**上 海 对 外 经 贸 大 学**

**数量经济学研究方法课程论文**

题 目 ： **基于隐马尔科夫链的沪深300研究**

姓 名 ： 潘科良

学 号 ：19331008

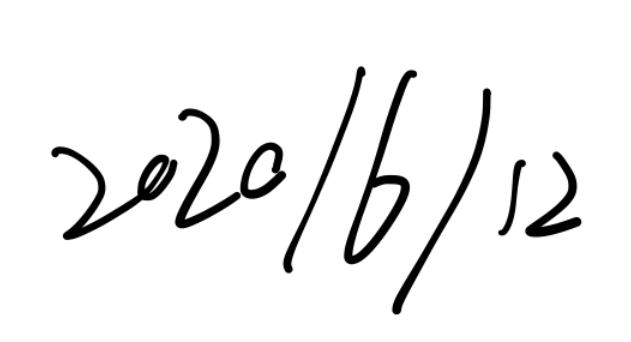
专 业 ：统计学（经济）

院 系 ：统计与信息学院

2020年 6 月 23 日

考试诚信承诺书

本人郑重承诺：在2019-2020学年第二学期课程考试中，严格遵守学校《学生考试规定》，独立完成考试（论文、报告、作业等），不违纪，不作弊，如有违反，按学校规定接受处理。

学生签名： 日期：

**基于隐马尔科夫链的沪深300研究**

**1.摘要**

新冠疫情的影响导致全球股市上蹿下跳，而股票的大幅波动在历史上不是个例，怎么预测股票，从而保存自己的资产甚至从中获得收益就是一件很严肃的事情。但股票价格具有不确定性,如果投资者能够事先预测价格走势,就能够规避价格波动风险,减少投资损失,甚至获取超额收益。随着大数据的热潮,越来越多的统计学习模式应用到股票价格预测中,隐马尔科夫模型就是其中一个。隐马尔科夫模型在马尔科夫链的基础上发展而来,用来研究一组隐藏状态。该模型是一个双随机过程,由两部分构成:马尔科夫链和一般随机过程,分别用来描述状态之间的转移关系和状态、观测值之间的关系。本文基于隐马尔科夫模型对股票价格预测、股票市场状态等问题进行了实证研究,以股票价格指数——沪深300为研究对象,讨论了隐马尔科夫模型的相关理论在预测股票市场状态方面的可行性,从而构建出适用于我国国情的股价预测模型。本文实证过程主要包括数据选取、隐状态数目确定、预测等步骤。本文通过所运用的隐马尔科夫模型构建出一个在不同隐藏状态下买卖的策略，并通过马尔科夫链模拟进行未来状态的预测。实证结果表明,本文模型能够预测股票市场状态，股票市场处于某一状态(熊市或者牛市)会持续一段时间,从一个状态转移到另一个状态较为困难，本文的模型和策略对实际投资有一定的指导意义。

关键词： 沪深300指数，隐马尔科夫链，买卖策略

**2.研究背景**

1960年，随着世界上第一只量化投资基金的产生，量化投资方式开始进入投资领域。2004年，涉及量化领域的公募基金在我国首次出现2010年以后，随着股指期货的上市，量化投资开始在我国快速发展。近几年，大数据以及统计学习模型的大量应用再一次掀起了量化投资领域相关研究的热潮，我国的量化投资也在逐渐向成熟化发展。量化投资通常指釆用数学模型并利用计算机技术去实现收益的投资方式，该投资方式本质上属于主动性投资，建立在非完全有效市场的假设上。

股票价格指数是描述市场价格水平及其变化的指标，该指数通过选取具有代表性的一组股票，再由加权平均和一定计算得到。股票价格涨跌不定，投资者面临价格风险，因此理性投资者会在投资前了解股价相关信息。但是，对投资者来说，了解个股涨跌趋势较为容易，但了解多种股票甚至整个市场走向却较困难。因此，一些金融机构为满足这些需求，充分利用市场上的公开信息和专业知识及能力，通过一定的规则编制股票价格指数，以帮助投资者更清晰地了解股票市场股票价格指数一般综合各个行业具有代表性的股票，帮助投资者以此为基准评价投资效果，也有助于投资者对市场做出预测判断。另外，股票市场在一定程度上反映了当前经济形势，因而政府、企业等机构将股票价格指数作为参考指标，并以此判断当前经济形势。优质的股票价格指数能够准确地反映出市场信息，在投资界应用广泛。以下给出几种著名的股票价格指数。

1.道琼斯股票价格平均指数

道琼斯股票价格平均指数又称道式指数，该指数包含运输业、工业、公用事业三个板块的65家公司的股票价格。道琼斯股票价格指数以1928年10月1日为基期，用当前时刻65家公司股票价格的算数平均数与上文所述基期的比值表示，并在纽交所的交易时间内每三十分钟公布一次。

2.标准普尔指数

标准普尔指数与道琼斯股票价格指数一样，在全球有很深远的影响，该指数由美国标准普尔公司于1923年开始编制。目前所使用的标准，普尔指数在1957年的版本上扩展而成，并通过500种股票加权平均计算得到。

3.恒生指数

恒生指数由香港恒生银行编制，是香港股票市场上历史最久且影响最大的股票价格指数。该指数于1969年11月开始发布，并以1964年7月31日为基期，选取了 33家经济实力雄厚且具有代表性的香港上市公司股票作为指数的成分股，包含金融业、公共事业、地产业和其他工商业四大类股票。

4.纽约证券交易所股票价格指数

纽约证券交易所股票价格指数由纽交所编制而成，以1965年12月31日为基期，选取1570种股票作为成分股，包含工业、运输业、金融业等行业，可以全面地反映股票市场变动情况。

5.日经指数

日经指数由日本经济新闻社编制，包含金融保险业、不动产业、建筑业、矿业等多种行业，综合反映出日本股票市场的变动情况，在日本有着深远的影响。

我国金融机构编制了适合我国国情的价格指数，帮助投资者了解我国股票市场，进行投资决策。其中上证股票指数和深证综合股票指数历史最为悠久。上证指数由上交所编制，并于1990年12月19日正式发布。该指数包含所有在上交所挂牌上市的股票。深证综合股票指数由深交所编制，并以1991年4月3日为基期，包含所有在深交所挂牌上市的股票。

本文选取的股票价格指数为沪深300,该指数横跨沪、深两个市场，包含300只质地优良的成分股。选取沪深300而非其它指数进行实证研究的原因为：（1)上证指数和深证指数仅仅代表沪、深两个市场各自的行情走势，没有反映出两个市场整体走势，而沪深300的成分股来自两个市场，可以反映出沪深两个市场的整体走势；（2)与其他跨市场股票价格指数相比，沪深300由沪、深两大交易所编制，这两大交易所编制指数经验丰富，且借鉴了国际编制技术，使得编制出的指数更具有市场代表性。

随着大数据时代的到来，越来越多的统计学习模型被应用于金融领域。与传统的人工方式相比，统计学习模型具有极大的优势。统计学习模型在金融领域中的一个重要应用就是股票市场预测。在进行股票交易时，理性投资者通常会对历史数据等做大量分析，然后进行股票买卖。然而，这种人工分析数据的方式耗时费力，且很容易产生很多错误，从而加大投资风险，导致投资损失。统计学习模型能够对大量数据做深入且精确地分析，并在较短时间内给出更准确的预测结果。与人工分析相比，统计学习模型提高了投资效率和准确率。统计学习模型不仅可以处理人工无法处理的数据量级，还能通过计算机技术浏览新闻、社交网站等，收集并处理与投资有关的更多信息，在增加投资者信息来源的同时，节省投资者的时间。

近几年，支持向量机、神经网络、遗传算法等统计模式识别算法被广泛地应用于股票市场预测。统计模式识别算法在股票价格预测中的应用成为当前金融界的研究热点，很多学术研究人员和金融从业者都对此进行过深入的研究与探讨。大量结论表明，统计学习模型对股价的预测是有效的，人工智能与金融领域结合是未来的发展趋势。

**3.文献综述**

隐马尔科夫模型（Hidden Maikov Model，HMM)的发展源头可追溯到20世纪60年代后半期，Baum等人在一系列的统计学论文中给出该模型的原始雏形。1966年，Baum[1]发表了有关隐马尔科夫模型最初原型的相关论文，给出了关于有限状态马尔科夫链的概率函数的统计推断。随后，Baum[2][3]在1970年和1972年分别给出了一种关于马尔科夫链概率函数的最大化技术和马尔科夫过程概率函数统计估计中的相关不等式和最大化技术，该技术是对1966年论文的进一步拓展，也为之后隐马尔科夫模型的相关工作奠定了基础。1973年，Ryan和Nudd总结给出了 Viterbi算法相关内容以及应用领域，并且对箅法进行拓展。1983年，Levinsom等人给出了隐马尔科夫链概率函数理论在自动语音识别中的应用，将理论与实际问题结合，给出适用于孤立词识别的模型J986年,Rabiner和Juang[4]在前人的基础上总结给出了 HMM的存关内容。2010年，Krishnalal，Rengarajan和Srinivasagaii将隐马尔科夫模型与支持向量机结合，将其运用到文本挖掘、新闻分类中。

在金融领域中，HMM也有广泛的应用。2005年，Hassan和Nath首先将HMM应用到股票价格预测中，给出了一种预测股票价格的新方法。该方法以股票价格的开盘价、收盘价、最高价和最低价作为模型输入，通过参数估计、状态解码等步骤预测股票价格。2008年Srivastava等人应用到信用欺诈检测中，实证结果表明，利用HMM进行信息欺诈检测是有效的.2009年，HassanM将HMM与模糊模型相结合，并把该模型运用到股票价格预测屮.该模型通过使用HMM进行数据模式识別，然后利用模糊逻辑获得预测，并能够对来自不同行业的股票市场进行检验&实证结果表明，改进后的模型预测准确性有了明显的提升。

隐马尔科夫链模型作为统计模型的一种，最初主要应用于语音识别，后来逐渐应用到金融领域，本文将着重探讨HMM在股票价格预测中的应用。

**4.模型介绍**

**4.1.马尔科夫链**

马尔可夫链是一组具有马尔可夫性质的离散随机变量的集合。具体地，对概率空间() 内以一维可数集为指数集（index set）的随机变量集合，若随机变量的取值都在可数集内：，且随机变量的条件概率满足如下关系：



则X被称为马尔可夫链，可数集s被称为状态空间（state space），马尔可夫链在状态空间内的取值称为状态。这里定义的马尔可夫链是离散时间马尔可夫链（Discrete-Time MC, DTMC），其具有连续指数集的情形虽然被称为连续时间马尔可夫链（Continuous-Time MC, CTMC），但在本质上是马尔可夫过程（Markov process）。常见地，马尔可夫链的指数集被称为“步”或“时间步（time-step）”。

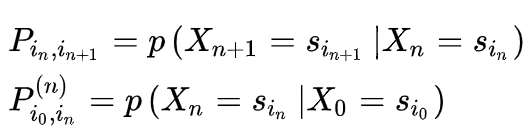
上式在定义马尔可夫链的同时定义了马尔可夫性质，该性质也被称为“无记忆性（memorylessness）”，即t+1步的随机变量在给定第t步随机变量后与其余的随机变量条件独立（conditionally independent）：。在此基础上，马尔可夫链具有强马尔可夫性（strong Markov property），即对任意的停时（stopping time），马尔可夫链在停时前后的状态相互独立。

n-阶马尔可夫链拥有n阶的记忆性，可视为马尔可夫链的推广。类比马尔可夫链的定义，n-阶马尔可夫链满足如下条件：



按上式，传统的马尔可夫链可以认为是1-阶马尔可夫链。由马尔可夫性质可得，将多个马尔可夫链作为组元可以得到n-阶的马尔可夫链。

马尔可夫链中随机变量间的条件概率可定义为如下形式的（单步）转移概率和n-步转移概率：



**4.2.隐马尔科夫模型**

HMM是在马尔科夫链的基础上发展而来的，用来研究一组隐藏（即不可见)状态。该模型由一个隐藏的马尔科夫链生成不可见的状态序列，再由该不可见的状态序列生成一组可见序列（上述两个序列生成过程均是随机的）。其中隐藏的马尔科夫链生成的序列称为状态序列，描述状态之间的关系，用状态转移概率分布来表示；状态序列生成的可见的序列称为观测序列，描述状态与观测值之间的关系，用观测概率分布表示。因此，HMM由状态转移概率分布、观测概率分布以及初始概率分布决定该模型具有以下特点：

1. HMM是一个双随机过程，由两部分构成：马尔科夫链和一般随机过程，分别用来描述状态之间的转移关系和状态、观测值之间的关系；

2. HMM的状态是隐藏的，也就是说状态是不可见的，它通过一组观测序列表现出来；

3. HMM的观测序列和状态序列并非一一对应，而是通过一组概率分布相联系.

设Z为所有可能状态的集合，N为上述集合包含的状态数；G为所有可能观测的集合，M为相应可能的观测数，从而有.

设S为长度为T的状态序列，O为相应的观测序列，则有.

设A为状态转移概率矩阵，,其中，

表示t时刻处于状态下，在t+1时刻状态转移到的概率。

设B是观测概率矩阵，

其中，

表示在t时刻处于状态条件下生成观测值的概率。

设是初始状态概率分布，

其中

表示初始时刻t=1时处于状态（即初始状态）的概率。

HMM由状态转移矩阵A、观测概率矩阵B和初始概率分布为HMM的三要素。其中初始概率分布和状态转移概率矩阵决定马尔科夫链以及隐藏状态的生成，观测概率矩阵决定状态如何生成观测值，即观测序列的产生。

一般的，可以用λ=(A,B,π)三元组来简洁的表示一个隐马尔可夫模型。隐马尔可夫模型实际上是标准马尔可夫模型的扩展，添加了可观测状态集合和这些状态与隐含状态之间的概率关系。

由上文可知，HMM有以下两个假设;

a.其次马尔科夫假设

其次马尔科夫假设是指未来时刻t时的状态仅依赖于前一时刻的状态，与其他时刻的状态以及该模型的观测无关，与所处时刻也无关，即有下式成立

b.观测独立性假设

观测独立性假设是指某一时刻的观测状态仅依赖于该时刻的马尔科夫链的状态，与其他观测和状态无关，即有下式成立

本文选取 HMM 预测股票价格指数，对于隐马尔科夫过程的“无后效性”以及全局预测的特点，本文认为可以结合其他算法或者对模型进行优化解决。由于股票市场状态特征比较明显，可以根据每个状态的均值判断出状态数字的实际含义，所以状态预测的不显著性对本文模型预测的影响可以忽略不计。另外，对于一阶隐马尔科夫模型可能产生的拟合度不够的情况，可以考虑更高阶的模型或者与其他算法结合进行模型优化。最后关于HMM黑箱、不直观的缺点，由于这是机器学习算法通常具有的缺陷，所以目前不予考虑，本文主要侧重考虑模型的适用性。

综上所述，本文认为利用HMM对股票价格指数进行预测是合理且适用的。

**5.实证研究**

**5.1．初步建模**

该实验的数据为沪深300指数2010年初到2019年年底10年的交易日数据，数据来源于joinquant平台，数据的内容包括开盘价，收盘价、日高值、日低值以及交易量。隐藏状态数目各个学者都有过不小的争议，比如隐藏状态为3,4或者6，但基于本文选取的数据，该数据显示15年从2000点一直暴涨到5000多点，如果选择3或4的话，模型不能很好的解释除了15年之外其他年份明显的涨或跌，因此隐藏状态更好的数目应该设置为6，即暴涨，暴跌，涨，跌，震荡涨，震荡跌这6个状态。

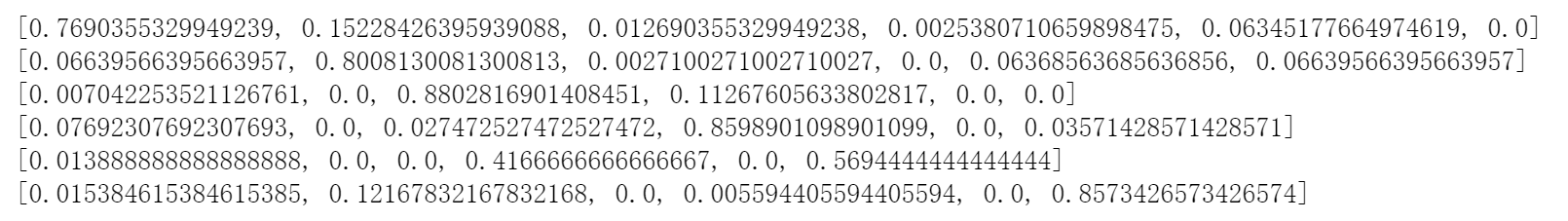
模型选取的特征状态选3个，分别是日高值与日低值的对数差，每五日的指数对数收益差，每五日的沪深指数成交量的对数差。将状态代入后，可得结果如图。



其中，隐藏状态序列在算出来之后，将状态代入对应的交易日当中，可以得到如上的交易状态散点图。图中可以清晰的看出暴涨，暴跌状态以及其他状态对应的颜色。我们可以再图中看到状态3：红色，暴涨；状态2，绿色，暴跌；状态5，棕色，涨；状态0，蓝色，跌，以及其他两个震荡的状态。

**5.2.构建策略**

隐藏状态的转移概率矩阵可以通过已获得的隐藏状态序列计算得到，如下图所示。



其中，转移概率矩阵的每一行对应着隐藏状态的每一个状态，第一行表示状态0，第二行表示状态1，以此类推。

我们从中可以发现，状态0,1,2,3和5在下一步维持自身状态的概率比较大，而状态4转移到3或5的概率比较大。

可以构建一个简单但有效的策略：在获得当日的状态之后，第二日立即买入。通过这种方式，可以查看每一个状态在第二天买入之后具体的盈利，如下图所示。



其中，我们可以看到，暴涨的状态第二天买入也是暴涨的，收益率明显比其他情况要高，而暴跌的状态第二天买入的收益率反而不是最低的，是因为暴跌转移到其他状态的概率也比较高，因此他的收益并不是最差的。其他的状态也类似。

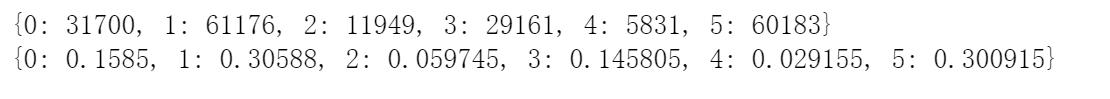
通过对每个状态第二天买入的收益率分析，我们可以进一步拓展我们的策略，即在涨的状态下第二天买入，在跌的状态下第二天卖出，我们可以构建出如下的收益图。



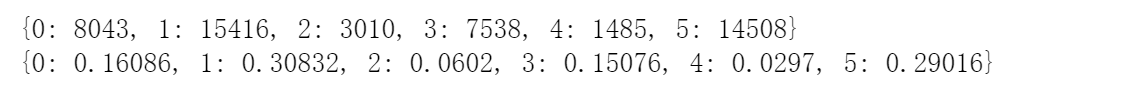
如图所示，只要始终遵循着简单的涨买，跌卖原则，收益率在10年的期限内便能涨到900%，中间虽然会有平坦和回撤，但是在总体上却是稳步向上的。

**5.3.随机模拟**

本文用马尔科夫链模拟对隐藏状态进行模拟，并以此预测了之后20万个状态，如下图所示。

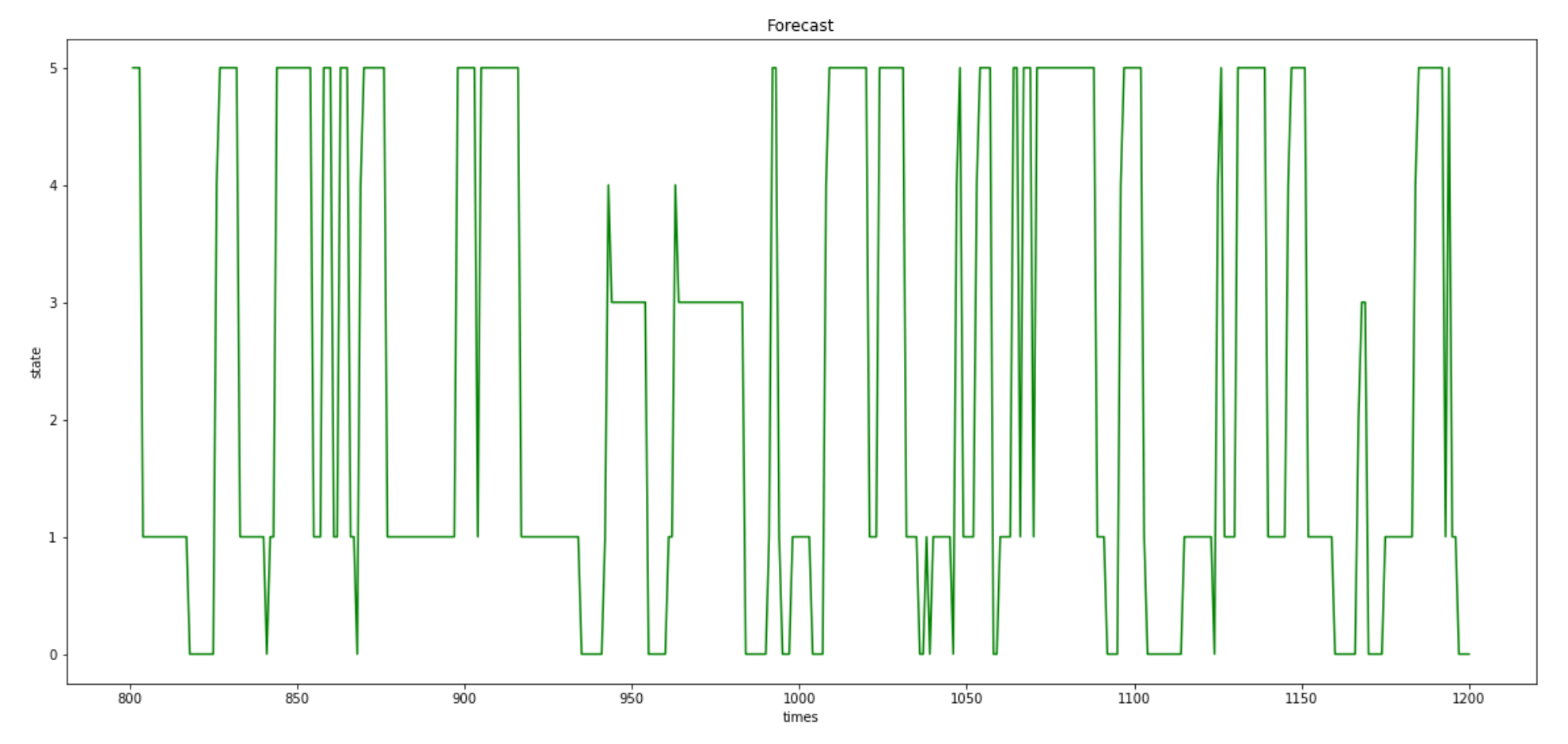


在去掉前15万个状态后的频率如下。



其中，第一行是频数，第二行是频率。可以看到，无论是20万数据状态的频率还是只获取最后5万数据的频率都是近似的。

将预测的状态中第801到1200的状态画出如下图。



其中，纵轴代表状态，横轴代表时间。

**6. 总结及展望**

虽然根据隐马尔科夫模型所做的策略效果不错，但是还是有很多地方需要改进。

1.HMM对于中长期的预测准确性较差，因而不适宜中长期的预测；

2.HMM预测从全局出发，不适合局部预测；

3.HMM算法比较复杂，具有不直观、黑箱等特点；

4.马尔科夫过程显著特征是价格的“无后效性”，即给定时初始状态，t时的状态只依赖于初始状态，而与之前的状态无关，而该特性适合当前的价格或者某状态己经完全反应了市场的所有信息，但是真实的市场情况通常不是充分且有效的。

5.最后在算整个收益的情况的时候，是先将整个模型分为买入和卖出两个部分分别计算之后，再做收益率的求和变化。但实际交易过程中肯定是不可能这么简单的分的。

6.本文假设了无限本金的模型。比如说，这里有可能出现需要做买入某个指数，和卖出某个指数的操作，但实际做期货交易的时候，都是需要提交一定的佣金的。本文并没有讨论这个问题。

**7.思政元素发掘**

1.选这门课以及做论文需要有一种经世济民和学无止境的态度，否则的话无法有足够的内驱力让你完成这些困难的任务。

2.在理解问题和获得答案的时候，需要拥有古代工匠精神，在探索的时候需要专注和创新，这样才能不放过任何一个机会。

3.在进行实验的时候，需要实事求是，调查研究。对问题的假设和结论需要考证，否则的话会得到不严谨的结果。

4.分析问题时需要有全面、历史的看问题的方法，否则，在研究和分析时会总是漏这漏那，逻辑不清晰，思维不严谨。

5.而在整个课程当中，需要拥有长征精神，不畏艰难，迎难而上，不惧怕挑战。这样方可达到学习的目的。

**8.参考文献**

[1]Leonard E. Baum, Ted Petrie. Statistical Inference for Probabilistic Functions of Finite State Markov Chains. 1966, 37(6):1554-1563.

[2] Leonard E. Baum, Ted Petrie, George Soules, et al. A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains. 1970, 41(1):164-171.

[3] Leonard E. Baum. An inequality and associated maximization technique in statistical estimation of probabilistic functions of a Markov process[j]. Inequalities, 1972, 3:1-8

[4]Schuster-Böckler Benjamin,Bateman Alex. An introduction to hidden Markov models.[J]. Current protocols in bioinformatics,2007,Appendix 3.

**9.附件**

# “随机过程”课程论文评分表

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **序号** | **学号** | **姓名** | **概念方法与结论** | **随机模拟** | **思政教育元素** | **表达与规范** | **总评** |
|  |  |  | 20% | 30% | 20% | 30% | 100% |
| 1 | 19302002 | 胡佳垚 | 15 | 20 | 15 | 20 | 70 |
| 2 | 19331001 | 李会会 | 19 | 27 | 18 | 28 | 92 |
| 3 | 19331002 | 刘娟 | 18 | 25 | 18 | 28 | 89 |
| 4 | 19331006 | 何炜成 | 17 | 26 | 18 | 28 | 89 |
| 5 | 19331007 | 李健 | 19 | 29 | 19 | 29 | 96 |
| 6 | 19331008 | 潘科良 | 20 | 30 | 20 | 30 | 100 |
| 7 | 19331009 | 周驰 | 18 | 27 | 18 | 28 | 91 |
| 8 | 19332002 | 李圣洁 | 18 | 28 | 18 | 28 | 92 |
| 9 | 19332011 | 薛涛 | 19 | 25 | 18 | 28 | 90 |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |

评分人：潘科良

评分时间：2020/6/18

**随机模拟程序**

def transition(k):

a = 0

b = 0

c = 0

d = 0

e = 0

f = 0

for i in range(len(list1)-1):

if list1[i] == k:

if list1[i+1] == None:

continue

elif list1[i+1] == 0:

a += 1

elif list1[i+1] == 1:

b+=1

elif list1[i+1] == 2:

c+=1

elif list1[i+1] == 3:

d+=1

elif list1[i+1] == 4:

e+=1

elif list1[i+1] == 5:

f+=1

return [a/(a+b+c+d+e+f), b/(a+b+c+d+e+f), c/(a+b+c+d+e+f), d/(a+b+c+d+e+f), e/(a+b+c+d+e+f), f/(a+b+c+d+e+f)]

for i in range(6):

print(transition(i))

mat = []

for i in range(6):

mat.append(transition(i))

print(mat)

i = 2

state = []

count = 200000

np.random.seed(666)

ran = np.random.random\_sample([count])

for t in range(count):

r = ran[t]

if r <= mat[i][0]:

i = 0

elif mat[i][0] <= r <= mat[i][0] + mat[i][1]:

i = 1

elif mat[i][0] + mat[i][1] <= r <= mat[i][0] + mat[i][1] + mat[i][2]:

i = 2

elif mat[i][0] + mat[i][1] + mat[i][2] <= r <= mat[i][0] + mat[i][1] + mat[i][2] + mat[i][3]:

i = 3

elif mat[i][0] + mat[i][1] + mat[i][2] + mat[i][3] <= r <= mat[i][0] + mat[i][1] + mat[i][2] + mat[i][3] + mat[i][4]:

i = 4

else:

i = 5

state.append(i)

countDict = dict()

proportitionDict = dict()

for i in range(6):

countDict[i] = state.count(i)

proportitionDict[i] = state.count(i)/len(state)

print(countDict)

print(proportitionDict)

Astate = state[150000:]

countDict1 = dict()

proportitionDict1 = dict()

for i in range(6):

countDict1[i] = Astate.count(i)

proportitionDict1[i] = Astate.count(i)/len(Astate)

print(countDict1)

print(proportitionDict1)

plt.figure(figsize=(20, 9))

plt.title('Forecast')

plt.plot(range(801,1201), state[800:1200], color='green')

plt.xlabel('times')

plt.ylabel('state')

plt.show()

**原始数据**

参考附件