Imputing missing data from stock time series

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

1. Overview:

Motivation: Missing data is a typical problem in science research. For example, in clinical trials, wearable sensors might lose signal due to battery. A telescope's line of sight might get covered by a cloud at random times. Errors in measuring instruments often leading to a gap in time series. Naively dropping missing data can remove important information. This project is about imputation of missing financial times series data in particular stock price time series. Though missing data in the daily stock prices is rare, in this project, we analyse a toy problem where we delete a few data points by hand and attempt to impute it through various methods. The goal is to see which methods and what market indicators work best for such a dataset. The completeness of stock data allows us to test how well a model predicts missing data. Analyzing imputation for such time series could therefore yield insight on correlations in international markets and the relevant models and market predictors to use for the more practical problem of making forecast in stock price movements.

Stakeholders: Investors, financial markets

2. Data and Feature Extraction:

We use AAPL (Apple Inc.) stock price data from January 1, 2023 until December 31, 2023. There are 250 data points because trading market closes on weekends and national holidays. We artificially created missing data by manually deleting Close values for 7 different windows of [5, 4, 3, 2, 1] consecutive days. The goal is to impute these "missing" data though various modelling techniques. We also used 2023 data for NVDA, MSFT, TSM, META, GOOG stocks for performing cross-sectional analysis using Linear Regression and Vector Auto Correlation. For this purpose it is convenient to work with the relative change in close prices Y_t given by

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

where X_t is the close value on day t.

3. Exploratory Data Analysis:

Examining the empirical cumulative distribution function of closing price differences reveals that price fluctuations largely follows a normal distribution across a large range of differenced values. This suggests that stock price movements follow a random walk pattern. The excess kurtosis of differenced values is positive indicating that stock price movements sample a fat-tailed distribution rather than a normal distribution - a fact that is empirically known to be true for stock prices. Given the random walk behaviour, the simplest imputing procedure would be Last Observation Carried Forward that yields an MSE of 11.38 for the missing window of length 5. In what follows we explore regression and time series models in search of other better performing techniques.

4. Models:

a. Baseline model: It turns out that linear interpolation give a lower MSE than LOCF. We therefore take this as our baseline model.

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.

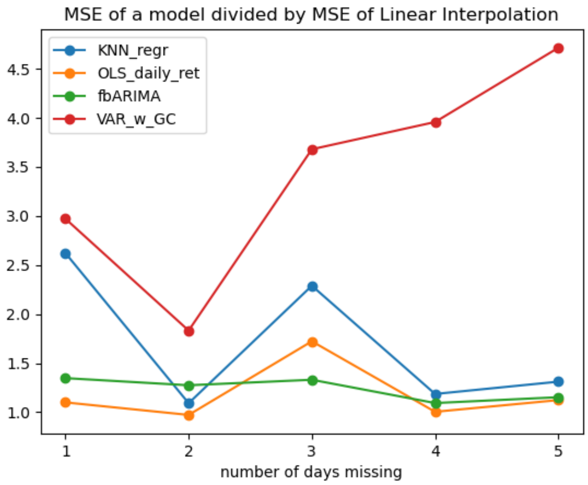
d. SARIMA: The method incorporates seasonal patterns, trends, and autoregressive components to make predictions. Can use average of both forward and backward forecasting for imputing missing data in the middle of the time series.

e. KNN: The method uses other predictors from AAPL stocks such as open stock prices and dates to make predictions.

f. Linear Regression: Companies with high correlations of the closing daily returns with AAPL are used as regression predictors.

g. Vector Auto Regression: Companies that *Granger-causes* AAPL close differences are used for fitting a vector auto regression model which is then used for imputation.

5. Results:



Above, we plot the ratio of the MSE of a model to the MSE of linear interpolation against consecutive missing days. We can see that in all circumstances of missing values no model performs any better than linear interpolation on this dataset. Consequently, linear interpolation remains a robust choice for both small and large gaps in stock time-series data compared to more sophisticated imputation methods.

6. Next steps:

- In a future work we would like to incorporate other predictors that include trading volume, derived indicators, industry trends, market indices and interest rates that affect stock price movement.
- We would like to systematically explore the circumstances under which methods we used 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 we believe there is room for fine tuning.