QOLMAT : THE TOOL FOR DATA IMPUTATION

Julien Roussel[1] | Hông-Lan Botterman | Mikaìl Duran[1] | Firas Dakhli[1]
Rima Hajou | David Medernach | Anh Khoa Ngo Ho[1]
Guillaume Saës[1] | Vianney Taquet | Nicolas Brunel[1,2]

1 : Quantmetry, 52, rue d'Anjou, 75008, Paris, France
2 : Laboratoire de Mathématiques et de Modélisation d'Evry, ENSIIE, Université Paris Saclay

Quantmetry
Part of Capgemini Invent

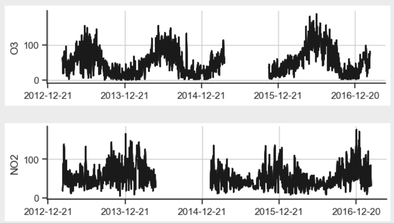
QOLMAT

1 MOTIVATION

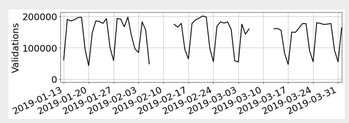
Missing data imputation is a critical step in machine learning pipelines, having a major impact on the final **performance**. Qolmat brings together dispersed imputation methods, allows to compare them on any dataset and puts a focus on the imputation of multivariate time series.

Incomplete data	
Age	IQ score
25	
26	121
29	91
30	
30	110
31	
44	118

Tabular Datas



Multivariate Time Series



Seasonal Time Series

2 OBJECTIVES

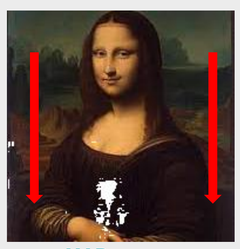
Qolmat is a Python library which can be seamlessly **integrated** into **standard** data processing **pipelines.**

Standardization Qolmat provides a scikit-learn compatible library for many imputation methods, from the more standard to the more advanced. It is now much easier to try state of the art or custom approaches on your missing data.

Benchmarks Qolmat incorporates a Comparator class, which assesses several reconstruction metrics for a variety of imputation methods by cross-validation, on the provided dataset.

3 METHODOLOGY

MCAR
Missing Completely At Random

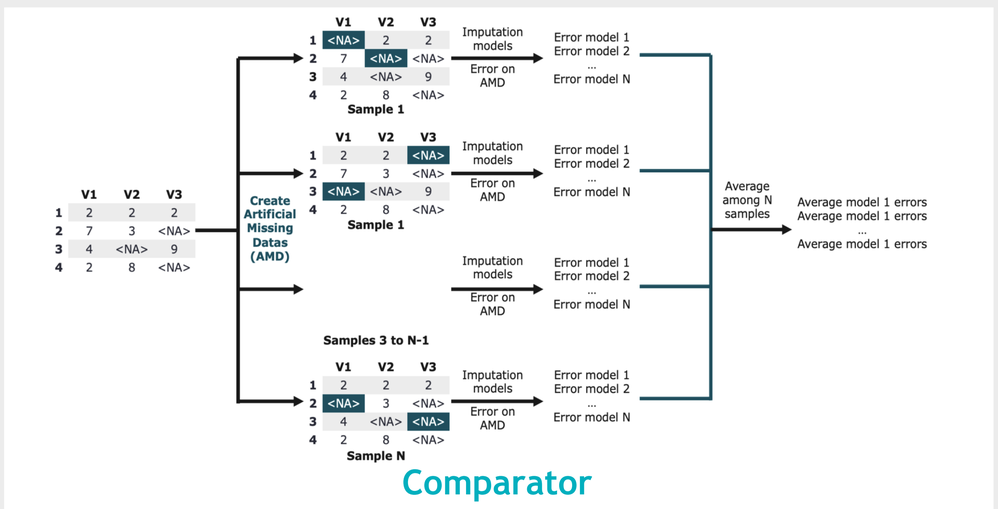


MAR
Missing At Random

MNAR
Missing Not At Random

Missingness distribution Missing data rarely follows a random uniform distribution but is rather auto-correlated (MCAR), correlated to observed values (MAR) or even correlated to unobserved ones (MNAR). Estimating an imputer performance requires to « generate new holes » according to an approximate missingness distribution. Moreover the traditional approaches (mean, interpolation, …) tend to fail for non MCAR missing data.

Performance metrics For each imputation method several performance metrics can be computed, including **elementwise reconstruction errors** such as MAE and **distribution errors** assessing the discrepancy between reference and generated data.



Comparator

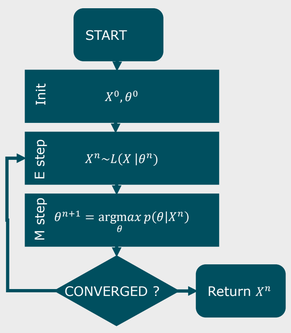
4 IMPUTATION EXAMPLES

Robust PCA (RPCA) imputation[1-3] separates the low-rank part of the data X from the outliers A, and uses the former to impute missing data. Recent developments take into account time correlations and can mimic fluctuations in the data.

$$\min_{\substack{X,A \\ D=X+A}} \|X\|_* + \lambda\|A\|_1$$

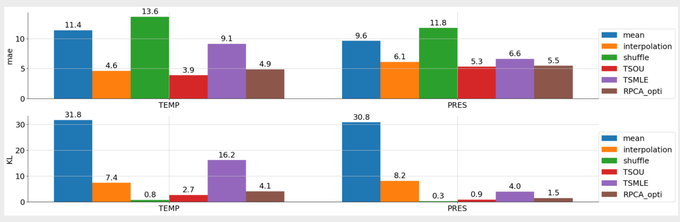
Nuclear norm:
$\|X\|_* = \sum_{\sigma \in Sp(X)} |\sigma|$

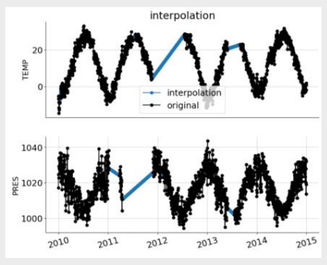
Multivariate Expectation Maximisation imputation for Gaussian or VAR models[4] iteratively estimates the parameters θ describing the distribution of the data X. This imputation takes into account linear correlations between variables or subsequent measures, and can preserve the data variability.

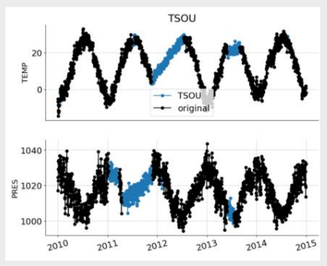


5 BENCHMARK EXAMPLE

When imputing missing data on an athmospheric dataset we can check that the EM imputation we proposed (TSOU) provides a smaller mean absolute error than simpler approaches while achieving a small Kullback-Leibler divergence with the reference data.



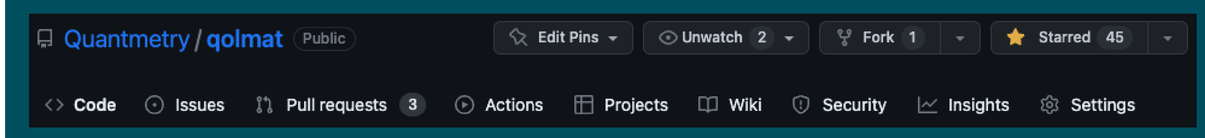




5 OPEN SOURCE LIBRARY

pip install qolmat

https://github.com/Quantmetry/qolmat



GitHub · GitHub Actions · Black · mypy · Read the Docs · pytest 78% · pypi

FUTURE DEVELOPMENTS

- Enrich the hole generation module with new distributions
- Enrich the metrics module with scalable distribution metrics
- Implement deep learning approaches such as diffusion models
- Diagnostic and deal with MNAR data
- Imputation uncertainty using conformal methods

[1] Candès, Emmanuel J., et al. "Robust principal component analysis?." Journal of the ACM (JACM) 58.3 (2011): 1-37
[2] Wang, Xuehui, et al. "An improved robust principal component analysis model for anomalies detection of subway passenger flow." Journal of advanced transportation 2018 (2018)
[3] Chen, Yuxin, et al. "Bridging convex and nonconvex optimization in robust PCA: Noise, outliers, and missing data." arXiv preprint arXiv:2001.05484 (2020)
[4] Borman, Sean. "The expectation maximization algorithm-a short tutorial." Submitted for publication 41 (2004)
[5] Gui, Y., Barber, R. F., & Ma, C.. Conformalized matrix completion. arXiv preprint arXiv:2305.10637 (2023)