

基于VMD-PCA-BiLSTM的股票价格预测

王亚琪, 孙德山

辽宁师范大学数学学院, 辽宁 大连

收稿日期: 2026年6月17日; 录用日期: 2026年7月11日; 发布日期: 2026年7月21日

摘要

针对股票时间序列非平稳、高噪声导致预测效果受限的问题, 本文提出VMD-PCA-BiLSTM股票价格预测模型。结合平均样本熵构建自适应变分模态分解, 对收盘价序列做分解处理, 实现频域降噪与多尺度特征解耦; 同时采用主成分分析, 对量价数据及技术指标进行降维, 消除变量多重共线性, 重构低维时域特征空间。模型搭建双分支双向长短期记忆网络, 分别提取频域、时域特征并深度融合以完成预测。实验表明, 该模型有效改善传统方法的平滑效应与时滞缺陷, 均方根误差降至28.55, 决定系数达97.95%, 预测精度与鲁棒性提升显著。

关键词

股价预测, 变分模态分解, 主成分分析, 双向长短期记忆网络

Stock Price Prediction Based on VMD-PCA-BiLSTM

Yaqi Wang, Deshan Sun

School of Mathematics, Liaoning Normal University, Dalian Liaoning

Received: June 17, 2026; accepted: July 11, 2026; published: July 21, 2026

Abstract

To address the limitations in forecasting performance caused by the non-stationary nature and high noise levels of stock time series, this paper proposes the VMD-PCA-BiLSTM stock price forecasting model. By combining average sample entropy with adaptive variational mode decomposition, the model decomposes closing price sequences to achieve frequency-domain denoising and multi-scale feature decoupling. Simultaneously, principal component analysis is employed to reduce the dimensionality of volume-price data and technical indicators, eliminating multicollinearity among variables and reconstructing a low-dimensional time-domain feature space. The model employs a dual-

branch bidirectional long short-term memory (BiLSTM) network to extract frequency-domain and time-domain features, respectively, and deeply integrate them to perform forecasting. Experimental results demonstrate that this model effectively addresses the smoothing effects and time lag issues inherent in traditional methods, reducing the root mean square error (RMSE) to 28.55 and achieving a coefficient of determination of 97.95%, thereby significantly improving forecasting accuracy and robustness.

Keywords

Stock Price Forecasting, Variational Modal Decomposition, Principal Component Analysis, Bidirectional Long Short-Term Memory Network

Copyright © 2026 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

股票市场是金融体系的核心, 股价波动不仅反映宏观经济态势, 也深刻影响投资者的资产配置与风险管控。受宏观政策、市场情绪、企业经营、国际资本流动等多重因素影响, 股价序列表现出高噪声、非线性、非平稳的特点, 使得股价预测成为金融工程与计算机科学交叉领域的难点问题[1]。

传统自回归移动平均、马尔可夫链等计量模型, 仅能解释市场线性趋势与短期状态变化[2]。因其理论假设偏于理想化, 无法有效挖掘股市深层非线性时序规律, 预测效果并不理想。随着大数据与人工智能技术发展, 深度学习逐步成为股价预测的主流手段[3]。黄建华等人[4]针对 LSTM 超参数优化难题, 采用改进粒子群算法调参, 并结合特征选择削减数据冗余。但单一深度学习模型难以同时学习全局趋势与局部高频波动, 容易产生过拟合、预测滞后等问题。

为此, 诸多研究引入信号分解方法优化模型: 黄后菊[5]等人运用变分模态分解拆分股价序列, 结合麻雀搜索算法优化 LSTM, 有效削弱了序列非平稳性带来的干扰; 白军成[6]等人融合聚类与集成经验模态分解, 结合相空间特征选择构建多神经网络集成预测框架, 实现分时序子序列差异化建模预测。

现有研究仍存在明显短板: 传统 VMD 依赖人工设定参数, 自适应能力弱, 易出现模态混叠或过分解; 多数模型未区分股价频域特征与技术指标的属性差异, 特征融合效率偏低。本文基于平均样本熵优化 VMD 算法, 实现频域降噪与多尺度特征解耦; 利用 PCA 对量价及技术指标降维, 消除多重共线性。同时搭建双分支 BiLSTM 并行框架, 充分挖掘时频特征与非线性关联, 实现多维特征高效互补。

2. 模型介绍

2.1. 自适应变分模态分解

利用传统经验模态分解及其衍生算法处理非平稳信号, 常面临模态混叠以及端点效应等问题。针对上述局限, Konstantin Dragomiretskiy 等提出一种完全非递归信号处理方法 - 变分模态分解。该方法的核心是将信号分解过程转化为一个约束变分问题的求解过程, 原始信号 $f(t)$ 自适应分解为 K 组本征模态函数(Intrinsic Mode Functions, IMF), 各 IMF 均具有特定中心频率 ω_k 、有限带宽 $u_k(t)$ 。

具体而言, 通过 Hilbert 变换计算各模态 $u_k(t)$ 单边频谱, 获得其解析信号。并利用指数项 $e^{-j\omega_k t}$ 实现解析信号频谱混合, 完成中心频率调制; 进而, 通过计算解调信号梯度平方 L^2 范数实现带宽估计。约束变分问题如式(1)所示:

$$\min_{\mathbf{u}, \boldsymbol{\omega}} \left\{ \sum_{k=1}^K \left\| \partial_t \left[\left(\delta(t) + \frac{j}{\pi t} \right) * u_k(t) \right] e^{-j\omega_k t} \right\|_2^2 \right\} \quad (1)$$

$$\text{s.t. } \sum_{k=1}^K u_k(t) = f(t)$$

式中: $\mathbf{u} = \{u_1, \dots, u_K\}$ 、 $\boldsymbol{\omega} = \{\omega_1, \dots, \omega_K\}$ 分别表示各模态及其中心频率; K 表示 IMF 数目; $\partial_t(\cdot)$ 表示时间偏导函数; $\delta(t)$ 表示狄拉克分布; $*$ 表示卷积运算; $j/\pi t$ 表示希尔伯特变换核; $\|\cdot\|_2^2$ 表示范数平方。

给定候选值 K , 针对原始信号 $f(t)$ 执行 VMD 分解操作得到 IMF 分量。为综合评价分解效果, 使用平均样本熵作为目标函数, 如式(2)所示。

$$\Gamma(K, \alpha) = \frac{1}{K} \sum_{k=1}^K \text{SampEn}(u_k, m, r) \quad (2)$$

式中: $\Gamma(K, \alpha)$ 表示给定候选模态数以及惩罚因子 (K, α) , 计算所得平均样本熵目标函数值; u_k 表示第 k 个本征模态函数; $\text{SampEn}(u_k, m, r)$ 表示 u_k 样本熵值; m 、 r 分别表示样本熵计算嵌入维度以及相似容限阈值。

基于此, VMD 参数自适应选择可形式化为如式(3)所示:

$$\{K^*, \alpha^*\} = \arg \min_{K \in [K_{\min}, K_{\max}], \alpha \in [\alpha_{\min}, \alpha_{\max}]} \Gamma(K, \alpha) \quad (3)$$

式中: $K_{\min}$ 、 $K_{\max}$ 分别表示模态数搜索空间最小值以及最大值; $\alpha_{\min}$ 、 $\alpha_{\max}$ 分别表示惩罚因子搜索空间最小值以及最大值。

2.2. 主成分分析

本节引入基础量化指标(开盘价、最高价、最低价、成交量)以及技术衍生指标(5 日移动平均线、10 日移动平均线、布林带中轨、平均真实波幅、波动率、平均趋向指标)。引入主成分分析(Principal Component Analysis, PCA)方法, 通过正交变换可实现原始特征重新组合获得线性无关主成分(Principal Components, PC), 并最大化保留原始数据集总方差。其核心数学过程如下:

给定经标准化处理所得特征矩阵 $\mathbf{X} \in \mathbb{R}^{n \times m}$ (其中 n 表示样本数, m 表示特征数), 计算其协方差矩阵 $\mathbf{C}$ 。其数学表达式如式(4)所示:

$$\mathbf{C} = \frac{1}{n-1} \mathbf{X}^T \mathbf{X} \quad (4)$$

同时, 针对高频金融数据协方差矩阵的奇异性问题, 引入 Ledoit-Wolf 收缩估计进行矩阵正则化处理。其数学表达式如式(5)所示:

$$\mathbf{C}' = (1-\delta)\mathbf{C} + \delta \text{Tr}(\mathbf{C}) \mathbf{I}_m / m \quad (5)$$

式中: $\mathbf{C}'$ 表示正则化协方差矩阵; δ 表示最优收缩系数, 通过最小化均方误差估计获得; $\text{Tr}(\cdot)$ 表示矩阵迹运算; $\mathbf{I}_m$ 表示 m 维单位矩阵。

通过特征分解求解协方差矩阵 $\mathbf{C}'$ 特征值 $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_m \geq 0$ 及其相应单位正交特征向量 $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_m$, 并定义各主成分累计方差贡献率。其数学表达式如式(6)所示:

$$\mathbf{C}' \mathbf{v}_k = \lambda_k \mathbf{v}_k, \quad k=1, 2, \dots, m \quad (6)$$

$$\Psi_d = \sum_{k=1}^d \psi_k = \frac{\sum_{k=1}^d \lambda_k}{\sum_{i=1}^m \lambda_i} \quad (7)$$

为保留原始数据主要信息结构, 遵循 Kaiser 准则, 即选取特征值大于 1 的主成分构建投影矩阵, 原始特征空间至主成分空间投影变换如式(7)所示:

$$\mathbf{T} = \mathbf{XV}_d \quad (8)$$

式中: $\mathbf{V}_d = [\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_d]$ 为前 d 个特征向量所构成投影矩阵; $\mathbf{T} \in \mathbb{R}^{n \times d}$ 表示降维特征矩阵。

2.3. 双向长短期记忆网络

LSTM 记忆单元 t 时刻更新过程可分为遗忘门计算、输入门与候选细胞状态生成、细胞状态更新、输出门与隐藏层输出生成阶段。

遗忘门决定由时刻 $t-1$ 细胞状态选择性丢弃信息比例, 用以实现历史冗余信息自适应遗忘。其数学表达式如式(9)所示:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad (9)$$

式中: $\sigma(\cdot)$ 表示 Sigmoid 激活函数, 输出范围为 $(0, 1)$, 用于将输入映射为门控的概率权重; W_f 表示遗忘门权重矩阵; $[h_{t-1}, x_t]$ 表示时刻 $t-1$ 隐藏层输出 h_{t-1} 以及时刻 t 输入向量 x_t 所构成拼接向量; b_f 表示遗忘门偏置向量; f_t 表示时刻 t 遗忘门输出。

输入门动态决定时刻 t 输入信息 x_t 纳入新细胞状态比例, 用以实现新增信息选择性保留。候选细胞状态则生成时刻 t 待纳入细胞状态候选信息, 包含当前输入以及历史隐藏层输出非线性变换结果。其数学表达式如式(10)~(11)所示:

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad (10)$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \quad (11)$$

式中: W_i 、 W_C 分别表示输入门、候选细胞状态权重矩阵; b_i 、 b_C 分别表示输入门、候选细胞状态偏置向量; $\tanh(\cdot)$ 表示双曲正切激活函数, 输出范围为 $(-1, 1)$, 用于将输入映射为归一化候选信息; i_t 、 $\tilde{C}_t$ 分别表示时刻 t 输入门输出、候选细胞状态向量。

细胞状态更新阶段则结合遗忘门信息丢弃结果以及输入门信息纳入结果, 完成细胞状态的动态更新, C_t 实现历史信息、当前信息融合, 其数学表达式如式(12)所示:

$$C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t \quad (12)$$

式中: $\odot$ 表示 Hadamard 逐元素积, 用于实现门控输出与状态向量逐元素加权; C_t 、 C_{t-1} 分别表示时刻 t 以及时刻 $t-1$ 细胞状态向量。

输出门动态控制当前细胞状态输出至隐藏层信息比例, 用以实现记忆信息选择性暴露。隐藏层输出则作为当前时刻网络输出, 同时传递至下一时刻 LSTM 单元。其数学表达式如式(13)~(14)所示:

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (13)$$

$$h_t = o_t \odot \tanh(C_t) \quad (14)$$

式中: W_o 表示输出门权重矩阵; b_o 表示输出门偏置向量; o_t 表示时刻 t 输出门输出; h_t 表示时刻 t 隐藏层输出向量。

同时, 采用双向长短期记忆网络构建时序特征提取模块, 其网络结构并行设置正向 LSTM 分支以及反向 LSTM 分支。

针对时刻 t 输入 x_t , 通过并行前向、后向 LSTM 分支实现双向时序特征提取, 其前向隐藏层状态 $\overrightarrow{h_t}$

和后向隐藏层状态 $\bar{h}_t$ 分别计算, 其数学表达式如式(15)所示:

$$\begin{cases} \bar{h}_t = \text{LSTM}_{\text{fwd}}(x_t, \bar{h}_{t-1}) \\ \bar{h}_t = \text{LSTM}_{\text{bwd}}(x_t, \bar{h}_{t+1}) \end{cases} \quad (15)$$

式中: $\bar{h}_t$ 、 $\bar{h}_t$ 分别表示时刻 t 前向隐藏层、后向隐藏层状态; LSTM_{fwd} 、 LSTM_{bwd} 分别表示前向、后向 LSTM 单元, 各自包含独立门控权重矩阵、偏置向量, 可实现双向时序信息独立建模。

隐藏层输出 H_t 则由前向隐藏层、后向隐藏层状态融合输出, 其数学表达式如式(16)所示:

$$H_t = [\bar{h}_t \oplus \bar{h}_t] \quad (16)$$

式中: $\oplus$ 表示特征维度拼接操作; H_t 表示隐藏层输出。

3. VMD-PCA-BiLSTM 股票预测模型框架

针对股票市场多源异构数据(基础量价数据、多维技术指标)非线性、非平稳特性, 以及单一模型难以同时适配高频波动捕捉、长期趋势建模、指标冗余剔除等局限性, 本文提出一种频域-时域并行学习框架, 并构建 VMD-PCA-BiLSTM 混合预测模型。VMD-PCA-BiLSTM 混合预测模型具体实施步骤如下。

步骤 1: 数据预处理与特征集划分

针对原始股票交易数据执行清洗操作, 并划分为目标序列、辅助特征集, 其中目标序列即为股票收盘价序列, 用于后续频域分解; 辅助特征集则包含开盘价、最高价、最低价、成交量及其他技术衍生指标。上述所有输入变量执行 Z-Score 标准化处理, 映射至标准正态分布, 以消除量纲差异影响。

步骤 2: 目标序列自适应 VMD 分解

针对 VMD 算法模态数 K 以及惩罚因子, 引入平均样本熵作为评价指标。通过遍历搜索不同参数组合相应模态分量熵值, 以序列复杂度最小化、模态区分度最大化作为寻优目标, 确定最佳参数组合。利用最优参数实现收盘价序列变分模态分解, 获取若干本征模态函数, 多尺度频域特征向量即作为频域分支输入。

步骤 3: 辅助特征集 PCA 降维重构

构建包含基础量价数据、多维技术指标高维辅助特征矩阵。利用主成分分析方法针对该矩阵进行正交变换, 通过计算协方差矩阵特征值以及累计贡献率, 提取主要成分以重构低维特征空间, 降维特征矩阵即作为时域分支输入。

步骤 4: 双分支 BiLSTM 时序特征提取

构建并行双分支 BiLSTM 网络结构。分支 1 以分解所得 IMF 序列作为输入, 利用双向循环结构捕捉股价自身演化多尺度时频特征; 分支 2 以 PCA 降维所得主成分序列作为输入, 重点提取外部市场环境、技术指标间非线性耦合关系。

步骤 5: 多维特征融合与回归预测

提取上述 BiLSTM 分支 1、分支 2 隐藏层状态向量, 执行特征级拼接, 实现时/频信息深度融合。融合特征向量随即输入 Dropout 层抑制过拟合风险, 进而通过全连接层进行非线性映射, 获得收盘价预测值。

4. 算例验证分析

4.1. 仿真实验设置

为全面验证本文所提 VMD-PCA-BiLSTM 模型有效性, 本节基于真实股票市场数据开展系统性仿真

实验。实验环境基于 Python3.8、TensorFlow2.10.1 框架。选取上证指数作为实证研究对象, 该实验数据通过 Akshare 财经数据接口获取, 采样区间为 2018 年 1 月 1 日至 2024 年 3 月 8 日(约 1500 个交易日样本), 覆盖牛市、熊市以及震荡完整市场周期行为。数据集遵循时间轴顺序进行划分: 前 80% 作为训练集, 用于模型权重学习、超参数微调以及 VMD 分解参数自适应寻优; 后 20% 作为测试集, 用于最终性能评估。

4.2. 数据预处理分析

为实现收盘价序列频域自适应分解, 首先基于训练集收盘价序列针对自适应 VMD 关键参数进行多目标寻优, 确定最优参数组合为 $K = 3$ 、 $\alpha = 2000$; 随后, 利用该组最优参数, 针对测试集收盘价序列进行分解, 所得各阶 IMF 序列如图 1 所示。

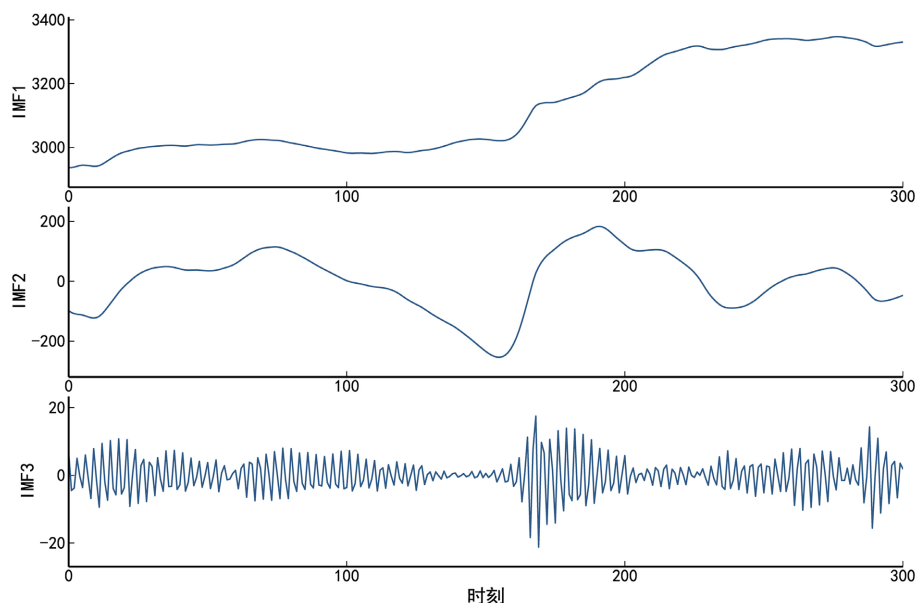


Figure 1. Results of VMD modal decomposition for stock time series
图 1. 股票序列 VMD 模态分解结果

由图 1 可见, 原始收盘价序列经自适应 VMD 分解, 可解耦为高频、中频、低频子序列。其中, 高频 IMF 分量呈现出高频、低幅随机震荡特征, 该部分分量主要对应股票市场高频交易噪声以及短期情绪扰动, 微观结构信息丰富, 同时其单独可预测性较弱, 容易放大局部残差。中频 IMF 分量呈现周期性波动规律, 具备稳定时频尺度, 承载序列中等周期尺度振荡模态信息, 该部分分量主要对应市场波段性调整、行业轮动周期或季度性资金流动特征。低频 IMF 分量呈现趋势项平滑分布特征, 幅值较大且惯性强, 表征原始序列全局缓变特征, 可表明宏观经济基本面驱动市场长期运行轨迹。

进而, 结合金融学含义深入分析, 高频 IMF 分量对应波段主要反映交易噪声、短期情绪冲击以及流动性扰动, 中频分量对应波段主要反映资金轮动、风险偏好切换等中期市场结构; 低频分量对应波段主要反映宏观经济基本面驱动市场长期运行轨迹, 提供模型识别指数长期运行方向重要依据。根据上述分解方式可见, 自适应 VMD 能够实现不同时频尺度金融价格信号精准剥离, 有效分离噪声、中期周期以及长期趋势驱动因素, 用于提供后续多尺度融合预测模型清晰分层输入。

PCA 特征值分布以及方差解释比例如图 2 所示, 该结果可为 VMD-PCA-BiLSTM 预测模型提供数据支撑。由图 2 可见, 共有 2 项主成分特征值满足筛选条件, 其数值分别为 6.41、2.07, 累计方差贡献率达

84.78%。上述数据说明保留 PC1、PC2, 已有效覆盖原始高维特征空间核心信息, 实现高维技术指标降维解耦, 进而大幅降低后续模型计算复杂度。此外, 结合主成分载荷金融含义分析, PC1 主成分主要源自开盘价、最高价、最低价以及移动均线等市场状态因子, 反映指数运行趋势延续性; PC2 主成分主要源自成交量、平均真实波幅、波动率、平均趋向指标等波动强度/交易活跃度信息, 用作补充解释市场风险偏好变化。

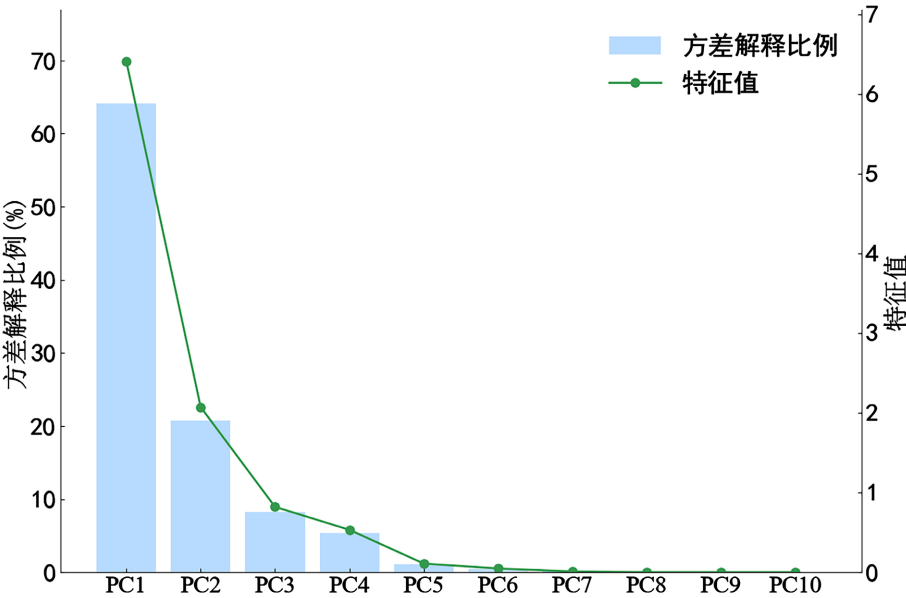


Figure 2. PCA eigenvalue distribution and percentage of variance explained
图 2. PCA 特征值分布以及方差解释比例

4.3. 模型预测性能对比分析

基于测试集数据, BiLSTM、PCA-BiLSTM、VMD-BiLSTM 以及本文 VMD-PCA-BiLSTM 模型所得各项评价指标如表 1 所示。本文选取以下评价指标全面评估模型性能: 均方根误差(RMSE)、平均绝对误差(MAE)、决定系数(R^2)、平均绝对百分比误差(MAPE)。

Table 1. Comparison of performance evaluation metrics for different models
表 1. 不同模型预测性能评价指标对比

模型	RMSE	MAE	MAPE	R^2
BiLSTM	42.02	32.02	9.99	95.62
PCA-BiLSTM	38.90	23.74	7.53	96.25
VMD-BiLSTM	34.93	24.47	7.72	96.97
VMD-PCA-BiLSTM	28.55	20.28	6.47	97.95

由表 1 可见, BiLSTM 模型作为基础对照组, 其各项误差指标均为最高, RMSE、MAE、MAPE 分别为 42.02、32.02、9.99, 同时决定系数 R^2 为 95.62。表明由于该模型缺乏数据预处理机制, 针对包含噪声、冗余以及非平稳特征原始数据, 难以剥离冗余信息, 导致预测结果存在较大偏差。其次, PCA-BiLSTM、VMD-BiLSTM 模型分别引入单一预处理策略, 可缓解原始数据复杂性, 但仍存在特定结构性弱点。其中, PCA-BiLSTM 通过线性降维机制, 可实现 RMSE 降低至 38.90、 R^2 微幅提升至 96.25, 然而 RMSE 仍显著

高于最优模型, 表明仅通过主成分分析降低计算冗余过程易忽略部分非线性特征, 导致信息利用率不充分。VMD-BiLSTM 模型通过模态分解机制, 可实现 RMSE 优化至 34.93, 然而其 MAE 指标为 24.47 略高于 PCA-BiLSTM, 表明仅通过信号分解导致子序列中仍残留冗余特征或高频噪声, 限制模型精度进一步收敛。

同时, BiLSTM、PCA-BiLSTM、VMD-BiLSTM 以及本文 VMD-PCA-BiLSTM 模型所得预测曲线如图 3 所示。VMD-PCA-BiLSTM 模型 RMSE 降低至 28.55、MAPE 降低至 6.47, 相较于基准模型 MAPE 普遍处于 7.53%至 9.99%区间, 说明 BiLSTM 模型通过结合股票收盘价 VMD 分解、基础量价数据及其技术衍生指标 PCA 降维处理, 可融合感知股票市场微观波动细节以及宏观趋势信息, 进而可提升股票收盘价预测精度。

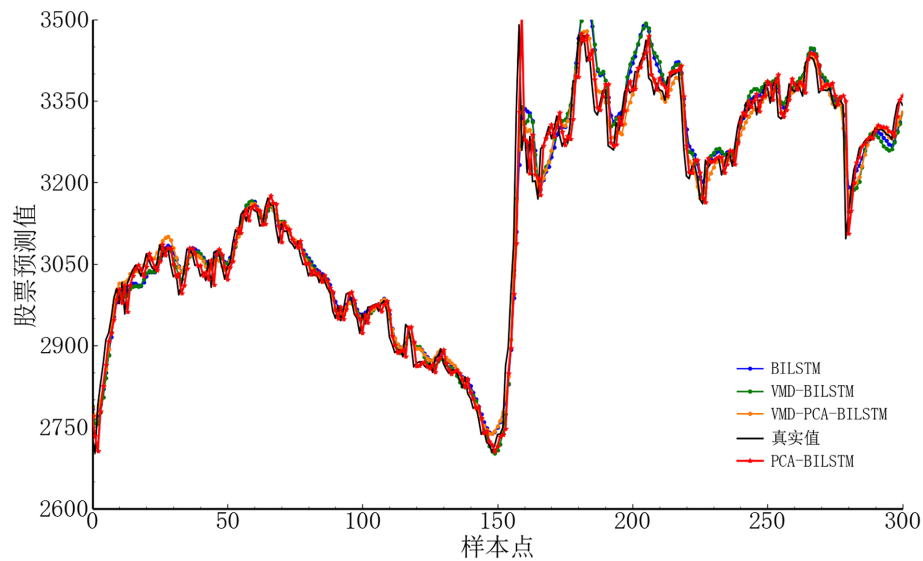


Figure 3. Comparison of prediction curves from different models
图 3. 不同模型预测曲线对比

基于图 3 预测分析结果所示, 真实收盘价序列约在样本点 150 处发生结构性突变, 呈现显著非平稳高频震荡特征。针对该处特征, 单一 BiLSTM 模型存在严重平滑效应、跟踪时滞问题, 针对局部极值捕捉呈现欠拟合现象。PCA-BiLSTM 模型、VMD-BiLSTM 模型可在一定程度修正趋势偏差, 瞬态高波动区间仍呈现跟踪时滞现象。相比之下, 本文 VMD-PCA-BiLSTM 模型可在市场平稳期以及波动期均保持高拟合精度, 并且未出现平滑效应、跟踪时滞问题, 即可说明本文双分支特征提取策略可有效剥离噪声并重构多尺度特征, 显著提升模型动态适应性以及鲁棒性。

4.4. 模型预测误差分析

针对复杂市场环境预测模型失效边界, 本文 VMD-PCA-BiLSTM 模型预测误差较大交易日如表 2 所示。由表 2 可见, 最大误差交易日集中分布于 2024 年 9 月末至 10 月上旬, 以及 2025 年 4 月上旬。

Table 2. The VMD-PCA-BiLSTM model had large prediction errors on trading days

表 2. VMD-PCA-BiLSTM 模型预测误差较大交易日

交易日	序号	真实值	预测值	误差	相对误差
2024-09-26	155	3000.95	2910.98	-89.98	3.00%

续表

2024-09-30	157	3336.50	3166.16	-170.34	5.11%
2024-10-08	158	3489.78	3261.02	-228.76	6.56%
2024-10-15	163	3201.29	3278.42	77.13	2.41%
2025-04-07	279	3096.58	3227.84	131.26	4.24%

结合具体市场背景, 2024 年 9 月 24 日金融支持经济高质量发展新闻发布会释放降准、降息以及创设支持股票市场稳定发展等信号, 市场风险偏好迅速修复。针对测试样本, 2024 年 9 月 26 日、9 月 27 日、9 月 30 日以及 10 月 8 日上证指数收益率分别为 3.61%、2.88%、8.06%以及 4.59%, 其中 9 月 30 日 A 股全天成交超 2.6 万亿元并创历史新高, 10 月 8 日节后首日上证指数收于 3489.78 点、沪市成交额达 15105.13 亿元, 沪深两市成交额超 3.45 万亿元。同时, 结合市场波动率变化, 2024 年 10 月 9 日上证指数回落 6.62%, 5 日滚动收益率标准差由 9 月 26 日 1.88%上升至 10 月 9 日 5.47%, 表明政策行情高位震荡已经显著超出前期常态波动区间。进而, 上述此类成交量异常放大以及价格跳升情况, 导致基于历史窗口学习所得模型难以及时更新, 导致模型预测误差加大。

4.5. 模型泛化性能分析

针对本文预测模型跨市场样本泛化能力, 本节选取 2020 年 4 月 9 日至 2026 年 6 月 17 日深证指数交易日样本进行扩展验证。相较于上证指数, 深证指数成分股中成长型、科技型以及中小市值公司权重较高, 指数波动弹性、风格切换特征显著, 因此能够支撑检验模型鲁棒性。实验采用 VMD-BiLSTM、PCA-BiLSTM、VMD-PCA-BiLSTM 以及 BiLSTM 预测模型进行对比分析, 评价指标如表 3 所示, 深证指数不同模型预测曲线对比如图 4 所示。

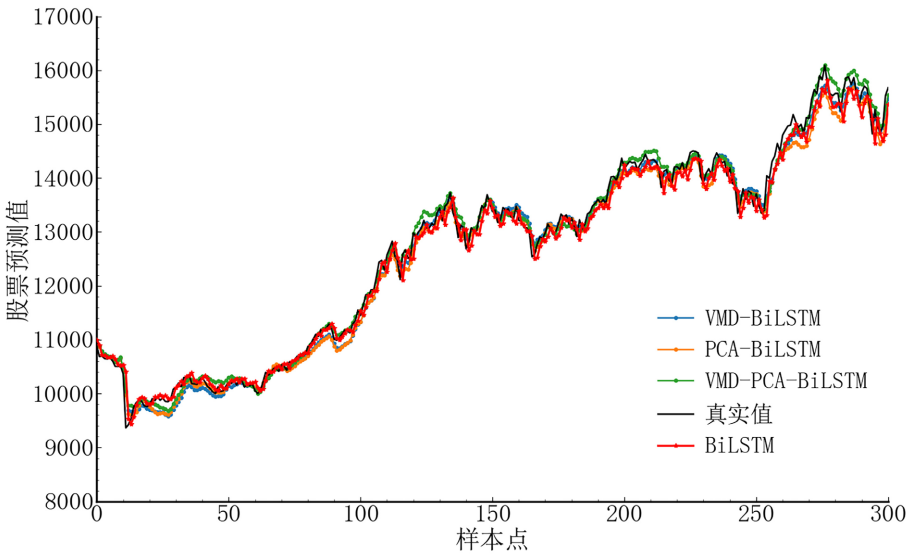


Figure 4. Comparison of prediction curves from different models of the Shenzhen Stock Index
图 4. 深证指数不同模型预测曲线对比

由表 3 可见, 关于深证指数样本, VMD-PCA-BiLSTM 模型预测效果最优, 其 RMSE、MAE 和 MAPE 分别为 146.22、110.50、0.87%, 决定系数达 99.37%。说明面向市场结构、波动幅度、风格特征变化场景, 频域分解、主成分降维联合建模框架预测性能较为稳定。

Table 3. Comparison of performance evaluation indicators for different models of the Shenzhen index
表 3. 深证指数不同模型预测性能评价指标对比

模型	RMSE	MAE	MAPE	R^2
BiLSTM	208.69	153.03	1.17	98.71
PCA-BiLSTM	191.49	149.71	1.16	98.92
VMD-BiLSTM	160.75	128.17	1.03	99.24
VMD-PCA-BiLSTM	146.22	110.50	0.87	99.37

由图 4 可见, 对于深证指数样本, 各模型均能够跟踪总体趋势。然而, 单一 BiLSTM 高波动区间存在平滑效应, PCA-BiLSTM 局部峰谷跟踪不足, VMD-BiLSTM 波动幅度变化较大区间易产生局部偏差。相比之下, VMD-PCA-BiLSTM 各类场景均能保持趋势贴合以及波动响应, 具有跨指数预测泛化能力。

4.6. 模型时序交叉验证回测

面向真实交易场景, 本节采用滑动窗口思想进行时序回测, 保持样本时间顺序不变, 基于上证指数测试区间划分得到 5 组连续子窗口, 并分别统计各窗口 VMD-PCA-BiLSTM 模型样本外预测误差, 上证指数测试区间滑动窗口回测结果如表 4 所示。基于上述方法即可检验真实市场环境模型预测稳定性, 避免单一整体误差掩盖局部阶段性风险。

Table 4. Shanghai composite index test range sliding window backtest results
表 4. 上证指数测试区间滑动窗口回测结果

窗口	序号范围	样本数	RMSE	MAE	MAPE
窗口 1	0~60	61	25.62	19.67	0.66
窗口 2	61~120	60	16.41	13.42	0.45
窗口 3	121~180	60	50.04	33.02	1.06
窗口 4	181~240	60	27.40	21.54	0.65
窗口 5	241~300	60	25.57	16.40	0.50

由表 4 可见, 本文模型 5 组连续窗口 MAPE 均处于 1.10% 以内, 说明整体预测稳定性较好。其中, 第 2 窗口 RMSE 仅为 16.41, 表明趋势较平稳阶段模型能够实现较高精度拟合; 第 3 窗口 RMSE 升至 50.04, 显著高于其他窗口, 对应前文第 4.4 小节分析所得样本点 157~158 附近局部剧烈波动现象, 说明市场状态快速切换显著影响模型预测性能。第 4、第 5 窗口误差重新回落, 表明模型具备自我修正能力。

5. 结语

本文依托股票量价与技术指标数据, 构建 VMD-PCA-BiLSTM 预测模型, 采用频域、时域并行学习框架。研究结合平均样本熵优化 VMD 参数, 解决传统方法模态混叠、过分解缺陷, 实现股价序列频域解耦与降噪; 利用 PCA 对多维指标降维重构, 有效消除多重共线性。模型通过双分支 BiLSTM 分别提取频域、时域特征并完成融合, 适配数据特性实现协同建模。实验表明, 该模型关于上证指数预测 RMSE 为 28.55, 决定系数达 97.95%, 预测精度、鲁棒性与动态适应性优异, 改善了平滑效应与时滞问题, 可快速响应价格波动, 具备实际应用价值。

参考文献

- [1] Kehinde, T.O., Chan, F.T.S. and Chung, S.H. (2023) Scientometric Review and Analysis of Recent Approaches to Stock Market Forecasting: Two Decades Survey. *Expert Systems with Applications*, **213**, Article 119299. <https://doi.org/10.1016/j.eswa.2022.119299>
- [2] 穆阳, 李多全. 基于马尔可夫链方法的股票价格预测研究[J]. 金融发展研究, 2024(8): 89-92.
- [3] Dave, D., Sawhney, G. and Chauhan, V. (2025) Multi-Agent Stock Prediction Systems: Machine Learning Models, Simulations, and Real-Time Trading Strategies.
- [4] 黄建华, 钟敏, 胡庆春. 基于改进粒子群算法的 LSTM 股票预测模型[J]. 华东理工大学学报(自然科学版), 2022, 48(5): 696-707.
- [5] 黄后菊, 李波. 基于 VMD-CSSA-LSTM 组合模型的股票价格预测[J]. 南京信息工程大学学报, 2024, 16(3): 332-340.
- [6] 白军成, 孙秉珍, 郭誉齐, 等. 融合三支聚类与分解集成学习的股票价格预测模型[J]. 运筹与管理, 2024, 33(8): 213-218.