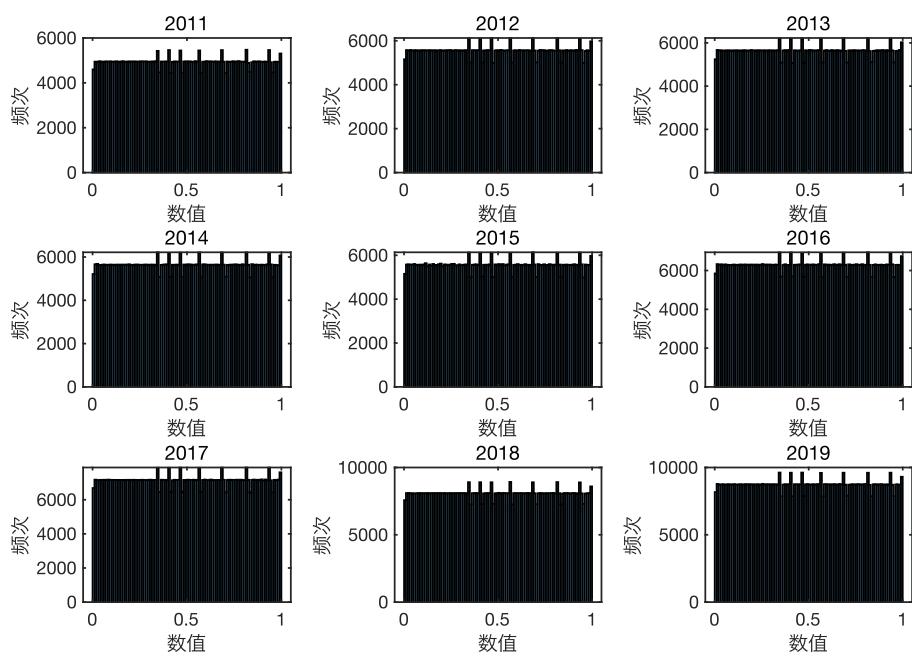


图 3.34 因子十分布图

值得注意的是，从上述基本的分析，可以对因子库，进行简单的删选，得到潜在合理的因子。这里需要注意的是，上述的分析只适合于本文的预测模型，而非普适性的方法。在传统的金融工程中，需要在因子共线性检测做完之后，做各类的行业中性化等等处理。但是在本文中采用的深度学习模型，一些处理可以被深度学习自己所“学习”到。因此不做过多的处理。同时对于为什么还要做共线性处理，笔者的看法是，在深度学习的预测模型中，的确可以不做共线性处理，使用大量的因子，虽然从结果上，深度学习的模型最终可以过滤一些共线性高的因子，也就是择其一而用，但是大量的共线性因子本身就增加了模型训练的时间，这是对于模型不利的。其次，大量共线性的因子，本身也是对于模型的一种干扰。因此，本文的选择因子的逻辑还是遵从，共线性处理，加上后续简单的因子分布形态分析。

3.8 因子优化

对于一些分布形态存在一定问题，但是有效性较为明显的因子，本文做优化处理。典型的，在先前文章中提到了大量的分布因子，但是这些因子没法在上述的分布形态测试中表现的很好或者反应出有价值的信息。但是可以通过对因子定义的变化，形成相对符合要求的分布形态。接下来以 alpha56 因子为例，进行优化和解释。该因子的具体含义为：IF 主力合约的昨日 $dhigh$ 是否大于今日 $dhigh$ ，是则，因子数值为 1，否则因子数值为 0。本身是一个二分类的因子，并且有效性较高。但是在 因子形态分布上，较为混乱。采用如下的措施：将因子定义由 IF 主力合约的昨日 $dhigh$ 是否大于今日 $dhigh$ 改为 IF 主力合约的昨日 $dhigh$ 与今日 $dhigh$ 的比值。

由此，得到的因子分布图为：

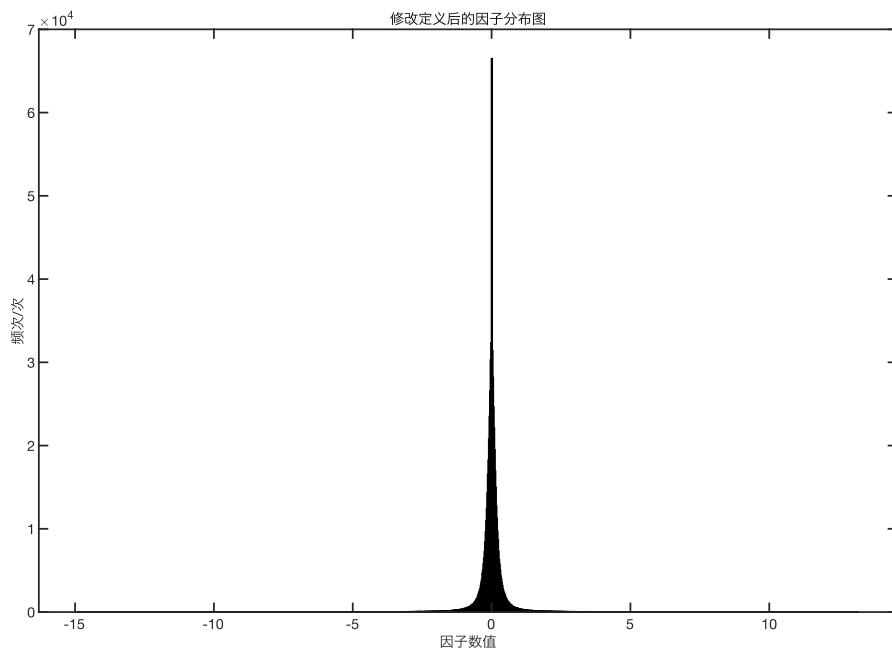


图 3.35 未优化因子分布图

明显，这一因子是存在尖峰现象的，且缺失数值 0 也分布较多。故此对其的取得绝对值，在开开更好进行开更号的数据处理且排除掉 0 数值的情况下；得：

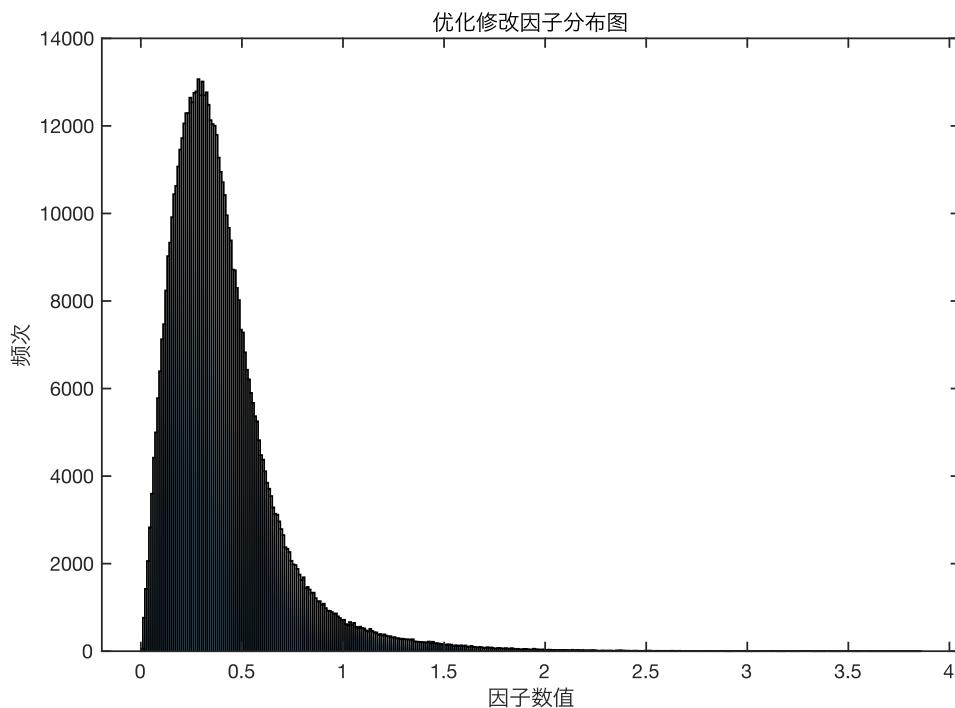


图 3.36 优化后因子分布图

由此，得到的分布，更像是偏态的正态分布。因此，这一因子分布基本满足分布的需求了。同理，大量的分布因子被转化为回归类因子。由于因子数量巨大，暂不在正文内展示。

3.9 因子删选结果

终根据分析结果，本文采用了以下因子：

表 3.5 最终入选因子表（节选）

序号	名称	序号	名称
1	Index_SKDJ_R_1	143	Shape_KLINE_R_10
2	Index_ATR_R_1	144	Shape_KLINE_R_11
3	Label_TIME_R_1	145	Micro_KLINE_C_18
4	Label_SPACE_R_1	146	Micro_KLINE_C_19
5	Shape_KLINE_R_1	147	Micro_VWAP_C_1
6	Shape_ATR_R_1	148	Micro_VWAP_C_2
7	Shape_COR_R_1	149	Micro_KLINE_R_10
8	Shape_TURNOVERRATE_R_1	150	Coupling_KLINE_C_1
9	Shape_POWER_R_1	151	Coupling_KLINE_C_2
10	Shape_POWER_R_2	152	Coupling_KLINE_C_3
11	Shape_POWER_R_3	153	Coupling_KLINE_C_4
12	Shape_POWER_R_4	154	Coupling_KLINE_R_9
13	Shape_COR_R_2	155	Coupling_KLINE_R_10
14	Micro_VP_R_1	156	Coupling_KLINE_R_11
15	Shape_RANK_R_1	157	Coupling_KLINE_R_12
16	Shape_RANK_R_2	158	Micro_KLINE_C_20
17	Shape_RANK_R_3	159	Micro_KLINE_C_21
18	Shape_RANK_R_4	160	Micro_KLINE_C_22
19	Shape_RANK_R_5	161	Micro_KLINE_C_23
20	Shape_RANK_R_6	162	Micro_KLINE_R_11
21	Shape_RANK_R_7	163	FAnalysis_CLASSIFY_C_2
22	Shape_RANK_R_8	164	FAnalysis_CLASSIFY_C_3
23	Shape_COR_R_3	165	FAnalysis_CLASSIFY_C_4
24	Index_MARKET_VALUE_R_1	166	Label_TIME_R_3
25	Shape_ATR_R_2	167	Micro_VWAP_R_1
26	Shape_ATR_C_1	168	FAnalysis_PROFIT_R_2
27	Shape_ATR_C_2	169	FAnalysis_PROFIT_R_3

4 LSTM 模型建立

4.1 预测逻辑概述

在模型预测之前，值得说明的是，由于是对于指数数据进行一个预测，但是指数数据预测，应用深度学习模型的时候存在数据量较小的问题。为此，本文进行了如下处理。先对指数的成分股做预测，在通过指数的成分股的预测结果合成相应的指数的预测结果。

在这里可以对此预测逻辑进行一个简单的验证。假设对个股的预测结果收益(pre)为：

$$pre = return + erros \quad 4.1$$

其中 $return$ 为真实的个股收益， $erros$ 为预测误差， $erros$ 的期望均值为 0。同时又可以将个股的真实收益转化为，个股的 α 收益和 β 收益，且个股的 α 的累加和为市场的 β 收益：具体公式如下：

$$return = \alpha + \beta \quad 4.2$$

由此，将全市场的个股预测收益进行累加，得到了如下的公式：

$$\sum_{n=1}^{all} pre_n = \sum_{n=1}^{all} \alpha_n + \sum_{n=1}^{all} \beta_n + \sum erros \quad 4.3$$

在误差假设均值为 0，且个股 α 收益总和应为全市场的 β 的情况下，公式又可以被转化为如下：

$$\sum_{n=1}^{all} pre_n = \beta + \sum_{n=1}^{all} \beta_n \quad 4.4$$

当 all 比较大的时候，也就是说，则可以将由 α 收益累加构成的 β ，将其忽略，从而对于公式作出一种近似，得到的公式如下：

$$\sum_{n=1}^{all} pre_n = \sum_{n=1}^{all} \beta_n \quad 4.5$$

因为 β 收益是未知但是固定的，所以在通过简单的移项，得到了个股预测的预测值累加和比上使用预测个股的数量可以得到如下的精确解和近似解：

精确解：

$$\frac{\sum_{n=1}^{all} pre_n}{all + 1} = \beta \quad 4.6$$

近似解：

$$\frac{\sum_{n=1}^{all} pre_n}{all} = beta \quad 4.7$$

对于近似解，本质就是一个求均值的过程，即如下：

$$mean(pre, n) = beta \quad 4.8$$

其中，n 为预测 pre 的个数。

因此，根据上述分析，当我应用成分股的时候，单个的预测误差可能会被减小。并且预测的结果可能更加贴合 beta 预测。上述预测是是指的是对于全市场的个股都进行预测，通过预测的 pre 相加而形成的。但是在实际操作中，由于沪深 300 指数本身就可以理解成一种对于真正的市场 beta 收益的近似。因此，实际上按照沪深 300 指数的成分股进行预测，在进行累加，形成的预测是可以接近预测沪深 300 指数的。同样的，既然可以预测了沪深 300 指数，也反应了市场 beta，因此，使用沪深 300 指数的收益来近似市场 beta 的做法也是可以得到支持的。

由此设计预测模型，直接导入沪深 300 指数的成分股的因子数据，以及权重信息，将其用以预测沪深 300 指数对数化收益的标签。但是在这里还有这更加完善的模型，倘若牺牲模型的训练时间问题，可以在网络结构层中多加入一层。也就是说，输入 300 个沪深 300 指数的成分股因子数据，接下来，在第一层，输出结果为，预测 300 个成分股的次日收益 pre。第二层，将这些 pre 进行输入，整合，在经过训练后，输出唯一的预测结果。该结果为指数的预测收益。这一做法，看上去逻辑上是更加完善的。但是也是存在一定问题，接下来开始进行论证，首先在增加一层的情况下，模型的训练时间会变长。其次，尽管，从理论上预测 300 个股的次日收益这一层，300 个预测结果应该是彼此独立的。但是潜在的也是存在一定的问题。大部分情况下，单独使用一个模型预测输入不同的数据，进行 300 次预测的结果，和一次输入 300 组不同的数据，得到的 300 个结果，还是有偏差的。这就说明第一层预测结果可能会增大误差。最后，对于第二层，输入 300 个股的预测数据，输出指数收益的设计而言，是否能有更加明显的效果还是存疑的。并且，这一层的出现无疑是增加了训练时间，并且也增大了潜在的误差。因此，不宜直接采用这类方法。

在考虑到上述情况，本文决定还是采用将 300 个沪深 300 指数的成分股个

股选出，将其相匹配的因子集中。接下来，根据同期的沪深 300 编织权重，富裕不同的因子数据不同的权重，在进行相加。得到一个初步过滤的结果，最后使用过滤后的因子数据，直接投入 LSTM 模型中，进行预测标签。这种做法，不仅在一定程度上增大了训练数据，可以利于深度学习的应用，另一方面也控制了模型的训练时间。

4. 2 建立 LSTM 模型

4. 2. 1 数据匹配与切分

在建立 LSTM 模型之前，首先需要进行数据的预处理工作。准确的来说，数据的预处理部分分为 2 个环节。首先是，数据的匹配。在上文中已经将标签做了处理并且将因子数据也给予了索引数值，需要将这二者的数据进行匹配。由于上文的处理过程中，已经给予了处理。这里只需要将二者的时间日期索引数值相匹配即可。需要注意的是，本身不应该匹配相同时间日期索引，这会导致未来函数的鹅产生。但是在标签部分，本文已经将所有的标签日期索引向前提前了一日，故此，在这一环节，直接相互匹配即可。

其次，需要划分训练数据与测试数据，故此，这里手动给定参数训练比例 0.7/0.3。当然，这一手动给予也并不是完全依靠个人主观经验。大部分轻量级的数据预测模型，会使用这一切分比率。考虑到指数数据，并没有规模很大。因此，采用这一比率也是合理的。这一比率的具体意思为使用全局历史的 70% 的数据作为样本训练数据，用以验证数据，使用剩下的 30% 的数据作为测试数据。这里暂时无需过度思考训练比例的控制，在后续的结果分析中，有更加具体和详细的论证过程。

4. 2. 2 删除恒定值与缺失数值处理

对于删除恒定值，由于本文采用的时间为日级别，但是一些因子涉及使用的是，五分钟的数据，这就容易出现一些因子数据，在日级别的时间步骤中保

持不变的，这可能会对训练产生负面影响。查需要通过找具有相同最小值和最大值的数据，然后逐一分析，是否存在这一现象，进行排除。

对于为什么有这种现象的产生，有如下的考虑。虽然因子在全局上体现出，对数正态分布的状态，但是在深度学习训练模型的过程中，是每次分批取得一小部分的数据，进行训练的。这就造成了一个问题，当在进行日级别的预测时候，使用的因子中含有从五分钟或者小于日级别的周期，所提取的时，就可能因为个股本身的原因或者其他原因，导致个股的这一些因子数值，恒定。这并不是说个股的因子失效，只是说，这些因子在这一个训练轮上失效，可以理解为局部失效。大多数情况下，这种局部失效似乎是可以忽略的，但是为了进一步提高模型的预测准确度，故此，需要对于这种现象进行处理。

对于缺失数据，一般有以下的几种来源，1. 同花顺金融数据库的数据缺失，2. 由于因子设计的本身，存在一些缺失的部分。比如 20 日均线的 19 日应该是没有数据的。通常来说，第一种情况较少，第二种情况较多。因此，首先默认为第一种情况不太会发生，因此，暂时不做考虑。同时，重点针对第二种情况。第二种情况采用给予 8888 来解决。在模型执行前，设置检索是否数据等于 8888 来处理，若数据为 8888，则进行跳过处理。这样子就可以避免错误的产生了。

这样子操作下，就会带来另外一个问题，很有可能由于 8888 的存在，导致某一天的因子数据有 300 个因子构成，但是另外一天却出现了少于 300 个因子构成的因子数据的情况。故此，这里采用基本 MINSIZE 来解决这一问题。这一函数规定了最小的训练规模，当训练规模不到 300 时，可以自动填充，从而保证了每一次的训练数据的规模稳定。值得进一步思考的是，这种默认的使用 0 作为填充的方法，和之前提到的直接将默认缺失值为 0，有什么区别。因此，最为稳妥的方法，就是直接舍去，数据不完整的日期。但是这就容易造成指数数据的不连续和不完整，影响模型的预测。同时由于本身指数数据就并不是十分多，相对于其他领域的 大数据而言，指数数据规模十分的小，在这种背景下，若直接舍去，会对预测模型造成更加恶劣的影响。因此，直接舍去的方法也是存在一定的问题，需要思考。

综上所述，本文进行另外一种处理。由于大部分的因子数据，缺失情况是发生在全局样本数据的初期的。这是因为因子结构造成的，比如，一些均线因子等，需要初始的数据造成的。只有少量的因子数据的缺失是在非初期的。这写一般是因为个股停牌或者其他特殊原因导致的。因此，本文就对全局样本数据的初期的数据，若发生缺失现象，则直接丢弃，以减少不连续现象。对于少部分的非初期因子缺失情况，本文采用基本的 MINSIZE 进行解决。

4. 2. 3 数据的特征缩放

接下来，需要做的是数据的特征缩放处理。这里需要着重解释一下的是，特征缩放（Feature scaling）是一种特征工程。特征工程对于机器学习十分重。因为在原始的资料中，各变量的的范围大且不相同。对于某些机器学习的算法，若没有做过特征工程，目标函数会无法适当的运作。举例来说，多数的分类器利用两点间的距离计算两点的差异，若变量中其中一个特征具有非常广的适用范围，那两点间的差异就会被该特征左右，因此，所有的特征都该被标准化，这样才能大略的使各特征依比例影响距离。另外一个支持特征工程的理由是他能使加速梯度下降法的收敛。

主要是为了在机器学习中，梯度下降的时候，起到加速的作用。一般的特征缩放有以下四种（为了避免混淆，目前把暂时不使用翻译的中文名称，以避免混淆）：

- 1) Rescaling (min-max normalization) 有时简称 normalization

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)} \quad 4.8$$

- 2) Mean normalization

$$x' = \frac{x - \text{mean}(x)}{\max(x) - \min(x)} \quad 4.9$$

- 3) Standardization (Z-score normalization)

$$x' = \frac{x - \text{mean}(x)}{\sigma} \quad 4.10$$

- 4) Scaling to unit length

$$x' = \frac{x}{\|x\|} \quad 4.11$$

一般而言把第一种叫做归一化，第三种叫做标准化。无论哪种方法，其实都需要取决于数据的是否稀疏，或者说数据是否有明确的界限。由于主流采用第一种和第三种计算方法。本文不妨来比较这两种方法：

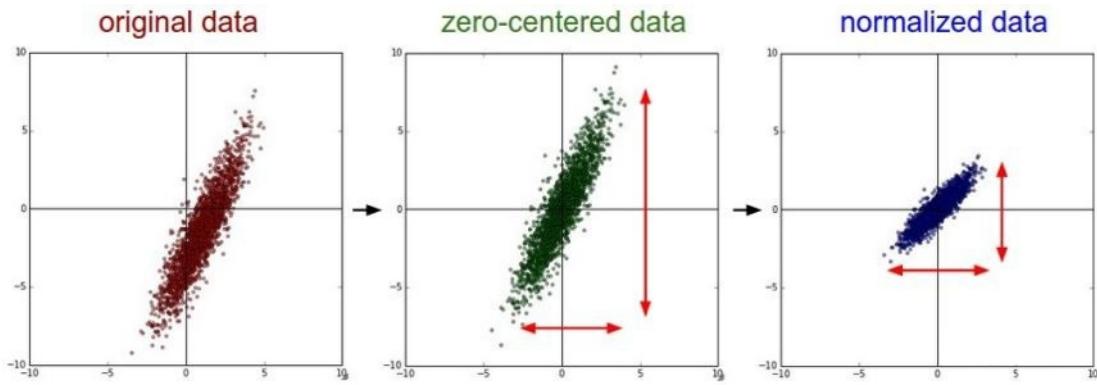


图 4.1 归一化与标准化比较

根据这幅图所显示，显然可以看到归一化和标准化都将数据分布中心移到原点，归一化没有改变数据分布的形状，而标准化使样本数据的分布近似为某种分布（通常为正态分布）。考虑到本文的因子数据，在上文中，进行初步作出因子分布形态的测试与分析了，使得因子数据本身就偏向于是正态分布。因此，在这里采用归一化的处理，这就使得归一化后的数据保留了原始数据的分布形态。当然，使用归一化还有一种考量，这取决于在散点图中，大量的因子部分都是较为集中的，再进一步压缩，反而会使得数据分辨程度不高，而导致模型的预测结果不理想。

4.2.4 确定模型与定义网络架构

首先，关于确认模型部分，这里指的是选择多步预测模型还是单步预测模型。至于是分类预测模型还是回归预测模型，已经在标签处理处，有了论证。LSTM 单步预测模型指的是：LSTM 可用于对单变量时间序列预测问题进行建模。这些都是由单个观测值序列组成的问题，需要一个模型来从一系列先前观测值中学习，以预测序列中的下一个值。在对单变量序列进行建模之前，必须先进行数据准备。LSTM 模型将学习将过去的观测序列作为输入映射到输出观测的函数。

因此，观察序列必须转换为 LSTM 可以学习的多个序列。可以将序列分为多个称为样本的输入模式，其中多个时间步长用作输入，一个时间步长用作输出，用于正在学习的单步预测。

LSTM 多步预测模型指的是：需要对未来的多个时间步长进行预测的时间序列预测问题。具体地说，这些是预测范围或间隔超过一个时间步长的问题。与单步预测一样，用于多步时间序列预测的时间序列必须分为具有输入和输出成分的样本。其中输入和输出组件都将包含多个时间步长，并且可能具有也可能不具有相同数量的步长。

根据上述定义与描述，本文需要的是预测每一日的沪深 300 指数的涨跌幅，因此，这里文章需要的是单步预测模型。首先，在这一章的模型的参数选择是任意，并且没有针对本文的预测标签进行任何优化，在后续的模型分析中有着一定的分析。

接下来开始定义网络结构：

- 1) 指定 LSTM 的隐藏单元，指定 LSTM 层具有 200 个隐藏单元。
- 2) 指定训练选项，将求解器设置为，常用的' adam'，并训练 100 次。同时考虑到梯度爆炸问题，为了避免这一问题的发生，这里将梯度阈值设置为 1。同时，指定初始的学习率为 0.05，然后在训练 40 次后，乘以 0.2 来降低学习率。
- 3) 这里使用均方根误差进行衡量模型预测效果。均方根误差（RMSE）是残差的标准偏差（预测误差）。残差用于衡量距回归线数据点有多少远；RMSE 衡量这些残差分布的程度。换句话说，它解释了数据在最佳拟合线附近的集中程度。均方根误差通常用于预测和回归分析中以验证试验结果。

上述定义和设置，是一个最基本的 LSTM 单步预测模型的配置，在后续的模型分析部分，将对这些参数进行分析和测试。

4.2.5 模型初步结果与评估函数建立

基于上述，得到了初始的预测结果为：

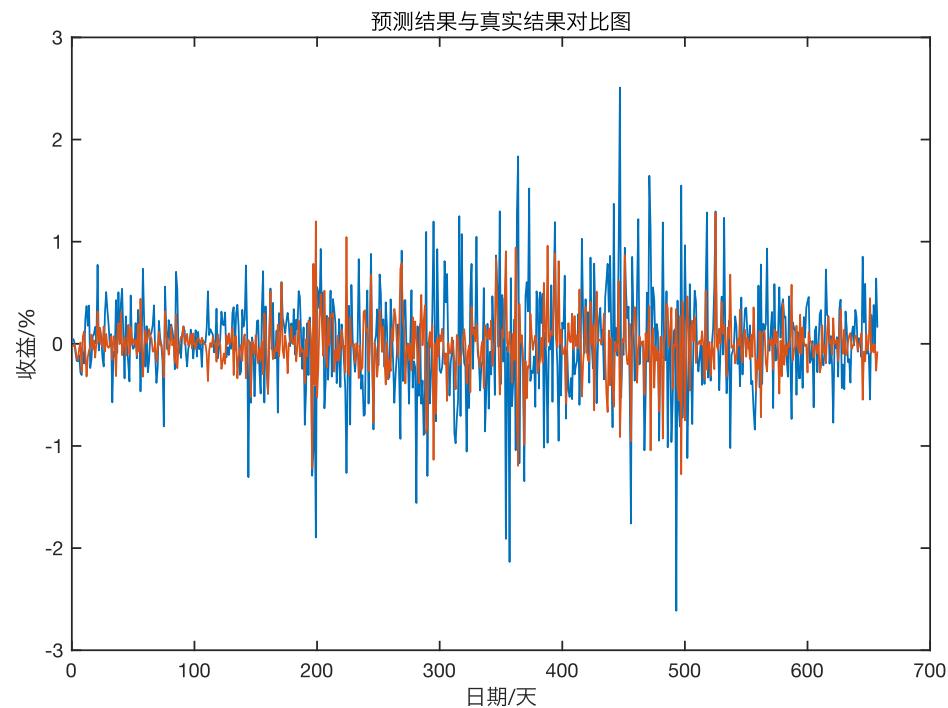


图 4.2 初始预测结果对比图

初始预测的 RMSE 为 16.30。根据初始的预测误差结果，和通过上述预测结果和真实结果对比图来看，初始的预测结果还是相对准确的。为了进一步分析预测结果，接下来定于一种评估函数：

同时在这里初步定义一种预测准确率的估计方法，将预测的数值和真实数值正负号一致的情况，认为 1，而不一致的情况认为 0，统计全局范围内 1 出现的次数，将其比上全局的测试天数。该数据可以反应预测模型的预测准确性。这一模型的预测准确率为 0.40（保留两位小数）。显然这一初始的模型预测准确率是不足以投入使用的。

再结合本文的预测目标，对于金融风险的评估。因此，应该针对指数收益表现较差和很差的时候，进行预测。因此，根据这一目标，提出两种新的评估函数：

第一种：统计测试集上，所有的指数收益为负的日期总和，记做 a，统计测试集上，所有指数收益为负的日期中，预测收益收益为负的日期综合，记做 b。 $c = b/a$ ，c 就为模型的风险预测准确率。

第二种，统计测试集上，所有的指数收益为极端负的日期总和，记做 a，统计测试集上，所有指数收益为极端负的日期中，预测收益收益为极端负的日期综合，记做 b。 $c = b/a$ ，c 就为模型的风险预测准确率。其中指数收益极端负的定义，为当天收益为过去十天内的最小收益。预测收益也同理。

上述两种评估函数，基本上可以满足本文对于金融风险的预测结果的评估。在这里对设计评估函数的逻辑进行论证：

实际上，这一评估函数的设计，类似统计学的第一类与第二类误差，只不过在这里进行了进一步的拓展：首先将测试集的指数收益划分为四类：指数收益盈利较大，指数收益盈利一般，指数收益亏损一般，指数收益亏损较大。这四类预测标签对应了，下表中的十六种打勾标记的预测结果可能性。

表 4.1 预测结果表

实际结果 /	预测结果			
	收益盈利较大	收益盈利一般	收益亏损一般	收益亏损较大
收益盈利较大	√			
收益盈利一般	√			
收益亏损一般	√			
收益亏损较大	√			
收益盈利较大		√		
收益盈利一般		√		
收益亏损一般		√		
收益亏损较大		√		
收益盈利较大			√	
收益盈利一般			√	
收益亏损一般			√	
收益亏损较大			√	
收益盈利较大				√
收益盈利一般				√
收益亏损一般				√
收益亏损较大				√

其中，对于本文，应该关注的是指数收益亏损较大和亏损一般的两种标签特征。而其中会对投资造成极大的损失的有如下几种情况：1. 预测收益盈利较

大，实际亏损较大；2. 预测收益盈利较大，实际亏损一般；3. 预测收益盈利一般，实际亏损一般；4. 预测收益盈利一般，实际亏损较大；5. 预测收益亏损一般，实际盈利较大；6. 预测收益亏损一般，实际盈利一般；7. 预测收益亏损较大，实际盈利较大；8. 预测收益亏损较大，实际盈利一般。

综合概括来说，就是预测结果与真实结果异向相背离。注意是异相背离！由此，进行提取，有了上述的三种预测结果评估函数。

基于上述评估函数，得到了下述的初始预测结果模型信息：

表 4.2 初始预测结果评估表

评估函数	结果
EMSE	16.3
第一类评估函数	0.4
第二类评估函数	0.4
第三类评估函数	0.17

根据对预测结果的评估函数分析，基本可以发现，初始的预测结果全局预测准确率不高，为 0.4，对于指数收益亏损预测准确率为 0.4，说明初始预测模型在预测指数收益为正的时候准确率为 0.6，要比预测亏损要准确。但是这是不符合本文的预测需求的。而对于极端亏损的指数收益预测，准确率只有 0.17，预测效果极度不理想。初步来看，初始预测结果是不怎么符合需求的，需要进一步非优化与分析

4.3 模型分析与优化

4.3.1 训练比率

训练比例一般是基础的机器学习的参数。通常取得机器学习的训练比率为 6:2:2，其中 8 为训练比例，2 为测试比例。但是当训练数据的规模特别大的时候，在上万级别的时候，这三者的训练比例可以变为 98:1:1。由于本文的标签是

是指数收益，指数的特征就是数据量小，并没有特别大的数据量。因此，在这里还是不能采用 98:1:1 的模式进行训练的。

这里本文将验证比例尽心固定，为 2，将训练比例和测试比例进行轮动。形成以下的图，来分析训练比例与预测模型的结果关系。

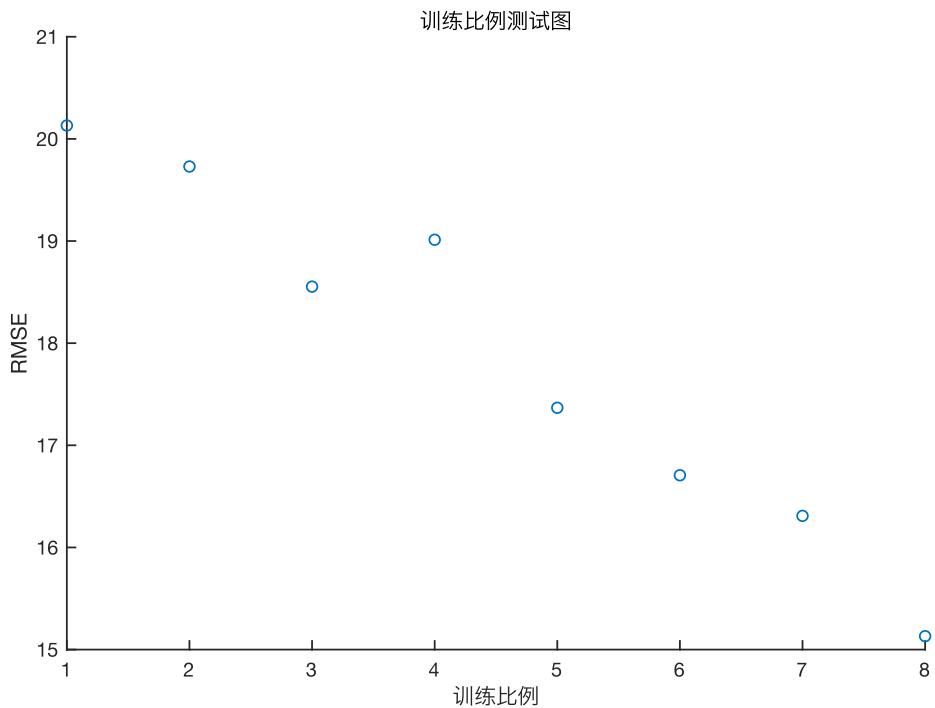


图 4.3 训练比例测试图

根据上述结果分析，基本上符合训练比例越大，RMSE 的数值越小的特点。虽然样本数据比较少，但是还是基本可以观察到，在训练比例在 5, 6, 7 这一附近时，RMSE 的缩小程度有所减缓，因此，将维持训练比例 7 的设置。

这一部分还是有进一步优化的空间的，可以通过更加小的间距切分，以获得更加准确的和精细的结果。但是这同样会造成，过拟合现象的产生。故此，暂时不做过多深入的测试。

4.3.2 学习率

机器学习的一种方法为梯度下降法，即对损失函数数进行缩小。因此，在这

一缩小过程中，学习率控制的是参数的更新速度，学习率设置过大，会导致损失函数在极小值附近来回变化，不收敛，学习率设置过小，会导致学习速度变慢。这里本文设置一个较大的学习率和一个衰减系数，让损失函数在刚开始以的较快的速度下降，随着训练次数增加，学习率也逐渐变小，使用损失函数以一个比较慢的速度收敛。

这里测试了初始学习率从 0.5 到 0.05 的范围内，每 0.05 个变动下的散点图。

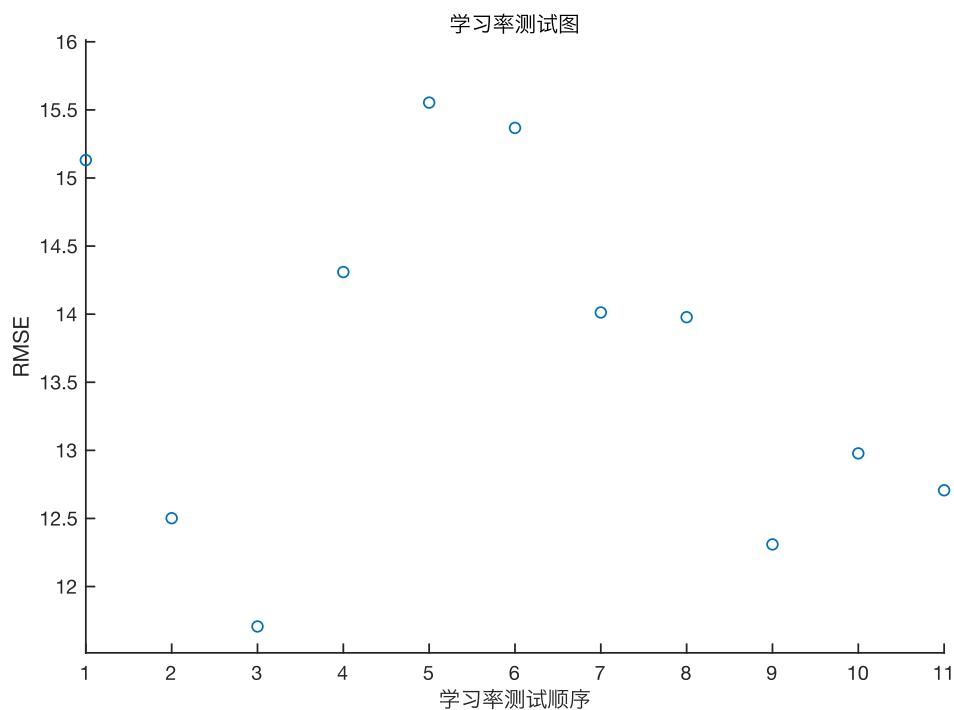


图 4.4 学习率测试图

基本上，可以观察到序号的代表的学习率 0.15 开始的学习率是对于 RMSE 的下降是有着明显的改善的。并且通过改变学习率，带来的性能提升是最大的。因此选择这一学习率来代表初始的学习率。后续虽然随着学习率低，RMSE 也会在降低，但是因为样本数量比较少，可能存在一定的收敛问题。同样的，这一部分的分析也受到样本抽样的数量的干扰。更为准确的做法应该是，减小每一个测试学习率低的间距，多进行测试。但是这就会导致模型训练的时间不断增多。而且还是无法排出过拟合的问题。

4.3.4 迭代次数

大部分的情况下，迭代次数意味着训练的深度。通常增加迭代次数，可以提升模型的效果，但是过度的迭代，容易导致出现过拟合的问题。因此，这里比较了，20 次迭代到 160 次迭代的结果散点图。

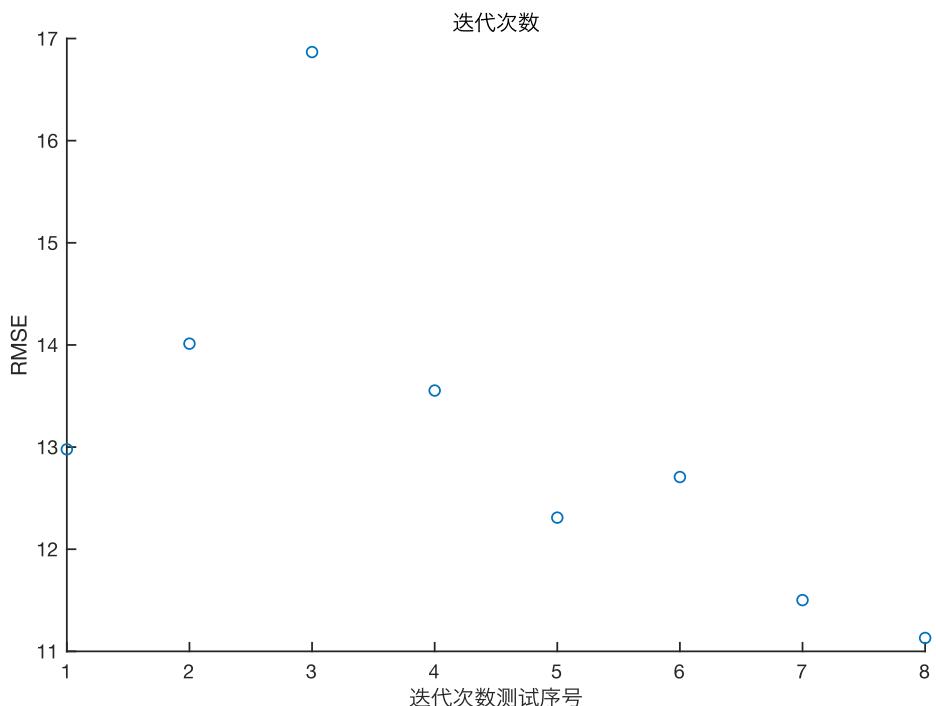


图 4.5 迭代次数测试图

从上述图分析，随着迭代次数的增加，RMSE 是不断减小的，但是减小的程度是逐渐减小的。这也说明迭代次数的增加，带来的模型预测性能的提升正在逐渐减小。因此，这里选择序号 4 所对应的迭代次数作为优化测试的学习率。理由基于此：迭代次数的增加在序号 4 处，带来的性能提升是最为明显的。

4.3.5 初始结果与优化结果对比

根据上述三个进行参数优化，先给出对比的参数表。初始参数与优化参数表如下：

表 4.3 初始与优化使用参数表

参数	初始数值	优化数值
学习率	0.05	0.15
迭代次数	40	80
训练比例	6/4	6/4

这里也给出预测结果表，初始预测结果和优化后预测结果表：

表 4.4 初始与优化预测结果评估表

评估对象	初始结果	优化结果
RMSE	16.3	12.8
第一类评估函数	0.4	0.65
第二类评估函数	0.4	0.63
第三类评估函数	0.17	0.31

最后，在这里在给出初始预测结果表与优化后预测结果对比图：

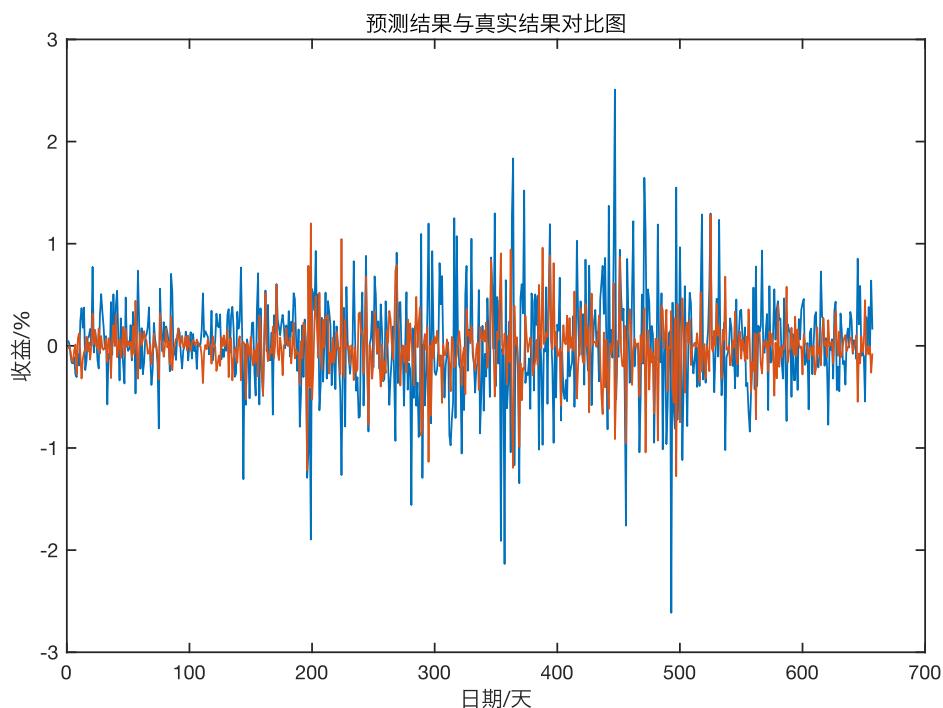


图 4.6 初始预测结果测试图

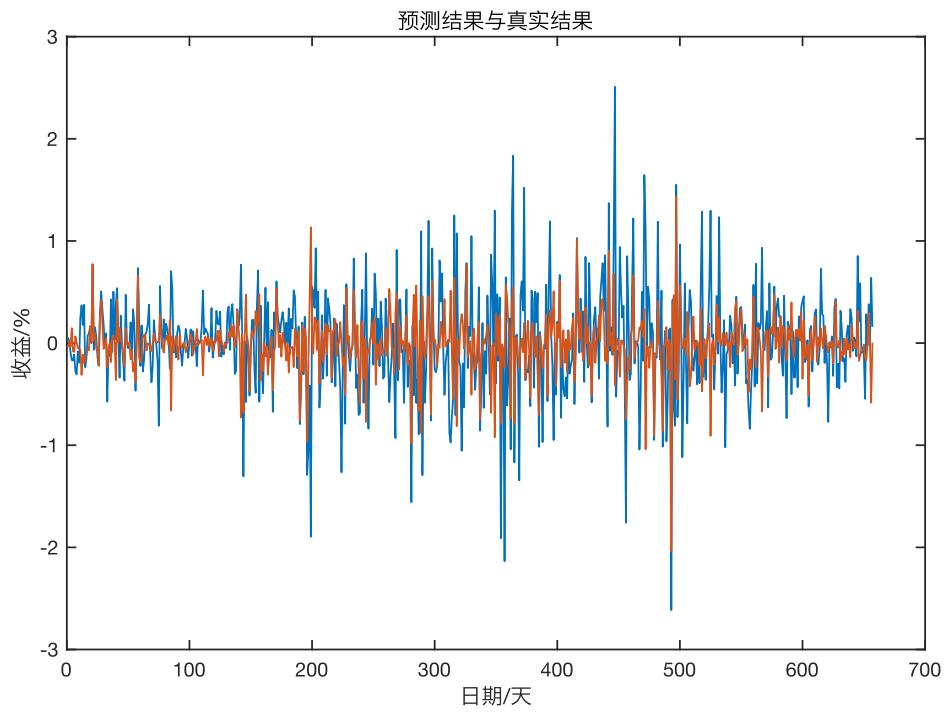


图 4.7 优化预测结果测试图

根据两张预测结果与真实结果对比图与预测结果评估表，基本可以发现，优化后的预测模型，全局预测准确率有一定的提升，从 0.4 上升到了 0.65，并且模型的 RMSE 的从 16 降低到 12。更为重要的一个特点是，对于指数收益为负的预测准确率从 0.4 上升到了 0.63，这就意味着，模型对于指数收益为负的预测情况，从预测不准，到预测较准有了提升。说明预测模型满足了本文的预测目标。此外，对于极端亏损指数收益的情况，优化后的预测准确率为 0.3 这是一个极大的提升。

综上所述，预测模型的优化是比较成功的，优化后的预测模型，对本文定义的金融风险有了较好的预测。并且对于极端的亏损情况，也就是风险极端高的情况，有了预测性能的提升。因此，这一优化后的预测模型是十分有效的，有应用的价值。

4.3.6 全因子预测与因子删选预测结果对比

在这里也需要对因子删选的作用进行论证，直接使用全因子模型和删选后因子进行对比，可以论证是否因子删选有作用。首先给出全因子和删选后因子参数：

表 4.5 全因子与删选后因子参数表

参数	删选后因子数值	全因子数值
学习率	0.05	0.05
迭代次数	40	40
训练比例	6/4	6/4
因子数量	删选后	全因子

进一步给出，预测结果的评估结果表与预测结果对比图：

表 4.6 全因子与删选后因子预测结果评估表

评估对象	删选后结果	全因子结果
RMSE	16.3	17.57
第一类评估函数	0.4	0.3
第二类评估函数	0.4	0.27
第三类评估函数	0.17	0.11

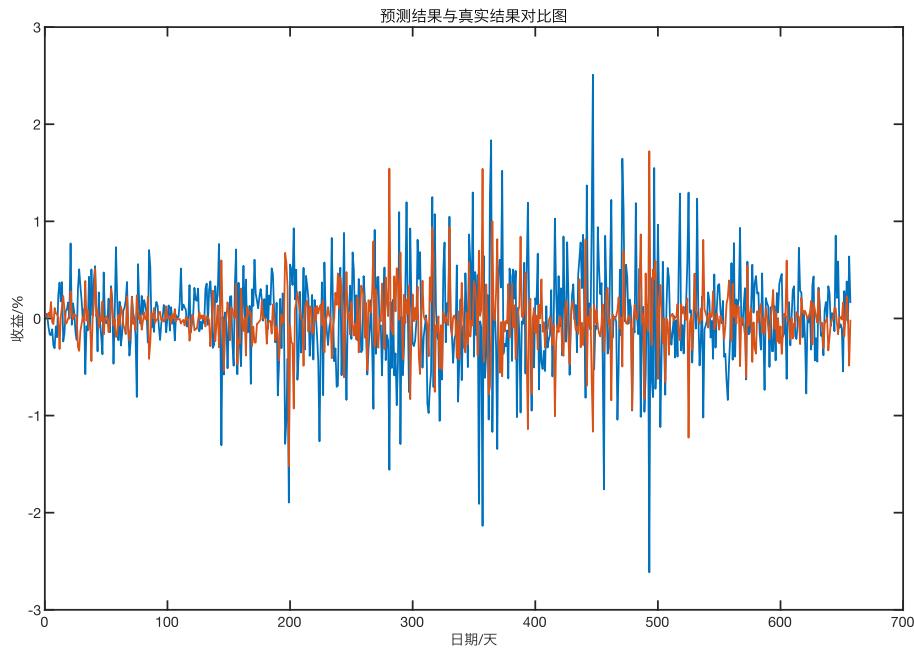


图 4.8 删选后因子预测结果测试图

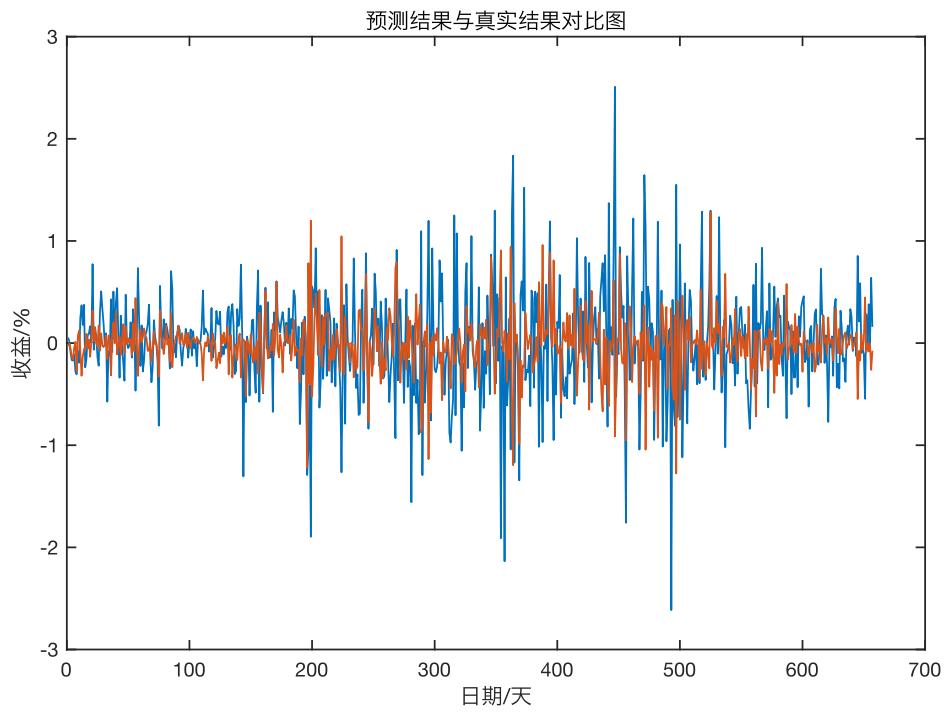


图 4.9 全因子预测结果测试图

根据两张预测结果与真实结果对比图与预测结果评估表，可以看到删选后因子的预测准确率无论是在全局预测准确性上（0.4），还是在收益为负的时候的预测准确性（0.4），还是在指数大跌日期的预测（0.17）上，都是优于全因子预测的。并且在模型的 RMSE 的计算上，删选后的因子明显是减小了误差的。因此，基本可以得到因子删选的过程是有意义的。对于提高模型的性能而言，这是有正向的积极作用的。

综上所述，预测模型的因子删选的过程是有积极作用的，但是目前并不能对于因子删选的三个步骤中的哪个步骤对于预测模型产生了更大的作用是未知的。从严格意义上讲，有效性分析、共线性处理、分布形态测试这三个步骤都可能与预测模型紧密相关。由于本文重点研究是风险预测，并非因子预先处理过程对于预测模型的作用。因此，在这里不对这三个步骤的影响程度做重点分析。但是本文给出以下的基本猜想和支撑依据：

首先，有效性分析、共线性处理、分布形态测试这三个步骤，在本文的背景的下，对预测模型产生的影响从大到小的排序应该为：因子有效性分析、分布形态测试、共线性处理。之所以给出以上的猜想，基本依据如下：

已知因子的有效性是任何预测的前提。正如任何人无法利用一堆无效的海量数据，对一个标签进行准确预测一样。只有因子和标签的相互联系（不只是因果关系）越密切，预测模型的准确率才会越高。倘若因子和标签的联系十分清晰且密切，那么即使不使用复杂的模型工具，也可以通过简单分析，得到预测。

其次，分布形态测试对于预测模型的性能与本文所采用的归一化处理有关系。本文并没有采用标准化的理由，已经在上述文章中得以论述。而归一化是保留了数据的原始分布形态的，那么分布形态测试就对预测模型的性能产生了影响。

对于共线性，其实在大部分的学者已经证实：在预测模型中，只要拟合的恰当，因子之间的共线性基本可以对预测模型不产生干扰。故此，对于有效性分析、共线性处理、分布形态测试这三个步骤对于预测模型的影响，本文暂时作此猜测。

4. 3. 7 对数化收益和非对数化收益的对比

在这里也需要对标签的特征进行论证，直接使用对数化标签模型和非对数化标签模型进行对比，可以论证对数化标签的作用。首先给出二者的参数：

表 4.7 对数化与非对数化标签参数

参数	非对数化标签模型数值	对数化标签模型数值
学习率	0.15	0.15
迭代次数	80	80
训练比例	6/4	6/4
标签类型	非对数化	对数化

进一步给出，预测结果的评估结果表与预测结果对比图：

表 4.8 对数化与非对数化预测结果评估表

评估对象	非对数化结果	对数化结果
RMSE	35.7	12.8
第一类评估函数	0.49	0.65
第二类评估函数	0.44	0.63
第三类评估函数	0.25	0.31

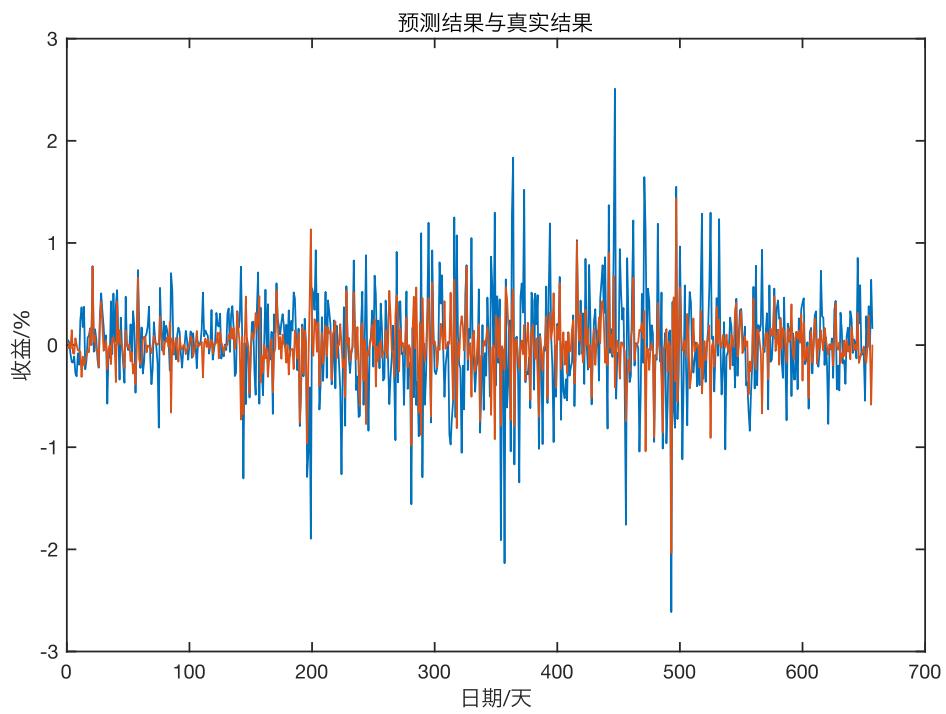


图 4.10 优化后对数化收益预测结果

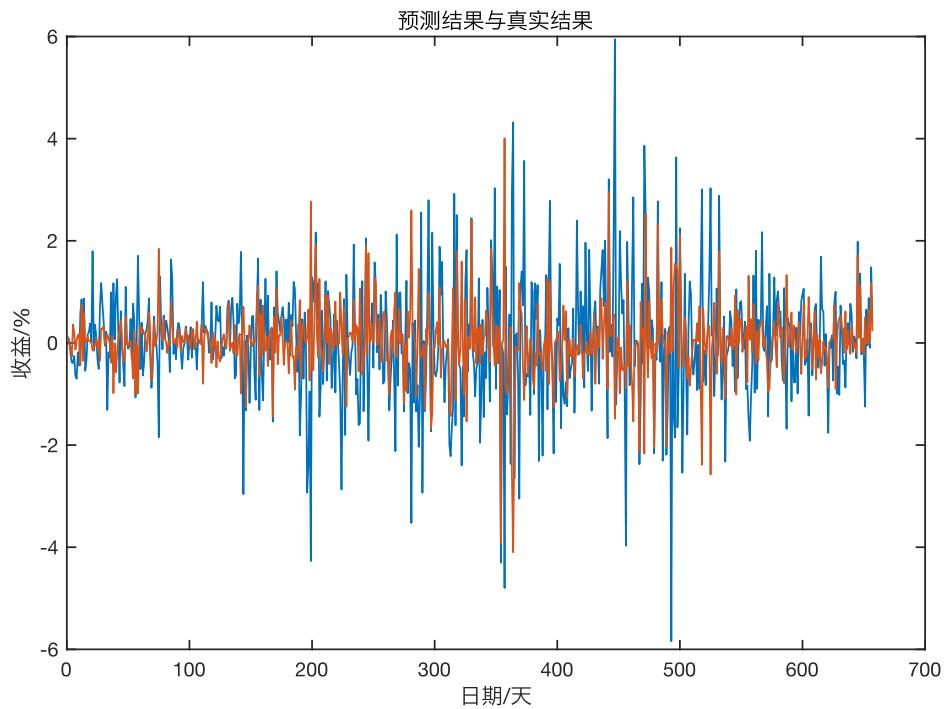


图 4.11 优化后非对数化收益预测结果

根据两张预测结果与真实结果对比图与预测结果评估表，可以看到对数化只是收益后的模型的预测准确率无论是在全局预测准确性上（0.65），还是在收益为负的时候的预测准确性（0.63），还是在指数大跌日期的预测（0.3）上，都是优于非对数化收益的模型的预测的结果。但是这二者最大的差异在于，二者的 RMSE，对数化后的并且在模型的 RMSE 的计算上，远小于对数化后的预测模型的 RMSE。因此对数化指数收益的标签显然是具有更好的预测效果的。对于为何非对数化后的指数收益的标签的 RMSE 的表现远大于对数化后的，本文也给出了如下的猜想。对数化后的标签，极差被减小了，整体标签序列也趋向于平稳，适合模型的学习和应用预测。

4.3.8 对比多元非线性回归

最后，在文章这里将优化后的结果和一般的多元非线性回归（多项式拟合）的预测结果进行比较，预测模型的结果如下：

表 4.9 LSTM 模型与多元非线性回归预测结果评估表

评估对象	多元非线性回归	LSTM
RMSE	16.1	12.8
第一类评估函数	0.35	0.65
第二类评估函数	0.36	0.63
第三类评估函数	0.17	0.31

根据两张预测结果与真实结果对比图与预测结果评估表，可以看到对优化后的预测模型无论是在全局预测准确性上（0.65），还是在收益为负的时候的预测准确性（0.63），还是在指数大跌日期的预测（0.3）上，都是优于多元非线性回归（多项式拟合）。这说明，本文设计且优化后的模型的预测性能是优于一般的模型的。因此，本文所建立的 LSTM 预测模型的性能是优秀，且预测准确率是可以做到应用的。

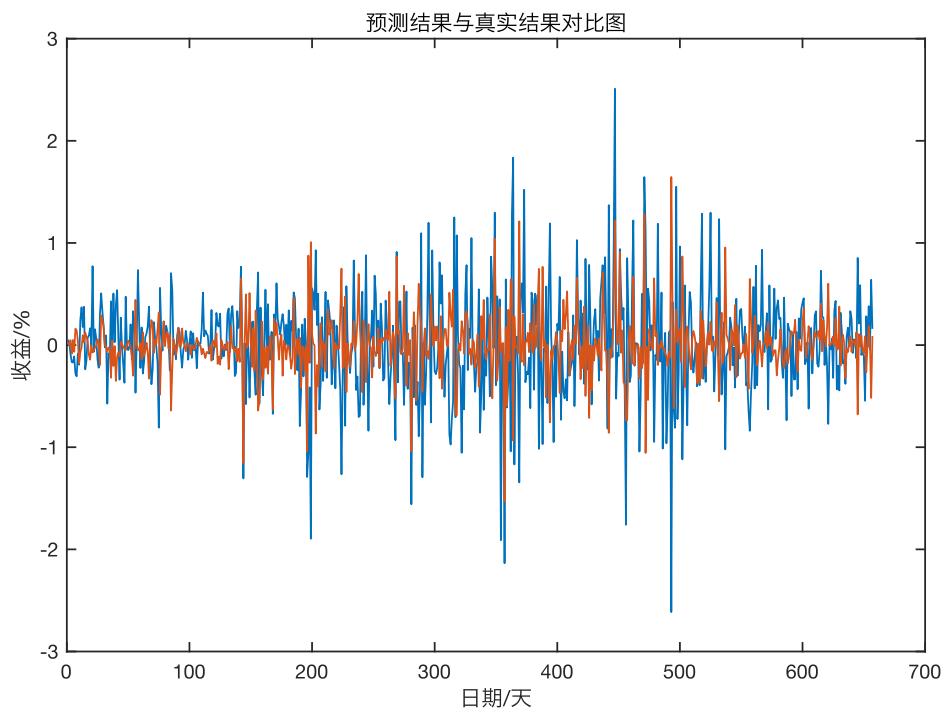


图 4.12 多元非线性回归预测结果

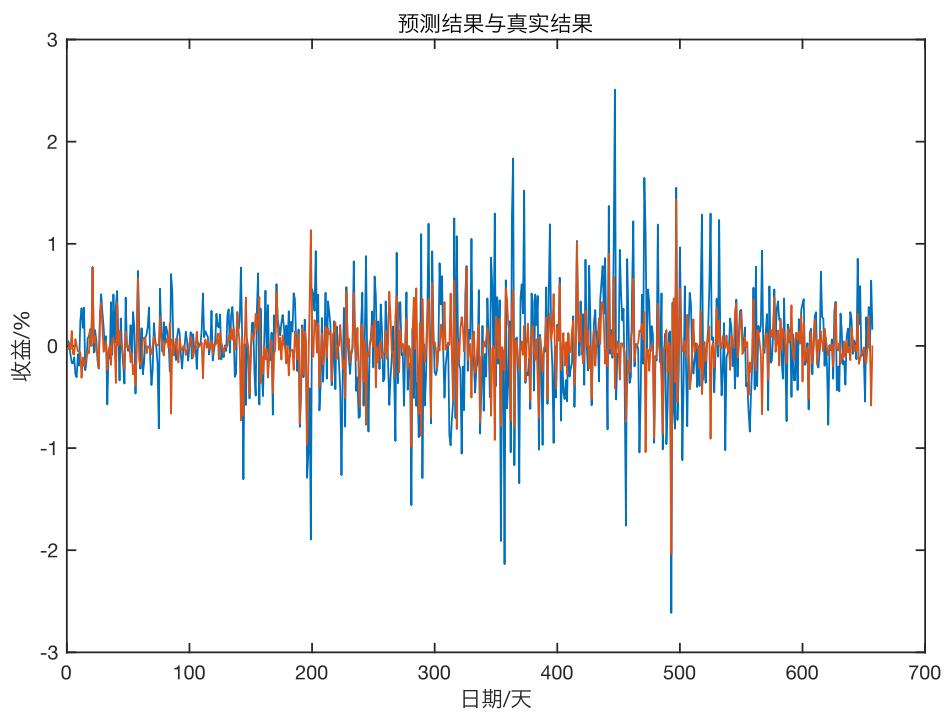


图 4.13 最优 LSTM 预测结果

4.3.9 思考与不足

本文的模型也是存在一定的问题的，可以分为三个部分来进行反思：

首先，训练比例层面上，有进一步优化的空间的，文章虽然可以通过更加小的间距切分，以获得更加准确的和精细的结果。但是这潜在会造成，过拟合现象的产生。

学习率测试的过程中，因为样本数量比较少，可能存在一定的收敛问题。同样的，学习率的分析也受到样本抽样的数量的干扰。更为准确的做法应该是，减小每一个测试学习率低的间距，多进行测试。但是这就会导致模型训练的时间不断增多。而且还是无法排出过拟合的问题。

上述的两个部分，提到的过拟合和缩短间距进行更加精准的测试之间的矛盾，在文章中，并没有很好的解决。

其次，和所有的机器学习模型一样，都势必会遇到过拟合的问题。诚然，本文对于模型的过拟合的处理，显然还是缺少措施的。过拟合通常指的是所选模型的复杂度比真模型更高；学习时选择的模型所包含的参数过多，对已经数据预测得很好，但是对未知数据预测得很差的现象。通常过拟合一般特点：高方差，低偏差；而导致过拟合原因很多有：训练数据不够，模型进行过度训练(overtraining)等一般有以下的几种措施进行避免：

- 1) Early stopping (适当的 stopping criterion) : Early stopping便是一种迭代次数截断的方法来防止过拟合的方法，即在模型对训练数据集迭代收敛之前停止迭代来防止过拟合。
- 2) 数据集扩增 : 数据机扩增即需要得到更多的符合要求的数据，即和已有的数据是独立同分布的，或者近似独立同分布的。一般方法有：从数据源头采集更多数据，复制原有数据并加上随机噪声，重采样，根据当前数据集估计数据分布参数，使用该分布产生更多数据等
- 3) 正则化方法：一般有 L1 正则与 L2 正则等。
- 4) Dropout: 正则是通过在代价函数后面加上正则项来防止模型过拟合的。而在神经网络中，有一种方法是通过修改神经网络本身结构来实现的，其名为 Dropout。

但是在文章中只是选择了，正则化，并且也没有对于这四种方法进行讨论分析与

论证。

最后，对于调参部分，本文是简单的遍历调参。实际上，目前学术界应用广泛的是贝叶斯优化（自动机器学习超参数调整）贝叶斯优化通过基于目标函数的过去评估结果建立替代函数（概率模型），来找到最小化目标函数的值。贝叶斯方法与随机或者遍历的不同之处在于，它在尝试下一组超参数时，会参考之前的评估结果，因此可以省去很多无用功。

超参数的评估代价很大，因为它要求使用待评估的超参数训练一遍模型，而许多深度学习模型动则几个小时几天才能完成训练，并评估模型，因此耗费巨大。贝叶斯调参则使用不断更新的概率模型，通过推断过去的结果来“集中”有希望的超参数。

但是，对于本文，还是比较合适的。但是本文未进行探索。

4.4 模型测试结论

综上所述，本文所建立的深度学习的模型还是积极有效的。优化后的模型预测性能要比多元非线性回归（多项式拟合）的效果更加出色。同时模型测试阶段实证发现，对于指数收益的对数化处理，可以提升模型的性能。这潜在与对数化后，标签，也就是预测对象更加平稳有关系。

同时在分析与测试模型的阶段，证明了预测模型的因子删选的过程是有积极作用的，但是目前并不能对于因子删选的三个步骤中的哪个步骤对于预测模型产生了更大的作用是未知的。但是从初步测试来说，有效性分析、共线性处理、分布形态测试这三个步骤都可能与预测模型紧密相关。

其次，本文根据文章预测的需求，对于金融市场（股票市场）进行风险评估。也就是指数的价格风险的预测。设计了选择了 RMSE 和三组自行设计的评估函数。三组自行设计的评估函数，分别针对预测模型的全局预测准确率、指数收益为负的情况下预测准确率、指数收益极端差的预测准确率。这三组评估函数很好的评估了模型应用价值。

其次，根据模型的最优参数配置的预测结果而言，虽然对于极端差样本的预测准确率还是不高，但是对于指数收益为负的预测准确率已经相对较为准确。

最后，文本的模型也是存在一定的问题的。对于过拟合应对上，没有进行深入的分析。在参数优化的时候，尚未使用超参进化等方法。但是，初步得到的模型的价格还是很高的。

5 结论

本文采用的预测和评估方法是利用因子来预测成分股，再利用成分股个股的预测结果合成指数的预测结果。在文章中，先给出了生成个股因子数据的数据来源的解释和说明，并且对这些原始的数据进行了符合文章数据要求的处理。这里，处理指的是对于原始股票数据的后复权的数据处理。

其次，对于设计标签而言。本文最终预测的是沪深 300 的指数数据。因此，预测的对象应该是指数。但是由于本文是通过预测个股合成指数，再来预测指数的。因此，个股也是本文的预测对象。但是，对于机器学习的模型而言，标签应该是只有一个的，那就是指数的收益。

对于标签——指数收益，应该需要考虑了以下几个方面，使用传统的当天收盘价（close）到当天开盘价（open）的计算方式是存在指数而言存在跳空等现象，容易出现收益率不连续的情况。这一弊端的影响，类似个股使用未复权数据进行预测的情况。因此，为了将跳空情况考虑在内，本文使用计算指数收益率的时间范围为，当天指数的收盘价（close）到次日的收盘价（close）的指数收益变化的时间。

在给定计算指数收益的之间之后，同样给予指数收益的计算方法给予定义，这里采用两种计算方法，一种为对数化的计算收益方式，另一种为非对数化的计算收益，具体定义在上文的公式中给出。同时，在计算标签的时候，还需要给予标签一个索引数值，这是为了将标签和因子数据最终匹配起来。这里采用较为方便且好分辨的时间作为索引数值。值得注意的是，根据本文定义的指数收益率计算时间范围内，索引数值应该被定义为次日的收盘价的时间，也就是次日的日期为这一收益率的日期索引数值。但是本文涉及的是一个预测模型，因此，模型的因子数据与标签数据的时间上，应该被要求错开。也就是：标签数据为 T1 日，而因子数据为 T0 日。这样子才可以避免未来函数的影响，提高模型泛化能力。

在实际操作中考虑到因子数据的庞大，而标签的数量的较少的情况，且为了方便后续代码程序中的操作，故此，本文将标签的索引值集体向前移动一天。这也就意味着，标签原先的日期索引值为次日的日期，现在变为了当日的日期。这样子，在后续的预测模型的因子数据和标签的匹配中，只需要将具有相同的日期索引值的数据匹配在一起就可以了，这既避免了模型的未来函数的出现，提高了泛化能力，也方面了程序实现的代码工作量被大大减小了，并且代码错误风险降低了。

在规范定义且计算了标签之后，将进行因子设计。在设计因子之前，在本文将从以下几个方面阐述了进行因子设计的逻辑。

第一类：单一指数类，单一指数类。

第二类，跨指数的对比，跨指数的对比具体而言就是沪深 300 与中证 500 的对比。

第三类，期现对比，也就是指数与其对应的期货的对比。

第四类，期货的远近合约对比，远近期货合约对比主要是依靠价差来反映一个基本的市场情绪，从而对预测现货的价格起到积极的作用。

第五类，高波动率的个股统计，高波动率的个股反映的是市场中情绪最为激烈的票，因此，对于预测指数会有着积极的作用。

第六类，特殊个股统计也就是 st、近期复牌、与沪深 300 指数低相关的个股。

第七类，国际市场特殊品种，国际市场的特殊品种有包括黄金、白银、原油这三个主要品种。

第八类，国际指数。

在上述因子设计的逻辑下，进行因子设计后，可以得到一些列的初始因子。但是这些因子并不是都可以使用的。因此，进行因子有效性检验，主要采用三种计算方法进行验证，分别为 xy 图，相关性检测，线性拟合的 K 数值。通过这三种方法得出初步的可用因子的列表。

之后，进行共线性检验。排除一些类具有高度线性相关的因子，减少代码的工作量以及模型的运算负荷。考虑到通常模型要是仅仅用于预测的话，一般可以不处理多重共线性问题，因为这并不会极大程度的影响预测结果。但是考虑到本文的因子数量较多，如果不做处理，算力负荷过大。因此还是进行处理。

值得申明的是，需要先做的是因子有效性检测，在进行的是因子共线性检测，因为倘若先进性因子共线性，存在将一些有效因子错误排除的情况。具体情况会在下文中有详细的说明。

最后需要对一些因子的分布形态做进行的删选，将一些因子分布形态上并不接近对数正态分布的因子进行排除。原因是股票市场的回报率基本符合对数正态分布，因此采用相近的形态的因子进行预测，或许在准确性上有所提升。同时，对于接近对数正态分布的定义为：正态分布、卡方分布等，因为这些因子分布基本可以通过模型的迭代得到函数变化，转为对数正态分布。而对于一些被排除的因子，可以进行简单的优化得到对数正态分布。

之后将因子数据与标签进行匹配，同时根据不同特征缩放的特点，采用了归一化的特征缩放方法，原因是其保留了原始的因子形态。之后进行了简单的网络结构的定义，并且从训练比率、学习率、迭代次数三个层面不断进行测试和调参。

之后，将优化后的结果与未优化、多元非线性回归模型进行比较分析。得到了本文的所建立的深度学习的模型还是积极有效的结论。具体体现为：1) 优化后的模型预测性能要比多元非线性回归（多项式拟合）的效果出色。2) 对于指数收益的对数化处理，可以提升模型的性能。3) 预测模型的因子删选的过程是有积极作用的，但是目前并不能对于因子删选的三个步骤中的哪个步骤对于预测模型产生了更大的作用是未知的。4) 根据模型的最优参数配置的预测结果而言，虽然对于极端差样本的预测准确率还是不高，但是对于指数收益为负的预测准确率已经相对较为准确。

最后，文本的模型也是存在一定的问题的。对于过拟合应对上，没有进行深入的分析。在参数优化的时候，尚未使用超参进化等方法。但是，初步得到的模型的价格还是很高的，并且还有更大的提升空间。

参考文献

- [1] KAUFMAN G G. Banking and currency crises and systemic risk: Lessons from recent events [J]. *Economic Perspectives*, 2000, 24(3): 9-28.
- [2] 高国华. 基于系统性风险的银行资本监管及其宏观经济效应 [D]; 上海交通大学, 2013.
- [3] 马君潞, 范小云, 曹元涛. 中国银行间市场双边传染的风险估测及其系统性特征分析 [J]. *经济研究*, 2007, 01): 68-78+142.
- [4] DRAKOS A A, KOURETAS G P. Bank ownership, financial segments and the measurement of systemic risk: An application of CoVaR [J]. *International Review of Economics & Finance*, 2015, 40(127-40).
- [5] 包全永. 银行系统性风险的传染模型研究 [J]. *金融研究*, 2005, 08): 72-84.
- [6] 张晓朴. 系统性金融风险研究:演进、成因与监管 [J]. *国际金融研究*, 2010, 07): 58-67.
- [7] KANE E J. Redefining and Containing Systemic Risk [J]. *Atlantic Economic Journal*, 2010, 38(3):
- [8] 徐廷玮. 独立董事网络、企业非效率投资与非系统性风险 [D]; 浙江工商大学, 2015.
- [9] BERNAL O, GNABO J-Y, GUILMIN G. Assessing the contribution of banks, insurance and other financial services to systemic risk [J]. *Journal of Banking and Finance*, 2014, 47(
- [10] ELYASIANI E, JIA J. Relative performance and systemic risk contributions of small and large banks during the financial crisis [J]. *Quarterly Review of Economics and Finance*, 2019, 74(
- [11] BALUCH F, MUTENGA S, PARSONS C. Insurance, Systemic Risk and the Financial Crisis [J]. *The Geneva Papers on Risk and Insurance - Issues and Practice*, 2011, 36(1):
- [12] 孙培源, 施东晖. 基于 CAPM 的中国股市羊群行为研究——兼与宋军、吴冲锋先生商榷 [J]. *经济研究*, 2002, 02): 64-70+94.
- [13] STOIMENOV P. Philippe Jorion, Value at Risk, 3rd Ed: The New Benchmark for Managing Financial Risk [J]. *Statistical Papers*, 2011, 52(3):
- [14] 江婕, 王正位. 系统性市场风险度量指标的测算与评价 [J]. *中山大学学报(社会科学版)*, 2015, 55(06): 187-95.
- [15] ARTZNER P, DELBAEN F, EBER J, et al. Coherent measures of risk, mathematical Finance 9 (3): 203-228 [M]. Blackwell Publishers Inc. 1999.
- [16] BATTAGLIA F, GALLO A. Securitization and systemic risk: An empirical investigation on Italian banks over the financial crisis [J]. *International Review of Financial Analysis*, 2013, 30(
- [17] 高国华, 潘英丽. 银行系统性风险度量——基于动态 CoVaR 方法的分析 [J]. *上海交通大学学报*, 2011, 45(12): 1753-9.
- [18] BOUCHER C M, KOUONTCHOU P S, MAILLET B B, et al. The Co-CoVaR and some other fair systemic risk measures with model risk corrections [J]. Available at:

- www.org/papers/selectedworks/BoucherKouontchouMailletScaillet_2013_fullpaper pdf, 2013,
- [19] ENGLE R F, SIRIWARDANE E N. Structural GARCH: the volatility-leverage connection [J]. *The Review of Financial Studies*, 2018, 31(2): 449-92.
- [20] ENGLE R, JONDEAU E, ROCKINGER M. Systemic Risk in Europe* [J]. *Review of Finance*, 2015, 19(1): 145-90.
- [21] 陈奕延, 李晔. 投资组合对非系统性风险的发散作用——基于单调非增次模集函数的证明 [J]. 首都师范大学学报(自然科学版), 2018, 39(06): 1-4.
- [22] KIM J-B, LI Y, ZHANG L. Corporate tax avoidance and stock price crash risk: Firm-level analysis [J]. *Journal of Financial Economics*, 2011, 100(3): 639-62.
- [23] KIM Y, LI H, LI S. Corporate social responsibility and stock price crash risk [J]. *Journal of Banking & Finance*, 2014, 43(1-13).
- [24] KIM J B, ZHANG L. Accounting conservatism and stock price crash risk: Firm-level evidence [J]. *Contemporary Accounting Research*, 2016, 33(1): 412-41.
- [25] KIM J B, WANG Z, ZHANG L. CEO overconfidence and stock price crash risk [J]. *Contemporary Accounting Research*, 2016, 33(4): 1720-49.
- [26] ROBIN A J, ZHANG H. Do industry-specialist auditors influence stock price crash risk? [J]. *Auditing: A Journal of Practice & Theory*, 2015, 34(3): 47-79.
- [27] ANDREOU P C, LOUCA C, PETROU A P. CEO age and stock price crash risk [J]. *Review of Finance*, 2017, 21(3): 1287-325.
- [28] CALLEN J L, FANG X. Religion and stock price crash risk [J]. *Journal of Financial and Quantitative Analysis*, 2015, 50(1-2): 169-95.
- [29] YUAN R, SUN J, CAO F. Directors' and officers' liability insurance and stock price crash risk [J]. *Journal of Corporate Finance*, 2016, 37(173-92).
- [30] AN H, ZHANG T. Stock price synchronicity, crash risk, and institutional investors [J]. *Journal of Corporate Finance*, 2013, 21(1-15).
- [31] CHANG X, CHEN Y, ZOLOTOY L. Stock liquidity and stock price crash risk [J]. *Journal of financial and quantitative analysis*, 2017, 52(4): 1605-37.
- [32] XU N, LI X, YUAN Q, et al. Excess perks and stock price crash risk: Evidence from China [J]. *Journal of Corporate Finance*, 2014, 25(419-34).
- [33] CALLEN J L, FANG X. Short interest and stock price crash risk [J]. *Journal of Banking & Finance*, 2015, 60(181-94).
- [34] XU N, JIANG X, CHAN K C, et al. Analyst coverage, optimism, and stock price crash risk: Evidence from China [J]. *Pacific-Basin Finance Journal*, 2013, 25(217-39).
- [35] LI X, WANG S S, WANG X. Trust and stock price crash risk: Evidence from China [J]. *Journal of Banking & Finance*, 2017, 76(74-91).
- [36] SU L. Multivariate local polynomial regression with application to Shenzhen component index [J]. *Discrete Dynamics in Nature and Society*, 2011, 2011(
- [37] MA X, LIU Z, WANG Y. Application of a novel nonlinear multivariate grey Bernoulli model to predict the tourist income of China [J]. *Journal of Computational and Applied Mathematics*, 2019, 347(84-94).
- [38] JADITZ T, RIDDICK L A, SAYERS C L. Multivariate nonlinear forecasting using financial information to forecast the real sector [J]. *Macroeconomic Dynamics*, 1998,

- 2(3): 369-82.
- [39] CHU J, NADARAJAH S. A statistical analysis of UK financial networks [J]. *Physica A: Statistical Mechanics and its Applications*, 2017, 471(
- [40] BOGINSKI V, BUTENKO S, PARDALOS P M. Mining market data: A network approach [J]. *Computers and Operations Research*, 2005, 33(11):
- [41] MANTEGNA R N. Hierarchical structure in financial markets [J]. *The European Physical Journal B - Condensed Matter and Complex Systems*, 1999, 11(1):
- [42] BOGINSKI V, BUTENKO S, PARDALOS P M. Statistical analysis of financial networks [J]. *Computational Statistics and Data Analysis*, 2004, 48(2):
- [43] HUANG W-Q, ZHUANG X-T, YAO S. A network analysis of the Chinese stock market [J]. *Physica A: Statistical Mechanics and its Applications*, 2009, 388(14):
- [44] MA J, YANG J, ZHANG X, et al. Analysis of Chinese Stock Market from a Complex Network Perspective: Better to Invest in the Central; proceedings of the 第三十四届中国控制会议, 中国浙江杭州, F, 2015 [C].
- [45] J-P O, A C, K K, et al. Dynamics of market correlations: taxonomy and portfolio analysis [J]. *Physical review E, Statistical, nonlinear, and soft matter physics*, 2003, 68(5 Pt 2):
- [46] DIMITRIOS K, VASILEIOS O. A Network Analysis of the Greek Stock Market [J]. *Procedia Economics and Finance*, 2015, 33(
- [47] 马骏. 中国股票市场复杂网络性质的动态研究 [D]; 上海交通大学, 2015.
- [48] TSE C K, LIU J, LAU F C M. A network perspective of the stock market [J]. *Journal of Empirical Finance*, 2010, 17(4): 659-67.
- [49] LI H, AN H, FANG W, et al. Global energy investment structure from the energy stock market perspective based on a Heterogeneous Complex Network Model [J]. *Applied Energy*, 2016,
- [50] BILLIO M, CASARIN R, COSTOLA M, et al. An entropy-based early warning indicator for systemic risk [J]. *Journal of International Financial Markets, Institutions & Money*, 2016, 45(
- [51] NGUYEN-HUY T, DEO R C, MUSHTAQ S, et al. Copula statistical models for analyzing stochastic dependencies of systemic drought risk and potential adaptation strategies [J]. 2019, 33(3):
- [52] 刘傲琼, 刘新宇. 传染性渠道、相关性渠道与银行系统风险——基于网络结构视角 [J]. 武汉金融, 2018, 06): 60-6.
- [53] NAMAKI A, SHIRAZI A H, RAEI R, et al. Network analysis of a financial market based on genuine correlation and threshold method [J]. *Physica A: Statistical Mechanics and its Applications*, 2011, 390(21-22):
- [54] TOMOMICHI N, TOSHIHIRO T, MICHAEL S. Constructing networks from a dynamical system perspective for multivariate nonlinear time series [J]. *Physical review E*, 2016, 93(3):
- [55] 杨子晖, 陈里璇, 陈雨恬. 经济政策不确定性与系统性金融风险的跨市场传染——基于非线性网络关联的研究 [J]. 经济研究, 2020, 55(01): 65-81.
- [56] 姚小强, 侯志森. 基于树结构长短期记忆神经网络的金融时间序列预测 [J]. 计算机应用, 2018, 38(11): 3336-41.