



ENTERPRISE RISK MANAGEMENT  
MISSING DATA IMPUTATION  
DATA CHALLENGE

Group I

A TEAM OF 5 DATA SCIENTISTS TO HELP YOU CREATE VALUE FROM YOUR DATA



**Aicha BOKBOT**

Data Scientist

*aicha.bokbot@hec.edu*



**Leonardo NATALE**

Data Scientist

*leonardo.nadale@hec.edu*



**Bruno YZEIRI**

Data Scientist

*bruno.yzeiri@hec.edu*



**Guillaume LE FUR**

Data Scientist

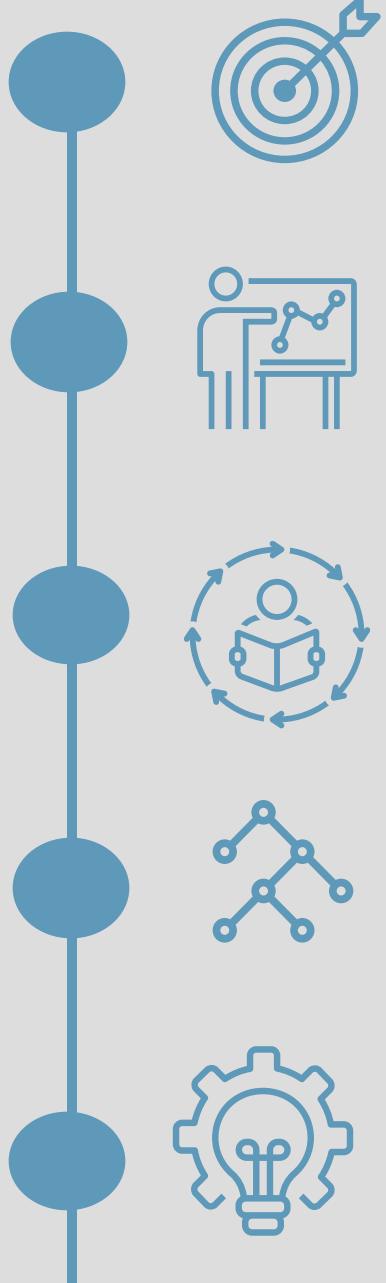
*guillaume.le-fur@hec.edu*



**Constantin VODE**

Data Scientist

*constantin.vode@hec.edu*



## PROJECT PRESENTATION & OBJECTIVES

## DATA ANALYSIS

## PRE-PROCESSING

## MODELS & RESULTS

## SUGGESTIONS, LIMITATIONS & NEXT STEPS



## KEY OBJECTIVES AND EXPECTED BENEFITS



### PROJECT OBJECTIVES

- Establish an optimal, robust method to impute missing data in Financial Time-Series



### BUSINESS IMPACT

- Increase the accuracy and robustness of Risk Management Models
- Increase the performance of predictive models, enterprise-wide



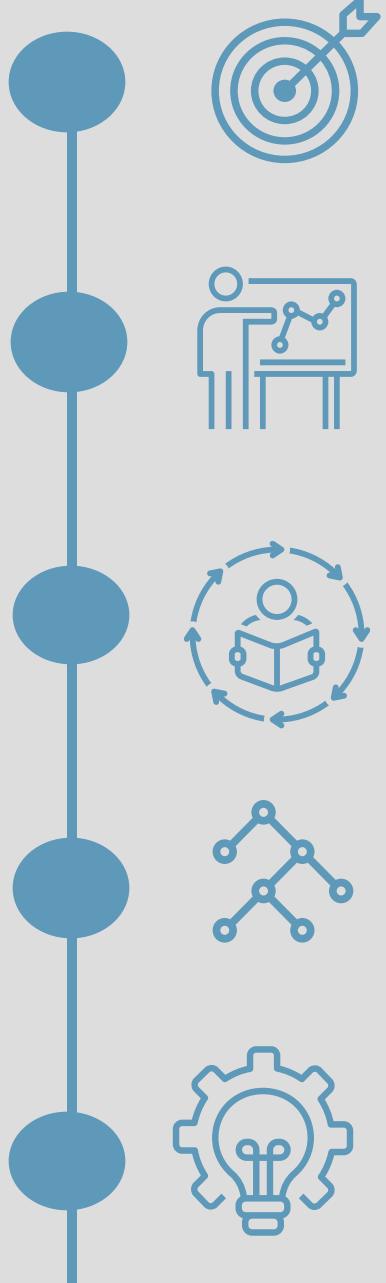
### OUR APPROACH

- Sound, research-backed implementations of state-of-the-art Algorithms



### DATA AVAILABLE

- 1504 Time-Series with daily granularity across 6 different asset classes



PROJECT PRESENTATION & OBJECTIVES

DATA ANALYSIS

PRE-PROCESSING

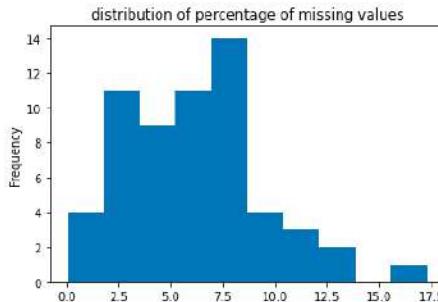
MODELS & RESULTS

SUGGESTIONS, LIMITATIONS & NEXT STEPS

One type fits-all solution for structurally different asset classes

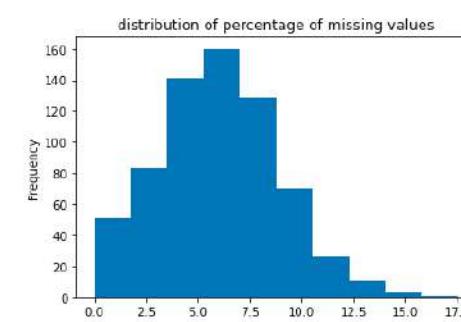
## OVERVIEW OF THE DATA

### Bonds 59 series



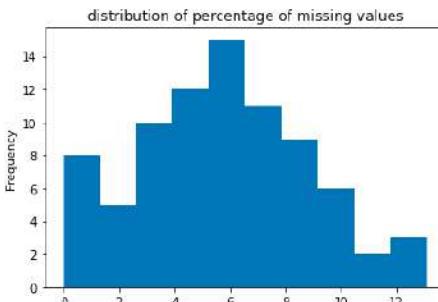
Average correlation:  
62%

### CDS Spreads 675 series



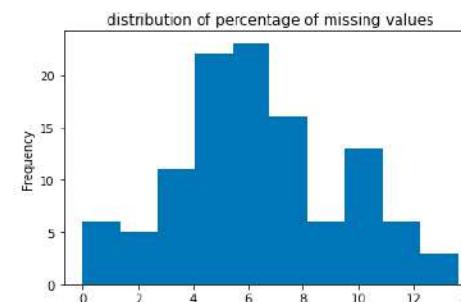
Average correlation:  
58%

### Commodities 81 series



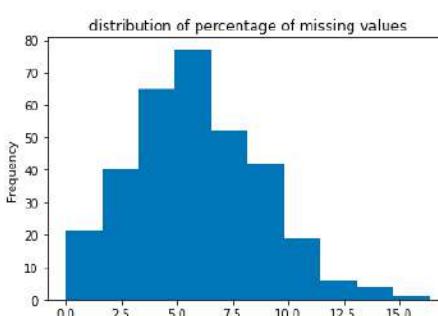
Average correlation:  
57%

### FX Rates 111 series



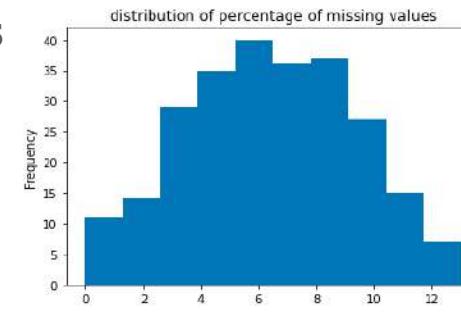
Average correlation:  
7.4%

### Stocks 327 series

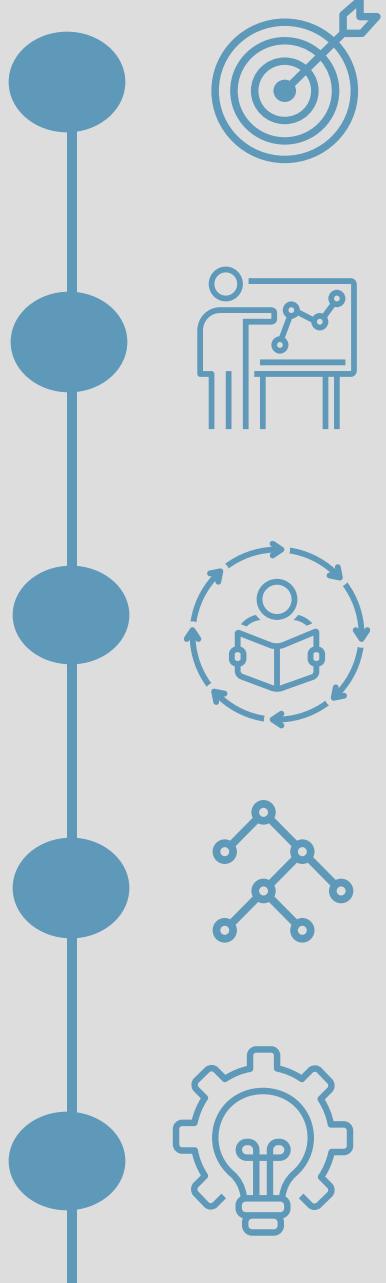


Average correlation:  
30%

### Yield curves 251 series



Average correlation:  
73%



**PROJECT PRESENTATION & OBJECTIVES**

**DATA ANALYSIS**

**PRE-PROCESSING**

**MODELS & RESULTS**

**SUGGESTIONS, LIMITATIONS & NEXT STEPS**

Introducing new missing data allows us to create a validation set

## CREATING A VALIDATION SET

0.0934	0.1097	0.5960	0.3092	0.0774
0.2098	0.9227	0.9253	NaN	0.2140
0.6428	NaN	0.0266	0.3097	0.7496
0.5883	0.6325	NaN	0.0640	0.4914
0.2636	0.6916	0.6133	0.3173	0.3894

Original data

	NaN			NaN
NaN		NaN		
				NaN

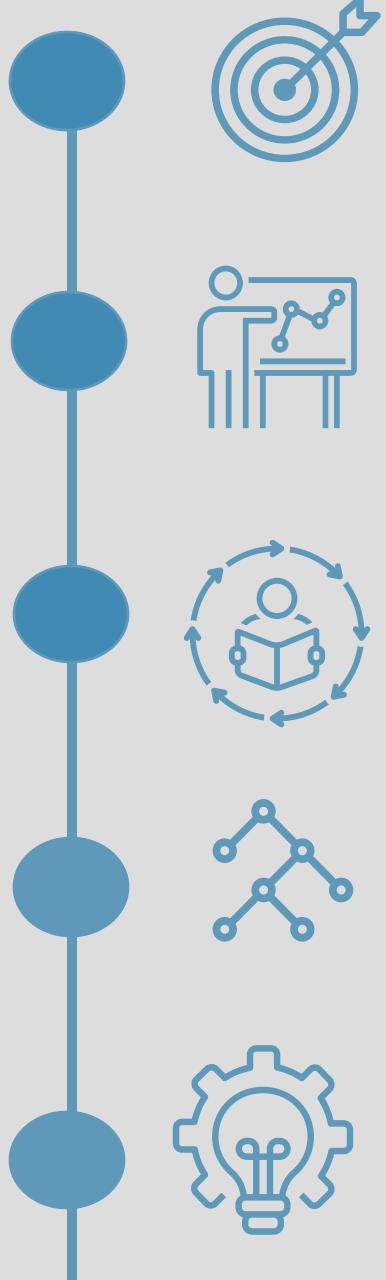
NaN mask



Element-wise  
multiplication

0.0934	0.1097	0.5960	0.3092	0.0774
0.2098	NaN	0.9253	NaN	NaN
0.6428	NaN	0.0266	0.3097	0.7496
NaN	0.6325	NaN	0.0640	0.4914
0.2636	0.6916	0.6133	0.3173	NaN

"Training" data



**PROJECT PRESENTATION & OBJECTIVES**

**DATA ANALYSIS**

**PRE-PROCESSING**

**MODELS & RESULTS**

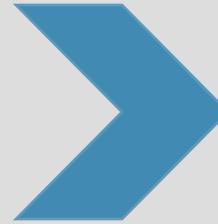
**SUGGESTIONS, LIMITATIONS & NEXT STEPS**

## Our baseline method

## Last Value Carried Forward

0.0934	0.1097	0.5960	0.3092	0.0774
0.2098	0.9227	0.9253	NaN	0.2140
0.6428	NaN	0.0266	0.3097	NaN
0.5883	NaN	NaN	0.0640	0.4914
0.2636	0.6916	0.6133	0.3173	0.3894

Original data



0.0934	0.1097	0.5960	0.3092	0.0774
0.2098	0.9227	0.9253	0.3092	0.2140
0.6428	0.9227	0.0266	0.3097	0.2140
0.5883	0.9227	0.0266	0.0640	0.4914
0.2636	0.6916	0.6133	0.3173	0.3894

Predicted data

- Infer missing values by using the last non missing value.
- Works well with time series that do not vary a lot.
- Can lead to very inaccurate predictions in case of long batches of missing values.

# LSS impute is a recognized missing values imputation technique

## LSS Impute

### Research

*CBN Journal of Applied Statistics* Vol. 10 No. 1 (June, 2019)

51-73

### Imputation of Missing Values in Economic and Financial Time Series Data Using Five Principal Component Analysis Approaches

BIOINFORMATICS

ORIGINAL PAPER

Vol. 21 no. 2 2005, pages 187–198  
doi:10.1093/bioinformatics/bth499



### Missing value estimation for DNA microarray gene expression data: local least squares imputation

Hyunsoo Kim<sup>1</sup>, Gene H. Golub<sup>2</sup> and Haesun Park<sup>1,3,\*</sup>

### Methodology to impute missing values of a given gene

#### Selection of k-nearest genes

Based on Pearson correlation coefficient

#### Local Least Square Imputation

$$\min_{\mathbf{x}} \|\mathbf{A}^T \mathbf{x} - \mathbf{w}\|_2. \quad (2)$$

Then, the missing value  $\alpha$  is estimated as a linear combination of first values of genes

$$\alpha = \mathbf{b}^T \mathbf{x} = \mathbf{b}^T (\mathbf{A}^T)^{\dagger} \mathbf{w}, \quad (3)$$

where  $(\mathbf{A}^T)^{\dagger}$  is the pseudoinverse of  $\mathbf{A}^T$ .

#### Optimization of parameter k

Using a validation set

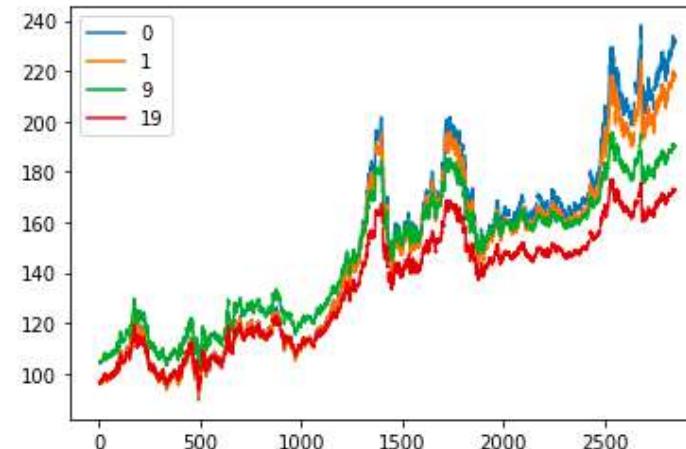
**LLS Impute**

Methodology to impute missing values of time-series “0” of type BONDS

**Selection of k-nearest time-series**

Pearson coefficients are computed between “0” and other bonds (rows with NA dropped)

Let  $k=3$ . The k-nearest time-series of “0” are “1”, “9” and “19”

**Local Least Square Imputation**

	0	1	9	19	
53	NaN	98.2305	106.9925	98.8760	
54	98.0455	98.2885	107.0710	98.9570	
55	97.6295	NaN	106.7160	98.6410	
56	NaN	98.0175	106.8045	NaN	
57	98.7190	98.9550	107.6640	99.4950	
58	99.1200	99.3565	108.0010	99.8055	
59	99.0270	99.2550	107.9245	99.7240	
60	98.3235	98.5610	107.2780	99.1405	
61	NaN	98.0715	106.8720	98.7580	

NA imputed by linear interpolation on column “19”

w

A<sup>T</sup>

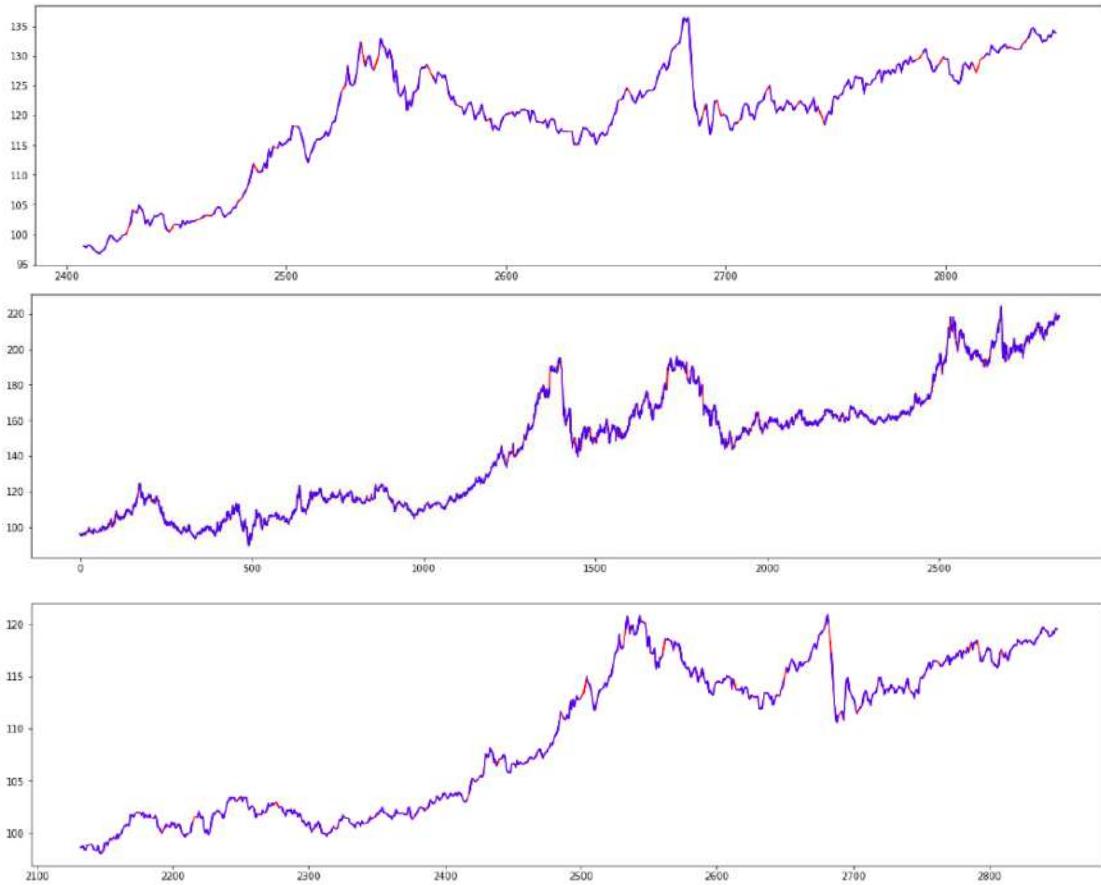
b

$$\text{missing\_value} = b^T (A^T)^+ w = 97.501$$

LSS Impute shows satisfying results but no real break-through

## LLS Impute

LLS Impute on a selection of time series  
(in red imputation of missing values)



Scores

submission_name	metric_name	score
Group1_baseline	nrmse	0,096079856
Group1_baseline	cov	4,005684866
Group1_interpolation	nrmse	0,063992136
Group1_interpolation	cov	1,653697404
Group1_LLS	nrmse	0,067356061
Group1_LLS	cov	1,653691329

LSS Impute shows satisfying score, but it does not outperform a simple imputation using linear interpolation

# A robust method that couples Principal Component Regression with Expectation-Maximization

## Bayesian PCA

### Research

*CBN Journal of Applied Statistics Vol. 10 No. 1 (June, 2019)*

51-73

### Imputation of Missing Values in Economic and Financial Time Series Data Using Five Principal Component Analysis Approaches

Multiple imputation for continuous variables using a Bayesian principal component analysis

VINCENT AUDIGIER<sup>1</sup>, FRANÇOIS HUSSON<sup>2</sup> AND JULIE JOSSE<sup>2</sup>

Applied Mathematics Department, Agrocampus Ouest, 65 rue de Saint-Brieuc, F-35042 RENNES Cedex, France

[audigier@agrocampus-ouest.fr](mailto:audigier@agrocampus-ouest.fr)  
[husson@agrocampus-ouest.fr](mailto:husson@agrocampus-ouest.fr)  
[josse@agrocampus-ouest.fr](mailto:josse@agrocampus-ouest.fr)

*Computationally Unfeasible*

### Methodology

#### Init

- Calculate the Matrix of means
- Center the data
- Estimate initial parameters with PCA

#### Loop

- Impute the centered matrix with a random imputation
- Add back the matrix of means
- Calculate the new matrix of means
- Evaluate posterior parameters
- Draw new parameters from the posterior distribution

$$\text{draw } \tilde{x}_{ij}^{[\ell]} \text{ from } \mathcal{N} \left( \hat{x}_{ij}^{rPCA[\ell]}, \frac{\hat{\sigma}^{2[\ell]} \sum_{\ell=0}^{\ell} \hat{\phi}_{ij}^{[\ell]}}{\min(n-1, p)} \right).$$

#### Return

- Imputed Matrix
- RMSE (if validation data is provided)
- We stop if improvement in RMSE is < 1e-6

An efficient way to impute the missing values based on modeling the time series with an autoregressive (AR) model, convenient to model log-prices and log-volumes in financial data

ImputeFin

Research

# Parameter Estimation of Heavy-Tailed AR Model With Missing Data Via Stochastic EM

Junyan Liu , Sandeep Kumar, and Daniel P. Palomar , *Fellow, IEEE*

Methodology to impute missing values

## The idea

Each point as a noisy linear combination of the previous steps

$$y_t = \varphi_0 + \sum_{i=1}^p \varphi_i y_{t-i} + \varepsilon_t,$$

## The solution in context

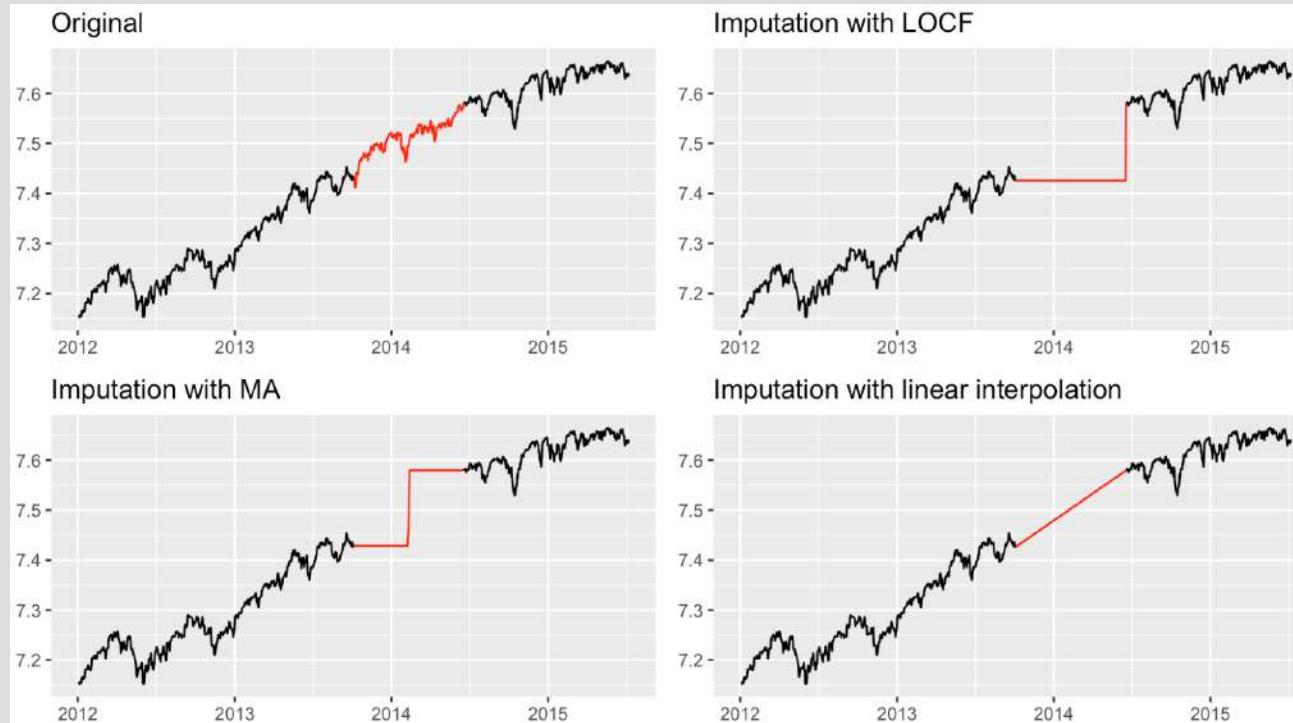
- Parameters are estimated using Stochastic Expectation-Maximization
- Heavy-tailedness of the errors ensures robustness to outliers
- Convergence is proven
- Computationally cheap

## Imputation

Values are imputed by drawing samples from the conditional distribution of the missing values.

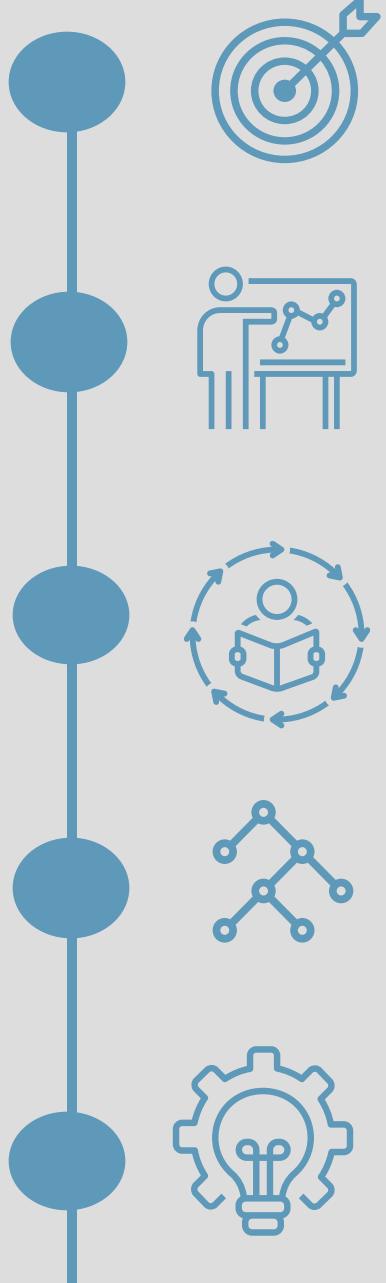
Compared to other imputation techniques, our results look extremely realistic

## ImputeFin



## Our Prediction





**PROJECT PRESENTATION & OBJECTIVES**

**DATA ANALYSIS**

**PRE-PROCESSING**

**MODELS & RESULTS**

**SUGGESTIONS, LIMITATIONS & NEXT STEPS**



There is a trade off between exploring new solutions or building on top of what has already been implemented

## OUR SUGGESTIONS

### Suggestion 1

- Invest more time in fine tuning (e.g. assuming that the residuals follow a Student's t distribution) and make our implementations more efficient.

### Suggestion 2

- Build asset-class specific solutions.

### Suggestion 3

- Conduct more thorough literature review (e.g. exploring advancements in Neural Controlled Differential Equations for Irregular Time Series).



## EXECUTIVE SUMMARY

### KEY OBJECTIVE



**Predict as accurately as possible the missing values of time series for different asset classes.**

### OUR APPROACH



**Apply state of the art methods extracted from a literature review.**

### MODEL SELECTED



**Impute Fin is the best performing model we tried.**

### RESULTS



**Compared to simple interpolation techniques, our results neither look artificial nor destroy the time series statistics**