基于 EMD-SE-LSTM 模型的股指日内已实现波动率预测

——以中证 500 指数为例

刘 传，陈彦晖

（上海海事大学 经济管理学院，上海 201306）

摘要：由于股指波动率具有非平稳、高嘈杂、非线性等特征，而传统的预测模型在建模时要求数据平稳、线性或近似线性，所以很难精准预测股指波动率。为提高股指波动率的预测效果，采用经验模态分解（EMD）、样本熵（SE）和长短期记忆网络（LSTM）构建的模型对股指日内已实现波动率进行预测。以中证 500 指数为例，经过 EMD 分解得到一系列分量，再根据分量的样本熵大小进行重构，最后利用 LSTM 对重构后的各序列进行预测。结果表明，EMD 算法对 LSTM 模型的预测精度有很大的提升，相较于传统模型，EMD-SE-LSTM 模型在预测股指波动率时精度更高，拟合优度更好。

关键词：经验模态分解（EMD）；长短期记忆网络（LSTM）；已实现波动率；股票指数

中图分类号：F830.91 **文献标志码**：A **文章编号**：1671－1807(2022)08－0385－07

随着中国金融市场的蓬勃发展，作为金融市场中重要组成部分的股票市场逐渐成为企业融资者筹集资金的重要渠道，同时也给投资者进行资金管理和实现投资收益提供了重要途径。在股票市场中，股票价格指数作为整个股票市场的股票总指数，反映了整体股票价格水平以及整体走势。股指波动率像是一个方向标，在波动的股市当中起着重要的作用。在即将遇到风险时，投资者利用股指期货将其对整个股票市场价格指数的预期风险转移至期货市场，以此来规避风险。股指期货也是对股票未来价格预期，深受股票指数的影响。股指期货的基础标的是股票指数。股指的波动率情况则是对股指未来走势的影响因素。而股票价格波动变化极其复杂，并没有明确的规律。因此，准确地预测股指的波动率及走势不仅可以有效地实现高额的投资回报还可以有效地规避投资风险。

股票指数的波动性是极其复杂没有明确规律的，想要从复杂的股指波动中洞悉股票指数的走势和波动情况，从而实现高额的投资回报，一直以来是人们关注的焦点问题。而股票指数数据具有非线性、不平稳、数据量很大、非常复杂等特点，增加了预测难度。而传统的预测金融数据模型则要求数据必须是平稳的、线性或近似线的，在预测股指走势和波动方面，其准确性和精度并不高。众所周知，经验模态分解是一种自适应性强的时间序列数据分解算法，能够对非线性、非平稳的时间序列数据进行分解，非常适合对像股指波动率这样的金融高频数据进行分解。面对庞大的金融时间序列数据，深度学习算法脱颖而出，它可以从大量复杂的数据中提取特征，无须依赖先验知识，非常适合预测高频金融时间序列的波动率，在所有深度学习算法中，长短期记忆神经网络因其循环结构和链状结构，具有长记忆性，可作为复杂的非线性单元构造更大型的神经网络。因此，长短期记忆神经网络更适合预测金额高频时间序列数据。

国内外对波动率的研究可以追溯到 1982 年，Engle 提出并采用自回归条件异方差（ARCH）模型对金融资产收益率方差进行统计并有效地拟合了收益率的波动性，研究发现金融资产波动率具有高度的相关性[1]。随后其他传统模型也被纷纷应用于

收稿日期：2022-03-28

基金项目：国家自然科学基金青年项目（71701127）。

作者简介：刘传（1996—），男，安徽铜陵人，上海海事大学经济管理学院，硕士研究生，研究方向为航运金融；陈彦晖（1984—），女，山西临县人，上海海事大学经济管理学院，副教授、硕士研究生导师，研究方向为航运金融、时间序列分析。

股指波动率预测，如 ARMA 模型、ARIMA 模型、GARCH 模型以及由 GARCH 模型改进的 TGARCH、EGARCH、IGARCH 等诸多 GARCH 族模型[2-5]。而后随着计算机技术的飞速发展，机器学习、深度学习等算法逐渐应用于高频金融时间序列数据分析当中。Ghosh 等采用随机森林和 LSTM 网络作为训练方法对标普 500 指数成分股进行预测，结果表明，使用 LSTM 网络的多特征设置提供了 0.64%的日回报要高于随机森林 0.54%的日回报[6]。杨青和王晨慰在研究全球股票指数预测中，实证表明 LSTM 神经网络具有很强的泛化能力，预测效果非常稳定，与其他模型对比，LSTM 模型预测精度很高且能够有效控制误差[7]。Zhang 等使用长短期记忆网络模型来预测股价走势，通过采用投资者注意力的代理变量作为市场变量的补充，实证结果表明，LSTM 模型相比其他的人工神经网络(ANNs)在处理非线性、非平稳和复杂的金融时间序列更合适，且其预测精度更高[8]。而经验模态分解将时间序列数据根据自身的时间尺度特征分解成不同周期、不同频率的本征模函数和残差项，无须提前设定任何基函数，也不要求数据是线性、平稳的。刘海飞和李心丹使用 EMD 分解算法对股票价格进行预测和小波分析预测方法做比较，实证研究表明，使用经验模态分解方法预测结果精度更高、拟合度更优、预测功能更强、模型更加稳定[9]。Luo 等通过构建 EMD-Copula-CoVaR 模型来衡量国际股票市场多尺度的金融风险传染力，实证结果表明，EMD-Copula-CoVaR 模型在衡量金融风险传染在所有时间尺度上都是有效的，金融风险传染主要贡献者是高频成分。同时还实证了除英国外，在原始和中频分量下，美国金融市场对其他金融市场输出的风险要高于接受的风险[10]。Wei 等为了能够准确地预测海浪情况提出了 EMD-LSTM 模型，通过分析不同预报时间的预报效果，实证表明，EMD 分解算法可以有效降低 LSTM 的误差且预报时间在相同的容忍度下可以提前一倍以上[11]。刘铭和单玉莹在预测股指时发现，在预测沪深 300 指数收盘价和深证成指收盘价时，EMD 和 LSTM 组合模型有较好的预测效果[12]。

梳理前人的研究成果发现：长短期记忆网络在预测高频金融时间序列数据方面，相比传统模型准确性和精确度都更高，预测过程也比较简单；经验模态分解算法对数据自身的特征提取有着很好的表现。基于以上讨论，本文提出一种基于经验模态分解和长短期记忆神经网络的组合模型。

1 基本理论

1.1 经验模态分解

经验模态分解(empirical mode decomposition，EMD)是由 Huang 等于 1998 年提出的一种全新的自适应性强的时间序列数据分析算法[13]。EMD 算法有 3 个假设条件：①原序列至少含有一个极大值和一个极小值；②特征时间尺度由极大值和极小值之间时间差决定；③若原序列无极值点，但有拐点，可通过求导求其极值。EMD 分解对于任意时间序列 $y_{(t)}$ 计算流程如下：

步骤 1 找出原序列 $y_{(t)}$ 的所有的局部极大值和极小值，再用三次样条插值画出 $y_{(t)}$ 的上下包络线分别为 $m_{(t)}$ 和$n_{(t)}$，求其均值：

$$u_{(t)} = (m_{(t)} + n_{(t)})/2 \tag{1}$$

步骤 2 从 $y_{(t)}$ 中减去均值包络线 $u_{(t)}$，得到一个新序列 $d_{(t)}$，即

$$d_{(t)} = y_{(t)} - u_{(t)} \tag{2}$$

步骤 3 判断新序列 $d_{(t)}$ 是否满足 IMF 的两个条件：①在整个时间尺度内，$d_{(t)}$ 所有的局部极值点的个数和零点个数要么相等，要么最多相差一个；②在整个时间尺度范围任何时间点上，其上、下包络线均值恒为 0。若满足，则 $d_{(t)}$ 是原始时间序列的一阶本征模函数，即 $d_{(t)} = \mathrm{IMF}_1$，若不满足，将 $d_{(t)}$ 看作是原始时间序列，重复步骤上述步骤，直到 $d_{(t)}$ 满足 IMF 的两个条件为止。

步骤 4 从原始时间序列 $y_{(t)}$ 中剔除 IMF_1，得到新的序列，重复以上步骤，得到 IMF_2、IMF_3… 和一个残差项 $r_{(t)}$。则原始序列 $y_{(t)}$ 可表示为

$$y_{(t)} = \sum_{i}^{n} \mathrm{IMF}_i + r_{(t)} \tag{3}$$

1.2 样本熵

样本熵(sample entropy，SE)是 Richman 和 Moornan 在近似熵原理的基础上提出的一种改进的衡量时间序列数据自身波动复杂程度的度量方法[14]。时间序列自身前后的波动的重复性和周期性，即该时间序列数据前后自相似性的概率大小。若测得一个时间序列数据的样本熵值很大，那么意味着该序列中有很多的杂乱的信号，该时间序列数据本身在震荡前后的相似度就越低，就有很大概率产生新模式，因而序列本身就越复杂。对任意一个包含有 n 个数据的时间序列 $X = \{x_1, x_2, \cdots, x_n\}$ 样本熵的计算方法如下：

步骤 1 按序号构成 $(n-m+1)$ 组 m 维向量空

间时间序列,可表示为一个 $m(n-m+1)$ 的矩阵。

步骤 2　计算任意两组向量 $X_{m,i}$ 和$X_{m,j}$ 的距离 $d[X_{m,i},X_{m,j}]$。在任意两组向量一一对应的元素中,对应元素差值的绝对值最大的那一组对应元素的差值绝对值即为这两组向量的距离,即

$$d[X_{m,i},X_{m,j}]=\max|x_{i+k}-x_{j+k}| \tag{4}$$

式中:$k=0,1,2,\cdots,m-1;1\leqslant i,j\leqslant n-m+1$。

步骤 3　根据步骤 2 的计算任意两组向量的距离值,统计向量 $X_{m,i}$ 和$X_{m,j}$ 的距离 $d[X_{m,i},X_{m,j}]$ 小于或等于相似容限 r 的数目,记为 B_i。然后计算 B_i 与向量总数 $n-m+1$ 的比值,记为 $B_i^m(r)$。

$$B_i^m(r)=\frac{B_i}{n-m+1} \tag{5}$$

步骤 4　计算 $n-m+1$ 个$B_i^m(r)$ 的平均值,记为 $B^m(r)$,即

$$B^m(r)=\frac{1}{n-m+1}\sum_{i=1}^{n-m+1}[B_i^m(r)] \tag{6}$$

步骤 5　将原向量组的维数 m 提升到$m+1$,再对 $m+1$ 维向量组进行重复上面步骤,得到 $B^{m+1}(r)$。

步骤 6　该时间序列数据的样本熵值 SE 为

$$\mathrm{SE}=-\ln\frac{B^{m+1}(r)}{B^m(r)} \tag{7}$$

经大量文献研究发现,向量组位数 m 一般取值小于 10,相似容限 r 一般取原始时间序列得 0.1～0.25 得标准差。在本文实证研究当中,选取 6 作为向量维数 m 值,相似容限 r 取 0.15 $\sqrt{\mathrm{Var}(x_i)}$。

1.3　长短期记忆网络

长短期记忆网络(long short-term memory,LSTM)算法是一种改进的递归循环神经网络模型(RNN)。LSTM 改进之处在于在原来的 RNN 结构中增加了“输入门”“遗忘门”“输出门”和隐藏单元控制门,能够及时有效地增加某些重要信息、剔除无关的信息以及处理时间或事件的影响。改进之后,LSTM 模型有效地缓解了传统循环神经网络的梯度消失和梯度爆炸等问题。LSTM 算法结构如图 1 所示。

步骤 1　输入时间序列数据 X_t 和隐藏层单元状态 h_{t-1} 经过“遗忘门”,得到此门细胞状态 f_t 值,其计算公式为

$$f_t=\sigma[\boldsymbol{W}_{\mathrm{f}}(h_{t-1},X_t)+b_{\mathrm{f}}] \tag{8}$$

步骤 2　当输入时间序列数据 X_t 和隐藏层单元状态 h_{t-1} 经过“输入门”时,得到“输入门”细胞状态 i_t 和待输入的细胞状态 $\tilde{C}_t$,其表达式为

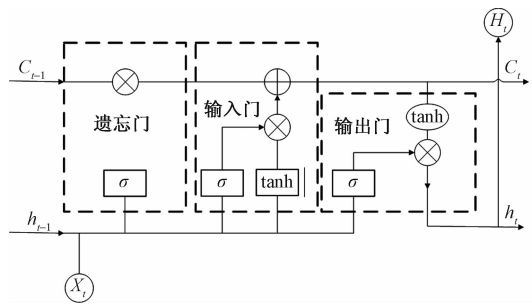

图 1　LSTM 算法结构

$$i_t=\sigma[\boldsymbol{W}_{\mathrm{i}}(h_{t-1},X_t)+b_{\mathrm{i}}] \tag{9}$$

$$\tilde{C}_t=\tanh[\boldsymbol{W}_{\mathrm{i}}(h_{t-1},X_t)+b_{\mathrm{c}}] \tag{10}$$

步骤 3　输入时间序列数据 X_t 和隐藏单元层状态 h_{t-1} 进入“输出门”,得到待输出结果 o_t 和 t 时刻细胞状态 C_t,待输出结果 o_t 再经过细胞状态 C_t 的筛选得到最终的输出结果 h_t。

$$o_t=\sigma[\boldsymbol{W}_{\mathrm{o}}(h_{t-1},X_t)+b_{\mathrm{o}}] \tag{11}$$

$$C_t=f_tC_{t-1}+i_t\tilde{C}_t \tag{12}$$

$$h_t=o_t\tanh C_t \tag{13}$$

式中:σ 为 sigmod 函数;$\boldsymbol{W}_{\mathrm{f}}$ 和 b_{f} 为“遗忘门”的权值矩阵和偏置系数;$\boldsymbol{W}_{\mathrm{i}}$ 和 b_{i} 分别为“输入门”的权值矩阵和偏置系数;$\boldsymbol{W}_{\mathrm{c}}$ 和 b_{c} 分别为细胞状态更新后的权值矩阵和偏置系数;$\boldsymbol{W}_{\mathrm{o}}$ 和 b_{o} 分别为“输出门”的权值矩阵和偏置系数;C_{t-1} 表示 $t-1$ 时刻的细胞状态。tanh 为双曲正切激活函数,取值范围为 $[-1,1]$。

1.4　EMD-SE-LSTM 预测模型

结合各种算法的优势,构建 EMD-SE-LSTM 组合预测模型,从而更加精准预测股指波动率,其模型框架如图 2 所示。

由图 2 可知,首先通过 EMD 算法将日内已实现波动率数据进行分解,得到不同频率、不同周期的本征模函数(IMF)序列和残差序列(Res)。再将这些 IMF 根据样本熵的大小分别重构成高频、中频和低频序列。最后通过 LSTM 算法进行滑动预测。将分解后的 IMF 序列和残差序列作为模型的输入数据集,经大量数据的训练,设定好模型的参数,得到一系列预测值,合并成最终的预测结果。

2　实证分析

2.1　数据处理与模型设定

以中证 500 指数为例,选取数据时间跨度为 2019 年 1 月 2 日到 2021 年 5 月 24 日每一分钟收盘

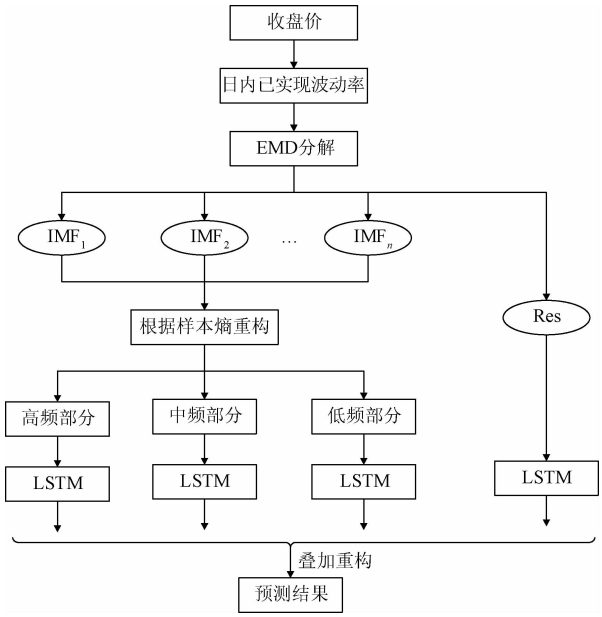

图 2　EMD-SE-LSTM 预测模型框架

价，共有 138 960 个有效数据。为了避免隔夜效应对已实现波动率的影响，剔除了每日开盘第一分钟的收盘价。计算每分钟的对数收益率（即 $r_t = \ln P_t - \ln P_{t-1}$，其中 P_t 为第 t 时刻收盘价，P_{t-1} 为第 $t-1$ 时刻收盘价），并采用 1 分钟和 5 分钟日内对数收益率平方和近似代替日内已实现波动率，最终形成 579 条有效样本数据。全部样本数据分为两部分，第一部分作为预测的训练集，取前 522 条数据；第二部分作为测试集，取后 57 条数据。全文以 1 分钟日内已实现波动率（图 3）样本为例进行详细介绍。

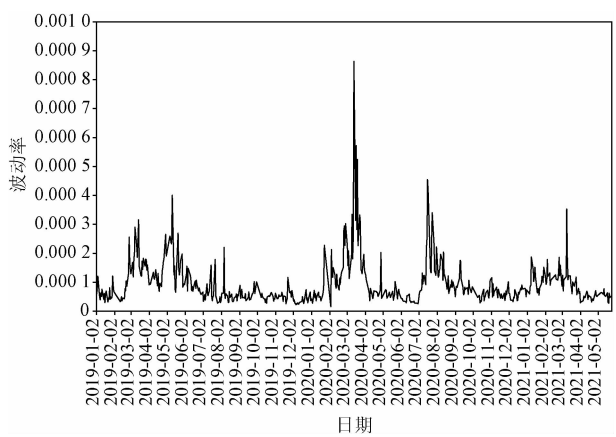

图 3　1 分钟日内已实现波动率

在 EMD-SE-LSTM 模型中，LSTM 模型结构选择单层 GPU，以均方根误差（RMAE）作为损失函数。LSTM 层含有 300 个隐含单元，在指定训练项，将求解器设置为 Adam 算法并进行 500 轮训练。使用动态学习算法，初始学习效率为 0.005，在进行 125 轮训练后，通过乘以因子 0.2 来逐渐衰减学习效率。设定 ARMA 模型时，先对原始时间序列进行了单位根检验，再进行数据平稳化处理。在构建 ARMA(p,q)模型参数设置时，经过多次调试，根据信息准则 AIC、SC 和 HQ 最小原理，在进行 1 分钟和 5 分钟日内已实现波动率建模时，分别选择了 ARMA(2,2)和 ARMA(1,3)。

2.2　EMD 分解

使用 MATLAB 软件实现 EMD 算法对数据正交分解，根据数据自身的时间尺度分解成不同频率的 7 个 IMF 序列和一个残差序列 Res。各序列统计指标见表 1，各序列走势如图 4 所示。

表 1　各序列统计指标

序列	平均周期	均值/10^{-7}	方差/10^{-10}	方差占比/%
IMF_1	3.74	−15.47	11.67	19.66
IMF_2	6.12	5.75	5.09	8.58
IMF_3	12.21	−8.91	7.03	11.84
IMF_4	23.11	−3.69	7.18	12.10
IMF_5	43.50	8.24	5.42	9.13
IMF_6	116.28	−57.21	20.13	33.91
IMF_7	272.78	81.22	18.83	31.91
Res	4 820.05	932.55	5.53	9.31

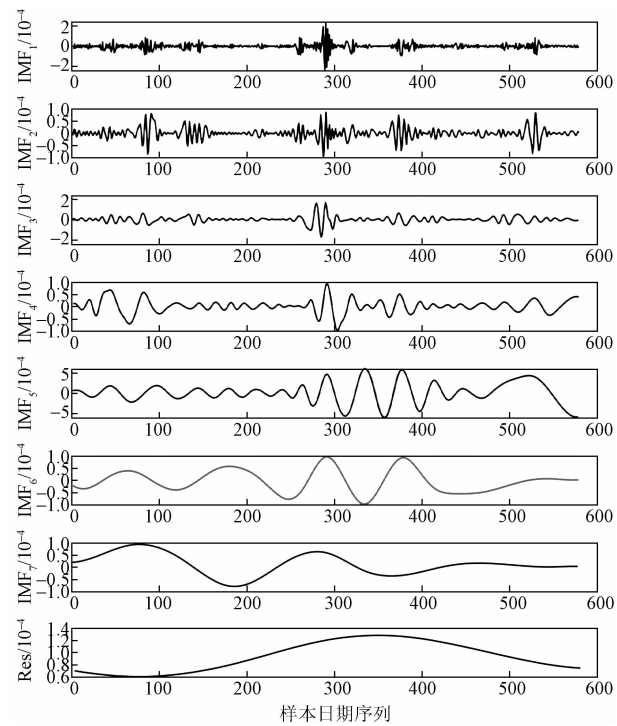

图 4　EMD 分解的 IMF 和残差序列 Res

2.3　基于 SE 的 IMF 重构

本文研究的 IMF 分量较多，如果对于每个 IMF

分量都分别进行 LSTM 算法预测,由于在建模过程中每个 IMF 分量都会产生相应的误差,IMF 分量越多所产生的误差就会越大,最后在合并预测结果的时候,所累积的误差就越大,最后在很大程度上影响了预测结果的精度。因此,本文提出了在将 IMF 分量进行 LSTM 建模前进行样本熵重构处理。计算出原始序列、IMF_1～IMF_7 以及残差项 Res 的样本熵分别为 0.218、2.309、1.432、0.731、0.517、0.328、0.062、0.048 和 0.027。其中 IMF_1～IMF_5 的样本熵都是大于原始序列的样本熵,IMF_6 和 IMF_7 都是小于原始样本熵。因此本文将 IMF_1 和 IMF_2 合并成高频序列,IMF_3～IMF_5 合并成中频序列,IMF_6 和 IMF_7 合并成低频序列。

2.4　模型评价指标

为了客观量化地评价各个模型的拟合水平,本文选取了均方根误差(RMSE)、平均绝对误差(MAE)、平均绝对百分比误差(MAPE)和纳什效率系数(R^2)4 个指标来评价模型的拟合优度。计算公式如下:

$$\mathrm{RMSE}=\sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i-y_i)^2}{n}} \tag{14}$$

$$\mathrm{MAE}=\frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i-y_i| \tag{15}$$

$$\mathrm{MAPE}=\frac{1}{n}\sum_{i=1}^{n}\left|\frac{\hat{y}_i-y_i}{y_i}\right|\times 100\% \tag{16}$$

$$R^2=1-\frac{\sum_{i=1}^{n}(\hat{y}_i-y_i)^2}{\sum_{i=1}^{n}(\bar{y}_i-y_i)^2} \tag{17}$$

式中:$\hat{y}_i$ 为预测值;y_i 为真实值;$\bar{y}_i$ 为样本均值;n 表示原序列的长度。RMSE、MAE 和 MAPE 越小,代表模型的预测效果越好,结果越准确。纳什效率系数 R^2 接近 1 时,拟合优度越佳,拟合的能力越强。

2.5　EMD-SE-LSTM 预测结果

2.5.1　1 分钟日内已实现波动率预测结果

通过比较 ARMA(2,2)、LSTM 和 EMD-SE-LSTM 模型的均方根误差、平均绝对误差、平均绝对百分比误差和纳什系数可知,EMD-SE-LSTM 的 4 项统计指标均优于其他模型,可以说明 EMD-SE-LSTM 模型的预测准确性、预测精度和模型的拟合优度均是最好的。各模型预测统计指标见表 2,结果走势如图 5 所示。

表 2　各模型 1 分钟日内已实现波动率预测统计指标值

模型	RMSE/10^{-5}	MAE/10^{-5}	MAPE/%	R^2
ARMA(2,2)	4.03	2.46	30.91	0.948 3
LSTM	9.06	6.40	66.87	0.921 6
EMD-SE-LSTM	3.59	1.21	27.23	0.983 4

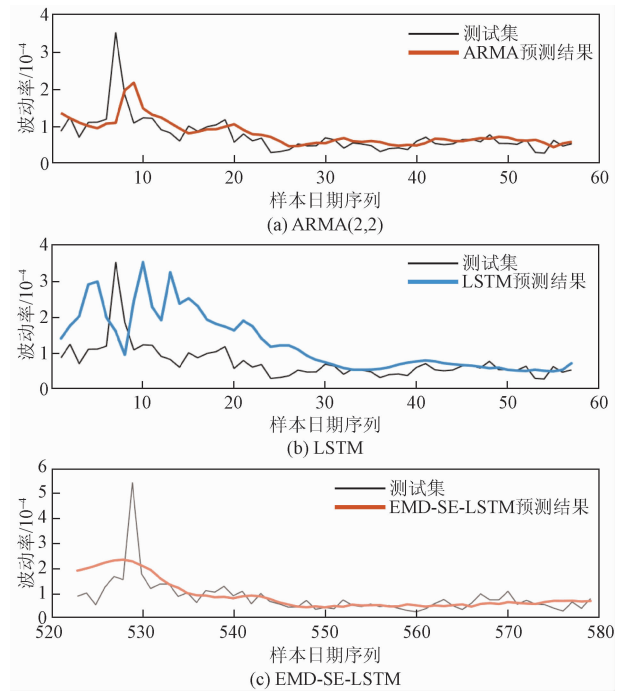

图 5　各模型 1 分钟日内已实现波动率预测结果

通过 EMD-SE-LSTM 模型与单独的 LSTM 模型预测结果统计指标对比,使用 EMD 分解算法后的 LSTM 模型,精确度评估指标 RMSE 从 9.06×10^{-5} 降低到 3.59×10^{-5},均方根误差减少了 5.47×10^{-5},MAE 从 6.40×10^{-5} 降低至 1.21×10^{-5},平均绝对误差减少了 4.19×10^{-5},MAPE 从 66.87% 降低至 27.23%,降低了 39.64 个百分点,而拟合优度从 0.921 6 提升至 0.983 4。这足以表明本文引用的 EMD 分解算法能够有效地提取股票指数波动率的特征,提高了 LSTM 模型的预测精度和拟合优度。

从预测结果图 5 来看,EMD-SE-STM 模型的预测效果明显比 ARMA 模型和 LSTM 模型好,ARMA 模型预测效果次之,LSTM 模型预测值与真实值有很明显的误差并且预测值的走势与真实值的延迟输出很类似,延迟的大小在两个工作日左右。通过对比 EMD-SE-LSTM 模型和单独的 LSTM 模型预测走势图,可以直观地知道 EMD-SE-LSTM组合模型的预测值和真实值拟合得更好,预测值更贴合真实值的走势,误差更小,延迟效果

也更小了。从而可知，EMD 分解算法提高 LSTM 预测模型的效果。在面对股指波动率出现异常值方面，EMD-SE-LSTM 模型很好地克服了波动率异常值的影响，使得预测结果更加平滑。

2.5.2　5 分钟日内已实现波动率预测结果

5 分钟日内已实现波动率实证过程与 1 分钟日内已实现波动率一致，故不再进行详细介绍，只给出最终结果，如表 3 和图 6 所示。

表 3　各模型 5 分钟日内已实现波动率预测统计指标值

模型	RMSE/10^{-5}	MAE/10^{-5}	MAPE/%	R^2
ARMA(1,3)	6.42	5.17	45.74	0.923 7
LSTM	11.78	8.92	104.71	0.872 3
EMD-SE-LSTM	5.78	4.91	37.82	0.963

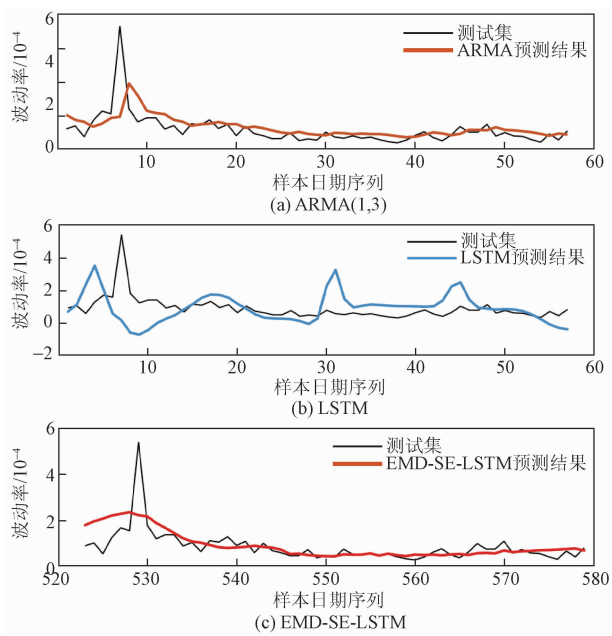

图 6　各模型 5 分钟日内已实现波动率预测结果

由最终结果对比分析，可以得出 EMD-SE-LSTM 模型在衡量模型精确度和拟合优度的 4 个指标评估下同样表现最佳。总体而言，针对不同频率的已实现波动率，不管是从统计指标来看，还是预测结果走势图对比分析来看，EMD-SE-LSTM 模型均能表现出最佳的预测效果。同样也可知，EMD 分解算法对于 LSTM 模型预测效果有很大的提升。

3　结论

在股票市场中，由于股票指数波动具有高度嘈杂、非线性、动态、非平稳等特点，预测股指波动率的变动显得格外棘手。面对传统的预测模型预测的结果并不那么理想，因此本文提出了 EMD-SE-LSTM 组合模型对股指波动率进行预测。实证结果表明：EMD-LSTM 组合预测模型在预测精确度和模型的拟合优度方面均超越其他模型，非常适合股票指数波动率的金融高频数据预测；此外，EMD 算法通过有效提取股指波动率的特征，提升了 LSTM 模型的预测效果，同时也体现了 EMD 算法对动态、非平稳数据处理的良好效果。本文提出的 EMD-SE-LSTM 组合预测模型为研究金额高频时间序列数据预测提供了新思路，为进一步预测国内外股指波动率奠定了基础。

参考文献

[1] ENGLE R E. Autoregressive conditional heteroskedasticity with estimates of the variance of United Kingdom inflation[J]. Econometrica，1982，50：987-1007.

[2] LING S，ZHU K，YEE C C. Diagnostic checking for non-stationary ARMA models with an application to financial data[J]. North American Journal of Economics & Finance，2013，26：624-639.

[3] PANIGRAHI S，PATTANAYAK R M，SETHY P K，et al. Forecasting of sunspot time series using a hybridization of ARIMA，ETS and SVM methods[J]. Solar Physics，2021，296：1-19.

[4] KIM C B. Leverage effect of HRCI volatility and the volatility impact on Korean export container volume before and after the global financial crisis：application of ARIMA-EGARCH and GIRF[J]. Asian Journal of Shipping and Logistics，2008，34：227-233.

[5] CURTO J D，PINTO J C，TAVARES G N. Modeling stock markets' volatility using GARCH models with Normal，Student' s and stable Paretian distributions[J]. Statistical Papers，2009，50：311-321.

[6] GHOSH P，NEUFELD A，SAHOO J K. Forecasting directional movements of stock prices for intraday trading using LSTM and random forests[J]. Finance Research Letters，2021，227：176-186.

[7] 杨青，王晨蔚. 基于深度学习 LSTM 神经网络的全球股票指数预测研究[J]. 统计研究，2019，36(3)：65-77.

[8] ZHANG Y J，CHU G，SHEN D H. The role of investor attention in predicting stock prices：the long short-term memory networks perspective[J]. Finance Research Letters，2021，38：1554-1584.

[9] 刘海飞，李心丹. 基于 EMD 方法的股票价格预测与实证研究[J]. 统计与决策，2010，26(23)：131-134.

[10] LUO C Q，LIU L，WANG D. Multiscale financial risk contagion between international stock markets：evidence from EMD-Copula-CoVaR analysis [J]. The North American Journal of Economics and Finance，2021，58：1062-1085.

[11] WEI H,SUN X F,WANG C Y,et al. A hybrid EMD-LSTM model for non-stationary wave prediction in offshore China[J]. Ocean Engineering,2022,246:527-556.

[12] 刘铭,单玉莹. 基于 EMD-LSTM 模型的股指收盘价预测[J]. 重庆理工大学学报(自然科学版),2021,35(12):269-276.

[13] HUANG N E,SHEN Z,LONG S R,et al. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis[J]. Proceedings of the Royal Society of London. Series A:Mathematical,Physical and Engineering Sciences,1998,454:903-995.

[14] RICHMAN J S,MOORMAN J R. Physiological time-series analysis using approximate entropy and sample entropy[J]. AmericanJournal of Physiology-Heart and Circulatory Physiology,2000,278(6):2039-2049.

Intraday Realized Volatility Forecasting for Stock Indices Based on EMD-SE-LSTM Model: The CSI 500 index as an example

LIU Chuan, CHEN Yanhui

(School of Economics and Management,Shanghai Maritime University,Shanghai 201306,China)

Abstract: It is difficult to predict stock index volatility accurately because of its non-stationary, highly noisy and non-linear, while traditional forecasting models require smooth, linear or approximately linear data in modeling. To improve the forecasting effect of stock index volatility, a model constructed by empirical modal decomposition (EMD), sample entropy (SE) and long short-term memory network (LSTM) is used to forecast the intra-day realized volatility of stock index. Taking the CSI 500 index as an example, a series of components are obtained after EMD decomposition, and are reconstructed according to the sample entropy magnitude of the components, and finally the LSTM is used to forecast each reconstructed series. The results show that the EMD algorithm also improves the prediction accuracy of the LSTM model, and the EMD-SE-LSTM model has higher accuracy and better fit superiority in predicting the stock index volatility compared with the traditional model.

Keywords: empirical mode decomposition(EMD); long short-term memory(LSTM); realized volatility; stock index