上海交通大學

SHANGHAI JIAO TONG UNIVERSITY

学士学位论文

THESIS OF BACHELOR

论文题目：基于大数据的股票板块轮动量化模型研究

学生姓名：曹韶光

学生学号：5130309421

专　　业：计算机科学与技术

指导教师：邓倩妮

学院(系)：电子信息与电气工程学院

基于大数据的股票板块轮动量化模型研究

摘要

在股票市场中，由于受到经济周期、国家政策、投资者心理等因素的影响，市场热点在各行业板块间来回轮转，导致不同行业间的相对强弱走势存在着此消彼长的现象。如果投资者能够准确把握住其中的行业板块轮动的规律，预测出未来相对表现强势的行业，势必能获得超越市场的超额收益。

本文基于申万一级行业分类标准，将三千多只 A 股股票划分成二十八个行业，借助各行业的日收益率指数和一些表征当前市场状态的常用因子作为观测向量，构建隐马尔科夫模型，预测市场隐含状态序列并计算得到观测值序列在该参数模型下的相对应的似然值序列。标注出每天日涨幅排名较前的行业板块，将连续的行业标记作为行业轮动特征。通过匹配市场隐含状态、行业轮动特征以及似然值序列，找出与当天行为模式最接近的交易日，并以此预测下一阶段上涨概率较大，表现较为强势的行业。

结果显示，在二十八个一级行业中给出一个候选行业、两个候选行业、三个候选行业的预测胜率分别可以达到 33.3%、52.0%、64.9%，相比基础预测概率有了显著的提高，具有一定的应用价值与现实意义，也证实了行业板块间的轮动效应的存在性以及本文所用预测方法的有效性。

关键词：股票，行业板块轮动，隐马尔科夫模型，匹配

A RESEARCH ON QUANTITATIVE MODEL OF SECTOR ROTATION BASED ON BIG DATA

ABSTRACT

In the stock market, due to the impact of the economic cycle, the national policy and the psychology of investors, the hot spot of market is back and forth between the plates, which we named it the phenomenon of sector rotation. If the investors can accurately grasp the law of the sector rotation, and then predict the plate with highest rise, it is bound to be able to get excess earnings beyond the market.

The sample data of this research is based on the SWS industry classification standard, which is divided more than three thousand stocks into twenty-eight sectors. We use these data as observation to constructing hidden Markov model. And then, we can use the hidden Markov model, the parameter of which is calculated, to predicted the implied state sequence of the market and to calculate the likelihood sequence corresponded to the observed value series. By the way, we mark the plate with the largest daily increase and the continuous industry mark will be treated as a feature of sector rotation. By matching the market implied state, the feature of sector rotation and the likelihood value sequence, we can find a trading day which has the closest behavior patterns with the day will be predicted, and then use it to predict the plate with the largest daily increase.

The result shows that the win rate of the prediction is nearly twice of the random probability. The result also confirmed the existence of phenomenon of sector rotation and the effectiveness of the prediction method.

Key words: stock market, sector rotation, HMM, match

目 录

第一章 绪论

1.1 研究背景

量化投资在西方发达市场已经有三四十年的发展历史了，它兴起于上个世纪七十年代，在九十年代的美国市场大行其道。再到如今成为了美国证券市场上最主流的交易方式。因其具有快速高效、客观理性、准确及时等优点，得到了越来越多投资者的认同与支持，占据的市场份额也得以不断地增长。詹姆斯·西蒙斯便是量化投资领域的领军代表人物。他不仅是一名伟大的数学家，在 1976 年便摘得全美维布伦（Veblen）奖这一数学界的皇冠。他同时还曾是全球收入最高的对冲基金经理，在文艺复兴科技公司成立的大奖章基金在 1989-2009 年期间达到了 35%以上的平均年回报率，创造了一个又一个神话。他也因此被人们奉之为对冲基金之王、量化鼻祖，被人们捧上神坛。

与在美国的大红大紫不同，虽然国内投资者早就了解到了量化投资的这个股市神兵，但当时我国股票市场证券行业的发展时间较短，成长还很有限，金融产品十分匮乏，又加上有着诸多外界因素的限制，因此在最初一段时间并不具备量化投资生长的土壤。而随着我国金融市场的不断发展与完善，投资者越来越理性，股票交易市场也逐渐趋于有效，直到 2010 年 4 月，股指期货正式的出台，量化投资才真正的显示出来国内证券市场的发展潜力。

大家应该都听说过股神巴菲特，他本人最擅长的就是基本面投资和价值投资。他通过对相关企业的财务报表进行研究分析，从而从中挖掘出其内在价值，然后进行投资。相比之下量化投资则是利用计算机技术，借用数学统计学模型来预测股票市场的走势。在如今的互联网时代，随着计算能力不断增强，大数据技术一时风头无二，备受推崇。而同时，“阿尔法狗”与李世石的人机世纪之战也逐渐引爆了人工智能与深度学习的话题。借此风口，国内各大量化团队也都开始了对这类“黑科技”的探索。

1.2 研究内容与目的

1.2.1 行业板块轮动释义

板块轮动是指由于受到经济周期、国家政策、投资者心理等因素的影响，市场热点在各行业板块间来回轮转，导致不同行业间的相对强弱走势存在着此消彼长的现象[1]。在股市的任何时期，不管是牛市、熊市、震荡市，都会有某些行业板块的市场表现能够跑赢大盘指数，这就是市场的投资热点所在。然而没有只涨不跌的股票，随着时间的推移，那些因为成为市场投资热点而被股民疯狂涌入进而使得股价远超其市场价值的板块，最终会逐渐回归正常水平。而同时，由于不同行业之间存在着上下游关系以及内在的联系，从上一个热点板块中退出来的庞大的资金洪流也将涌进与当前热点相关的下一个板块，股市热点也就因此在不同行业板块之间来回轮转，形成了行业板块轮动现象。

1.2.2 本文研究主要内容与目的

行业板块轮动是我国股市的一个明显特征，也是股市投资者常用的一种选股投资手段。本文希望通过利用一些数学、统计学模型以及计算机技术方法来对股票行业板块轮动现象进行合理描述、量化分析，进而从多年历史数据中找到探索把握各行业轮动规律，对下一阶段市场表现较为强势，收益率较高的行业板块进行预测，以便进行更优的资产配置。

1.3 研究意义

虽然在我国目前的A股市场中，行业轮动现象相对明显，也有越来越多的投资者利用板块轮动的规律进行股票的选择与投资，可是目前仍是缺乏相对有效的方法来对板块轮动现象进行量化建模，从而把握其中的板块轮动规律。通过对国内行业轮动相关策略的研究，我们可以明显发现，在利用行业板块轮动规律进行投资的策略中，大多数都是采用基本面分析手段，依据经济周期理论，对当前热门行业的一些公司的基本面数据进行分析，再结合当前投资热点以及国家政策，进而对下一阶段市场表现较好的行业进行判断。而事实上，这些策因为需要人为的进行参与，因而带有了过多的主观性因素，这也导致了在很大程度上制约了它们在市场上的表现以及稳定性[1]。这些策略市场表现的好坏更多的取决于投资者个人的经验与以及对板块轮动的具体表现方式和持续时间的分析把握能力。

本文试图借助数学、统计学以及计算机科学知识，对股票行业板块轮动现象进行量化建模。如果我们能够从多年历史数据中准确把握住各行业的轮动规律，预测出未来相对表现强势的行业，进而在股票板块轮动开始阶段及早以较低价格进行资产配置，在本轮板块轮动结束前及时抽身变现，这样我们也势必能获得超越市场的超额收益。

1.4 论文组织结构

全文共分为以下六个很章节：

第一章，绪论部分，讲述了研究的社会背景，介绍了股票市场行业板块轮动现象以及主要的研究内容和目的，并且阐明了本课题研究的意义所在。

第二章，国内外研究现状综述，既阐述了国内外与板块轮动相关的学术研究贡献，还简要说明了国内大型券商研究报告中一些基于板块轮动的投资策略。总结了前人在行业板块轮动现象研究中做的一些工作与不足。

第三章，理论基础，从多个角度对板块轮动现象产生的原因进行了合理的解释。并对隐马尔科夫模型理论做了比较详细的讲解，为后文建立行业轮动预测模型打好了理论基础。

第四章，研究方法与思路，讲述了对股票市场行业板块轮动现象建模的基本思路，并主要介绍了对市场隐含状态、行业轮动特征以及似然值进行匹配的具体方法以及由于隐马尔可夫模型用来建模的样本数不宜过长而采用的滑动样本窗口法。最后提出对板块轮动现象进行量化的两个难点和做出的改进措施。

第五章，实证研究及结果分析，介绍了实际研究过程中的样本数据、具体步骤以及相应的模型参数选择情况，并对最终预测结果进行对比分析。

第六章，总结与展望，对本文研究内容做简短总结，并提出根据不足之处提出了几点改进方向。

第二章 研究现状综述

2.1 国外研究现状

在国外成熟市场，行业板块轮动多年来早已作为一种选股投资策略被广大投资经理们运用于投资实践中。Sam Stoval 在 1995 年任美国标普公司分析师时，通过对美国上个世纪七十年代到九十年代中间近三十年的经济周期以及行业情况的分析研究，他指出在经济由繁荣转为萧条的衰退时期，大多会出现市场资金短缺，导致股价下跌的现象。在这个阶段，医药卫生、食品饮料这些防守型的行业是较好的投资选择。而在经济活动收缩向下的萧条时期，政府部门则常常为了刺激经济而采取下调利率等诸多手段。因此在这个阶段，有着高资产负债率的行业与利率敏感型的行业往往会有不错的表现。在经济接近触底时，对家电、房地产等消费品行业的市场需求不断增加，这些板块也随之上涨。在经济由萧条转向繁荣的复苏时期，交通、餐饮、旅游等服务业会有超越大盘的表现。而在经济活动扩张向上的繁荣时期，各种化工、金属、原材料以及能源等行业板块会在前一个时期的下游行业的带动下逐步上涨。这样，股票市场各行业板块便会随着实体经济周期的变化，出现投资热点不断轮动的现象[2]。

虽然 Sam Stoval 并未在文章中提出投资时钟和板块轮动的说法，但为接下来的行业轮动现象的研究起到了重要的基础作用。美国知名的美林证券投资公司在 2004 年提出了投资时钟理论。通过对美国股市长达 30 多年的数据的研究分析，研究报告中认为经济周期可以大致分为四个时期，分别是复苏、过热、滞胀以及衰退。并且在经济周期不同的时期中，根据其各自特点，分别采取不同的资产配置策略，最终获得了超越市场表现的超额收益[3]。

Paolo Sassetti 和 Massimiliano Tani 在 2006 年对多个基金基于板块轮动策略的市场表现情况的进行来详细地分析研究，最终通过实验结果数据证明了板块轮动规律的存在性以及板块轮动策略的有效性 [4]。可是虽然他们通过实验得出超过银行基础理论的收益率，但因为在试验中未消除标注普尔 500 指数的影响，从而众多研究者对此提出了质疑。Jeffrey Stangl 和 BenJacobsen 在 2009 的论文中则对比了板块轮动策略和股票市场中其他投资组合，较好的避开了 Paolo Sassetti 和 Massimiliano Tani 实验中的系统性的影响，最终也得出了板块轮动策略能取得优于市场大盘表现的收益率[5]。

2.2 国内研究现状

2.2.1 国内行业轮动相关学术文献综述

目前来说，国内对于股票市场行业板块轮动现象的研究主要在两个方面，或是对行业轮动现象的解析，或是对行业轮动现象的识别。除此之外还有较少的一些研究则是探究股票行业板块轮动规律在投资策略中的应用，以及对应的收益效果。

何诚颖在 2001 年的对中国股市 “板块现象” 的分析一文中认为，股票轮动现象是在某一时段中，和一些特别事件相关的这些股票的涨跌有一定的一致性，因而出现的同涨同跌或者轮动上涨（下跌）的一种现象。文章中还通过相对收益率提出了能够代表当前板块市场表现的量化指标 [6]。

彭艳、张维在 2003 年的我国股票市场的分板块投资策略及其应用的研究中，通过对各类股票组合指数和基准指数间的超额收益增长趋势、累计超额收益率以及风险调整后的收益指标进行对比分析，对我国 A 股市场板块轮动现象的特点进行了展开说明，对板块轮动投

资策略的有效性和可行性进行了论证[7]。

杜伟锦、何桃富在 2005 发表的我国证券市场的板块联动效应及模糊聚类分析一文中，通过对多个行业板块进行模糊聚类，最终得出通信类、汽车类、能源类和钢铁类各板块之间关联性较低，相对独立；而地产类、商业类、公用类、工业类以及综合类行业板块之间走势相关性相对较强的结论[8]。

伍军在 2010 年的国际对冲基金的行业轮动投资一文中，对中美两国的股票行业轮动现象通过行业轮动投资理论 LIP 模型进行对比。最终结果显示，这两个股票市场因为相互之间的机制、成熟度已经行业自身特点的差异，所以两者的行业板块轮动特征有着很大的不同。本文作者还指出利用盈利性、流动性以及通货膨胀三个因子来划分经济周期，这与实体经济周期不同，对模型最终结果产生的影响较大[9]。

张福芬在 2010 年的中国股票市场板块轮动的机理研究一文中，通过对 A 股 2008 年到 2009 年期间的行业板块数据进行研究分析，对行业轮动现象出现的原因做出了合理解释，还提出了能够充分利用板块轮动规律获取超额收益的投资组合[10]。

苏民、逯宇铎在 2011 年的对我国股市行业轮动现象及相应策略的探讨一文中，从经济周期和货币周期角度出发行业板块轮动现象出现的原因。作者通过研究认为诸多行业在经济周期的不同时段市场表现相异是不同行业对市场经济状态改变、利率变化的敏感度不同所造成的。虽然阶段性的数据表明了行业轮动的确存在，但是本文认为行业轮动现象是特殊历史时段的产物。如果从长期来看，行业的兴衰替代是必然现象[11]。

何韬在 2011 年的证券市场行业轮动问题研究一文中，将股票市场按走势分为调整初期、中继、末期三个阶段，将申万 23 个一级行业分为强周期集群、中周期集群、轻工业集群以及消费类集群四个集群，进而探究四个集群的行业轮动规律[12]。

丁军广在 2011 年的我国 A 股市场行业轮动规律研究一文中，通过对 2000 年到 2010 年的股票行业数据进行聚类分析和相关性分析，进而探索股票市场行业板块轮动的现象。该文章最终的实验结果也同样证实了 A 股市场行业板块轮动现象的存在性。研究还显示轮动的效果与市场的状态趋势相关，牛市中的板块轮动现象要明显于熊市[13]。

2.2.2 国内行业轮动相关投资策略简述

除了上述与行业板块轮动相关的学术研究贡献，国内各种各样金融机构也利用股票板块轮动现象做了一些相关的量化投资策略，下面是一些关于板块轮动的投资策略的研究报告。

兴业证券在《基于不同市场情境下的行业轮动策略》的研究报告将对市场状态划分为强市、弱市和震荡市三种状态，并探究多种因子在不同市场状态下的表现，最终选出行业指数上期收益率（MoM-Bfficiency）、市盈率（PE_TTM）、实现率（PCF）、基于成交量的 Amihud 流动性（ILLiquidity_Amount）以及根据收益率调整后的换手率（Ajusted Turn_Over）五个因子建立行业轮动模型，最终收益在强市与弱市表现较好，在震荡市表现较差，不太稳定[14]。

数库量化组在《基于事件驱动思想的行业轮动模型》一文中，认为行业轮动规律可能在某个时间区间上并不显著，而在某些特定时间点上表现突出。于是当行业 A 收益率达到五日最高点，则认为触发事件，然后统计下一个交易日上涨的行业。然后按照如此方法统计回测时间区间内所有上涨行业出现次数，挑出上涨次数最多的行业 B，认为 A 与 B 在特定事件发生后存在轮动规律。最终策略年化收益率 31.9%，年华波动率 20.4%，最大回撤 15.5%[15]。

同样是数库量化组在《基于行业轮动规律的预测策略》一文中，将历史数据中每一个交易日按各个行业收益率大小排序，将排序向量作为行业轮动特征向量；同样计算 t 时刻的行业轮动特征向量，然后搜索历史向量空间，寻找与当前行业轮动特征向量空间距离最小的 m 个历史行业轮动特征向量。在匹配出的 m 个历史行业轮动特征向量之中，观察下一个交易日领涨的行业板块，选择出现次数最多的行业板块作为 t+1 时刻预测领涨概率最大的行业板块。最终策略年化收益率 29.1%，年华波动率 33.9%，最大回撤 68.9%[16]。

2.3 本章小结

本章通过对与股票市场行业板块轮动有关文献的引申，既阐述了国内外与板块轮动相关的学术研究贡献，还简要说明了国内大型券商研究报告中一些基于板块轮动的投资策略。我们通过这些研究情况，可以看出随着对我国 A 股市场行业轮动现象研究的不断深入，相关研究人员对股票市场板块轮动现象的理解更加深入，对行业板块轮动的规律的研究手段不断增多。但是因为我国 A 股市场与实体经济间的联系依然比较低，当前的自身机制还不够完善，板块轮动的规律还不够稳定，因此对于股票市场行业板块轮动规律的探究大多还是停留在定性的理解层面，而缺少系统的定量的研究。

第三章 理论基础

3.1 板块轮动现象理论解释

目前关于我国股票市场行业板块轮动现象有着诸多解释，有人从国家政策以及政府的决策等国家政治层面进行解读，也有人依据宏观实体经济周期理论展开分析，还有人从大众心理以及投资者的行为方式方面进行解释。不过总体来对板块轮动现象的理论解释还是要分为行为金融和实体经济两个方面。股票市场价格曲线的波动实质上是实体经济的一种反映，这种反映不单指当前实体经济的发展状态，更为重要的是反映未来的发展形势。当然，这种未来的发展形势也只是投资者根据自己的经验基于市场形势而产生的主观判断，不一定与未来实际发展情况相同。股票市场虽然是虚拟化的资本市场，但是它也应该是对国家实体经济的一种反映。只有这样，股票市场才能使有本之木，有源之水。综合来讲，主要有以下三种对股票市场行业板块轮动现象的理论解释[17]。

（1） 国家政策因素。

虽然改革开放三十多年来，我国经济实力得到了突飞猛进的发展，综合国力也得以逐步加强，但我国依然处于社会主义初级阶段，很多方面还需要进一步的完善与发展。实际来说，我国经济依然在很大程度上受到国家政府的控制，并没有实现理想中完全市场化的局面。这也是为什么那么多欧美国家并不完全认可我们国家的市场经济的地位的重要原因。在我国证券市场近二十多年的逐步发展的进程中，我们可以明显的感受到在我国证券市场中，国家的种种经济政策大都会对其产生决定性的影响，甚至是翻天覆地的改变。然而同时，我国证券市场尚不成熟，仍然存在着的结果矛盾、供需矛盾需要政府来进行调控与监管。

我国股票市场总是带着“政策市”的帽子。这也充分说明了国家经济政策对市场经济的调控，尤其是短时间突发性的重大政策事件，会给股票市场带来极大的冲击与影响。尽管这些年来，证券市场不断在发展进步，市场化的改革不断取得重大成果，国家政策对股票市场的调控也逐渐弱化，但是政府调控依然是引发股票市场价格波动的一个重要的因素。实际上，国际政策对股市过度的影响也是量化投资在我国发展进程中的一个重大阻碍。我们很难利用量化手段去判断国家政策会向那个方向发展，也因此会造成极大的不可预判性。

（2） 行业成长周期因素。

一般来说，一个行业的成长周期大致可以分为引入期、成长期、成熟期和衰退期四个阶段。而且在不同的时期，各个行业大都有着相似的发展形势。当行业处于引入期时，整个行业最近才刚刚出现在人们的视野之中，关键技术还未成熟，经营模式尚且稚嫩，而且缺少成功的经验。也因此当行业在这个时期在股市中的表现往往不够亮眼，但同时这也是这个行业潜力最大的一个时期。当行业到达成长期，行业中的各个企业发展迅速，产品的市场需求快速增长，行业的收入升高，盈利增加，行业规模也是急剧扩大。在这个时期，行业中的各个企业随着不断扩张，必然会增大相互之间的竞争压力，这也是这个行业投资风险相对较高的一个时期。在行业的成熟期，行业中的小型企业大都破产或被收购，只剩下少数几个大型企业分割市场。在这个时期，行业的生产技术和经营模式都相当成熟，行业生长最为迅速昌盛。而当行业进入到衰退期，行业的技术难以进步，利润开始下降，行业发展逐步停止甚至慢慢衰退。

在行业的不同发展时期，不同行业大都有着相似的发展形势。而在股票市场的同一时期，

不同的行业处于各自不同的成长阶段，也因此有着不同的市场表现。比如在我国当前的市场阶段，生物工程和医药相关的行业还处于各自的引入期，潜力很大但股票市场表现并不那么抢眼；IT行业以及旅游、服务业正处于成长期，增值空间相对较大，但也同时面临着巨大的竞争压力。商业、家电板块则处于行业的成熟期，生产技术和经营模式都已经非常成熟，公司的品牌、口碑以及规模是公司竞争力的决定性因素。而造纸印刷以及纺织行业则处于衰退期，行业发展停滞不前，难以有重大突破。

（3） 投资者投资理念因素。

在我国早期的证券市场，大众股民热衷投机，投资极不理性，导致市场常常波动很大，造成很大的投资风险。而随着二十多年的不断发展，投资者逐渐趋于理性，不再像以前那些盲目、任性。随着投资者的素质的提高和经验的提升，投资者的投资理念也发生了巨大的转变，中国股民也开始认可海外成熟市场上流行的以基本面分析入手的“价值投资理念”。

从这世纪初开始，我国股票市场经过新一轮股票价格调整，股民们对业绩更好的板块往往给予跟多的投资关注。股票市场的主体力量也从之前的盲目投机转变为理性的价值投资。这样就造成了朝阳行业的的蓝筹股和绩优股都出现了明显的上涨，进而形成了新一轮的板块轮动现象。

目前我们对于行业板块轮动现象的相关研究往往长于定性短于定量，仍是缺乏相对有效的方法来对板块轮动现象进行量化建模。虽然定性的分析结果不能直接被拿来当作投资策略，但是能为依靠板块轮动规律建立的预测模型提供了理论指导，使得我们在统计上得到的规律可以获得经济学上的肯定[18]。

3.2 隐马尔科夫模型理论

马尔科夫模型是自然语言处理中的基本模型之一，这些年来被人民广泛应用于语音识别、语言处理以及文本挖掘等诸多领域。从上个世纪九十年代开始，隐马尔科夫模型也逐渐被应用于视频图片信号处理以及视频信号处理等方面的研究。截至到目前，隐马尔科夫模型在语音识别、音频检索、视频内容分析等领域取得了巨大的成功[19]。

3.2.1 马尔科夫过程

马尔科夫过程本质上是一种随机过程，是有俄国有机化学家Markov于十九世纪七十年代年提出的。我们假设一个系统有S_1，S_2，…，S_N，N个状态。那么随着过程的进行，系统从某一个状态迁移到另一个状态，设t时刻的状态为q_t，那么当系统在t时刻时处于状态S_j的概率与系统在时间l，2，…，t-l时刻的状态有关，概率如公式2-1所示：

$$P(q_t = S_j \mid q_{t-1} = S_i, q_{t-2} = S_k, ...) \qquad \text{（公式2-1）}$$

如果系统的“将来”的状态仅依赖现在“现在”的状态而与“过去”的状态无关，那么此我们认为这个过程具有马尔科夫性。假设系统在时间t处的状态仅和系统在时间t-1时刻的状态有关，那么这个系统就形成了一个马尔科夫链，如公式2-2所示：

$$P(q_t = S_j \mid q_{t-1} = S_i, q_{t-2} = S_k, ...) = P(q_t = S_j \mid q_{t-1} = S_i) \qquad \text{（公式2-2）}$$

我们也就把具有这种马尔科夫性的随机过程称为马尔科夫过程。

3.2.2 隐马尔科夫模型

在马尔科夫过程中，某个时期的状态是可以直接观测到的，我们在事先是可以确定的。但是在一些问题中，比如说在股票市场中，相同的股价并不对应着相同的市场状态，而且市场的状态我们我们是无法直接观测到无法确定的，这里我们就需要用到隐马尔科夫模型。

隐马尔科夫模型是由一个是隐含状态的转换序列以及一个与隐含状态对应的观测值序列来共同组成的[21]。而且在上述的两个随机过程中，我们只能直接看到隐含状态序列对应

的观测值序列，而不能直接观测到隐含状态及其状态转移过程，但我们可以对观测值序列进行分析推断，进而得到对应的隐含状态序列。我们也因此把这个模型称为隐马尔科夫模型。在组成隐马尔科夫模型的两个随机过程中，隐含状态的转移是一个马尔科夫链，各个隐含状态间的转移概率用转移概率矩阵来表示，而隐含状态到观测值的概率，我们用混淆矩阵也就是发射矩阵来表示。隐马尔科夫模型如图2-1所示：

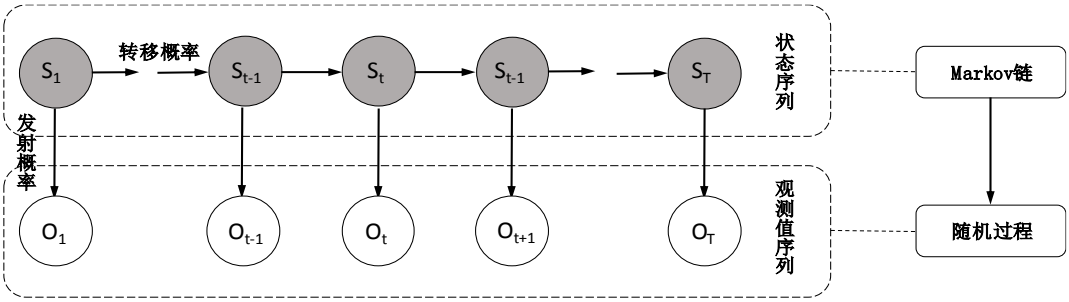

图2-1 隐马尔科夫模型组成示意图

通常我们用5个模型参数来表示一个完整的隐马尔科夫模型，分别记为{N，M，A，B，π}：

（1）N表示马尔科夫模型的隐含状态数目。我们用S来表示隐含状态的集合，如果模型中包含N个隐含状态，那么我们可以将隐含状态集表示为$S=\{S_1, S_2, S_3, \ldots, S_N\}$。我们设模型在t时刻的隐含状态为$q_t \in S$，其中$1 \le t \le T$，T为观察值序列的长度。模型的隐含状态序列我们可以记为$Q=\{q_1, q_2, \ldots, q_t\}$。

（2）M表示隐含状态对应的观察值数目。我们用V来代表输出观测值的集合，如果模型包含M个输出观测值，那么我们可以将观测值集合记为$V=\{V_1, V_2, V_3, \ldots, V_M\}$。

（3）A为隐含状态相互之间转移的概率分布矩阵，称为转移矩阵。我们用矩阵$A=\{a_{ij}\}$来表示隐含状态的转移概率分布，我们设$a_{ij}=P\{q_{t+1}-S_j|qt=Si)$，其中$1 \le i, j \le N$，且满足条件$a_{ij} \ge 0$，$a_{ij}$求和为1，$a_{ij}$表示系统从t时刻的隐含状态$S_t$转换到t+l时刻的隐含状态$s_j$的概率。

（4）B为隐含状态对应的观测值的概率分布，称为发射矩阵或混淆矩阵。我们如果假设K为观测值向量的样本空间，那么我们可以用矩阵$B=\{b_i(k)\}$来表示隐含状态到观察值的概率分布，我们设$b_i(k)=P\{O_t=V_k|q_t=S_i)$，其中$1 \le i \le N$，$l \le k \le M$ ，且满足$b_i(k) \ge 0$，$b_i(k)$求和为l。$b_i(k)$代表隐含状态为S_i时对应的输出观测值为V_k的概率

（5）π代表初始状态的概率分布矩阵。系统中初始时刻的状态概率分布我们可以用矩阵$\pi=(\pi_i)$来表示，我们设$\pi_i=P(q_1=S_i)$，其中$1 \le i \le N$。π_i则代表隐含状态S_i作为初始状态的概率，

同时隐马尔科夫模型也需要满足以下三条重要假设：

（1）马尔科夫假设（隐含状态序列构成一阶马尔科夫链）

$$P(q_t = S_j \mid q_{t-1} = S_i, q_{t-2} = S_k, \ldots) = P(q_t = S_j \mid q_{t-1} = S_i) \quad \text{（公式2-2）}$$

（2）不动性假设（状态与具体时间无关）

$$P(q_t = S_j \mid q_{t-1} = S_i) = P(q_m = S_j \mid q_{m-1} = S_i), \forall i, j \quad \text{（公式2-3）}$$

（3）输出独立性假设（输出仅与当前状态有关）

$$P(O_t = V_k \mid O_1 = V_i, \ldots, O_{t-1} = V_j, q_1 = S_i, \ldots, q_t = S_j) = P(O_t = V_k \mid q_t = S_j) \quad \text{（公式 2-4）}$$

3.2.3 基于隐马尔科夫模型的三类典型问题

（1）观测序列的概率（评估问题）

给定观测序列$O=O_1O_2\ldots O_T$和模型参数$\lambda=(A,B,\pi)$，计算出模型在该参数下生成观测序列O的概率，即$P(O|\lambda)$。我们可以认为是衡量一个模型和对应观测序列的匹配程度，我们通常使用Forward-backward算法来解决这类问题。

（2）预测隐马尔科夫隐含状态（解码问题）

给定观测序列$O=O_1O_2\ldots O_T$和模型参数$\lambda=(A,B,\pi)$，求出最优的隐含状态序列，$Q=q_1q_2\ldots q_T$。也可以理解成为对输出观察的最佳“解码”。常用的解决方法是Viterbi算法。

（3）估计模型的参数（学习问题）。

给定模型初始参数λ，通过调整模型参数$\lambda=(A,B,\pi)$，使得对于一个给定的观测序列$O=O_1O_2\ldots O_T$，使得$P(O|\lambda)$达到最大。此问题主要是通过参数优化，最佳的描述观测序列是如何得来的，即学习观测序列。常用的解决方法是Baum-Welch 算法。

3.2.4 解决三类典型问题的常用算法描述

（1）Forward-backward算法

Forward-backward算法使用的是动态规划的原理，在这里用于计算给定参数下隐马尔科夫模型生成特定观察序列的概率，如果用M代表隐含状态的数目，用T代表马尔科夫序列的长度，那么我们可以算出Forward-backward算法的算法复杂度为$O(M^2T)$。

Forward-backward算法常被我们拿来解决评估问题，算法描述如下：

我们定义前向变量为$\alpha_t(i)=P(O_1O_2\ldots O_t，q_t=S_i|\lambda)$，表示了在给定了参数的隐马尔科夫模型下，在t时刻我们观测到的输出值序列为$O_1O_2\ldots O_t$，对应的隐含状态为S_i的概率大小。

前向算法：

1．初始化：

$$\alpha_1(i)=\pi_i b_i(O_1),1\le i\le N \qquad \text{（公式2-5）}$$

2．递推：

$$\alpha_{t+1}(j)=\left[\sum_{i=1}^{N}\alpha_t(i)a_{ij}\right]b_j(O_{t+1}),1\le t\le T-1,1\le j\le N \qquad \text{（公式2-6）}$$

3．终止：

$$P(O|\lambda)=\sum_{i=i}^{N}\alpha_T(i) \qquad \text{（公式2-7）}$$

后向算法与前向算法在思想上基本一致，只是除了两者在递推的方向上相反外，在算法上两者也有少许不同之处。

我们定义后向变量为$\beta_t(i)=P(O_{t+1}O_{t+2}\ldots O_T|q_t=S_i,\lambda)$，表示了在给定了参数的隐马尔科夫模型下，在t时刻对应的隐含状态为S_t的情况下，从t+1时刻到最后的T时刻期间的我们观测到的输出值序列为$O_{t+1}O_{t+2}\ldots O_T$的概率大小。与前向算法相似，我们可以利用递推来得到所有的$\beta t(i)$:

1．初始化：

$$\beta_1(i)=1,1\le i\le N \qquad \text{（公式2-8）}$$

2．递推：

$$\beta_t(i)=\sum_{i=1}^{N}a_{ij}b_j(O_{t+1})\beta_{t+1}(j),t=T-1,T-2,\ldots,1,1\le i\le N \qquad \text{（公式2-9）}$$

3．终止：

$$P(O|\lambda)=\sum_{t=1}^{N}\beta_t(i) \quad \text{（公式2-10）}$$

（2）Viterbi算法

Viterbi算法的目的是在给定隐马尔科夫模型参数λ=(M，N，A，B，π)，找出时生成观测序列$O=O_1O_2\ldots O_T$概率最大的隐含状态序列$Q=q_1q_2\ldots q_T$，也既是解码问题。如果用M代表隐含状态的数目，用T代表马尔科夫序列的长度，那么我们可以算出Viterbi算法的算法复杂度为$O(M^2T)$。

我们定义$\delta_t(i)=\max P(q_1q_2\ldots q_{t-1},q_t=i,O_1O_2\ldots O_t|\lambda)$，则算法描述如下：

1．初始化：

$$\delta_1(i)=\pi_i b_i(O_1),\phi_1(i)=0,1\le i\le N \quad \text{（公式2-11）}$$

2．递推：

$$\delta_t(j)=\max_{1\le j\le N}\left[\delta_{t-1}(i)a_{ij}\right]b_j(O_j),2\le t\le T,1\le j\le N \quad \text{（公式2-12）}$$

$$\phi_t(j)=\arg\max_{1\le j\le N}\left[\delta_{t-1}(i)a_{ij}\right],2\le t\le T,1\le j\le N \quad \text{（公式2-13）}$$

4．终止：

$$P^*=\max_{1\le j\le N}[\delta_T(i)] \quad \text{（公式2-14）}$$

$$q_T^*=\arg\max_{1\le j\le N}[\delta_T(i)] \quad \text{（公式2-15）}$$

5．求S序列：

$$q_1^*=\phi_{t+1}(q_{t+1}^*),t=T-1,T-2,\ldots,1 \quad \text{（公式2-16）}$$

（3）Baum-Welch算法

Baum-Welch算法是对期望最大化算法(EM算法，Expectation Maximization)进行了改进后的一种迭代算法，在算法刚开始的时候是由用户自己依靠自己的经验给出对隐马尔科夫模型各项参数的估计初始值，然后通过多次迭代计算，不断的优化参数以达到收敛，最终逐步得到更加合理的表现较好的参数值。本文在运用隐马尔科夫模型进行预测建模时，第一步便是先利用样本观察值序列对隐马尔科夫模型的参数进行学习训练，得到与观察值序列最为匹配的模型。

定义：

$\xi_t(i,j)$：t时状态为S_i以及t+1时状态为S_j的概率

$\gamma_t(i)=\sum_{j=1}^{N}\xi_t(i,j)$：t时处于状态$S_i$的概率

$\sum_{t=1}^{T-1}\gamma_t(i)$：整个状态中从$S_i$转出次数的预期

$\sum_{t=1}^{T-1}\xi_t(i,j)$：从$S_i$跳转到$S_j$次数的预期

Baum-Welch算法描述如下：

1. 初始化：

$$\pi_i = \gamma_1(i), \lambda=(A0,B0,\pi) \quad \text{（公式2-17）}$$

2. 递推关系式：

$$\xi t(i,j) = \frac{P(qt=i, qt+1=j, O \mid \lambda)}{P(O \mid \lambda)} = \frac{\alpha_t(i) a_{ij} b_j(O_{t+1}) \beta_{t+1}(j)}{P(O \mid \lambda)}$$
$$= \frac{\alpha_t(i) a_{ij} b_j(O_{t+1}) \beta_{t+1}(j)}{\sum_{i=1}^{N} \sum_{j=1}^{N} \alpha t(i) a_{ij} b_j(O_{t+1}) \beta_{t+1}(j)} \quad \text{（公式2-18）}$$

3. 重估公式：

$$\tilde{a}_{ij} = \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)} \quad \text{（公式2-19）}$$

$$\tilde{b}_j(k) = \frac{\sum_{\substack{t=1 \\ O_t=V_k}}^{T} \gamma_t(j)}{\sum_{t=1}^{T} \gamma_t(j)} \quad \text{（公式2-20）}$$

4. 终止条件：

$$\left|\log P(O \mid \lambda) - \log P(O \mid \lambda_0)\right| < \varepsilon, \varepsilon \text{为预设阈值} \quad \text{（公式2-20）}$$

3.3 本章小结

本章先是从国家政策、行业成长周期以及投资者理念变化三个方面对板块轮动现象产生的原因进行了合理的理论分析。随后引出了本文采用的对隐马尔科夫模型，并对马尔科夫过程、隐马尔科夫结构、三类典型问题及其常用解决算法做了比较详细的讲解，为后文建立隐马尔科夫模型打好了理论基础。

第四章 研究方法与思路

根据前文的一些描述，我们已经了解到，股票市场中的行业板块轮动往往是具有一定的规律的。然而，这些轮动的规律并不能很轻易直观地被我们探索把握到。根据有效市场理论，资产价格反映市场中的所有信息。因此如果行业板块轮动存在一定的规律，那么这些规律也一定会反映在行业板块的价格变动中。于是我们可以从历史价格来挖掘行业轮动的规律。将每天行业板块的涨跌情况标记为行业轮动的特征，建立合适的股票板块轮动模型，进而预测未来的行业轮动情况。

4.1 本文基本研究思路

本文分别采用了中国 A 股市场各行业板块指数每日收益率以及与市场表征有关的常见的指数因子作为两种观测向量建立隐马尔科夫模型，解码出最有可能的隐状态序列，并计算得到再该模型下观测向量对应的似然值序列。通过对市场隐含状态、行业轮动特征以及对数似然值进行层层匹配，从多年的历史数据中找出与当前行为模式最为接近的交易日，进而对下一阶段涨幅较大市场表现较好的的行业板块进行预测。

4.1.1 市场隐含状态匹配

金融市场中在不同的阶段有着不同的市场状态，比如说“牛市”、“熊市”以及“震荡市”。而且显然不同的市场状态对应着不同概率的上涨下跌，其中的行业轮动的规律也不尽相同；只有当市场处于相同的状态时，行业内在联系才能表现为一定的行业轮动特征。也因此，我们在匹配当前行业轮动特征与历史行业轮动特征时，不得不考虑到当前市场状态与历史市场状态的一致性。

我们通常认为金融市场的资产价格具有马尔科夫性，即当前的市场状态与过去的状态无关。我们也因此可以借助隐马尔科夫模型确定市场状态。本文分别采用了中国A股市场每日各行业板块指数收益率以及与市场表征有关的常见的指数因子作为两种观测向量（用作对比），借用历史数据确建立隐马尔科夫模型，然后解码出最有可能的隐状态序列，遍历历史状态序列，匹配出与当前市场状态相同的历史时段，进而可以更好的进行行业轮动特征的匹配，以达到更加精确预测的目的。

4.1.2 行业轮动特征匹配

首先利用股市历史数据，求取并标记历史上每日收益率最大的行业板块，将连续两天的行业标记作为行业轮动的特征。遍历历史数据，将当前行业轮动特征与和当前市场状态一致的历史行业轮动特征相匹配，进而挑选出与当前交易日 i 具有相同市场隐含状态相同行业轮动特征的交易日 j，匹配方式是令第 i 天和第 i-1 天的日收益率最大的行业板块分别与第 j 天和第 j-1 天的日收益率最大的行业板块相同，以便进行下一步的似然值模式匹配。

4.1.3 似然值匹配

在利用历史数据建立隐马尔科夫模型之后，计算得到样本观测值序列在通过数据学习的到的模型中的对应的对数似然值序列。遍历在前两步中得到的与当前相同市场隐含状态相同行业轮动特征的历史数据，从中挑选出与当日对应的对数似然值最为接近的N个样本数据。进而对下一时刻上涨概率较大的行业板块进行预测[21]。

4.2 滑动样本窗口法

由于隐马尔可夫模型是一种体现短期自相关的非线性模型，因此我们不应该用过长的样本数量拿来建模，而且模型也只是对接下来较短的时间序列预测。为提升模型预测效果，解决上述两个问题，本文决定采用一种滑动样本窗口的方法来进行建模。这种方法每次从总样本中动态采取用于建模的长度为T的样本，设预测周期为t，则每隔t天删除样本序列中前t天数据，并同时添加后t天观测数据，利用建立的模型对接下来t天进行预测，即每隔t天需要对模型进行一次更新[21]。

在每个预测周期中，我们利用选择出来的T个样本进行建模，并预测得到T个样本对应的市场隐含状态序列，接下来我们利用对样本学习得到的模型参数计算得出T个样本对应的对数似然值序列。在进行预测的过程中，我们首先得到第i个交易日的观测值向量组在该参数下模型中对应的对数似然值K。

（1）单日预测

遍历T个样本观测向量对应的对数似然值序列，然后从中找出一个最接近于K的似然值L，也就是找出历史中的一个与第i个交易日行为模式最接近的交易日j，以第j日之后一天，既j+1日的各行业板块指数市场表现情况来预测第i+1日的上涨概率较大的行业板块。简而言之，就是从众多历史样本中找出与需要预测的交易日行为模式最接近的一天来对下一阶段市场表现最好的行业板块进行预测。

（2）多日加权预测

由于市场具有相对随机性，因此单日预测方法的波动性过大，很容易受到误差影响。因此我们可以从众多历史样本中选取多个与需要预测的交易日行为模式相近的交易日，对i+1日的各行业板块涨跌幅度进行不同程度的加权，从而我们预测得到一个相对来说跟家稳定、不宜受到误差影响的的预测值。例如，我们通过从历史数据中选取与第i日行为模式最接近的n个交易日作为预测参照，令每个交易日在预测行业板块指数涨跌的加权中的权重反比于其观测值的对数似然值和预测日的对数似然值的差值，也就是说对数似然值越接近的历史样本所赋予的权重越大。这样就可以很好地避免了市场随机性带来的误差。

4.3 板块轮动量化难点及改进方法

4.3.1 量化建模难点

股票市场的行业板块轮动现象是受到经济周期、国家政策、投资者心理等诸多因素影响而表现出来的一种宏观经济现象，也因此对板块轮动现象的量化建模也存在着两个方面难点[18]。

（1）板块轮动因果关系难以把握。一个行业板块指数的上扬很有可能同时拉动多个板块出现不同程度地上涨，这些行业板块之间的逻辑关系我们很难直接进行判断，或者说在很多时候他们并不存在很明确的逻辑关系。究其原因，很大程度上是因为普通投资者的注意力容易被表现最好的板块而吸引，形成羊群效应，从而令真正有意义的连续的板块轮动难以被观察到。

（2）板块轮动的持续时间难以确定。行业板块轮动的持续时间很大程度上与行业本身的特点息息相关，不同的板块也就因此可能会面临不同的长度的轮动时间，这会对极大地干扰到观察者对行业板块轮动规律的判断。行业板块不同的轮动持续时间必然会打乱板块轮动的因果次序，可能会使观察者将行业板块之间的因果逻辑顺序在时间上发生了背离，致使统计指标分析和传统的计量模型分析丧失意义甚至是给出错误的判断。

4.3.2 改进方法

由于板块轮动的量化建模有以上两个难点，在对实际数据的预测中我们需要对上述预测

方法进行近一步改进，以便能更加契合实际股市中的板块轮动现象。

对于第二个问题问题，由于本文的研究使用的是各行业板块指数的日涨跌数据，采用了模式匹配的方法来进行预测，因此很大程度上减少了不同板块轮动时间不同带来的负面影响。而对于第一个问题，由于板块轮动引起的行业板块指数上涨程度我们不能确定，所以我们做出以下两个方面的改进。

（1）对于行业轮动特征，我们将不再只标记历史上每日收益率最大的行业板块，而是改为将历史数据中每天行业板块指数中收益率排行前三的3个板块都予以标记，以便减少板块轮动引起的行业板块指数上涨程度的不确定性造成的误差影响，另一方面也减少了行业轮动特征匹配的难度。

（2）对于最终的加权预测，本文自行选择一种更符合实际情况的加权方式。首先，我们通过市场隐含状态匹配以及行业轮动特征匹配，进一步选择出与当前相同市场隐含状态相同行业轮动特征，并且似然值最为接近的5个交易日。然后选出这5个交易日之后一天的收益率排行前三的3个行业指数。并根据选出的5天历史交易日与当天似然值的接近程度依次赋予不同权重。例如表4-1，对于下表匹配结果，我们分别进行如下所示权重分配：

表4-1 行业板块权重及分配比例

日期（似然值接近程度依次递减）	收益率最高（权重）	收益率第二（权重）	收益率第三（权重）
2015-5-13	农林牧渔（5*1/2）	钢铁（5*1/3）	电子（5*1/6）
2015-6-3	电子（4*1/2）	有色金属（4*1/3）	化工（4*1/6）
2014-12-23	钢铁（3*1/2）	农林牧渔（3*1/3）	化工（3*1/6）
2015-1-14	化工（2*1/2）	农林牧渔（2*1/3）	电子（2*1/6）
2015-3-3	餐饮旅游（1*1/2）	建筑材料（1*1/3）	农林牧渔（1*1/6）

本文选择将5天中几个收益率最高的行业作为下一阶段上涨概率最大的行业的预测候选集，并按照上表分配的权重比对候选集中行业板块进行投票，进而选取出综合权重最大的板块进行预测。

4.4 本章小结

本章讲述了对股票市场行业板块轮动现象建模的基本思路和主要方法。先是介绍了对市场隐含状态、行业轮动特征以及似然值进行匹配的具体方法。又由于隐马尔可夫模型是一种体现短期自相关的非线性模型，用来建模的样本数不宜过长的原因，采用了一种滑动样本窗口法来进行建模。最后提出对板块轮动现象进行量化的两个难点以及做出的改进措施。

第五章 实证研究及结果分析

5.1 样本选取与说明

在行业轮动规律的探究过程以及在行业轮动特征的匹配过程中，行业的分类与选择都是其成功与否的关键所在。行业指数的选取要求所选择的行业既要具有内在的关联，又要有一定的区分度：内在的关联使得所选择行业存在一定的轮动规律，是策略成功的基础；一定的区分度要求所选择行业的收益率相关性不能过大，否则策略难以准确的捕捉到行业的轮动规律。因此我们对于行业板块的划分也需要适当，板块数目不宜过多也不能过少。过多的板块数目会加大行业特征匹配的困难，而过大的板块划分会掩盖部分行业轮动规律，进而影响到预测准确率。

目前来说，上海证券交易所和深圳证券交易所的合资子公司中证指数所发布的行业指数的“血统”最为高贵。而且中证指数的行业分类标准与标准普尔500指数为代表的国际通行的行业分类保持一致，满身都是国际范儿。可是一旦过于国际范了，就会忽视掉中国的特有国情，因此中证指数的行业板块分类一遇到中国特色的A股市场就变得水土不服。比如在国外股票市场中，房地产类的股票很多市值并不算高，对整体股票市场的影响也不大，基本上是不会划分为一级行业，可是A股市场的情况与国外却是大不一样，毕竟“招金万保”四大房地产龙头(招商地产、金地集团、万科集团、保利地产)可是有着赫赫威名。就凭这点，中证指数的一级行业分类也就很自然的在中国玩不转了。

所以虽然中证指数的行业分类根正苗红，但是一来不太适合中国A股市场，二来一级行业数目太少，板块过大而导致轮动规律不明显，所以本文最终选用分类更细致，更受欢迎也更有利于热点跟踪的的申万指数体系。申万一级行业将A股3000多只股票划分成28个一级行业，分别为采掘、农林牧渔、化工、有色金属、钢铁、电子、家用电器、公用事业、食品饮料、纺织服装、医药生物、商业贸易、轻工制造、房地产、休闲服务、建筑装饰、建筑材料、电气设备、交通运输、国防军工、传媒、计算机、通信、非银金融、银行、汽车、综合、机械设备。本文选取了2007年1月8日到2013年12月31日期间的历史数据来进行股票行业板块轮动规律的研究。选取了2014年2月24日到2017年4月25日期间的数据来对两种不同观测向量下的预测效果的进行对比分析。

5.2 预测步骤详解

（1）数据处理

根据历史行业板块指数相关数据，我们首先处理得到每日的各板块收益率指数，并由此标记出每日收益率排行前五的行业。由于选择每日28个行业板块收益率指数作为隐马尔科夫模型的观测向量，为了更好的学习效果，我们采用主成分分析（PCA）技术来进行降维处理。

(2)模型参数确定

本文选用样本窗口大小为350个交易日，预测窗口大小为一个交易日。对于隐马尔科夫模型中的隐含状态数量，不宜过多，也不能太少。过少的隐含状态会导致当前状态与相似的历史状态无法准确的匹配，过多的隐含状态则会导致匹配上的困难。因此在本文中，我们选择了效果最好的三种隐含状态。

对隐马尔科夫模型的建模学习以及隐状态的预测，本文是运用Python中的hmmlearn0.2.1

库来实现的。其中混合高斯模型GaussianHMM有四种不同的混淆矩阵选择(隐状态与观测向量间的发射概率矩阵)，分别如下：

1．Spherical 对应的混淆矩阵表示了在每个隐含状态下，观测值向量的所有特性分量使用相同的方差值。对应协方差矩阵中非对角处的数据为零值，并且对角处的值相等，即表现出球面特性，如图 5-1 所示。

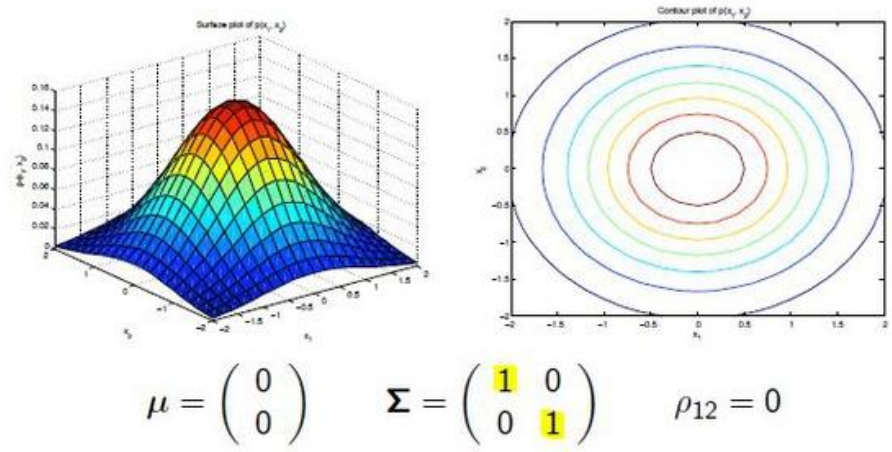

图 5-1 spherical 对应混淆矩阵特征

2．diag 对应的混淆矩阵表示了在每个隐含状态下，观测值向量使用对角协方差矩阵。对应协方差矩阵中非对角出数据均为零值，而其对角线处的值不相等，如图 5-2 所示。

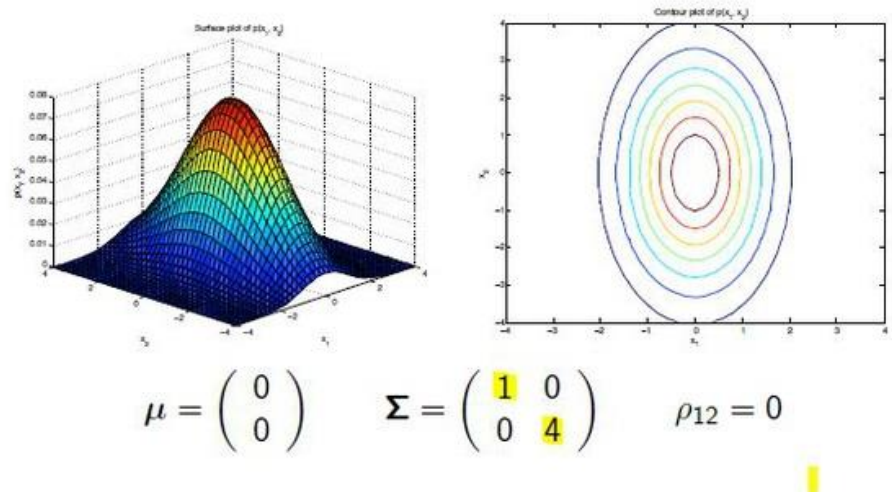

图 5-2 diag 对应混淆矩阵特征

3．full 对应的混淆矩阵表示了在每个隐含状态下，观测值向量使用完全协方差矩阵。对应的协方差矩阵里所有元素均不为零值，如图 5-3 所示。

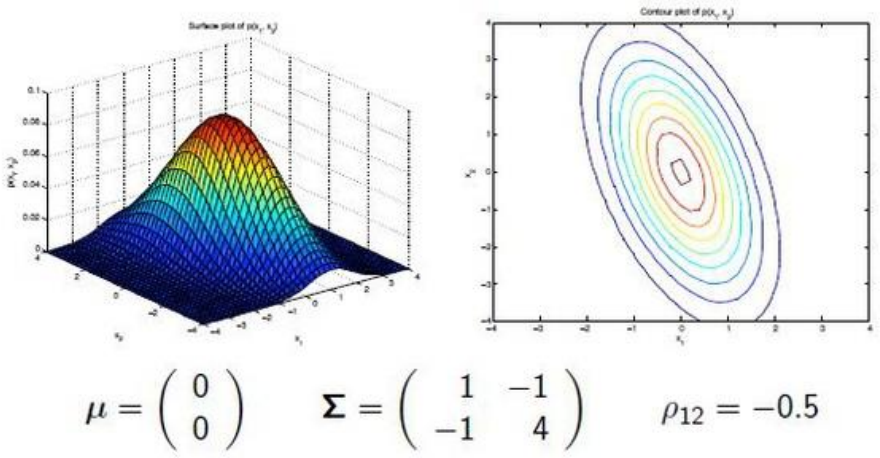

图 5-3 diag 对应混淆矩阵特征

4．tied 表示所有的马尔可夫隐含状态都使用相同的完全协方差矩阵。

在上述四种类型中，前三种分别代表不同类型的高斯分布概率密度函数，而第四种则是高斯隐马尔科夫模型的特有形式。在这四种类型中，full 是模拟能力最强大，也是需要最多的数据来进行的参数估计，如果数据量不够则会得到很差的参数模型；spherical 是四个中相对来说最简单的类型，大多数情况下不需要太多的数据就能达到相对表现较好的参数；而 diag 则是这上述两者一个折中，是隐马尔科夫包中的默认类型。本文则采用实际效果更好的 spherical 模式的混淆矩阵。

（3）隐状态序列预测

将处理好后的观测向量放到上述隐马尔可夫模型中进行训练，并利用训练好的模型预测出观测向量序列对应的最有可能的隐状态序列。

在本文中我们将市场隐含状态分为三种，为了比较清楚地说明这三种市场状态分别是什么，我们利用 2005 年一月年到 2015 年 12 月这 11 年期间数据建立隐马尔科夫模型，并预测出每一天的市场隐含状态，得到下图 5-4

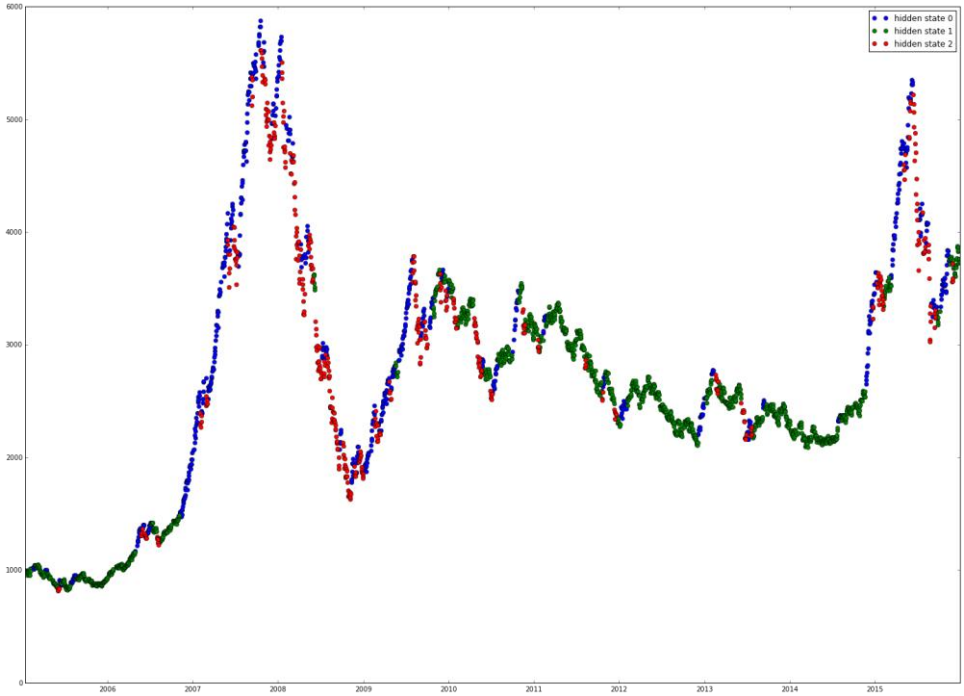

图 5-4 三种隐含状态在股票市场走势图中的标注

由图 5-4，我们可以清晰地看出三种隐含状态分别对应着股票市场上的“牛市”、“熊市”以及“震荡市”。

（4）求取对应似然值序列

在训练得到隐马尔科夫模型后，我们计算出在该参数模型下观测向量序列对应的似然值序列，我们认为似然值越接近的观测向量的数据模式也越相似。比如 2016 年 1 月 11 日当天对应的对数似然值为-26.76，同时，从数据集对应的对数似然值中，寻找与-26.76 最接近的值对应的且具有相同隐含状态，相同行业轮动特征的日期。通过遍历似然值序列，我们可以迅速找到 2014 年 12 月 21 日的对数似然值-26.86 与其最 相近且满足条件，两者数据模式对比如表 5-1、图 5-5 所示

表 5-1 2016 年 1 月 11 日与 2014 年 11 月 11 日数据对比表

行业板块（收益率百分比）	农林牧渔	采掘	化工	钢铁	有色金属	电子	家用电器
2016/1/11	-5.45	-5.94	-6.60	-5.69	-5.69	-7.96	-5.83
2014/11/11	-1.56	-1.03	-3.11	-2.64	-2.45	-3.57	-2.19

续表 5-1

行业板块（收益率百分比）	食品饮料	纺织服装	轻工制造	医药生物	公用事业	交通运输	房地产
2016/1/11	-5.03	-7.36	-6.84	-6.89	-6.21	-5.15	-5.74
2014/11/11	-1.00	-3.32	-3.32	-2.52	-2.63	-3.01	-1.24

续表 5-1

行业板块（收益率百分比）	商业贸易	休闲服务	综合	建筑材料	建筑装饰	电气设备	国防军工
2016/1/11	-6.69	-4.57	-6.80	-6.48	-4.47	-7.63	-8.89
2014/11/11	-3.72	-1.72	-3.49	-2.69	-3.53	-2.18	-4.85

续表 5-1

行业板块（收益率百分比）	计算机	传媒	通信	银行	非银金融	汽车	机械设备
2016/1/11	-7.07	-6.65	-6.67	-3.46	-7.28	-6.88	-6.73
2014/11/11	-3.56	-2.94	-1.95	2.99	0.29	-3.01	-3.44

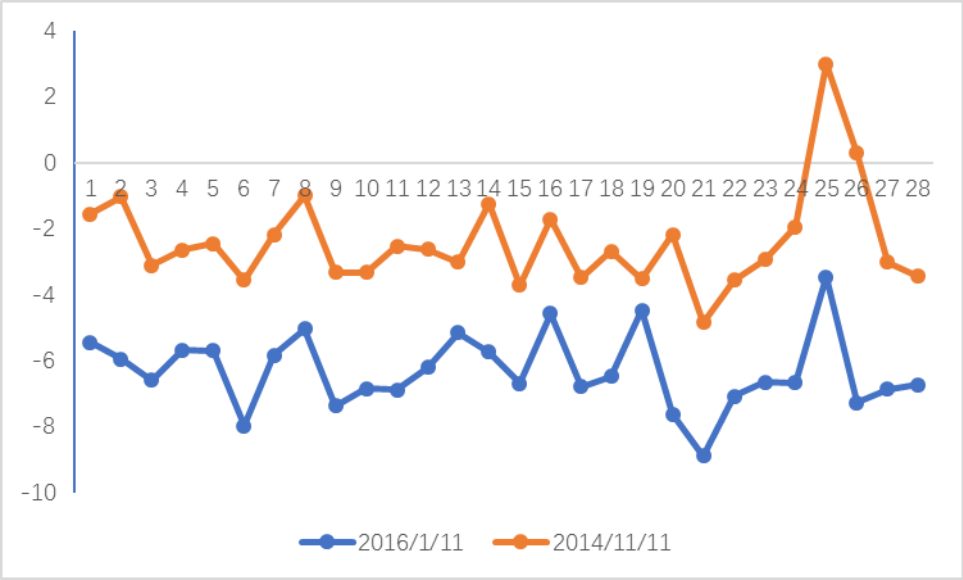

图 5-5 2016 年 1 月 11 日与 2014 年 11 月 11 日数据对比图

（5）模式匹配

在利用历史数据建立隐马尔科夫模型之后，利用学习得到的参数预测预测样本观测值序列对应的隐含状态序列，并计算出对应的对数似然值序列。遍历样本空间，选择出与当前交易日i具有相同市场状态和相同板块轮动特征的交易日j。轮动特征的匹配方式是令第i日收益率最高板块在第j日所有行业板块中收益率排行前三；并且第i-1日收益率最高板块在第j-1日所有行业板块中收益率排行前三。最终在与当前具有相同市场状态和相同板块轮动特征的若干交易日中选取与当前交易日对应的似然值最为接近的5个交易日，以便进行最后的加权预测。

（6）加权预测

根据模式匹配得到的与当前最为接近的5个交易日，并求得每个交易日收益率排行前三的行业，按照表4-1所示赋予各自的权重。

表4-1 行业板块权重及分配比例

日期（似然值接近程度依次递减）	收益率最高（权重）	收益率第二（权重）	收益率第三（权重）
2015-5-13	农林牧渔（5*1/2）	钢铁（5*1/3）	电子（5*1/6）
2015-6-3	电子（4*1/2）	有色金属（4*1/3）	化工（4*1/6）
2014-12-23	钢铁（3*1/2）	农林牧渔（3*1/3）	化工（3*1/6）
2015-1-14	化工（2*1/2）	农林牧渔（2*1/3）	电子（2*1/6）
2015-3-3	餐饮旅游（1*1/2）	建筑材料（1*1/3）	农林牧渔（1*1/6）

接下来我们选择将5天中每天收益率最高的行业作为下一阶段上涨概率最大的行业的预测候选集，如上表中的农林牧渔、电子、钢铁以及化工四个行业，然后按照上表分配的权重比对候选集中行业板块进行投票，进而统计出几个行业各自占据权重大小。同样以表4-1为例：

农林牧渔行业权重：$W_1 = 5\times\frac{1}{2}+3\times\frac{1}{3}+2\times\frac{1}{3}+1\times\frac{1}{6}=4\frac{1}{3}$

电子行业权重：$W_2 = 5\times\frac{1}{6}+4\times\frac{1}{2}+2\times\frac{1}{6}+1\times\frac{1}{2}=3\frac{2}{3}$

钢铁行业权重：$W_3 = 5\times\frac{1}{3}+3\times\frac{1}{2}=3\frac{1}{6}$

化工行业权重：$W_4 = 4\times\frac{1}{6}+3\times\frac{1}{6}+2\times\frac{1}{2}=2\frac{1}{6}$

进而依靠个行业板块的权重大小对下一阶段上涨概率较大大的板块进行预测。

对于最后的预测准确率（胜率），本文分为以下几种情况：

（1） 综合权重最高的行业在预测当天28中行业中涨幅排行前五为事件A，统计概率为P(A)。

（2） 综合权重第二的行业在预测当天28中行业中涨幅排行前五为事件B，统计概率P(B)。

（3） 综合权重第三的行业在预测当天28中行业中涨幅排行前五为事件C，统计概率P(C)。

（4） 综合权重排行前二的行业在预测当天28中行业中涨幅排行均前五为事件$A\cap B$，统计概率P($A\cap B$)。

（5） 综合权重排行前二的行业在预测当天28中行业中有一个涨幅排行前五为事件$A\cup B$，统计概率P($A\cup B$)。

（6） 综合权重排行前三的行业在预测当天28中行业中涨幅排行均前五为事件$A\cap B\cap C$，统计概率P($A\cap B\cap C$)。

（7） 综合权重第一与第二的行业在预测当天28中行业中有一个涨幅排行前五为事件$A\cup B\cup C$，统计概率P($A\cup B\cup C$)。

5.3 结果分析

5.3.1 两种不同观测向量对比分析

上文中，我们选择了将每日的各行业板块收益率指数作为隐马尔科夫模型的观测向量，来区分市场中不同的隐含状态。而对于市场状态的判断，我们通常会选用表征市场状态的一些常用因子来联合进行分析。在本文中，我们也将两类观测向量在试验中的实际效果做了对比与分析。

本文中采用对数收益率，相对强弱指标，乖离率，平均真实波幅四种因子来区分市场中不同的隐含状态。其中：

（1）对数收益率（LogReturn）属于收益率指标，表示价格的变动情况，计算方法见公式 4-1：

$$R_t = \ln(\frac{P_t}{P_{t-1}}) \qquad \text{(公式 4-1)}$$

其中 R_t 为第 t 日的对数收益率，P_t 为第 t 日的收盘价

（2）相对强弱指标（RSI）属于反趋向指标，利用供求平衡的理论，将一段时间内股票上涨的部分与下跌的部分进行对比，用来衡量买方与卖方之间的强弱，进而对之后股票走势具有一定指示作用。计算方法见公式 4-2：

$$\text{RSI}=100\times\frac{A}{A+B} \qquad \text{（公式 4-2）}$$

其中 A 为 M 天内股价上涨总幅度，B 为 M 天内股票下对总幅度，一般 M 取 5，10，14 为一周期，另外也有以 6，12，24 为计算周期的。本文选取 M=6。

（3）乖离率（BIAS）也是属于反趋向指标，表示了股价在市场波动过程中与移动平均线之间相互偏离的程度大小。计算方法见公式 4-3：

$$BIAS=100\times\frac{\text{收盘价-收盘价的N日平均值}}{\text{收盘价的N日平均值}} \quad \text{（公式 4-3）}$$

其中 N 一般定为 5、6、10、12、24、30 和 72，本文选用 N=6

（4）平均真实振幅（ATR）属于波动指标，主要表示一段时间内市场波动强烈程度。ATR 是真实波幅(TR)的 N 日移动平均值，本文中 N 取 14 天，其中 TR 为以下三个值中的最大者：

1. 当前交易日的最高价与最低价间的波幅
2. 前一交易日收盘价与当前交易日最高价间的波幅
3. 前一交易日收盘价与当前交易日最低价间的波幅

三种情况分别如图 5-4 所示：

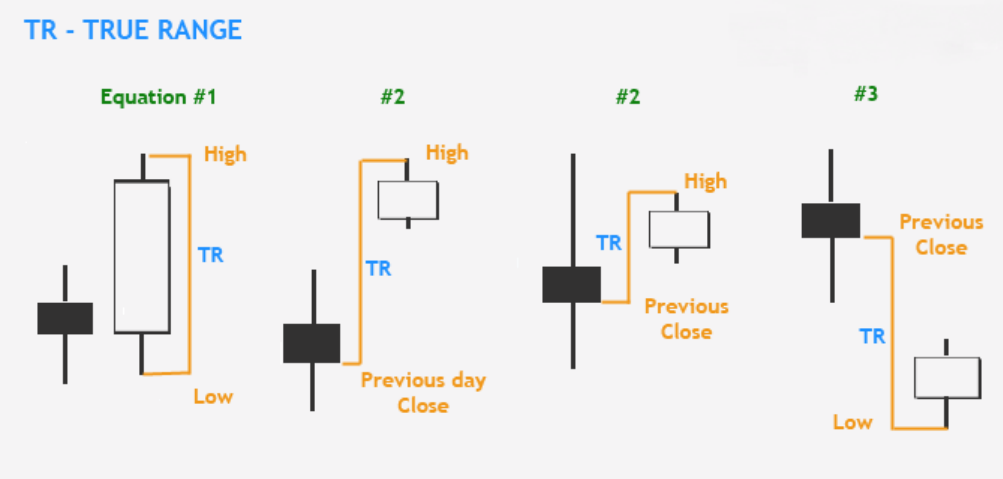

图 5-6 真实振幅的三种情况

对于大盘，我们选用了沪深 300 指数来进行表征。沪深 300 指数是从上海和深圳证券市场中选取 300 只最具代表性的 A 股股票作为样本而编制成的成分股指数。然而虽然沪深 300 指数所选取的样本股票的市值已经超过了沪市和深市两个股票市场总市值的百分之六十，有着很出色的的市场代表性，但是它主要表征的是主板市场的市场行情。而创业板市场则虽然是地位次于主板市场的二板市场，但同时也对 A 股市场影响巨大。因此我们同时使用了沪深 300 指数以及创业板指数的对数收益率（LogReturn），相对强弱指标（RSI），乖离率（BIAS），平均真实波幅（BIAS）四种因子来作为隐马尔可夫模型的观测向量。

使用与市场表征有关的指数因子作为隐马尔科夫模型的观测向量预测结果如 5-1 表所示：

表 5-2 使用指数因子作观测向量预测结果

	$P(A)$	$P(B)$	$P(C)$	$P(A\cap B)$	$P(A\cup B)$	$P(A\cap B\cap C)$	$P(A\cup B\cup C)$
胜率	33.3%	24.0%	18.6%	5.5%	51.9.0%	0%	60.2%

相对的，使用每日的各行业板块收益率指数作为观测向量预测结果如表 5-2 所示：

表 5-3 使用日收益率作观测向量预测结果

	$P(A)$	$P(B)$	$P(C)$	$P(A\cap B)$	$P(A\cup B)$	$P(A\cap B\cap C)$	$P(A\cup B\cup C)$
胜率	32.4%	24.5%	16.5%	7.2%	50.1%	0.07%	59.0%

从预测结果来看，使用与市场表征有关的指数因子作为观测向量与直接采用行业板块收益率指数相比，相差并不大，除去一些随机因素，两者预测准确率大致相似。而且在预测过程中我们还发现，前者在匹配难度上还要稍大于后者。经过分析，我认为问题主要存在以下

两点：

（1）对于作为观测向量的因子的选择有问题。我们只是使用了一些常见的表征市场状态的因子，但我们无法确定这些因子是否也能适用于区分市场隐状态。我们应该使用更为科学的手段，以便能挑选出真正合适的因子。

（2）我们建立隐马尔科夫模型，一来是为了市场隐含状态的匹配，二来是进行似然值匹配。直接使用各行业板块的日收益率作为观测向量，可以使似然值相似的交易日拥有较为相似的跟行业板块的市场表现，也更有利于我们对下一阶段上涨概率最大的板块进行预测

5.3.2 模型预测效果分析

将 2007 年 1 月 8 日到 2013 年 12 月 31 日期间的股票各行业板块历史数据放到搭建的模型中去运行，得到预测结果如表 5-4 所示：

表 5-4 模型对历史数据的预测胜率

	$P(A)$	$P(B)$	$P(C)$	$P(A\cap B)$	$P(A\cup B)$	$P(A\cap B\cap C)$	$P(A\cup B\cup C)$
胜率	33.3%	25.8%	31.0%	7.2%	52.0%	2.4%	64.9%

其中 P(C)概率大于 P(B)的主要原因是对于预测候选集行业数目小于 3 的时候进行了对综合权重排行前二的行业的预测准确率进行了额外的判断。

然后本文将预测正确的点用红点表示，错误的则用黄点表示，股市走势用蓝线表示，得到图 5-7、5-8、5-9：

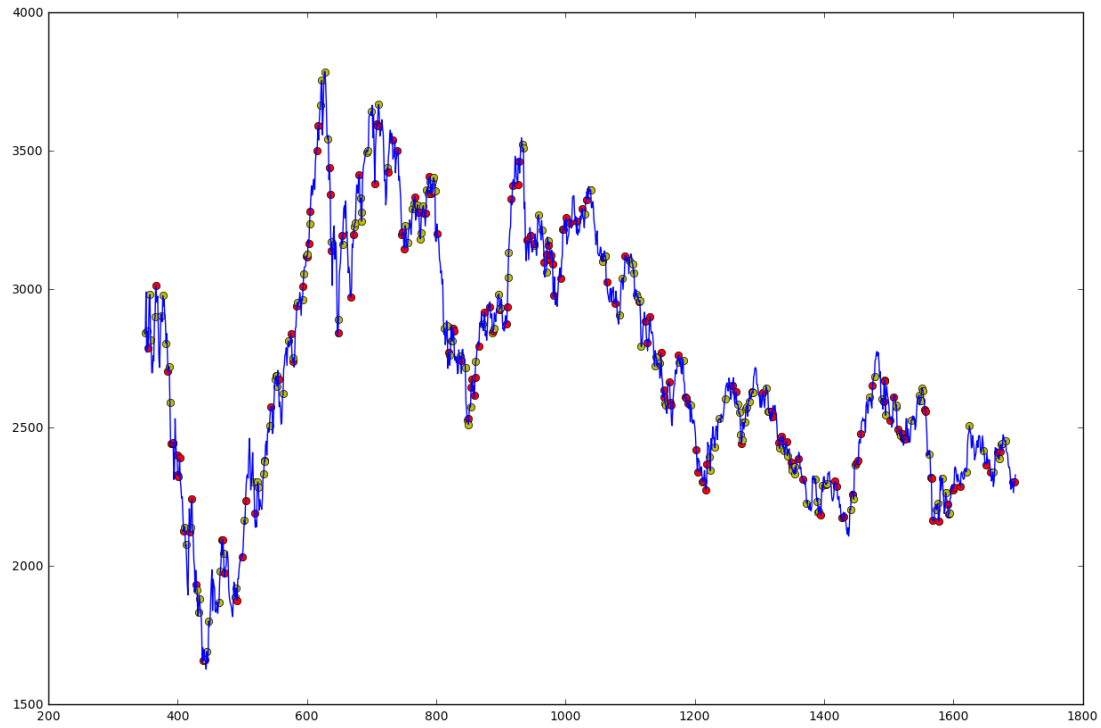

图 5-7 一个候选行业预测结果在股票市场走势图中的

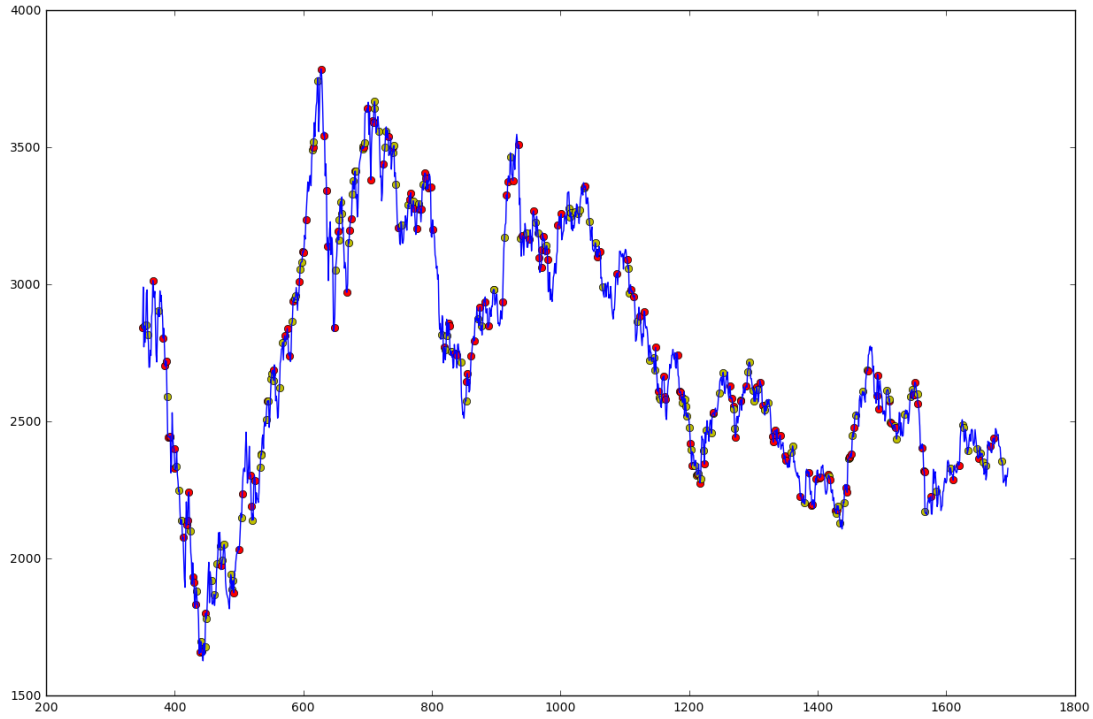

图 5-8 两个候选行业预测结果在股票市场走势图中的标注

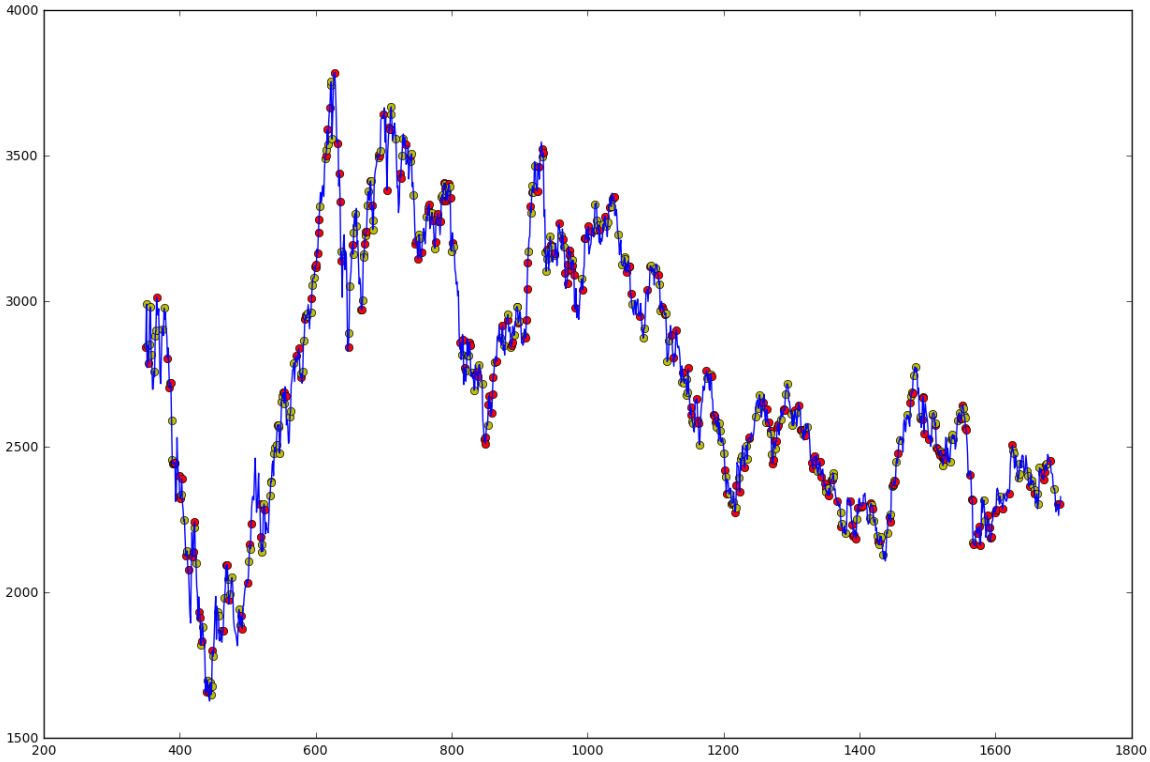

图 5-9 三个候选行业预测结果在股票市场走势图中的标注

在上面三幅图中我们可以看到在实际预测中不论是选取一个候选行业，两个候选行业还是三个候选行业，我们可以看出，本文建立的预测模型的整体表现还是不错的，不论是在“牛市”、“熊市”还是“震荡市”，预测的模型总是能保持一定的稳定性，不是说在熊市就会出现预测准确率大幅下降的情况。而且从后两幅图中我们开可以明显看到，在很多次股市从跌转涨以及从涨转跌的历史上的关键性时刻，模型预测表现相当出色，说明在这些时段的板块轮

动现象更为明显，轮动规律更易把握，模型也能很容易的匹配到合适的历史节点。

5.4 本章小结

本章主要介绍了实际预测过程中的样本的选取、数据的处理、模型参数的选择以及详细的预测步骤。最后还对以两种不同观测向量分别建立预测模型最终效果进行了对比，对模型的预测效果以及跟股市走势关系进行了分析。

第六章 总结与展望

6.1 总结

股票市场的行业板块轮动现象是受到经济周期、国家政策、投资者心理等诸多因素影响而表现出来的一种宏观经济现象，对板块轮动现象的量化建模也存在着板块轮动因果关系难以把握和板块轮动持续时间难以确定两大难点。也因此目前国内对行业轮动的策略研究大都是从基本面分析入手，而缺少一种有效的数量化手段对其进行建模与预测。本文则借助隐马尔科夫模型，试图对行业板块轮动现象进行尽可能合理的描述，利用多年历史数据，探索板块轮动的内在规律。

本文分别采用了中国 A 股市场各行业板块指数每日收益率以及与市场表征有关的常见的指数因子作为两种观测向量建立隐马尔科夫模型，解码出最有可能的隐状态序列，并计算得到再该模型下观测向量对应的似然值序列。通过对市场隐含状态、行业轮动特征以及对数似然值进行层层匹配，从多年的历史数据中找出与当前行为模式最为接近的交易日，进而对下一阶段涨幅较大的行业板块进行预测。

为了使模型进一步拟合股票市场中真实的行业板块轮动现象，本文一是拓宽了行业轮动特征区间，不再仅仅针对每日行业板块收益率最高的板块，而是对上涨幅度排行前三的行业板块都予以行业轮动特征标记，有效的减小了行业轮动特征的匹配难度。二是对最终的行业板块预测提出一种更为符合市场实际情况的加权方式。两种举措都在很大程度上降低了板块轮动引起的行业板块指数上涨程度的不确定性造成的误差影响，提升了预测的准确率。

最终预测结果如前文给出的那样，当我们在二十八个行业板块中给出一个候选行业时，该行业在预测日当天有三分之一的概率能够处于收益率排行前五的行业之中。当给出两个候选行业时，可以得到超过百分之五十的胜率；当给出三个候选行业时，则能达到百分之六十多的胜率。总体来说，本文对行业板块轮动规律的探究表现还是不错的，具有一定的实用性和现实意义，既对基金经理们关于股票的选择给予很好的指示作用，从而进行更优的资产配置，还可以结合现有量化策略，进一步提高策略收益率。

6.2 展望

本文对股票市场行业板块轮动现象的建模与预测是之前国内量化策略中很少涉及的一个研究方向。建立的模型很多地方的参数设置还比较粗糙，仍存在则不足之处，接下来的研究可以从以下几个方面进行改进：

（1）从前人的一些研究中，我们可以得知，不同的行业在板块轮动中的表现是不一样的，有些板块轮动效应比较明显，规律比较容易把握；也有一些板块则基本不受板块轮动效应的影响。而我们在本文中，对 28 个行业并未做区别对待，这样就会产生一定的影响。而且过多的行业也会导致似然值匹配不精确，误差太大。因此在后续的改进中，我们可针对某些特殊板块单独进行行业轮动规律的探究，相信一定会取得更好的效果。

（2）关于隐马尔科夫模型参数选择。在本文中，我们直接选择匹配难度较小且有一定区分度的三种隐含状态。以及对于隐状态的发射概率，我们选择了 spherical 作为高斯分布概率密度函数，即最简单的球面特性。在接下来的研究中我们可以使用更为科学的方法进行对隐状态数量、隐状态发射概率分布以及迭代算法初始值等参数的选择。比如使用 AIC 准

则、BIC 准则或 OEHS 准则等惩罚性似然值检验标准来选择拟合效果最好的模型

（3）对于使用表征市场状态的因子来进行区分市场隐状态，我们可以尝试用更科学的手段选择真正适合的因子，以取得更好的效果

（4）结合金融学知识，对股市行业板块轮动现象进行更好更贴切的描述，毕竟把握行业轮动规律前提是模型能够将行业轮动现象量化出来。

参考文献

[1] 黄智烨. 基于神经网络技术的行业配对量化投资策略研究[D].复旦大学,2012.
[2] Stovall S. Standard & Poor's guide to sector investing[M]. McGraw-Hill, 1995.
[3] Greetham T, Hartnett M. The Investment Clock[R]. Merrill Lynch. 10 November, 2004.
[4] Sassetti P, Tani M. Dynamic asset allocation using systematic sector rotation[J]. The Journal of Wealth Management, 2006, 8(4): 59-70.
[5] Stangl J, Jacobsen B, Visaltanachoti N. Sector rotation over business-cycles[J]. Massey University, 2009.
[6] 何诚颖. 中国股市“板块现象”分析[J]. 经济研究, 2001, 12: 82-87.
[7] 彭艳, 张维. 我国股票市场的分板块投资策略及其应用[J]. 数量经济技术经济研究, 2003 (12): 148-151.
[8] 杜伟锦, 何桃富. 我国证券市场的板块联动效应及模糊聚类分析[J]. 商业研究, 2005 (22): 41-45.
[9] 伍军. 国际对冲基金的行业轮动投资：理论与实践[J]. 世界经济研究所博士学位论文, 2010
[10] 张福芬. 中国股票市场板块轮动的机理研究[J]. 科协论坛 (下半月), 2010, 4: 139-141.
[11] 苏民, 逯宇铎. 我国股市行业轮动现象及相应策略的探讨——从经济周期和货币周期角度出发[J]. 中国证券期货, 2011 (2): 36-40.
[12] 何韬. 证券市场行业轮动问题研究[D]. 苏州大学, 2011
[13] 丁军广. 我国 A 股市场行业轮动规律研究[D]. 西南财经大学, 2011.
[14] 兴业证券. 基于不同市场情境下的行业轮动策略[R]. 行业轮动系列专题报告, 2015
[15] 数库量化组. 基于事件驱动思想的行业轮动模型[R]. 研究报告, 2016
[16] 数库量化组. 基于行业轮动规律的预测策略[R]. 研究报告, 2016
[17] 孙守坤. 基于沪深 300 的量化选股模型实证分析——多因子模型与行业轮动模型的综合运用[D]. 复旦大学, 2013.
[18] 廖田锐. 中国股票市场板块轮动投资策略的实证研究[D]. 对外经济贸易大学, 2011.
[19] 唐心亮, 王靖, 王震洲. 基于马尔科夫链模型的论文格式审查系统[J]. 河北科技大学学报, 2012, 33(5):434-438.
[20] 段康康. 基于隐马尔科夫模型的文本分类器的设计与实现[D]. 北京交通大学, 2016.
[21] 姜文卿. 基于隐马尔科夫模型的股指预测和股指期货模拟交易研究[D]. 上海师范大学, 2013.

谢辞

随着毕业论文的完成，我在交大的四年本科生活也即将接近尾声。四年时光匆匆而过，在这交大四年中，我们经历了很多，也收获了很多、成长了很多。最后能用这样一篇毕业设计论文来为四年的本科生涯画上一个句号，也足以给我们的人生留下深刻的记忆。

在完成毕设的这 5 个月里，我首先需要感谢的就是指导我毕业设计的邓倩妮老师。邓老师为人亲和，在我陷入难题时总是能够积极帮我们进行解答，并给予亲切的鼓励。谢谢邓老师对我给予的信任与鼓励，以及对我们生活上的关心与照顾。

在完成毕设的过程中，我还需要感谢鸣石投资公司给我们提供的帮助。陈总等公司成员不仅为我们提供了良好的工作环境，还不知疲倦的为我们这些对金融证券所知寥寥的门外汉解答一个又一个幼稚的问题。没有邓老师和陈总他们的保驾护航、把握方向，我们也不会顺利的完成毕业设计的内容。再次感谢他们的帮助与指导。

最后我要感谢爸爸妈妈在我状态不好的时候的耐心开导，感谢他们对我的付出与培养；我要感谢和我一起去鸣石做毕业设计的四个小伙伴，感谢他们的陪伴；我要感谢那些科研道路上孜孜不倦的科研工作者们，谢谢他们所做工作对我的毕业设计的帮助；最后，我还要感谢上海交通大学，感谢交大对我们学子的培养与教导。毕业之后，我也要离开校园，踏上社会。但我会始终铭记交大校训“饮水思源，爱国荣校”，始终保持一颗不断学习的心。

A RESEARCH ON QUANTITATIVE MODEL OF SECTOR ROTATION BASED ON BIG DATA

In the stock market, due to the impact of the economic cycle, the national policy and the psychology of investors, the hot spot of market is back and forth between the plates, which we named it the phenomenon of sector rotation. At any time in the stock market, whether it is bull market, bear market or not bull and not bear market, there will be certain sectors of the market performance to outperform the market index, which is the market investment hot spot. However, there is no stock can only rise. As time goes by, those plate which become the hot spot of the market and has the stock price nuch higher than the real market value will gradually return to normal levels. At the same time, because of the existence of the internal relations between different industries, large amount of money withdrawn from the hot plate will also be poured into the next plate. Therefore the hot spot of the stock market is back and forth between the plates, forming the phenomenon of sector rotation.

Although in China's current stock market, the sector rotation phenomenon is quite obvious. And more and more investors use the law of sector rotation to invest in stocks. However, there is still a lack of effective methods to build a quantitative model to grasp the law of sector rotation. Therefore, this paper hopes to use some mathematical models, statistical models and computer technology to establish a suitable quantitative model for sector rotation.

If the we can accurately grasp the law of the sector rotation, and predict the plate with highest rise. Then, we can buy the shares at a lower price at the beginning of the sector rotation and sell the shares promptly before the end of the round. In this way, it is bound to be able to get excess earnings beyond the market.

There are three main reasons for plate rotation. The first is the national policy factors. China's stock market, due to the regulation of the country, does not achieve a real effective market. And the state regulation is one of the most important cause of fluctuation in stock market. The changes in national policy will bring serious impact on the stock market. Because we can not predict the country's policies, this has brought great difficulties to our quantitative work. The second is the industry cycle factor. Different industries have similar industry cycles. The industry cycle can be approximately divided into four stages: introduction period, growth period, maturity stage and decline stage. In the different stages, different industries have the same development. And at the same time of the stock market, different industries are in different stages of the industry cycle. Thus, sector rotation occurs. The third is Psychological changes in investment. As the time goes by, People's investment psychology began to change. As experience grows, investment also becomes more rational.

In this paper, we use hidden Markov model to quantify sector rotation phenomenon. Markov model is a basic model of Natural Language Processing. It has been widely applied to speech processing, speech recognition and text mining in recent years. From the early 90s, HMM began to be applied to image signal processing and video signal processing. So far, HMM and its

extended forms have achieved great success in speech recognition, video content analysis, audio retrieval and other fields.

The hidden Markov model has three assumptions. The first is Markov hypothesis, which means that the hidden state sequence make up a Markov chain. The second is immobility hypothesis, which means that state has no relationship with the specific time. The third is output independence hypothesis, which means that output is only related to the current state.

HMM has three typical problems. The first is evaluation problem. Given a sequenced column, O=O1O2…OT and model parameter λ= (A, B, π), we need to calculate the probability that the model generates the observation sequence O under this parameter, that is P (O| λ). The evaluation problem is commonly solved by the Forward-backward algorithm. The second one is the decoding problem. Given a sequenced column, O=O1O2…OT and model parameter λ= (A, B, π), we need to find the optimal implicit state sequence, Q=q1q2... QT. The decoding problem is commonly solved by the Viterbi algorithm. The third is the learning problem. Given the initial parameters of the model, we need to adjust the model parameters, λ= (A, B, π), to make P (O | λ) maximum with the given observation sequence., O=O1O2…OT. The learning problem is commonly solved by the Baum-Welch algorithm.

The sample data of this research is based on the SWS industry classification standard, which is divided more than three thousand stocks into twenty-eight sectors. Based on historical sector index data, we first deal with daily plate yields, and thus mark the top five in the daily yield. Since we choose 28 industry sectors' yield index as the observation vector of hidden Markov model, we use principal component analysis (PCA) technology to reduce dimension processing for better learning effect.

After data cleaning, we use these processed data as observation to construct hidden Markov model. And then, we can use the hidden Markov model, the parameter of which is calculated, to predicted the implied state sequence of the market and to calculate the likelihood sequence corresponded to the observed value series. By the way, we mark the plate with the largest daily increase and the continuous industry mark will be treated as a feature of sector rotation. By matching the market implied state, the feature of sector rotation and the likelihood value sequence, we can find a trading day which has the closest behavior patterns with the day will be predicted, and then use it to predict the plate with the largest daily increase.

The hidden Markov model is a nonlinear model of the self-reflection of the short-term correlation, so we should not use too large samples for modeling. And the model is used to predict the next short time sequence. In order to improve the model prediction effect and solve the above two problems, this paper decided to adopt a sliding sample window method to model. Every time we dynamically sample T samples and set forecast period is t. Every t days we delete first t days sample data, and also add following t days observation data. We use the model established to predict the plate with the largest daily increase in the following t days, namely every t days the model need an update.

The phenomenon of industry sector rotation in stock market is a macroeconomic phenomenon influenced by the economic cycle, national policy, investors' psychology and many other factors. Therefore there are two aspects of difficulties when we quantify the sector rotation phenomenon. One is that it is difficult to grasp the relation between the plates in sector rotation. The boom of a plate is likely to drive multiple plates at the same time and rise this plates in varying degrees. The logical relationship between them is difficult to judge, and in many cases

there is no clear logical relationship between them. The reason is that the attention of ordinary investors is easily attracted by the best performing plates, creating a herd effect that makes real meaningful continuous plate movements difficult to observe. The other is that the duration of plate rotation is difficult to ascertain. The duration of sector rotation is often associated with the characteristics of the plate itself. Different plates face different length of duration, which will have an bad impact on the judgment of the observer.

In order to improve the quantitative model, two methods are adopted in this paper. First is broaden the industry characteristic interval and no longer only mark the highest daily yield of the plate. This method effectively reduces the difficulty of the match. Second is put forward a better weighting way to industry sector forecast. The two measures have greatly reduced the error effects caused by the uncertainty of index rise caused by plate rotation, and improved the prediction accuracy.

The final prediction results have achieved good effects. The result also confirmed the existence of phenomenon of sector rotation and the effectiveness of the prediction method.