Extracting the Low Coverage Juice

Exploiting Information of Alternative data with High Missing Ratio for Financial Forecasting

A. Tytgat – Pr. A. Shestopaloff – Pr. C. Vande Kerckhove – Pr. M. Cucuringu – E. Bregasi – A. Peek

25 June 2024, Workshop on Complex Networks in Banking and Finance

Outline

Extracting information from low coverage features to forecast fundamental values

Multi-View imputation to reconstruct incomplete data

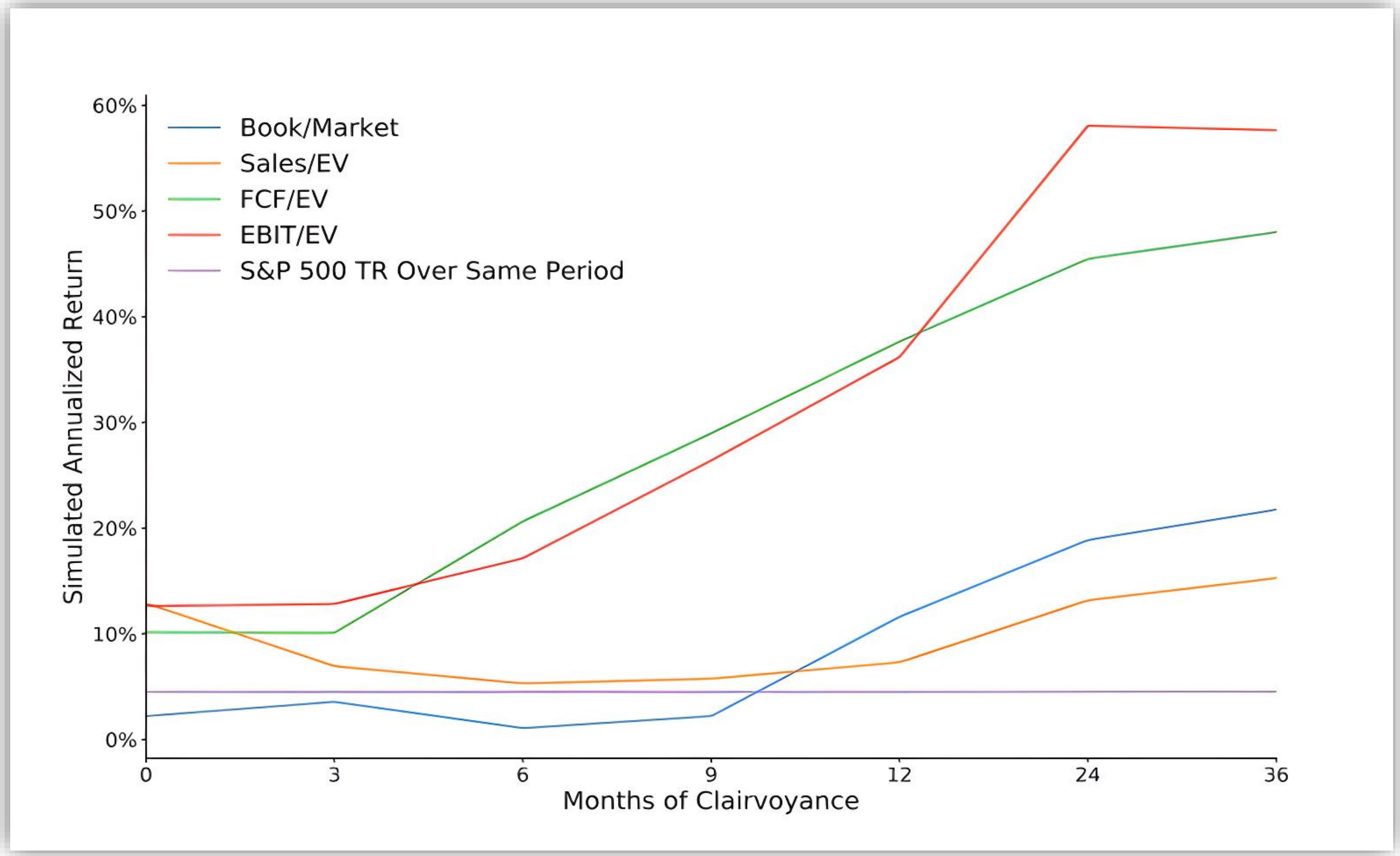
Benchmarking Multi-View imputation on datasets with various structures

Further work

Conclusion

Extracting information from low coverage features to forecast fundamental values

Forecasting fundamental data is a strong portfolio management strategy (Chauhan, L., Alberg, J. and Lipton, Z., 2020 [1])



Incorporating alternative data can improve the prediction performance

Benefits
- Timeliness
- Uncovering Hidden Insights
- Competitive Edge

Challenges

Incorporating alternative data can improve the prediction performance

Benefits
- Timeliness
- Uncovering Hidden Insights
- Competitive Edge

Challenges
- Integration with Traditional Data
- Costly

Incorporating alternative data can improve the prediction performance

Benefits

- Timeliness
- Uncovering Hidden Insights
- Competitive Edge

Challenges

- Integration with Traditional Data
- Costly

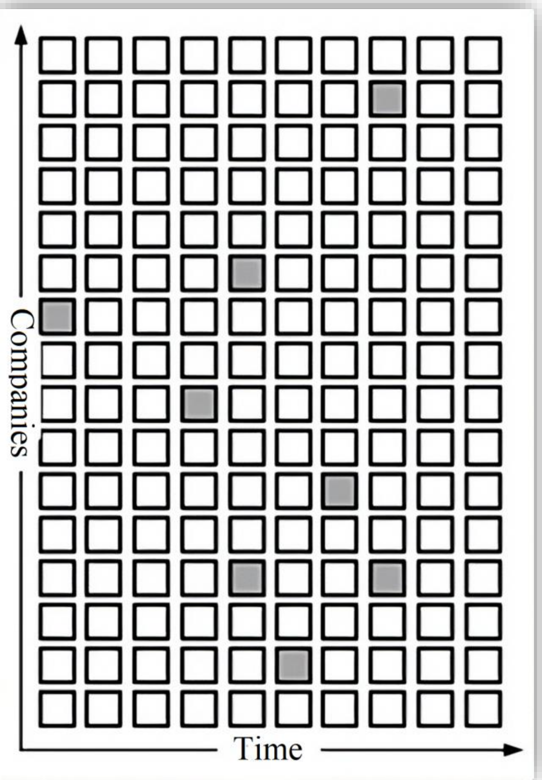
Q1 : How can we extract the most information from small sample of alternative data ?

Q2 : How can a small sample help guide the acquisition of additional observations of alternative data?

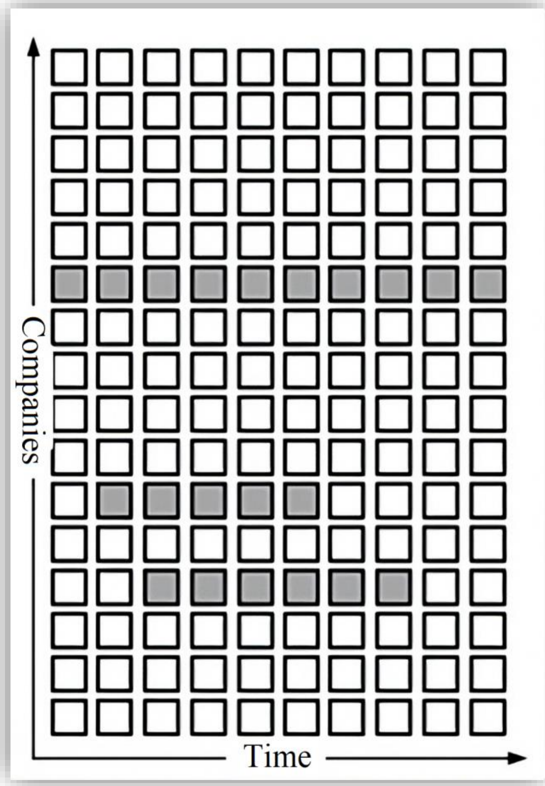
Aggregating alternative data leads to high missing ratios

Period	Company	Location	Traffic	Patent Filings	Consumer Sentiment	Average temperature	Earning Surprise
01/01/2023	Coca-Cola	coca-cola.com	Missing	3	Missing	13	+2%
01/04/2023	Coca-Cola	coca-cola.com	Missing	2	Missing	16	-3%
01/07/2023	Coca-Cola	coca-cola.com	Missing	6	Missing	15	+5%
01/01/2023	Netflix	Netflix.com/CA/	6,000,000	Missing	-0.1	-12	-1%
01/01/2023	Netflix	Netflix.com/UK/	Missing	Missing	Missing	Missing	-2%
01/04/2023	Netflix	Netflix.com/CA/	2,000,000	Missing	0.4	0	+1%
01/01/2023	Apple	Apple Store, New York	Missing	Missing	Missing	3	+3%
01/04/2023	Apple	Apple Store, New York	Missing	Missing	Missing	8	+2%
01/07/2023	Apple	Apple Store, New York	Missing	Missing	Missing	23	+1%

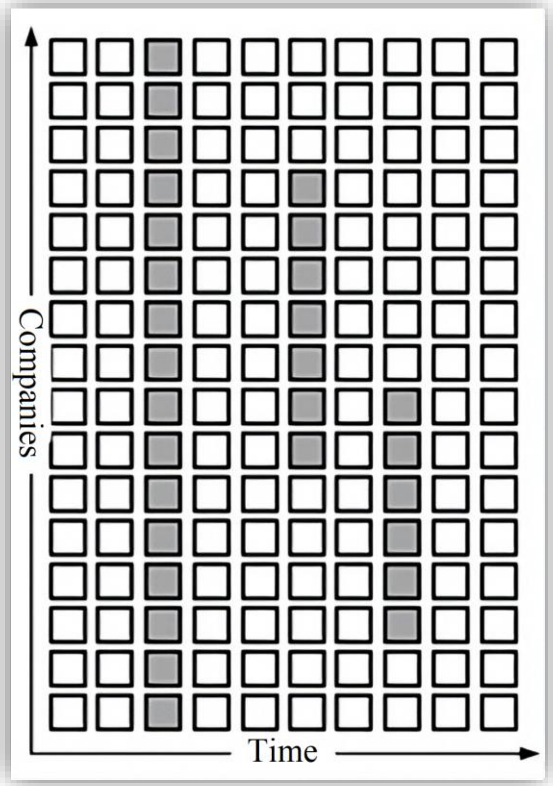
Four types of missing values patterns [3]



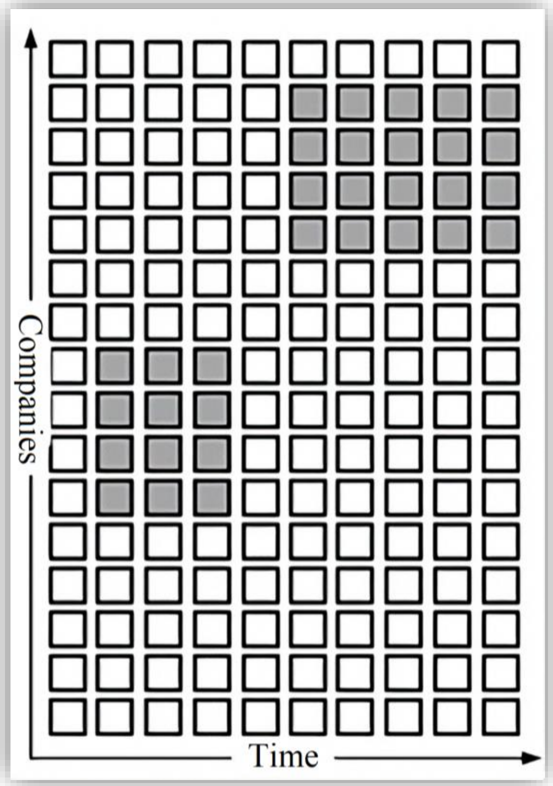
Type 1: Random



Type 2: Temporal Monotone



Type 3: Entity Monotone



Type 4: Block Monotone

Multi-View imputation to reconstruct incomplete data

Standard imputation techniques

Simple	**Advanced**
Mean, median, mode Most frequent value Forward/Backward Fill	Matrix completion Maximum likelihood Cluster-based Regression Interpolation

Limitations:

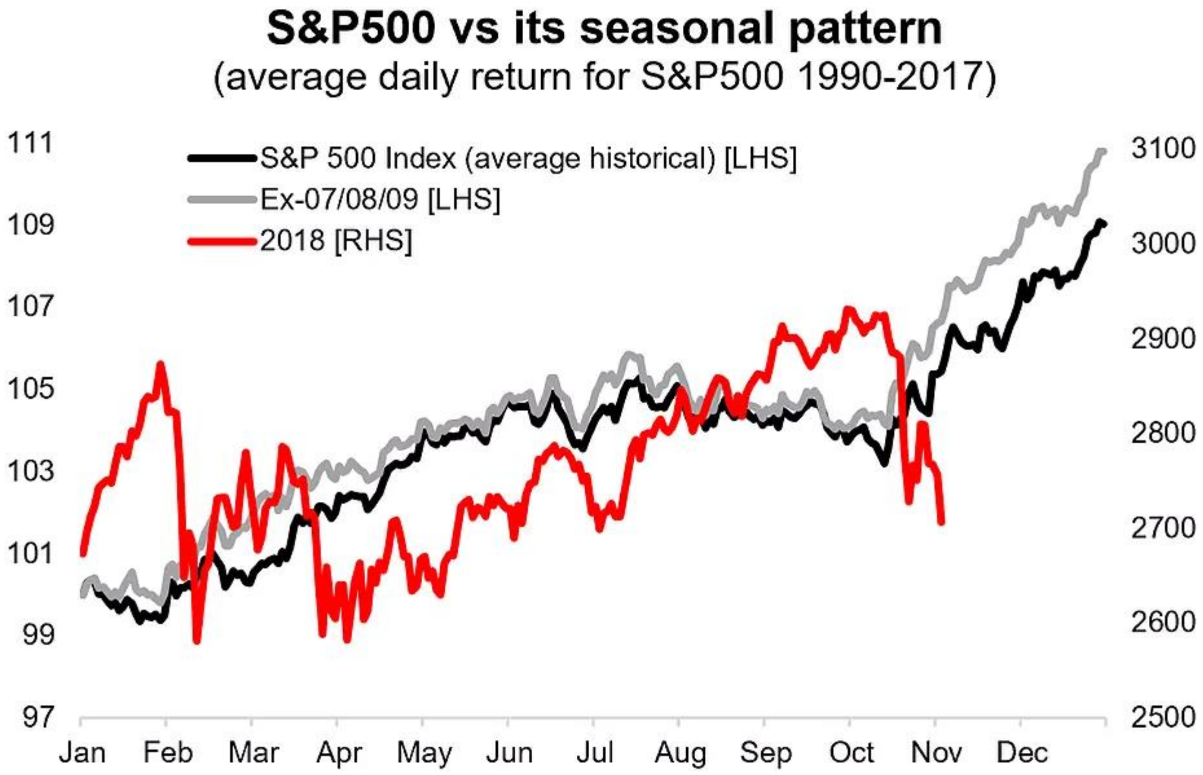
Standard imputation techniques

Simple	Advanced
Mean, median, mode Most frequent value Forward/Backward Fill	Matrix completion Maximum likelihood Cluster-based Regression Interpolation

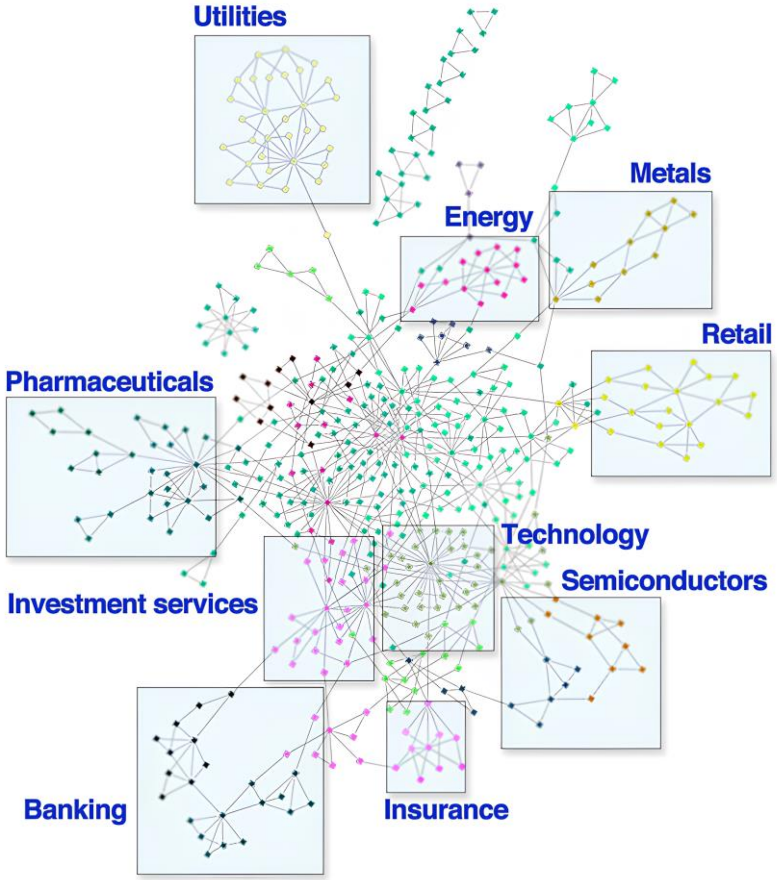
Limitations:

- Not robust to all missingness scenarios.
- Do not fully leverage the structure/pattern present in the data.

Standard methods fail to leverage all the structures present in the data



Seasonality of the S&P 500 [3]

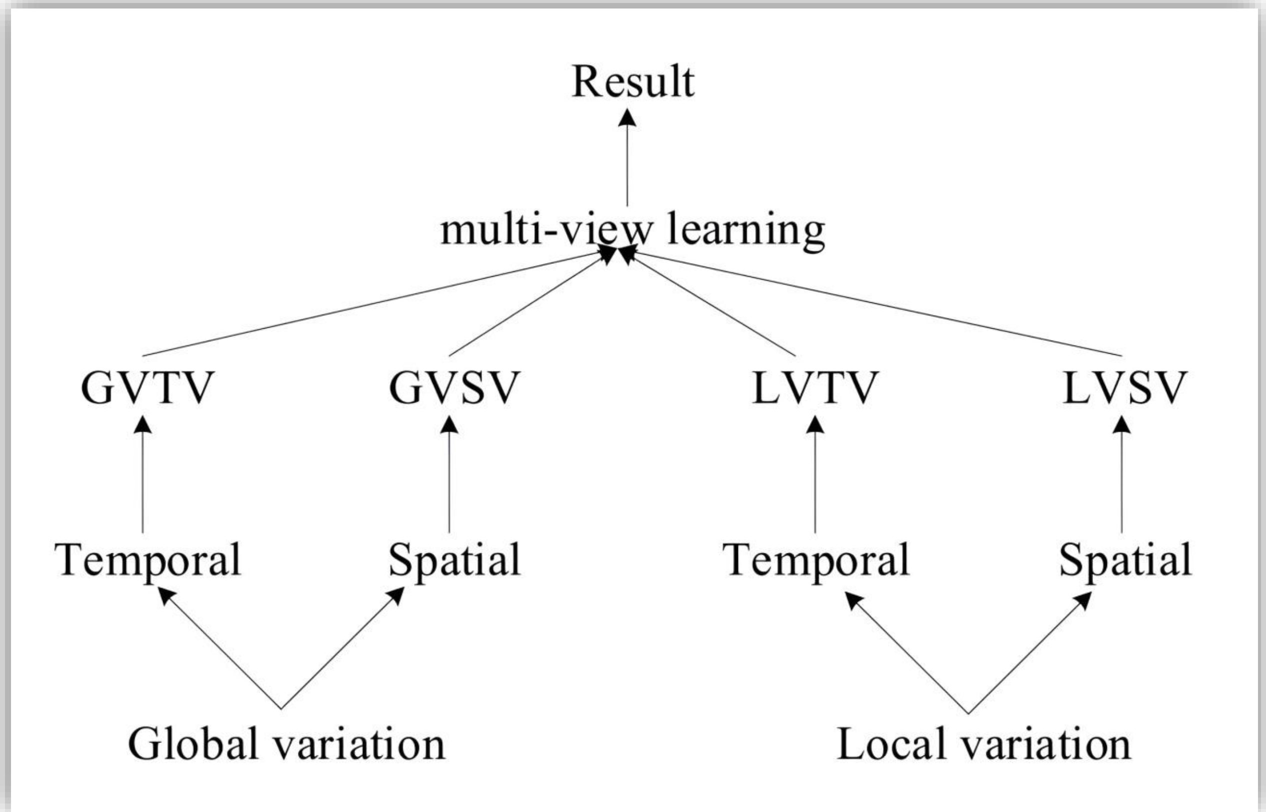


Clustering of the S&P 500 [4]

A more informed technique: Multi-View imputation (Li, L., Zhang, J., Wang, Y. and Ran, B., 2018. [3])

Ensemble of four views:

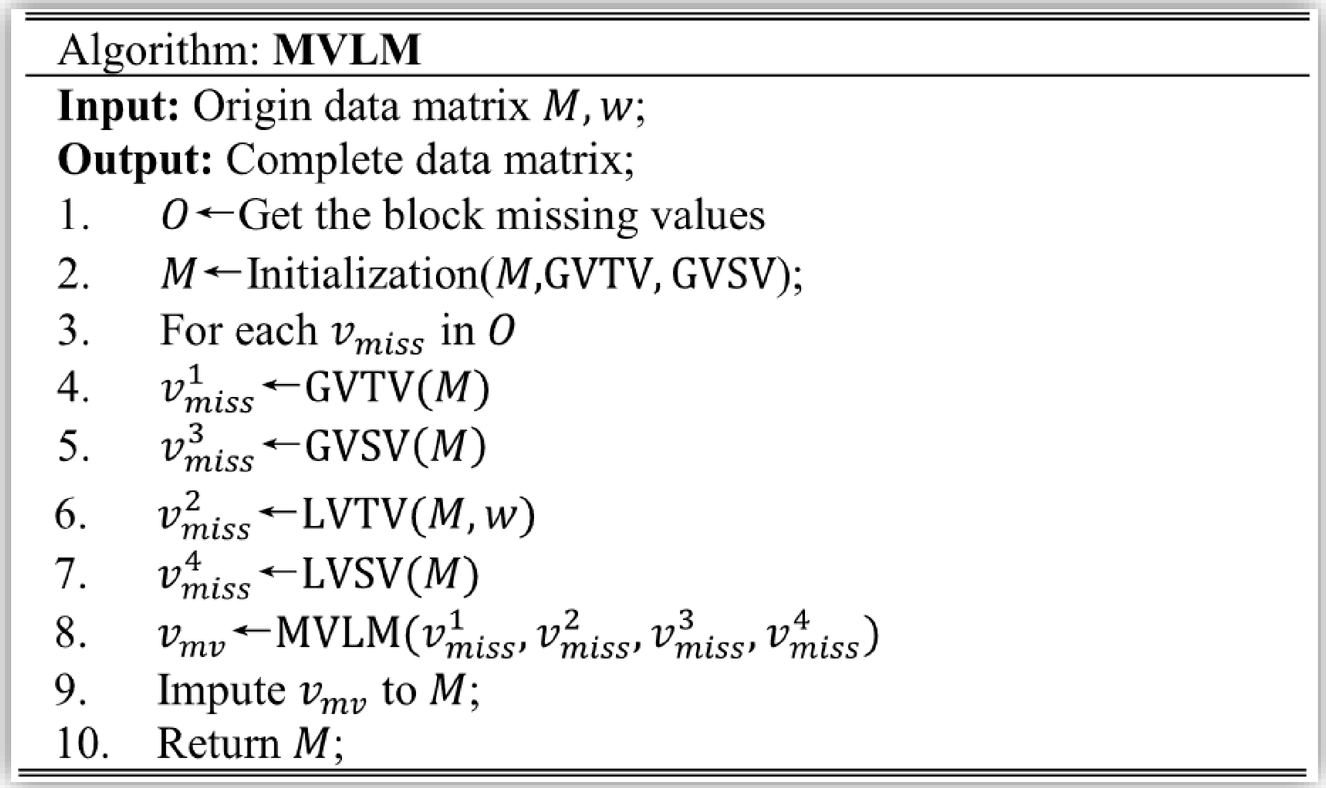
- *Local temporal* : captures the local variations over short periods.
- *Local spatial*: captures the local structural variations between entities.
- *Global temporal*: captures the global variation over long periods.
- *Global spatial*: captures the global structural relationships between different entities.



A more informed technique: Multi-View imputation (Li, L., Zhang, J., Wang, Y. and Ran, B., 2018. [3])

Advantage over standard methods

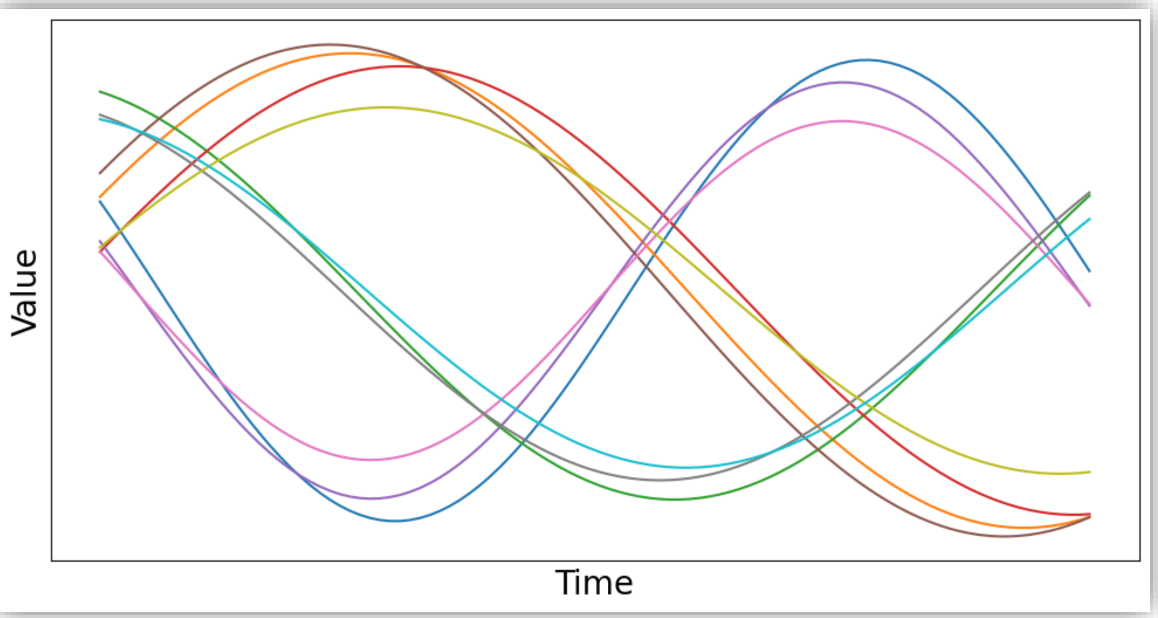
- Robust across different missingness patterns and ratio levels.
- Adaptable to different data structures.



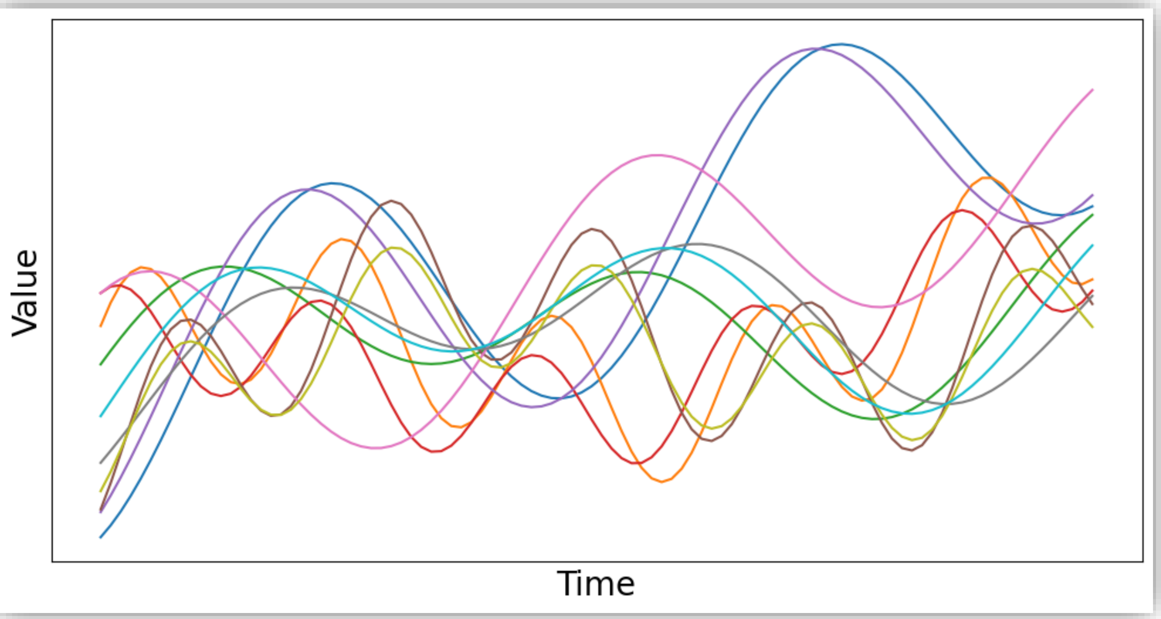
```
Algorithm: **MVLM**
**Input:** Origin data matrix $M, w$;
**Output:** Complete data matrix;
$O$ ← Get the block missing values
$M$ ← Initialization($M$,GVTV, GVSV);
For each $v_{miss}$ in $O$
$v_{miss}^{1}$ ← GVTV($M$)
$v_{miss}^{3}$ ← GVSV($M$)
$v_{miss}^{2}$ ← LVTV($M, w$)
$v_{miss}^{4}$ ← LVSV($M$)
$v_{mv}$ ← MVLM($v_{miss}^{1}, v_{miss}^{2}, v_{miss}^{3}, v_{miss}^{4}$)
Impute $v_{mv}$ to $M$;
Return $M$;
```

Benchmarking Multi-View imputation on datasets with various structures

Generating panel datasets with different structures

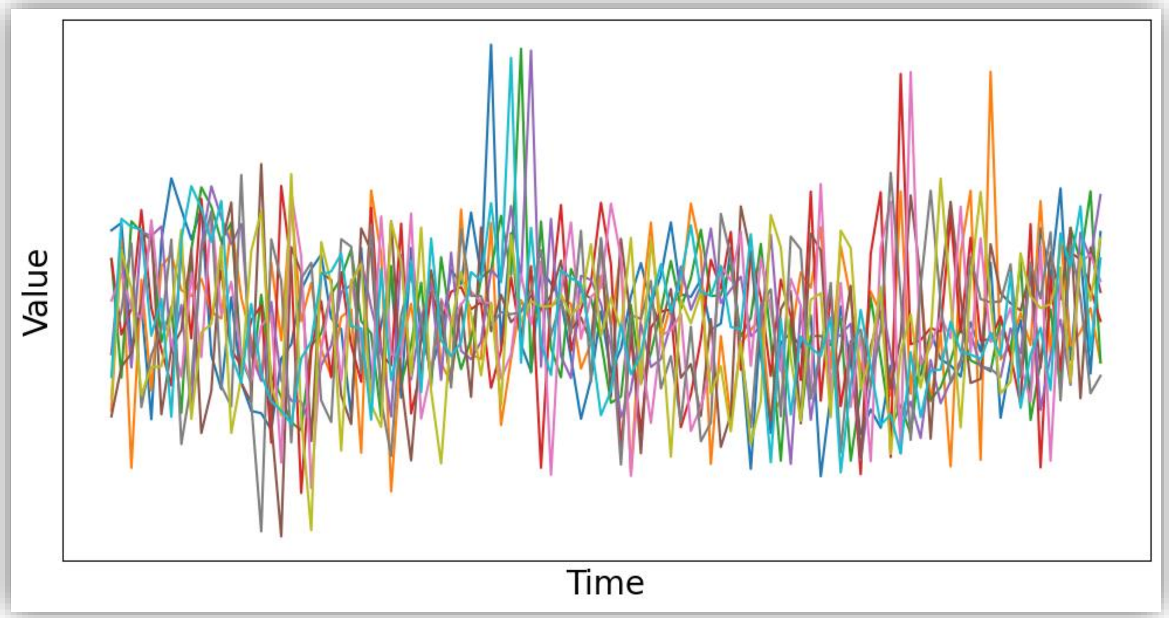


(i) Sine Wave Clusters

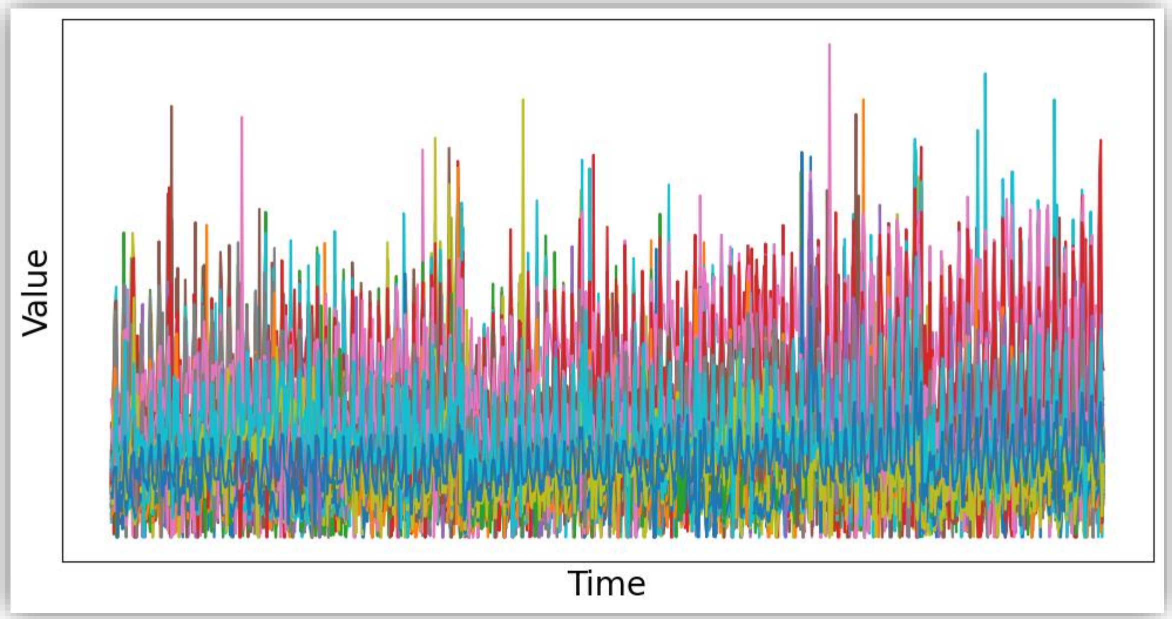


(ii) Clusters of Additive Component Time Series

Generating panel datasets with different structures

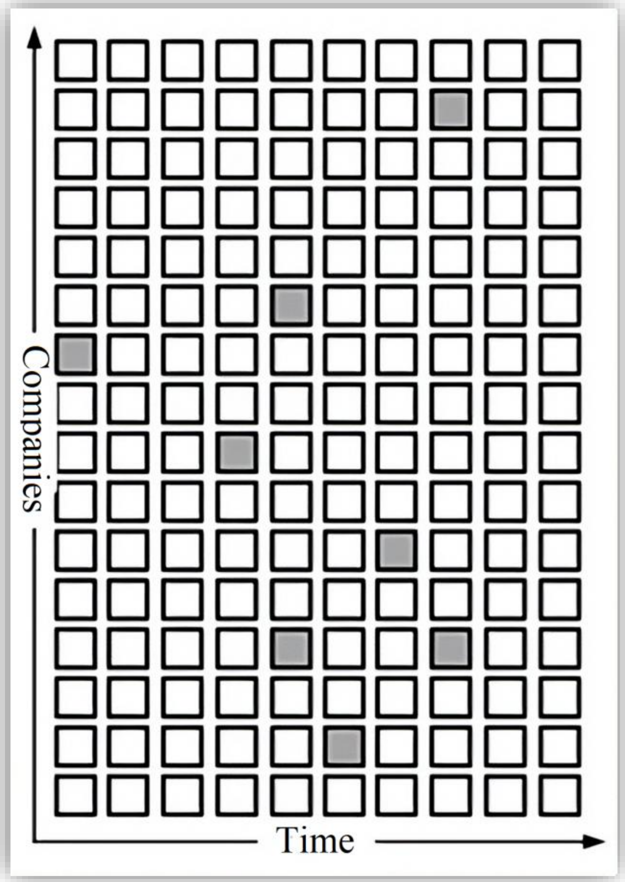


(iii) Clusters of AR(3) series

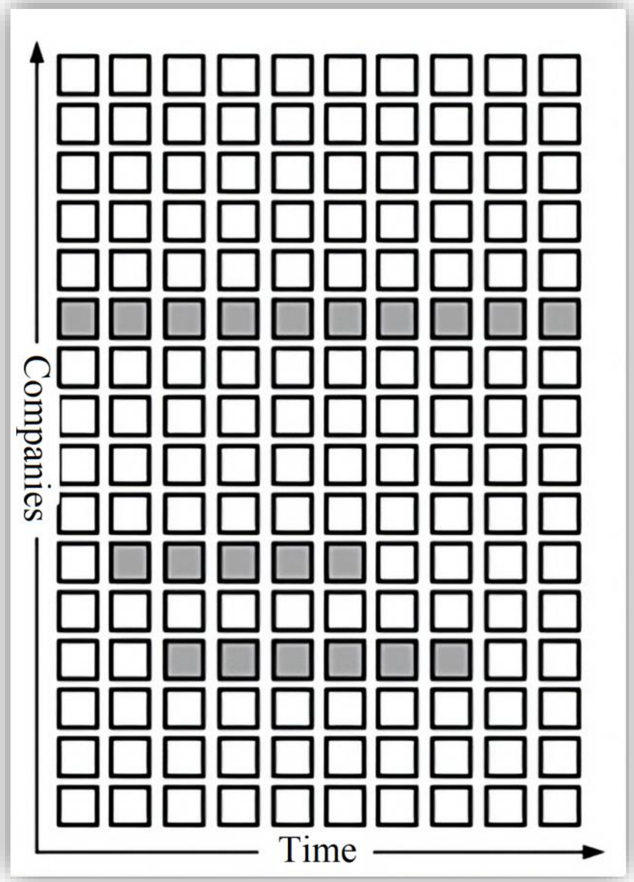


(iv) NN5 (UK ATMs cash withdrawal)

Generating panel datasets with different structure and missing values patterns



Type 1: Random



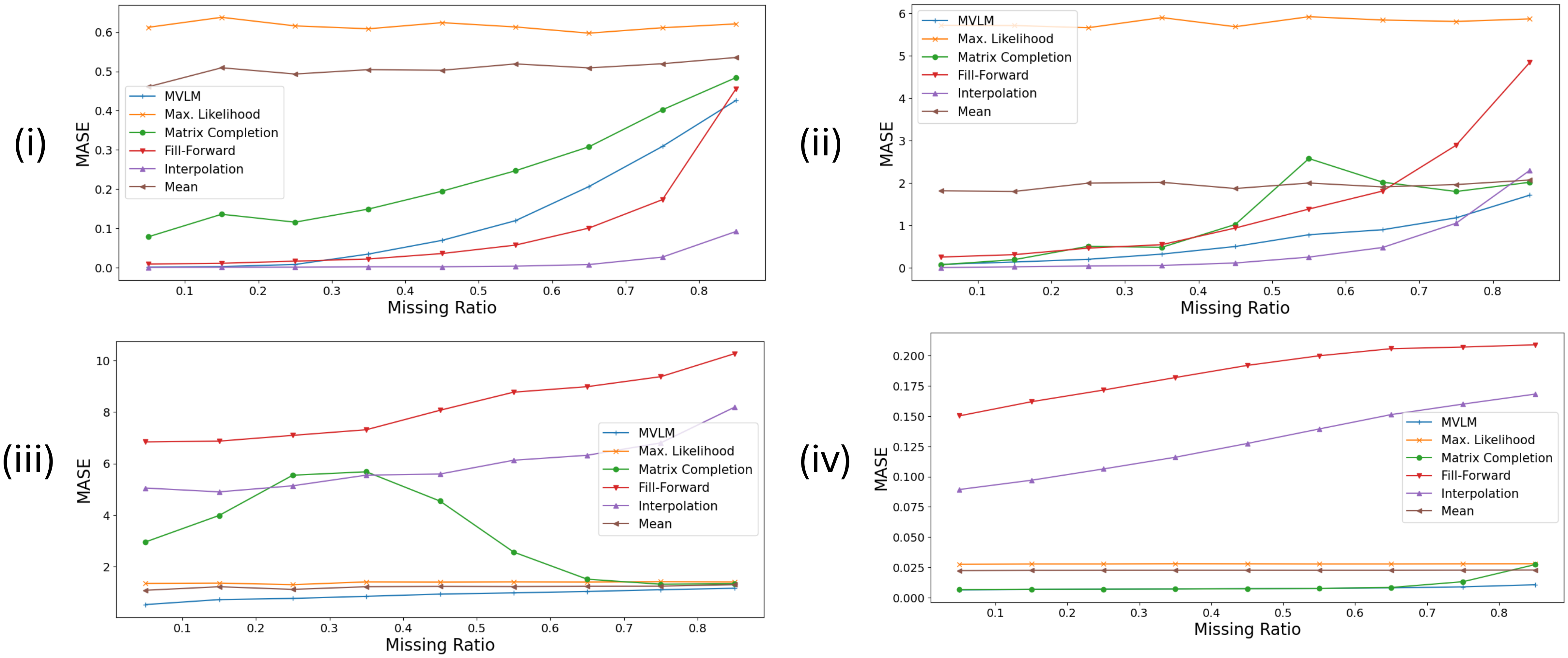
Type 2: Temporal Monotone

Performance metric: Mean Absolute Scaled Error (MASE)

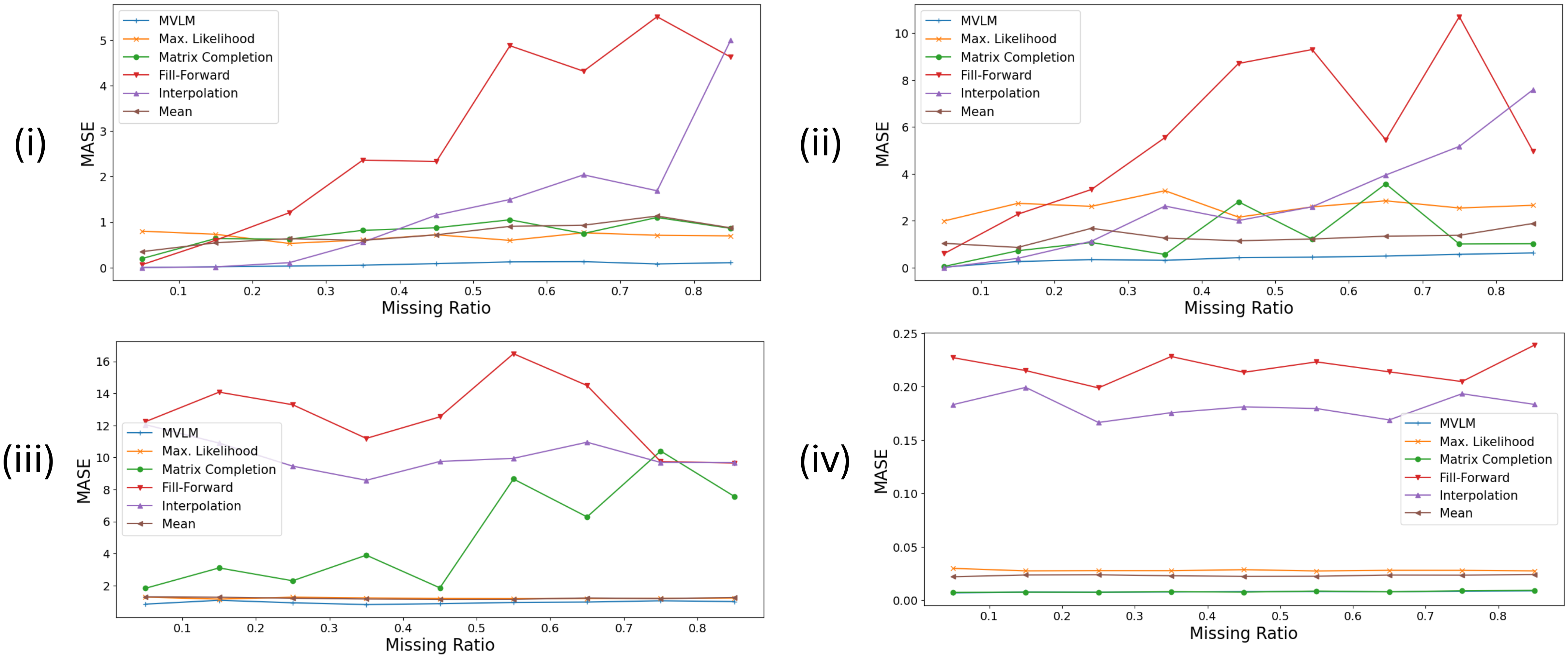
$$\mathrm{MASE} = \mathrm{mean}\left(\frac{|e_j|}{\frac{1}{T-1}\sum_{t=2}^{T} |Y_t - Y_{t-1}|} \right)$$

- e_j is the forecast error for a given period.
- The denominator is the Mean Absolute Error of the one-step *naïve forecast method* on the training set of length *T*.

Multi-View is competitive on random missing values across all datasets



Multi-View is superior on monotone missing values across all datasets



Further work

I. Benchmark of Multi-View Imputation
   a) Extend to datasets with different structures and missingness patterns.
   b) Quantify how much of the structure is well-modelled by the different models within Multi-View.

Further work

I. Benchmark of Multi-View Imputation

   a) Extend to datasets with different structures and missingness patterns.

   b) Quantify how much of the structure is well-modelled by the different models within Multi-View.

II. Apply Multi-View to the problem of fundamental forecasting with alternative features

   a) Explore the structures present datasets of conventional data mixed with alternative data.

   b) Compare the performance between models using Multi-View imputation and other imputation methods

Summary

I. Predicting fundamental values is a challenging task, yet it holds the potential to create highly effective portfolio strategies.

II. Incorporating alternative data could enhance predictions, though such data is often limited and/or not easily accessible.

III. Multi-View imputation is a robust method capable of accurately estimating missing values in structured data.

IV. Future work will focus on assessing the improvements this method brings to the fundamental forecasting problem by addressing the sparsity of alternative data.

Thank you for your attention

References

1. L. Chauhan, J. Alberg, and Z. C. Lipton, “Uncertainty-Aware Lookahead Factor Models for Quantitative Investing,” *arXiv.org*, Jul. 15, 2020. https://arxiv.org/abs/2007.04082 (accessed Jun. 24, 2024).
2. L. Li, J. Zhang, Y. Wang and B. Ran, "Missing Value Imputation for Traffic-Related Time Series Data Based on a Multi-View Learning Method," in IEEE Transactions on Intelligent Transportation Systems, vol. 20, no. 8, pp. 2933-2943, Aug. 2019, doi: 10.1109/TITS.2018.2869768.
3. C. Thomas, “S&P500 Seasonality - Year End Rally Time?,” *topdowncharts*, Oct. 23, 2018. https://www.topdowncharts.com/post/2018/10/24/sp500-seasonality-year-end-rally-time (accessed Jun. 25, 2024).
4. Nonparametric Sparsification of Complex Multiscale Networks - Scientific Figure on ResearchGate. Available from: https://www.researchgate.net/figure/Clustering-of-the-S-P-500-network-a-0-003-into-22-clusters-using-the-spectral_fig4_49967500 [accessed 25 Jun, 2024]