Project Report of Stock Data Imputation

Yi Liu[1],
[1]Department of Computing, The Hong Kong Polytechnic University
joeylau.liu@connect.polyu.hk,

1. A Personal Understanding

This project aims to impute the missing inventory data on the public holidays. Since the holidays are different in different places, the stock data from other places can be used as a condition to generate the invisible stock trend. Since we use global stock, the trend we rely on is a global economic trend that is more significant and easier to grasp from a long-term perspective. Based on this understanding, the following program design is created.

2. Dataset Preprocessing

The dataset under investigation consists of stock data from the current Dow Jones 30, EuroStoxx 50 and Hang Seng indices over the last decade. In the following subsections, a detailed explanation of the preprocessing procedures applied to the dataset will be provided.

2.1. Dimension

The formation of the dataset requires the determination of its dimensional characteristics. The dataset encompasses a total of 156 companies from the Dow Jones 30, EuroStoxx 50, and Hang Seng indices. Each company possesses 7 features, obtained from the yfinance package, including ”Open”, ”High”, ”Low”, ”Close”, ”Volume”, ”Dividends”, and ”Stock Splits”. The time granularity of the data can be adjusted through the interface provided by the yfinance package.

Given the 10-year scope of this task, the objective is to uncover long-term patterns and the impacts of global economics. To achieve this goal, we chose a daily granularity as the most appropriate option for forming the data. To reveal these long-term patterns, a representative feature was required for each day. We utilized the ”High” and ”Low” features from yfinance and calculated the median as the representative value.

After aligning all data based on the date index, the resulting dataset had a size of 3655×156. The first dimension represents the time length and the second dimension represents the various companies. We then inspected the first 10 rows of the data and any companies with None values in the first 10 rows were considered to have insufficient historical data over the 10-year period and were subsequently filtered out. This resulted in a final dataset of size 3655×131.

2.2. Standarization

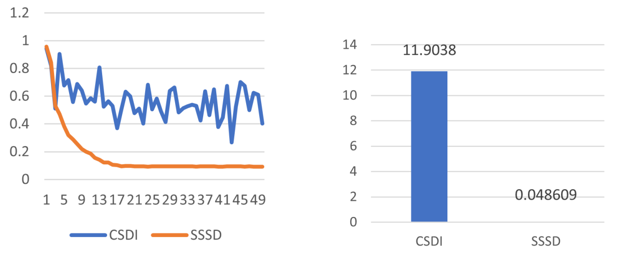
As the prices of different stocks can vary greatly, excessive variations in price ranges could negatively impact performance. To mitigate this, we standardized the dataset in a column-wise manner. This resulted in each company’s stock data having an average value of 0 and a standard deviation of 1. Following standardization, any remaining None values were filled with 0.

2.3. Data Split

We divided the entire dataset into 11 smaller batches along the time dimension. Each batch had a shape of 332×131. Ten of these batches were used for training purposes and one was used for testing. Given the relatively small volume of the data, a batch size of 1 was utilized. In order to address the potential for overfitting, a random mask was generated each time data was sampled.

2.4. Configuration

In the interest of fairness, we utilized the recommended diffusion configurations provided by both the SSSD [1] and CSDI [3] frameworks. The training configuration involved a learning rate of 0.0002 with a step scheduler, applied to both methods. The training phase comprised 50 epochs for both SSSD and CSDI. During the evaluation phase, since only one batch of test data was available, we randomly generated 5 masks, and the test performance was computed 5 times. The performance was evaluated using the Root Mean Squared Error (RMSE) between the generated data and the original data in the masked positions. We generated 100 samples and use their average as the predicted value. The results of the experiment are presented in the subsequent section.



(a) Train Loss

(b) Test RMSE

Figure 1. Performance Comparison. a) The training loss curve. The x represents epoch index. b) Test RMSE.

3. Experiment Result

3.1. Performance Comparison

In this study, we conducted a direct comparison of the performance between the SSSD and CSDI methods. A random masking strategy was employed to simulate holidays, whereby 10% of the visible data in each column of each stock was randomly masked. Subsequently, the masked values were generated by both the SSSD and CSDI methods based on the unmasked data.

In this study, we limit our comparison to the case of random missing data (RM). This is because we believe that the distribution of holidays among workdays is more akin to RM compared to missing not at random (MNR), blackout missing (BM), and time series forecasting (TF). The training loss and test RMSE results are depicted in Figures 1a and 1b, respectively.

The training loss curve demonstrates a significant advantage of SSSD over CSDI, with SSSD converging at an early stage, while CSDI fails to converge. The test RMSE results further confirm this conclusion, with CSDI having a high RMSE of 11.9, while SSSD achieved a low RMSE of 0.0486. Additionally, the test loss of CSDI was found to be much higher than its training loss, indicating a high level of overfitting, which was not observed in the case of SSSD. Overall, our findings suggest that SSSD outperforms CSDI by a significant margin in the random missing task of stock data.

3.2. A Further Exploration: Holiday Effect

To determine whether investors' behavior would change if holidays were suddenly declared as workdays, we propose a hypothesis based on the weekend effect.

The weekend effect, also known as the Monday effect, refers to the tendency for stock prices to experience a decline on Mondays compared to the closing prices on the preceding Fridays. The existence of the weekend effect in stock markets has been widely documented in academic literature, with numerous studies attempting to explain the phenomenon [2].

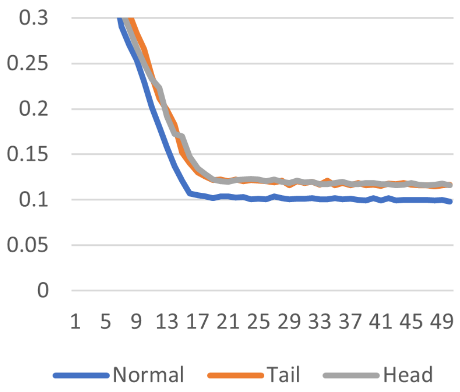


Figure 2. Train Loss

Figure 3. Experiment on holiday effects. The training loss curve of different categories of days.

One explanation for the weekend effect is rooted in investor psychology, where it is posited that individuals tend to make impulsive investment decisions that are not based on sound analysis or research over the weekends. This behavioral phenomenon, known as a "gambling mindset," can increase market volatility and heighten the risk of losses for investors.

Additionally, the weekend effect has been attributed to the slowdown in trading activity and a lack of market-moving news releases over the weekends. This reduction in market activity may result in increased volatility, as investors react to market events without the benefit of the full range of market information.

Based on our analysis, it is our belief that holidays may have a similar impact on the stock market. We therefore put forth the "holiday effect" hypothesis, which postulates that the market behavior observed during weekends is also applicable to holidays. It is suggested that the day prior to a holiday may display comparable characteristics to a typical Friday, and the day following a holiday may demonstrate similar characteristics to a typical Monday.

In an effort to provide further proof for the existence of the "holiday effect," we conducted an additional experiment. This experiment involved classifying working days into three categories: head days, normal days, and tail days. Head days refer to the first workday following a holiday, tail days refer to the workday preceding a holiday, and normal days refer to the workdays that fall between two workdays.

To better complete the stock data imputation task, instead of using a universal imputer, we believe that there should be three imputers to handle each case with a trained model. We then trained SSSD on three different cases, as shown in Fig.3, we can notice that the loss of the tail days and head days are higher than that of normal days by a noticeable margin, which represents that the behavior of investors is harder to predict in the head days and tail days, confirming the existence of holiday effects. After the training, the average test RMSE between the three models on three categories of data is 0.04499, which is lower than the previous hybrid training method.

Returning to the question above, if a holiday were suddenly canceled and became a working day, the behavior would change. Their pattern will switch from the tail day pattern to a normal day pattern since tomorrow is not a holiday anymore.

4. Forecasting Problem

This question is discuss whether it is possible to forecast HK stock fluctuation range using the stock data fo the USA from last night. We believe such an approach is unfeasible for several reasons: 1) Short-term stock data is relatively unstable. There are many factors that can affect it and the global economy is very irrelevant. 2) The time-space difference will increase the difference between investor patterns in the short term.

Also, we have no expectations for SSSD to accomplish this task. SSSD works with conditional data imputation. The condition of the diffusion model is critical to performance. Stock data from other time-space is not a decent condition.

References

[1] Juan Miguel Lopez Alcaraz and Nils Strodthoff. Diffusion-based time series imputation and forecasting with structured state space models. *arXiv preprint arXiv:2208.09399*, 2022. 1

[2] Donald B Keim and Robert F Stambaugh. A further investigation of the weekend effect in stock returns. *The journal of finance*, 39(3):819–835, 1984. 2

[3] Yusuke Tashiro, Jiaming Song, Yang Song, and Stefano Ermon. Csdi: Conditional score-based diffusion models for probabilistic time series imputation. *Advances in Neural Information Processing Systems*, 34:24804–24816, 2021. 1