Huang *et al. Financial Innovation* (2024) 10:97
https://doi.org/10.1186/s40854-024-00622-6

Financial Innovation

RESEARCH Open Access

A structural VAR and VECM modeling method for open-high-low-close data contained in candlestick chart

Wenyang Huang[1], Huiwen Wang[2,3] and Shanshan Wang[2,4*]

*Correspondence: sswang@buaa.edu.cn

[1] College of Economics and Management, China Agricultural University, Beijing, China
[2] School of Economics and Management, Beihang University, Beijing, China
[3] Key Laboratory of Complex System Analysis, Management and Decision (Beihang University), Ministry of Education, Beijing, China
[4] Beijing Advanced Innovation Center for Big Data and Brain Computing, Beijing, China

Abstract

The structural modeling of open-high-low-close (OHLC) data contained within the candlestick chart is crucial to financial practice. However, the inherent constraints in OHLC data pose immense challenges to its structural modeling. Models that fail to process these constraints may yield results deviating from those of the original OHLC data structure. To address this issue, a novel unconstrained transformation method, along with its explicit inverse transformation, is proposed to properly handle the inherent constraints of OHLC data. A flexible and effective framework for structurally modeling OHLC data is designed, and the detailed procedure for modeling OHLC data through the vector autoregression and vector error correction model are provided as an example of multivariate time-series analysis. Extensive simulations and three authentic financial datasets from the Kweichow Moutai, CSI 100 index, and 50 ETF of the Chinese stock market demonstrate the effectiveness and stability of the proposed modeling approach. The modeling results of support vector regression provide further evidence that the proposed unconstrained transformation not only ensures structural forecasting of OHLC data but also is an effective feature-extraction method that can effectively improve the forecasting accuracy of machine-learning models for close prices.

Keywords: OHLC data, Structural modeling, Unconstrained transformation, Candlestick chart, VAR, VECM

Introduction

Technical analysis emerges as a preeminent investment analysis method in financial markets with the purpose of detecting price trends at an early stage to seize and profit from trading opportunities. The efficient market hypothesis (EMH) argues that financial prices comprehensively encapsulate all available information, rendering consistent market outperformance an implausible endeavor through equity selection and timing trades-that is, technical analysis is ineffective (Fama 1970). However, the premises of EMH are often disrupted by reality. For instance, the EMH assumes that financial prices are random walks, whereas there are often historical repetitions in investor behavior, leading to regularity in stock price fluctuations. Typical examples include the weekend

(Doyle and Chen 2009), calendar (Ariss et al. 2011), and week-of-the-year effects (Levy and Yagil 2012). After decades of development, numerous scholars have confirmed the feasibility of technical analysis, and various methodologies have flourished (Hsu and Kuan 2005; Smith et al. 2016; Ilham et al. 2022).

Currently, candlestick chart analysis has become the most intuitive and extensively employed technical analysis methodology for analyzing the price movements of financial products (Romeo et al. 2015) owing to its precise definition (Caginalp and Laurent 1998), efficient representation of market signals, and cogent reflection of investors' overall motivation (Tsai and Quan 2014). The candlestick chart encapsulates the four-dimensional price data of a particular financial product during a given period, including the open, high, low, and close prices, collectively termed OHLC data (Huang et al. 2022b). The academic community has spared no effort in studying the candlestick charts and the OHLC data contained within them, channeling their explorations into two domains: graphical analysis and numerical analysis.

Scholars engaged in graphical analysis ardently endeavor to forecast future price trends by identifying repetitive patterns in candlestick charts. Notably, Caginalp and Laurent (1998) investigated the candlestick charts of the S &P 500 from 1992 to 1996 and confirmed the significant predictive power of the 8-day reversal pattern using an out-of-sample test. The pattern demonstrated an empirical capacity to earn a profit of nearly 1% over a 2-day holding period. Goo et al. (2007) compared to the average returns across different patterns and holding days based on daily data of 25 blue chip stocks in Taiwan from 1997 to 2006. The empirical results show that investors can earn an average return of 9.99% by holding the Bullish Harami pattern for 10 days. Lan et al. (2011) employed fuzzy logic theory to define the sequence of symptoms before the appearance of reversal points and identified the reversal patterns of candlestick charts in Chinese stock markets. Lu et al. (2012) identified three bullish and three bearish reversal patterns based on the Taiwan Top 50 Tracker Fund data in 2002–2008. Cutting-edge research on the graphical analysis of candlestick charts based on machine-learning methods and artificial neural networks can be found in Ramadhan et al. (2022), Santur (2022), Cagliero et al. (2023), Chen et al. (2023), and Varghese et al. (2023).

However, the predictability and profitability of the patterns derived from the graphical analysis tend to lack universality, evincing notable disparities across different markets and periods. For instance, Shiu and Lu (2011) validated the excellent predictive power of the Bullish Harami pattern by exhaustively analyzing electronic securities data from the Taiwan Stock Exchange from 1998 to 2007. In contrast, Marshall et al. (2008) concluded that the Bullish Harami pattern exhibited negligible predictive ability across 59 stocks within the TOPIX Large 70 Index and 41 stocks within the TOPIX Mid 400 Index during the expansive temporal ambit of 1975–2004. Although graphical analysis commands widespread prominence, an unequivocal consensus eludes the academic domain regarding the profitability of candlestick-chart patterns (Tharavanij et al. 2017). Furthermore, graphical analysis is unable to establish a quantitative relationship between financial prices and their explanatory indicators, thereby warranting supplementary augmentation through a numerical analysis.

Numerical analysis is about buying low and selling high through short-term forecasts of financial prices. Although ubiquitous price data in financial markets always possess an

OHLC data structure, most literature concerning numerical analysis concentrates solely on the close price, using historical time series of close prices to forecast future close prices (García-Martos et al. 2013; Sun et al. 2016; García and Jaramillo-Morán 2020; Liu and Shen 2020; Xu and Zhang 2023). Some studies consider the optimal portfolio for a group of stocks based on close-price returns (Mehrjoo et al. 2014; Mahmoudi et al. 2021). A superior approach to modeling OHLC data is to treat it as interval data considering the interval consisting of low and high prices (Mager et al. 2009; von Mettenheim and Breitner 2012). For instance, Arroyo et al. (2011) utilized multilayer perception (MLP), K-nearest neighbor (KNN) algorithm, autoregressive integrated moving average (ARIMA) model, vector autoregression (VAR), vector error correction model (VECM), and exponential smoothing to perform regressions on the interval-based Dow Jones Industrial Average index and Euro-dollar exchange rate.They adopted two classical modeling methodologies for interval data: (1) the Center and Range method and (2) the Min and Max method (Hu and He 2007; Guo et al. 2012; Hao and Guo 2017). These two interval time-series modeling approaches achieved breakthroughs in the structural modeling of binary complex data. Although the modeling object has been expanded from unary to binary, these two methods can only consider low and high prices within the OHLC data.

For OHLC data, in addition to low and high prices, the open and close prices between the two boundaries possess strong explanatory power for future price movements and warrant careful consideration in forecasting models (Cheung 2007; Liu and Wang 2012; Huang et al. 2022a). Regrettably, open and close prices are beyond the scope of the Center and Range and Min and Max methods. As an extension of interval data, a novel structural modeling procedure for OHLC data should be investigated. Only a few studies have endeavored to utilize OHLC data for modeling. However, these studies only utilize OHLC data as model inputs, and their output objective remains to forecast only the close price (Liu and Wang 2012; Luo and Chen 2013; Qiu et al. 2020; Staffini 2022) or to forecast trading signals (Chang et al. 2011; Ahmadi et al. 2018; Chen and Hao 2020; Mahmoodi et al. 2023a, b). Table 1 summarizes the current state of research on OHLC data-forecasting techniques, demonstrating that the existing literature lacks structural modeling of OHLC data. This motivated us to propose a unified structural modeling framework for OHLC data.

Structural modeling of OHLC data is significant in finance practice, from which we may benefit compared to partial information modeling methodologies, especially for investors in financial markets. First, forecasting OHLC data can substantially assist investors in developing profitable investment plans. Specifically, according to the traditional forecast of the close price, investors can only try to buy a particular financial product at the closing quotation of period t and sell it at the closing quotation of period $(t+1)$ if an upward trend is forecasted (close-to-close strategy) (Dunis et al. 2011). If OHLC data are forecasted, investors can buy financial products at a price near the forecasted low price and sell them at a price near the forecasted high price to obtain excess profits (low-to-high strategy) (von Mettenheim and Breitner 2012). Furthermore, investors can sell promptly at the opening quotation to gain stop-loss profits when a bear market is forecasted (Huang et al. 2022a). In summary, trading based on OHLC data allows investors to obtain overnight returns (Cooper et al. 2008), reduce fund holding

Table 1 A summary of literature review

Model input	Model output	Methodologies and references
Daily close prices	Next day's close price	Autoregressive integrated moving average model (García-Martos et al. 2013); Spiking neural networks (Sun et al. 2016); Multilayer perceptron (García and Jaramillo-Morán 2020); Gated recurrent unit (Liu and Shen 2020); Feedforward network (Xu and Zhang 2023)
Low and high prices	Low and high prices, or center and range	MLP, KNN, ARIMA, VAR,VECM, and exponential smoothing (Arroyo et al. 2011); Historically consistent neural network (von Mettenheim and Breitner 2012); A constrained center and range joint model (Hao and Guo 2017)
Daily OHLC data	Next day's close price	Legendre neural network (Liu and Wang 2012); Weighted support vector machine (Luo and Chen 2013); Long-short term memory neural network (Qiu et al. 2020); Deep convolutional generative adversarial network (Staffini 2022)
Indicators based on daily OHLC data	Buy, sell or no-action signal	Piecewise linear representations and artificial neural networks (Chang et al. 2011); Piecewise linear representation and feature weighted support vector machine (Chen and Hao 2020); Support vector machine and heuristic algorithms (Ahmadi et al. 2018; Mahmoodi et al. 2023a, b)

periods (Dunis et al. 2011), lower overnight exposure (Kelly and Clark 2011), and derive better profits from high-low price range trades (von Mettenheim and Breitner 2012). Second, candlestick charts can be drawn according to the forecasted OHLC data, whose patterns can reveal the power of demand and supply in financial markets and reflect market conditions and investor sentiment (Nison 2001; Tsai and Quan 2014). A graphical analysis can provide further investment advice based on the patterns of the forecasted candlestick chart, such as up and down indications (Marshall et al. 2006, 2008). Third, a comprehensive information set of OHLC prices can enhance the reliability and explanatory ability of the research (Huang et al. 2022a). As pointed out in Fiess and MacDonald (2002) and Cheung (2007), OHLC prices have proven to possess significant power in explaining price fluctuations and future trends. In addition, Rogers and Satchell (1991) and Magdon-Ismail and Atiya (2003) noted that a more stable and valid estimate of return volatility can be obtained by considering the daily high, low, and open prices in addition to the traditionally used close prices. Finally, forecasting OHLC data offers the possibility of applying a wide range of multivariate modeling techniques to explore the dynamic and structural relationships between the components of multidimensional vectorized complex data (Fiess and MacDonald 2002; Huang et al. 2022b).

The challenge of structurally modeling OHLC data lies in the proper handling of inherent constraints. The three inherent constraints of OHLC data are as follows: (1) the low price should be higher than 0, (2) the high price should be greater than the low price, and (3) the open and close prices should be within the interval consisting of low and high prices. Some studies on interval data have attempted to ensure that the upper boundary is greater than the lower boundary by adding additional conditions to the models (Neto and De Carvalho 2010; González-Rivera and Lin 2013). However, this approach increases model complexity and is not suitable for OHLC data with multiple constraints. Other studies have attempted to model the four prices of OHLC data separately without considering the inherent constraints of OHLC data (Manurung et al. 2018; Kumar and Sharma 2019). The disadvantage of this method is that the modeling results are likely to

destroy the OHLC data structure. The three typical modeling failures originating from separate forecasts of open, high, low, and close prices are as follows: (1) the forecasted low price becomes negative (see Fig. 1a), (2) the forecasted high price is lower than the forecasted low price (see Fig. 1b), and (3) the forecasted open price (or forecasted close price) breaks through the boundaries consisting of the forecasted low and forecasted high prices (see Fig. 1c). These misleading forecasting results disrupt investors' plans and significantly undermine their confidence in their investments (Huang et al. 2022a).

To this end, we propose a novel unconstrained transformation method to transform OHLC data from an original four-dimensional constrained subspace into a four-dimensional real domain full space. The unconstrained transformation, along with its explicit inverse transformation, ensures that the subsequent forecasting models obtain meaningful OHLC prices. As an example of combining multivariate time-series analysis, we illustrate the detailed procedure of VAR and VECM modeling for OHLC data and support vector regression (SVR) as a special application of machine learning. Ample simulation experiments under different forecast periods, time-period basements and signal-to-noise ratios were conducted to validate the effectiveness and stability of the unconstrained transformation method. Furthermore, three financial datasets from the Kweichow Moutai, CSI 100 index, and 50 ETF from their advent in the Chinese stock market to June 14, 2019, were employed to demonstrate the empirical utility of the proposed method. The results showed a satisfactory modeling effect.

Compared with the existing literature, this study offers three contributions. First, the proposed unconstrained transformation method can properly handle the inherent constraints of OHLC data. Simulation experiments and empirical analysis demonstrate that the constraints inherent in OHLC data are satisfied throughout the numerical modeling process without increasing the complexity of the model, which enables more interpretable results. Second, this study proposes the first unified forecasting framework for OHLC. Within this framework, VAR and VECM modeling of OHLC data was implemented. The modeling procedure can take full advantage of the information contained in OHLC data, including open, high, low, and close prices. This enables a more efficient analysis and provides satisfactory predictive accuracy on the three datasets of the Kweichow Moutai, CSI 100, and CSI 50 ETF. Third, the method for dealing with unconstrained transformed variables can be generalized to all types of statistical models. The

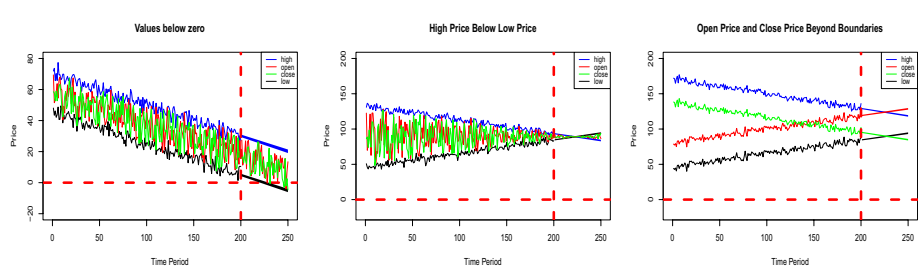


Fig. 1 Meaningless modeling results caused by ignoring the inherent constraints in OHLC data (The original data contains 200 periods, and 50 periods are forecasted forward by the linear models. The red dotted line perpendicular to the vertical axis indicates zero value, and the red dotted line perpendicular to the horizontal axis indicates the 200th period, whose right side are forecasted values with a confidence interval of 95% confidence level)

results from the extended SVR provide evidence that the proposed unconstrained transformation is an effective pre-processing technique for machine learning models and can significantly improve the forecasting accuracy of close prices compared with the direct use of raw OHLC data. From this perspective, this study provides a novel, effective, and scalable alternative to OHLC data analysis, thereby enriching existing literature on structural modeling for complex data.

The remainder of this study is organized as follows. In "Preliminaries" section, we introduce the mathematical definition of OHLC data and its inherent constraints. In "Methodology" section, we propose transformation and inverse transformation formulas to handle the inherent constraints of OHLC data and illustrate the VAR and VECM modeling processes for OHLC data. In "Simulations" section presents the simulation experiments, and "Empirical analysis" section demonstrates the empirical application of the proposed method in the real financial market. Finally, we conclude the study with a brief discussion in "Conclusions" section.

Preliminaries

To obtain an intuitive depiction of the candlestick chart, we use the daily candlestick chart in the U.S. stock market as an example (all candlestick charts refer to daily candlestick chart in this study), as shown in Fig. 2. Obviously, a daily candlestick chart can not only record the open, high, low, and close prices of a particular stock on that day but also visually reflect the difference between any two prices.

Generally, a candlestick chart is divided into two categories, as shown in Fig. 2. Specifically, Fig. 2a indicates that the close price is greater than the open price, which corresponds to a bull market, while Fig. 2b corresponds to a bear market with the close price being lower than the open price. In the U.S. stock market, green and red are habitually used to mark the real body of the candlestick chart of bull and bear markets, respectively. If daily candlestick charts are arranged in chronological order, a sequence reflecting the historical price changes of a particular financial product is formed, called the candlestick chart series, and the corresponding data are termed OHLC series.

The essence of OHLC series is a four-dimensional time series of stock prices with three inherent constraints. First, all four prices in OHLC data should be positive, because the values of the OHLC data in the financial market cannot be less than zero. Second, the high price must be higher than the low price on the same day. Third, open and close prices should fall within the boundaries of low and high

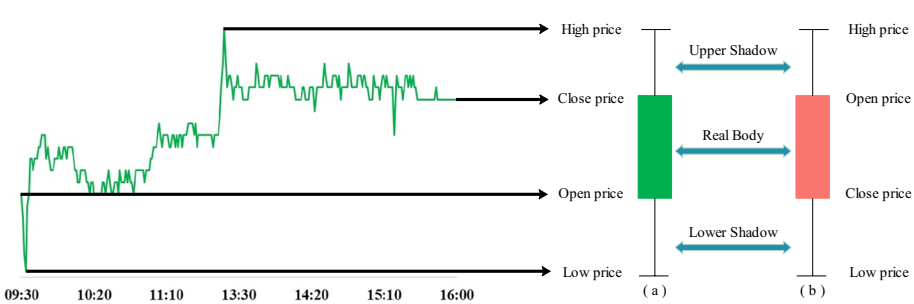


Fig. 2 An example of daily candlestick chart

prices. To represent the constraints mathematically for any time period t, we provide the following definition of OHLC data:

Definition 1 A four-dimensional vector $\boldsymbol{X}_t = (x_t^{(o)}, x_t^{(h)}, x_t^{(l)}, x_t^{(c)})^T$ is typical OHLC data if it satisfies

1. $x_t^{(l)} > 0$;
2. $x_t^{(l)} < x_t^{(h)}$;
3. $x_t^{(o)}, x_t^{(c)} \in \left[x_t^{(l)}, x_t^{(h)}\right]$.

Here, $x_t^{(o)}$ is the t-period daily open price, $x_t^{(h)}$ is the t-period daily high price, $x_t^{(l)}$ is the t-period daily low price, and $x_t^{(c)}$ is the t-period daily close price.

For the $\mathcal{T} = [1, T]$ period, the collection of $\boldsymbol{X}_t$ for any $t \in \mathcal{T}$ forms the OHLC series, denoted by

$$\boldsymbol{S} = \{\boldsymbol{X}_t\}_{t=1}^{T}.$$

Compared with the ordinary real domain vector, the biggest difference between the vectors in $\boldsymbol{S}$ is that there are intrinsic constraint formulas between its four components, which poses a significant challenge to classical statistical analysis. To establish a time-series model of OHLC series, the most difficult problem is ensuring that the calculation process and forecasting results are also subject to these constraint formulas. Otherwise, the modeling results may be meaningless. That is, after obtaining the forecasting results in the forecasting period $(T + m)$ $(m \in \mathbb{R}^+)$ using time-series modeling, it must be ensured that

$$\begin{aligned} \hat{x}_{T+m}^{(l)} &> 0, \\ \hat{x}_{T+m}^{(l)} &< \hat{x}_{T+m}^{(h)}, \\ \hat{x}_{T+m}^{(o)}, \hat{x}_{T+m}^{(c)} &\in \left[\hat{x}_{T+m}^{(l)}, \hat{x}_{T+m}^{(h)}\right]. \end{aligned}$$

These constraints are not guaranteed to be valid if we directly apply the time-series forecasting methods to the original four time series of OHLC data. To address this problem, a common practice is to remove these inherent constraints via proper data transformation. Then, we can freely forecast the transformed time-series data. Finally, we can obtain the forecaster for the original OHLC data using the corresponding inverse transformation.

Methodology

In "Data-transformation method" section, we first propose a flexible transformation method along with its inverse transformation method for OHLC data as well as a model-independent framework for modeling OHLC data. Then, we use VAR and VECM as implementations of the framework and present the corresponding forecasting procedure in "The VAR and VECM modeling process for OHLC data" section.

Data-transformation method

From Definition 1, the first constraint is $x_t^{(l)} > 0$, which can be relaxed via a commonly used logarithmic transformation. That is,

$$y_t^{(1)} = \ln x_t^{(l)}. \tag{1}$$

It is quite clear that the transformed data $y_t^{(1)}$ in Eq. (1) satisfies $-\infty < y_t^{(1)} < +\infty$ with no positive constraints. Moreover, it preserves a positive relative relationship between the original data, as the logarithm transformation is a monotonically increasing function, but also compresses the scale of the data, which reduces the absolute values of the original data and makes them somewhat more stable.

Second, to guarantee the second constraint $x_t^{(l)} < x_t^{(h)}$, that is, $x_t^{(h)} - x_t^{(l)} > 0$, the same practice as that in Eq. (1) yields

$$y_t^{(2)} = \ln\left(x_t^{(h)} - x_t^{(l)}\right), \tag{2}$$

where $y_t^{(2)}$ is also free of any constraints, which can be modelled easily.

Finally, the last constraint is $x_t^{(o)}, x_t^{(c)} \in [x_t^{(l)}, x_t^{(h)}]$, implying that both the open and close prices must be higher than the low price and lower than the high price. Without properly processing the raw data, it is highly likely that the forecasted open or close prices are beyond the boundaries. To remedy this situation, based on the concept of a convex combination, we introduce two proxy datasets, $\lambda_t^{(o)}$ and $\lambda_t^{(c)}$, which are formulated as

$$\lambda_t^{(o)} = \frac{x_t^{(o)} - x_t^{(l)}}{x_t^{(h)} - x_t^{(l)}} \quad \text{and} \quad \lambda_t^{(c)} = \frac{x_t^{(c)} - x_t^{(l)}}{x_t^{(h)} - x_t^{(l)}}. \tag{3}$$

There is $0 \leqslant \lambda_t^{(o)}, \lambda_t^{(c)} \leqslant 1$ and the original data $x_t^{(o)}$ and $x_t^{(c)}$ can be obtained as follows:

$$x_t^{(o)} = \lambda_t^{(o)} x_t^{(h)} + \left(1 - \lambda_t^{(o)}\right) x_t^{(l)}, \tag{4}$$

$$x_t^{(c)} = \lambda_t^{(c)} x_t^{(h)} + \left(1 - \lambda_t^{(c)}\right) x_t^{(l)}. \tag{5}$$

Thus, the original constraint $x_t^{(o)}, x_t^{(c)} \in [x_t^{(l)}, x_t^{(h)}]$ reduces to $0 \leqslant \lambda_t^{(o)}, \lambda_t^{(c)} \leqslant 1$ if we deal with the proxy data $\lambda_t^{(o)}$ and $\lambda_t^{(c)}$ instead of $x_t^{(o)}$ and $x_t^{(c)}$. Moreover, $\lambda_t^{(o)}$ and $\lambda_t^{(c)}$ are reasonable. Specifically, a larger $\lambda_t^{(o)}$ indicates that the open price $x_t^{(o)}$ is closer to the high price $x_t^{(h)}$, whereas a smaller $\lambda_t^{(o)}$ indicates that the open price $x_t^{(o)}$ is closer to the low price $x_t^{(l)}$. Similarly, an explanation for $\lambda_t^{(c)}$ can be obtained.

To further remove the constraint $0 \leqslant \lambda_t^{(o)}, \lambda_t^{(c)} \leqslant 1$ on $\lambda_t^{(o)}$ and $\lambda_t^{(c)}$, following the idea of logistic regression, we propose the logit transformation to obtain the unconstrained data $y_t^{(3)}$ and $y_t^{(4)}$ as follows:

$$y_t^{(3)} = \ln \frac{\lambda_t^{(o)}}{1 - \lambda_t^{(o)}}, \tag{6}$$

$$y_t^{(4)} = \ln \frac{\lambda_t^{(c)}}{1-\lambda_t^{(c)}}. \tag{7}$$

Until now, via the transformation process, the raw OHLC data $\boldsymbol{X}_t = (x_t^{(o)}, x_t^{(h)}, x_t^{(l)}, x_t^{(c)})^T$ are transformed to the unconstrained four-dimensional data $\boldsymbol{Y}_t = (y_t^{(1)}, y_t^{(2)}, y_t^{(3)}, y_t^{(4)})^T$. In summary, the proposed transformation method can be described as follows:

$$\boldsymbol{Y}_t = \begin{pmatrix} y_t^{(1)} \\ y_t^{(2)} \\ y_t^{(3)} \\ y_t^{(4)} \end{pmatrix} = \begin{pmatrix} \ln x_t^{(l)} \\ \ln\left(x_t^{(h)} - x_t^{(l)}\right) \\ \ln\left(\frac{\lambda_t^{(o)}}{1-\lambda_t^{(o)}}\right) \\ \ln\left(\frac{\lambda_t^{(c)}}{1-\lambda_t^{(c)}}\right) \end{pmatrix}, \tag{8}$$

where $\lambda_t^{(o)}$ and $\lambda_t^{(c)}$ are defined by Eq. (3). The transformation of Eq. (8) not only ranges from $-\infty$ to $+\infty$ and is an explicit inverse for the values in its range but also shares the flexibility of the well-known log and logit transformation. Furthermore, the components in $\boldsymbol{Y}_t$ have fruitful economic relevance. Specifically, $y_t^{(1)}$ measures the basic price level of a specific financial product, $y_t^{(2)}$ denotes the trading price range and measures intraday volatility, and $y_t^{(3)}$ and $y_t^{(4)}$ can be used to reflect the long- and short-game dynamics in the financial market, as they describe the relative positions of the open and close prices among the boundaries consisting of low and high prices, respectively. The larger the $y_t^{(3)}$ or $y_t^{(4)}$ is, the closer the open or close price is to the high price, respectively. Smaller $y_t^{(3)}$ or $y_t^{(4)}$ suggests that the open or close price is closer to the low price, respectively. The relative sizes of $y_t^{(3)}$ and $y_t^{(4)}$ can also reflect the bullish and bearish attributes of the market with $y_t^{(3)} > y_t^{(4)}$ implying a bearish market and $y_t^{(3)} < y_t^{(4)}$ pointing toward a bullish market. In summary, the unconstrained transformation process can effectively extract feature information from the original OHLC data, including the strength of intraday trends and price volatility. As indicated by Fiess and MacDonald (2002), the relationship between trends and volatility provides the underlying information about future price developments.

Therefore, the forecasting model of the OHLC series $\{\boldsymbol{X}_t\}_{t=1}^T$ can be transformed into forecasts for the unconstrained series $\{\boldsymbol{Y}_t\}_{t=1}^T$ with the entire real number domain and variance stability, which provides significant convenience for subsequent statistical modeling. That is, we can apply classical forecasting models (ARIMA, VAR, VECM, etc.) or machine-learning models (KNN, MLP, SVR, etc.) to $\{\boldsymbol{Y}_t\}_{t=1}^T$. After obtaining the forecaster of $\boldsymbol{Y}_t$ ($\widehat{\boldsymbol{Y}}_t = (\hat{y}_t^{(1)}, \hat{y}_t^{(2)}, \hat{y}_t^{(3)}, \hat{y}_t^{(4)})^T$, which may contain the results of m-step forecasts, $m \in \mathbb{R}^+$), we can obtain the corresponding forecaster of $\boldsymbol{X}_t$ ($\widehat{\boldsymbol{X}}_t = (\hat{x}_t^{(o)}, \hat{x}_t^{(h)}, \hat{x}_t^{(l)}, \hat{x}_t^{(c)})^T$) via the inverse transformation as follows:

$$\widehat{\boldsymbol{X}}_t = \begin{pmatrix} \hat{x}_t^{(o)} \\ \hat{x}_t^{(h)} \\ \hat{x}_t^{(l)} \\ \hat{x}_t^{(c)} \end{pmatrix} = \begin{pmatrix} \hat{\lambda}_t^{(o)}\left(\exp\{\hat{y}_t^{(1)}\} + \exp\{\hat{y}_t^{(2)}\}\right) + \left(1-\hat{\lambda}_t^{(o)}\right)\exp\{\hat{y}_t^{(1)}\} \\ \exp\{\hat{y}_t^{(1)}\} + \exp\{\hat{y}_t^{(2)}\} \\ \exp\{\hat{y}_t^{(1)}\} \\ \hat{\lambda}_t^{(c)}\left(\exp\{\hat{y}_t^{(1)}\} + \exp\{\hat{y}_t^{(2)}\}\right) + \left(1-\hat{\lambda}_t^{(c)}\right)\exp\{\hat{y}_t^{(1)}\} \end{pmatrix}, \tag{9}$$

where

$$\hat{\lambda}_t^{(o)} = \frac{\exp\left\{\hat{y}_t^{(3)}\right\}}{1+\exp\left\{\hat{y}_t^{(3)}\right\}} \quad \text{and} \quad \hat{\lambda}_t^{(c)} = \frac{\exp\left\{\hat{y}_t^{(4)}\right\}}{1+\exp\left\{\hat{y}_t^{(4)}\right\}}. \tag{10}$$

The unconstrained transformation expressed in Eq. (8) and the inverse transformation in Eq. (9) provide a new perspective for forecasting OHLC data, which makes the forecasting results obey the three inherent constraints listed in Definition 1 and thus realizes the structural modeling of OHLC data. Basically, the unified structural forecasting process for OHLC data can be summarized as a trilogy: (1) transform $\{\boldsymbol{X}_t\}_{t=1}^T$ into $\{\boldsymbol{Y}_t\}_{t=1}^T$ using the unconstrained transformation, (2) model $\{\boldsymbol{Y}_t\}_{t=1}^T$ using various time-series models to obtain $\widehat{\boldsymbol{Y}}_t^{(m)}$; and (3) conduct inverse transformation on $\widehat{\boldsymbol{Y}}_t^{(m)}$ to derive $\widehat{\boldsymbol{X}}_t^{(m)}$. A detailed unified modeling framework for OHLC data is introduced in Algorithm 1. Furthermore, the proposed method is highly feasible and can be easily generalized to any type of positive interval data with minimum and maximum boundaries greater than zero and multi-valued sequences between the two boundaries. For example, the salaries of groups of people or rainfall in different districts.

Algorithm 1 Unified forecasting framework for OHLC data

1: Get the raw candlestick charts with T periods from the financial market;
2: Extract the four-dimensional time series data in candlestick charts, record as $\{\boldsymbol{X}_t\}_{t=1}^T$;
3: Conduct transformation method to $\{\boldsymbol{X}_t\}_{t=1}^T$ and obtain $\{\boldsymbol{Y}_t\}_{t=1}^T$ according to Eq. (8);
4: Model $\{\boldsymbol{Y}_t\}_{t=1}^T$ via forecasting models to get m-step ($m \in \mathbb{R}^+$) forecasting results $\widehat{\boldsymbol{Y}}_t^{(m)}$;
5: Conduct inverse transformation method to $\widehat{\boldsymbol{Y}}_t^{(m)}$ according to Eq. (9), and the original forecaster of OHLC data $\widehat{\boldsymbol{X}}_t^{(m)}$ can be derived.

It should be noticed that, in the unconstrained transformation process, we assume that $x_t^{(o)}, x_t^{(h)}, x_t^{(l)}$, and $x_t^{(c)}$ are not equal (except for $x_t^{(o)} = x_t^{(c)}$). In other words, $x_t^{(h)} \neq x_t^{(l)} \neq 0$ and $\lambda_t^{(o)}, \lambda_t^{(c)} \notin \{0, 1\}$. However, such assumptions are inevitably spoiled in real financial markets. Here, we list the circumstances that render these assumptions invalid and provide a measure to deal with them accordingly. (1) When the subject is in trade suspension and all prices are equal to 0, namely, $x_t^{(o)} = x_t^{(h)} = x_t^{(l)} = x_t^{(c)} = 0$, we exclude these extreme cases from the raw data. (2) When $\lambda_t^{(o)}$ or $\lambda_t^{(c)}$ is equal to 0, it corresponds to $x_t^{(o)} = x_t^{(l)}$ or $x_t^{(c)} = x_t^{(l)}$, respectively. We add a random term to $x_t^{(o)}$ or $x_t^{(c)}$ and make $\lambda_t^{(o)}$ or $\lambda_t^{(c)}$ slightly greater than zero. In practice, determining the magnitude of this random term is difficult. In this study, it was set to one percent of the magnitude of the original data. In the future, the size of this random term can be treated as a model parameter for optimization. (3) When $\lambda_t^{(o)}$ or $\lambda_t^{(c)}$ is equal to 1, it indicates that $x_t^{(o)} = x_t^{(h)}$ or $x_t^{(c)} = x_t^{(h)}$, respectively. We subtract a random term from $x_t^{(o)}$ or $x_t^{(c)}$ to make $\lambda_t^{(o)}$ or $\lambda_t^{(c)}$ slightly smaller than 1. (4) When a particular financial product reaches a limit-up or limit-down as soon as the opening quotation, there is $x_t^{(o)} = x_t^{(h)} = x_t^{(l)} = x_t^{(c)} \neq 0$. If a limit-up occurs, we first multiply $x_t^{(c)}$ and $x_t^{(h)}$ by 1.1 to make a relatively large interval. If a limit-down occurs, we first multiply $x_t^{(o)}$ and $x_t^{(h)}$ by 1.1. We then conduct the measurements given in circumstances (2) and (3). This model is designed to comply with the 10% limit of the Chinese stock market. At the same time, a 10% daily fluctuation is sufficient

to reflect strong changes in financial markets. For financial markets without stop limits, the interval magnification can be appropriately increased. (5) In extreme cases, financial markets can produce negative low prices, that is, $x_t^{(l)} < 0$. For instance, the downturn in the crude oil market due to the COVID-19 pandemic caused May U.S. WTI crude oil futures to plummet, eventually closing at − 37.63 dollars per barrel on April 20, 2020 (the last trading day before the delivery date). When modeling a time series that includes such extremes, the removal of these data should be considered. This is because such extreme prices are subject to rapid adjustments, and severely distorted extremes lose their ability to forecast future price movements. For instance, the price of WTI crude oil futures on April 21, 2020, switched to futures with a delivery date of June 21, 2020, which quickly returned to positive values and closed at 10.01 dollars per barrel. Investigating alternatives to such extreme OHLC data will be a future research direction.

The VAR and VECM modeling process for OHLC data

Here, we employ the VAR and VECM as examples of the framework proposed in Algorithm 1 and present the corresponding procedure for forecasting OHLC data.

VAR for OHLC data

As one of the most widely used multiple time-series analysis methods, VAR, proposed by Sims (1980), has become an important research tool in economic studies with the advantage of capturing the linear interdependencies among multiple time series (Pesaran and Shin 1998). According to Algorithm 1, we embed the VAR into the modeling process of unconstrained four-dimensional time-series data $\{\boldsymbol{Y}_t\}_{t=1}^T$. Without loss of generality, we first assume that all time series in $\boldsymbol{Y}_t$ are stationary. Then, a p-order ($p \in \mathbb{R}^+$) VAR, denoted by VAR(p), can be formulated as

$$\boldsymbol{Y}_t = \boldsymbol{\alpha} + \boldsymbol{A}_1\boldsymbol{Y}_{t-1} + \cdots + \boldsymbol{A}_p\boldsymbol{Y}_{t-p} + \boldsymbol{w}_t = \boldsymbol{\alpha} + \sum_{j=1}^{p}\boldsymbol{A}_j\boldsymbol{Y}_{t-j} + \boldsymbol{w}_t, \quad t = (p+1), \ldots, T \tag{11}$$

where $\boldsymbol{Y}_{t-j}$ is the j-th lag of $\boldsymbol{Y}_t$; $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \alpha_3, \alpha_4)^T$ is a four-dementional vector of intercepts; $\boldsymbol{A}_j$ stands for the time-invariant 4×4 coefficient matrix; and $\boldsymbol{w}_t = (w_t^{(1)}, w_t^{(2)}, w_t^{(3)}, w_t^{(4)})^T$ is a four-dimensional error term vector satisfying:

(1) Mean zero: $E(\boldsymbol{w}_t) = \boldsymbol{0}$;
(2) No correlation across time: $E(\boldsymbol{w}_{t-k}^T\boldsymbol{w}_t) = 0$, for any non-zero k.

Writing Eq. (11) in the concise matrix form yields

$$\boldsymbol{Y} = \boldsymbol{B}\boldsymbol{Z} + \boldsymbol{U}, \tag{12}$$

where $\boldsymbol{Y} = [\boldsymbol{Y}_{p+1}, \boldsymbol{Y}_{p+2}, \cdots, \boldsymbol{Y}_T]$ is a $4 \times (T-p)$ matrix; $\boldsymbol{B} = [\boldsymbol{\alpha}, \boldsymbol{A}_1, \cdots, \boldsymbol{A}_p]$ is a $4 \times (4p+1)$ coefficient matrix; $\boldsymbol{U} = [\boldsymbol{w}_{p+1}, \boldsymbol{w}_{p+2}, \cdots, \boldsymbol{w}_T]$ is a $4 \times (T-p)$ error term matrix; and

$$\boldsymbol{Z} = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ \boldsymbol{Y}_p & \boldsymbol{Y}_{p+1} & \cdots & \boldsymbol{Y}_{T-1} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{Y}_1 & \boldsymbol{Y}_2 & \cdots & \boldsymbol{Y}_{T-p} \end{pmatrix}$$

is a $(4p+1) \times (T-p)$ matrix. Then, we can solve for the coefficient matrix $\boldsymbol{B}$ using a least-squares estimation (Lütkepohl 2005):

$$\widehat{\boldsymbol{B}} = \left(\boldsymbol{Z}^T \boldsymbol{Z}\right)^{-1} \boldsymbol{Z}^T \boldsymbol{Y}. \tag{13}$$

VECM for OHLC data

The reliability of the VAR estimation is closely related to the stationarity of the variable sequences. If this assumption does not hold, we may use a restricted VAR, that is, the VECM, in the presence of the cointegration among variables. Otherwise, the variables must first be differenced by d times until they can be modelled by VAR or VECM. As evidenced in Cheung (2007), in U.S. stock markets, stock prices are typically characterized by $I(1)$ processes, and the daily highs and lows follow a cointegration relationship. This implies that the VECM may be a more practically relevant model than the VAR in the context of forecasting the OHLC series. Here, we use the augmented Dickey–Fuller (ADF) unit root test to examine the stationarity of each variable and the Johansen test (Johansen 1988) to determine the presence or absence of the cointegration relationship.

Assuming that $\boldsymbol{Y}_t$ is integrated to order one, the corresponding VECM takes the following form:

$$\Delta \boldsymbol{Y}_t = \sum_{j=1}^{p-1} \boldsymbol{\Gamma}_j \Delta \boldsymbol{Y}_{t-j} + \boldsymbol{\gamma} \boldsymbol{\beta}^T \boldsymbol{Y}_{t-1} + \boldsymbol{\alpha} + \boldsymbol{w}_t, \tag{14}$$

where Δ denotes the first difference, and $\sum_{j=1}^{p-1} \boldsymbol{\Gamma}_j \Delta \boldsymbol{Y}_{t-j}$ and $\boldsymbol{\gamma} \boldsymbol{\beta}^T \boldsymbol{Y}_{t-1}$ are the VEC components of the first difference and the error correction term, respectively. Here, $\boldsymbol{\Gamma}_j$ is a 4×4 matrix representing the short-term adjustments among the variables across the four equations at the j-th lag. Two matrices, $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$, are of dimension $4 \times r$, where r is the order of cointegration, $\boldsymbol{\gamma}$ denotes the speed of adjustment, and $\boldsymbol{\beta}$ represents the cointegrating vector, which can be obtained using the Johansen test (Johansen 1988; Cuthbertson et al. 1992). $\boldsymbol{\alpha}$ is a 4×1-constant vector representing a linear trend, p is the lag structure, and $\boldsymbol{w}_t$ is a 4×1-vector of the white noise error term.

For the VECM in Eq. (14), Johansen (1991) employed the full-information maximum likelihood method to implement the estimation. Specifically, the main procedure consists of (1) testing whether all variables are integrated of order one by applying a unit-root test (Lai and Lai 1991), (2) determining the lag order p such that the residuals from each equation of the VECM are uncorrelated, (3) regressing $\Delta \boldsymbol{Y}_t$ against the lagged differences of $\Delta \boldsymbol{Y}_t$ and estimating the cointegrating vectors from the canonical correlations of the set of residuals from these regression equations, and (4) determining the order of cointegration r.

Discussion of parameter selection in VAR and VECM

Finally, we discuss the determination of the lag order p in the VAR model and the order of cointegration r in the VECM to model the OHLC data. First, for p, on one hand, it should be sufficiently large to fully reflect the dynamic characteristics of the constructed model; on the other hand, an increase in p will cause an increase in the parameters to be estimated; thus, the degree of freedom of the model decreases. A trade-off must be evaluated to choose p, and the commonly used criteria in practice are AIC, BIC, and Hannan–Quinn. We prefer AIC because of its conciseness, which is formulated as

$$\text{AIC}\ (p) = \ln \frac{\sum_{i=1}^{4} \sum_{j=1}^{T} \hat{u}_{ij}^2}{T} + \frac{2pK^2}{T}, \tag{15}$$

where T denotes the total period number of OHLC series, p is the VAR lag order, K is the VAR dimension, and $\hat{u}_{ij} = \hat{Y}_j^{(i)} - Y_j^{(i)} (1 \le i \le 4, 1 \le j \le T)$ represents the VAR residuals. The optimal p is obtained by minimizing AIC (p).

Second, in the order of cointegration, r indicates the dimension of the cointegrating space, and (1) if the rank of $\boldsymbol{\gamma}\boldsymbol{\beta}^T$ is 4, that is, $r = 4$ and $\boldsymbol{Y}_t$ are already stationary, the proper specification of Eq. (14) is without the error-correction term and degenerates into a VAR; (2) if $\boldsymbol{\gamma}\boldsymbol{\beta}^T$ is a null matrix, that is, $r = 0$, then there is no cointegration relation; and (3) if the rank of $\boldsymbol{\gamma}\boldsymbol{\beta}^T$ is between 0 and 4, that is, $0 < r < 4$, there exist r linearly independent columns in the matrix and r cointegration relations in the system of equations. Along the line of Johansen (1991), r is determined by constructing the "Trace" or "Eigen" test statistics, which are two widely used methods in Johansen test. For further details, please refer to Johansen (1995) and Lütkepohl (2005).

Unified modeling framework for OHLC data

In summary, as one of the popular econometric forecasting models in Step 4 of Algorithm 1, the main implementations of VAR and VECM can be summarized as Algorithm 2.

Algorithm 2 VAR and VECM framework

1: Divide $\{\boldsymbol{Y}_t\}_{t=1}^T$ into segments $\boldsymbol{Y}_i^{(q)}$; each segment involves q periods, and each segment rolls forward by one period.

2: Conduct ADF test on the four-dimensional time series data in $\boldsymbol{Y}_i^{(q)}$. If they are all stationary processes, proceed to Step 5; otherwise, proceed to Step 3.

3: Use Johansen method to perform the cointegration test on $\boldsymbol{Y}_i^{(q)}$, and determine r. If there is a cointegration relationship, proceed to Step 6; otherwise, proceed to Step 4.

4: Take one-order difference on non-stationary time series in $\boldsymbol{Y}_i^{(q)}$, and then return to Step 2. The optimal situation is iterated until $\boldsymbol{Y}_i^{(q)}$ can be modelled using the VAR in Eq. (11) or VECM in Eq. (14). After establishing the VAR or VECM, the data are restored to the corresponding values before the difference.

5: Model $\boldsymbol{Y}_i^{(q)}$ with VAR and forecast m periods forwards to obtain $\widehat{\boldsymbol{Y}}_i^{(q,m)}$. Go to step 7.

6: Model $\boldsymbol{Y}_i^{(q)}$ with VECM and forecast m periods forwards to obtain $\widehat{\boldsymbol{Y}}_i^{(q,m)}$. Go to step 7.

7: Conduct inverse transformation method on $\widehat{\boldsymbol{Y}}_i^{(q,m)}$ to obtain $\widehat{\boldsymbol{X}}_i^{(q,m)}$; evaluate the forecasting accuracy and end.

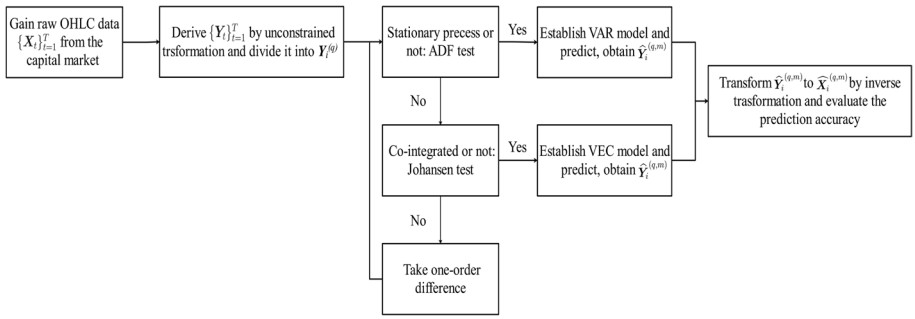


Fig. 3 An unified framework for modeling OHLC data based on VAR and VECM

By incorporating Algorithm 2 into Algorithm 1, we can obtain a unified framework for the statistical modeling of the OHLC series, as shown in Fig. 3.

Simulations

We assessed the performance of the proposed method using finite-sample simulations. We first describe the data construction in "Data construction" section, then provide five indicators to evaluate forecasting accuracy in "Measurements" section and finally report the simulation results in "Results of simulations" section.

Data construction

We generate the simulation data under the VAR structure in Eq. (11) as follows: (1) Assign the lag period p and the coefficient matrices $\boldsymbol{A}_1, \boldsymbol{A}_2, \cdots, \boldsymbol{A}_p$; (2) Generate an original four-dimensional vector $\boldsymbol{Y}_1 = [y_1^{(1)}, y_1^{(2)}, y_1^{(3)}, y_1^{(4)}]^T$; (3) Generate $\{\boldsymbol{Y}_t\}_{t=2}^T$ in a sequence via the VAR(p) model

$$\boldsymbol{Y}_t = \boldsymbol{A}_1 \boldsymbol{Y}_{t-1} + \cdots + \boldsymbol{A}_p \boldsymbol{Y}_{t-p} + \boldsymbol{w}_t,$$

where $\boldsymbol{w}_t$ follows the multivariate normal distribution with zero mean and covariance matrix $\Sigma_{\boldsymbol{w}}$. Finally, the simulated OHLC data $\{\boldsymbol{X}_t\}_{t=1}^T$ are generated by applying the inverse transformation formula in Eq. (9).

To evaluate the robustness of the proposed method, we considered the following scenarios with different variance component levels:

Scenario 1: $p = 1, T = 220, \boldsymbol{Y}_1 = [4,\ 0.7,\ -0.85,\ 0]^T$ and

$$\boldsymbol{A}_1 = \begin{pmatrix} 0.55 & 0.12 & 0.12 & 0.12 \\ 0.12 & 0.55 & 0.12 & 0.12 \\ 0.12 & 0.12 & 0.55 & 0.12 \\ 0.12 & 0.12 & 0.12 & 0.55 \end{pmatrix},$$

and $\Sigma_{\boldsymbol{w}}$ is a 4×4 diagonal matrix with diagonal element being 0.05^2, i.e.,

$$\Sigma_{\boldsymbol{w}} = \text{ diag } \{0.05^2, 0.05^2, 0.05^2, 0.05^2\}.$$

Scenario 2: $p, T, \boldsymbol{Y}_1$ and $\boldsymbol{A}_1$ follows Scenario 1 except that

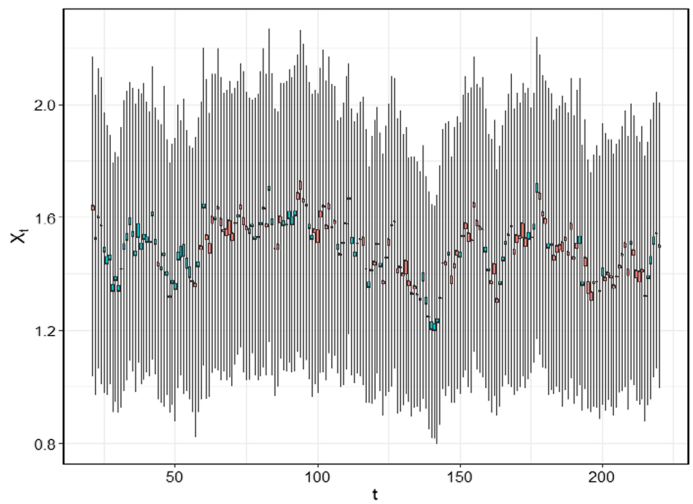


Fig. 4 Simulation OHLC data under Scenario 2

$$\Sigma_{\boldsymbol{w}} = \text{ diag } \{0.07^2, 0.07^2, 0.07^2, 0.07^2\}.$$

Scenario 3: $p, T, \boldsymbol{Y}_1$ and $\boldsymbol{A}_1$ follows Scenario 1 except that

$$\Sigma_{\boldsymbol{w}} = \text{ diag } \{0.03^2, 0.03^2, 0.03^2, 0.03^2\}.$$

All these scenarios present the transformed unconstrained time-series data $\{\boldsymbol{Y}_t\}_{t=1}^T$ that follow VAR(1), with different variance component levels according to the median (scenario 1), low (scenario 2), and strong (scenario 3) signal-to-noise ratios, respectively. A higher signal-to-noise ratio means that the information contained in the data comes more from the signal than from the noise, indicating a better quality of data. In contrast, a lower signal-to-noise ratio means that the noise carries more interference, indicating worse data quality.

Note that the raw simulation data have 220 periods; we only take 21–220 periods as the final simulated dataset, as the data generated initially may be highly volatile. Considering Scenario 2 as an illustration, Fig. 4 shows the simulated OHLC series $\{\boldsymbol{X}_t\}_{t=1}^T$.

Measurements

Based on the process illustrated in Fig. 3, the VAR and VECM are used to measure the statistical relationships between variables contained in $\boldsymbol{Y}_t$. As Corrado and Lee (1992) and Marshall et al. (2008) point out, short-term technical analyses can be more helpful for investors than long-term technical analyses. Therefore, we focused on a relatively short-term analysis. Specifically, q periods of the simulated data, namely the time period basement, were used to train the model and make out-of-sample forecasts ahead of m periods. We set q ranges from 30 to 70, and $m = 1, 2, 3$. For each setting (q, m), $\boldsymbol{Y}_i^{(q)}$ scrolls forward by one period and forecasts $(T - q - m + 1)$ times in total, as indicated in Fig. 5.

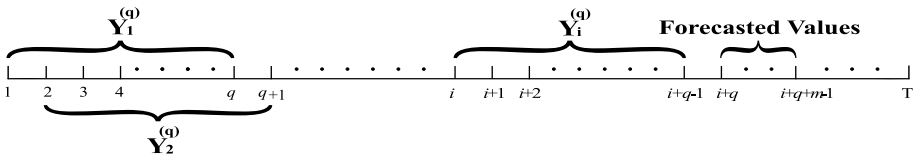


Fig. 5 Specific segment method

The forecasted $\widehat{Y_i}^{(q,m)}$ is first derived, and then the forecasted $\widehat{X_i}^{(q,m)}$ is obtained based on the inverse transformation formulas Eq. (9). We evaluated the effectiveness of forecasting in terms of five measurements, which are defined as follows:

- The mean absolute percentage error (MAPE)

$$\text{MAPE} = \frac{100\%}{k}\sum_{i=1}^{k}\left|\frac{x_i^{(*)} - \widehat{x}_i^{(*)}}{x_i^{(*)}}\right|,$$

  where $x_i^{(*)}$ and $\widehat{x}_i^{(*)}$ are the actual and forecasted values with $x_i^{(*)}$ indicating $x_i^{(o)}$, $x_i^{(h)}$, $x_i^{(l)}$, or $x_i^{(c)}$, respectively; k is the number of forecasted points.
- The standard deviation (SD) is defined as the empirical standard derivation of the forecasted values $\{\widehat{x}_i^{(*)}\}_{i=1}^{k}$, i.e.,

$$\text{SD} = \sqrt{\frac{1}{k-1}\sum_{i=1}^{k}\left(\widehat{x}_i^{(*)} - \bar{\widehat{x}}^{(*)}\right)^2},$$

  where $\bar{\widehat{x}}^{(*)} = \sum_{i=1}^{k}\widehat{x}_i^{(*)}/k$.
- The root mean squared error (RMSE) as defined in Neto and de Carvalho (2008)

$$\text{RMSE} = \sqrt{\frac{1}{k}\sum_{i=1}^{k}\left(x_i^{(*)} - \widehat{x}_i^{(*)}\right)^2}$$

- The RMSE based on the Hausdorff distance (RMSEH) is defined in De Carvalho et al. (2006)

$$\text{RMSEH} = \sqrt{\frac{1}{k}\sum_{i=1}^{k}\left(\left|\frac{x_i^{(h)} + x_i^{(l)}}{2} - \frac{\widehat{x}_i^{(h)} + \widehat{x}_i^{(l)}}{2}\right| + \left|\frac{x_i^{(h)} - x_i^{(l)}}{2} - \frac{\widehat{x}_i^{(h)} - \widehat{x}_i^{(l)}}{2}\right|\right)^2}$$

- The accuracy ratio (AR) is adopted in Hu and He (2007)

$$\text{AR} = \begin{cases} \frac{1}{k}\sum_{i=1}^{k}\frac{w(SP_i \cap \widehat{SP_i})}{w(SP_i \cup \widehat{SP_i})}, & \text{if} \quad (w(SP_i \cap \widehat{SP_i}) \neq 0) \\ 0, & \text{if} \quad (w(SP_i \cap \widehat{SP_i}) = 0) \end{cases},$$

  where $w(SP_i \cap \widehat{SP_i})$ and $w(SP_i \cup \widehat{SP_i})$ represent the length of the intersection and union between the observation interval $[x_i^{(l)}, x_i^{(h)}]$ and the forecasting interval $[\widehat{x}_i^{(l)}, \widehat{x}_i^{(h)}]$, respectively.

Smaller values of MAPE, RMSE, and RMSEH and a larger AR indicate a more accurate forecasting result, whereas a smaller SD indicates a more stable result.

Results of simulations

For Scenario 1, we summarize the results with $q = 40, 50, 70$ and $m = 1, 2, 3$ in Table 2. From this, we can observe that (1) the overall performance of the proposed method in terms of these five measurements is satisfactory and stable, and (2) for a fixed q, a smaller forecast period m makes the forecasted results more accurate with smaller values of MAPE, RMSE, and RMSEH and larger AR. There is no obvious pattern in terms of SD.

Moreover, we present additional results with q ranging from 30 to 70 and $m = 1, 2, 3$ to further demonstrate the performance of the proposed method. Specifically, Fig. 6 summarizes the results in terms of MAPE (the left panel), SD (the middle panel), and RMSE (the right panel), while Fig. 7 shows the RMSEH and AR of the forecasted results.

Basically from Fig. 6, under different q and m, the MAPE is between 3.08% and 7.13%, the SD is between 0.048 and 0.158, and the RMSE is between 0.051 and 0.148, indicating good forecasting accuracy and stability. As the forecast period m increases, these three indicators increase synchronously, decreasing forecasting accuracy. However, the forecasting performance shows a trend of getting better first and then getting worse as q increases. From Fig. 7, the RMSEH maintains a small value between 0.083 and 0.152. Meanwhile, AR is relatively high, varying from 0.842 to 0.903, which illustrates that the

Table 2 Simulation results for Scenario 1 when $q = 40, 50, 70$ and $m = 1, 2, 3$

Criterion	$q = 40$			$q = 50$			$q = 70$		
	$m = 1$	$m = 2$	$m = 3$	$m = 1$	$m = 2$	$m = 3$	$m = 1$	$m = 2$	$m = 3$
MAPE									
$x_t^{(o)}$	3.73%	4.95%	5.41%	4.19%	4.82%	5.91%	3.58%	4.67%	5.80%
$x_t^{(h)}$	3.75%	4.71%	4.92%	3.93%	4.65%	5.59%	3.43%	4.37%	5.23%
$x_t^{(l)}$	4.30%	5.28%	6.28%	4.80%	5.17%	6.21%	4.60%	5.64%	6.81%
$x_t^{(c)}$	3.69%	4.94%	5.53%	4.31%	4.87%	5.96%	3.58%	4.67%	5.70%
SD									
$x_t^{(o)}$	0.115	0.122	0.115	0.127	0.127	0.122	0.104	0.105	0.104
$x_t^{(h)}$	0.126	0.138	0.130	0.150	0.147	0.143	0.123	0.123	0.122
$x_t^{(l)}$	0.073	0.079	0.076	0.072	0.074	0.072	0.070	0.072	0.072
$x_t^{(c)}$	0.112	0.123	0.116	0.131	0.129	0.126	0.103	0.104	0.104
RMSE									
$x_t^{(o)}$	0.073	0.096	0.108	0.081	0.093	0.110	0.067	0.088	0.106
$x_t^{(h)}$	0.094	0.123	0.134	0.101	0.118	0.139	0.084	0.110	0.131
$x_t^{(l)}$	0.052	0.067	0.075	0.059	0.065	0.076	0.055	0.068	0.081
$x_t^{(c)}$	0.071	0.099	0.109	0.083	0.096	0.114	0.066	0.088	0.106
RMSEH	0.098	0.127	0.137	0.099	0.122	0.142	0.088	0.114	0.135
AR	0.891	0.868	0.858	0.886	0.872	0.849	0.895	0.871	0.848

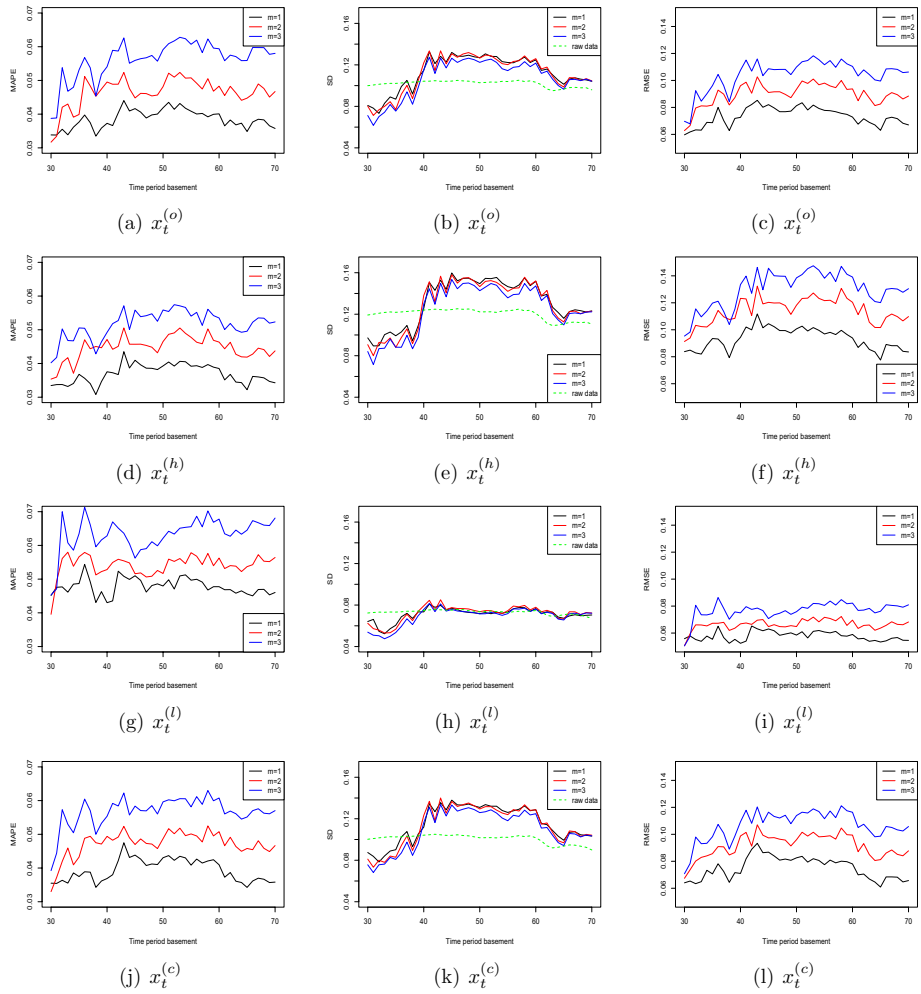


Fig. 6 MAPE (Left panel), SD (Middle panel) and RMSE (Right panel) of forecasted values for $x_t^{(o)}$ (The first row), $x_t^{(h)}$ (The second row), $x_t^{(l)}$ (The third row) and $x_t^{(c)}$ (The fourth row) with different q and m for Scenario 1 respectively

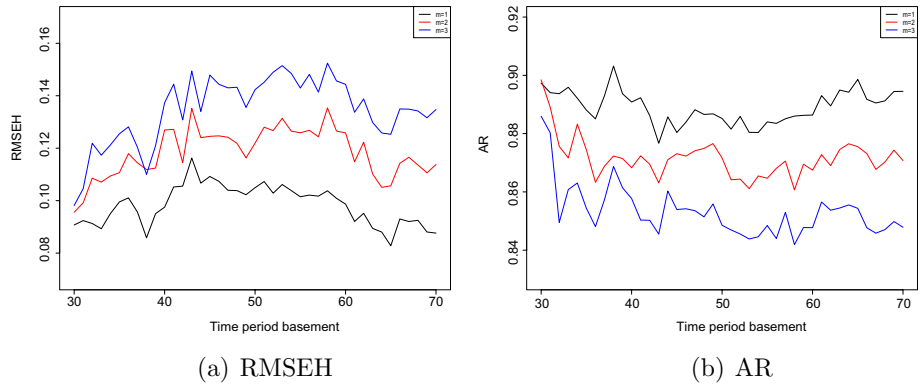
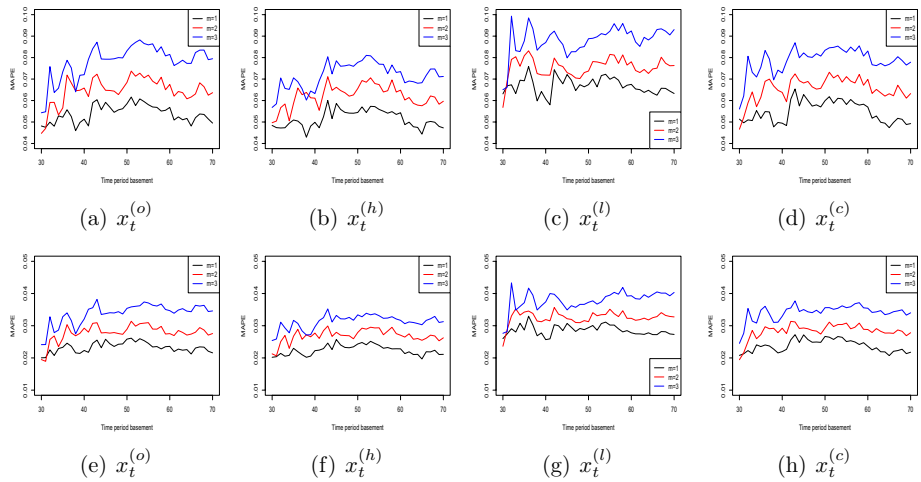


Fig. 7 RMSEH (Left) and AR (Right) of forecasted values for different q and m in Scenario 1



(a) $x_t^{(o)}$ (b) $x_t^{(h)}$ (c) $x_t^{(l)}$ (d) $x_t^{(c)}$

(e) $x_t^{(o)}$ (f) $x_t^{(h)}$ (g) $x_t^{(l)}$ (h) $x_t^{(c)}$

Fig. 8 MAPE of forecasted values with different q and m for Scenario 2 (The first row) and 3 (The second row), respectively

forecasting interval closely coincides with the observation interval, indicating a satisfactory forecasting effect.

For Scenarios 2 and 3, we conducted simulations in line with Scenario 1, and the results exhibited the same trend. Owing to space constraints, we only show the forecasted results in terms of MAPE in Fig. 8 for Scenario 2 with low signal-to-noise ratio (the first row) and Scenario 3 with high signal-to-noise ratio (the second row). The MAPE values in the first row of Fig. 8 is between 4.29 and 9.93%, while the corresponding MAPE in the second row of Fig. 8 is between 1.89 and 4.33%. The left panel of Fig. 6 corresponds to the MAPE with medium signal-to-noise ratio, whose MAPE (ranges from 3.08 to 7.13%) is between those in the second and the first row of Fig. 8, which indicates that the accuracy of the forecasted results decreases with the signal-to-noise ratio.

Empirical analysis

We illustrate the practical utility of the proposed method using three different types of real datasets: the OHLC series of Kweichow Moutai, CSI 100 index, and 50 ETF in the Chinese financial market. For each case, we first briefly describe the dataset in "Raw OHLC dataset description" section, then apply the proposed method with different forecasting basement period q and forecasting period m and report the performance in terms of MAPE, RMSE, RMSEH, and AR in in "Results of empirical analysis" section.

Raw OHLC dataset description

- *OHLC series of the Kweichow Moutai* The Kweichow Moutai is a well-known company in Chinese liquor industry, which has a long history, and its stamp (SH: 600519) is an important part of the China Securities Index (CSI 100). Here, we study its OHLC series with the time ranging from 27/8/2001 to 14/6/2019, yielding 4243 data in total.

- *OHLC series of the CSI 100 index* The CSI 100 index is one of the most important stock price indexes in China, reflecting the overall situation of the companies with the most market influence power in the Shanghai and Shenzhen stock markets. China Securities Index Co., Ltd. officially issued the CSI 100 index on 30/12/2005, and 1000 is its base data and base point. We collected the OHLC series of the CSI 100 index from 30/12/2005 to 14/6/2019 with a total of 3269 periods.
- *OHLC series of the 50 ETF* The 50 ETF (code: 510050) is China's first transactional exchange traded fund, compiled by the Shanghai Stock Exchange, whose base date and base point are 31/12/2003 and 1000, respectively. The investment objective of the 50 ETF is to closely track the Shanghai Stock Exchange 50 (SSE 50) index, minimizing tracking deviation and tracking error. This study collected 3481 OHLC data samples of the 50 ETF from 23/2/2005 to 14/6/2019.

Results of empirical analysis

Based on the results of the simulation experiments, we used the proposed method, as shown in Fig. 3 with q varying from 30 to 70 and $m = 1$, to realize the forecast of OHLC series of the Kweichow Moutai, CSI 100, and 50 ETF.

Specifically, we consider the first vector time series of $\{\boldsymbol{Y}_t\}_{t=1}^{90}$ (i.e., $\boldsymbol{Y}_1^{(90)}$ in Algorithm 2) of the 50 ETF as an example to illustrate our modeling process. At the significance level of $\alpha = 0.05$, the four time series are stationary except for $\{\boldsymbol{y}_t^{(1)}\}_{t=1}^{90}$ with the p-value of ADF test being 0.628, indicating that $\{\boldsymbol{Y}_t\}_{t=1}^{90}$ cannot be modelled by VAR. The ACF plots in Fig. 9 further demonstrate the distinct auto-correlation and non-stationary of $\{\boldsymbol{y}_t^{(1)}\}_{t=1}^{90}$. Then, the Johansen test is applied to examine the cointegration relationship between the four variables in $\{\boldsymbol{Y}_t\}_{t=1}^{90}$, and the essence of Johansen test

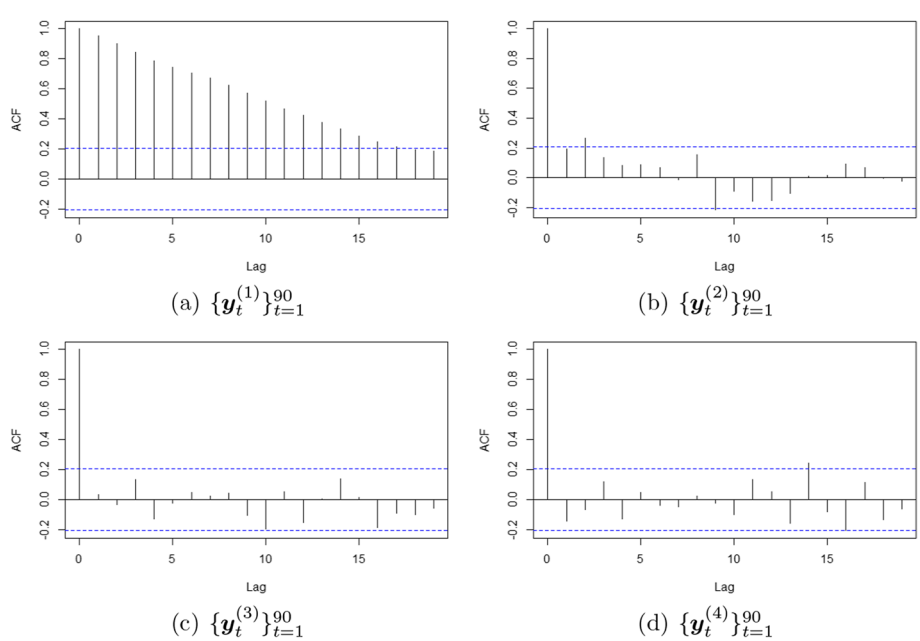


Fig. 9 ACF plots of the four time-series variables in $\boldsymbol{Y}_1^{(90)}$ of the 50 ETF

based on "Trace" investigates the number of cointegration vectors, which is recorded as r. The results show that the possibility of $r \leq 2$ is less than 1% and the possibility of $r \leq 3$ is greater than 10%; thus, r in VECM is determined to be 3. Finally, a VECM with order of cointegration $r = 3$ is established, and the one-step forecasting value $\widehat{\boldsymbol{Y}_1}^{(90,1)}$ is obtained using the regression function. Using the inverse transformation method, we obtain forecasted $\widehat{\boldsymbol{X}_1}^{(90,1)}$. Iterating through each vector time series $\{\boldsymbol{Y}_{t+l}\}_{t=1}^{90}$ $(l = 0, 1, \ldots, T - 91)$, the forecasting accuracy of the entire data samples can be evaluated.

The forecasting accuracy of the VAR and VECM models with unconstrained transformations proposed in this study is satisfactory with q varying from 30 to 70. (1) The average MAPE of Kweichow Moutai was between 1.247% and 1.369%, RMSE was between 32.759 and 38.046, RMSEH was between 39.842 and 45.818, and the AR was between 0.418 and 0.454. (2) The average MAPE of the 100 index was between 1.000% and 1.070%, RMSE was between 47.387 and 51.594, RMSEH was between 55.389 and 60.486, and the AR was between 0.395 and 0.429. (3) The average MAPE of the 50 ETF was between 0.981 and 1.114%, RMSE was between 0.039 and 0.045, RMSEH was between 0.046 and 0.055, and AR was between 0.403 and 0.442.

Furthermore, we take $q = 30$, $q = 50$, and $q = 70$ as three milestones and summarize the forecasting results in terms of MAPE, RMSE, RMSEH, AR, the ratios of VAR and VECM, and the numbers of the three types of forecasting failures. The forecasting results for the Kweichow Moutai, CSI 100, and CSI 50 ETF are presented in Tables 3,

Table 3 Results of the VAR and VECM for OHLC data of Kweichow Moutai

Criterion	$q = 30$			$q = 50$			$q = 70$		
	Naive	**Yes**	**No**	**Naive**	**Yes**	**No**	**Naive**	**Yes**	**No**
MAPE									
$x_t^{(o)}$	1.602%	1.002%	0.759%	1.602%	0.940%	0.677%	1.600%	0.831%	0.617%
$x_t^{(h)}$	1.378%	1.418%	1.348%	1.376%	1.353%	1.300%	1.377%	1.297%	1.267%
$x_t^{(l)}$	1.378%	1.269%	1.315%	1.377%	1.240%	1.207%	1.381%	1.191%	1.147%
$x_t^{(c)}$	1.486%	1.765%	1.773%	1.485%	1.709%	1.689%	1.488%	1.667%	1.626%
$\bar{x}_t$	1.461%	1.363%	1.299%	1.460%	1.310%	1.218%	1.461%	1.247%	1.164%
RMSE									
$x_t^{(o)}$	42.207	29.651	26.959	42.307	24.220	22.981	42.408	22.328	21.309
$x_t^{(h)}$	36.205	37.646	40.406	36.291	35.714	36.262	36.378	33.639	34.585
$x_t^{(l)}$	36.049	34.708	37.354	36.135	31.826	33.037	36.221	31.121	30.930
$x_t^{(c)}$	40.772	47.770	50.991	40.869	45.643	46.073	40.967	44.628	44.452
$\bar{x}_t$	38.808	37.443	38.928	38.900	34.351	34.588	38.994	32.929	32.819
RMSEH	44.329	44.639	47.777	44.435	42.153	42.657	44.541	39.988	40.560
AR	0.411	0.419	0.412	0.411	0.442	0.441	0.411	0.454	0.459
VAR Ratio	-	8.40%	5.53%	-	12.96%	5.98%	-	1.37%	2.35%
VECM Ratio	-	91.60%	94.47%	-	87.04%	94.02%	-	98.63%	97.65%
Failure 1	0	0	0	0	0	0	0	0	0
Failure 2	0	0	9	0	0	9	0	0	7
Failure 3	0	0	434	0	0	452	0	0	144

Table 4 Results of the VAR and VECM for OHLC data of CSI 100

Criterion	$q = 30$			$q = 50$			$q = 70$		
	Naive	**Yes**	**No**	**Naive**	**Yes**	**No**	**Naive**	**Yes**	**No**
MAPE									
$x_t^{(o)}$	1.310%	0.796%	0.641%	1.314%	0.728%	0.562%	1.317%	0.693%	0.512%
$x_t^{(h)}$	1.054%	0.996%	0.973%	1.057%	0.961%	0.919%	1.060%	0.937%	0.887%
$x_t^{(l)}$	1.177%	1.054%	1.061%	1.179%	1.003%	0.975%	1.182%	1.001%	0.935%
$x_t^{(c)}$	1.223%	1.436%	1.452%	1.227%	1.395%	1.357%	1.229%	1.378%	1.322%
$\bar{x}_t$	1.191%	1.070%	1.032%	1.194%	1.022%	0.953%	1.197%	1.002%	0.914%
RMSE									
$x_t^{(o)}$	61.400	40.557	35.927	61.586	36.477	31.117	61.770	34.748	28.773
$x_t^{(h)}$	48.781	46.347	45.321	48.928	43.774	42.415	49.076	43.027	40.248
$x_t^{(l)}$	55.653	51.616	51.660	55.820	47.804	47.266	55.988	47.685	45.184
$x_t^{(c)}$	57.773	67.857	68.388	57.947	64.767	64.498	58.122	64.260	62.213
$\bar{x}_t$	55.902	51.594	50.324	56.070	48.198	46.324	56.239	47.430	44.104
RMSEH	64.021	60.486	60.158	64.212	55.994	55.272	64.406	55.719	52.732
AR	0.370	0.395	0.403	0.370	0.417	0.431	0.371	0.428	0.448
VAR Ratio	-	6.80%	5.94%	-	10.90%	3.61%	-	0.88%	2.63%
VECM Ratio	-	93.20%	94.06%	-	89.10%	96.39%	-	99.12%	97.37%
Failure 1	0	0	0	0	0	0	0	0	0
Failure 2	0	0	7	0	0	3	0	0	1
Failure 3	0	0	379	0	0	285	0	0	106

Table 5 Results of the VAR and VECM for OHLC data of 50 ETF

Criterion	$q = 30$			$q = 50$			$q = 70$		
	Naive	**Yes**	**No**	**Naive**	**Yes**	**No**	**Naive**	**Yes**	**No**
MAPE									
$x_t^{(o)}$	1.259%	0.773%	0.647%	1.261%	0.670%	0.539%	1.263%	0.645%	0.496%
$x_t^{(h)}$	1.051%	1.072%	1.066%	1.053%	1.004%	0.984%	1.054%	0.975%	0.943%
$x_t^{(l)}$	1.117%	1.090%	1.054%	1.120%	0.980%	0.952%	1.120%	0.952%	0.911%
$x_t^{(c)}$	1.208%	1.482%	1.496%	1.210%	1.376%	1.349%	1.211%	1.357%	1.306%
$\bar{x}_t$	1.159%	1.104%	1.066%	1.161%	1.015%	0.956%	1.162%	0.982%	0.914%
RMSE									
$x_t^{(o)}$	0.048	0.035	0.031	0.048	0.031	0.025	0.048	0.028	0.024
$x_t^{(h)}$	0.040	0.044	0.041	0.040	0.039	0.037	0.040	0.037	0.035
$x_t^{(l)}$	0.044	0.045	0.044	0.044	0.039	0.038	0.044	0.038	0.037
$x_t^{(c)}$	0.047	0.058	0.058	0.047	0.053	0.052	0.047	0.052	0.051
$\bar{x}_t$	0.045	0.045	0.043	0.045	0.040	0.038	0.045	0.039	0.037
RMSEH	0.052	0.055	0.053	0.052	0.049	0.047	0.053	0.047	0.045
AR	0.387	0.403	0.399	0.387	0.427	0.429	0.387	0.440	0.449
VAR Ratio	-	6.56%	6.51%	-	6.31%	3.59%	-	1.56%	2.62%
VECM Ratio	-	93.44%	93.49%	-	93.69%	96.41%	-	98.44%	97.38%
Failure 1	0	0	0	0	0	0	0	0	0
Failure 2	0	0	9	0	0	2	0	0	0
Failure 3	0	0	380	0	0	271	0	0	123

4, and 5, respectively. Failure 1 refers to the forecasted low price becoming negative; that is, $\hat{x}_t^{(l)} < 0$. Failure 2 indicates that the forecasted high price is lower than the forecasted low price, that is, $\hat{x}_t^{(h)} < \hat{x}_t^{(l)}$. Failure 3 is for the forecasted open price (or forecasted close price) to break through the forecasted high-price and forecasted low-price boundaries, that is, $\hat{x}_t^{(o)}, \hat{x}_t^{(c)} \notin [\hat{x}_t^{(l)}, \hat{x}_t^{(h)}]$.

Meanwhile, we compare the VAR and VECM with an unconstrained method (marked as "Yes" in Tables 3–5) with two other methods: (1) The Naive method proposed by Arroyo et al. (2011), which takes the price of the previous day as the price of the day. (2) The VAR and VECM under non-unconstrained method, which employ the raw OHLC data as input (marked as "No" in Tables 3–5). The results demonstrate the following patterns:

(1) With the increase of q, the forecasting results of the VAR and VECM under unconstrained and non-unconstrained methods become more accurate; MAPE, RMSE, and RMSEH decrease, and AR increases.
(2) Regarding MAPE and RMSE, the VAR and VECM under unconstrained and non-unconstrained methods possess better forecasting accuracy for $x_t^{(o)}$, $x_t^{(h)}$, and $x_t^{(l)}$ than the Naive method while the Naive method has better forecasting accuracy for $x_t^{(c)}$ than the VAR and VECM. As for RMSEH and AR, the VAR and VECM under unconstrained and non-unconstrained methods are superior to the Naive method.
(3) The proportion of utilizing the VECM model is significantly greater than that of the VAR model, which indicates that the Chinese stock market has the same characteristics as the U.S. stock market. That is, stock prices are usually non-stationary, and a cointegration relationship exists between the quaternary OHLC prices or unconstrained variables (Cheung 2007).
(4) The modeling of the VAR and VECM under the non-unconstrained method results in a large number of forecasting failures, while the VAR and VECM under the unconstrained method can always guarantee a meaningful forecast of OHLC data. The forecasting failures of the non-unconstrained method are mostly $\hat{x}_t^{(o)}, \hat{x}_t^{(c)} \notin [\hat{x}_t^{(l)}, \hat{x}_t^{(h)}]$. This is because it is unlikely that $\hat{x}_t^{(l)} < 0$ or $\hat{x}_t^{(l)} > \hat{x}_t^{(h)}$ will occur under accurate forecasting conditions.
(5) For Kweichow Moutai, the average MAPE, RMSE, RMSEH, and AR of the VAR and VECM under unconstrained method are 10.543%, 10.23%, 4.89%, and 0.71% better than the Naive method, respectively. For CSI 100, the average MAPE, RMSE, RMSEH, and AR of the VAR and VECM under the unconstrained method are improved by 13.62%, 12.48%, 10.61%, and 1.02% compared to the Naive method, respectively. For 50 ETF, the average MAPE, RMSE, RMSEH, and AR of the VAR and VECM under the unconstrained method are 10.94%, 8.15%, 3.82%, and 1.16% more optimized than in the Naive method, respectively.
(6) The overall forecasting performance of the VAR and VECM under the non-unconstrained method is better than the VAR and VECM under the unconstrained method. However, the forecasting failures of the VAR and VECM under the non-unconstrained method can cause confusion for investors and significantly undermine their investment confidence (Huang et al. 2022a). The results are consistent

with our original intention to ensure the integrity of the forecasted OHLC data structure under the possibility of losing a certain forecast accuracy.

To obtain a clear depiction of the forecasted performance, we also compare the actual and forecasted stock values of the Kweichow Moutai from November 5, 2003, to June 22, 2004 (left panel); the CSI 100 index from May 11, 2011, to December 16, 2011 (middle panel); and the 50 ETF from August 9, 2007, to March 24, 2008 (right panel) in Fig. 10. The data forecasted by the VAR and VECM combined with the unconstrained transformation are in line with reality. Specifically, for the Kweichow Moutai, the continuous rise that exists before April 9, 2004, and the subsequent fall are perfectly forecasted; for the CSI 100 index, the overall downward trend and two rebounds around July 4, 2011, and November 9, 2011, are also fully reflected; and, for the 50 ETF, two spikes around October 16, 2007, and January 15, 2008, coincide precisely.

This study was complemented by a machine-learning approach for modeling OHLC data using SVR. The selected SVR employs a linear kernel function with a constant of the regularization term in the Lagrange formulation set to 1 and epsilon in the insensitive-loss function set to 0.1. SVR performs out-of-sample forecasting using the first 80% of the data as the training set and the last 20% as the testing set. Table 6 demonstrates the forecasting accuracy of SVR. Several patterns can be found. (1) The overall forecasting accuracy of SVR is significantly improved compared with that of VAR and VECM modeling. Under the unconstrained transformation method, the MAPEs obtained by SVR on the Kweichou Moutai, CSI 100, and CSI 50 ETF datasets are 26.30%, 33.83%, and 15.48% lower compared to those obtained by VAR and VECM modeling, respectively. In particular, the accuracy of SVR in forecasting close prices improved significantly with the close price MAPEs for the three datasets decreasing by 74.33%, 71.84%, and 64.11%, respectively, compared to those obtained from VAR and VECM modeling. At the same time, the high- and low-price MAPEs obtained by SVR are, on average, 32.36%

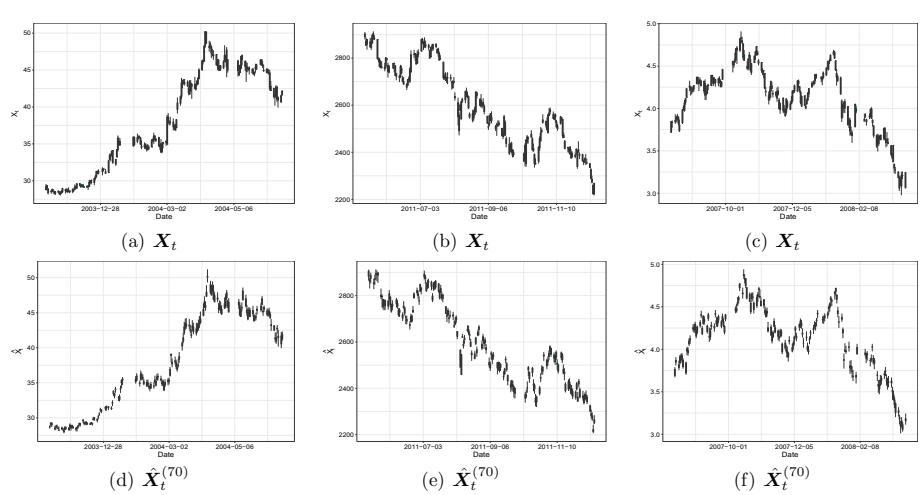


Fig. 10 Comparison of the real values (top row) and forecasted values (bottom row) of the Kweichow Moutai (left panel), CSI 100 index (middle panel), and 50 ETF (right panel)

Table 6 Results of the SVR for OHLC data of Kweichou Moutai, CSI 100, and 50 ETF

Criterion	Kweichou Moutai		CSI 100		50 ETF	
	Yes	No	Yes	No	Yes	No
MAPE						
$x_t^{(o)}$	1.465%	0.482%	0.976%	0.342%	1.073%	0.278%
$x_t^{(h)}$	0.833%	1.040%	0.575%	0.629%	0.754%	0.625%
$x_t^{(l)}$	0.948%	0.937%	0.715%	0.600%	1.008%	0.571%
$x_t^{(c)}$	0.428%	1.430%	0.388%	0.830%	0.487%	0.813%
$\bar{x}_t$	0.919%	0.972%	0.663%	0.600%	0.830%	0.572%
RMSE						
$x_t^{(o)}$	77.143	39.891	46.055	21.048	0.043	0.016
$x_t^{(h)}$	77.143	39.891	46.055	21.048	0.043	0.016
$x_t^{(l)}$	50.224	59.633	33.466	32.082	0.039	0.028
$x_t^{(c)}$	26.451	85.870	18.087	43.486	0.020	0.039
$\bar{x}_t$	53.374	65.544	32.904	33.188	0.034	0.029
RMSEH	56.652	80.311	35.904	38.711	0.043	0.035
AR	0.503	0.461	0.421	0.461	0.361	0.471
Failure 1	0	0	0	0	0	0
Failure 2	0	0	0	0	0	0
Failure 3	0	0	0	1	0	0

and 14.36% lower than those obtained by VAR and VECM modeling, respectively. The forecasting accuracy of the open price by VAR and VECM is better than that of SVR modeling with an average decrease of 37.39% in the three datasets. (2) The overall forecasting accuracy of the Kweichou Moutai under the unconstrained transformation method is higher than that of the non-unconstrained method, while it performs slightly worse on CSI 100 and 50 ETF. Interestingly, the SVR under the unconstrained method has significantly better forecasting accuracy for the close price than the SVR under the non-unconstrained method for all three datasets of the Kweichou Moutai, CSI 100, and CSI 50 ETF with MAPE reduced by 70.070%, 53.25%, and 40.10% and RMSE reduced by 69.20%, 58.41%, and 48.72%, respectively. The unconstrained transformation not only ensures the structural forecasting of OHLC data but also is a manual feature-extraction method that can effectively enhance the forecasting accuracy of machine-learning models for close prices. Given the importance of the close price in various trading strategies, this study demonstrates the significance of the proposed unconstrained method. (3) With high forecasting accuracy, the three types of forecasting failures rarely occur. The SVR under the non-unconstrained method produced a forecasting failure in the CSI 100 dataset.

Conclusions

To solve the structural modeling issues of the OHLC data contained in the candlestick chart, we proposed a novel unconstrained transformation method to relax the inherent constraints of OHLC data along with its explicit inverse transformation. The proposed methodology facilitates the subsequent establishment of various forecasting models and

guarantees meaningful, structurally forecasted OHLC prices. The unconstrained transformation method not only extends the range of modeling variables to $(-\infty, +\infty)$ but also shares the flexibility of the well-known log and logit transformations. Based on this unconstrained transformation, we established a flexible and efficient framework for modeling OHLC data with full utilization of the information. As an application of the multivariate time series, we demonstrated a detailed modeling procedure using VAR and VECM.

The proposed unconstrained transformation has high practical utility owing to its flexibility, simple implementation, and straightforward interpretation. For instance, it is applicable to various positive interval data with internal variables, and the selected model can be generalized to other econometric or machine-learning models. From this perspective, the proposed method provides a novel and useful alternative for OHLC data analysis, thereby enriching existing literature on the structural modeling of complex data.

We documented the finite performance of the OHLC data modeling process via extensive simulation studies on various measurements. The simulation experiments demonstrated that the VAR and VECM under unconstrained transformation obtained stable and satisfactory results with different forecast periods, time period basements, and signal-to-noise ratios, verifying the effectiveness and robustness of the proposed modeling approach. The analysis of OHLC data for three different types of financial products in the Chinese financial market-the Kweichow Moutai, CSI 100 index, and 50 ETF-also illustrated the utility of the unconstrained method. Using raw OHLC data directly as input to the VAR and VECM resulted in a large number of forecasting failures. In contrast, unconstrained and inverse transformations guaranteed structural modeling of OHLC data at the cost of a small loss in forecasting accuracy.

As a complement to machine learning, this study further employed SVR for modeling OHLC data in the empirical analysis section. SVR modeling demonstrated superior performance in forecasting OHLC data. Under unconstrained transformation, SVR can achieve higher forecast accuracy than VAR and VECM while ensuring OHLC data structure. In addition, the SVR under the unconstrained method had significantly better forecasting accuracy for the close price than the SVR under the non-unconstrained method in all three datasets of the Kweichou Moutai, CSI 100, and CSI 50 ETF. The proposed unconstrained method proved to be an effective pre-processing technique for machine learning models. These results provide new evidence for the practical utility and extensibility of the unconstrained method.

Future research can embed various time-series forecasting models into the proposed unified forecasting modeling framework for OHLC time-series data based on unconstrained transformation and its inverse transformation to achieve the structural forecasting of OHLC data for various financial products. In particular, artificial neural networks can be employed to accurately forecast OHLC data. This can help investors manage and hedge their portfolios to earn profits and reduce risks (Huang et al. 2022a). Specifically, based on OHLC data forecasting results, investors can achieve overnight returns (Cooper et al. 2008), reduce fund-holding periods (Dunis et al. 2011), lower overnight exposure (Kelly and Clark 2011), and derive better profits from high-low price range trades (von Mettenheim and Breitner 2012).

Abbreviations

OHLC	Open-high-low-close
EMH	Efficient market hypothesis
VAR	Vector autoregression
VECM	Vector error correction model
SVR	Support vector regression
COVID-19	Coronavirus disease 2019
MAPE	Mean absolute percentage error
SD	Standard deviation
RMSE	Root mean squared error
RMSEH	RMSE base on the Hausdorff distance
AR	Accuracy ratio

Acknowledgements
We are grateful for the grants and would like to express our sincere gratitude to the reviewers who provided suggestions to our article.

Author contributions
WH: Methodology, software, formal analysis, writing—original draft. HW: Conceptualization, methodology, funding acquisition. SW: Validation, methodology, writing—review and editing.

Funding
The authors are grateful for the financial support from the Beijing Natural Science Foundation (Grant No. 9244030) and the National Natural Science Foundation of China (Grant Nos. 72021001, 11701023).

Availability of data and materials
Available on request.

Declarations

Competing interests
The authors have declared that no competing interests exist.

Received: 29 September 2022 Accepted: 18 January 2024
Published online: 05 March 2024

References

Ahmadi E, Jasemi M, Monplaisir L, Nabavi MA, Mahmoodi A, Jam PA (2018) New efficient hybrid candlestick technical analysis model for stock market timing on the basis of the support vector machine and heuristic algorithms of imperialist competition and genetic. Expert Syst Appl 94:21–31
Ariss RT, Rezvanian R, Mehdian SM (2011) Calendar anomalies in the gulf cooperation council stock markets. Emerg Mark Rev 12(3):293–307
Arroyo J, Espínola R, Maté C (2011) Different approaches to forecast interval time series: a comparison in finance. Comput Econ 37(2):169–191
Caginalp G, Laurent H (1998) The predictive power of price patterns. Appl Math Finance 5(3–4):181–205
Cagliero L, Fior J, Garza P (2023) Shortlisting machine learning-based stock trading recommendations using candlestick pattern recognition. Expert Syst Appl 216:119493
Chang P-C, Liao TW, Lin J-J, Fan C-Y (2011) A dynamic threshold decision system for stock trading signal detection. Appl Soft Comput 11(5):3998–4010
Chen Y, Hao Y (2020) A novel framework for stock trading signals forecasting. Soft Comput 24(16):12111–12130
Chen J, Wen Y, Nanehkaran YA, Suzauddola MD, Chen W, Zhang D (2023) Machine learning techniques for stock price prediction and graphic signal recognition. Eng Appl Artif Intell 121:106038
Cheung Y-W (2007) An empirical model of daily highs and lows. Int J Finance Econ 12(1):1–20
Cooper MJ, Cliff MT, Gulen H (2008) Return differences between trading and non-trading hours: like night and day. Available at SSRN 1004081
Corrado CJ, Lee S-H (1992) Filter rule tests of the economic significance of serial dependencies in daily stock returns. J Financ Res 15(4):369–387
Cuthbertson K, Hall SG, Taylor MP (1992) Applied econometric techniques. P. Allan
De Carvalho FAT, de Souza RMCR, Chavent M, Lechevallier Y (2006) Adaptive Hausdorff distances and dynamic clustering of symbolic interval data. Pattern Recognit Lett 27(3):167–179
Doyle JR, Chen CH (2009) The wandering weekday effect in major stock markets. J Bank Finance 33(8):1388–1399
Dunis CL, Laws J, Rudy J (2011) Profitable mean reversion after large price drops: a story of day and night in the S &P 500, 400 midcap and 600 smallcap indices. J Asset Manag 12:185–202
Fama EF (1970) Efficient capital markets: a review of theory and empirical work. J Finance 25:383–417
Fiess NM, MacDonald R (2002) Towards the fundamentals of technical analysis: analysing the information content of high, low and close prices. Econ Model 19(3):353–374
García A, Jaramillo-Morán MA (2020) Short-term European union allowance price forecasting with artificial neural networks. Entrep Sustain Issues 8(1):261

García-Martos C, Rodríguez J, Sánchez MJ (2013) Modelling and forecasting fossil fuels, CO2 and electricity prices and their volatilities. Appl Energy 101:363–375
González-Rivera G, Lin W (2013) Constrained regression for interval-valued data. J Bus Econ Stat 31(4):473–490
Goo Y, Chen D, Chang Y et al (2007) The application of Japanese candlestick trading strategies in Taiwan. Invest Manag Financ Innov 4(4):49–79
Guo J, Li W, Li C, Gao S (2012) Standardization of interval symbolic data based on the empirical descriptive statistics. Comput Stat Data Anal 56(3):602–610
Hao P, Guo J (2017) Constrained center and range joint model for interval-valued symbolic data regression. Comput Stat Data Anal 116:106–138
Hsu P-H, Kuan C-M (2005) Reexamining the profitability of technical analysis with data snooping checks. J Financ Econom 3(4):606–628
Hu C, He LT (2007) An application of interval methods to stock market forecasting. Reliab Comput 13(5):423–434
Huang W, Wang H, Qin H, Wei Y, Chevallier J (2022a) Convolutional neural network forecasting of European union allowances futures using a novel unconstrained transformation method. Energy Econ 110:106049
Huang W, Wang H, Wang S (2022b) A pseudo principal component analysis method for multi-dimensional open-high-low-close data in candlestick chart. Commun Stat Theory Methods. https://doi.org/10.1080/03610926.2022.2155787
Ilham RN, Sinta I, Sinurat M (2022) The effect of technical analysis on cryptocurrency investment returns with the 5 (five) highest market capitalizations in Indonesia. J Ekon 11(02):1022–1035
Johansen S (1988) Statistical analysis of cointegration vectors. J Econ Dyn Control 12(2–3):231–254
Johansen S (1991) Estimation and hypothesis testing of cointegration vectors in Gaussian vector autoregressive models. Econom J Econom Soc 59:1551–1580
Johansen S (1995) Likelihood-based inference in cointegrated vector autoregressive models. Oxford University Press, Oxford
Kelly MA, Clark SP (2011) Returns in trading versus non-trading hours: the difference is day and night. J Asset Manag 12:132–145
Kumar G, Sharma V (2019) Stock market index forecasting of nifty 50 using machine learning techniques with ANN approach. Int J Mod Comput Sci (IJMCS) 4(3):22–27
Lai KS, Lai M (1991) A cointegration test for market efficiency. J Futures Mark 11(5):567–575
Lan Q, Zhang D, Xiong L (2011) Reversal pattern discovery in financial time series based on fuzzy candlestick lines. Syst Eng Procedia 2:182–190
Levy T, Yagil J (2012) The week-of-the-year effect: evidence from around the globe. J Bank Finance 36(7):1963–1974
Liu H, Shen L (2020) Forecasting carbon price using empirical wavelet transform and gated recurrent unit neural network. Carbon Manag 11(1):25–37
Liu F, Wang J (2012) Fluctuation prediction of stock market index by Legendre neural network with random time strength function. Neurocomputing 83:12–21
Lu TH, Shiu Y-M, Liu T-C (2012) Profitable candlestick trading strategies-the evidence from a new perspective. Rev Financ Econ 21(2):63–68
Luo L, Chen X (2013) Integrating piecewise linear representation and weighted support vector machine for stock trading signal prediction. Appl Soft Comput 13(2):806–816
Lütkepohl H (2005) New introduction to multiple time series analysis. Springer, Berlin
Magdon-Ismail M, Atiya AF (2003) A maximum likelihood approach to volatility estimation for a Brownian motion using high, low and close price data. Quant Finance 3(5):376
Mager J, Paasche U, Sick B (2009) Forecasting financial time series with support vector machines based on dynamic kernels. In: IEEE conference on soft computing in industrial applications
Mahmoodi A, Hashemi L, Jasemi M, Laliberté J, Millar RC, Noshadi H (2023a) A novel approach for candlestick technical analysis using a combination of the support vector machine and particle swarm optimization. Asian J Econ Bank 7(1):2–24
Mahmoodi A, Hashemi L, Jasemi M, Mehraban S, Laliberté J, Millar RC (2023b) A developed stock price forecasting model using support vector machine combined with metaheuristic algorithms. OPSEARCH 60(1):59–86
Mahmoudi A, Hashemi L, Jasemi M, Pope J (2021) A comparison on particle swarm optimization and genetic algorithm performances in deriving the efficient frontier of stocks portfolios based on a mean-lower partial moment model. Int J Finance Econ 26(4):5659–5665
Manurung AH, Budiharto W, Prabowo H (2018) Algorithm and modeling of stock prices forecasting based on long short-term memory (LSTM). Int J Innov Comput Inf Control (ICIC) 12:12
Marshall BR, Young MR, Rose LC (2006) Candlestick technical trading strategies: Can they create value for investors? J Bank Finance 30(8):2303–2323
Marshall BR, Young MR, Cahan R (2008) Are candlestick technical trading strategies profitable in the Japanese equity market? Rev Quant Finance Account 31(2):191–207
Mehrjoo S, Jasemi M, Mahmoudi A (2014) A new methodology for deriving the efficient frontier of stocks portfolios: an advanced risk-return model. J AI Data Min 2(2):113–123
Neto EAL, de Carvalho FDAT (2008) Centre and range method for fitting a linear regression model to symbolic interval data. Comput Stat Data Anal 52(3):1500–1515
Neto EAL, De Carvalho FDAT (2010) Constrained linear regression models for symbolic interval-valued variables. Comput Stat Data Anal 54(2):333–347
Nison S (2001) Japanese candlestick charting techniques: a contemporary guide to the ancient investment techniques of the Far East. Penguin
Pesaran MH, Shin Y (1998) An autoregressive distributed-lag modelling approach to cointegration analysis. Econom Soc Monogr 31:371–413
Qiu J, Wang B, Zhou C (2020) Forecasting stock prices with long-short term memory neural network based on attention mechanism. PLoS ONE 15(1):e0227222

Ramadhan A, Palupi I, Wahyudi BA (2022) Candlestick patterns recognition using CNN-LSTM model to predict financial trading position in stock market. J Comput Syst Inform (JoSYC) 3(4):339–347
Rogers LCG, Satchell SE (1991) Estimating variance from high, low and closing prices. Ann Appl Prob 1:504–512
Romeo A, Joseph G, Elizabeth DT (2015) A study on the formation of candlestick patterns with reference to Nifty index for the past five years. Int J Manag Res Rev 5(2):67
Santur Y (2022) Candlestick chart based trading system using ensemble learning for financial assets. Sigma J Eng Nat Sci 40(2):370–379
Shiu Y, Lu T (2011) Pinpoint and synergistic trading strategies of candlesticks. Int J Econ Finance 3(1):234–244
Sims CA (1980) Macroeconomics and reality. Econom J Econom Soc 48:1–48
Smith DM, Wang N, Wang Y, Zychowicz EJ (2016) Sentiment and the effectiveness of technical analysis: evidence from the hedge fund industry. J Financ Quant Anal 51(6):1991–2013
Staffini A (2022) Stock price forecasting by a deep convolutional generative adversarial network. Front Artif Intell 5:8
Sun G, Chen T, Wei Z, Sun Y, Zang H, Chen S (2016) A carbon price forecasting model based on variational mode decomposition and spiking neural networks. Energies 9(1):54
Tharavanij P, Siraprapasiri V, Rajchamaha K (2017) Profitability of candlestick charting patterns in the stock exchange of Thailand. SAGE Open 7(4):2158244017736799
Tsai CF, Quan Z-Y (2014) Stock prediction by searching for similarities in candlestick charts. ACM Trans Manag Inform Syst 5(2):1–21
Varghese AA, Krishnadas J, Satheesh KR (2023) Candlestick chart based stock analysis system using ensemble learning. In: 2023 International conference on networking and communications (ICNWC). IEEE, pp 1–7
von Mettenheim H, Breitner MH (2012) Forecasting and trading the high-low range of stocks and ETFS with neural networks. In: International conference on engineering applications of neural networks. Springer, pp 423–432
Xiaojie X, Zhang Y (2023) Coking coal futures price index forecasting with the neural network. Miner Econ 36(2):349–359

Publisher's Note