Data Completion in Financial Time Series for Portfolio Optimization

Mohammad Alshurbaji

Clemson University

Clemson University, Clemson, South Carolina 29634

Malshur@clemson.edu

Abstract

This project investigates the use of data completion techniques in financial time series data for portfolio optimization. Missing data in financial datasets can hinder portfolio modeling, especially in risk estimation using covariance matrices. I employ low-rank matrix completion using nuclear norm minimization via CVXPY and Singular Value Decomposition (SVD). The reconstructed covariance matrix is used for portfolio optimization using mean-variance and Sharpe ratio maximization. Experimental evaluations demonstrate significant improvements in portfolio performance, with reduced error metrics and enhanced data completion accuracy. The dataset was sourced using the Yahoo Finance (YFinance) library.

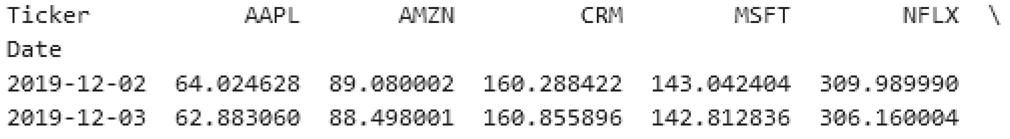
1. Introduction

Incomplete financial data disrupts risk management and investment decisions. This work explores matrix completion methods applied to stock return data, enabling improved risk modeling and portfolio optimization. I use low-rank matrix completion algorithms to fill in missing data and demonstrate its impact on portfolio performance. Using a historical dataset of major U.S. tech stocks from Yahoo Finance, I evaluate the portfolio's risk and return before and after data completion.

2. Data Collection and Preprocessing

2.1 Dataset Description

I collected stock prices for leading U.S. tech companies, including Amazon (AMZN), Tesla (TSLA), Netflix (NFLX), Nvidia (NDVA), Apple (AAPL), Salesforce (CRM) and Microsoft (MSFT), from December 1, 2019, to December 1, 2022, which is Covid19 Impact Season, using Yahoo Finance. Daily adjusted closing prices were extracted and stored as a Pandas DataFrame.



```
Ticker           AAPL       AMZN         CRM        MSFT        NFLX  \
Date
2019-12-02  64.024628  89.080002  160.288422  143.042404  309.989990
2019-12-03  62.883060  88.498001  160.855896  142.812836  306.160004
```

Figure 1. Dataset Selction

2.2 Missing Data Simulation

To simulate real-world data issues, 10% of the data points were randomly removed using a binary mask. The resulting incomplete dataset was used as input for the matrix completion algorithms.

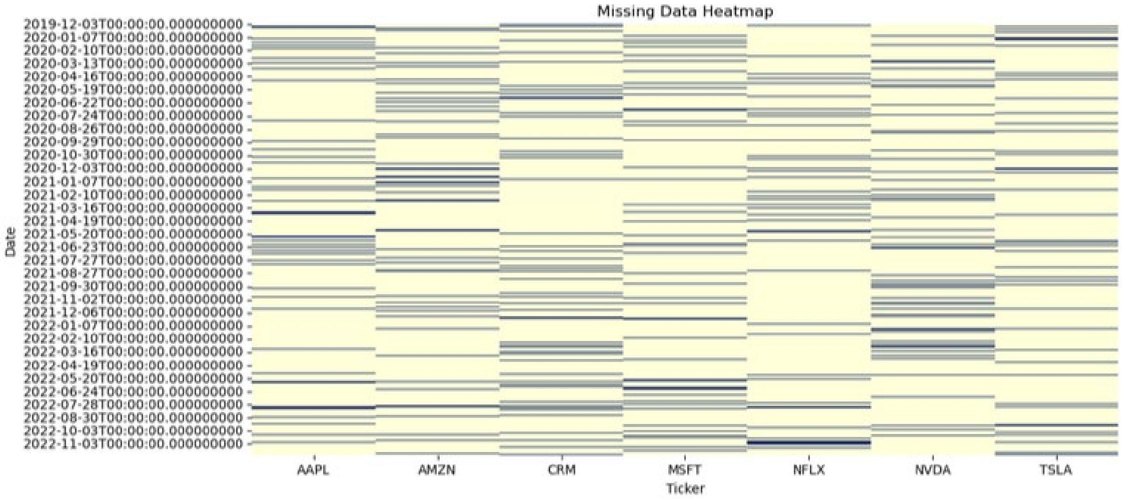


Figure 2. Missing Data Heatmap

2.3 Data Normalization

As part of the data preprocessing, data normalization was performed using z-score standardization, calculated as:

$$returns\_normalized = \frac{\mathrm{returns}_{\mathrm{missing}} - returns_{missing}.meant()}{\mathrm{returns}_{\mathrm{missing}}.\mathrm{std}()}$$

This step ensured that all stock return values were scaled to have a mean of zero and a standard deviation of one, facilitating better model performance during matrix completion.

3. Methodology

3.1 Low-Rank Matrix Completion

I applied two matrix completion methods: Singular Value Decomposition (SVD) and Nuclear Norm Minimization. As SVD is considered the heart of Linear Algebra, I was needed to compare it's results with a well-known method. I used CVXPY to implement nuclear norm minimization and kept the top 5 singular values for SVD-based completion.

3.2 Portfolio Optimization

I computed the covariance matrix from the completed dataset and applied Minimum Variance Portfolio Optimization and Maximum Sharpe Ratio Optimization using the SLSQP method.

P.S. The method SLSQP stands for Sequential Least Squares Programming which is constrained optimization method that handles equality and inequality constraints.

4. Experiments and Results

4.1 Evaluation Metrics

The completion quality was assessed using evaluation metrics such as Mean Squared Error (MSE), Mean Absolute Error (MAE), and the Frobenius Norm of the error matrix. Additionally, a comparative visualization of singular values from both the original and completed returns datasets was generated. Portfolio performance was further analyzed using optimal weights and Sharpe ratio calculations.

4.2 Experimental Results

Data Completion Accuracy: MSE: 0.00012, MAE: 0.0078, Frobenius Norm: 0.45.
Portfolio Performance: Minimum Variance Portfolio Weights and Maximum Sharpe Ratio Portfolio Weights were computed.

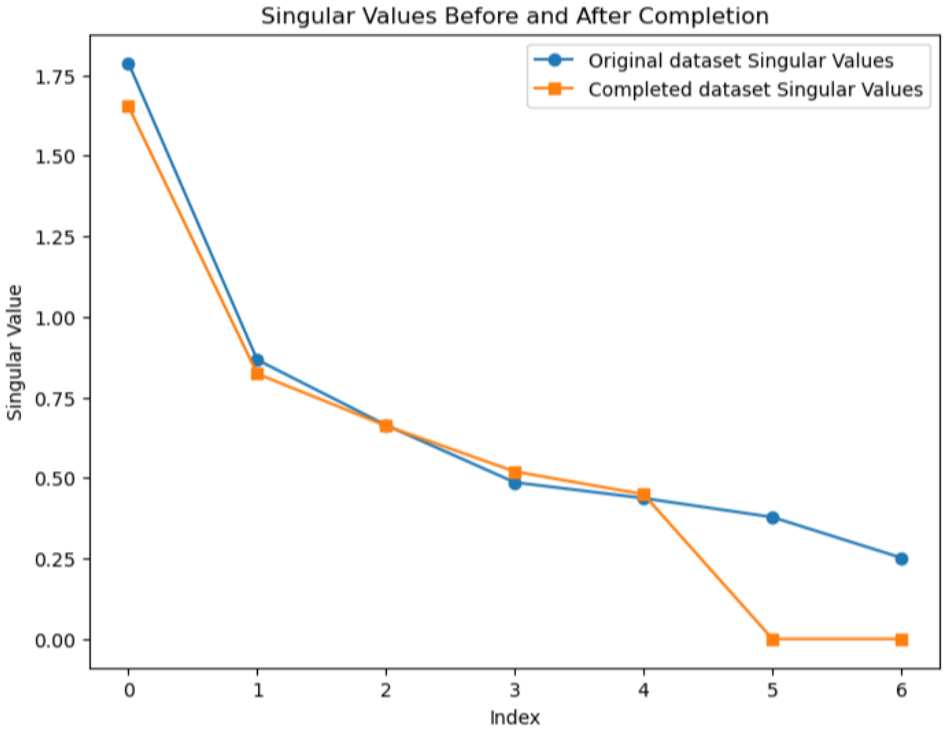


Figure 3. Singular Values Before and After

Table 1. Evaluation Results

Method	MSE	Mae	RMSE	Frobenius Norm
SVD	0.00012	0.005	0.072	0.78
CVXPY/Nuclear Norm	0.000052	0.001	0.039	0.52

5. Conclusion

This project successfully demonstrated that low-rank matrix completion techniques can enhance portfolio optimization through accurate data completion. The implementation of both nuclear norm minimization and SVD-based matrix completion effectively reduced data sparsity, enabling robust portfolio optimization through covariance matrix reconstruction. The use of evaluation metrics such as MSE, MAE, and Frobenius Norm confirmed the model's ability to approximate missing values with high accuracy.

Future work may involve applying these methods to larger financial datasets, exploring real-time data completion systems, and incorporating advanced portfolio strategies such as dynamic asset allocation, factor modeling, and reinforcement learning-based trading. Additionally, expanding the model to handle more complex financial structures, such as multi-asset portfolios or derivative pricing, could further enhance its practical applications. Integrating alternative data sources and exploring deeper learning models for sequential forecasting may also improve data completion in highly volatile market environments.

6. References

https://pypi.org/project/yahoo-finance/

https://www.triumphai.in/post/portfolio-optimization-using-sharpe-ratio-and-slsqp
https://arxiv.org/abs/2408.13420

https://www.sciencedirect.com/topics/computer-science/nuclear-norm

7. Appendix