股指与国家经济

摘要

近年来我国资本市场和证券交易规模不断扩大，越来越多的资金投资于证券市场，与此同时市场价格的波动也十分剧烈，而波动作为证券市场中最本质的属性和特征，对于人们风险收益的分析、股东权益最大化和监管层的有效监管有着至关重要的作用。本文针对题中所给的数据进行时间序列分析，通过建立不同的模型，给出了合理的选股方案和投资组合方案，并提出有关投资建议和策略。

针对问题一（1），成交量数据是一种时间序列数据，对数据的几项关键指标进行计算分析，据此选择合适的时间序列模型进行拟合和预测，从而得到较为可信的预测值。

针对问题一（2），通过对成分股数据的聚类分析，每个类别中选择历史成交量排名第一的股票，利用蒙特卡洛模拟法，通过建立误差跟踪模型，分别从价格时间序列角度使投资组合与目标指数的跟踪误差最小、从月化收益率时间序列角度使投资组合与目标指数的超额收益率最大，再将得到的两个权重求取平均值，从而得到稳定性好且具有良好的超额收益率的选股投资组合。

针对问题二：将打分法、仅聚类选股等传统模型设为对照组，从模型功能和模型给出结果等方面，通过选取的评价指标对比，说明问题一中模型的性能。

针对问题三：首先通过股价指数公式进行计算，然后据图分析，选择合适的时间序列模型，通过 SPSS 软件进行一年内的股指预测，通过问题一（2）中的模型，对投资方案进行说明。

，

关键词：ARIMA 模型、ward 聚类分析、蒙特卡罗模拟、误差跟踪模型、时间序列预测分析

一、问题重述

1.1 问题的背景

自 1990 年 12 月 19 日上海证券交易所挂牌成立，经过 30 年的快速发展，中国证券市场已经具有相当规模，在多方面取得了举世瞩目的成就，对国民经济的资源配置起着日益重要的作用。截至 2019 年年底，上海和深圳两个证券交易所交易的股票约 4000 种。目前，市场交易制度、信息披露制度和证券法规等配套制度体系已经建立起来，投资者日趋理性和成熟，机构投资者迅速发展已具规模，政府对证券市场交易和上市公司主体行为的监管已见成效。

随着近年来我国资本市场的发展和证券交易规模的不断扩大，越来越多的资金投资于证券市场，与此同时市场价格的波动也十分剧烈，而波动作为证券市场中最本质的属性和特征，市场的波动对于人们风险收益的分析、股东权益最大化和监管层的有效监管都有着至关重要的作用，因此研究证券市场波动的规律性，分析引起市场波动的成因，是证券市场理论研究和实证分析的重要内容，也可以为投资者、监管者和上市公司等提供有迹可循的依据。

1.2 问题的相关信息

根据题目提供的相关信息，可知如下数据条件：

附件中给出了十支股票的参数，包含了 10 支股票从 2019 年 9 月至 2020 年 3 月每天的开盘收盘价格、最高最低价格和当天的股票成交量。

1.3 需解决的问题

根据上述题目背景及数据，题目要求建立数学模型解决以下问题：

问题一：

（1）在附件数据的分析和处理的过程中，补全附件中缺失的数据。

（2）对于附件的成分股数据，通过建立模型，给出合理选股方案和投资组合方案。

问题二：给出合理的评价指标来评估问题一中的模型，并给出分析结果。

问题三：通过附件股指数据和补充的数据，对当前的指数波动和未来一年的指数波动进行合理建模，并给出合理的投资建议和策略。

二、问题分析

2.1 问题一（1）的分析

问题一（1）要求分析处理附件数据，并补全附件中缺失的数据。

首先对附件数据进行分析，可以看出数据中的空缺可以分为两种：个别交易日数据空缺、最后一天成交量的空缺。查阅资料讨论得到：个别交易日的数据空缺是由节假日股市不交易、因为社会因素导致股票停盘等因素导致，我们对这类

空缺的数据不进行补全操作，因此问题便转化为跟据已知数据对 abc001-abc008 三月二十六日缺失的成交量进行预测。成交量数据是一种时间序列数据，考虑到股票数据具有多变性，所以首先对数据的几项关键指标进行计算分析，据此选择合适的时间序列模型进行拟合和预测，从而得到较为可信的预测值。解决问题一（1）的思路流程图如下：

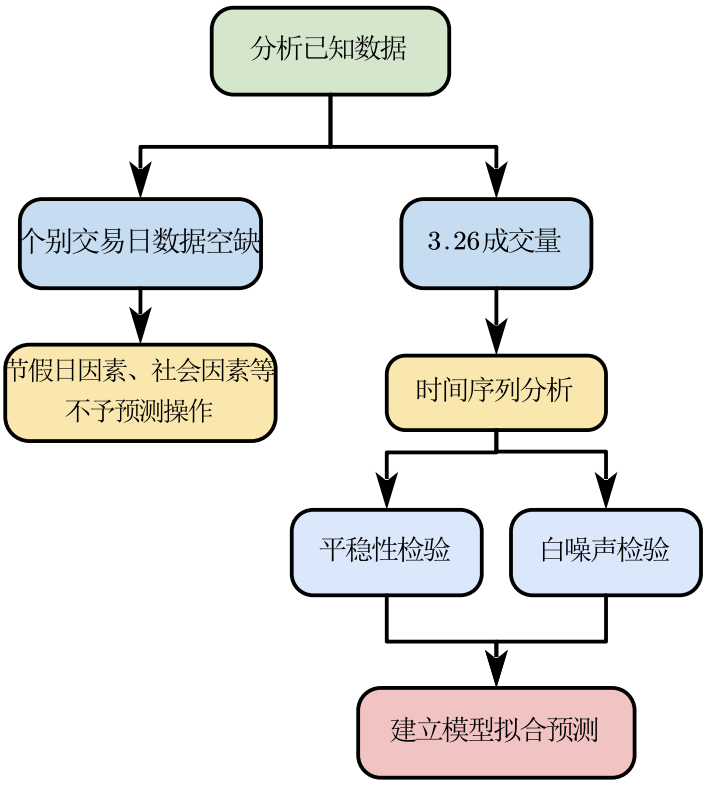


图 2-1　问题一（1）的流程图

2.2 问题一（2）的分析

问题一（2）要求对于附件的成分股数据，通过建立模型给出合理选股方案和投资组合方案。

跟踪一个指数往往能稳定地得到收益，但完全复制的成本太高，通常需要合理选择指数中有代表性的几只股票进行指数跟踪。通过对成分股数据的聚类分析，每个类别中按照历史平均成交量进行排序，选择成交量最大的一只股票，达成选择指数中有代表性的良好成分股的选股目标。考虑到选取的各成分股对指数影响效果不同，考虑对选中的成分股利用蒙特卡洛模拟法分配权重进行投资，通过建立数学模型，分别从价格时间序列角度使投资组合与目标指数的跟踪误差最小、从月化收益率时间序列角度使投资组合与目标指数的超额收益率最大，再将得到的两个权重求取平均值，从而得到稳定性好且具有良好的超额收益率的选股投资组合。解决问题一（2）的思路流程图如下：

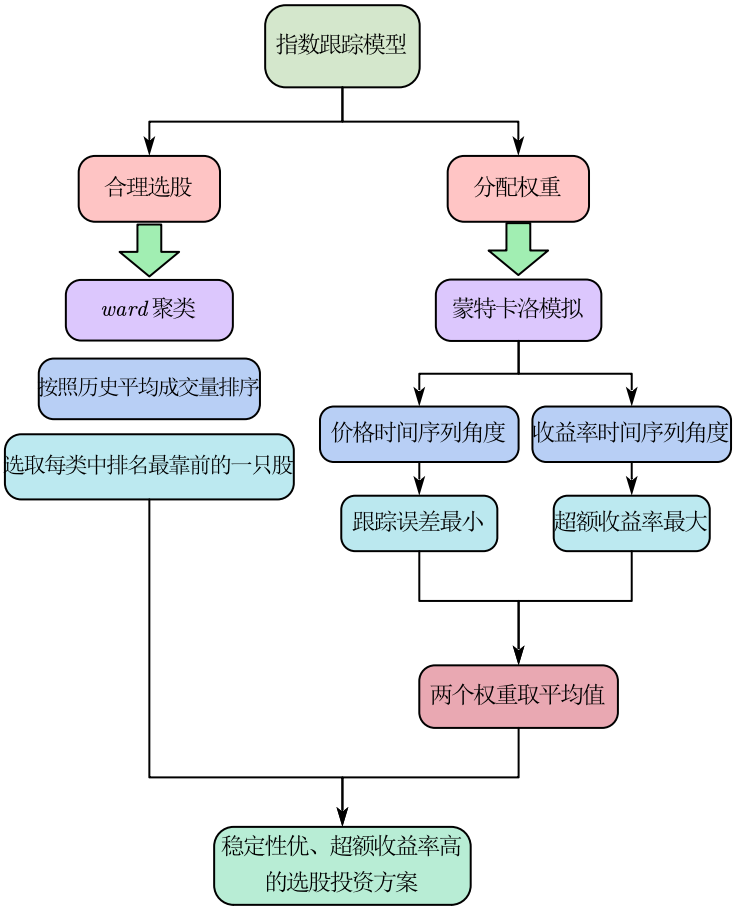


图 2-2　问题一（2）的流程图

2.3 问题二的分析

问题二要求给出合理的评价指标来评估问题一中的模型，并给出分析结果。

问题一中的模型功能是合理选股和赋权，从而给出优良的选股投资方案，将打分法、仅聚类选股等传统模型设为对照组，从模型功能和模型给出结果等方面，通过选取的评价指标对比，说明问题一中模型的性能。

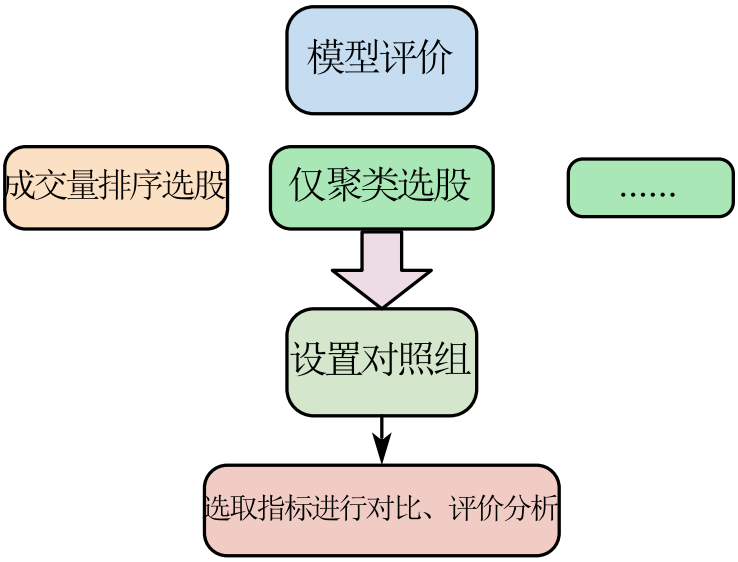


图 2-3　问题二的流程图

三、模型假设和符号说明

3.1 模型的假设

1、投入的资产较少。

2、不考虑成分股的股息、分红和指数跟踪产品管理费用等其他因素。

3、在跟踪投资组合中持有现金资产为零，跟踪投资组合成分股的种类和权重不变。

4、附件给出的十支股为一只指数且走势良好。

3.2 符号的说明

表 3-1 符号说明

符号	意义
N	目标指数中的成分股个数
W_i	基于股票价格时间序列得到的跟踪组合中的成分股 i 的权重
X_i	基于股票收益率时间序列得到的跟踪组合中的成分股 i 的权重
Z_i	0-1 变量，若为 1，代表第 i 只股在组合中被持有
R_t	目标指数在时期 t（1，2，3，……，T）的收益率
r_t	跟踪组合在时期 t（1，2，3，……，T）的收益率
r^*	超额收益率
H_t	每个时期对跟踪误差的贡献
P_{it}	股票 i 在时期 t（1，2，3……，T）的每股收盘价
I_t	目标指数在时期 t（1，2，3……，T）的价格

注：未列出符号及重复的符号以出现处为准

四、模型的建立与求解

4.1 问题一的模型建立与求解

4.1.1 数据的分析

附件数据给出了 abc001-abc010 成分股 280 个交易日的股票数据，其中包含：交易日时间、开盘股价、最高股价、最低股价、收盘股价及成交量，但跟据金融学公式我们可以得到收益率、成分股指数等其他的潜在数据，第一问（1）中主要研究成交量，所以我们将成交量时间序列单独分析。

首先是股票成交量平稳性的检验，随机选取 abc003 股为例，我们借助 Eviews 软件首先做出其成交量时序图，进行整体判断。平稳序列的时序图应该显示出序列始终围绕一个常数值进行波动，且波动的范围不大。

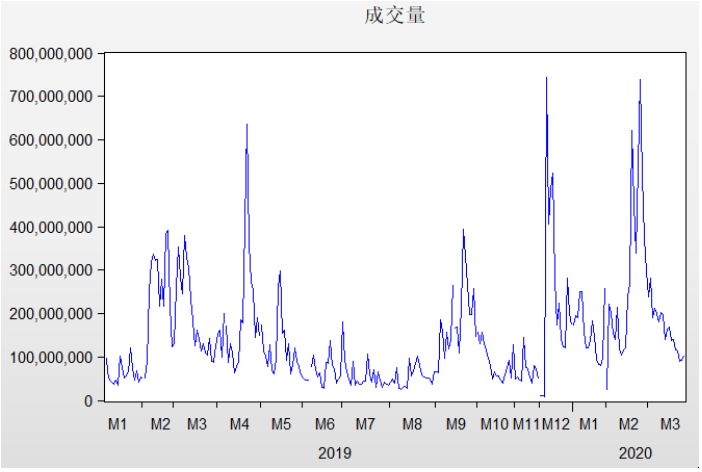


图 4-1　abc003 股成交量时序图

由图 1 可以看出，在这段时间内 abc003 股的成交量波动较大，考虑进行 ADF 单位根检验，通过统计量来进一步判断该序列的平稳性。对其进行单位根检验（见表 1），若未通过检验则考虑进一步对原始数据差分平稳化处理。如果时间序列是不平稳的，可以通过低阶差分、对数差分等方法，使其变成平稳序列。

表 4-1　abc003 股成交量序列 ADF 检验图

		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		-6.302893	0.0000
Test critical values:	1% level	-3.453737	
	5% level	-2.871731	
	10% level	-2.572273	

从表 4-1 可以看出 p 值为 0，在 99%的置信水平下拒绝原假设，即该序列平稳，可以为后续建模使用。若用传统的 Eviews 分布建模的方法，我们还需要进行白噪声检验以及通过计算自相关系数和偏自相关系数来为 ARIMA 模型定阶，为了简化预测过程，我们对平稳序列利用 SPSS 软件进行专家建模。

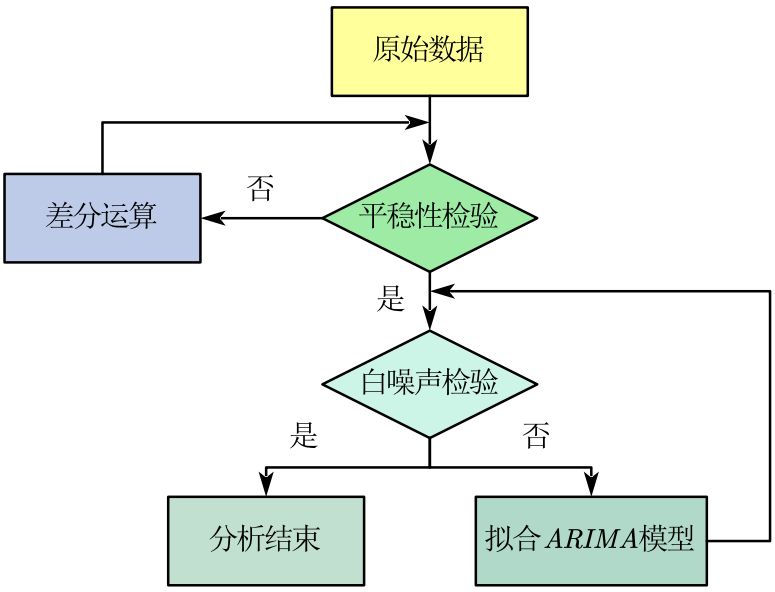


图 4-2　SPSS 中 ARIMA 模型建模流程

表 4-2　各成分股拟合参数表

	001	002
ARIMA	（0，1，2）	（1，0，0）
Q	0.246	0.338
R 方	0.259	0.403
方程	$L_t=\varepsilon_1-0.616\varepsilon_{t-1}+0.2\varepsilon_{t-2}$	$L_t=27614229.77+0.590X_{t-1}$
	003	004
ARIMA	（1，0，0）	（0，1，2）
Q	0.526	0.888
R 方	0.560	0.712
方程	$L_t=144412238.8+0.746L_{t-1}$	$L_t=\varepsilon_t-0.559\varepsilon_{t-1}+0.174\varepsilon_{t-2}$
	005	006
ARIMA	（1，1，1）	（0，1，1）
Q	0.290	0.230
R 方	0.153	0.369
方程	$L_t=0.432X_{t-1}+\varepsilon_t-0.791\varepsilon_{t-1}$	$L_t=\varepsilon_t-0.749\varepsilon_{t-1}$
	007	008
ARIMA	（0，1，1）	（1，0，0）
Q	0.011	0.408
R 方	0.703	0.274

方程	$L_t = \varepsilon_t - 0.739\varepsilon_{t-1}$	$L_t = 18393968.68 + 0.528L_{t-1}$

从表中我们可以看出 Q 值均大于 0.05，所以每个残差项的概率值均大于 0.05，说明 Q 统计量均小于检验水平为 0.05 的卡方分布临界值。说明此时的残差项为白噪声序列，可以考虑使用 ARIMA 模型，以下是各股拟合情况图示：

根据所建立的模型对 abc003 于 2020 年 3 月 26 日的股票发行量进行预测，预测结果如下：

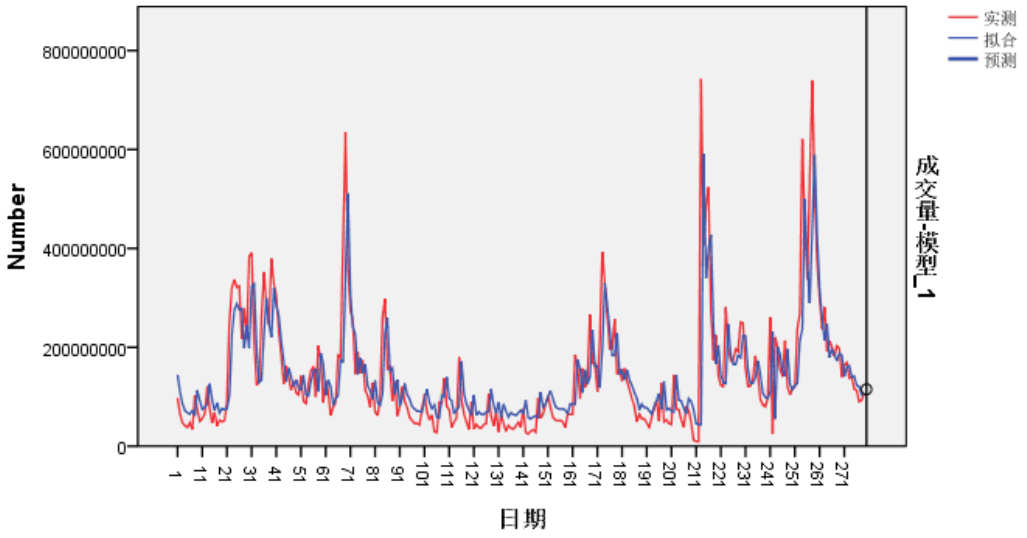
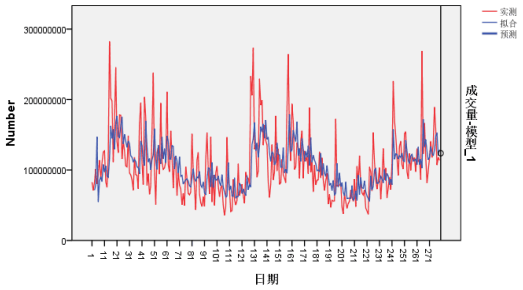
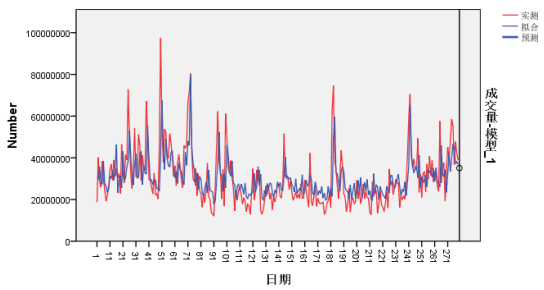


图 4-3　abc003 拟合效果图

利用 ARIMA（1,0,0）模型对 2020 年 3 月 26 日的股票发行量进行预测，绘制预测值与实际值比较图，真实数据和拟合数据的时序图大致重合，得到每天的股票发行量的平均误差为 4.8%，控制在 5%之内，预测效果较好。在 2019 年预测值与实际值之间的差距较大，到 2020 年时，二者之间的差距逐渐缩小，说明模型的预测效果逐渐增强，从预测可知，abc003 股票发行量随着时间的推移呈现出明显的下降趋势，且下降速度逐年加快。





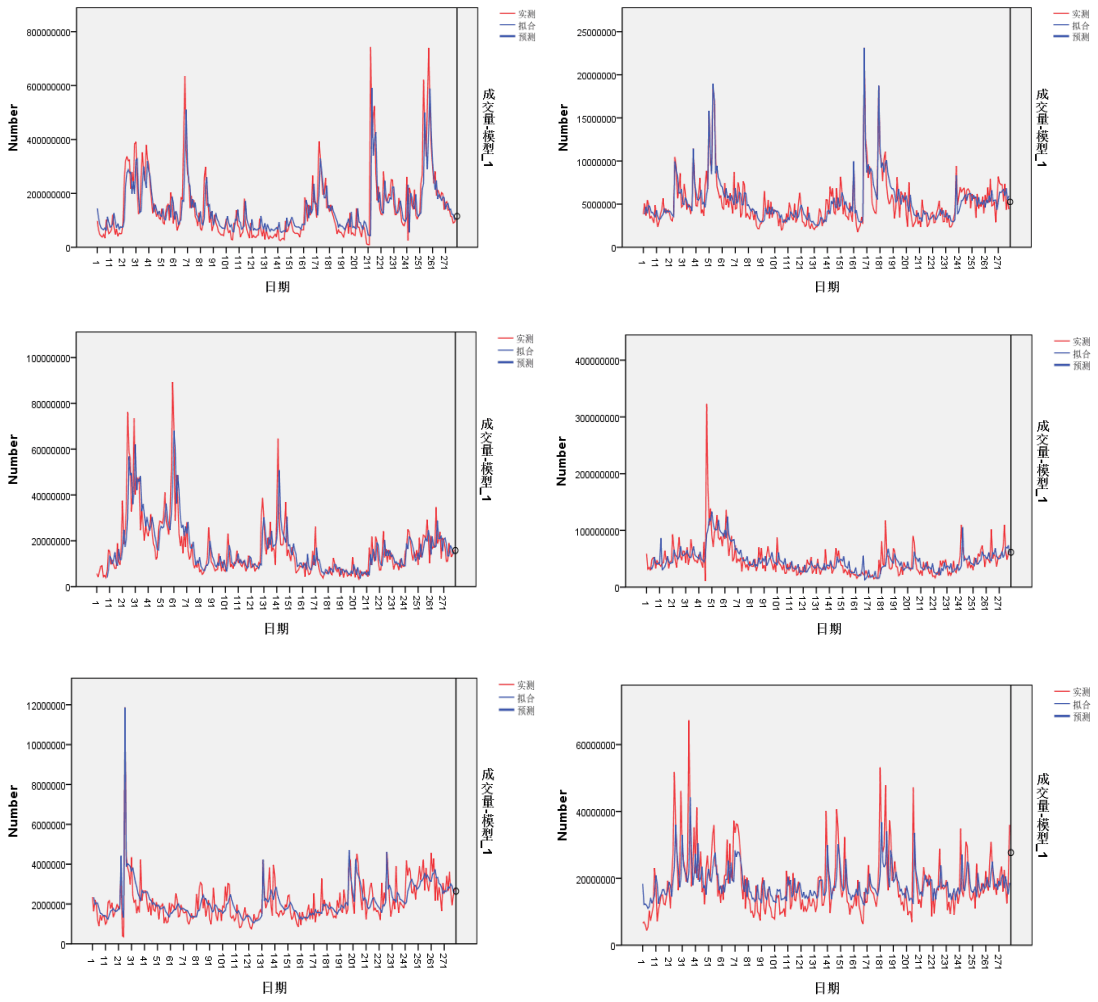


图 4-4　Abc001-abc008 拟合图示

本次模型选择了 ARIMA 对股票发行量进行预测，从预测的结果看，和实际的差别较小，这个模型在短期预测结果比较好。以下是补全的数据列表

表 4-3　成交量预测值统计表

001	002	003	004	005	006	007	008
123828473	35066829	115134125	5260722	15769119	61495563	2644230	27690957

4.2 问题一（2）的建模与求解

首先运用聚类分析研究目标指数成分股的特征，然后找到最能代表目标指数的一定数量的成分股，然后对选中的成分股利用蒙特卡洛模拟法分配权重进行投资，通过建立数学模型，分别从价格时间序列角度使投资组合与目标指数的跟踪误差最小、从月化收益率时间序列角度使投资组合与目标指数的超额收益率最大，再将得到的两个权重求取平均值，从而得到稳定性好且具有良好的超额收益率的选股投资组合。

4.2.1 基于 Ward 聚类选股

利用谱系聚类把题中所给的 10 支成分股分成具有相同特征的类别，用“ward”（离差平方和距离来度量类别之间的距离）把附件中给出的 10 支股票的价格采取聚类分析进行处理，最终我们将十只股票分成了有代表性的五类成分股，结果如图 4-5 所示。

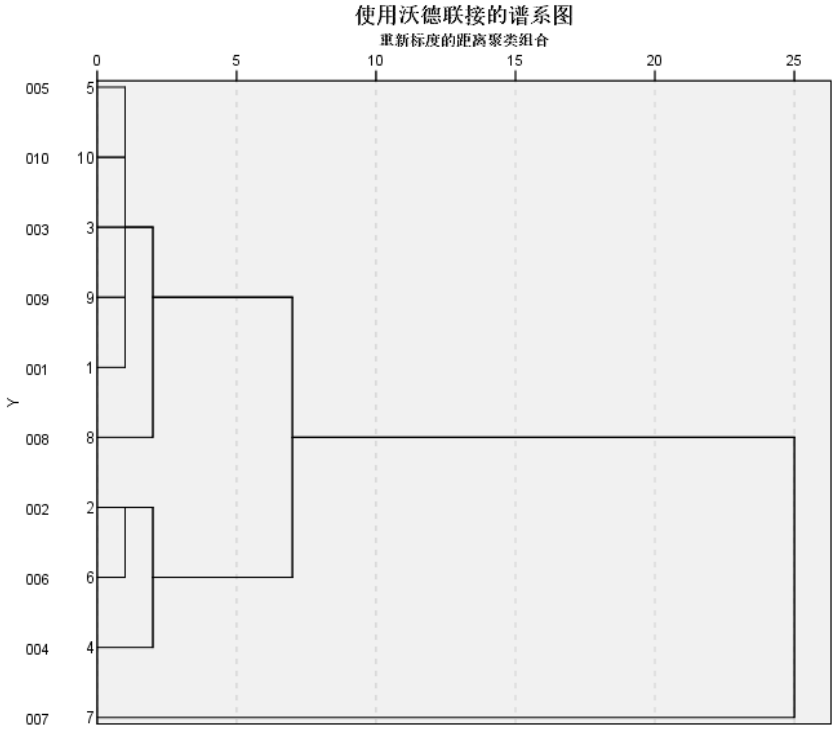


图 4-5　聚类谱系图

表 4-4　十只成分股通过聚类分析得到的五个类别

类一	abc001，abc003，abc005，abc009，abc010
类二	abc008
类三	abc002，abc006
类四	abc007
类五	abc004

据此，从这五个类别中的每个类别里按照成交量大小分别排序，选取每类成交量最大的一只股票，总共五只股票来作为投资组合方案中的股票，之后建立模型及算法确定每只股票的权重。

4.2.2 跟踪误差的非线性优化模型

指数跟踪的优化目标和主要关注对象是跟踪误差，本文建立了跟踪误差优化模型。指数跟踪主要解决两个问题：第一，应该选择哪些种类的股票构建跟踪组合；第二，在跟踪组合中应该如何配置这些股票的权重。我们主要围绕这两个问题来建立跟踪模型。

1、模型参数

跟踪误差 TE

$$TE=\left(\sum_{t=1}^{T}H_t\left(\left|\sum_{i=1}^{n}P_{it}w_i-I_t\right|\right)^{\alpha}\right)^{1/\alpha}$$

TE 表示了我们跟踪的投资组合与目标函数之间的跟踪误差，随着模型参数值的改变，它既可以表示成线性跟踪误差，又可以表示二次跟踪误差，这是模型的一个优势。即若 α=1，那么它就是一个线性跟踪误差；但若 α =2，那么它就是一个二次跟踪误差。H_t 代表每个时期对跟踪误差的贡献，可以根据需要赋予不同时期以不同的权重。

这里我们定义 N 是目标指数中的成分股个数，将附件中的十只股票每只都按月份分期，共分为 15 期，P_{it} 指股票 i 在时期 t（1，2，3……，T）的每股收盘价，按每月末的股票收盘价来表示；I_t 是目标指数在时期 t（1，2，3……，T）的价格，计算公式为：$I_t=\frac{\sum_{i=1}^{10}P_{it}}{10}$

超额收益率r^*

$$r^*=\sum_{t=1}^{T}(r_t-R_t)$$

其中 r*为跟踪组合的收益率与目标指数的收益率之差，定义为超额收益率；R_t 指目标指数在时期 t（1，2，3，……，T）的收益率，其计算公式为：$R_t=\ln\left(\frac{I_t}{I_{t-1}}\right)$，$r_t$ 为跟踪组合在时期 t（1，2，3，……，T）的收益率，计算公式为：$r_t=\ln\left(\left(\sum_{i=1}^{N}P_{it}x_i\right)/\left(\sum_{i=1}^{N}P_{i(t-1)}x_i\right)\right)$

4.2.3 模型的建立

目标函数：$\min \ \ TE=\left(\sum_{t=1}^{T}H_t\left(|\sum_{i=1}^{n}P_{it}w_i-I_t\right)^{\alpha}\right)^{1/\alpha}$

$$\max \ r^*=\sum_{t=1}^{T}(r_t-R_t)$$

约束条件：

$$Z_i\in\{0,1\},\ \ \mathrm{i}=1,...,N$$

$$\sum_{i=1}^{N}Z_i=K$$

$$w_i\geq 0, i=1,2,.....,N$$

$$x_i\geq 0, i=1,2,.....,N$$

4.2.4 确定投资组合成分股权重

为了规避只采取平稳的收益率时间序列带来的局限性，减低时间序列有关信息的丢失，在构建投资组合的过程中，我们同时采取非平稳的价格时间序列构建投资组合，利用简单平均加权法得到综合的跟踪投资组合成分股的权重。

故综合权重如下：

$$w^*=\delta w+(1-\delta)x$$

其中，w 是基于价格时间序列得到的投资组合成分股的权重；X 是基于收益率时间序列得到的投资组合成分股权重。δ 为上述两种权重的加权系数（本文采用简单平均加权法，此处的δ 值为 0.5）。W^*即综合考虑价格时间序列和收益率时间序列信息后的新的投资权重。

4.2.5 指数跟踪遗传算法优化参数设计

五、问题二的模型建立与求解

5.1 设置对照组

本文采用实证分析的方法对问题一建立的模型进行评价，通过设置对照组，从收益率均值与方差、跟踪误差、与目标指数的相关系数、累计收益率、累计超额收益率等几个方面，利用 SPSS 软件，对几种方案的指数跟踪效果进行对比来评价问题一所建模型的可靠性。本文设计了以下几种实证方案进行对比优化计算：

表 5-1　方案组情况概述表

方案一	平均股价成交量排序，等权重投资
方案二	仅聚类分析，等权重投资
方案三	成交量排序与蒙特卡洛赋权结合
方案四	蒙特卡洛算法，基于收益率时间序列用聚类分析
方案五	蒙特卡洛算法，基于价格时间序列聚类分析
方案六	方案四与方案五的简单平均

5.2 实证评价结果

实证分析主要从收益率均值与方差、跟踪误差、与目标指数的相关系数、累计收益率、累计超额收益率六个方面展开，对 6 种方案的指数跟踪效果进行对比来评价问题一所建模型的可靠性，对比结果如表 5-2 到表 5-6 所示。

表 5-2　跟踪组合与目标函数的相关系数

	方案一	方案二	方案三	方案四	方案五	方案六
相关系数	0.7127	0.8521	0.8206	0.9691	0.9353	0.9607

根据表 5-2 计算的投资跟踪组合和目标指数的相关系数来看，这 6 种方案都与目标指数有很高的相关性，其中方案 4、5、6 一组明显比传统的跟踪方法相关性要好， 但从相关系数方面对比，最优的组合是方案四。

表 5-3　跟踪组合的收益率均值与方差

	方案一	方案二	方案三	方案四	方案五	方案六
$E(r_t)$	0.2992	0.3688	1.8756	1.2543	1.1035	1.8996
σ	0.7886	0.5139	0.8661	0.7998	0.5712	0.5946

注：由于收益率均值和收益率标准差数量级较小，故上表中的每组数据的第一行表示投资组合 $E(r_t)$ 的收益率均值乘以 1000 后的结果，每组数据的第二行表示投资组合的收益率标准差 σ 乘以 100 后的结果。

根据表 5-3 里的投资组合的收益率均值和标准差,可以知道基于价格时间序列的跟踪组合的收益率均值要好于基于收益率时间序列的值;通过聚类分析得到的股票去构建的跟踪组合得到的收益率均值要好于用平均股票成交量排名得到的股票所构建的组合的结果;蒙特卡洛算法得到的跟踪组合的收益率均值要好于传统算法;虽然基于价格时间序列所得到的组合的收益率均值高，但是收益率的标准差也大，这说明它的不稳定性也高，即风险也大。总之，基于收益率时间序

列并且采用聚类分析得到的股票运用蒙特卡洛算法所得到的组合收益率均值最大,从这个指标来看，方案四也仍然是最好的。

表 5-4　累计收益率

	方案一	方案二	方案三	方案四	方案五	方案六
累计收益率	0.2918	0.1623	0.7265	0.7987	0.1157	0.2213

表 5-5 累计超额收益率

	方案一	方案二	方案三	方案四	方案五	方案六
累计超额收益率	-0.1645	-0.1482	0.4161	0.4883	-0.1278	-0.0892

从表 5-4 和表 5-5 中可以看出,基于价格时间序列的累计收益率和累计超额收益率高于基于收益率时间序列采用同等计算方法下的值;采用聚类分析确定跟踪组合成分股的方法比采用平均股票成交量排名的方法能获得更高的累计收益率和累计超额收益率;同样运用蒙特卡洛算法也比运用传统的最优化方法得到的累计收益率和累计超额收益率要高。总之，基于价格时间序列，采用聚类分析技术，然后运用蒙特卡洛算法，仅仅从累计收益率和累计超额收益率这两个指标来看，方案六最好。

表 5-6 投资组合的日均跟踪误差

	方案一	方案二	方案三	方案四	方案五	方案六
跟踪误差	3.763E-06	4.657E-06	6.213E-06	6.555E-06	2.231E-06	4.594E-06

通过表 5-6 可以很明显的看出基于价格时间序列的指数跟踪误差要远大于基于收益率时间序列的指数跟踪误差,也就是说采用价格时间序列可以扩大了指数跟踪的误差;采用聚类分析确定跟踪组合成分股的方法比采用平均股票成交量排名的方法能获得较小的日均跟踪误差;运用传统方法计算的跟踪误差在同等条件下比蒙特卡洛算法计算的跟踪误差要略小。总之，基于收益率时间序列,采用聚类分析技术，然后运用蒙特卡洛算法，可以获得比其它方法更小的日均跟踪误差，仅仅从这个角度来看，方案六跟踪效果最好。

由以上分析可知，方案四和方案五在上述指标的分析中相对前六种方案较优,也就是把蒙特卡洛算法运用到聚类分析中得到的跟踪组合成分股的跟踪效果最好。而为了充分运用这两个时间序列的信息，问题一中我们将这两种方案得到的跟踪组合中成分股的权重通过简单的平均加权法获得新的权重,也就是我们的第六种方案，对比结果如下表 5-7 和表 5-8 所示。

表 5-7 方案七和方案八中投资组合成分股的权重

股票代码	abc003	abc004	abc007	abc006	abc008
基于价格	0.0088	0.0021	0.0053	0.0045	0.9793
基于收益率	0.1185	0.3121	0.1876	0.1408	0.2410
基于价格与收益率	0.0636	0.1571	0.9645	0.0727	0.6102

综上，方案六与目标指数的相关系数要高于方案四和方案五的相关系数，也就是说综合收益率和价格两种时间序列的跟踪组合与目标指数的相关性要优于单个时间序列得到的相关性，说明了提取时间序列中更多的信息可以提高跟踪组合和目标指数的相关系数。通过简单平均法综合收益率时间序列和价格时间序列后，方案六的跟踪组合的收益率向基于价格时间序列的收益率收缩，同时降低了收益率的波动性，也就是收益率的标准差减少了。同时，综合了价格时间序列和收益率时间序列的指数跟踪组合跟踪的误差较单独的时间序列得到的误差要小，说明更多的时间序列信息可以明显地提高模型指数跟踪的精度。

通过比较基于收益率时间序列和基于价格时间序列的指数跟踪效果，我们建立的综合两种时间序列信息能够进一步优化跟踪效果，并且从总体来看,运用聚类分析技术和蒙特卡洛算法相结合的优化方法在各个评价指标方面都可以获得较好的结果;基于价格时间序列虽然比基于收益率时间序列导致更大的日均跟踪误差,但比基于收益率时间序列得到更好的累计收益率和累计超额收益率;在此基础上的综合价格时间序列和收益率时间序列这两方面的信息，可以降低日均收益率的波动，提高跟踪组合与目标指数的相关系数，降低跟踪误差，改善了指数跟踪的效果。因此，基于聚类选股的蒙特卡洛算法优化模型，利用价格时间序列和收益率时间序列的数据综合信息，可以得到比单一时间序列信息、普通选股的传统算法优化模型更好的解题效果，故问题一中建立的模型是合理的、可靠的。

五、问题三的模型建立与求解

5.1 股票指数的计算

股票价格指数分为全部上市股票价格指数和成分股价格指数，题目数据只包括成分股一段时期内的时点价格，所以用成分股价格指数代替目标质数的股票指数。

股票价格平均数含义：反应一定时点上市股票价格的绝对水平，计算的方法分分加单算术估价平均数方法、修正的股价平均数、加权股价平均数。其中加权股价平均数更能反映出目标指数的变化趋势，权重的计算方法一般依据成交股数(成交量)、股票总市值以及股票发行量，结合题目已有数据，我们依据数据包含

的时期中每一个成分股的所有成交量的和来计算出每一个成分股的权重，依此计算出目标指数的股票价格。

选取 10 个成分股的起始日期最大值 2019/01/29 为基期，基期指数的确定不同的交易所有不同的数值，其中沪深 300 的基期指数为 1000，美国道琼斯基期指数液位 1000，在第三问的模型中，参考以上两种，确定基期目标股指数为 1000，在基期之后的每一个时点，都依据计算公式计算出目标指数的时点指数

股市的指数影响因素很复杂，结合近年来中美贸易战重要事件，主要有如下：2019 年 5 月 8 日美国贸易代表室(USTR)宣布将 2000 亿美元中国商品税率从 10%提高至 25%，2019 年 5 月 10 日正式实施，美国贸易代表室（USTR）公布对华约 3000 亿美元商品拟加 35%关税。同日中国政府对美国 5140 项进口商品提高加征关税税率。

反映到目标指数的股价指数的图形上也不难看出这一规律。

图 6-1 目标指数股指

考虑到股市在节假日等时间不进行交易，对下一年的预测即预测 250 个工作日的走势，基于时间序列对未来一年进行预测。

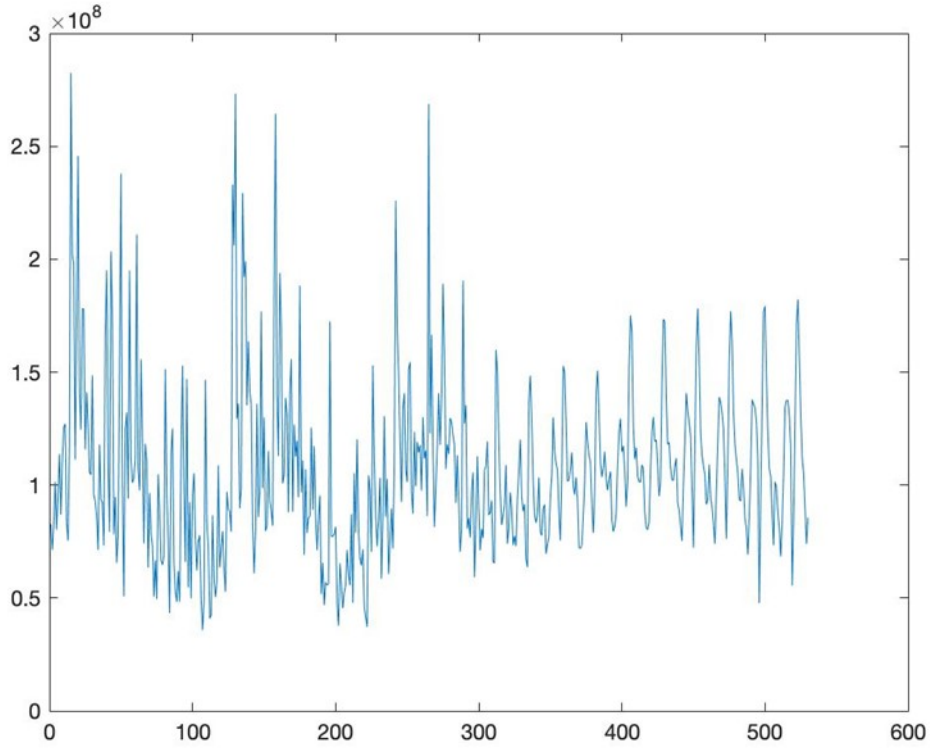


图 6-2 股指预测

结合第二问的组合优化方式求得一个最优方式的年收益率为 1.7454，总的收益率为 1.8619，在这种情况下 10 只股票的权重分别为：0.0005、0.0426、0.0001、0.0220、0.0006、0.0010、0.4172、0.3369、0.1791、0.0001。

七、模型的评价

7.1 模型的优点

1.本文根据跟踪误差的内涵，建立基于股票价格和收益率时间序列的规划模型，明确定义基本参数、约束条件和目标函数，行文具有逻辑性，清晰完整。

2.大多数学者都采用随机抽样的方法构建股票投资组合，而本文认为通过聚类分析技术得到股票投资组合，选择聚类分析方法作为选股的基本手段是可行的合理的可靠的。并同时考虑了股票收益率与价格，具有创新性，得到了更好的指数跟踪效果，比较合理。

3.本文针对传统优化算法的不足，在模型求解的过程中采用了遗传算法，这种算法从初始群体开始搜索，得到最优解的概率要大于点到点的优化算法，得到的结果更加准确。

7.2 模型的缺点

1.本文采用简单的平均法加权收益率和价格序列得到的权重，具有一定的局限性。在综合时间序列和收益率时间序列这两方面信息的时候考虑不同比例的权重，做到最大程度地降低日均收益率的波动，提高投资组合与目标指数的相关系数，降低跟踪误差，改善指数跟踪的效果。

2.本文仅简单地采用附件中所给十只股票算出一个目标指数，大市场的指数可能与其存在一定的误差。

3.本文虽然将智能算法运用到指数跟踪优化求解的过程中，但求解的模型仍然是传统的指数跟踪静态模型，下一步的研究应考虑动态模型。

八、参考文献

[1]林杰,朱家明,陈富媛. 基于 ARMA 模型对股票“青岛海尔”成交量的分析预测票成交量对投资组合优化影响的研究[J].哈尔滨师范大学自然科学学报,2016.

[2]姜乐.基于时间序列的股票价格分析研究与应用[D].大连理工大学,2015.

附录

```
clc
clear
P=xlsread('各股 Pit.xlsx');
I=xlsread('目标指数 It.xlsx');
H=xlsread('Ht.xlsx');

alpha=2;
alpha_=1/alpha;

N=100000000;
min_ans = +inf;
min_w=zeros(1,5);

for rnd=1:N
    x=rand(1,5);
    y=sum(x);
    W=x/y;
    now=0;
    for t=1:14
        tmp=0;
        for i=1:5
            tmp=P(t,i)*W(i);
        end
        now=now+H(t) * power(abs(tmp-I(t)),alpha);
    end
    now=power(now,alpha_);
    if min_ans>now
        min_ans=now;
        min_w=W;
    end
end

min_ans
min_w
```

求第二个

```
clc
clear

```

```
R=xlsread('Rt.xlsx');
P=xlsread('各股 Pit.xlsx');

N=100000000;
max_ans=-inf;
max_x=zeros(1,5);
for n=1:N
    x_=rand(1,5);
    y_=sum(x_);
    x=x_/y_;
    now=0;
    for t=2:14
        rt=0;
        t1=0;
        t2=0;
        for i =1 : 5
            t1=P(t,i)*x(i);
            t2=P(t-1,i)*x(i);
            rt=rt+log(t1/t2);
        end
        now=now+rt-R(t-1);
    end
    if max_ans<now
        max_ans=now;
        max_x=x;
    end
end

max_ans
max_x
```

第三问时间序列预测

```
clc;clear
load('date.mat')
yuce=gupiao_2;
for i=1:10
    a=size(yuce{i},1);
    yuce{i}=[yuce{i};zeros(250,6)];
    yuce{i}(a+1:end,1)=((43916+1):(43916+250))';
    b=size(yuce{i},1);
    for j=a+1:b
        for k=2:size(yuce{i},2)
            yuce{i}(j,k)=dopredict(yuce{i}(:,k),j,5,23);
```

```
        end
    end
end
plot((1:530),yuce{1}(:,6))
average_income=zeros(10,1);
day_income=cell(1,10);
for i=1:10
    a=length(yuce{i}(:,5))-1;
    clear income
    income_rate=zeros(a,1);
    for k=2:length(yuce{i}(:,5))
        income_rate(k-1)=yuce{i}(k,5)/yuce{i}(k-1,5);
    end
    day_income_rate{i}=income_rate;
    average_income_rate(i)=mean(income_rate);
end
save average_income_rate.mat average_income_rate
for i=1:10
    for j=1:length(day_income_rate{i})
        risk{i}=day_income_rate{i}-average_income_rate(i);  % 第 i 支股票的日风险=每天的收益率-平均收益率
    end
end
for i=1:10
same_date(i)=length(risk{i});
end
max_date=min(same_date);
max_date=max(same_date);
global day_risk
day_risk=zeros(max_date,10);
for i=1:10
day_risk(end-length(risk{i})+1:end,i)=cell2mat(risk(i));
end

for i=1:size(day_risk,1)
    if sum(day_risk(i,:)==0)==0
        day_risk(1:i-1,:)=[];
        break
    end
end


A=ones(1,10);
b=[1];
```

```
x0=[0.1;0.1;0.1;0.1;0.1;0.1;0.1;0.1;0.1;0.1];
L=zeros(10,1);
global q
q=0.1;
[x,favl]=fmincon(@R,x0,[],[],A,b,L,[],@solve);
max_in_rate=-favl;
year_in_rate= (max_in_rate).^250
sum_in_rate= (max_in_rate).^279
save answer.mat q x max_in_rate
load 'answer.mat'
```

函数 1

```
function do=dopredict(order,index,n,m)
order=order(1:index-1);% 参考的数据 order
L=length(order);
A=floor(L/m);
B=mod(L,m);
y=[];
for i=1:A
    y=[y;order(B+m*i)];
end

L1=length(y);
n=floor(A/2); %让 n 动态变化

for i=1:L1-n+1
    get1(i)=sum(y(i:i+n-1))/n;
end
L2=length(get1);
for i=1:L2-n+1
    get2(i)=sum(get1(i:i+n-1))/n;
end
ans1=2*get1(end)-get2(end);
ans2=2*(get1(end)-get2(end))/(n-1);

do=ans1+ans2;

end
```

函数 2：

```
function [c,ceq]=solve(x)
global ri_feng_xian
```

```
E=ri_feng_xian;
c=sum((E*x).^2)-q;
ceq=[];
```