中图分类号：　　学校代号：11845

UDC：　　密级：　　学　　号：2112101234

广东工业大学硕士学位论文

（工学硕士）

标题

张三

指导教师姓名、职称：	李四　教授
校外指导教师姓名、职称：	无
学科（专业）或领域名称：	控制科学与工程
学生所属学院：	自动化学院
答辩委员会主席：	赵六　教授
论文答辩日期：	2020 年 5 月 25 日

A Dissertation Submitted to Guangdong University of Technology

for the Degree of Master

(Master of Engineering Science)

Title

Candidate: ZHANG San

Supervisor: LI Si

May 25, 2020

School of Automation

Guangdong University of Technology

Guangzhou, Guangdong, P. R. China, 510006

摘要

摘要

关键词：关键词 1；关键词 2；关键词 3；关键词 4

ABSTRACT

Abstract

Key words: keywords 1; keywords 2; keywords 3; keywords 4

目录

CONTENTS

第一章　绪论

1.1 课题来源

本课题来源于???

1.2 本课题研究背景及研究意义

当前股票市场瞬息万变，价格变动剧烈。因此，科学的分析市场状态，合理的选择入场时机不必可少。在量化投资领域，有一位传奇人物——詹姆斯 · 西蒙斯 (James Simons)。他在 1988 年成立了大奖章基金 (Medallion Fund)，除了成立的第二年，在 1988 年至 2010 年期间净年均收益率超过 35%，远远超过标普 500 指数的年化收益率。而人们相信背后的秘密模型正是其成立初期的成员之一——莱昂纳多 · 鲍姆 (Leonard Baum) 参与提出的隐马尔可夫模型 (Hidden Markov Model，HMM)，此外统计学中著名的 Baum-Welch 算法可应用在 HMM 训练中也是由他的名字命名。西蒙斯在 2019 年的演讲中表示语音识别和股价预测的过程存在许多相似之处。在对某个词语进行识别时，会把词语的发音分割成一系列连续的音素，换言之一个词语的发音是可以由几个连续的音素组成。在股票市场波动的过程中，也是具有一系列与价格相关的因素参与发挥着影响股价上涨或下跌的作用。比如在左侧上涨行情中能识别出整体上升趋势，从而推测后续继续上涨的概率更大。HMM 模型可以在一系列特征组成的连续时间序列数据中，挖掘到更多内在状态的模式，能更好地拟合描绘当前时间段股价的动态变化。

HMM 最早应用于语音识别[1, 2] 领域，并慢慢发展成为语音识别中最常用的方法之一。随后 HMM 在生物 DNA 序列分析[3, 4]、图像处理[5, 6]、模式识别[7, 8] 等领域也有了成熟的应用。在金融领域中，Hassan[9] 等人早已被证实 HMM 可用于推测市场状态。数据挖掘的技术工具很多，包括 SVM、神经网络、HMM 等，都在各个领域中发挥重大的作用，但也有效果优劣的差别[10, 11]。李嵩松[12]、Badge[13] 等人认为在股价预测中使用 HMM 更加合适，主要考虑的是股价的变化来自于许多未知力量的价量推动，所谓的推动是一个动态的过程，HMM 相比其他工具明显具备优势。因此本课题选用 HMM 模型刻画股价动态变化的过程。

此外，现在有的多数研究和应用当中，更多的是利用了 HMM 模型对股价预测的能力，从而制定择时交易策略，在选股方面的研究较少。孙守坤[14] 在分析 Alpha 收益中认为，要想获取超越市场的收益，大致可以从选股、择时 (包括事件驱动) 以及衍生

品方面进行探讨。考虑到中国市场炒题材、炒板块等行业轮动现象明显[15]，赵静[16]经过实证分析证实了行业轮动策略的有效性，所以在策略的制定中也应充分考虑行业的因素。因此本课题从对市场状态的模式识别出发，利用训练的模型和识别的状态进行选股和择时的混合策略。旨在结合模式识别的思想应用隐马尔可夫模型进行不同行业股票的选取，充分发挥 HMM 在选股盈利方面的能力。

1.3 国内外相关研究现状

1.3.1 国外研究现状

根据最大熵模型理论[17–19]结合贝叶斯网络 Lafferty[20]提出了被 Bhaskaran[21]等人认为有无向图思想的条件随机场 (CRF)，避免了先验观测分布的复杂计算，并且具有更灵活的框架来全面描述观测之间的相关关系，同时保留了 HMM 能够实现状态转移的属性。Zemel[22]提供了单隐状态的情形下多重马氏链传递过程的 HMM 模型，这属于创新型的 HMM 模型，解决了多特征因素合力共同影响传递过程的模型估计问题。Li[23]等研究者根据理论框架结构进一步申明与论述了多特征传递过程的模型结构的属于特征分布的混合高斯分布密度的算法，最终的实证分析结果展示了此类型模型结构比较有意义且有效果的用于多方向多传递者的语音识别模型中。Liporace[24]根据前人总结的离散分布的思想以及其缺陷性和现实世界数据的复杂性提出了连续分布的 HMM 模型。针对连续 HMM 模型，Wellekens[25]针对传递性为相邻的观测点的信息向量给出了一种高斯概率密度函数，从而能够更好的解决观测数据之间具有互相识别的特征信息。Kenny[26]等人对于解释观测序列的相关性特点尝试性的用简单的线性预测模式来对此进行参数化，从而挖掘出相对于单观测序列的简化思维多观测序列具有的更广泛或者所隐藏的联系性信息。单个观察序列来建立模型可能会造成低识别能力，从而降低预测准确度。Rabiner[27]将原始观察序列 O 分成一组 M 个短序列，并得出重估计公式从而进行多观测序列的训练。

隐马尔可夫模型作为机器学习常用的方法之一，最早由 Baum 等[28, 29]提出。隐马尔可夫模型是一种基于统计的概率模型，从可观测序列中推测未知隐含状态的分布，在推测的结果上进行预测或状态解码。在证券交易市场中时间序列无处不在，对于股票、期权、期货等基本数据以外还延伸出各种指标，均为随时间变动的可观测序列数据。而影响这些观测值的内因数据如基本面、政策面、甚至是量化程序交易带来的高

频波动等，基本都是不可观测的。从这一性质考量，研究员陆续把注意力放到隐马氏模型上，对证券市场数据进行分析建模。Zhu[30] 将隐马尔可夫模型 (HMM)、人工神经网络 (ANN) 和粒子群优化 (PSO) 相结合，提出了一种用于预测金融市场行为的融合模型 APHMM 在 APHMM 中，使用 ANN 将每日股价转换为独立的值集，并成为 HMM 的输入。然后利用 PSO 优化 HMM 的初始参数。训练后的 HMM 用于识别和定位历史数据中的相似模式。Hassan 等[31] 也结合人工神经网络 (ANN) 和遗传算法 (GA) 提出 ANN-GA-HMM 模型进行类似的训练过程优化。实对一些股票的预测表明，这些改进的 HMM 算法都是可行的。Yu[32] 选取了 ROA、ROE、销售净利润率、经营活动净收益等因子进行个股差异化排序处理结合所提出的 PRHMM 进行训练，并采用 MAPE、走势准确率指标来比较预测性能。Ingle 等[33] 讨论了公司相关新闻对其股票价格的影响，通过收集网上新闻并对其提取关键词来训练 HMM。最后利用维特比算法 (Viterbi Algorithm) 预测股票的收盘价。Liu 和 Wang[34] 通过建立含有三个隐状态的 HMM 对中国近 10 年来的股市进行分析，实验结果发现通货膨胀、PMI、汇率变化与股市的市场条件相关。

1.3.2 国内研究现状

郭平[35, 36] 等人在高斯混合模型中使用改进的 EM 算法进行参数估计，利用 BIC 准则计算比较出高斯混合成分个数。郭庆[37, 38] 等人依赖于时间的状态转移概率来模型化状态停留时间，用非线性的概率近似公式拟合模型的条件概率函数，修改后的模型称为 MHMM，融合更多的音素跳转信息提高未参加过的训练人语音数据识别率。余文利等[39] 改进了 BIC(Bayesian Information Criterion) 算法用于自动化确定 HMM 隐状态个数。徐朱佳等[40] 利用 k-means 算法对混合高斯模型训练的初始值进行较优定值。

在因子选股的方法论和系统性方面，国内结合机器学习的量化研究时机也基本成熟，且有大量根据国内市场特性的针对性研究。陈亮[41] 基于 SVM 机器学习提出情绪的传播与股票高频交易相关性的研究方法。苏治[42] 等人结合遗传算法优化了带 KPCA 核函数的支持向量回归机，在中长期数据中表现更佳。王淑燕[43] 和焦健[44] 等人使用指标相关性分析方法提出了从六因子到八因子的选股模型指标体系，结合随机森林建立的选股模型在市场表现中也取得成果。李斌[45, 46] 等人设计了结合 SVM、ANN 以及 Adaboost 的混合预测模型，实验选出一系列指标制定量化策略的效果也是非常明显。他

们的大量研究表明结合机器学习的因子选股都优于基础的股票观测特征，并且发现价值相关的特征表现不如动量相关的因子。但在蒋志强[47]的研究中发现在 A 股中流动性的指标相比动量因子的预测能力要强。国信证券[48, 49]基于小市值股投资逻辑提出小市值因子刻画小市值股票池，其策略也获得明显的超额回报，并对动量类因子进行了更全的解析。国内著名的券商机构广发证券[50]创新性地基于 HMM 识别股票上涨和下跌趋势在股票池中进行因子选股也给本文研究提供新的思路。

1.4 本文主要研究工作

第二章　基于高斯混合的隐马尔可夫模型描述

2.1 高斯分布

概率统计模型能够用来研究和揭示随机现象的统计规律，在许多领域中有着广泛的应用，包括气象预报、水文预报、生物统计、保险、金融投资等领域。当中有着许多著名的数学分布如：伯努利分布、二项式分布、几何分布、泊松分布、伽玛分布以及高斯分布等。其中，高斯分布 (Gaussian Distribution)，又称正态分布 (Normal Distribution)，其一维概率密度函数表示如下：

$$p(x,\mu,\sigma^2)=\frac{1}{\sqrt{2c}\sigma}\exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} \tag{2.1}$$

这里，μ 是均值，σ 是标准差。在一维高斯分布的基础上，可定义两个独立一维正态分布随机变量的二维高斯分布，其概率密度函数为：

$$\begin{aligned}p(x_1|(\mu_1,\sigma_1^2),x_2|(\mu_2,\sigma_2^2)) &= \frac{1}{\sqrt{2c}\sigma_1\sigma_2}\exp\left\{-\frac{1}{2}\left[\frac{(x_1-\mu_1)^2}{\sigma_1^2}+\frac{(x_2-\mu_2)^2}{\sigma_2^2}\right]\right\} \\ &= p(x_1,\mu_1,\sigma_1^2)p(x_2,\mu_2,\sigma_2^2)\end{aligned} \tag{2.2}$$

同样地，有 D 维高斯分布的概率密度函数

$$\begin{aligned}p(\boldsymbol{x},\boldsymbol{\mu},\boldsymbol{\Sigma}) &= \frac{1}{(2c)^{D/2}}\frac{1}{\sqrt{|\boldsymbol{\Sigma}|}}\exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})^T\right\} \\ &= \prod_{n=1}^{D}p(x_n,\mu_n,\sigma_n^2)\end{aligned} \tag{2.3}$$

其中，$\boldsymbol{\Sigma}$ 是大小为 $D\times D$ 的协方差矩阵，$|\boldsymbol{\Sigma}|$ 表示为 $\boldsymbol{\Sigma}$ 的行列式。对于给定观察序列 $x=(x_1,x_2,...,x_D)$，根据公式 (2.3)，可计算其高维概率密度。此外，对于多维高斯在本文的应用在于拟合多维的特征向量，D 维数应与特征因子数相同。为了简化计算，便于计算机处理防止精度溢出，一般采用对数概率 (似然率)。

2.2 最大期望与高斯混合模型

2.2.1 最大期望算法

最大期望算法 (Expectation-maximization，EM)，又译期望最大化算法，在统计中被用于寻找依赖不可观察的隐性变量的概率模型中，参数的最大似然估计。在统计计算中，EM 算法是在概率模型中寻找参数最大似然估计或者最大后验估计的算法，其中

概率模型依赖于无法观测的隐性变量。最大期望算法经常用在机器学习和计算机视觉的数据聚类领域。EM 算法的标准计算框架由 E 步和 M 步交替组成，算法的收敛可以确保迭代至少逼近局部极大值。EM 算法被广泛应用于处理数据的缺失值，以及很多机器学习算法，包括高斯混合模型和隐马尔科夫模型的参数估计。EM 算法是高斯混合模型 (Gaussian mixture model, GMM) 的基础。因此，在本小节中，介绍 EM 算法的基本步骤与原理。

EM 算法包含两步计算，分别称为 E 步计算与 M 步计算。给定的 m 个训练样本 $\{x(1), x(2), \ldots, x(m)\}$，样本间独立，为找出样本的模型参数 θ，有如下极大化模型分布的对数似然函数：

$$\theta = \arg\max_{\theta} \sum_{i=1}^{m} \ln\left(P\left(x^{(i)}; \theta\right)\right). \tag{2.4}$$

假定样本数据中存在隐含数据 $z = \{z(1), z(2), \ldots, z(k)\}$，此时极大化模型分布的对数似然函数修正为：

$$\begin{aligned} l(\theta) &= \arg\max_{\theta} \sum_{i=1}^{m} \ln\left(P\left(x^{(i)}; \theta\right)\right) \\ &= \arg\max_{\theta} \sum_{i=1}^{m} \ln\left(\sum_{z=1}^{m} P\left(x^{(i)}, z^{(i)}; \theta\right)\right). \end{aligned} \tag{2.5}$$

z 是隐随机变量，为方便进行参数估计，令 z 的分布为 $Q(z;\theta)$，并且 $\mathrm{Q}(z;\theta) \geq 0$。然后，计算 $l(\theta)$ 的下界，求该下界最大值，重复该过程，直到收敛到局部最大值，即

$$\begin{aligned} l(\theta) &= \sum_{l=1}^{m} \ln \sum_{z} p(x, z; \theta) \\ &= \sum_{i=1}^{m} \ln \sum_{z} Q(z;\theta) \cdot \frac{p(x, z; \theta)}{Q(z;\theta)} \\ &= \sum_{i=1}^{m} \ln\left(E_Q\left(\frac{p(x, z; \theta)}{Q(z;\theta)}\right)\right) \\ &\geq \sum_{i=1}^{m} E_l\left(\ln\left(\frac{p(x, z; \theta)}{Q(z;\theta)}\right)\right) \\ &= \sum_{i=1}^{m} \sum_{z} Q(z;\theta) \ln\left(\frac{p(x, z; \theta)}{Q(z;\theta)}\right), \end{aligned} \tag{2.6}$$

其中，根据詹森不等式的特性，当 $\frac{p(x,z;\theta)}{Q(z;\theta)} = c$ 时，$l(\theta)$ 函数可取得等号。此时，

$$\begin{aligned} l(\theta) &\geq \sum_{i=1}^{m} E_Q \left(\ln \left(\frac{p(x,z;\theta)}{Q(z;\theta)} \right) \right) \\ \sum_z Q(z;\theta) &= 1 \\ Q(z,\theta) &= \frac{p(x,z;\theta)}{c} = \frac{p(x,z;\theta)}{c \cdot \sum_{z'} Q\left(z^i;\theta\right)} = \frac{p(x,z;\theta)}{\sum_{z'} c\left(z';\theta\right)} \\ &= \frac{p(x,z;\theta)}{\sum_{z'} p\left(x,z';\theta\right)} = \frac{p(x,z;\theta)}{p(x;\theta)} = p(z|x;\theta). \end{aligned} \tag{2.7}$$

最后得到参数估计函数如下：

$$\begin{aligned} l(\theta) &= \arg\max_{\theta} l(\theta) = \arg\max_{\theta} \sum_{i=1}^{m} \sum_{z} Q(z;\theta) \ln \left(\frac{p(x,z;\theta)}{Q(z;\theta)} \right) \\ &= \arg\max_{\theta} \sum_{i=1}^{m} \sum_{z} Q(z|x;\theta) \ln \left(\frac{p(x,z;\theta)}{Q(z|x;\theta)} \right) \\ &= \arg\max_{\theta} \sum_{i=1}^{m} \sum_{z} Q(z|x;\theta) \ln(p(x,z;\theta)). \end{aligned} \tag{2.8}$$

根据公式 (2.8)，EM 算法流程如下：

表 2-1 EM 算法流程

Table 2-1 Processing of EM algorithm

EM 算法	
初始化参数值:	随机初始化模型参数 θ 的初始值 θ^0
算法迭代处理:	1. E 步——计算联合分布的条件概率期望:
	$l(\theta) = \sum_{i=1}^{m} \sum_{z} Q(z|x;\theta^j) \ln(p(x,z;\theta^j))$
	2. M 步——极大化估计函数，得到 θ^{j+1}，$\theta^{j+1} = \arg\max l(\theta)$
	3. 判断——如果 θ_{j+1} 收敛，则算法结束，输出参数，否则继续迭代处理

2.2.2 高斯混合模型

高斯混合模型 (Gaussian mixture model, GMM) 在许多领域中广泛应用，如模式识别、图像分割、数据压缩和异常检测等。它能够对复杂的数据结构进行建模，并且具有良好的灵活性和表示能力。通过适当选择模型的参数和聚类数量，可以有效地捕捉数据的分布特征，并提供有关数据集的有用信息。GMM 是一种常用的概率模型，用于对

数据进行建模和聚类分析。它基于高斯分布的概念，将数据集表示为多个高斯分布的加权组合。每个高斯分布代表一个聚类中心或模式，加权表示每个聚类的重要性。模型的训练过程常使用 EM 算法进行参数估计。

假定 GMM 由 N 个 Gaussian 分布线性叠加而成，那么概率密度函数如下:

$$p(x) = \sum_{k=1}^{N} p(k)p(x|k) = \sum_{k=1}^{N} c_k p(x; \mu_k, \Sigma_k), \tag{2.9}$$

c_k 表示第 k 个高斯分布的权重，满足 $\sum_{k=1}^{K} c_k = 1$。$p(x; \mu_k, \Sigma_k)$ 是 D 维高斯分布的概率密度函数 (如公式 (2.3)所示)，μ_k 是第 k 个高斯分布的均值向量，Σ_k 是对应的协方差矩阵。那么，公式 (2.9)所对应的对数似然函数为:

$$l(c, \mu, \sigma) = \sum_{i=1}^{N} ln(\sum_{k=1}^{k} c_k p(x^i; \mu_k, \Sigma_k)). \tag{2.10}$$

在估计高斯混合模型参数中，需事先设定高斯模型个数 N。为此，在算法求解之初，我们采用 K-means 聚类算法获取 N 值，K-means 聚类算法步骤如下:

表 2-2 K-means 算法流程

Table 2-2 Processing of K-means algorithm

K-means 算法	
初始化:	把训练数据（特征向量）平均分配 N 组，计算每组的高斯均值 μ_k
最近邻分类:	针对每个特征向量 x_n，通过计算欧式距离， 寻找与之最近的第 k 个高斯分布，并把该特征向量归于此高斯分布
更新中心点:	通过求平均，更新每个分布的中心点，得到对应高斯的均值
迭代:	重复步骤 2 和 3，直到整体的平均距离低于预设的阈值。

对于所求参数均值向量 μ_k、协方差矩阵 Σ_k 以及权重 c_k 的迭代，构造 GMM 的似然函数 (公式 (2.10))。根据权重系数 c_k 所满足的条件，加入拉格朗日算子 λ，分别对 μ_k，Σ_k，c_k 求最大似然函数。

$$l(\mathbf{x} \mid c, \mu, \Sigma, \lambda) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{N} c_k p\left(x_n \mid \mu_k, \Sigma_k\right) \right\} + \lambda \left(\sum_{k=1}^{N} c_k - 1 \right) \tag{2.11}$$

因难以对上述似然函数求取最大值，此时，我们就可借助上小节所述的 EM 算法

来估计参数值。首先，计算 EM 算法中 E 步概率与 M 步似然函数

$$\begin{aligned} w_j^{(i)} &= Q_i\left(z^{(i)}=j\right)=p\left(z^{(i)}=j|x^{(i)};c,\mu,\Sigma\right) \\ l(c,\mu,\Sigma) &= \sum_{i=1}^{m}\sum_{z^n} Q_i\left(z^{(i)}\right)\ln\left(\frac{p\left(x^{(i)},z^{(i)};c,\mu,\Sigma\right)}{Q_i\left(z^{(i)}\right)}\right) \\ &= \sum_{i=1}^{m}\sum_{j=1}^{k} Q_i\left(z^{(i)}=j\right)\ln\frac{p\left(x^{(i)}|z^{(i)}=j;\mu,\Sigma\right)\cdot p\left(z^{(i)}=j;c\right)}{Q_i\left(z^{(i)}=j\right)} \\ &= \sum_{i=1}^{m}\sum_{j=1}^{k} w_j^{(i)}\ln\frac{\left(\frac{1}{(2c)^{\frac{n}{2}}|\Sigma_j|^{\frac{1}{2}}}e^{\frac{-\frac{1}{2}(x^{(i)}-\mu_j)^T\cdot(x^{(i)}-\mu_j)}{\Sigma_j}}\right)\cdot c_j}{w_j^{(i)}}. \end{aligned} \tag{2.12}$$

然后，对均值和方差求偏导，如公式 (2.13)所示。最后，使用拉格朗日乘子法求解。

$$\begin{aligned} \frac{\partial l}{\partial \mu_l} &= -\frac{1}{2}\sum_{i=1}^{m} w_l^{(i)}\left(x^{(i)^T}\Sigma_l^{-1}x^{(i)}-x^{(i)^T}\Sigma_l^{-1}\mu_l-\mu_l^T\Sigma_l^{-1}x^{(i)}+\mu_l^T\Sigma_l^{-1}\mu_l\right) \\ &= \frac{1}{2}\sum_{i=1}^{m} w_l^{(i)}\left(\left(x^{(i)^T}\Sigma_l^{-1}\right)^T+\Sigma_l^{-1}x^{(i)}-\left(\left(\Sigma_l^{-1}\right)^T+\Sigma_l^{-1}\right)\mu_l\right) \\ &= \sum_{i=1}^{m} w_l^{(i)}\left(\Sigma_l^{-1}x^{(i)}-\Sigma_l^{-1}\mu_l\right), \\ \frac{\partial l}{\partial \mu_l} &= 0\rightarrow\mu_l=\frac{\sum_{i=1}^{m} w_l^{(i)}x^{(i)}}{\sum_{i=1}^{m} w_l^{(i)}}, \\ \frac{\partial l}{\partial \Sigma_l} &= \frac{1}{2}\sum_{i=1}^{m} w_i^{(i)}\left(\Sigma_l-\left(x^{(i)}-\mu_j\right)\left(x^{(i)}-\mu_j\right)^T\right) \\ \frac{\partial l}{\partial \Sigma_l} &= 0\rightarrow\Sigma_l=\frac{\sum_{i=1}^{m} w_l^{(i)}\left(x^{(i)}-\mu_l\right)\left(x^{(i)}-\mu_l\right)^T}{\sum_{i=1}^{m} w_l^{(i)}} \end{aligned} \tag{2.13}$$

高斯混合模型算法流程如表 2-3。

2.3 隐马尔可夫模型描述

隐马尔可夫模型（Hidden Markov Model，HMM）是一种描述具有未知参数的马尔可夫过程的统计模型，常用于对具有隐含状态的序列数据进行建模。该模型包含两个随机序列，即不可观测的状态序列 (隐藏状态序列) 和可观测序列。状态序列可以影响观测序列，但无法直接观察或获取。在研究市场状态和行为时，市场的状态就是隐藏序列。可观测序列在状态序列的影响下可以直接观察或获取。在研究市场状态和行为时，可以从市场中获取的各种历史数据，即可观测序列。在 HMM 中，隐藏序列 J 的

表 2-3 GMM 算法流程

Table 2-3 Processing of GMM algorithm

高斯混合算法	
初始化:	定义高斯数 K，采用 K-mean 算法，对每个高斯设置 μ_k， Σ_k,c_k 的初始值
E 步:	根据当前的 μ_k, Σ_k, c_k, 计算后验概率 $\gamma(n,k)=\frac{c_k p(x_n\|\mu_k,\sum_k)}{\sum_{k=1}^{K} c_k p(x_n\|\mu_k,\sum_k)}$
M 步:	根据 E 步中计算的 $\gamma(n,k)$，更新 μ_k,Σ_k,c_k
计算对数似然函数:	$\ln p(\mathbf{x}\mid c,\mu,\Sigma)=\sum_{n=1}^{N}\ln\left\{\sum_{k=1}^{K} c_k p\left(x_n\mid\mu_k,\Sigma_k\right)\right\}$
迭代:	求解，检查对数似然函数是否收敛，若不收敛，则返回第 2 步。

形式为：$\boldsymbol{J}=(j_1,j_2,\ldots,j_T)$，每个隐含状态都对应着不同的概率值 q_k，且 $\sum_{k=1}^{N} q_k=1$。可观测序列 O 的形式为：$O(o_1,o_2,\ldots,o_T)$，每个观测状态 o_k 也对应着不同的值 v_k，且 $\sum_{k=1}^{N} v_k=1$。

在 HMM 中，隐藏马尔可夫链可随机生成一个状态序列 $\boldsymbol{I}$，然后从状态序列的每个状态生成观测序列 O。HMM 可以用初始概率分布 π、状态转移矩阵 $\boldsymbol{A}$ 和发射矩阵 $\boldsymbol{B}$ 来表示。其中，初始概率分布 π 表示不可观测状态的初始概率分布，即 $\pi=(p_1,p_2,\ldots,p_N)$。状态转移矩阵 $\boldsymbol{A}$ 表示在时间 t 时隐藏状态为 q_m 的情况下，时间 $t+1$ 时隐藏序列为 q_n 的概率：$\boldsymbol{A}=[a_{mn}], a_{mn}=Pj_{t+1}=q_n|j_t=q_m$。发射矩阵 $\boldsymbol{B}$ 表示在时间 t 时隐藏状态为 q_n 的情况下，观测序列为 v_k 的概率：$\boldsymbol{B}=[b_{nk}], a_{nk}=Po_t=v_k|j_t=q_n$。

对于 HMM，存在两个重要的假设：第一个是齐次马尔可夫假设，这个假设指的是隐藏状态序列满足马尔可夫性质，即隐藏序列在任意时刻的状态只与前一个状态的隐藏序列状态有关，独立于其他因素；第二个是观测独立假设，表示在任意时刻，观测序列的状态只与该时刻的隐藏状态有关，独立于其他因素。

2.4 高斯混合-隐马尔可夫模型的建立与训练

HMM 的三个基本问题包括：

1. 评估问题（Evaluation）：给定一个模型和观测序列，如何计算给定观测序列的概率。

2. 解码问题（Decoding）：给定一个模型和观测序列，找到最可能的对应状态序列。

3. 学习问题（Learning）：给定观测序列，估计模型的参数，包括状态转移概率和观测概率。

怎么解决这三个问题

在 GMM-HMM 中，每个隐含状态 j 表示为若干函数 $p(o_t)$ 的线性组合，$p(o_t)$ 是连续高斯概率密度函数，则可表示成 $b_j(o_t)=\sum_{k=1}^{N} c_{jk}p(o_t \mid \mu_{jk},\Sigma_{jk})$。

模型的训练主要是一个参数重估计的过程，需要重估的参数有：起始概率、转移概率、各状态中不同高斯混合概率密度函数（pdf）的权重以及各状态中不同 pdf 的均值和方差。当观察概率为混合高斯分布形式，即 $b_j(o_t)=\sum_{k=1}^{K} c_{jk}N(o_t \mid \mu_{jk},\Sigma_{jk})$，定义中间变量

$$\begin{aligned}\gamma_t^c(j,k)&=\left[\frac{\alpha_t(j)\beta_t(j)}{\sum_{j=1}^{N}\alpha_t(j)\beta_t(j)}\right]\left[\frac{c_{jk}p\left(\mathbf{o}_t^c,\mu_{jk},\Sigma_{jk}\right)}{\sum_{k=1}^{N}c_{jk}p\left(\mathbf{o}_t^c,\mu_{jk},\Sigma_{jk}\right)}\right]\\&=\begin{cases}\frac{1}{P(\mathbf{0}|\lambda)}\pi_j\beta_1(j)c_{jk}p\left(o_1^c,\mu_{jk},U_{jk}\right), & t=1\\ \frac{1}{P(\mathbf{0}|\lambda)}\sum_{i=1}^{N}\alpha_{t-1}(i)a_{ij}\beta_t(j)c_{jk}p\left(o_t,\mu_{jk},U_{jk}\right), & t>1.\end{cases}\end{aligned} \tag{2.14}$$

四个参数的估计均跟前向和后向概率关联，是一种软判决，即分配到状态的观察值多少使用概率来调节的。这样的分配机制跟普通 HMM 的训练过程是类似的，即均为 Baum-Welch 算法。参数重估后有以下形式：

$$\begin{aligned}a_{ij}&=\frac{\sum_{c=1}^{C}\sum_{t=1}^{T_c-1}\xi_t^c(i,j)}{\sum_{c=1}^{C}\sum_{t=1}^{T_c-1}\gamma_t^c(i)}\\c_{jk}&=\frac{\sum_{c=1}^{C}\sum_{t=1}^{T_c}\gamma_t^c(j,k)}{\sum_{k=1}^{K}\sum_{c=1}^{C}\sum_{t=1}^{T_c}\gamma_t^c(j,k)}\\\mu_{jk}&=\frac{\sum_{c=1}^{c}\sum_{t=1}^{T_c}\gamma_t^c(j,k)\mathbf{o}_t^c}{\sum_{c=1}^{C}\sum_{t=1}^{T_c}\gamma_t^c(j,k)}\\\Sigma_{jk}&=\frac{\sum_{c=1}^{c}\sum_{t=1}^{T_c}\gamma_t^c(j,k)\left(\mathbf{o}_t^c-\mu_{jk}\right)\left(\mathbf{o}_t^c-\mu_{jk}\right)'}{\sum_{c=1}^{C}\sum_{t=1}^{T_c}\gamma_t^c(j,k)}\end{aligned} \tag{2.15}$$

其中，C 为训练条数，即进行多观测序列训练，Rabiner 通过将原始观测序列 O 分成一组 M 个短序列。假设每一个观测序列是相互独立的，要最大化的概率函数 $P(O|\mu)$ 变成：$P(O|\mu)=\prod_{m=1}^{M}P(O^{(m)} \mid \mu)$，在 $O^{(m)}$ 序列分割的基础上对上述参数再次进行重估计。

2.5 参数选取准则

2.5.1 AIC 与 BIC 准则

AIC(Akaike Information Criterion) 是一种常用的参数选择准则，用于在统计模型中选择最佳模型。AIC 准则基于信息论的概念，旨在平衡模型的拟合优度和模型的复杂度。其计算公式如下：

$$\mathrm{AIC} = -2ln(L) + 2k, \tag{2.16}$$

其中，L 是模型的最大似然估计值，k 是模型的参数个数。AIC 准则的基本思想是，通过将模型的最大似然估计值和参数个数同时考虑，来选择最佳的模型。在 AIC 准则中，模型的最大似然估计值越高表示模型对数据的拟合程度越好，而参数个数越多表示模型越复杂。AIC 准则通过将模型的最大似然估计值和参数个数进行权衡，惩罚参数个数较多的模型，以避免过度拟合。较小的 AIC 值表示模型能够更好地平衡拟合优度和模型复杂度。需要注意的是，AIC 准则仅可以用于比较同一类模型之间的相对优劣，而不能用于判断单个模型的绝对优劣。因此，在使用 AIC 准则时，需要将不同模型的 AIC 值进行比较，并选择具有最小 AIC 值的模型作为最佳模型

BIC(ayesian Information Criterion) 准则类似于 AIC 准则，用于在统计模型中选择最佳模型。BIC 准则基于贝叶斯统计学的概念，旨在平衡模型的拟合优度和模型的复杂度。其计算公式如下：

$$\mathrm{BIC} = -2ln(L) + kln(n) \tag{2.17}$$

其中，n 是样本的大小。BIC 准则与 AIC 准则类似，都是通过将模型的拟合优度和模型复杂度进行权衡来选择最佳模型。然而，BIC 准则在惩罚参数个数方面比 AIC 准则更为严格。BIC 准则中的参数惩罚项 $kln(n)$ 中的 $ln(n)$ 部分对参数个数的惩罚更大，特别是在样本较小的情况下。相较于 AIC 准则，BIC 准则更加倾向于选择更简单的模型。在 BIC 准则中，模型的最大似然估计值越高表示模型对数据的拟合程度越好，而参数个数越多表示模型越复杂。通过引入 $ln(n)$ 这一惩罚项，BIC 准则更加强调模型的复杂度，以避免过度拟合。较小的 BIC 值表示模型能够更好地平衡拟合优度和模型复杂度。与 AIC 准则类似，BIC 准则也仅用于比较同一类模型之间的相对优劣，并不能用于判断单个模型的绝对优劣。因此，在使用 BIC 准则时，需要将不同模型的 BIC 值进行比

较，并选择具有最小 BIC 值的模型作为最佳模型。

2.5.2 交叉检验准则

交叉验证 (Cross-validation) 准则是在交叉验证过程中使用的评估指标或准则，用于选择最佳的模型或调整模型的超参数。常见的交叉验证准则包括：留一交叉验证、K 折交叉验证、Stratified K 折交叉验证以及 Bootstrap 法。最常用的交叉检验方法是 K 折交叉验证，即将数据集分成 K 个大小相等的子集（折），依次将每个子集作为验证集，剩余的 K-1 个子集作为训练集，重复这个过程 K 次。最后将 K 次的评估结果取平均作为模型的性能评估。K 折交叉验证是一种常用且计算成本相对较低的交叉验证方法。在交叉验证过程中，使用交叉验证准则对模型进行评估，可以帮助选择最佳的模型或调整模型的超参数。常见的交叉验证准则包括准确率、均方误差、对数损失、F1 分数等。

对于高斯混合-隐马尔可夫模型 (GMM-HMM) 中的隐含状态数量 Y_h 的选择，可使用交叉检验方法寻找最优状态个数。假设 Y_h 的数量有 r 个，依据交叉检验，产生 r 个 GMM-HMM 模型，即 $\{\mathrm{GMM-HMM}_s\}_{s=1,\ldots,r}$。对数据进行 K 折拆分，计算每个 GMM-HMM 模型的平均似然值 L，似然值最小的模型则为隐含数量参数最优的模型。

2.6 变量与参数指代说明

在本文所建立的研究对象中，一部分参数用“变量”指代，而另一部分则用“参数”指代。这种指代区别方式与程序设计思维相似，变量在程序运行前主动声明且赋值。而参数则表明输入到特定函数、模型中进行运行计算结果，得到相应的强关联的输出。

用“变量”指代参数，说明此参数设定的可解释性较弱或者其设定大小数值依托于实际的策略表现，与数理推导过程不太相关。此类参数一般为隐含状态个数的选取、特征因子数的设定以及特征因子构建方式等，因其主观性设定占比较大也能理解为对问题的描述以及研究环境范围的主观限定。

用“参数”指代的参数，一般指模型训练或其原理所涉及数理推导公式中较为严谨核心的参数变量，是由计算所得而非主观设定。如本章第 3 小节所示，混合高斯模型训练时所需初始化的参数为〖均值矩阵 μ_k、协方差矩阵 Σ_k、混合高斯成分的权重分配 c_k，均由前述 K-means 算法训练计算所得。而隐马尔可夫模型中的状态转移矩阵同

样作为“参数”，由 EM 算法训练所得。这些“参数”均在本文程序实验中模型训练部分所求得，也是程序运行时间占比的最大构成部分。

2.7 本章小结

第三章　量化策略建立所涉及的基本概念

3.1 概述

量化投资策略与以往人工决策投资相比，最大的特点是理性。具体表现为依赖大量真实的数据对金融问题进行建模和计算，将策略程序化以后通过模拟盘操作进行数据回测，在实盘中不断验证和改进，进行持续性地迭代，是一套规则固化严格执行的流程，克服人性的贪婪与恐惧的弱点。同时一个完整的投资体系包括选择标的、择时、对冲以及风险控制等，这是主观投资难以系统性地分析和评估的。

在策略收益或者目标中，涉及贝塔 (Beta) 收益和阿尔法 (Alpha) 收益。贝塔 (Beta) 收益，指的是对照标的投入后不进行干预随市场波动产生的收益。阿尔法 (Alpha) 收益，在策略中往往是与对照标的的相比超过或少于的那部分收益。而量化策略所追求的目标之一正是最大化阿尔法收益，追求超越市场波动以外的收益也是量化的意义所在。

在股票的对照标的中，以指数为标的时，策略的工作体现在指数股票池中选出要投资的股票。而对于特定股票或指数等资产，策略的工作则体现在择时买入卖出以及仓位的控制中。整体的策略则是在设定的风险阈值范围内，如控制最大回撤率，建立模型制定选股或择时的迭代优化策略。而在模型的建立中涉及的特征工程对实验效果与预期的一致性是关键且必要的。在量化策略中，特征工程主要的工作包括特征因子的选取以及数据的清洗，都需要依赖行业的方法论设定相应的处理方法，而这些处理方法多样并且效果各异。

本章将围绕上述各个模块结合本文所构建的策略进行详细的论述与实践的说明。

3.2 特征工程

3.2.1 特征因子

借鉴经典多因子选股模型的建立过程，提炼出关于因子的以下几个步骤。一般遵从候选因子的选取、特征因子有效性检验、冗余因子的筛除、因子权重即对收益效果的贡献度、因子迭代替换适应市场风格切换。本节对关键的因子选取以及有效性检验展开详尽策略说明。

3.2.1.1 特征因子选取　Feature factor selection

根据指标的分类，基本分为基本面指标、技术面指标以及其他指标。

基本面指标，描述反映某个股票公司的财务状况，常见的基本面因子是由利润表、资产负债表以及现金流表中的数据直接计算的比率。通过这些财务报表可以构建出无数的财务比率以及财务报表变量的组合，以此来预测股票的收益率。一般分为 6 小类：估值因子、偿债能力因子、营运效率因子、盈利能力因子、财务风险因子以及流动性风险因子。

技术面指标，包括过去的价格、成交量以及其他可获得的金融信息等，技术面因子则由以上数据所构建。最大的优势在于能够持续更新，最新的技术指标最小单位可以是天甚至类似于现价等指标每隔几秒就可以获得。是高频交易策略最重要的因子。

其他指标包括一些流行的经济因子，如 GDP 增速、失业率以及通货膨胀率等宏观经济层面的数据，或者是其他一些事件驱动因子、情绪因子和自建的因子等。

换言之，除了使用市场常用的指标外，还可以根据策略特点基于这些财务数据建立自己的特征因子，不拘泥于通用的因子从而优化策略收益。考虑的因素在于随着市场风格变化以及投资者数量不断增加，有的因子会随着使用率上升以及市场风格切换逐渐失效。因此选取和构建合适的因子以提高模型对市场环境变化的适应能力。

本文基于 HMM 模型构建的策略场景属于中高频交易，换手率较高。考虑模型复杂度和数据时效性主要选取技术面因子进行模型训练，并在此基础上构建独有的 HMM 因子构建后续的交易策略。

3.2.1.2 特征因子有效性检验　Feature factor validity test

IC 法检验。信息系数 (Information Coefficient, 简称 IC)，代表因子预测股票收益的能力。IC 的计算方法是：计算全部股票在调仓周期期初排名和调仓周期期末收益排名的线性相关度 (Correlation)。IC 越大的因子，选股能力就越强。IR：信息比率 (Information Ratio, 简称 IR)= IC 的多周期均值/IC 的标准方差，代表因子获取稳定 Alpha 的能力。整个回测时段由多个调仓周期组成，每一个周期都会计算出一个不同的 IC 值，IR 等于多个调仓周期的 IC 均值除以这些 IC 的标准方差。所以 IR 兼顾了因子的选股能力（由 IC 代表）和因子选股能力的稳定性（由 IC 的标准方差的倒数代表）。

IC 最大值为 1，表示该因子选股 100% 准确，对应的是排名分最高的股票，选出来的股票在下个调仓周期中，涨幅最大；相反，如果 IC 值为-1，则代表排名分最高的股票，在下个调仓周期中，跌幅最大，是一个完全反向的指标。当 IC 的绝对值大于 0.05

时，因子的选股能力较强，当 IR 大于 0.5 时因子稳定获取超额收益能力较强。

在本文中，HMM 模型特征因子的选取以及对模型所构建的 HMM 因子有效性的验证中，也需要进行 IC 值计算。

3.3 选择标的（选股）

本文策略目标旨在获取超越市场波动以外的阿尔法收益，基本的选股逻辑为在沪深 300、上证 50、行业指数等股票池中挑选成分股，以此获得相对指数的超额收益。

3.4 择时

通过构建择时信号进行仓位的调整。本文策略中选股机会出现的同时也带来了择时信号，因此构建一套完整的交易逻辑。

3.5 行业轮动

行业可以大致分为周期性和非周期性行业。行业轮动是一种市场短期趋势的表现形式。对于周期性行业，在一个完整的经济周期中，有些是先行行业，有些是跟随行业。非周期性行业有可选、消费、信息、医药、电信和公用等。相较于周期性行业，规律更不容易把握。研究在一个经济周期中的行业轮动顺序，从而在轮动开始前进行配置，在轮动结束后进行调整，则可以获取超额收益。

本文基于 HMM 模型构建一个完整的选股择时策略的同时，再次利用 HMM 模型的能力制定了行业轮动的策略，从而在指定行业轮动信号出现时，对该行业指数股票池中进行选股择时策略的应用，进一步提高相对整个市场波动的超额收益。

3.6 投资组合评价指标

略的好坏需要我们用历史数据去检验。评价策略的好坏的准则一般有以下几个：年化收益率、最大回撤、夏普比、胜率等。而不同的策略只看这些指标还不够，需要有对比标的。如在行业指数选股策略中，则比较的是超越指数的收益，以及其他市场指数的阿尔法收益。

本文在策略效果评价中遵从这一逻辑，除计算基本评价指标外并进行基准的对比。

3.7 本章小结

第四章　模型与策略构建

4.1 数据清洗

去除空值和修正异常值。本文研究策略基于中国股市数据，对 2015 年 1 月 1 日到 2022 年 12 月 31 日这期间的股票市场进行建模分析。最小数据时间单位设定为每个交易日的收盘时，即剔除非交易日。因收盘前股价与当日收盘价相近，故策略执行当日交易时间设定在收盘前，使策略具有实盘实际操作的可行性。以下选取几个指数/成分股数据作为数据示例，如图 4-1 到 4-4。

4.2 数据处理

取股票因子处理成符合隐马尔可夫相关假设条件及市场有效性的训练数据。为了更好地体现隐马尔可夫模型对股票数据特征的拟合能力和效果，本策略简单利用上述基本数据构建符合市场有效性的特征因子，设定特征因子数为 5，如图 4-5 分别为：

对于构建后的因子值适当处理使其服从正态分布，更符合隐马尔可夫模型对数据的要求。

4.3 模型参数初始化与模型训练

4.3.1 变量选取

隐马尔可夫模型的其中一个特点是初始需要设定的变量少，当前应用中只需设定隐藏状态个数。所谓隐含状态映射到股票市场的情景中，大致归类为牛市、熊市、震荡市的状态。一方面看隐含状态数的设定，若为 5 时，还可以更具体地加入牛市上涨、牛市下跌以及对应的熊市上涨、熊市下跌等状态描述。随着隐含状态数设定的增多，还可以在上涨、下跌的具体程度上作模型解释性的映射描述。但需要注意的基本逻辑是，模型训练所得的隐含状态分布仅为程序对训练数据特征的分类拟合，是一种关于数据的客观描述。而上述所谓市场状态对应则是对不同隐含状态下策略执行的收益表现或相关因子曲线作的人为主观判断。换言之，并不是所有的隐含状态都能在常规的市场状态类别定义中找到映射，而能与真实市场状态匹配的隐含状态具体实验表现可为 0 个、1 个或多个。

由此可见，关于隐含状态数的设定，过少则无法很好地拟合市场状态的变化情况。而过多则模型对往期训练数据的过度拟合，虽然能在市场状态刻画中有深度细分，但

日期	开盘价	收盘价	最高价	最低价	成交量	成交额
2022-08-04	4114.83	4066.98	4150.98	4058.18	11666612400.0	268695834800.0
2022-08-05	4089.44	4101.54	4108.16	4064.66	8953264400.0	216898674900.0
2022-08-06	4109.63	4156.91	4159.48	4097.86	10257100400.0	247325902000.0
2022-08-09	4142.11	4148.07	4155.73	4135.08	9020813800.0	224080812100.0
2022-08-10	4143.67	4156.29	4161.81	4131.43	8413695900.0	214223692300.0
2022-08-11	4149.21	4109.74	4160.79	4092.55	8556583900.0	217447368500.0
2022-08-12	4130.74	4193.54	4193.81	4116.06	12138505100.0	269630998400.0
2022-08-13	4185.42	4191.15	4202.47	4175.05	10680829500.0	219208055300.0
2022-08-16	4180.0	4185.68	4219.92	4174.3	10589849500.0	245539801000.0
2022-08-17	4190.53	4177.84	4208.8	4170.52	9390613900.0	233818584700.0

图 4-1 沪深 300 指数 10 个交易日市场数据

Figure 4-1 ???

日期	开盘价	收盘价	最高价	最低价	成交量	成交额
2022-08-04	10110.32	10130.63	10294.23	10110.32	961919500.0	29738126800.0
2022-08-05	10198.0	10347.04	10350.96	10190.99	816748900.0	29682222600.0
2022-08-06	10383.79	10571.0	10571.22	10383.79	928369600.0	33874290200.0
2022-08-09	10583.63	10515.66	10633.29	10477.83	1095435700.0	34589836800.0
2022-08-10	10481.37	10439.99	10481.37	10365.91	764811300.0	24266195600.0
2022-08-11	10427.62	10269.91	10440.26	10215.53	760256000.0	26499408700.0
2022-08-12	10333.96	10554.43	10562.02	10310.52	916858300.0	32524496400.0
2022-08-13	10534.56	10577.22	10612.66	10469.67	758180400.0	24821378400.0
2022-08-16	10549.07	10427.04	10549.07	10410.48	722601300.0	25209819100.0
2022-08-17	10426.05	10316.32	10437.85	10293.27	699952700.0	25390305700.0

图 4-2 中证医药指数 10 个交易日市场数据

Figure 4-2 ???

日期	开盘价	收盘价	最高价	最低价	成交量	成交额
2022-08-04	326.7	321.99	334.85	318.89	18872311.0	6197191078.0
2022-08-05	328.0	326.8	331.47	322.1	18663677.0	6098858888.0
2022-08-06	330.0	324.49	331.89	318.28	15756385.0	5115620394.0
2022-08-09	322.11	321.35	324.9	319.0	12948786.0	4158491944.0
2022-08-10	321.18	319.72	323.8	318.15	11610661.0	3719885520.0
2022-08-11	319.7	308.35	319.7	305.0	25849901.0	7998810821.0
2022-08-12	311.43	315.02	315.89	305.6	16544481.0	5159742999.0
2022-08-13	315.02	311.93	316.28	310.0	10133763.0	3165436406.0
2022-08-16	312.2	315.21	319.23	311.22	13511926.0	4265356112.0
2022-08-17	315.8	318.92	321.89	315.5	15724331.0	5029761317.0

图 4-3 比亚迪 10 个交易日市场数据

Figure 4-3 ???

日期	市盈率(PE,TTM)	流通股本(万股)	换手率(%)	流通市值(亿元)
2022-08-04	263.0803	9513.6152	1.6024	3806.4507
2022-08-05	261.2207	9446.3672	1.3528	3779.5447
2022-08-08	258.6929	9354.958	1.1117	3742.9712
2022-08-09	257.3808	9307.5059	0.9968	3723.9854
2022-08-10	248.2277	8976.5088	2.2193	3591.5518
2022-08-11	253.5972	9170.6826	1.4204	3669.2415
2022-08-12	251.1097	9080.7275	0.87	3633.2502
2022-08-15	253.7501	9176.2139	1.1601	3671.4546
2022-08-16	256.7367	9284.2168	1.35	3714.6672
2022-08-17	263.99	9546.5107	1.8104	3819.6125

图 4-4 比亚迪 10 个交易日公司数据

Figure 4-4 ???

特征名	特征计算公式
特征1	每日换手率(仅针对个股)
特征2	(每日收盘价-每日开盘价)/每日开盘价
特征3	(每日最高价-每日开盘价)/每日开盘价
特征4	(每日开盘价-每日最低价)/每日开盘价
特征5	(T+1日收盘价 - T日收盘价)/ T日收盘价

图 4-5 特征计算表

Figure 4-5 ???

往往在新数据或策略实盘中表现欠佳，泛化能力较弱也会导致策略不可用。除此之外，随着隐含状态数的增多，模型复杂度变大，模型训练时间也会大幅增加。因此本文对隐含状态数这一参数根据“专家经验”进行范围主观限定，通过 AIC 和 BIC 准则及交叉检验的方法，以预训练的形式比较不同隐藏状态数的实验效果，从而设定适合本策略的具体数值。

4.3.2 GMM-HMM 的模型参数初始化

一般的离散型 HMM 包含的参数有：初始状态分布、隐含状态转移概率矩阵以及观测状态转移概率矩阵。由第二章的数理分析可知，GMM-HMM 最大的特征在于观测状态转移概率矩阵在连续型中，使用了混合高斯分布函数去拟合特定隐含状态到连续观测值的概率映射，而不再使用离散的矩阵来描述。

1. 初始状态分布

初始状态分布即为迭代初始模型 t_0 时刻隐含状态的分布概率。从策略效果角度可根据在特定研究时段中，A 股市场牛熊天数的统计，但很明显的缺陷是对隐含状态数设定的耦合度较高，且主观地把隐含状态与市场状态直接映射，具有较大的不准确性。此外，本文以探究隐马尔可夫模型对股市数据的拟合以及市场状态挖掘的能力，所得市场状态不完全可用证券市场语言准确描述，但对交易策略构建一定是有参考价值的，是属于实际存在的状态但未被常见市场状态所归类。因此本文策略对初始状态分布以平均概率 1/N（N 为隐含状态数）来设定，不加入先验知识，充分考察隐马氏模型对中国股票市场数据的建模能力与表现。

2. 隐含状态转移概率分布

与上述初始状态分布设定类似，在隐含状态转移中，某个状态向其他所有状态的转移概率设为均匀取值 1/N。在后续的滚动训练中数据日期在小范围滚动，直接把上一次训练所得隐含状态转移概率分布值作为下一次训练的初始值，以此提高训练效率。

3. 混合高斯模型参数初始化

GMM 训练时，为达到更好的拟合效果和减小训练模型收敛时间，需要初始化每个状态的混合高斯模型的均值矩阵 μ_k，协方差矩阵 Σ_k，权重 c_k。并且使用上一次所训练的参数结果作为下一次训练的初始值。具体流程如表 2-1。使用 K-mean 聚类算法对数据集进行分类，分类数为高斯混合成份数，分类以后对每个类别下的每个特征数据求

均值和方差，形成初始均值矩阵、协方差矩阵以及统计出初始权重。

4.3.3 GMM-HMM 的模型训练

使用 EM 算法和多观测序列下重估计公式，对 GMM 模型进行训练，在此过程中得到 HMM 的隐藏状态转移矩阵以及所得模型的均值矩阵 μ_k，协方差矩阵 Σ_k，权重 c_k 构造 GMM 的混合连续概率密度函数。对于多观测序列的概念体现为策略数据训练集构建时，以固定值的时间跨度为单位构建出单条多特征的观测序列。在研究时段中由多个时间跨度所得多个单条观测序列构成多观测序列训练数据集。多观测序列数据集的训练方式与常规单观测序列的模型训练方式不同，所以需要对公式进行重估计，推导过程在第二章有详细描述，此处不赘述。

4.4 策略数据逻辑与应用

选取不同指数（如上证 50）或行业（如新能源）成分股作为股票池子，其中可调试参数为观测长度（交易天数 $t-a$ 到 t 的数据）和预测期（$t+b$ 日）。获取每日向前滚动 15 天数据，比对当日与向前第 5 个交易日（设为 t 日）涨跌对比选出上涨或下跌的训练集。比如，比对当日与向前第 5 个交易日（t 日）收盘价为涨，则判断此中有上涨的模式，取 $t-10$ 日到 t 日之间 10 天的数据，作为上涨模型训练数据。把股票池中（如新能源行业）所有具备此状态的训练数据集成上涨模型训练集，下跌模型则使用相反操作。

使用多观测序列方法分别训练出 GMM-HMM 模型的上涨模型和下跌模型。每个成分股按当日向前滚动 10 天数据作为训练数据，输入模型输出给定模型中观测序列（这 10 天观测数据）的概率，设为该行业的 HMM 上涨或下跌因子。对此因子进行 IC 均值分析，对满足的进行成份股因子值分档，按照不同仓位权重进行交易或选取排行靠前的成分股进行交易。回测时按照全仓买入该指数进行长时间段的对比。

4.5 行业轮动模型训练

选取不同行业指数。选取 100 天的横截面数据，使用单高斯的 Gaussian-HMM 模型进行训练。利用 HMM 模型中的解码能力得出 100 天的隐含状态序列。输入近 5 日的横截面数据得出近 5 日的隐含状态序列，在前述所得 100 天的隐含状态序列中匹配相似度最大的子序列。使用向量的相似度获取距离最小的子序列 (此处补充距离算法描述)，回顾在此状态下 b 日（同上，如 5 日）后的状态进行参考。构建行业轮动信号，在

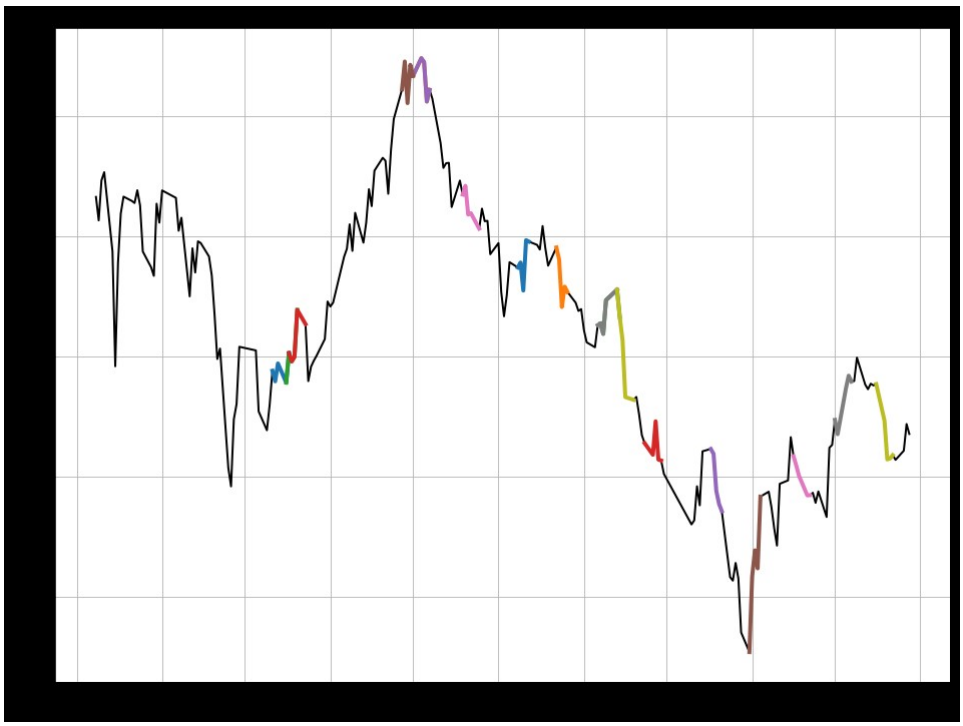

图 4-6 HMM 识别相同市场风格段图

Figure 4-6 ???

所选不同行业中对轮动信号进行排序，对信号明显的行业进一步执行上述选股择时策略。

如图 4-6 示例，对沪深 300 指数 2022 年 12 月 28 日往前 200 个交易日收盘价数据进行 HMM 拟合，利用所训练模型进行隐含状态解码。从右至左每 11 个（此处设定 11 为展示更直观减少相似段重叠）交易日进行一次搜索。如下图在最右边颜色段解码出长度为 5，近 5 日隐含状态向量，从前往后滚动搜索最相似的 5 日隐含状态向量，从而匹配到具体 5 日日期并进行高亮描粗。如图可观测颜色相同两段的相似性基本概括为两个特点，趋势相合以及拐点位置相似，通过不同股票指数的实验可验证 HMM 对股票数据内在状态建立分类的有效性。另外在状态稳定性测试方面，虽然每次模型训练所得隐含状态代表的具体市场实际映射不一定相同，但所匹配的位置是相同的，通过多次重复实验容易验证这一点。

由上述分析启发，除了利用 HMM 此能力进行市场行业风格轮动追踪预测外，把其视为特定行业指数投资的交易信号，也能构建基于 HMM 的择时策略，图 4-7。具体策略表现如图 4-8。

HMM模型	隐含状态数	特征因子	回测开始时间	回测结束时间	仓位变动周期
Gaussian-HMM	4	特征2-5	2020年1月1日	2022年12月28日	5天

图 4-7 HMM 择时策略参数表

Figure 4-7 ???

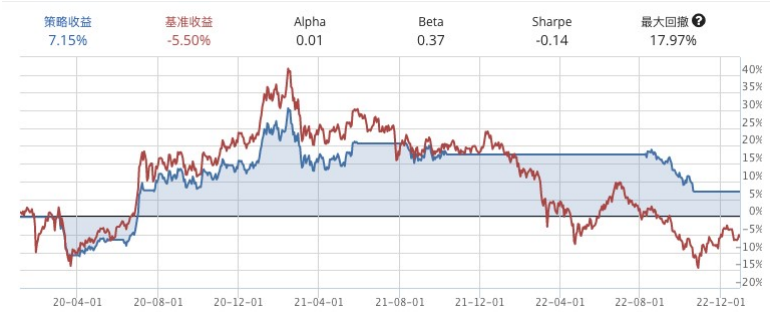

图 4-8 HMM 择时策略收益表现图

Figure 4-8 ???

4.6 本章小结

第五章　实证分析

5.1 因子相关性与 IC 检验

5.2 BIC 准则及交叉检验

5.3 模型的建立

5.4 混合高斯模型参数的初始化

5.5 HMM 因子的构建及 IC 有效性检验

5.6 策略效果回测

5.7 本章小结

结论与展望

研究结论

未来研究展望

参考文献

[1] ELOUAHABI S, ATOUNTI M, BELLOUKI M. Amazigh isolated-word speech recognition system using hidden markov model toolkit (htk)[C]//2016 International Conference on Information Technology for Organizations Development (IT4OD). [S.l.]: IEEE, 2016: 1-7.

[2] HO H, MAWARDI V C, DHARMAWAN A B. Question answering system with hidden markov model speech recognition[C]//2017 3rd International Conference on Science in Information Technology (ICSITech). [S.l.]: IEEE, 2017: 257-262.

[3] LEOS-BARAJAS V, GANGLOFF E J, ADAM T, et al. Multi-scale modeling of animal movement and general behavior data using hidden markov models with hierarchical structures[J]. Journal of Agricultural, Biological and Environmental Statistics, 2017, 22: 232-248.

[4] TOTTERDELL J A, NUR D, MENGERSEN K L. Bayesian hidden markov models in dna sequence segmentation using r: the case of simian vacuolating virus (sv40)[J]. Journal of Statistical Computation and Simulation, 2017, 87(14): 2799-2827.

[5] TAHERKHANI F, HEDAYATI R. An ensemble learning method for scene classification based on hidden markov model image representation[J]. arXiv preprint arXiv:1607.06794, 2016.

[6] AMINI M, SADREAZAMI H, AHMAD M O, et al. Multichannel color image watermark detection utilizing vector-based hidden markov model[C]//2017 IEEE International Symposium on Circuits and Systems (ISCAS). [S.l.]: IEEE, 2017: 1-4.

[7] YIYAN L, FANG Z, WENHUA S, et al. An hidden markov model based complex walking pattern recognition algorithm[C]//2016 fourth international conference on ubiquitous positioning, indoor navigation and location based services (UPINLBS). [S.l.]: IEEE, 2016: 223-229.

[8] LU L, YI-JU Z, QING J, et al. Recognizing human actions by two-level beta process hidden markov model[J]. Multimedia Systems, 2017, 23: 183-194.

[9] HASSAN M R, NATH B. Stock market forecasting using hidden markov model: a new

approach[C]//5th International Conference on Intelligent Systems Design and Applications (ISDA'05). [S.l.]: IEEE, 2005: 192-196.

[10] KIM K J. Financial time series forecasting using support vector machines[J]. Neurocomputing, 2003, 55(1-2): 307-319.

[11] CAO L J, TAY F E H. Support vector machine with adaptive parameters in financial time series forecasting[J]. IEEE Transactions on neural networks, 2003, 14(6): 1506-1518.

[12] 李嵩松. 基于隐马尔可夫模型和计算智能的股票价格时间序列预测 [D]. [出版地不详]: 哈尔滨工业大学, 2011: 37-38.

[13] BADGE J, SRIVASTAVA N. Comparative analysis of arima, fuzzy time series method and hidden markov model for stock market prediction[J]. Fuzzy Systems, 2010, 2(8): 23–27.

[14] 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]. [出版地不详]: 复旦大学, 2013: 7.

[15] 彭惠, 刘欣雨. 基于关联规则的中国股票市场行业轮动现象研究 [J]. 北京邮电大学学报: 社会科学版, 2016: 66-71.

[16] 赵静. 行业轮动多因子选股模型及投资效果实证分析 [D]. [出版地不详]: 东北财经大学, 2015: 1.

[17] SUTTON C, MCCALLUM A, et al. An introduction to conditional random fields[J]. Foundations and Trends® in Machine Learning, 2012, 4(4): 267-373.

[18] ALONSO A, VAN DER ELST W, MOLENBERGHS G. A maximum entropy approach for the evaluation of surrogate endpoints based on causal inference[J]. Statistics in Medicine, 2018, 37(29): 4525-4538.

[19] CARLI F P, CHEN T, LJUNG L. Maximum entropy kernels for system identification[J]. IEEE Transactions on Automatic Control, 2016, 62(3): 1471-1477.

[20] LAFFERTY J, MCCALLUM A, PEREIRA F. Conditional random fields: probabilistic models for segmenting and labeling sequence data[C]//Proceedings of the 8th International Conference on Machine Learning. [S.l.]: Morgan Kaufmann Publishers Inc, 2001: 282-289.

[21] BHASKARAN S K, SREEJITH C, RAFEEQUE P. Neural networks and conditional random fields based approach for effective question processing[J]. Procedia computer science, 2018, 143: 211-218.

[22] ZEMEL R S. A minimum description length framework for unsupervised learning.[D]. Toronto: University of Toronto, 1993.

[23] LI H, LIU Z, ZHU X. Hidden markov models with factored gaussian mixtures densities[J]. Pattern Recognition, 2005, 38(11): 2022-2031.

[24] LIPORACE L. Maximum likelihood estimation for multivariate observations of markov sources[J]. IEEE transactions on information theory, 1982, 28(5): 729-734.

[25] WELLEKENS C. Explicit correlation in hidden markov models for speech recognition[C]//Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing. [S.l.]: San Francisco, CA, USA: IEEE, 1987: 384-387.

[26] KENNY P, LENNIG M, MERMELSTEIN P. A linear predictive hmm for vector valued observations with application to speech recognition[J]. IEEE Transactions on Acoustics, Speech and Signal Processing, 1990, 38(2): 220-225.

[27] RABINER R L. A tutorial on hidden markov models and selected applications in speech recognition[J]. Proceedings of the IEEE, 1989: 258-273.

[28] BAUM L E, PETRIE T. Statistical inference for probabilistic functions of finite state markov chains[J]. Ann.math.stat, 1966, 37: 1554-1563.

[29] BAUM L E, EAGON J A. An inequality with applications to statistical estimation for probabilistic functions of markov processes and to a model for ecology[J]. Bull.amer.math.stat, 1967, 37(3): 360-363.

[30] ZHU J Y, HAI-YAN Y E, GAO Y. Fusion model of hidden markov model for stock market forecasting[J]. Computer Engineering and Design, 2009, 30(21): 4945-4948.

[31] HASSAN M R, NATH B, KIRLEY M. A fusion model of hmm, ann and ga for stock market forecasting[J]. Expert Systems with Applications, 2007, 33(1): 171-180.

[32] YU Y, XIE T, LIU C, et al. A study of stock price behavior based on feature selection and hmm[J]. Information Technology and Network Security, 2018, 3.

[33] INGLE V, DESHMUKH S. Hidden markov model implementation for prediction of stock prices with tf-idf features[C]//Proceedings of the International Conference on Advances in Information Communication Technology & Computing. [S.l.: s.n.], 2016: 1-6.

[34] LIU Z, WANG S. Decoding chinese stock market returns: three-state hidden semi-markov model[J]. Pacific-Basin Finance Journal, 2017, 44: 127-149.

[35] 平郭. 贝叶斯概率图像分割中混合模型参数高效计算的研究 [J]. 计算机科学, 2002, 29(8): 101-103.

[36] 郭平, 卢汉清. 贝叶斯概率图像自动分割研究 [J]. 光学学报, 2002, 22(12): 1479-1483.

[37] 郭庆, 柴海新, 吴文虎. 隐 Markov 模型中状态停留时间的模型化 [J]. 清华大学学报: 自然科学版, 1999, 39(5): 4.

[38] 郭庆, 吴文虎, 方棣棠. 隐马尔可夫模型中一种新的帧相关建模方法 [J]. 软件学报, 1999, 10(6): 631-652.

[39] 余文利, 廖建平, 马文龙. 一种新的基于隐马尔可夫模型的股票价格时间序列预测方法 [J]. 计算机应用与软件, 2010, 27(6): 186-190.

[40] 徐朱佳, 谢锐, 刘嘉, 等. 隐马尔科夫模型的改进及其在金融预测中的应用 [J]. 工程数学学报, 2017(5).

[41] 陈亮. 基于微博舆情的股票高频交易分析技术研究与实现 [D]. [出版地不详]: 复旦大学, 2014: 7-57.

[42] 苏治, 傅晓媛. 核主成分遗传算法与 SVR 选股模型改进 [J]. 统计研究, 2013, 30(5): 9.

[43] 王淑燕, 曹正凤, 陈铭芷. 随机森林在量化选股中的应用研究 [J]. 运筹与管理, 2016, 25(3): 7.

[44] 焦健. 我国上市公司盈利质量评价研究 [D]. [出版地不详]: 南京航空航天大学, 2011.

[45] 李斌, 林彦, 唐闻轩. ML-TEA: 一套基于机器学习和技术分析的量化投资算法 [C]// 中国系统工程学会第 19 届学术年会. [出版地不详: 出版者不详], 2016.

[46] 李斌, 邵新月, 李㊣阳. 机器学习驱动的基本面量化投资研究 [J]. 中国工业经济,

2019(8): 19.

[47] 蒋志强, 田婧雯, 周炜星. 中国股票市场收益率的可预测性研究 [J]. 管理科学学报, 2019, 22(No.178(04)): 97-114.

[48] 张欣慰. 聚焦小盘股-如何构建小市值股票投资策略？[R]. [出版地不详]: 研究报告国信证券, 2022.

[49] 张欣慰. 金融工程专题研究：动量类因子全解析 [R]. [出版地不详]: 研究报告国信证券, 2021.

[50] 罗军. 再探西蒙斯投资之道: 基于隐马尔科夫模型的选股策略研究 [R]. [出版地不详]: 研究报告广发证券, 2018.

攻读学位期间取得与学位论文相关的成果

发表和投稿与学位论文相关学术论文

[1] **张**三, 李四, 王五, 等. Jet electrochemical machining of micro dimples with conductive mask. Journal of Materials Processing Technology. 2018, 257:101-111. (SCI Impact Factor 3.647, WOS:000431161400010)

[2] 李四, **张**三, 王五, 等. Electrochemical direct-writing machining of micro- channel array. Journal of Materials Processing Technology. 2019, 265:138-149. (SCI Impact Factor 3.647, WOS:000451935100014)

申请发明专利

[1] 李四, **张**三, 王五. 一种微流道电解加工装置. 发明专利申请号: 201810467763.5.

学位论文独创性声明

本人郑重声明：所呈交的学位论文是我个人在导师的指导下进行的研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明，并表示了谢意。本人依法享有和承担由此论文所产生的权利和责任。

论文作者签名：　　日期：

学位论文版权使用授权声明

本学位论文作者完全了解学校有关保存、使用学位论文的规定：“研究生在广东工业大学学习和工作期间参与广东工业大学研究项目或承担广东工业大学安排的任务所完成的发明创造及其他技术成果，除另有协议外，归广东工业大学享有或特有”。同意授权广东工业大学保留并向国家有关部门或机构送交该论文的印刷本和电子版本，允许该论文被查阅和借阅。同意授权广东工业大学可以将本学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印、扫描或数字化等其他复制手段保存和汇编本学位论文。保密论文在解密后遵守此规定。

论文作者签名：　　日期：

指导教师签名：　　日期：

致谢