B 题　　量化投资问题

摘　　要

股市预测是量化投资策略的基础。但由于数字经济受宏观市场指标、政策和庞杂的市场因素等多方面的影响，从海量的市场信息中提取出有效指标并制订出可靠的交易策略，对股市走势预测的要求较高，提高金融时序预测精度一直是时间序列预测问题中的难点及热门研究问题。

针对问题一，为挖掘与“数字经济板块”高度相关的指标，利用皮尔逊相关系数法对共计 44 个指标分别与“数字经济板块”中的成交量和收盘价进行皮尔逊相关性分析。最终筛选出与成交量高度相关指标为 VMA、伦敦金融时报 100 指数、数字孪生和快手概念，与收盘价高度相关的指标为 BBI、MA、EXPMA 和 BOLL。经过计算与分析，并进行皮尔逊显著性检验，结果表明所筛选指标与“数字经济板块”具有较强的相关性。

针对问题二，需预测“数字经济板块”规定时间内的成交量，首先利用 VMD 分解成交量历史数据降低训练集非平稳性、混沌的特性。后利用 LSTM 对非线性数据良好的回归能力分别对分解后的各模态分量进行预测，并对预测出的各分量加和重构得到初步预测成交量。同时对成交量与其高度相关性指标进行多元线性相关拟合得到其关系式。为了使预测结果更加准确，通过粒子群优化算法联合寻找出最优的权重组合，结果表明修正过后的模型误差更小，更好地预测了股市的总体趋势。

针对问题三，需预测“数字经济板块”规定时间内的收盘价，利用问题二建立的基于 VMD-LSTM 的预测模型。基于与收盘价相关的主要指标对预测模型进行修正，为了更好地反应模型的泛化能力，采用了交叉验证对模型进行评判，在高度相关性指标的辅助下能够更好的捕捉股市涨跌趋势。结果表明模型的有效性良好及精度较高。

针对问题四，要根据问题三预测得到的收盘价序列来对“数字经济”板块每 5 分钟频率价格进行买卖交易。由于交易频率较高且股票的浮动较大，引入了相关强弱指标（RSI）、对数移动均线指标（LOGMA）、布林线指标（BOLL）及夏普比率四种指标。其中 RSI 能够排除风险因素对绩效评估的不利影响；在频繁大量的交易基础上需要参考 LOGMA 采取实际交易行为；同时 BOLL 反映了股市的波动情况；然后利用夏普比率有效评价预期报酬与规避风险的关系。最后形成组合投资策略，实现精准量化投资，最终获得盈利。

关键词:皮尔逊相关系数法;VMD 分解;LSTM 神经网络;交叉验证;粒子群算法;对数移动均线

目录

一：问题重述

1.1 问题背景

随着中国经济和股票市场的快速增长，积极推进改革开放资本市场的稳定发展，也为人们提供了更多的投资手段。作为中国市场经济的主要组成部分，股票市场在国家和个人中都扮演着重要的角色，但是它对人们的投资和理财也有着惊人的要求。因此定量投资的出现吸引了许多投资者的注意并成为一种趋势，成为许多投资者的金融投资体系中不可或缺的一部分，随着互联网技术的发展，定量投资在欧洲和美国以及其他地区发展迅速。定量投资已成为一种新的投资方式让投资者投资理财，它与计算机和数据模型相结合以便能更有效、更准确地投资。

但是目前的定量投资预测技术较多地只考虑历史数据的训练和预测市场走势[1]，并没有考虑庞杂的市场影响因素，而外在的市场信息变化可以直接影响股票市场，使股票走势更加难以预测和捕捉。如何从海量的市场信息中提取出有效指标并制订出可靠的交易策略，是一个非常有意义且具有挑战性的工作。

1.2 本题所给信息及数据

本题提供的主要信息为数字经济板块信息，其中包含了从 2021 年 7 月 14 日到 2022 年 1 月 28 日的开盘价、收盘价和成交量等关键数据，同时还提供了两种宏观市场指标和 2020 年 12 月 31 日至 2022 年 1 月 28 日区间的 12 项国内市场指标，2020 年 12 月 31 日至 2022 年 1 月 28 日时间段内的技术指标方面含有 VMA，ARBR 和 BOLL 等 14 项重要参考数据，国际市场指标包含道琼斯工业平均指数、恒生指数在内的 11 项数据，且日期区间为 2020 年 12 月 31 日至 2022 年 1 月 31 日。汇率方面提供了 2020 年 12 月 31 日至 2022 年 1 月 31 日区间的美元兑换人民币和欧元的汇率水平。除此之外，数字网络经济包括数字媒体、数字孪生、快手概念和互联网电商等信息也是比较重要的关联信息数据。

1.3 待解决的问题

在本文中，我们需要根据问题中所提供的附表数据，推导出“数字经济”板块信息与各项指标之间的数学关系，然后建立起成交量和收盘价结合指标间数学关系的模型，对成交量和收盘价进行预测，最后假设“数字经济”板块指数为交易对象，计算在 2022 年 1 月 4 日至 2022 年 1 月 28 日期间交易的总收益率、信息比率、最大回撤率。本文具体要解决的

问题如下：

问题 1：对所提供的各项指标进行分析，从中提取出与“数字经济”板块有关的主要指标。

问题 2：以 2021 年 7 月 14 日至 2021 年 12 月 31 日的每 5 分钟“数字经济”板块指数为训练集，以 2022 年 1 月 4 日至 2022 年 1 月 28 日的每 5 分钟“数字经济”板块指数为测试集。根据问题 1 提取出来的各项指标对“数字经济”板块指数每 5 分钟成交量进行预测。

问题 3：以 2021 年 7 月 14 日至 2021 年 12 月 31 日的每 5 分钟“数字经济”板块指数为训练集，以 2022 年 1 月 4 日至 2022 年 1 月 28 日的每 5 分钟“数字经济”板块指数为测试集。根据问题 1 和问题 2 建立模型对每 5 分钟的“数字经济”板块指数（收盘价）进行预测。

问题 4：假设以“数字经济”板块指数为交易对象（在实际交易中指数无法交易，只能交易其中的个股），给定初始资金 100 万元，交易佣金为 0.3%，根据问题 3 得到的结果对 “数字经济”板块每 5 分钟频率价格进行买卖交易，计算在 2022 年 1 月 4 日至 2022 年 1 月 28 日期间交易的总收益率、信息比率、最大回撤率。

二：模型假设

假设 1：假设处理空值和异常值后的数据是平滑的，能够满足后续模型的计算需求。

假设 2：假设所选取的与“数字经济”板块有关的主要指标具有很高的可靠性。

假设 3：假设股票的买入和卖出都是根据收盘价来操作的。

三：符号说明

表 3-1　符号说明

符号	意义
r	皮尔逊相关系数
s	皮尔逊相关系数的显著性水平
IMFs	VMD 分解产生的变分模态分量
$x(t)$	给定信号
$m_k(t)$	第 k 个模态分量
w_k	第 k 个模态的中心频率
K	分解得到的模态总数
α	惩罚参数
$h(t)$	LSTM 神经元的短期状态
$c(t)$	LSTM 神经元的长期状态
$\bar{x}$	输入变量的平均值
σ	输入变量的标准差
$x_i(t)$	t 时刻第 i 个粒子的位置
$v_i(t)$	t 时刻第 i 个粒子的速度

c_i	粒子群算法的学习因子
w	粒子群算法的惯性权重
MAE	预测值和真实值的平均系统性偏离程度
$RMSE$	预测值和真实值的平均绝对偏离程度
$MAPE$	预测值和真实值的偏离相对于真实值的平均偏离程度
J_{rmse}	粒子群算法的适应度函数
y_1	通过主要指标拟合的成交量表达式
y_2	通过主要指标拟合的收盘价表达式
V_{LSTM}	基于 VMD 和 LSTM 预测的结果
C_1	经过修正的成交量预测表达式
C_2	经过修正的收盘价预测表达式

四：问题一模型的建立与求解

4.1 问题一分析

由于问题二和问题三需要分别对“数字经济”板块指数的成交量和收盘价进行预测，所以我们提取出的与“数字经济”板块有关的主要指标，肯定是与成交量和收盘价相关性较高的指标。为保证分析的准确性与可靠性，需对附表中的原始数据进行处理。在缺失值处理中，采用保形样条插值法做插值拟合，在 Matlab 中使用 filliming 函数的 pchip 参数实现。

K 线图可以反映价格的走势，其一个线段内记录了多项讯息，广泛用于股票、期货、贵金属、数字货币等行情的技术分析[2]。选取 2022 年 1 月 28 日“数字经济”板块绘制 K 线图如图 4-1 所示。

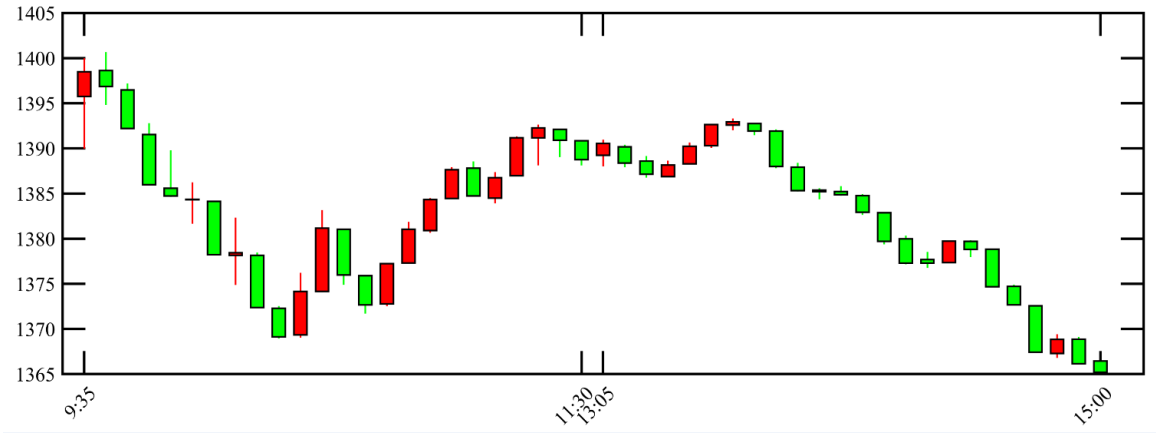


图 4-1　2022 年 1 月 28 日“数字经济”板块 K 线图

4.2 数据处理

各数据的时间范围与采集频次如表 4-1 所示。

表 4-1　各数据的采集时间范围和采集频次

数据类别	数据采集时间范围	数据采集频次
国内市场指标	2020/12/31—2022/1/28	1 天/次
数字经济板块	2021/7/14—2022/1/28	5 分钟/次
技术指标	2020/12/31—2022/1/28	1 天/次
国际市场指标	2020/12/31—2022/1/31	1 天/次
汇率	2020/12/31—2022/1/31	1 天/次
其他板块	2020/12/31—2022/1/28	1 天/次

4.2.1 异常数据的处理

由于问题需要我们对股市中的部分股票或者指数进行策略研究，因此数据所在的时间段应该为：交易日的上午 9:35—11:30 和下午 13:05—15:00，对于股市在星期六和星期天的数据不予考虑，并对这部分数据进行删除处理。

4.2.2 缺失数据的处理

在删除掉异常数据后，进行表格中缺失数据的补充。在此步中，我们使用 Matlab 中 filliming 函数的 pchip 参数进行插值拟合，填补缺失的数据，清洗后的数据见支撑材料，用清洗后的数据来进行后续建模，增强了模型的鲁棒性。

4.3 问题一模型的建立

4.3.1 相关系数 *r* 的计算

皮尔逊相关系数法是一种准确度量两个变量之间的关系密切程度的统计学的方法[3]。对于两个变量 x 和 y，通过试验可以得到若干组数据，记为$(x_i, y_i)(i=1,2,\cdots,n)$，则相关系数的数学表达式为：

$$r=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}} \qquad (4\text{-}1)$$

其中，$\bar{x}$、$\bar{y}$ 分别为 n 个试验值的均值。相关系数 r 的取值范围在-1 和+1 之间，即$|r|\le 1$。$|r|$越接近 1，则表明 x 与 y 线性相关程度越高。若$r=-1$，表明 x 与 y 之间为完全负线性相关关系；若$r=1$，表明 x 与 y 之间为完全正线性相关关系；若$r=0$，表明两者不存在线性相关关系。

一般情况下，r 的取值在（-1,1）之间，相关程度可分为以下几种情况：当$|r|\ge 0.8$时，可视为高度相关；$0.5\le|r|\le 0.8$时，视为中度相关；$0.3\le|r|\le 0.5$时，视为低度相关；当$|r|<0.3$时，说明两个变量之间的相关程度极弱，可视为非线性相关。

4.3.2 *r* 的显著性检验

相关系数 r 是通过样本数据计算而得，其值受到样本抽样的随机性、样本的数量等影

响，因此需要考察样本相关系数的可靠性，即进行显著性检验[4]。首先样本不相关推断的零假设为H_0；其次计算检验的统计量，通常情况下采用 t 分布检验，计算式为：

$$t = |r|\sqrt{\frac{n-2}{1-r^2}} \sim t(n-2) \qquad (4\text{-}2)$$

最后，根据给定的显著性水平s和自由度$d_f = n-2$，利用 t 分布表查出$t_{s/2}(n-2)$的临界值。若$|t| > t_{s/2}$，则拒绝原假设H_0，表明总体两个变量之间存在显著的线性关系。

4.4 问题一模型的求解及结果分析

利用上述提到的皮尔逊相关系数法，对国内市场指标中的 12 个指标、技术指标中的 14 个指标，国际市场指标中的 11 个指标、汇率中的 2 个指标、其他板块信息中的 5 个指标，共计 44 个指标分别对数字经济板块中的成交量和收盘价进行皮尔逊相关性分析。

44 个指标与成交量的皮尔逊相关系数如表 4-2 所示，所有指标的皮尔逊相关系数表见支撑材料。

表 4-2 44 个指标与成交量的皮尔逊系数

皮尔逊相关系数 r			
0.22941	0.425409	-0.42865	-0.36992
0.266416	-0.09904	-0.00019	0.35552
-0.10714	-0.09727	0.228327	-0.00446
0.723428	0.518608	-0.05608	0.013854
0.203631	-0.05758	0.211451	0.231953
0.168099	0.024272	0.223274	0.228412
0.189646	0.147711	-0.26042	0.063747
-0.33802	-0.52929	0.460136	0.10328
-0.58168	-0.45744	0.083454	0.396835
-0.40903	0.552381	0.481902	-0.35932
-0.59826	-0.63277	0.453683	-0.39012

根据表 4-2 中的各指标与“数字经济”板块的成交量的皮尔逊相关系数，选择$|r| > 0.58$的指标为影响成交量的主要指标，共有 4 个，分别是技术指标中的 VMA，国际市场指标中的伦敦金融时报 100 指数，其他板块中的数字孪生和快手概念。

44 个指标与成交量的皮尔逊相关性系数矩阵和皮尔逊显著性检验矩阵分别如图 4-2 和图 4-3 所示。

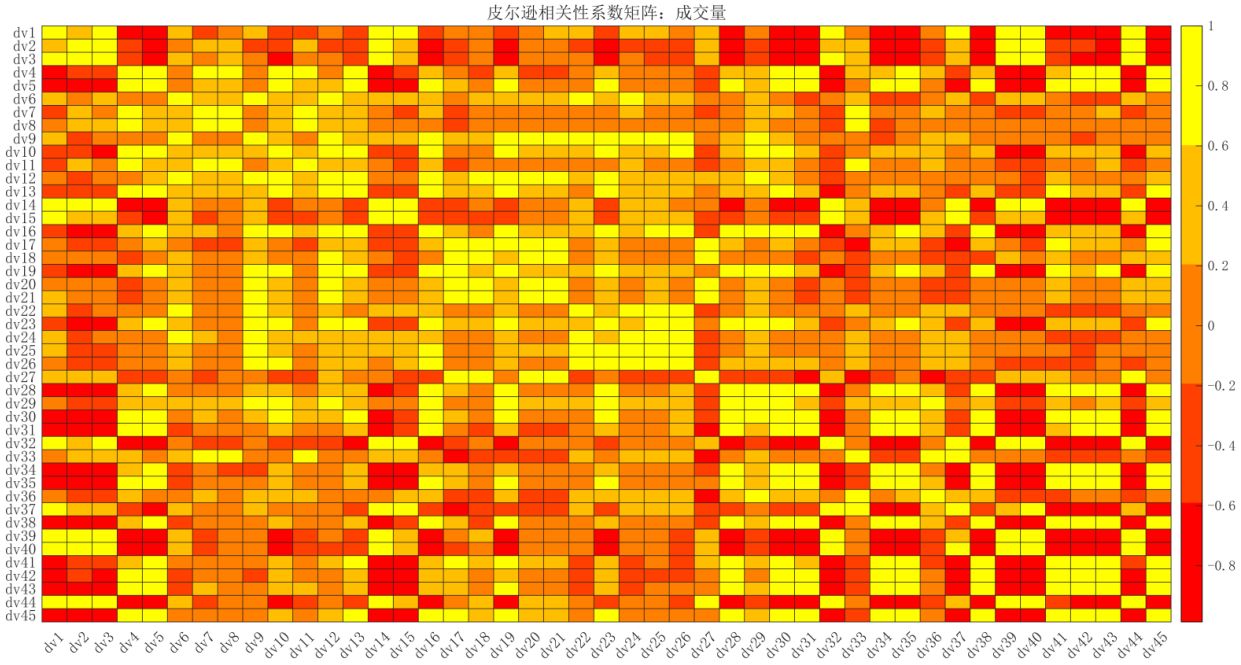


图 4-2 成交量皮尔逊相关性系数矩阵

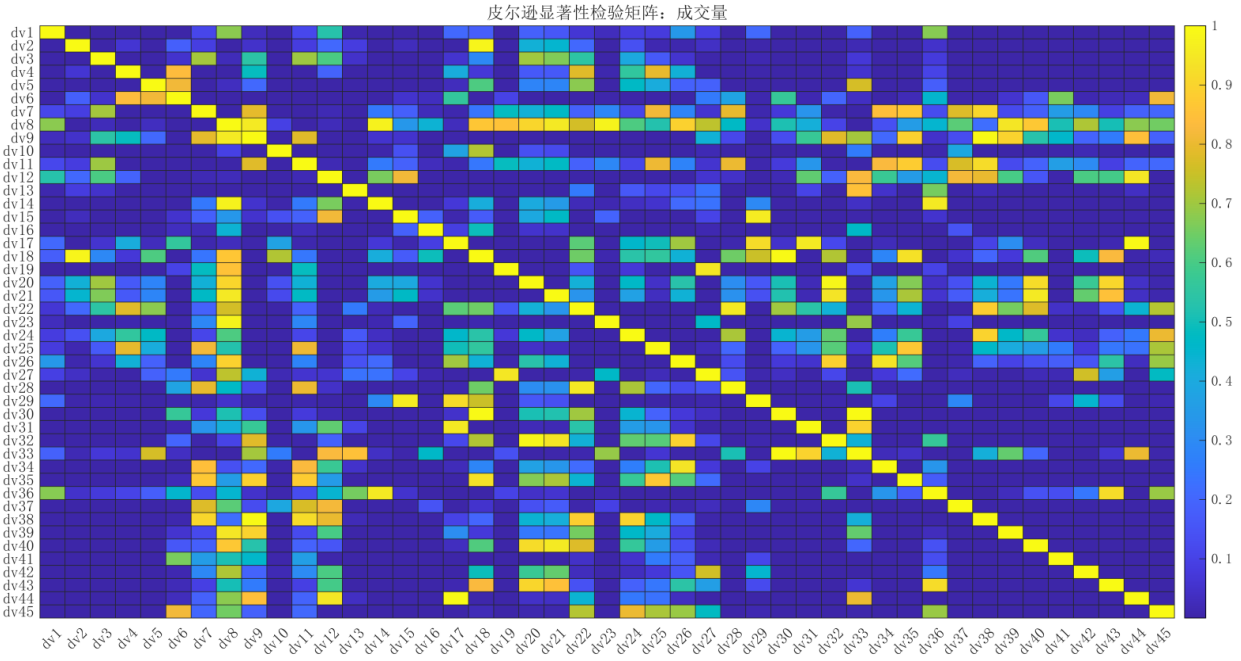


图 4-3 成交量皮尔逊显著性检验矩阵

以技术指标中的 VMA 为例，其与成交量的皮尔逊相关系数为 0.723428，显著性水平为 0.0165，表明 VMA 与成交量不相关的概率小于 1.65%，即两者之间存在很强的相关性。

44 个指标与收盘价的皮尔逊相关系数如表 4-3 所示，所有指标的皮尔逊相关系数表见支撑材料。

表 4-3 44 个指标与收盘价的皮尔逊系数

皮尔逊相关系数 r			
0.232142	0.442545	-0.23645	0.07444
0.532152	0.039854	0.203776	0.585188
0.09142	0.041466	0.674719	0.511068
0.439592	0.07633	0.165409	0.633406
0.885514	0.375518	0.867155	0.883865
0.223454	0.398074	0.214014	0.306305
0.169206	0.750526	0.174897	0.297253
0.089032	-0.64642	0.262164	-0.08238
-0.32926	-0.06776	-0.03166	-0.03859
-0.19972	0.46475	0.29691	0.397664
-0.06437	-0.21314	0.608859	0.310288

根据表 4-3 中的各指标与“数字经济”板块的收盘量的皮尔逊相关系数，选择$|r|>0.75$的指标为影响成交量的主要指标，共有 4 个，分别是技术指标中的 BBI、MA、EXPMA 和 BOLL。

44 个指标与收盘价的皮尔逊相关性系数矩阵和皮尔逊显著性检验矩阵分别如图 4-4 和图 4-5 所示。

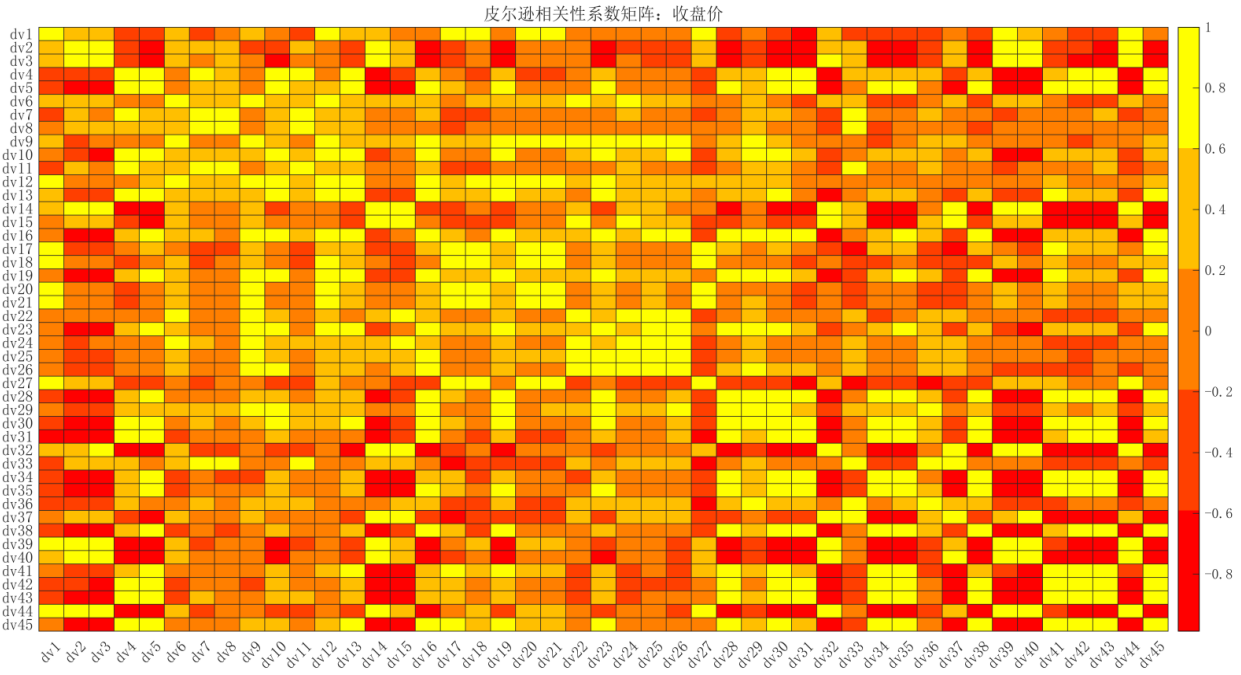


图 4-4 收盘价皮尔逊相关性系数矩阵

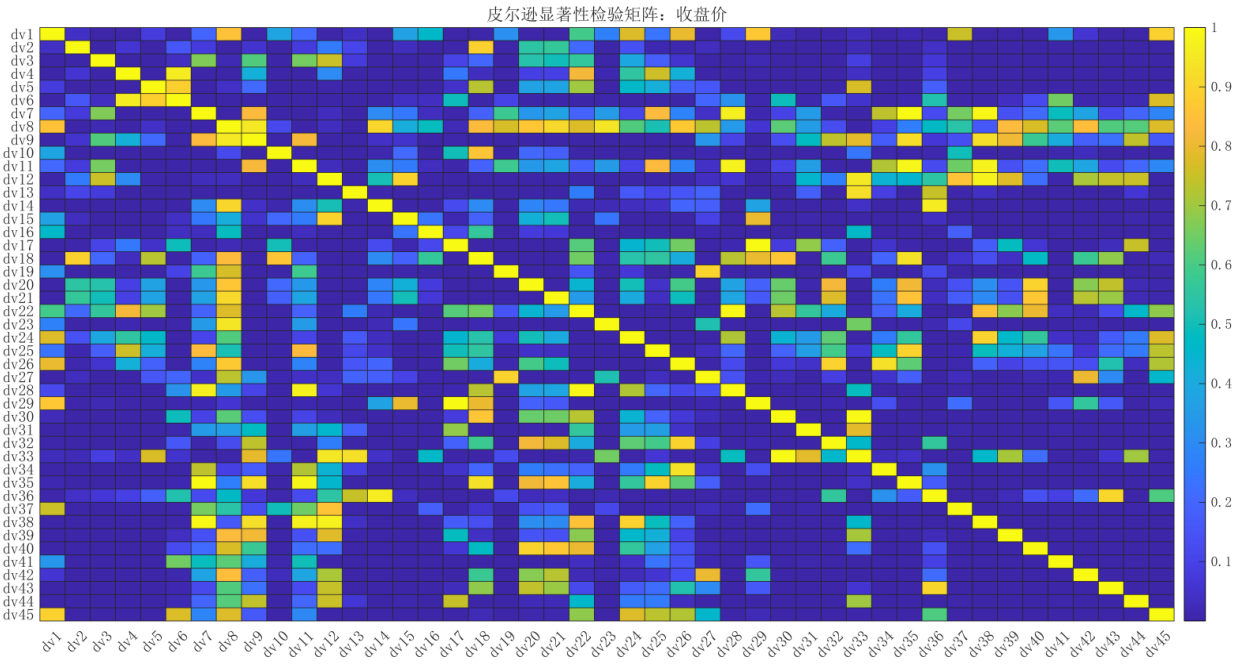


图 4-5 收盘价皮尔逊显著性检验矩阵

以技术指标中的 EXPMA 为例，其与收盘价的皮尔逊相关系数为 0.883865，显著性水平为 0.008，表明 EXPMA 与收盘价不相关的概率小于 0.8%，即两者之间存在很强的相关性。

综上所述，与“数字经济”板块成交量有关的主要指标有技术指标中的 VMA，国际市场指标中的伦敦金融时报 100 指数，其他板块中的数字孪生和快手概念；与“数字经济”板块收盘价有关的主要指标有技术指标中的 BBI、MA、EXPMA 和 BOLL。

五：问题二模型的建立与求解

5.1 问题二的分析

金融时序难以预测主要源自于其非平稳、混沌以及非线性的特性。为解决金融时序的非平稳和非线性两个特性，分别引入变分模态分解和机器学习算法，提出了基于 VMD-LSTM 的预测模型。首先，通过引入变分模态分解（VMD），解决了经典去噪方法无法将金融时间序列中包含的多尺度复杂模态有效分离的问题。其次，引入了长短期记忆神经网络（LSTM）解决了在时序预测的非线性问题。最后，通过粒子群算法对修正关系式的权重进行最优化处理。具体预测流程图如图 5-1 所示。

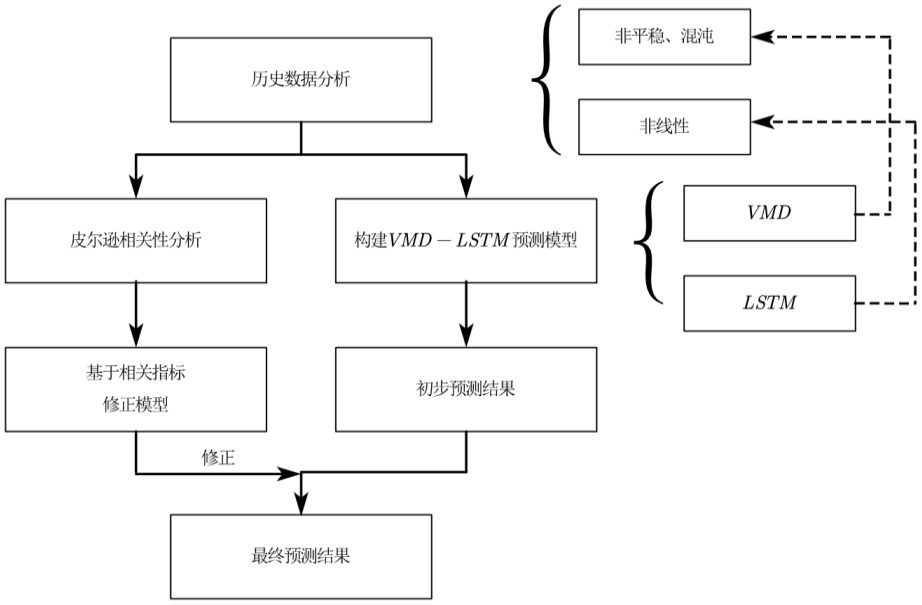


图 5-1 基于 VMD-LSTM 的预测流程图

5.2 问题二模型的准备

5.2.1 变分模态分解

VMD 是一种自适应准正交信号分解新方法[5]，与 EMD 的循环筛分求解方式不同的是，VMD 在变分框架内通过递归地求解变分问题实现信号分解，由变分问题的最优解即可确定各分量的中心频率和有限带宽。其求解过程具有较好的自适应性，突出了信号的局部特征。VMD 的求解过程包括变分问题的构造与求解两步[6]，涉及经典维纳滤波、希尔伯特变换和频率混合等概念。首先构造变分问题，假设 VMD 将给定信号分解为 K 个具有中心频率和有限带宽的模态分量，则变分问题的优化目标即寻求各模态分量的估计带宽之和最小。

对于第 k 个模态分量 $m_k(t)$，为得到其单边频谱，通过希尔伯特变化将其转换为解析信号：

$$\left(\delta(t)+\frac{j}{\pi t}\right)*m_k(t) \tag{5-1}$$

对 $m_k(t)$ 的解析信号混合—中心频率，将其频谱调制到相应的基频带：

$$\left[\left(\delta(t)+\frac{j}{\pi t}\right)*m_k(t)\right]e^{-jw_k t} \tag{5-2}$$

通过计算上述解调信号梯度的平方 L^2 范数即可估计出分量 $m_k(t)$ 的带宽。假设 $m_k(t)$ 与 w_k 分别对应给定信号 $x(t)$ 分解后第 k 个模态的时域信号和中心频率，则约束变分问题描述如下：

$$\min_{m_k,w_k}\left\{\sum_k\left\|\partial_t\left[\left(\delta(t)+\frac{j}{\pi t}\right)*m_k(t)\right]e^{-jw_kt}\right\|_2^2\right\}$$
$$s.t.\ \sum_{k=1}^{K}m_k(t)=x(t),\quad k=1,2,...,K \tag{5-3}$$

式（5-3）中，K 为分解得到的模态总数。

为求解该变分问题，引入二次惩罚项和 Lagrange 乘子，其中二次惩罚项用于降低高斯噪声的干扰，Lagrange 乘子则为增强约束的严格性，增广变分问题如下：

$$L(m_k,w_k,\beta)=\alpha\sum_k\left\|\partial_t\left[\left(\delta(t)+\frac{j}{\pi t}\right)*m_k(t)\right]e^{-jw_kt}\right\|_2^2+\left\|f(t)-\sum_k m_k(t)\right\|_2^2+\left\langle\beta(t),f(t)-\sum_k m_k(t)\right\rangle \tag{5-4}$$

式（5-4）中，α 为惩罚参数。

利用基于对偶分解和 Lagrange 法的交替方向乘子方法（ADMM）求解变分问题[7]，对 m_k、w_k 与 β 进行交替迭代寻优，可得如下迭代公式。

$$m_k^{n+1}(w)=\frac{f(w)-\sum_{i\neq k}m_i(w)+\frac{\beta(w)}{2}}{1+2\alpha(w-w_k)^2} \tag{5-5}$$

$$w_k^{n+1}=\frac{\int_0^\infty w|m_k(w)|^2dw}{\int_0^\infty|m_k(w)^2|dw} \tag{5-6}$$

$$\beta^{n+1}=\beta^n+\tau\left(f-\sum_i m_i\right) \tag{5-7}$$

对于给定求解精度 ε，满足式（5-8）时停止迭代：

$$\sum_k\left\|m_k^{n+1}-m_k^n\right\|_2^2<\varepsilon \tag{5-8}$$

VMD 的具体迭代求解过程如下：

Step1：初始化 m_k^1、w_k^1、β^1 与最大迭代次数 N，$n=0$；

Step2：更新 m_k、w_k；

Step3：更新 β，$n=n+1$；

Step4：判断收敛性，若不收敛且 $n<N$，则重复 Step2，否则停止迭代，得到最终模态函数 m_k 和中心频率 w_k。

5.2.2 长短期记忆神经网络

长短期记忆网络（LSTM）是在循环神经网络（RNN）的基础上发展出的机器学习模型[8]。在 RNN 中，每个神经元有两个权重向量 W_x、W_y，分别对应于输入 $x(t)$ 和输出 $y(t-1)$。式（5-9）为用向量化形式来表示的网络的整层输出：

$$Y(t)=\phi\left(X(t)\cdot W_x+Y(t)\cdot W_y+b\right) \tag{5-9}$$

从式（5-9）中可以看出，RNN 中的神经元在时间迭代 t 的输出包含了之前时间迭代内的所有输入信息。而包含之前时间迭代信息的神经网络被称为记忆单元。通常一个单元在时间迭代 t 的状态被记作 $h(t)$，表示在某个时间迭代的输入和它在前一时间迭代状态的函

数：

$$h(t)=f\left(h(t-1),x(t)\right) \tag{5-10}$$

在简单的基础单元中输出 $y(t)$ 等于状态 $h(t)$。LSTM 主要通过在每个神经元中加入记忆单元来实现过去信息的记忆，而每个记忆单元则由特殊的门所控制，这些门将决定历史信息是否被写入、读取或清空。

LSTM 单元将状态分成了短期状态 $h(t)$ 和长期状态 $c(t)$。LSTM 网络可以通过输入数据组学习在长期状态下储存什么，忘记什么以及从什么中进行读取。如图 5-2 所示，长期状态 $c(t-1)$ 首先经过忘记门，丢弃部分记忆，同时输入门会选择增加部分记忆，最后输出 $c(t)$。此外，长期状态 $c(t)$ 会被复制到 tanh 函数，然后被输出门过滤产生短期状态 $h(t)$。

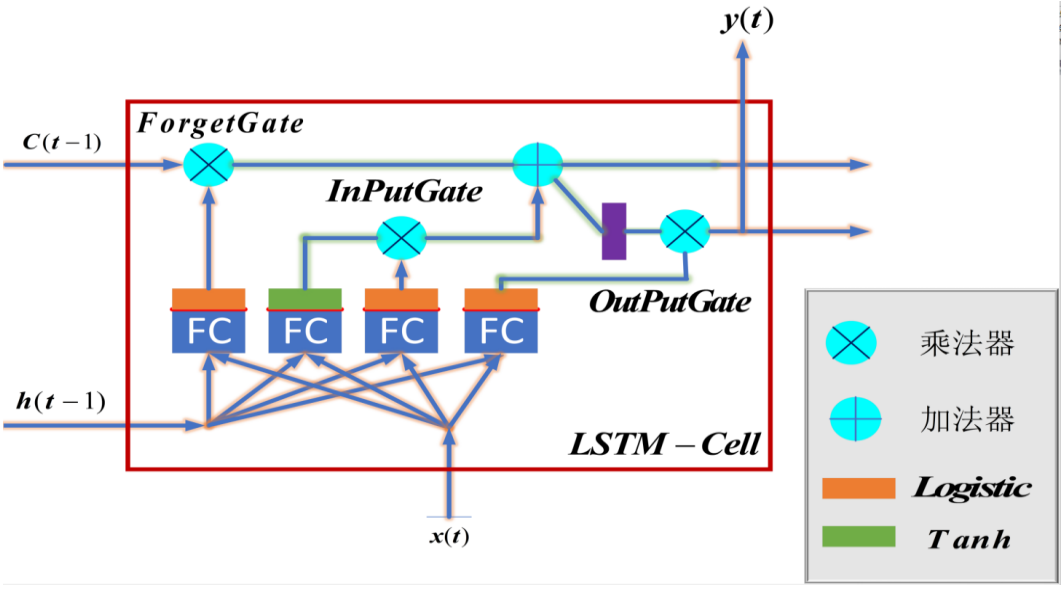


图 5-2　LSTM 网络神经元结构

综上所述，LSTM 的功能如下：（1）通过对输入数据的学习识别到重要输入信息；（2）将重要的信息存储到长期状态 $c(t)$ 中；（3）通过学习识别到何时对重要信息进行保存以及何时提取重要输入信息。由于 LSTM 具有这些强大的学习功能，故其可以成功地捕捉到输入序列中的长期模式，被广泛应用于时间序列预测领域。

5.2.3 数据的标准化处理

在 LSTM 神经网络模型中，输入变量之间的数量级差异会较大地影响模型对该输入变量的判断，并且很大程度地影响模型迭代时 loss 的收敛速度。而标准化处理过后的数据去除了量纲，使得输入变量之间具有可比性，对于模型而言，求解过程变得更平缓，更容易找到最优解。由于本文处理得到的待标准化处理数据均为连续化数据，所以可以使用 Z-score 标准化方法对输入数据进行标准化处理[9]，公式如（5-11）所示：

$$x_i^{'}=\frac{x_i-\bar{x}}{\sigma} \tag{5-11}$$

其中，$\bar{x}$ 为输入变量的平均值，σ 为输入变量的标准差。

5.2.4 粒子群算法

粒子群算法启源于鸟群觅食行为，鸟群在觅食过程中会经常改变飞行状态，时而散开时而聚集，有时还会改变飞行方向，但整个群体依然会保持统一协调性，从而高效有序的寻找到食物；每一个粒子代表着鸟群中的一只鸟，应用粒子群算法求解问题时，问题的解就是其中某一只鸟的位置[10]。设基本粒子群算法由 N 个粒子构成，在 t 时刻第 i 个粒子的位置向量为：

$$x_i(t)=[x_1(t),x_2(t),...x_N(t)],\ x_{\min}(t)\le x_i(t)\le x_{\max}(t) \tag{5-12}$$

每个粒子都具有决定移动方向和位置的速度，t 时刻第 i 个粒子的速度向量为：

$$v_i(t)=[v_1(t),v_2(t),...v_N(t)],\ v_{\min}(t)\le v_i(t)\le v_{\max}(t) \tag{5-13}$$

其中，$x_{\min}$、$x_{\max}$ 为粒子搜索空间的上下限，$v_{\min}$、$v_{\max}$ 为粒子移动速度的限值。

在迭代过程中，粒子通过追踪两个极值来更新：每个粒子移动过程中距离最优解最近的位置称为个体极值 $p_i(t)=(p_1,p_2,...p_N)$，整个粒子群体寻找到最好的解称为全局极值 $p_g(t)$。基本的粒子群算法具体更新公式如下：

$$\begin{cases} v_i(t+1)=v_i(t)+c_1r_1[p_i(t)-x_i(t)]+c_2r_2[p_g(t)-x_i(t)] \\ x_i(t+1)=x_i(t)+v_i(t+1) \end{cases} \tag{5-14}$$

式中，c_1、c_2 称为学习因子，r_1、r_2 为[0,1]随机产生的数。

在实际寻优过程中，希望粒子群能开拓搜索空间快速收敛到一定范围，需要较强的全局搜索能力，随之在该范围内细化搜索获得期望最优解，需要较强的局部搜索能力；因此在式(5-14)中添加一个惯性权重 w，更新公式变为：

$$\begin{cases} v_i(t+1)=wv_i(t)+c_1r_1[p_i(t)-x_i(t)]+c_2r_2[p_g(t)-x_i(t)] \\ x_i(t+1)=x_i(t)+v_i(t+1) \end{cases} \tag{5-15}$$

w 越大粒子群获得越强的全局搜索能力，w 越小粒子群获得越强的局部搜索能力；通常采用线性递减惯性权重：

$$w(it)=w_a-it\frac{w_a-w_b}{it_{\max}} \tag{5-16}$$

式中，it 为当前迭代数，$it_{\max}$ 为最大迭代次数，一般取 $w_a=0.9$、$w_b=0.4$，上述算法称为标准粒子群算法。

5.3 问题二模型的建立

针对问题二，需要对每 5 分钟的“数字经济”板块指数（成交量）进行预测。根据第一问得出的和成交量有关的主要指标分别为技术指标中的 VMA，国际市场指标中的伦敦金融时报 100 指数，其他板块中的数字孪生和快手概念。首先利用 VMD 对成交量数据进行分解，解决了金融时序数据非稳定的问题；然后引入了 LSTM 对 VMD 分解后的各模态分量进行预测，解决了金融时序数据非线性的问题；最后基于与成交量相关的主要指标对预测模型进行修正，得到最终的预测结果。

5.4 问题二模型的求解及结果分析

5.4.1 成交量数据的变分模态分解

在进行 VMD 分解前需要事前确定模态个数 K 值、惩罚因子 α、保真度系数 τ 以及收

敛停止条件 ε。选取 $\alpha=2500, \tau=0, \varepsilon=10^{-7}, K=5$，对成交量数据进行 VMD 分解得到如下 5 组变分模态函数，如图 5-3 所示。

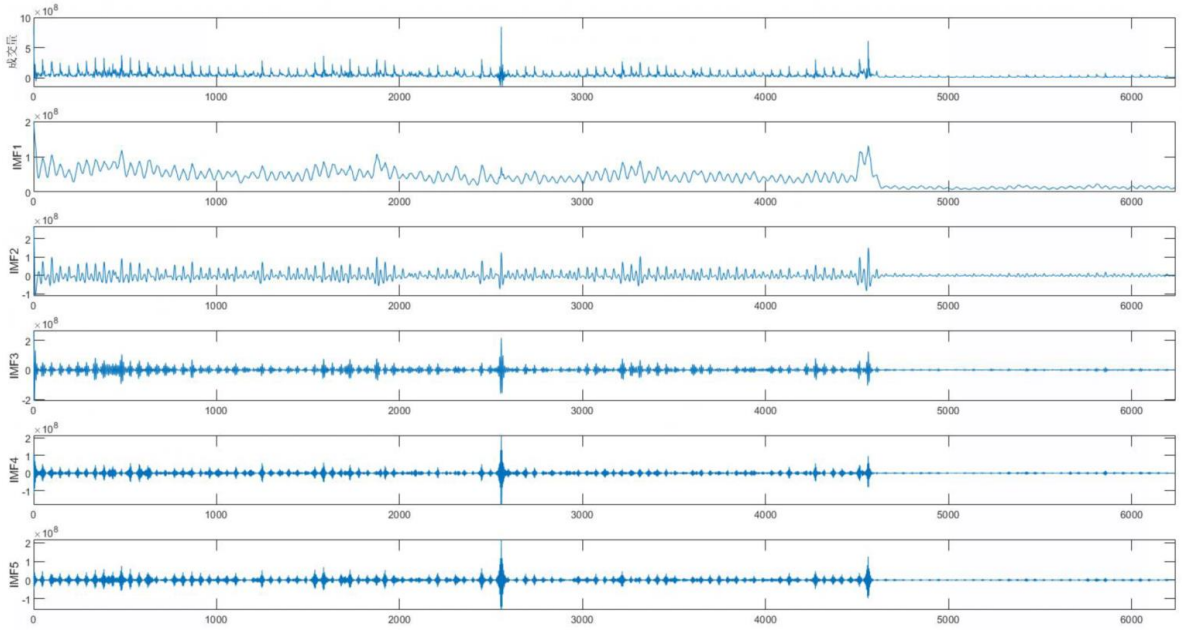
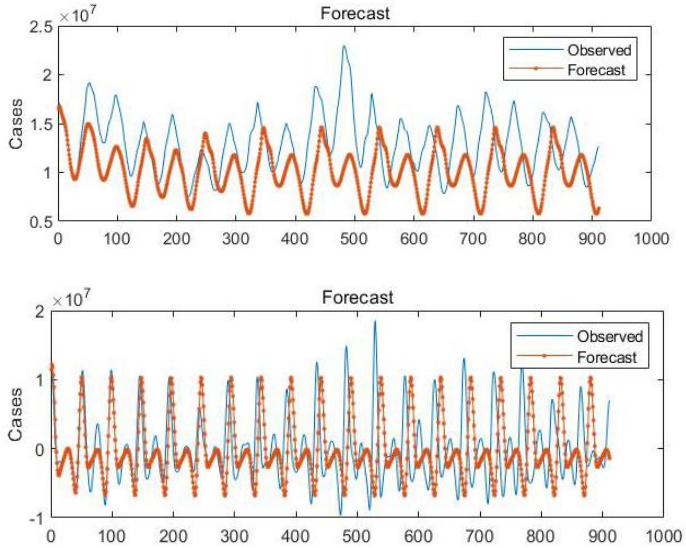


图 5-3　当 K=5 时，经 VMD 分解后的变分模态分量

5.4.2 各模态分量基于 LSTM 模型的预测

我们选取了从 2021-7-14 到 2021-12-31 日的每 5 分钟“数字经济”板块数据在模型中进行训练，并对 2022-1-4 到 2022-1-28 共计 25 天的每 5 分钟成交量数据进行预测，用 LSTM 算法对各本征模态序列进行预测，预测结果如图 5-4 所示。



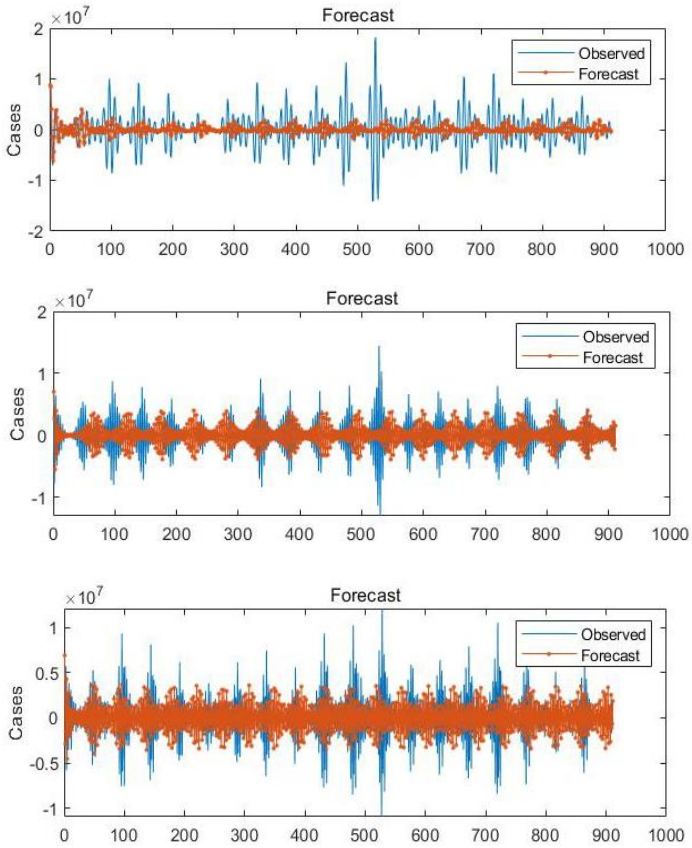


图 5-4 各分量序列基于 LSTM 模型的预测结果

从图 5-4 中可以看出，LSTM 模型对于 IMF1 和 IMF2 的预测效果相对较好，对于 IMF3、IMF4 和 IMF5，LSTM 模型无法很好的捕捉到大涨和大跌。

将上述各分量预测结果进行加总，得到基于 VMD-LSTM 的成交量预测结果如图 5-5 所示。

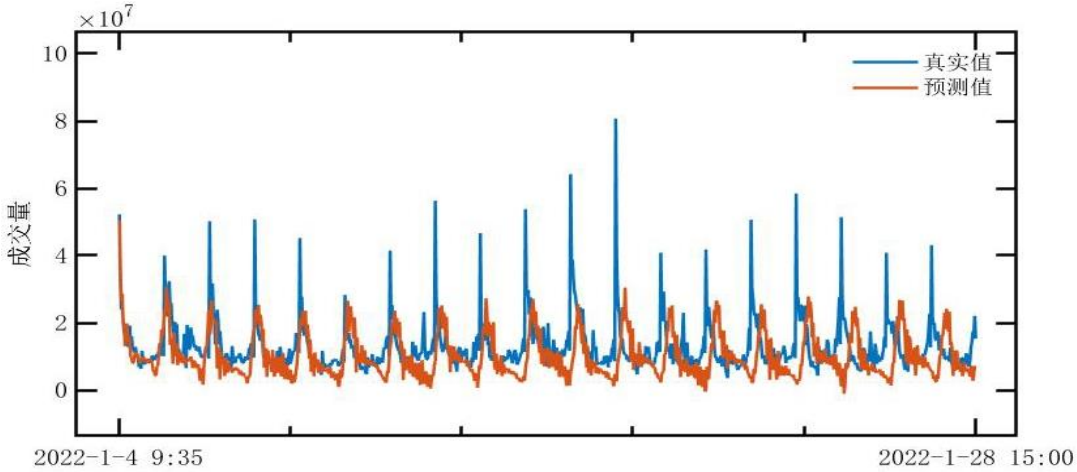


图 5-5 基于 VMD-LSTM 的成交量预测结果

从图 5-5 中可以看出，基于 VMD-LSTM 的成交量预测模型对于大涨、大跌的捕捉依然不显著。因此，在基于该预测模型构建投资策略时，需要针对尾部风险进行特别分析。

5.4.3 模型的评价指标

为了验证本文提出的模型的有效性，本文采用了三个常用的标准来评价所提出的模型的性能，即平均绝对误差（MAE）、均方根误差（RMSE）和平均绝对百分比误差（MAPE）。MAE 反映预测值与实测值的平均系统性偏离程度，RMSE 反映预测值与实测值的平均绝对偏离程度，MAPE 反映预测值与实测值的偏离相对于实测值的平均偏离程度。这三种常用准则的数学公式描述如公式（5-17）、（5-18）和（5-19）所示。

$$MAE=\frac{1}{N}\sum_{t=1}^{N}\left|\hat{y}(t)-y(t)\right| \tag{5-17}$$

$$RMSE=\sqrt{\frac{1}{N}\sum_{t=1}^{N}\left(\hat{y}(t)-y(t)\right)^{2}} \tag{5-18}$$

$$MAPE=\frac{1}{N}\sum_{t=1}^{N}\left|\frac{\hat{y}(t)-y(t)}{y(t)}\right| \tag{5-19}$$

式（5-17）、（5-18）和（5-19）中，N 为测试的样本数，$y(t)$ 是 t 时刻的真实值，$\hat{y}(t)$ 是 t 时刻的预测值。VMD-LSTM 模型的预测误差如表 5-1 所示。

表 5-1　VMD-LSTM 模型的预测误差

VMD-LSTM 模型	
MAE	0.7007
RMSE	1.0871
MAPE（%）	2.7626

该模型的 MAE、RMSE 和 MAPE 分别为 0.7007、1.0871 和 2.7626%，都是比较小的值，所以本文提出的模型适用于成交量的预测，并能取得较好的预测效果。

5.4.4 成交量相关性分析

对成交量和与成交量有关的四个主要指标进行相关性分析，通过 Matlab 拟合得到式（5-20）：

$$y_1=0.0498x_1+4087.5x_2-75836x_3+5110x_4 \tag{5-20}$$

式中，y_1 表示成交量，x_1,x_2,x_3,x_4 分别表示 VMA，伦敦金融时报 100 指数，数字孪生和快手概念。

为了使预测结果更加准确，将式（5-20）和 LSTM 预测的序列 V_{LSTM} 进行联合并通过粒子群优化算法寻找出最优的权重组合。

设置粒子群算法的学习因子 $c_1=0.4,c_2=0.6$，惯性权重 w 采用式（5-16）所示的线性递减策略；设置 100 个粒子在二维欧几里得空间[0,1]区间范围内进行搜索，选择适应度函数如下：

$$J_{rmse}=\sqrt{\sum_{i=1}^{T}\frac{1}{T}\left(y_i-y\right)^{2}} \tag{5-21}$$

在进行 100 次迭代之后，最终粒子群收敛的全局最优位置 k_1,k_2 即为 y_1 和 V_{LSTM} 的权重。可以得到修正预测的表达式如下：

$$C_1=0.5132y_1+0.351V_{LSTM} \tag{5-22}$$

5.4.5 预测结果的修正

我们选取了从 2021-7-14 到 2021-12-31 日的每 5 分钟“数字经济”板块数据在模型中进行训练，并对 2022-1-4 到 2022-1-28 共计 25 天的每 5 分钟成交量数据进行预测。将直接应用 VMD-LSTM 预测的结果与经过修正的 VMD-LSTM 预测的结果进行对比来检验所提方法的有效性，如图 5-6 所示。

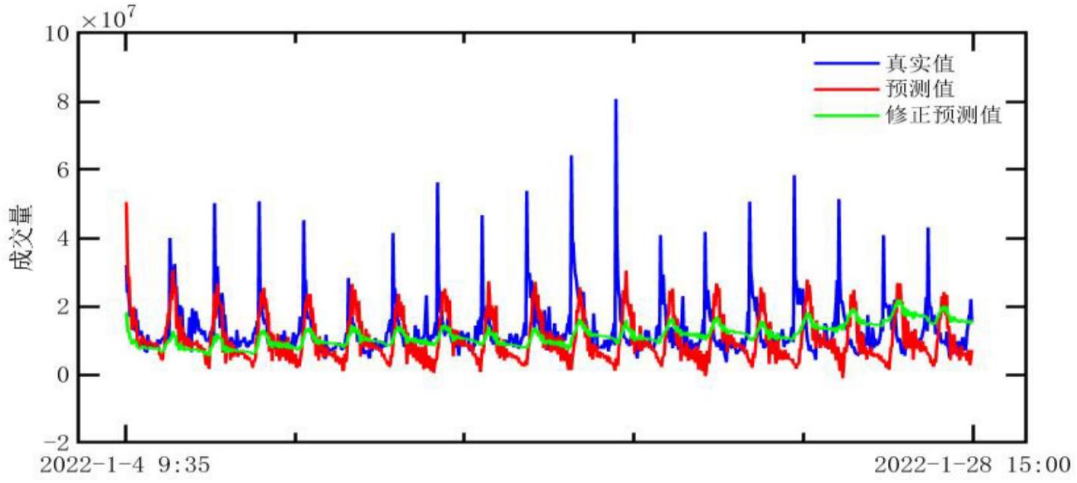


图 5-6 预测结果对比

图 5-6 较明显的反映出修正过后的模型更好地预测了股价的总体趋势。

六：问题三模型的建立与求解

6.1 问题三模型的建立

针对问题三，需要对每 5 分钟的“数字经济”板块指数（收盘价）进行预测。根据第一问得出的和收盘价有关的主要指标分别为技术指标中的 BBI、MA、EXPMA 和 BOLL。首先利用 VMD 对收盘价数据进行分解，解决了金融时序数据非稳定的问题；然后引入了 LSTM 对 VMD 分解后的各模态分量进行预测，解决了金融时序数据非线性的问题；最后基于与收盘价相关的主要指标对预测模型进行修正，得到最终的预测结果。

6.2 问题三模型的求解及结果分析

对收盘价数据进行 VMD 分解，在进行 VMD 分解前需要事前确定模态个数 K 值、惩罚因子α、保真度系数τ以及收敛停止条件ε。选取$\alpha=2500,\tau=0,\varepsilon=10^{-7},K=4$，对收盘价数据进行 VMD 分解得到如下 4 组本征模函数，如图 6-1 所示。

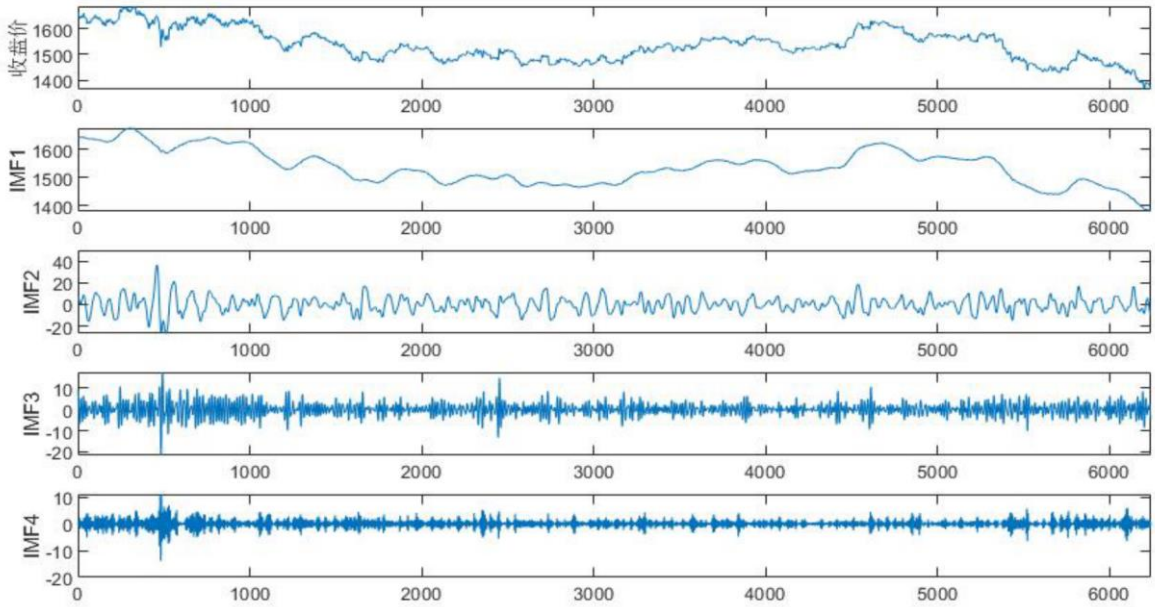


图 6-1 当 K=4 时，经 VMD 分解后的本征模态分量

输入数据经过 VMD 分解后，我们选取了从 2021-7-14 到 2021-12-31 日的每 5 分钟“数字经济”板块数据在模型中进行训练，并对 2022-1-4 到 2022-1-28 共计 25 天的每 5 分钟收盘价数据进行预测，用 LSTM 算法对各本征模态序列进行预测，预测结果如图 6-2 所示。

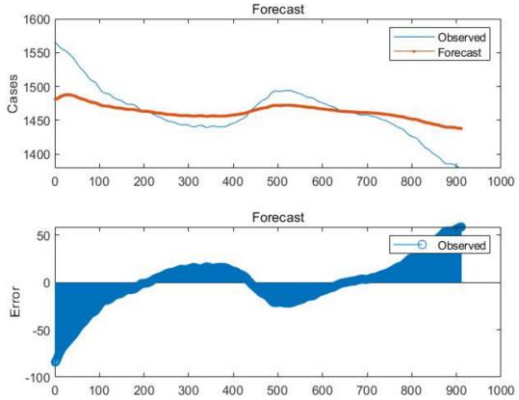


图 6-2（a）IMF1 基于 LSTM 模型的预测结果

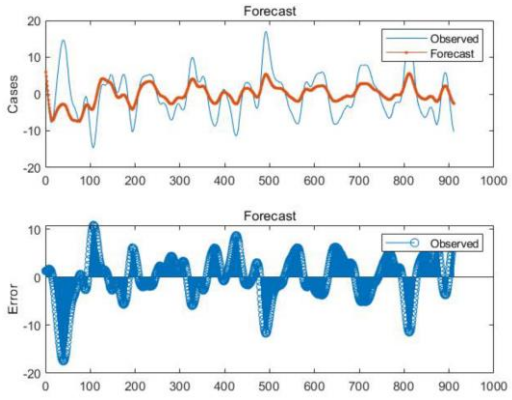


图 6-2（b）IMF2 基于 LSTM 模型的预测结果

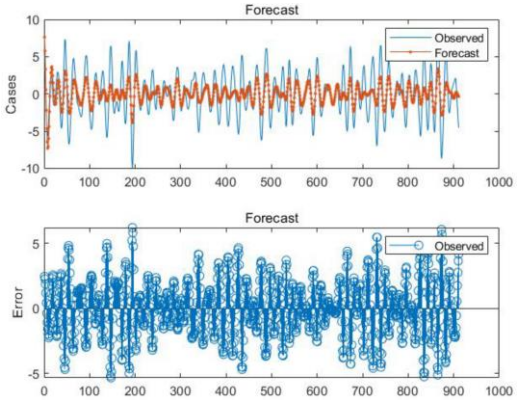


图 6-2（c）IMF3 基于 LSTM 模型的预测结果

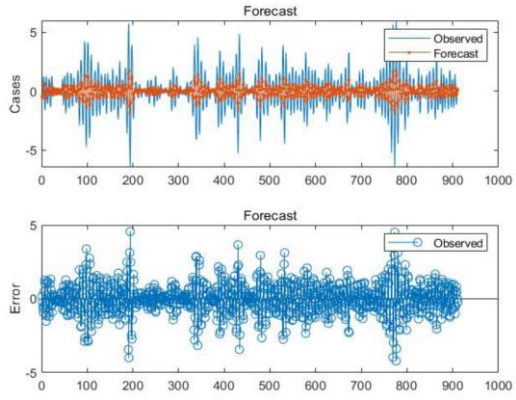


图 6-2（d）IMF4 基于 LSTM 模型的预测结果

图 6-2 各分量序列基于 LSTM 模型的预测结果

将上述各分量预测结果进行加总，得到基于 VMD-LSTM 的收盘价预测结果如图 6-3

所示。

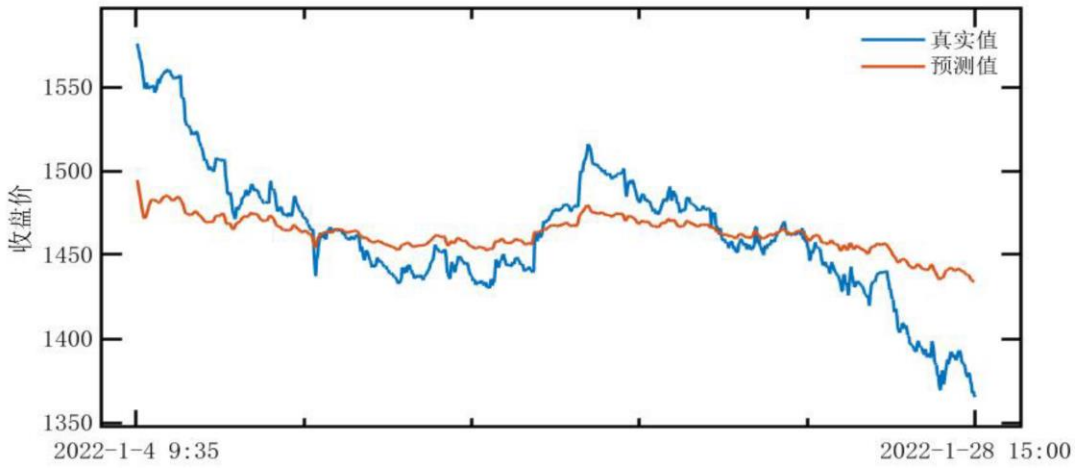


图 6-3 基于 VMD-LSTM 的收盘价预测结果

从图 6-3 中可以看出，基于 VMD-LSTM 的收盘价预测模型对于大涨、大跌的捕捉依然不显著。因此，在基于该预测模型构建投资策略时，需要针对尾部风险进行特别分析。

对收盘价和与收盘价有关的四个主要指标进行相关性分析，通过 Matlab 拟合得到式（6-1）：

$$y_2 = -0.3029x_5 + 1.2871x_6 + 0.04x_7 - 0.0123x_8 \quad (6\text{-}1)$$

通过粒子群算法求得式（6-1）和 LSTM 预测的序列V_{LSTM}的最优权重分别为 0.7538 和 0.2278，即有：

$$C_2 = 0.7538y_2 + 0.2278V_{LSTM} \quad (6\text{-}2)$$

选取了从 2021-7-14 到 2021-12-31 日的每 5 分钟“数字经济”板块数据在模型中进行训练，并对 2022-1-4 到 2022-1-28 共计 25 天的每 5 分钟收盘价数据进行预测。将直接应用 VMD-LSTM 预测的结果与经过修正的 VMD-LSTM 预测的结果进行对比来检验所提方法的有效性，如图 6-4 所示。

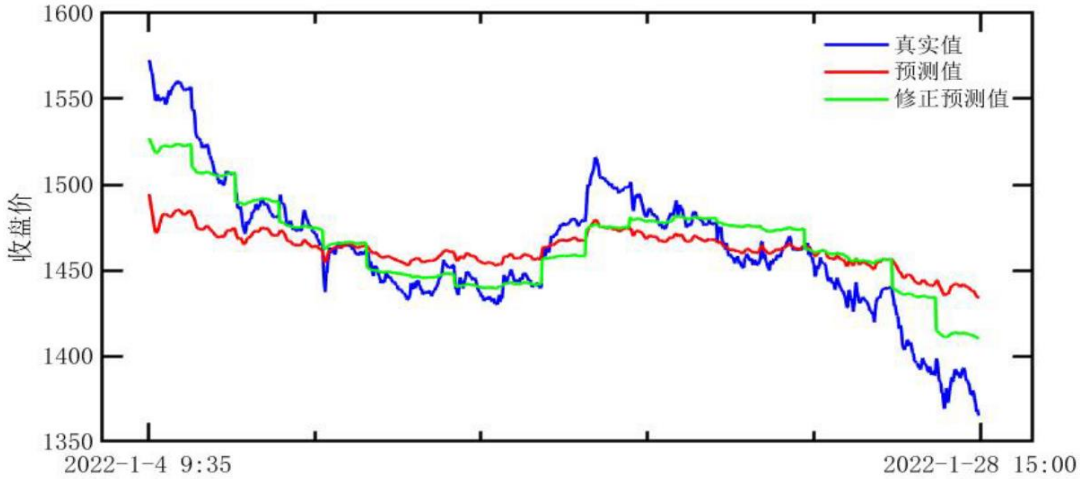


图 6-4 预测结果对比

图 6-4 较明显的反映出修正过后的模型更好地预测了股价的总体趋势。

6.3 模型的交叉验证

为了更好地反应模型的泛化能力，我们采用交叉验证的方式来对模型进行评判。交叉验证是一种没有任何前提假定直接估计泛化误差的模型选择方法，解决了同一数据集同时进行训练和预测带来的误差估计问题。基本思想为：在给定的模型样本中，取出大部分的样本来建模，即作为训练集；留下小部分样本用建立好的模型进行预测，即作为测试集。这样的过程持续进行，直至所有的样本都被预测了一次。

由于股票数据是时间序列数据，考虑采用时间序列交叉验证。将“数字经济”板块的6240个收盘价数据分为六组，分别为：1—1000、1001—2000、2001—3000、3001—4000、4001—5328、5329—6240。其中5329—6240为2022年1月4日到2022年1月28日每隔5分钟的收盘价数据。采用正向链接的方式，首先用第1组数据作为训练集，第2组数据作为测试集；然后选第1组和第2组数据作为训练集，第3组数据作为测试集；以此类推，直到第6组数据完成预测。具体示意图如图6-5所示。

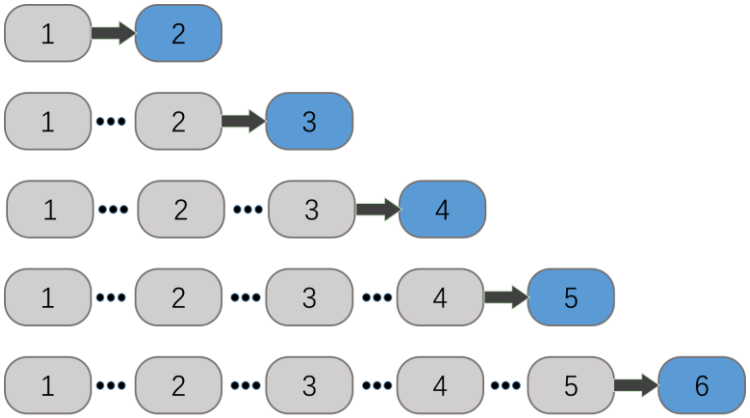


图6-5 时间序列交叉验证示意图

以收盘价数据进行VMD分解的第一个本征模态分量IMF1为例，其时间序列交叉验证结果如图6-6（a）~（e）所示。

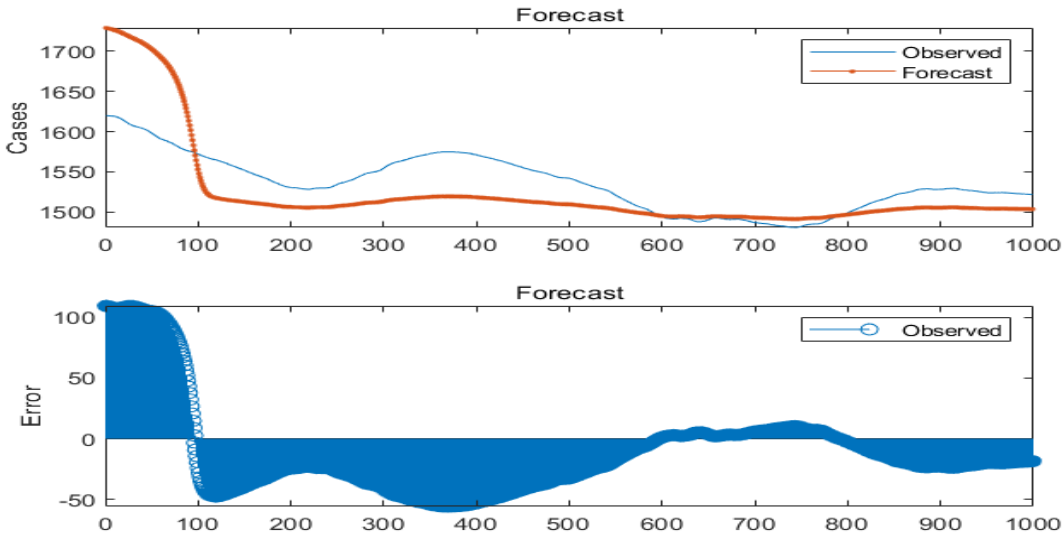


图6-6（a） 交叉验证（1—2）

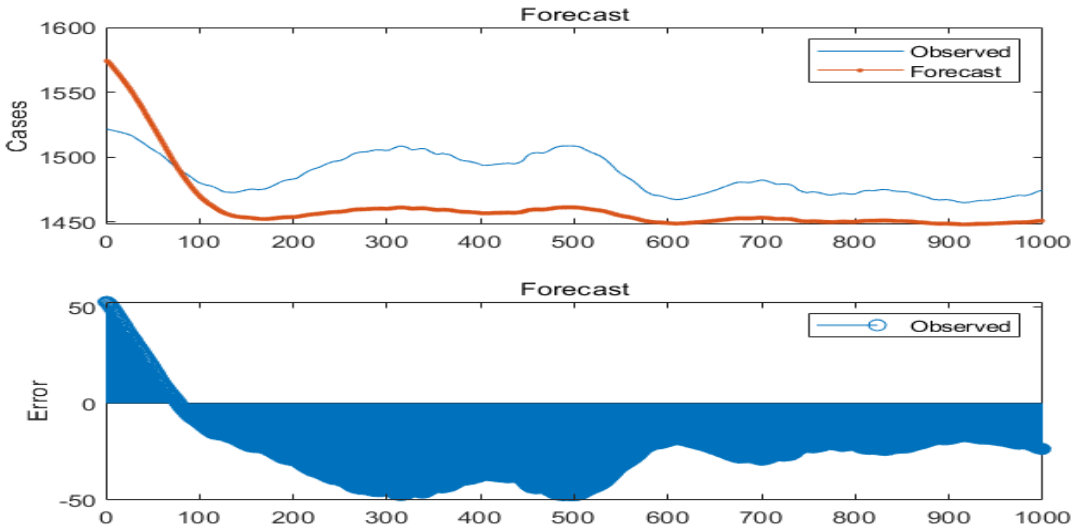


图 6-6（b） 交叉验证（1、2—3）

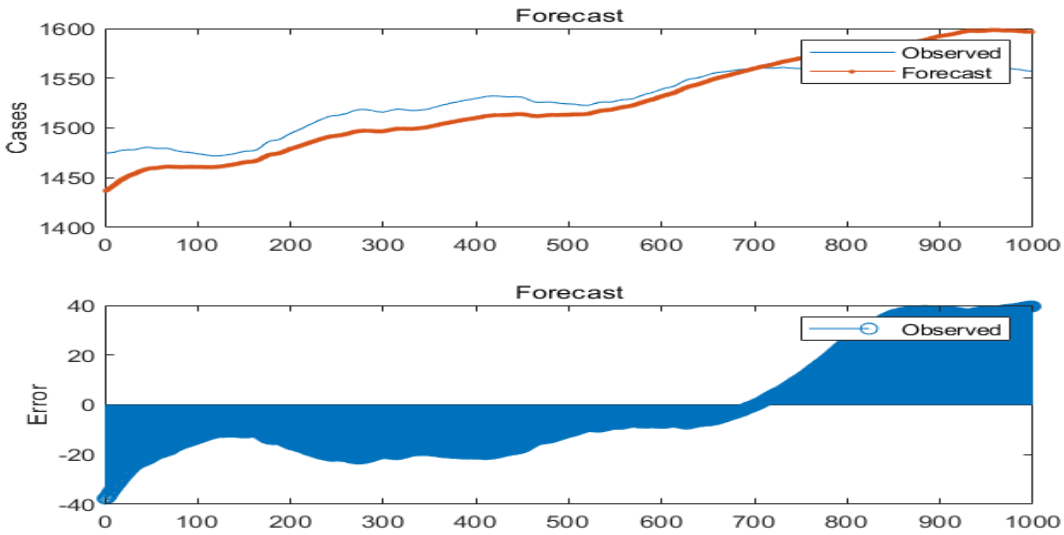


图 6-6（c） 交叉验证（1、2、3—4）

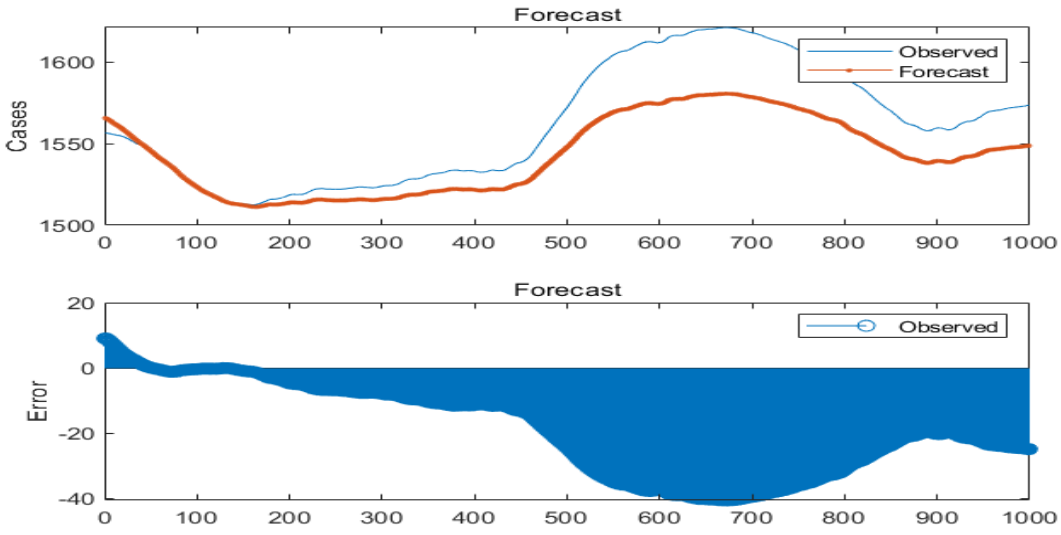


图 6-6（d） 交叉验证（1、2、3、4—5）

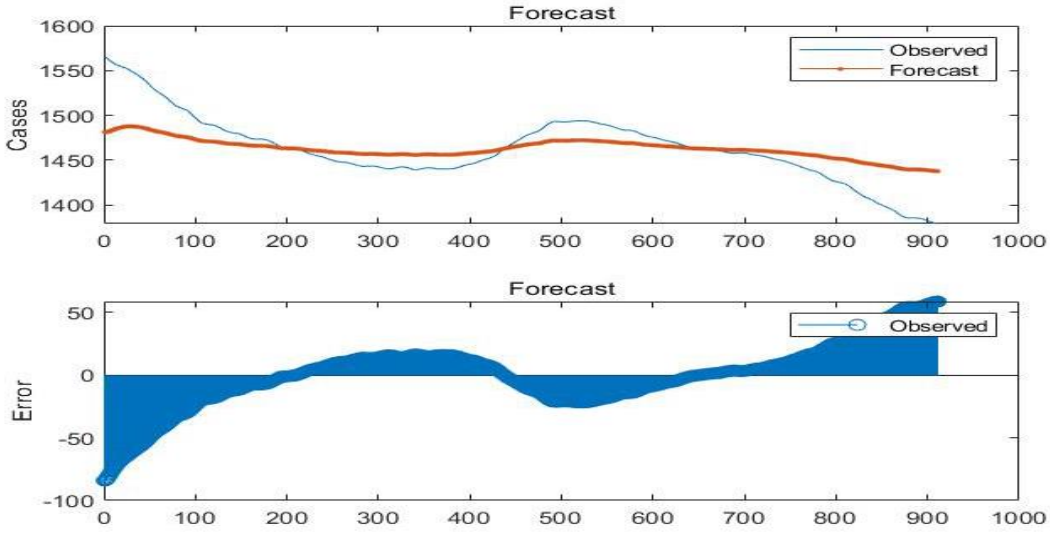


图 6-6（e）交叉验证（1、2、3、4、5—6）

七：问题四模型的建立与求解

7.1 问题四的分析

针对问题四，需要根据问题三预测得到的收盘价序列来对“数字经济”板块每 5 分钟频率价格进行买卖交易。由于交易频率较高且股票的浮动较大，给量化投资造成了一些困难。因此我们需要参考一些技术指标来进行量化投资。首先我们考虑引入相关强弱指标（RSI），其通过计算股价涨跌的幅度来推测市场运动趋势的强弱度，并据此预测趋势的持续或者转向。它显示的是股价向上波动的幅度占总的波动幅度的百分比，在很多情况下，市场很好的买卖讯号是：RSI 进入超买超卖区，然后又穿过超买或超卖的界线回到正常区域。需要得到价格方面的确认，才能采取实际的行动。因此我们也参考了对数移动均线指标（LOGMA）。与 RSI 指标一样，布林线（BOLL）指标也是股票市场实用的技术分析参考指标，它反映了股价的波动状况。为了能够排除风险因素对绩效评估的不利影响，引入了夏普比率，可以同时对风险和收益加以综合考虑。

7.2 问题四模型的建立

7.2.1 相对强弱指标

相对强弱指标是期货市场和股票市场中最为著名的摆动指标[11]，其原理就是通过计算股价涨跌的幅度来推测市场运动趋势的强弱度，并据此预测趋势的持续或者转向。实际上它显示的是股价向上波动的幅度占总的波动幅度的百分比，如果其数值大，就表示市场处于强势状态，如果数值小，则表示市场处于弱势。通常该指标称作 RSI 指标。这是一个测市的重要指标，但他的作用与股票和期货的市值大小有着密切的关系，盘子大的波动幅度小，盘子小的波动幅度大。其具体计算公式为：

$$RSI(n) = A/(A+B)\times 100\% \qquad (7\text{-}1)$$

其中，*A* 表示 *n* 天中股价向上波动的幅度大小，而 *B* 则表示 *n* 天中股价向下波动的大小。*A+B* 表示股价在此期间总的波动幅度大小。

运用原则如下：

1）不论价位如何变动，强弱指标的值均在 0 与 100 之间，强弱指标保持高于 50 表示为强势市场，反之低于 50 表示为弱势市场。

2）强弱指标多在 30 与 70 之间波动。当六日指标上升到达 80 时，表示股市已有超买现象，如果一旦继续上升，超过 90 以上时，则表示已到严重超买的警戒区，股价已形成头部，极可能在短期内反转回转。

3）当六日强弱指标下降至 20 时，表示股市有超卖现象，如果一旦继续下降至 10 以下时则表示已到严重超卖区域，股价极可能有止跌回升的机会。

指标数值：

RSI 的变动范围在 0-100 之间，强弱指标值一般分布在 20-80。

80-100 极强 卖出

50-80 强 买入

20-50 弱 观望

0-20 极弱 买入

7.2.2 对数移动均线指标

移动平均线是将特定时间段内的价格求出一个平均值，然后随着将整个时间段向后推移一个时间单位得出下一个平均值，以此类推，将算出来的平均值用线连接起来，就得到了我们所说的移动平均线[12]。在移动均线的基础上做对数化处理，优点在于：1.可以缩小数据的绝对值，方便计算；2. 对数值小的变化差异的敏感程度比数值大的变化差异敏感程度更高。

7.2.3 布林线指标

布林线（Boll）指标是股市技术分析的常用工具之一，通过计算股价的“标准差”，再求股价的“信赖区间”[13]。该指标在图形上画出三条线，其中上下两条线可以分别看成是股价的压力线和支撑线，而在两条线之间还有一条股价平均线，我们将布林线指标的参数设为 20，这样一来股价将运行在压力线和支撑线所形成的通道中。与 RSI 指标一样，BOLL 指标也是股票市场最实用的技术分析参考指标，它反映了股价的波动状况。

布林线一般的应用规则是，当股价向下击穿支撑线的时候买点出现，而向上击穿阻力线卖点出现。而平均线是考验一个趋势（无论是上升还是下降或是盘整）是否得以继续的重要支撑或阻力。

7.2.4 夏普比率

在现代投资理论的研究表明，风险的大小在决定组合的表现上具有基础性的作用。风险调整后的收益率就是一个可以同时对收益与风险加以考虑的综合指标，能够排除长期风险因素对绩效评估的不利影响。夏普比率就是一个可以同时对收益与风险加以综合考虑的经典指标[14]。投资中的一个常规特点，即投资标的的预期报酬越高，投资人所能忍受的波动风险越高；反之，预期报酬越低，波动风险也越低。目标是在固定所能承受的风险下，追求最大的报酬；或在固定的预期报酬下，追求最低的风险。

计算公式如下：

$$SharpRatio = \frac{E(R_p) - R_f}{\sigma_p} \qquad (7\text{-}2)$$

其中，$E(R_p)$为投资预期年化报酬率，R_f为年化无风险利率，σ_p为投资年化报酬率的标准差。

目的是计算投资组合每承受一单位总风险，会产生多少的超额报酬。比率依据资产配置线的观念而来，是市场上最常见的衡量比率。当投资内的资产为风险性资产时，适用夏普比率。夏普指数代表投资人每多承担一分风险，可以拿到几分超额报酬；若大于 1，代表基金报酬率高过波动风险；若为小于 1，代表基金操作风险大过于报酬率。

7.3 问题四模型的求解

预测收盘价的相对强弱指标曲线如图 7-1 所示。

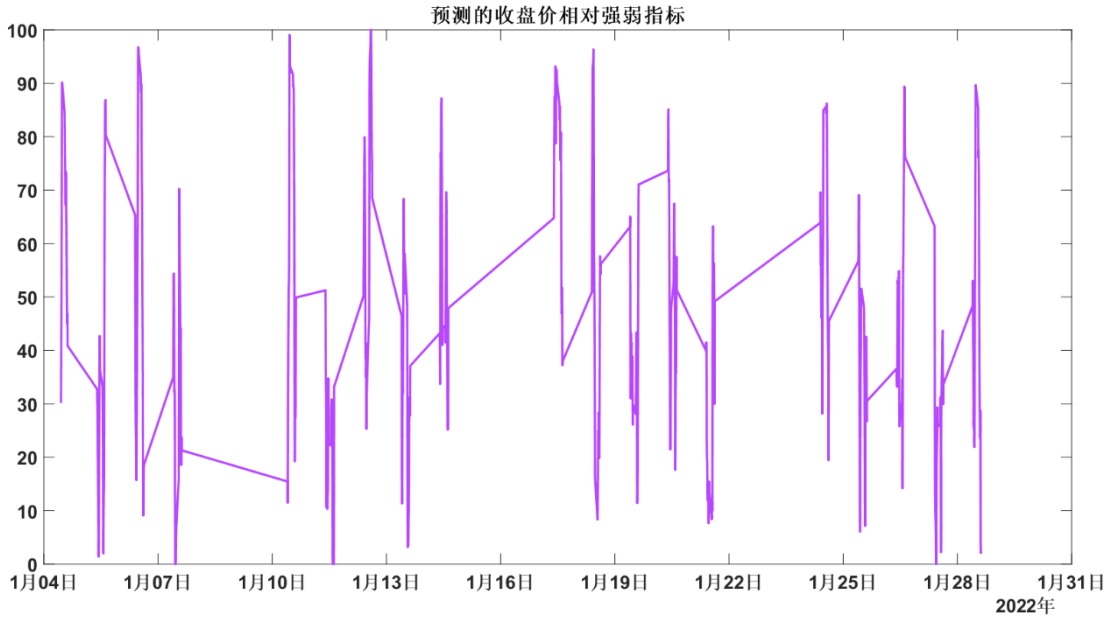


图 7-1 收盘价的相对强弱指标

除去 2022 年 1 月 8 号、9 号、15 号、16 号、22 号和 23 号这六天，当 RSI 指标大于 80 时有 115 次极强卖出信号；当 RSI 指标大于 50 小于 80 时有 228 次强买入信号；当 RSI 指标大于 20 小于 50 时有 399 次弱观望信号；当 RSI 指标小于 20 时有 155 次极弱买入信号。

在上述大量的买入卖出信号的基础上，我们还需要参考对数移动均线指标，对数移动均线如图 7-2 所示。

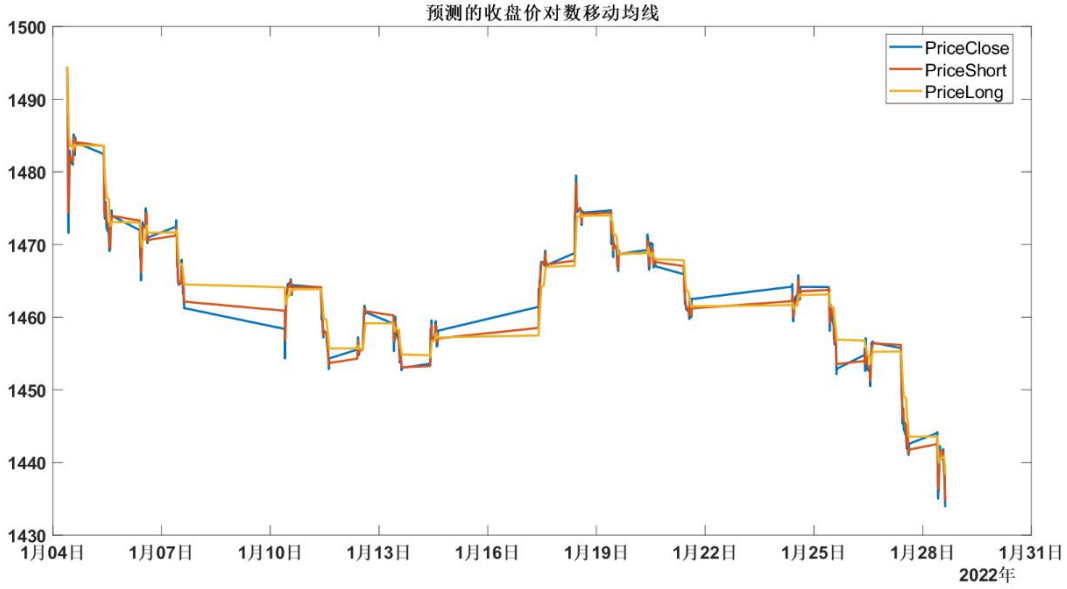


图 7-2 收盘价对数移动均线

当相对强弱指标出现强买和强卖信号时，我们再根据对数移动均线中的预测收盘价是否穿过短期线和长期线来参考是否采取买入或卖出的实际行动。当短期移动平均线向上穿过长期移动平均线时，是买入信号；当短期移动平均线向下穿过长期移动平均线时，是卖出信号。由图 7-2 可知，买入信号共 20 次，卖出信号共 19 次。

为了显示出股价的波动状况，绘制收盘价的布林曲线如图 7-3 所示。

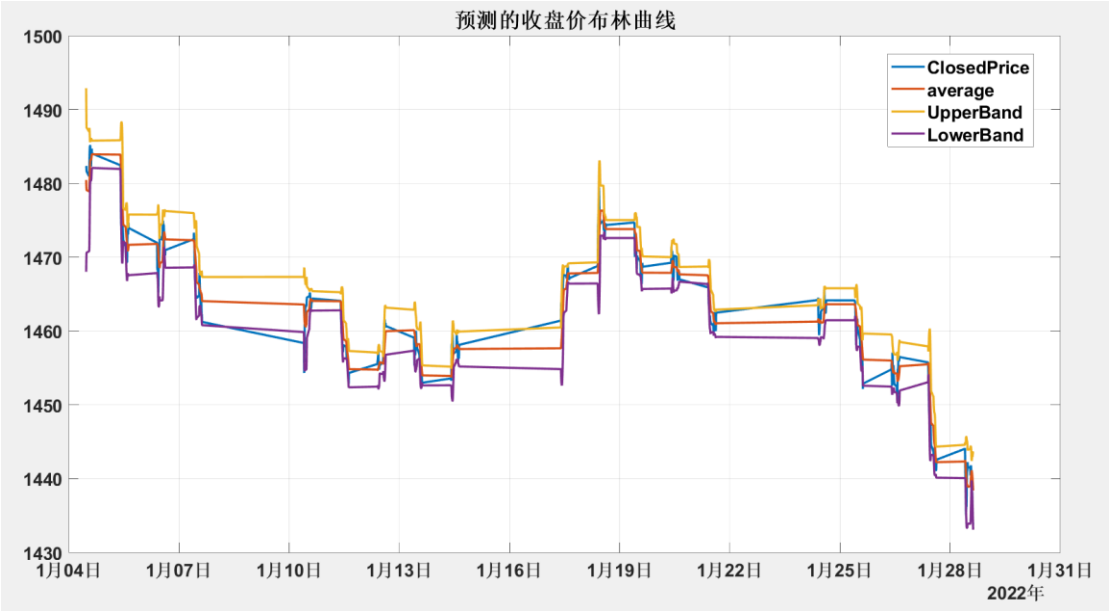


图 7-3 收盘价布林曲线

当股价向下击穿支撑线的时候买点出现，而向上击穿阻力线卖点出现。而平均线是考验一个趋势（无论是上升还是下降或是盘整）是否得以继续的重要支撑或阻力。我们采用的布林曲线策略如图 7-4 所示。

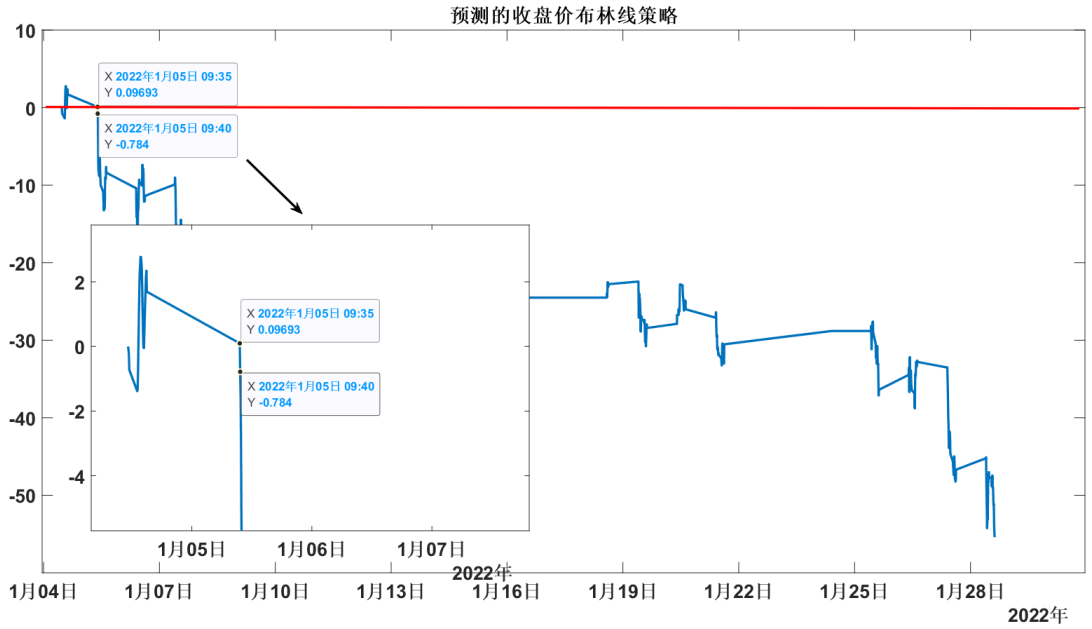


图 7-4 收盘价布林线策略

由图 7-4 可知， 1 月 5 日 9：40 之前不宜买入。

在上述买入卖出可以获得报酬的同时，我们还需要考虑规避风险，因此引入夏普比率指标。经过对预测收盘价进行夏普指数计算，得到夏普比率分布图如图 7-5 所示。

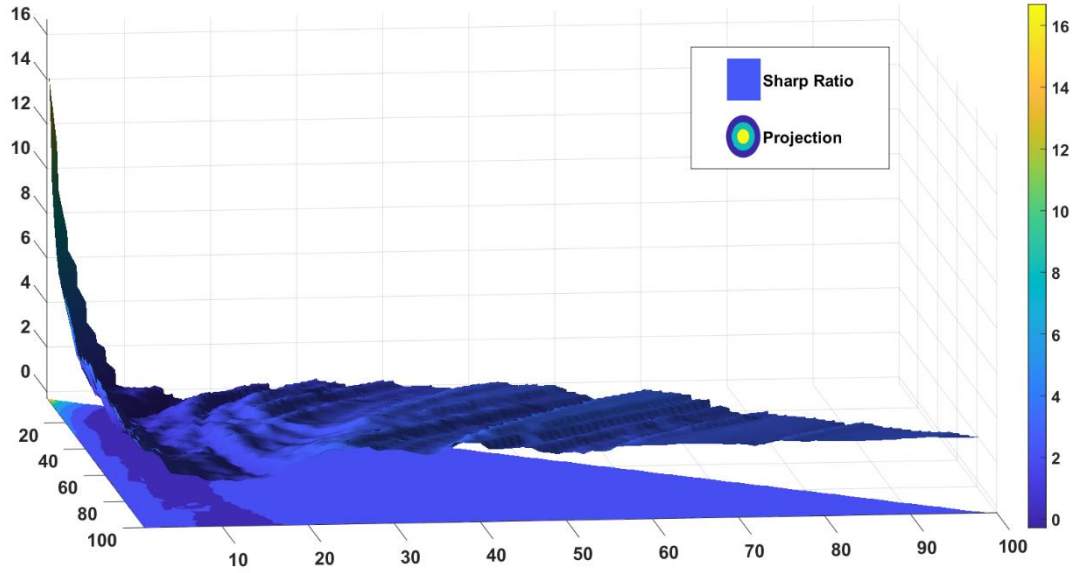


图 7-5 夏普比率分布

从图 7-5 中可以看出夏普比率指数分布大部分位于 1.6 附近，即说明报酬率高过波动风险，有利于投资。

因此结合上述的 RSI 指标、LOGMA 指标、BOLL 指标以及夏普比率，按照制定的投资策略，计算公式如式（7-3）~（7-6）所示：

$$C_{all}=\left[\prod_{i=1}^{19}\left(1+c_i\right)-1\right]\times 100\% \qquad (7\text{-}3)$$

其中，C_{all} 为总收益率，c_i 为第 i 日的当日收益率。

$$c_i = (p_i - p_{i-1}) / p_{i-1} \tag{7-4}$$

其中，p_i 为第 i 日收盘总资产市值。

$$I_i = \frac{c_i - c_{(i,500)} \times 90\%}{M \cdot \sigma_M} \tag{7-5}$$

其中，I_i 为第 i 日的信息比率，$c_{(i,500)}$ 为第 i 日中证 500 指数收益率，M 为交易日的天数，σ_M 为 M 日内的每日超额收益率序列求得的标准差。

设 D_i 为第 i 天的产品净值，D_j 是 D_i 后面第 j 天的净值，那么最大回撤率如式（7-6）所示：

$$Drawdown = \max \frac{D_i - D_j}{D_i} \tag{7-6}$$

按照上述指定的组合投资策略，得到收益曲线如图 7-6 所示：

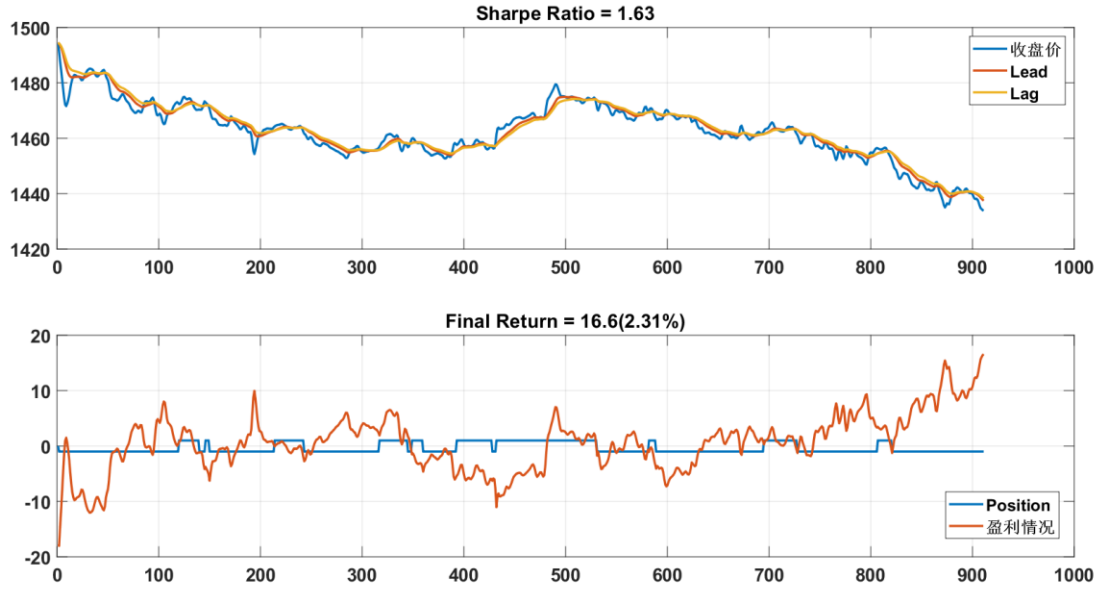


图 7-6 “数字经济”板块盈利情况

我们以每 5 分钟频率进行买卖交易，综合以上的组合投资策略，我们最终得到如表 7-1 所示的具体的交易节点。

表 7-1 投资方案

操作	时间	金额
买入	2022-01-05 10:05:00	100000
	2022-01-07 13:40:00	300000
	2022-01-18 09:45:00	150000
	2022-01-21 10:25:00	500000
卖出	2022-01-06 14:50:00	63845
	2022-01-13 10:40:00	275841

	2022-01-20 10:00:00	114757

结合表 7-1 和图 7-6 的操作，最终实现盈利，截至到 2022-01-28 15:00:00，最终回撤为 16.6，夏普比率为 1.63。

综上，得到 19 天的总收益率为 1.55%；19 日每日的信息比率分别为 -0.230047337595639 、 0.00143898938900844 、 -0.000671673156484313 、 -0.00590259930884163 、 0.00151419413891592 、 -0.00778979460776757 、 0.000327002870159887 、 -0.00237632498247652 、 -0.00181574877054584 、 -0.00127481133298093 、 0.000316258497999941 、 0.00271349367472684 、 -0.000325256693384910 、 -0.00304401531244568 、 0.00220102649398108 、 -0.00157839999680372 、 0.000743436253828925 、 -0.00145870266727996 、 0.000514958944579981；最大回撤率为 2.31%。

八：模型评价

8.1 模型的优点

经过模型的误差分析和交叉验证，本文所构建的模型具有一定的创新性、适用性和正确性，主要体现为如下：

1.数据处理部分。剔除了股市在星期六和星期天的数据，并使用 Matlab 中 filliming 函数的 pchip 参数进行插值拟合对缺失值进行了补充，使原始数据更接近真实数据。

2. 充分考虑了成交量与各指标之间的关系, 收盘价与各指标之间的关系。引入变分模态分解，解决了时间序列数据非平稳的问题；引入长短期记忆神经网络，解决了时间序列数据非线性的问题。

3. 与 BP 神经网络相比，LSTM 神经网络可以很好地解决长时依赖问题。

4. 使用粒子群算法来确定修正表达式的权重，提高了长短期记忆神经网络对成交量和收盘价的预测精度，进而得到高准确性的“数字经济”板块预测值。

8.2 模型的缺点

对原始数据进行 VMD 分解后，未考虑残差分量，可能会对预测精度造成一定的影响。

8.3 模型的推广

本文预测模型对时间序列数据有良好的预测精度，能够移植到对风速的预测中去。类似的，也可以满足诸如振动趋势预测、电力负荷预测等工程实际需要，具有很强的普适性。

8.3 模型的改进

为提高本文预测模型应对金融时间序列数据造成的模态混叠效应，可对数据变分模态分解后得到的本征模态分量，分别采用不同的机器学习方法来对各模态分量进行预测，最后针对各模态分量，分别选取预测效果最好的机器学习方法，以提高预测精度。

参考文献

[1] 唐成.基于 Attention-LSTM 深度学习方法的量化投资研究[J].商讯,2021(36):155-157.

[2] 章睿. 基于机器学习的股票特征 K 线图预测研究[D]. 大连理工大学,2021.DOI:10.26991/d.cnki.gdllu.2021.002712.

[3] 杨帆,冯翔,阮羚,陈俊武,夏荣,陈昱龙,金志辉.基于皮尔逊相关系数法的水树枝与超低频介损的相关性研究[J].高压电器, 2014,50(06): 21-25+31.DOI:10.13296/j.1001-1609. hva.2014.06.004.

[4] 纪德洋,金锋,冬雷,张姗,于坤洋.基于皮尔逊相关系数的光伏电站数据修复[J].中国电机工程学报,2022,42(04):1514-1523.DOI:10.13334/j.0258-8013.pcsee.211172.

[5] Kai Wang,Wenlong Fu,Tie Chen,Binqiao Zhang,Dongzhen Xiong,Ping Fang. A compound framework for wind speed forecasting based on comprehensive feature selection, quantile regression incorporated into convolutional simplified long short-term memory network and residual error correction[J]. Energy Conversion and Management,2020,222:

[6] Fu Wenlong,Wang Kai,Tan Jiawen,Zhang Kai. A composite framework coupling multiple feature selection, compound prediction models and novel hybrid swarm optimizer-based synchronization optimization strategy for multi-step ahead short-term wind speed forecasting[J]. Energy Conversion and Management,2020,205(C):

[7] Wenlong Fu,Kai Wang,Chu Zhang,Jiawen Tan. A hybrid approach for measuring the vibrational trend of hydroelectric unit with enhanced multi-scale chaotic series analysis and optimized least squares support vector machine[J]. Transactions of the Institute of Measurement and Control,2019,41(15):

[8] 李钦. 基于 LSTM 的技术指标量化投资策略设计[D]. 广东外语外贸大学,2021.DOI:10.27032/d.cnki.ggdwu.2021.000799

[9] 赵梦娜. 基于 SVM 和 BP 神经网络的量化策略研究[D]. 大连理工大学,2021.DOI:10.26991/d.cnki.gdllu.2021.001540

[10] 吕柏行,郭志光,赵韦皓,张凡.标准粒子群算法的优化方式综述[J].科学技术创新,2021(28):33-37.

[11] 陈杰. 中国股指期货市场基于相对强弱指标的日内趋势策略[D]. 上海交通大学,2017.DOI:10.27307/d.cnki.gsjtu.2017.001506.

[12] 徐崤笑. 一种基于大数据分析与移动均线寻找股票买入点的方法[D]. 沈阳师范大学,2020.DOI:10.27328/d.cnki.gshsc.2020.000398.

[13] 吴灿.股市中布林线指标的分析方法与应用研究[J].科技广场,2013(05):241-245.

[14] 赵霞,时雨,王佳琪.网络视角下基于夏普比率的投资组合策略[J]. 山东财经大学学报,2022,34(02):17-26.

附录

q1_1

简介:数据读取

```
clear,clc
%读取数据并储存
a=xlsread('附表.xlsx','国内市场指标');
[~,a0]=xlsread('附表.xlsx','国内市场指标','A4:A266');

b=xlsread('附表.xlsx','数字经济版块信息');

c=xlsread('附表.xlsx','技术指标');
[~,c0]=xlsread('附表.xlsx','技术指标','B2:B264');

d=xlsread('附表.xlsx','国际市场指标');
[~,d0]=xlsread('附表.xlsx','国际市场指标','A4:A290');

e=xlsread('附表.xlsx','汇率');
[~,e0]=xlsread('附表.xlsx','汇率','A4:A273');

f=xlsread('附表.xlsx','其他板块信息');
[~,f0]=xlsread('附表.xlsx','其他板块信息','A2:A264');

%指标日期对齐
zb0=e0;zb=e;
zb1=zeros(287,size(zb,2))*NaN;
for i=1:length(zb0)
    for j=i:287
        if strcmp(zb0{i},d0{j})
            zb1(j,:)=zb(i,:);
        end
    end
end
e=zb1;
```

Q1_2

简介：数据清洗

```
clear,clc
load('zb.mat')
%补全空值
for i=1:length(zb)
    a=zb{i};
    %使用保形样条插值
    for j=1:size(a,2)
        x = fillmissing(a(:,j),'pchip');
        a(:,j)=x;
    end
    %保存清洗后指标为 zbx.mat
    zbx{i}=a;
end
```

Q1_3

简介：相关性分析模型（用于成交量与收盘价）

```
clear,clc
load('zbx.mat')
%相关性分析
A=[];
for i=[1 3:length(zbx)]
    aa=zbx{i}(2:131,:);
    A=[A aa];
end
aa=zbx{2}(:,2);
[~,b0]=xlsread('附表.xlsx','数字经济版块信息','B2:B6241');
a='';bb=[];
for i=1:6240
    if ~strcmp(b0{i}(1:13),a)
        a=b0{i}(1:13);
        bb=[bb;aa(i)];
    end
end
aa=zbx{2}(:,5);
[~,b0]=xlsread('附表.xlsx','数字经济版块信息','B2:B6241');
a='';cc=[];
```

```
for i=1:6240
    if ~strcmp(b0{i}(1:13),a)
        a=b0{i}(1:13);
        cc=[cc;aa(i)];
    end
end
A1=[bb A];
A2=[cc A];

%收盘价 Pearson 相关性
B=corr(A1,'type','Pearson');

%成交量 Pearson 相关性
C=corr(A2,'type','Pearson');
```

Q1_4

简介:线性相关拟合(用于成交量与收盘价)

```
clear,clc
%相关关系
load('zbx.mat')
A=[];
for i=[1 3:length(zbx)]
    aa=zbx{i}(2:131,:);
    A=[A aa];
end
aa=zbx{2}(:,5);     %成交量  OR  收盘价
[~,b0]=xlsread('附表.xlsx','数字经济版块信息','B2:B6241');
a='';bb=[];
for i=1:6240
    if ~strcmp(b0{i}(1:13),a)
        a=b0{i}(1:13);
        bb=[bb;aa(i)];
    end
end
a=bb;
% A=A(1:20,:);a=a(1:20,:);
%高相关性指标
A=A(:,[14 34 42 43]);
b=regress(a,A);
```

```
plot(b'*A')
hold on
plot(a)
```

q2_1

简介:变分模态分解训练集

```
clear,clc
load('zbx.mat')
s=zbx{2}(:,2);
s=fliplr(s')';
k=4;%VMD 分解数
[u, u_hat, omega] = VMD(s, 2500, 0.3, k, 0, 1, 1e-7);
s1=sum(u);

subplot(k+1,1,1);
plot(1:size(s1,2),s1);
xlim([0 size(s1,2)+1])
ylabel(['收盘价']);
for n1=1:size(u,1)
    subplot(k+1,1,n1+1);
    plot(1:size(s1,2),u(n1,:));%输出 IMF 分量，
    %a(:,n)则表示矩阵 a 的第 n 列元素，
    %u(n1,:)表示矩阵 u 的 n1 行元素
    xlim([0 size(s1,2)+1])
    ylabel(['IMF' int2str(n1)]);
end
```

q2_2

简介:

```
clear,clc
load('u2.mat')
data =u(1,:) ; %对第 i 分量预测 u(i,:)
```

```
numTimeStepsTrain = 5328;

dataTrain = data(1:numTimeStepsTrain);
dataTest = data(numTimeStepsTrain+1:end);
%标准化数据
mu = mean(dataTrain);
sig = std(dataTrain);

dataTrainStandardized = (dataTrain - mu) / sig;
%准备预测变量和响应
XTrain=[];YTrain=[];
n=4;
for i=1:n
    XTrain = [XTrain dataTrainStandardized(i:end-n+i-1)];
    YTrain = [YTrain dataTrainStandardized(i+1:end-n+i)];
end
%创建 LSTM 回归网络。指定 LSTM 层有 200 个隐含单元。
numFeatures = 1;
numResponses = 1;
numHiddenUnits = 200;                    %超参数 2：神经元个数

layers = [ ...
    sequenceInputLayer(numFeatures)
    lstmLayer(numHiddenUnits)
    fullyConnectedLayer(numResponses)
    regressionLayer];

options = trainingOptions('adam', ...
    'MaxEpochs',170, ...                 %超参数 1：最大迭代次数
    'GradientThreshold',1, ...
    'InitialLearnRate',0.005, ...
    'LearnRateSchedule','piecewise', ...
    'LearnRateDropPeriod',125, ...
    'LearnRateDropFactor',0.2, ...
    'Verbose',0, ...
    'Plots','training-progress');
%训练 LSTM 网络
net = trainNetwork(XTrain,YTrain,layers,options);

dataTestStandardized = (dataTest - mu) / sig;
XTest = dataTestStandardized(1:end-1);
net = predictAndUpdateState(net,XTrain);
[net,YPred] = predictAndUpdateState(net,YTrain(end));
```

```
numTimeStepsTest = numel(XTest);
for i = 2:numTimeStepsTest
    [net,YPred(:,i)] = predictAndUpdateState(net,YPred(:,i-1),
    'ExecutionEnvironment','cpu');
end
%去标准化。
YPred = sig*YPred + mu;
YTest = dataTest(2:end);
rmse = sqrt(mean((YPred-YTest).^2))

% 使用预测值绘制训练时序。
figure
plot(data(1:5328))
xlim([0 6300])
hold on
idx = numTimeStepsTrain:(numTimeStepsTrain+numTimeStepsTest);
plot(idx,[data(numTimeStepsTrain) YPred],'.-')
hold off
xlabel("Month")
ylabel("Cases")

figure
subplot(2,1,1)
plot(YTest)
hold on
plot(YPred,'.-')
hold off
legend(["Observed" "Forecast"])
ylabel("Cases")
title("Forecast")
% 将预测值与测试数据进行比较。
subplot(2,1,2)
stem(YPred - YTest)
% xlabel("Month")
ylabel("Error")
title("RMSE = " + rmse)
title("Forecast")
legend(["Observed" "Forecast"])
```

q2_3

简介:初步预测值与实际值对比

```
clear,clc
%预测实际对比
load('zbx.mat')
y=zbx{2}(:,2);
y=fliplr(y')';
load('ypre.mat');%收盘价
ypre=sum(ypre);
plot(y(5329:end))
hold on
plot(ypre)
legend('真实值','预测值')
% xlim([0 1000])
plt=Plot();
plt.YMinorTick='off';
plt.XTick=[0 912];
plt.XTickLabel={'2022-1-4 9:35','2022-1-28 15:00'};
plt.YLabel='收盘价';
plt.FontName='宋体';
plt.LineWidth=[1.5 1.5];
```

q2_4

简介:根据相关指标修正预测值

```
clear,clc
load('zbx.mat')
load('b_sp.mat')
load('A_sp.mat')
bb=b;
A=fliplr(A(1:19,:)')';
%将指标扩充至每 5 分钟数据
AA=A;A=[];
for i=1:size(AA,1)
    A=[A;ones(48,1)*AA(i,:)];
end
A=A(2:end,:);

y=zbx{2}(:,2);     %成交量  OR  收盘价
```

```
y=fliplr(y')';
ypre=sum(ypre)';


y=y(5330:end);

%粒子群算法优化修正权重
c1=0.4;c2=0.6; %学习因子
wmax=0.6;wmin=0.4; %惯性权重
num=100; %粒子总数
n=2; %自变量个数
sub=[0 0]; %自变量下限  n
up=[1 1]; %自变量上限  n

for i=1:num
    for j=1:n
        x(i,j)=(up(j)-sub(j))*rand+sub(j); %初始化位置
    end
    if i==num
        x(i,:)=[1 0];
    end
    yypre=ypre*x(i,1)+A*bb*x(i,2);
    rmse = sqrt(mean((yypre-y).^2));
    fx(i,1) = rmse;
end
v=randn(num,n); %初始化速度

[bestf,a]=min(fx);%记录历史最优值
bestx=x(a,:);%记录历史最优解
trace(1)=bestf;
xx=[];ff=[];
for ii=1:100
    fave=mean(fx);
    fmin=min(fx);
    for i=1:num
        if fx(i)<=fave
            w=wmin+(fx(i)-fmin)*(wmax-wmin)/(fave-fmin);
        else
            w=wmax;
        end
        [~,b]=min(fx);
        best=x(b,:);%当前最优解
        v(i,:)=w*v(i,:)+c1*rand*(best-x(i,:))+c2*rand*(bestx-x(i,:));
        xx(i,:)=x(i,:)+v(i,:);
```

```
            yypre=ypre*x(i,1)+A*bb*x(i,2);
            rmse = sqrt(mean((yypre-y).^2));
            ff(i,1)=rmse;
            if ff(i,1)<fx(i,1)
                fx(i,1)=ff(i,1);
                x(i,:)=xx(i,:);
            end
        end
        if min(fx)<bestf %比较历史最优值
            [bestf,a]=min(fx);
            bestx=x(a,:);%更新历史最优解
        end
        trace=[trace;bestf];
end

%根据相关指标修正预测值
k1=bestx(1);k2=bestx(2);
yypre=ypre*k1+A*bb*k2;

plot(y,'b')
hold on
plot(ypre,'r')
plot(yypre,'g')
legend('真实值','预测值','修正预测值')
xlim([-50 970])
plt=Plot();
plt.YMinorTick='off';
plt.XTick=[0 912];
plt.XTickLabel={'2022-1-4 9:35','2022-1-28 15:00'};
plt.YLabel='成交量';
plt.FontName='宋体';
plt.LineWidth=[1.5 1.5 1.5];

%误差
a=mean(y);b=std(y);
y=(y-a)./b;yypre=(yypre-a)./b;
mae = mean(abs(yypre-y))
rmse = sqrt(mean((yypre-y).^2))
mape = mean(abs((yypre-y)./y))
```

q4_0

简介:计算总收益率、信息比率、最大回撤率

```
clear,clc
load('li.mat')%导入利润数据
load('zbx.mat')
li=li+18;
y=fliplr(zbx{1}(1:19,7));%中证 500
sy=[];
for t=1:19
    sy=[sy;sum(li((t-1)*48+1:t*48))];
end
sy=[sy(1);sy];
syl=[];z=1000000;
zz=[];
for i=1:19
    z=z+sy(i);
    zz=[zz;z];
end
for i=2:19
    syl=[syl (zz(i)-zz(i-1))/(zz(i-1))];
end
syl=[0 syl];
zhong=[];
for i=2:19
    zhong=[zhong (y(i)-y(i-1))/y(i-1)];
end
zhong=[1 zhong];

%总收益率
zsyl=prod(syl+1)-1;
%信息比率
c=syl-0.9*zhong;
xxbl=(c./19)/std(c);
%最大回撤率
drawdown=-10;
for i=1:18
    for j=i+1:19
        if ( zz(i)-zz(j) )/zz(i) > drawdown
            drawdown=( zz(i)-zz(j) )/zz(i);
        end
    end
end
```

plt_pearson

简介:皮尔逊相关性矩阵图

```
clc
clear all
%皮尔逊相关性
data=xlsread('SampleA1.xlsx',1,'A2:AS46')

%相关性分析
%默认类型为 Pearson 系数
[xiangguan,p_value]=corr(data)%等效于 xiangguan=corr(data,'Type','Pearson');

%x 轴和 y 轴的标签，要和数据的列数对应
index_name={'dv1','dv2','dv3','dv4','dv5','dv6','dv7','dv8','dv9','dv10','dv11','dv12','dv13','dv14',
'dv15','dv16','dv17','dv18','dv19','dv20','dv21','dv22','dv23','dv24','dv25','dv26','dv27','dv28','dv
29','dv30','dv31','dv32','dv33','dv34','dv35','dv36','dv37','dv38','dv39','dv40','dv41','dv42','dv43'
,'dv44','dv45'};
y_index = index_name;
x_index=index_name;
figure
%字号 12，字体宋体，可以随意改变 显示默认配色
H = heatmap(x_index,y_index, p_value, 'FontSize',12, 'FontName','宋体');
colormap(gca, 'parula')
H.Title = '皮尔逊显著性检验矩阵：收盘价';
figure
% 可以自己定义颜色块
H = heatmap(x_index,y_index, xiangguan, 'FontSize',12, 'FontName','宋体');
H.Title = '皮尔逊相关性系数矩阵：收盘价';
colormap(autumn(5))%设置颜色个数
```

q4_1

简介:计算指标 1

```
%% 常用指标的计算
clc, clear, close all
%load('matlab.mat')
data = xlsread('收盘价预测值正序.xlsx',1,'A2:C912')
```

```
CP = CP; % 提取收盘价序列
Dates = Time; % 提取日期
XH = XH;
%% 移动均线
% 简单均线
movavg(CP,5,30,0);
legend('Close','Short','Long','Location','Best')

% 对数均线
[Short1,Long1] = movavg(CP,5,30,'e');
figure, plot(Dates, [CP,Short1,Long1]), title('预测的收盘价对数移动均线');grid on
legend('PriceClose','PriceShort','PriceLong','Location','Best')
datetick('x', 'mmmm-yyy')

%%  相对强弱指标
figure
rsi = rsindex(CP,14);
plot(XH, rsi)
title('预测的收盘价相对强弱指标')
datetick('x', 'mmmm-yyy')

%% 最大回撤
[MaxDD, MaxDDIndex] = maxdrawdown(CP, 'return');
figure
plot(Dates,CP);
title('预测的收盘价最大回撤')
datetick('x', 'mmmm-yyy')
hold on
plot(Dates(MaxDDIndex(1):MaxDDIndex(end)),CP(MaxDDIndex(1):MaxDDIndex(end)),'r')
```

q4_2

简介:计算常用指标 2

```
%% 常用指标的计算
clc, clear, close all
load('matlab.mat')
CP = CP; % 提取收盘价序列
Dates = Time; % 提取日期
%% 布林曲线
[Movavgv, UpperBand, LowerBand] = bolling(CP, 20);
```

```
figure(1), plot(Dates(20:911,:), [CP(20:911,:),Movavgv, UpperBand, LowerBand]), grid on
legend('ClosedPrice','average','UpperBand','LowerBand','Location','Best')
datetick('x', 'mmmm-yyy')
title('预测的收盘价布林曲线')

%% 布林线策略
N =    size(CP(20:911,:));
s = ones(N);
for i = 2:N
      if s(i)==1 && CP(19+i-1)<UpperBand(i-1) && CP(19+i)> UpperBand(i)
            s(i)=1;
            s(i+1:end)=0;
      elseif s(i)==0 && CP(19+i-1)>LowerBand(i-1) && CP(19+i)<LowerBand(i)
            s(i)=-1;
             s(i+1:end)=1;
      end
end
r = [0; s(2:end).*diff(CP(20:911,:))];
figure(2)
plot(Dates(20:911,:), cumsum(r));
datetick('x', 'mmmm-yyy');
title('收盘价布林线策略')
%布林线
```

q4_3

简介:绘制收益图

```
clc,clear,close all
%% Load in some CSI931582_EOD
prepareData
CSI931582_EOD = CP;
Dates = Time; % 提取日期
testPts = floor (1*length(CSI931582_EOD));
CSIClose = CSI931582_EOD(1:testPts);
CSICloseV = CSI931582_EOD(testPts+1:end);
%%Develop a simple lead/lag technical indicator
%we'll use two exponentially weighted moving averages
[lead,lag]=movavg(CSIClose,20,30,'e');
plot(Dates(1:911,:),[CSIClose,lead,lag]),grid on
legend( 'close', 'Lead', 'Lag', 'Location', 'Best')
```

```
% datetick('x', 'mmmm-yyy')
%Develop a trading signal and performance meastres. We'll assume 250
% trading days per year.
s = zeros(size(CSIClose));
s(lead>lag)= 1;
%Buy(long)
s(lead<lag) = -1;
% Sell (short)
r = [0; s(1:end-1).*diff(CSIClose)];%Return
sh = sqrt(911)*sharpe(r,0);
%Annual Sharpe Ratio
%%
%Plot results
ax(1) = subplot(2,1,1);
plot([CSIClose,lead,lag]); grid on
% datetick('x', 'mmmm-yyy')
% legend( 'Close', ' Lead','Lag','Location','Best')
title(['First Pass Results，Annual Sharpe Ratio = ',num2str(sh,3)])
ax(2) = subplot(2,1,2);
plot([s,(cumsum(r)-18)]); grid on
title(['Final Return = ',num2str(sum(r),3),'(',num2str(sum(r)/CSIClose(1)*100,3),'%)'])
% legend('Position','Cumulative Return','Location','Best')
% datetick('x', 'mmmm-yyy')
linkaxes(ax,'x')
annualScaling=sqrt(911);
%‰% Estimate parameters over a range of values
% Return to the two moving average case and identify the best one.
sh = nan(100,100);
tic
for n = 1:100
for m= n:100
[~,~, sh(n,m)] = leadlag(CSIClose,n,m,annualScaling);
end
end
toc
%%
% Plot results
figure
surfc(sh),shading interp，lighting phong
view([80 35]),light('pos',[0.5,-0.9,0.05])
colorbar
%%
%Plot best Sharpe Ratio
[maxSH, row] = max(sh);%max by column
```

```
[maxSH,col] = max(maxSH); %max by row and colunn
leadlag(CSIclose,row(col), col, annualScaling)
%%Evaluate performance on validation CS1300
leadlag(CsIcloseV, row(col),col,annualScaling)
```