Imputing missing data from stock time series

Members: Khanh Nguyen, Yizhen Zhao, Evgeniya Lagoda, Himanshu Raj, Carlos Owusu-Ansah, Sergei Neznanov

1. **Overview:**

Motivation: The NYSE and NASDAQ average about 252 trading days yearly. What if someone accidentally deleted data from five (or more) trading days? Missing data is often ignored or removed when we analyze data, which could be problematic. One of the problems is that stock returns depend on that "missingness" (S. Bryzgalova, 2022).

Stakeholders: AAPL Inc., investors, financial companies

2. **Data:**

AAPL (Apple Inc.) data from January 1, 2023 until December 31, 2023. There are 250 data points because trading market closes on Saturdays, Sundays, and national holidays. We want to impute the Close values.

Daily return is computed from the relative change of close values:

$$Y_t = \frac{X_t - X_{t-1}}{X_{t-1}}$$

where Y_t is the daily return on day t, and X_t is the adjusted close on day t.

3. **Models:**

a. Baseline model: The linear interpolating method predicted the missing close values by taking the weighted sum of the two closest points.

b. Rolling average: The method uses the average of a fixed number of points to the left and right of the missing values to make a prediction.

c. Double Exponential Smoothing: The method considers the trends in the data for data imputation.

d. SARIMA: The method incorporates seasonal patterns, trends, and autoregressive components to make predictions.

e. KNN: The method uses other data from Apple's stocks such as open stock prices and dates to make predictions.

f. Linear Regression: Companies with high correlations of the closing daily returns with Apple will be used as regression predictors.

g. Vector AutoRegressive (VAR): Companies that Granger Causes Apple's close differencing will be implemented in VAR.

4. **Results:**

• Linear interpolation remains a robust choice for both small and large gaps in stock time-series data compared to more complicated interpolation methods.

• When there is a sufficiently high correlation between the price movements of two companies, one may be used to impute missing data in the other.

5. **Next steps:**

• We could also include other predictors, such as trade volume, that affect closing prices

• We could systematically explore the circumstances under which methods we evaluated outperform linear interpolation.

• We began exploring advanced techniques like State Space Models and Neural Network-based approaches whose applications go beyond the present context.

• Preliminary analysis in this direction suggested that linear interpolation still works better though there is room for fine tuning.