A New Method for Measuring Stock Trend Similarity: Integration of Sliding Window and Symbolic Statistics

Xiangliang Chen [1, a], Boxiang Liu [2, b]

[1]School of Information Technology and Engineering, Tianjin University of Technology and Education, Tianjin 300222, China;

[2]School of Information Technology and Engineering, Tianjin University of Technology and Education, Tianjin 300222, China.

[a]xiangliang_chen@163.com, [b]box0707@163.com

Abstract

The similarity analysis of stock time series plays a significant role in investment decision-making, market forecasting, and risk assessment. However, traditional similarity measurement methods (such as Euclidean distance and cosine similarity) have limitations when dealing with amplitude shifts, scaling, linear drift, and temporal axis deformations, making it difficult to accurately reflect the intrinsic trend similarity of stock prices. This paper proposes a new method for measuring the similarity of stock time series—the Integration of Sliding Window and Symbolic Statistics (ISWSS). This method calculates the logarithmic daily returns of stocks to eliminate price level differences and uses sliding window technology to segment and statistically analyze the symbol changes in two sequences. The local trend similarity is measured by the proportion of symbol matches within each window, and the overall similarity is obtained by averaging these values. Experimental results show that the ISWSS method effectively overcomes the impact of amplitude changes and temporal axis scaling. Its performance on synthetic and real stock data is superior to cosine similarity and Pearson similarity, closer to the actual trend similarity. Moreover, this method has good interpretability in mathematical probability statistics, intuitively reflects the trend consistency of time series, enhances the robustness and accuracy of similarity measurement, and provides a new approach for analyzing stock trend similarity.

Keywords

Stock Time Series, Similarity Measurement, Sliding Window, Symbolic Statistics, Trend Change.

1. Introduction

In the financial market, the fluctuation of stock prices is one of the core concerns of investors. The analysis of stock price time series is significant for understanding market dynamics, predicting price trends, and constructing investment portfolios. Among these analyses, the measurement of similarity in stock time series [1] is a crucial component, as it helps investors identify stocks with similar trends and provides strong support for investment decision-making.

Traditional methods for measuring the similarity of stock time series mainly include Euclidean distance, Pearson similarity, and cosine similarity. Euclidean distance [2] measures similarity by calculating the absolute differences between two sequences but is sensitive to amplitude changes in data, often misjudging trend similarity due to price level differences. Pearson similarity [3] measures the linear correlation between two sequences by correlation coefficients, eliminating the impact of scale but having limitations when dealing with non-linear changes in time series. Cosine similarity [4] assesses similarity by calculating the cosine of the angle between two sequence vectors and is sensitive to sequence length and amplitude changes. Moreover, these methods often overlook the economic significance of stock price fluctuations, focusing on numerical proximity rather than the actual upward or downward trends.

In recent years, with the increasing volume of financial market data and the rapid development of computing technology, researchers have begun to explore new methods for measuring time series similarity to better capture the intrinsic patterns of stock prices. Among these, symbolic sequence analysis has gained attention. This method simplifies data processing by converting stock price fluctuations into symbolic sequences and more intuitively reflects stock price trends. However, pure symbolic sequence analysis [5] may lose important information when dealing with long sequences, leading to imprecise similarity measurements.

To address the limitations of existing methods, this paper proposes a new method for measuring the similarity of stock time series—the Integration of Sliding Window and Symbolic Statistics (ISWSS). This method first calculates the logarithmic daily returns of stock time series to eliminate price level differences and highlight relative price changes. It then uses sliding window technology to segment and statistically analyze the symbol changes in two sequences, measuring local similarity by counting symbol matches within each window. Finally, the overall trend similarity is obtained by averaging these local similarity values. This method retains the trend sensitivity of symbolic sequence analysis while enhancing the ability to capture local features through sliding window technology and reducing computational complexity.

The ISWSS method proposed in this paper has the following innovations: (1) It segments the logarithmic daily returns of stocks using sliding window technology and counts the number of symbol matches within each window, capturing local trend features while avoiding the over-sensitivity of traditional methods to overall amplitude changes. (2) It uses symbolic statistics to directly reflect the consistency of stock price trends, rather than relying solely on numerical differences, significantly enhancing the sensitivity to trend similarity.

2. Research Methods

2.1. Sliding Window Method

The Sliding Window Method [6] is an effective approach for measuring the similarity of two time series through local analysis. Its core idea is to slide a fixed-size window over the time series, extract subsequences within the window for similarity calculations, and aggregate local similarity results to obtain a global similarity measure. The main implementation process is shown in Figure 1.

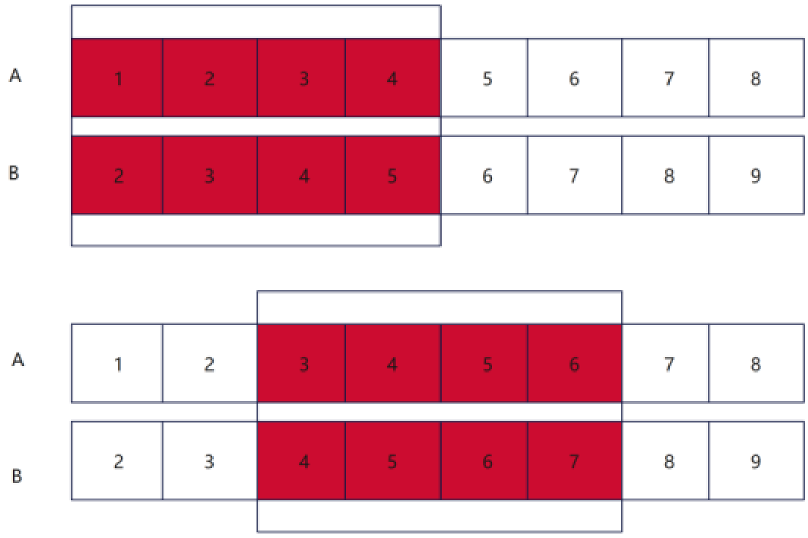


Figure 1. Sliding process of the sliding window technique.

Step 1: Before similarity measurement, time series data are typically preprocessed to ensure that the two sequences have the same length.

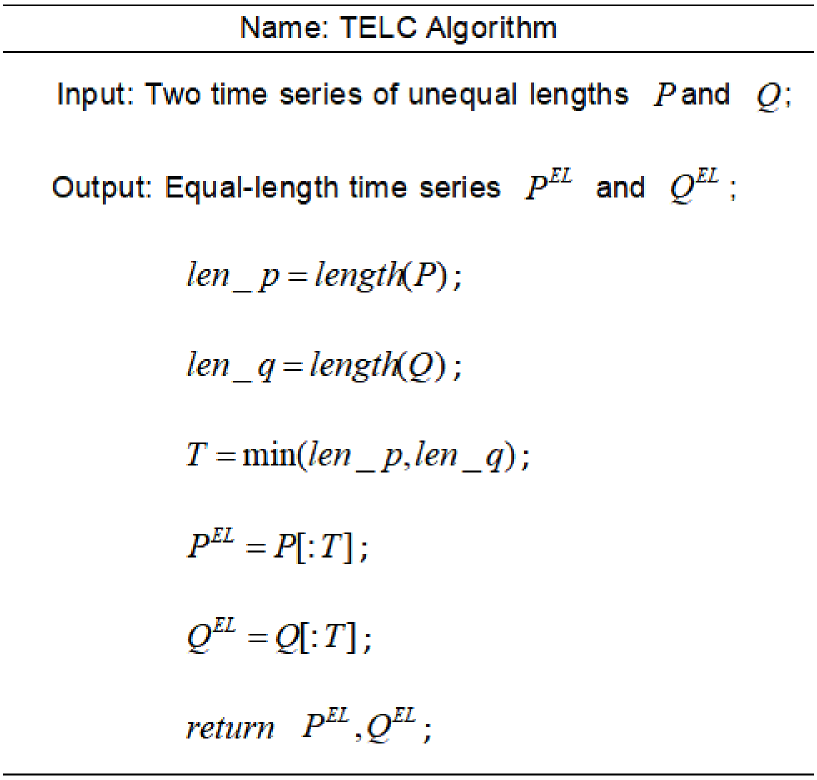
Step 2: Set the size of the sliding window w and the step length s.

Step 3: Starting from the beginning of the time series, cover the sequence with the sliding window and move it by one step length each time.

Step 4: At each window position, extract the subsequences from the two time series corresponding to the window.

Step 5: Calculate the similarity for each pair of subsequences within the window.

Step 6: Summarize the local similarity values obtained from each window and calculate the average to obtain the final global similarity measure.



```
Name: TELC Algorithm
Input: Two time series of unequal lengths $P$ and $Q$;
Output: Equal-length time series $P^{EL}$ and $Q^{EL}$;
    $len\_p = length(P)$;
    $len\_q = length(Q)$;
    $T = \min(len\_p, len\_q)$;
    $P^{EL} = P[:T]$;
    $Q^{EL} = Q[:T]$;
    return $P^{EL}, Q^{EL}$;
```

Figure 2. Pseudocode for the TELC algorithm.

2.2. Integration of Sliding Window and Symbolic Statistics (ISWSS)

This paper proposes a new method for measuring the similarity of stock time series based on the integration of sliding window and symbolic statistics. This method quantifies the trend similarity between two stock time series by statistically analyzing the symbol changes in daily return sequences and dynamically calculating the similarity of subsequences using sliding window technology. The core of this method is to capture the co-movement trends of stock prices while enhancing the sensitivity to local trends through sliding window technology.

2.2.1 Data Preprocessing

Before similarity measurement, the two stock time series are preprocessed to ensure data consistency and comparability. The specific steps are as follows:

Calculate the daily return series: For a given stock time series $P = \{P_1, P_2, \ldots, P_T\}$,compute its corresponding logarithmic daily return series $R = \{R_1, R_2, \ldots, R_{T-1}\}$,where

$$R_t = ln(\frac{P_{t+1}}{P_t}), \quad t = 1,2,\ldots,T-1 \tag{1}$$

The logarithmic return effectively reflects the relative changes in stock prices while avoiding the impact of absolute price differences.

Equal-Length Processing [7]: If the two stock time series have different lengths, they need to be processed to have the same length. Specific methods include interpolation, truncation, or padding, to ensure that the two daily return sequences have the same length T. In this paper, we adopt the truncation method to cut the excess part of the longer sequence to match the length of the shorter sequence. Figure 2 provides the pseudocode for the Truncation Equal Length Converter (TELC) algorithm based on truncation.

2.2.2 Sliding Window and Symbolic Statistics

After data preprocessing, the sliding window technique is used to measure the similarity of two equal-length daily return sequences. The specific steps are as follows:

Define the sliding window: Set a fixed-size sliding window with a positive integer size w, where $w < T$. The sliding window starts from the beginning of the sequence and moves to the right step by step. The sliding stops when the remaining sequence length is less than w.

Symbolic Statistics and Subsequence Similarity Calculation: In each sliding window position, extract the subsequences $R^{(p)}_{i:i+w}$ and $R^{(q)}_{i:i+w}$ from the two daily return sequences, where i is the starting position of the window. For each corresponding position $j(j = 1,2,\ldots,w)$ in the subsequences, compare the signs of the values. If the signs are the same (both positive, both negative, or both zero), increment the counter m by 1. The number of matches in signs m can be expressed as:

$$m = \sum_{j=1}^{w} I(sign(R^{(p)}_{i+j}) = sign(R^{(q)}_{i+j})) \tag{2}$$

where, I is an indicator function, and $sign(x)$ represents the sign of x (positive, negative, or zero), which corresponds to an increase, decrease, or no change in value.

The similarity S_i of the subsequence is defined as the ratio of the number of matches in signs to the window size.

$$S_i = \frac{m}{w} \tag{3}$$

where,The value of S_i ranges between [0,1], with values closer to 1 indicating higher trend similarity between the two subsequences.

Sliding Step and Global Similarity Calculation: The sliding window moves to the right with a fixed step size s, where s is a positive integer. After each move, repeat the above step (2) until the termination condition is met or the sliding window reaches the end of the sequence. Finally, calculate the overall trend similarity S between the two stock time series by averaging the similarity values of all windowed subsequences $\{S_1, S_2, \ldots, S_k\}$:

$$S = \frac{1}{k}\sum_{i=1}^{k} S_i \tag{4}$$

where, k represents the total number of sliding window iterations.

3. Experiments and Analysis

3.1. Experimental Data

This study uses two datasets for experimental validation. The first dataset consists of synthetic time series data, including five sequences of length 20. Sequence A is a randomly generated original sequence. Sequence B is obtained by shifting Sequence A. Sequence C is obtained by amplifying the rate of change of Sequence A by three times. Sequence D is obtained by sequentially applying shifting and amplifying the rate of change to Sequence A. Sequence E has the same trend as Sequence A except for specific points where the trend is opposite, as shown in Figure 3.

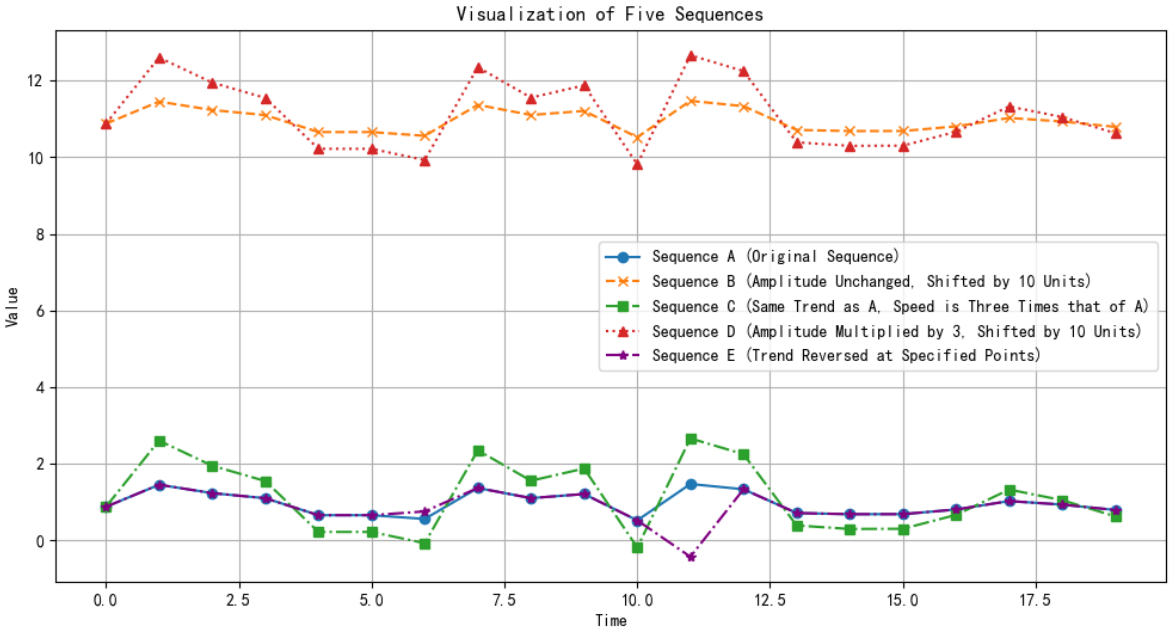


Figure 3. Amplitude scaling and shifting of sequences.

The second dataset consists of real stock data, selecting the closing prices of four stocks (Wanhua Chemical, Wantai Biological, Sany Heavy Industry, and 360) from the SSE 180 Index over 393 trading days from February 1, 2023, to September 2, 2024.

3.2. Experimental Design

This study designs two sets of experiments to systematically evaluate the performance of the proposed ISWSS method in measuring time series similarity and compare it with two traditional methods: cosine similarity and Pearson similarity.

Experiment 1: Based on the first dataset, the similarities between the five sequences (A, B, C, D, E) are compared pairwise to observe the sensitivity of different methods to amplitude scaling, linear drift, and local trend differences. The probability of two sequences having the same trend over the same period, set by probabilistic statistics, is used as the benchmark to evaluate the accuracy of similarity measurement by each method. The benchmark trend similarity between Sequence E and the other sequences (A, B, C, D) is 0.842. The sliding window size and step length for the ISWSS method are set to 4 and 2, respectively.

Experiment 2: To verify the accuracy and predictive power of the three similarity measurement methods in real financial time series, the closing price sequences of the four stocks in the second dataset are divided into 80% training sets and 20% testing sets. The similarities of the training sequences are calculated to compare the accuracy of the Pearson correlation coefficient, cosine similarity, and the ISWSS method (with a sliding window size of 20 and step length of 4). The probability of two sequences having the same trend over the same period in the testing set is used to validate the accuracy of the three methods in financial time series analysis.

3.3. Experimental Results Analysis

Experiment 1 simulates the impact of different types of sequence changes on three similarity measurement methods, with the results shown in Figure 4.

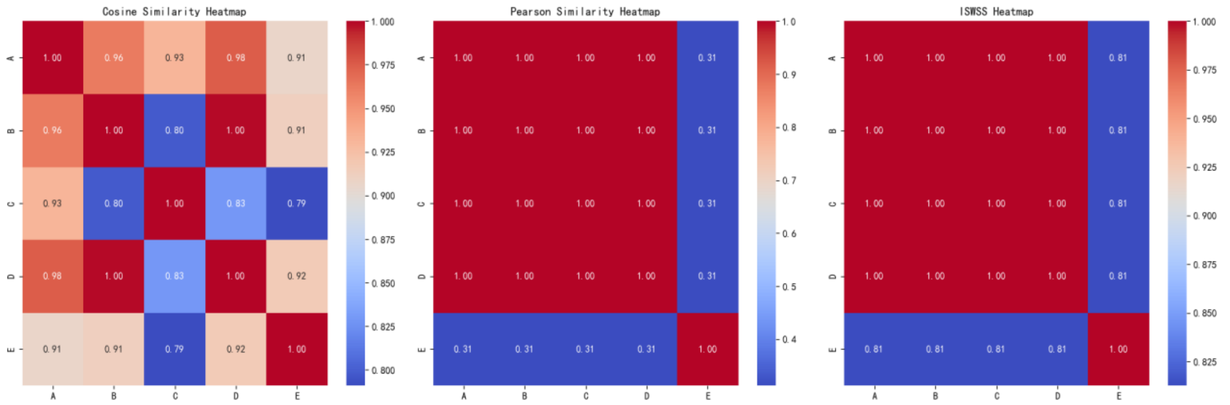


Figure 4. Similarity measurement results of the three methods.

From the experimental results in Figure 4, it can be observed that cosine similarity is highly sensitive to amplitude scaling and linear drift. This sensitivity limits its applicability in trend similarity analysis, as it fails to achieve the actual trend similarity value of 1 when measuring the similarity between Sequences A, B, C, and D. In contrast, both Pearson similarity and the ISWSS method effectively overcome the impact of amplitude scaling and linear drift, accurately identifying the trend consistency between sequences. When measuring the trend similarity between Sequence E and the other sequences (A, B, C, D), the ISWSS method yields a similarity value of 0.81, which is closer to the benchmark similarity of 0.842 than the result obtained by Pearson similarity. This finding indicates that the ISWSS method has a significant advantage in considering both overall trends and local features, more accurately reflecting the actual trend similarity between time series.

Experiment 2 uses the three similarity measurement methods to calculate the pairwise similarity of the training sequences of the four real stocks, with the results shown in Tables 1, 2, and 3.

Table 1. Cosine similarity between training sequences of the four stocks.

	Wanhua Chemical	Wantai Biological	Sany Heavy Industry	360
Wanhua Chemical	1.0000	0.9561	0.9993	0.9756
Wantai Biological	0.9561	1.0000	0.9590	0.9543
Sany Heavy Industry	0.9993	0.9590	1.0000	0.9769
360	0.9756	0.9543	0.9769	1.0000

Table 2. Pearson similarity between training sequences of the four stocks.

	Wanhua Chemical	Wantai Biological	Sany Heavy Industry	360
Wanhua Chemical	1.0000	0.5270	0.9303	0.5202
Wantai Biological	0.5270	1.0000	0.6021	0.4959
Sany Heavy Industry	0.9303	0.6021	1.0000	0.5636
360	0.5202	0.4959	0.5636	1.0000

Table 3. ISWSS similarity between training sequences of the four stocks.

	Wanhua Chemical	Wantai Biological	Sany Heavy Industry	360
Wanhua Chemical	1.0000	0.5912	0.6439	0.5743
Wantai Biological	0.5912	1.0000	0.5730	0.5905
Sany Heavy Industry	0.6439	0.5730	1.0000	0.6081
360	0.5743	0.5905	0.6081	1.0000

The probability of two sequences having the same trend over the same period in the testing set is shown in Table 4.

Table 4. Benchmark similarity between testing sequences of the four stocks.

	Wanhua Chemical	Wantai Biological	Sany Heavy Industry	360
Wanhua Chemical	1.0000	0.5769	0.5641	0.5512
Wantai Biological	0.5769	1.0000	0.4744	0.6538
Sany Heavy Industry	0.5641	0.4744	1.0000	0.5641
360	0.5512	0.6538	0.5641	1.0000

From the analysis of Tables 1 to 4, it can be concluded that cosine similarity has significant limitations in measuring stock trend similarity. The similarity values obtained in the training set differ greatly from the probability of having the same trend in the testing set, indicating its low applicability in actual stock trend analysis. Further comparison between Pearson similarity and ISWSS shows that, except for a few stock pairs (e.g., 360 and Sany Heavy Industry), the ISWSS measurement results are generally closer to the probability of having the same trend in the testing set, demonstrating higher accuracy and stability. Particularly in the case of Wanhua Chemical and Sany Heavy Industry, despite a Pearson similarity of 0.93 in the training set, the probability of having the same trend in the testing set is only around 0.564. This significant deviation further highlights the instability of Pearson similarity in certain stock trend measurements. In contrast, ISWSS shows stable performance in measuring the similarity between the training sequences of the four stocks, with results close to the actual probabilities in the testing set. This validates the applicability and superiority of ISWSS in measuring stock trend similarity.

4. Conclusion

This paper proposes a new method for measuring the similarity of stock time series based on the integration of sliding window and symbolic statistics (ISWSS), aiming to overcome the limitations of traditional methods in dealing with amplitude changes and temporal axis scaling. Through experimental validation, the ISWSS method demonstrates higher accuracy and stability on both synthetic and real stock data. It effectively eliminates the impact of amplitude changes on similarity measurement, enhances adaptability to temporal axis scaling, and more accurately reflects the co-movement trends of stock prices. Compared with cosine similarity and Pearson similarity, the ISWSS method shows significant advantages in handling complex market conditions, with results closer to actual trend similarity. This method provides a new and effective tool for analyzing stock trend similarity and offers new theoretical support and practical guidance for investment decision-making, market forecasting, and risk assessment.

References

[1] Zhao F, Gao Y, Li X, et al. A similarity measurement for time series and its application to the stock market[J]. Expert Systems with Applications, 2021, 182: 115217.

[2] da Silva M R, de Carvalho O A, Guimarães R F, et al. Wheat planted area detection from the MODIS NDVI time series classification using the nearest neighbour method calculated by the Euclidean distance and cosine similarity measures[J]. Geocarto International, 2020, 35(13): 1400-1414.

[3] Li Z, Wang L, Wang D. Analysis of music similarity based on Pearson correlation coefficient[J]. J]. Art and Performance Letters, 2021, 2(5).

[4] Dong Y, Sun Z, Jia H. A cosine similarity-based negative selection algorithm for time series novelty detection[J]. Mechanical Systems and Signal Processing, 2006, 20(6): 1461-1472.

[5] Yamano T, Sato K, Kaizoji T, et al. Symbolic analysis of indicator time series by quantitative sequence alignment[J]. Computational Statistics & Data Analysis, 2008, 53(2): 486-495.

[6] Yagoubi D E, Akbarinia R, Kolev B, et al. ParCorr: efficient parallel methods to identify similar time series pairs across sliding windows[J]. Data Mining and Knowledge Discovery, 2018, 32: 1481-1507.

[7] Farnoosh, A., Azari, B., & Ostadabbas, S. (2021). Robust explainer recommendation for time series classification. Proceedings of the AAAI Conference on Artificial Intelligence, 35(8), 7394–7403. DOI: 10.1609/aaai.v35i8.16907.