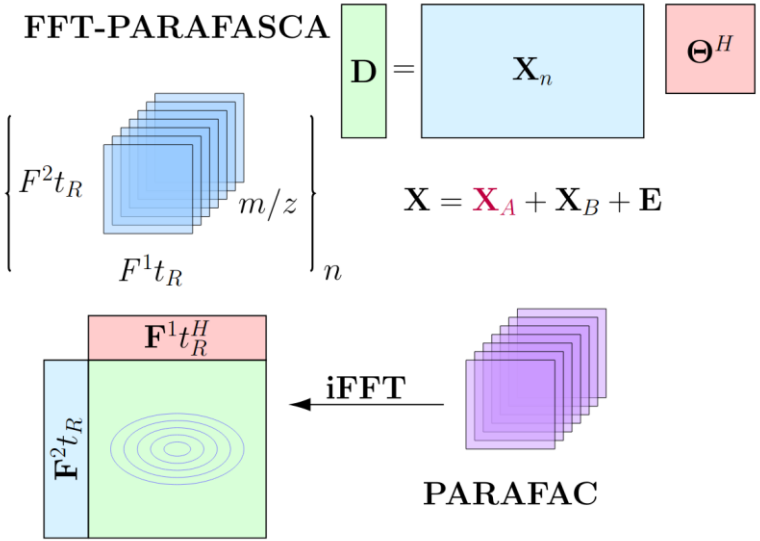
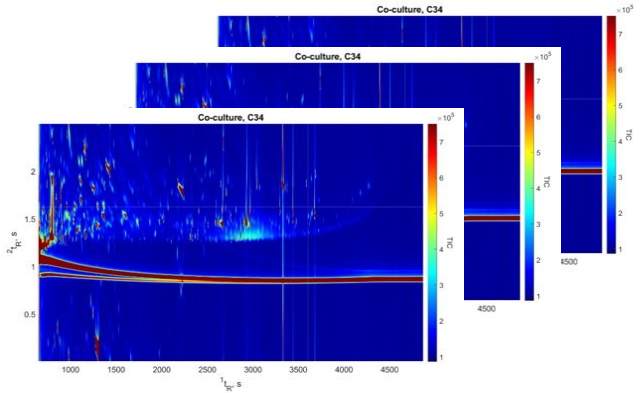
Comparative analysis of comprehensive 2D gas chromatography-time-of-flight mass spectrometry data in time and frequency domains

Michael Sorochan Armstrong [1]
José Camacho [1]
Court Sandau [2]



1-Computational Data Science Lab, Department of Signal Theory, Telematics and Communications, University of Granada, Spain
2-Chemistry Matters, Calgary, Alberta Canada

Raw Data



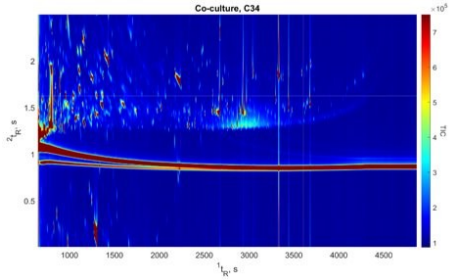
Peak Table

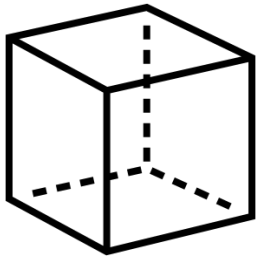
PKS	var1	var2	var3	var4	var5
cls1					
cls2					
cls1					
cls2					
cls1					

MINISTERIO DE CIENCIA, INNOVACIÓN Y UNIVERSIDADES

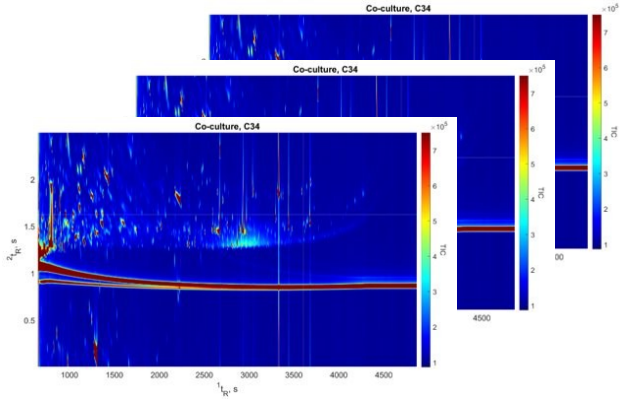
Proyecto MAHOD: 101106986
MSCA Information Science and Engineering

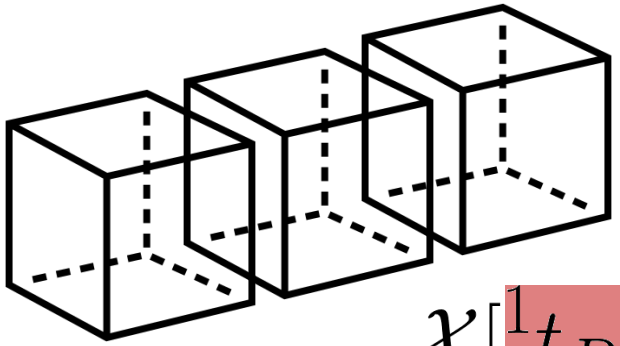
Data Structure

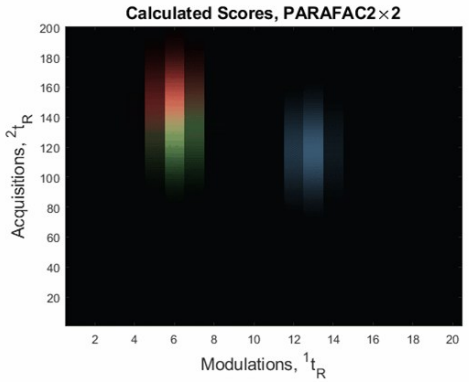




$$\mathcal{X}[{}^1t_R, {}^2t_R, m/z]$$





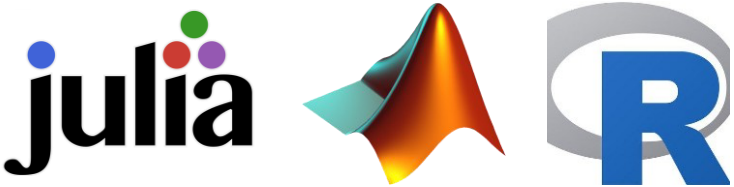


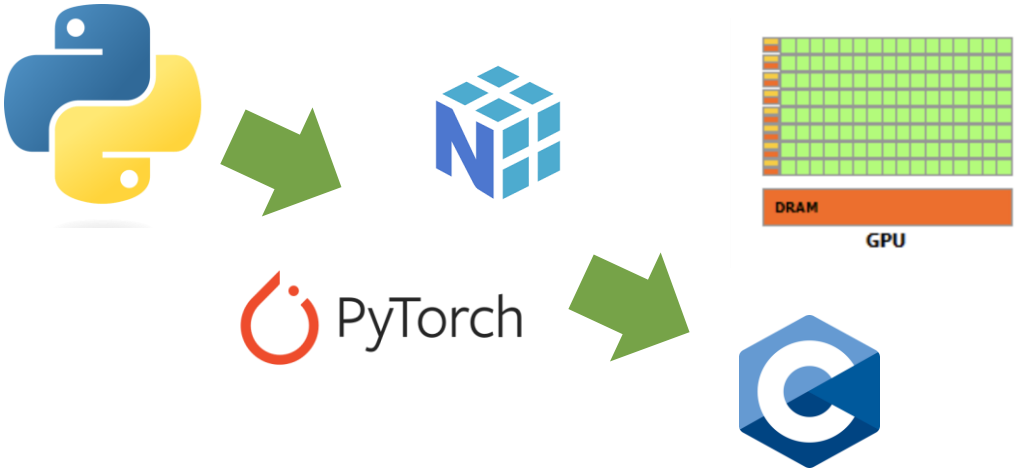
$$\mathcal{X}[{}^1t_R, {}^2t_R, m/z, smpl]$$

Proyecto MAHOD: 101106986
MSCA Information Science and Engineering

Why Python?

Python	Rating
Open Source	**yes**
Widespread	**10/10**
Ease of use	**5-9/10 ****
Efficiency	**5-9/10 ****



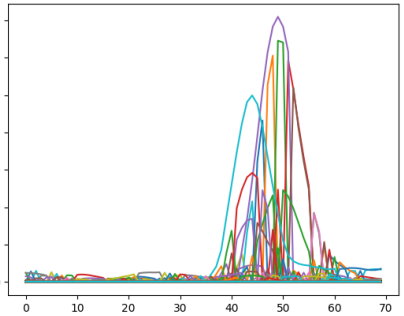


MINISTERIO DE CIENCIA, INNOVACIÓN Y UNIVERSIDADES
Proyecto MAHOD: 101106986
MSCA Information Science and Engineering

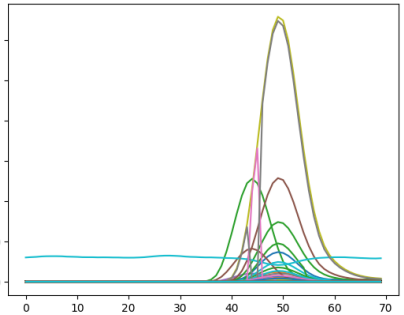
Mass alignment

At resolution m/z = +/- 0.001

Before alignment



After alignment

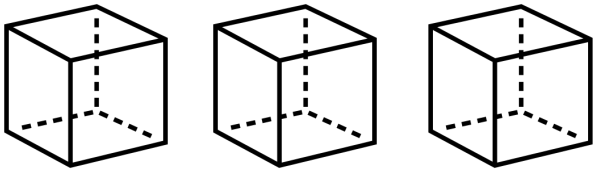


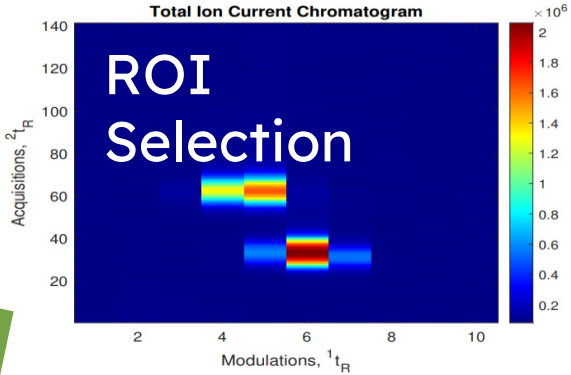
Proyecto MAHOD: 101106986
MSCA Information Science and Engineering

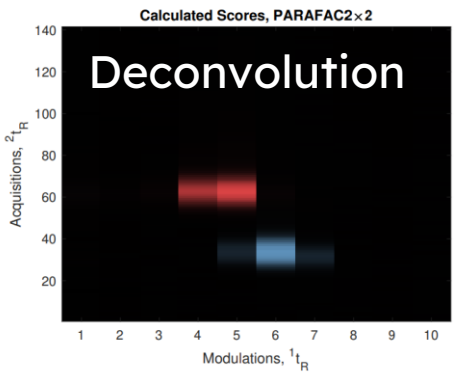
What's our proposed solution?

Feature engineering

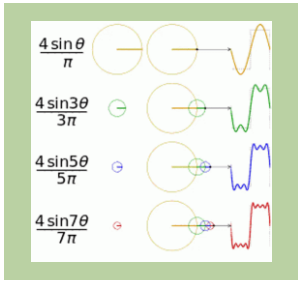
Load



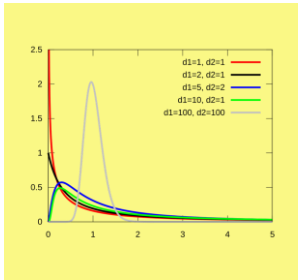




FFT



GLM



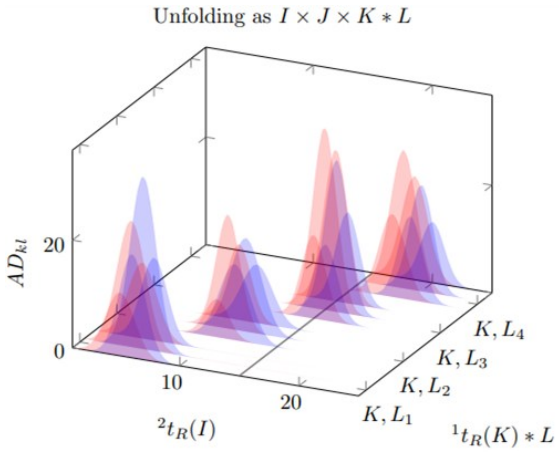
Factorization

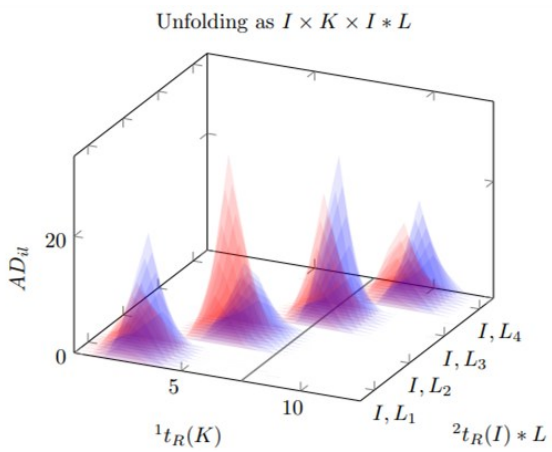
$$\mathbf{X} = \mathbf{X}_A + \mathbf{X}_B + \mathbf{E}$$

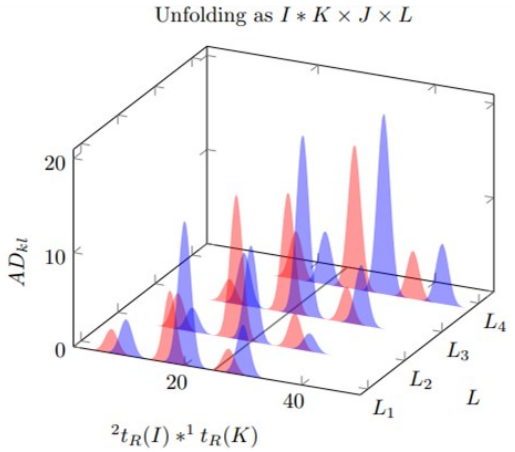
MINISTERIO DE CIENCIA, INNOVACIÓN Y UNIVERSIDADES

Proyecto MAHOD: 101106986
MSCA Information Science and Engineering

Deconvolution







ELSEVIER
Analytica Chimica Acta
Volume 1249, 8 April 2023, 340909

PARAFAC2×N: Coupled decomposition of multi-modal data with drift in N modes

Michael D. Sorochan Armstrong [a], Jesper Løve Hinrich [b], A. Paulina de la Mata [a], James J. Harynuk [a]

Show more

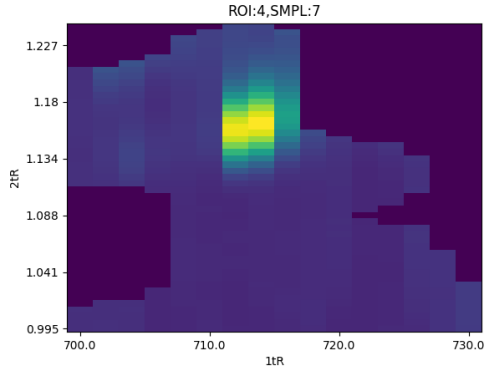
+ Add to Mendeley Share Cite

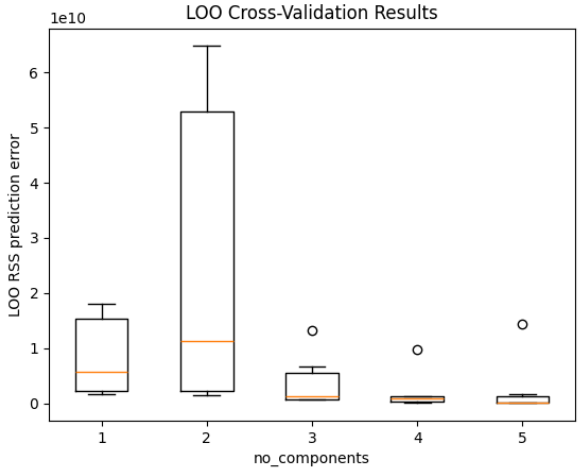
MINISTERIO DE CIENCIA, INNOVACIÓN Y UNIVERSIDADES

Proyecto MAHOD: 101106986
MSCA Information Science and Engineering

Component Number Selection (LOO)

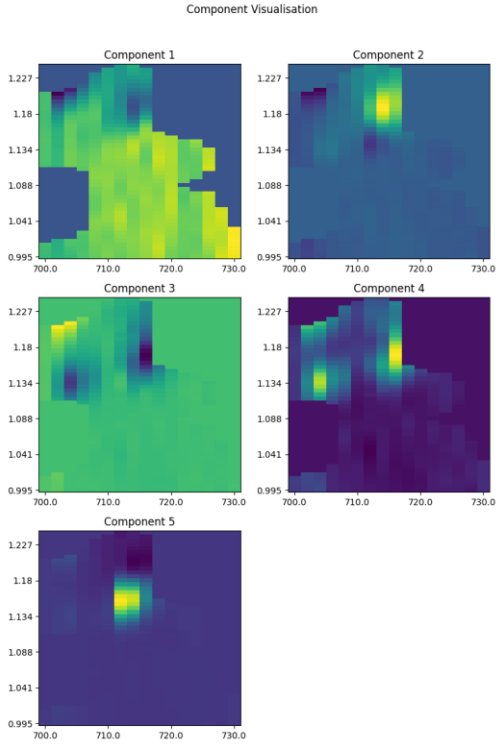
Figure 10: One iteration of a LOO cross-validation routine. The efficacy of the model with a particular component number is tested on the sum of squared residuals on the test set.

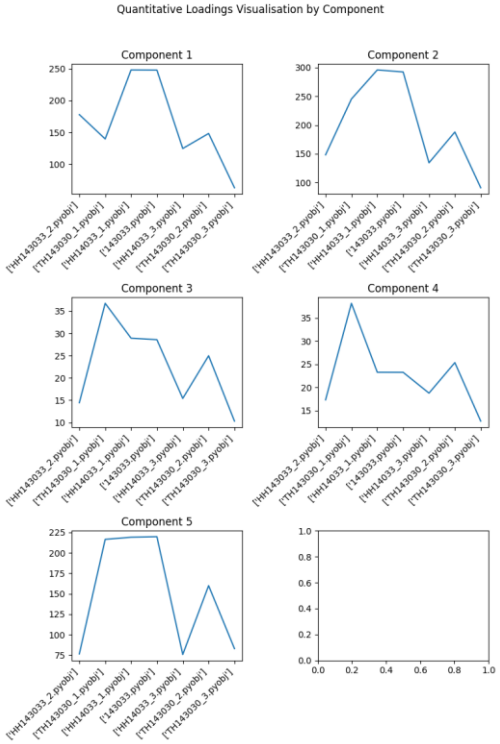


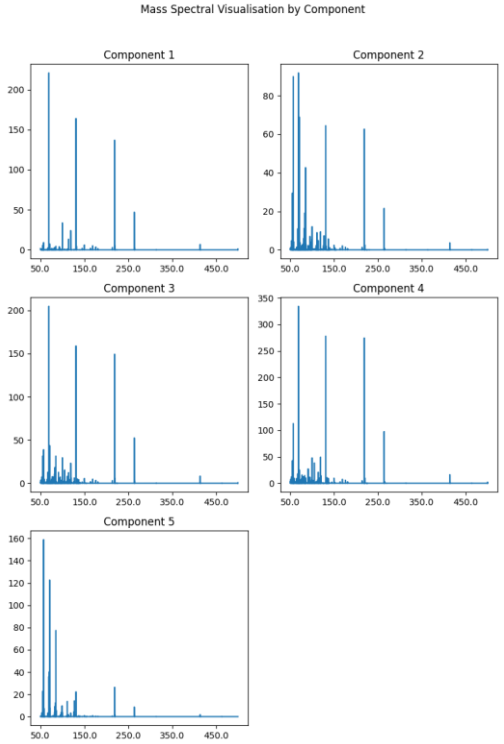


Proyecto MAHOD: 101106986
MSCA Information Science and Engineering

Deconvolution Results







MINISTERIO DE CIENCIA, INNOVACIÓN Y UNIVERSIDADES

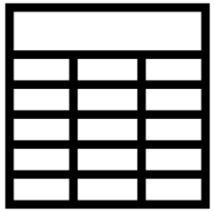
Proyecto MAHOD: 101106986
MSCA Information Science and Engineering

Peak Table Creation

raw2table.py

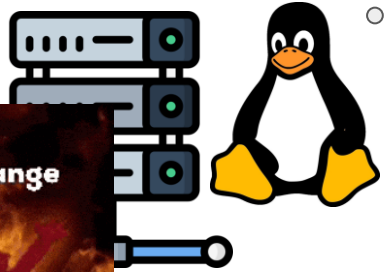
docker

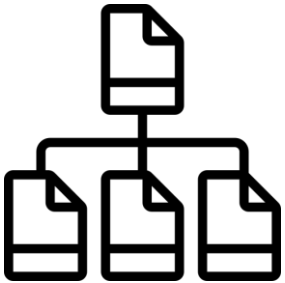
domdfcoding / pynist



pandas

Peak Table





File Structure

MINISTERIO DE CIENCIA, INNOVACIÓN Y UNIVERSIDADES

Proyecto MAHOD: 101106986
MSCA Information Science and Engineering

MS Search Results (Reverse)

roi1_1	roi1_2	roi1_3
Perfluorotributylamine (687)	Perfluorotributylamine (570)	Perfluorotributylamine (534)
Perfluoro(dibutylmethylamine) (679)	Perfluoro(dibutylmethylamine) (563)	Perfluoro(dibutylmethylamine) (507)
Tris(trifluoromethyl)phosphine (562)	Tris(trifluoromethyl)phosphine (439)	3-[4-Chromanyl]-2-oxazolidinone (431)
Nonafluoro-1-bromobutane (528)	Pentadecafluorooctyl butanesulfonate (431)	Chloromethyl-phenylethoxy-silacyclohexane (428)
Pentadecafluorooctyl butanesulfonate (524)	Nonafluoro-1-bromobutane (430)	3-Phenylpropionic acid, morpholide (419)

	roi001_1	roi001_2	roi001_3	roi001_4	roi001_5
TH143030_1.cdf	94.74	96.83	23.53	7.02	110.66
HH14033_3.cdf	89.73	159.22	29.13	4.38	64.36
HH143033_2.cdf	89.01	139.61	32.63	56.03	166.69
143033.cdf	89.01	139.61	32.63	56.03	166.69
HH14033_1.cdf	76.33	84.73	16.76	7.61	80.48
TH143030_2.cdf	102.50	130.82	24.66	6.51	66.17
TH143030_3.cdf	44.88	49.93	10.12	1.25	29.95
fvals	0.05	1.01	1.99	191.58	4.39
pvals	0.84	0.42	0.29	0.005	0.17

Quantitative Peak Table w/ 1-way ANOVA results

MINISTERIO DE CIENCIA, INNOVACIÓN Y UNIVERSIDADES

Proyecto MAHOD: 101106986
MSCA Information Science and Engineering

Better deconvolution w/ FFT

Computational Data Science (CoDaS) Lab
UNIVERSIDAD DE GRANADA

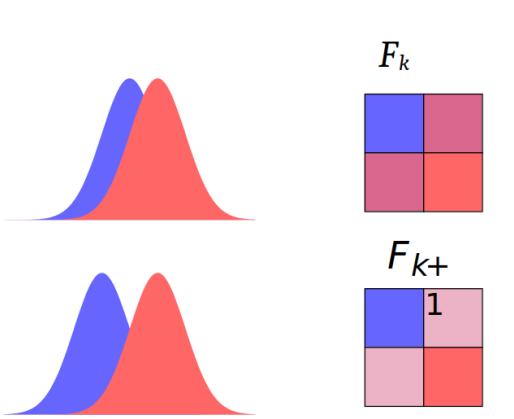
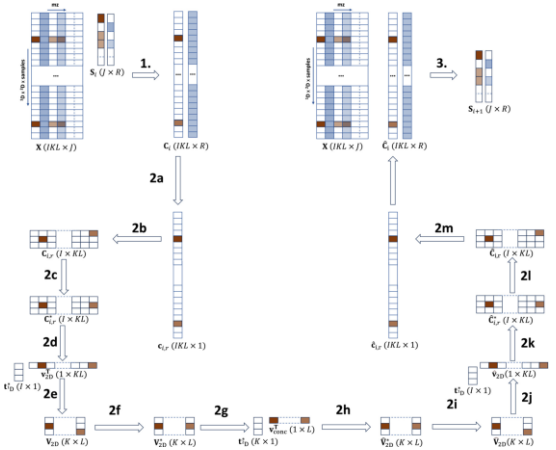
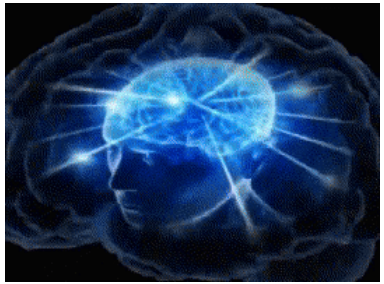


Figure 1: The effect of independent peak drift on the coupling matrix, F



Paul-Albert Schneide

Shift-Invariant Multi-Linearity



Journal of CHEMOMETRICS

RESEARCH ARTICLE | Full Access

Unlocking New Capabilities in the Analysis of GC × GC-TOFMS Data With Shift-Invariant Multi-Linearity

Paul-Albert Schneide, Michael Sorochan Armstrong, Neal Gallagher, Rasmus Bro

First published: 05 December 2024 | https://doi.org/10.1002/cem.3623

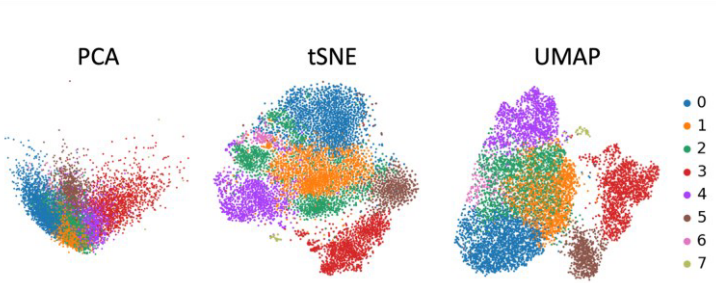
Funding: This work was supported by BASF SE. Michael Sorochan Armstrong is funded by an MSCA grant (number 101106986).

MINISTERIO DE CIENCIA, INNOVACIÓN Y UNIVERSIDADES

Proyecto MAHOD: 101106986
MSCA Information Science and Engineering

Topological Data Analysis

Computational Data Science (CoDaS) Lab
UNIVERSIDAD DE GRANADA



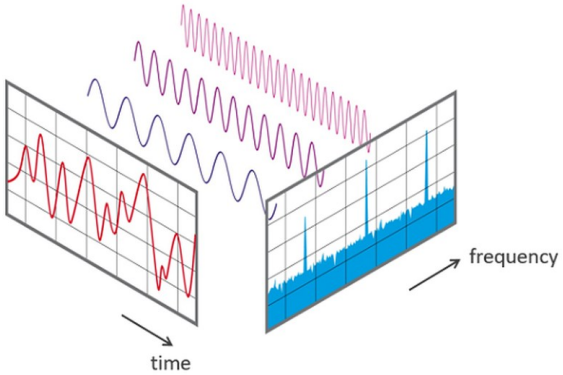
Manifold learning

- Slow
- More meta-parameters, more problems
-Non-unique solutions

- Wavelet selection

Wavelet transform

$$F(\tau, s) = \frac{1}{\sqrt{|s|}} \int_{-\infty}^{+\infty} f(t)\, \psi^* \left(\frac{t-\tau}{s} \right) dt$$



Fourier analysis

- Algebra is a little different

MINISTERIO DE CIENCIA, INNOVACIÓN Y UNIVERSIDADES

Proyecto MAHOD: 101106986
MSCA Information Science and Engineering

FFT-ASCA

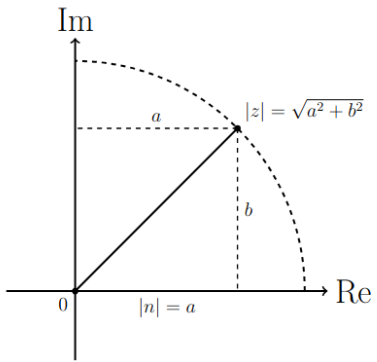
Journal of CHEMOMETRICS

RESEARCH ARTICLE | Full Access

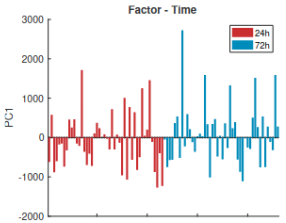
An Alignment-Agnostic Methodology for the Analysis of Designed Separations Data

Michael Sorochan Armstrong | José Camacho

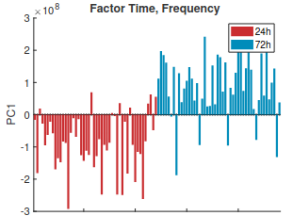
First published: 25 January 2025 | https://doi.org/10.1002/cem.70002



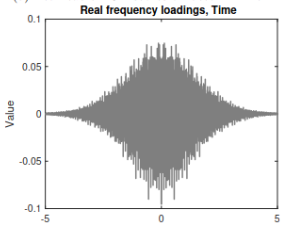
FFT, mean-centered	SumSq	PercSumSq	df	MeanSq	*F*	*p*-value
Mean	5.84e+14	5.09	1	5.84e+14	—	—
Time-1	1.54e+15	13.4	1	1.54e+15	19.5	**0.000999**
Treatment-2	7.79e+14	6.78	2	3.89e+14	4.92	**0.00599**
Sex-3	8.7e+14	7.57	1	8.7e+14	11	**0.00599**
Order-4	7.13e+14	6.2	7	1.02e+14	1.29	0.25
Interaction 1-2	4.51e+14	3.93	2	2.26e+14	2.85	0.0569
Residuals	6.49e+15	56.5	82	7.92e+13	—	—
Total	1.15e+16	100	96	1.2e+14	—	—



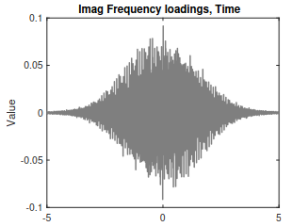
(a) Peak table PCA scores of factor "Time".



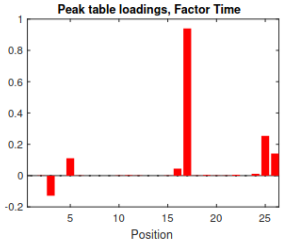
(b) Real-valued PCA scores for factor "Time" from FFT transformed data. Mean-centred.



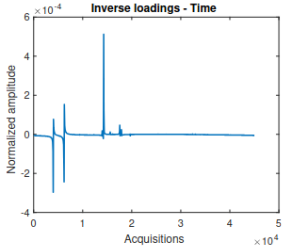
(c) Loadings for factor "Time" from the peak table representation of the data. 26 targeted metabolites were included in the analysis.



(d) Complex-valued loadings in the frequency domain for factor "Time".



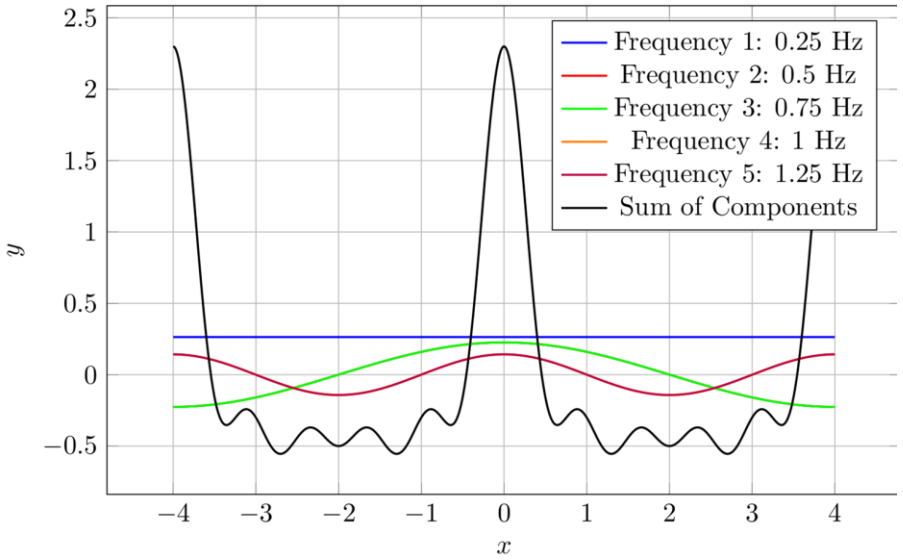
(e) Real-valued loadings in the time domain for factor "Time".



(f) Real-valued loadings in the time domain for factor "Time".

MINISTERIO DE CIENCIA, INNOVACIÓN Y UNIVERSIDADES

Proyecto MAHOD: 101106986
MSCA Information Science and Engineering



Amplitude - real part

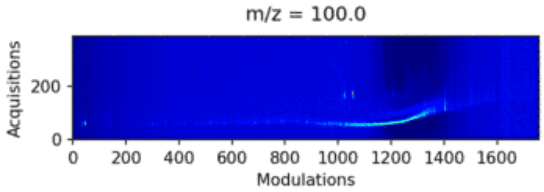
Shift - imaginary part

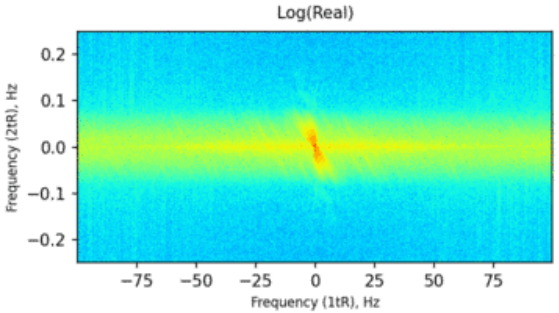
Both - complex domain

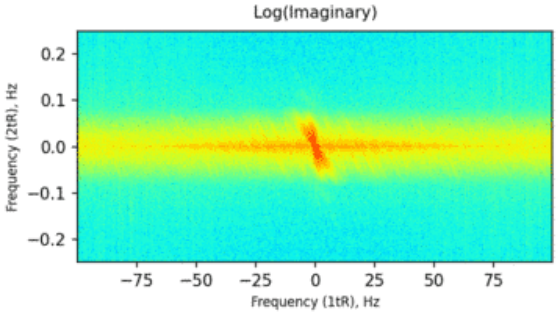
MINISTERIO DE CIENCIA, INNOVACIÓN Y UNIVERSIDADES

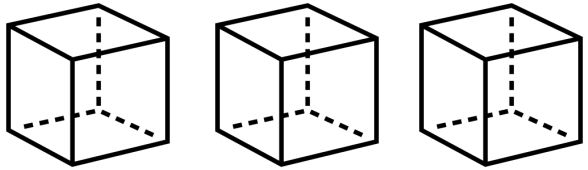
Proyecto MAHOD: 101106986

MSCA Information Science and Engineering

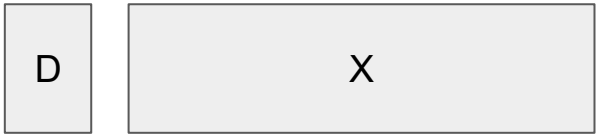








$$\mathcal{F}\left({}^{1}t_R, {}^{2}t_R\right) | \; m/z$$



$$\mathbf{X} = \mathbf{X}_A + \mathbf{X}_B + \mathbf{E}$$

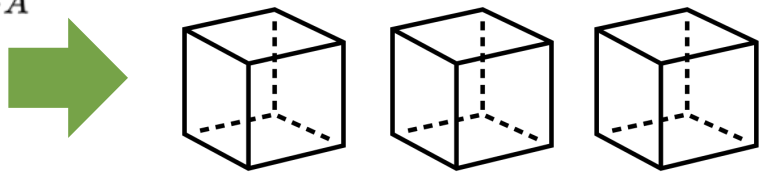
MINISTERIO DE CIENCIA, INNOVACIÓN Y UNIVERSIDADES

Proyecto MAHOD: 101106986

MSCA Information Science and Engineering

Tensor algebra - array indexing

$\mathbf{X}_A$



TensorLy
Tensor Learning in Python

$$\mathbf{X}_A = \mathbf{A}_{1t_R} \times \left(\mathbf{B}_{2t_R} \otimes \mathbf{C}_{m/z} \otimes \mathbf{D}_{smpls}\right)^H$$

$$\mathbf{X}_{[n]} = \mathbf{U}^{(n)}\mathbf{G}_{[n]}\left(\mathbf{U}^{(1)} \otimes \cdots \mathbf{U}^{(n-1)} \otimes \mathbf{U}^{(n+1)} \otimes \cdots \otimes \mathbf{U}^{(N)}\right)^T$$

Python

$$\mathbf{X}_{[n]} = \mathbf{U}^{(n)}\mathbf{G}_{[n]}\left(\mathbf{U}^{(N)} \otimes \cdots \mathbf{U}^{(n+1)} \otimes \mathbf{U}^{(n-1)} \otimes \cdots \otimes \mathbf{U}^{(1)}\right)^T$$

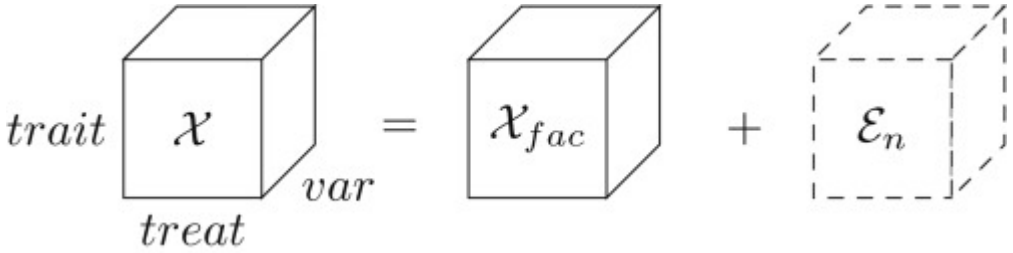
MATLAB, R

MINISTERIO DE CIENCIA, INNOVACIÓN Y UNIVERSIDADES

Proyecto MAHOD: 101106986
MSCA Information Science and Engineering

http://jeankossaifi.com/blog/unfolding.html

What else can we do with PARAFASCA?



Project the error tensor for uncertainty estimation

$$Z_{|I} = F_M \otimes F_L \otimes F_K$$
$$^{I}X_A^* \in \mathbf{R}^{I \times K * L * M}$$
$$F_I = ^{I} X_A^* Z_{|I} (Z_{|I}^T Z_{|I})^{-1}$$

Replicate analysis for uncertainty estimation in PARAFAC and PARAFASCA analyses of factorial metabolomics data

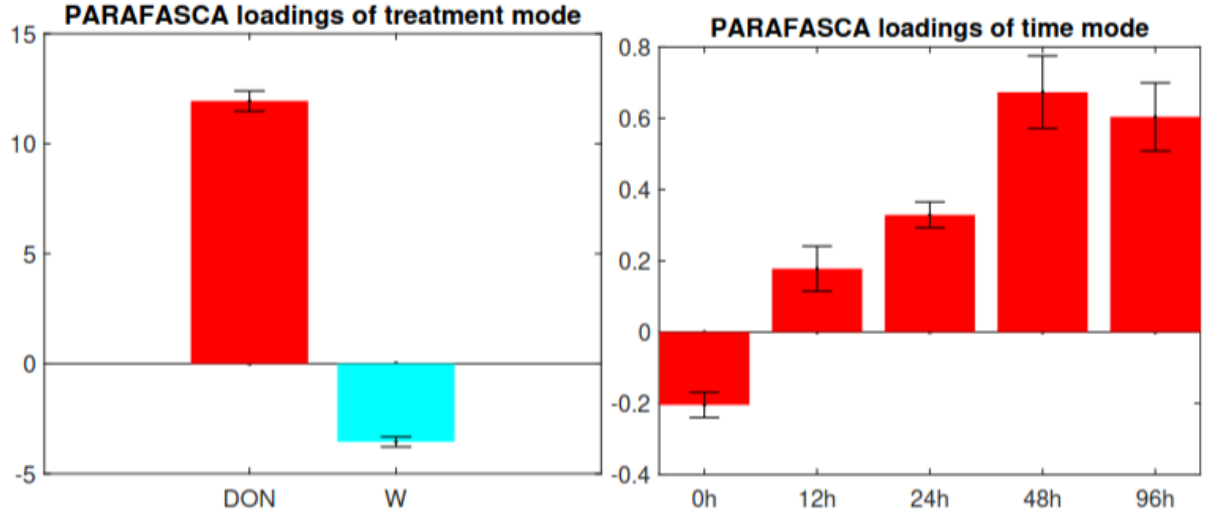
Michael Sorochan Armstrong[a,b], José Camacho[a]

[a] *Computational Data Science (CoDaS) Lab, C. Periodista Daniel Saucedo Aranda, Granada, 18071, Andalucia, Spain*
[b] *Lifewatch ERIC, Instituto Interuniversitario De Investigación Del Sistema Tierra En Andalucía, University of Granada, Av. del Mediterráneo, Granada, Granada, 18006, Andalucia, Spain*

MINISTERIO DE CIENCIA, INNOVACIÓN Y UNIVERSIDADES

Proyecto MAHOD: 101106986
MSCA Information Science and Engineering

PARAFASCA example



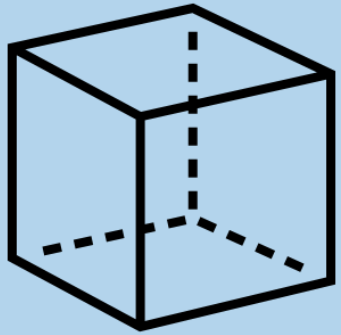
We can visualize the uncertainty per experimental level

MINISTERIO DE CIENCIA, INNOVACIÓN Y UNIVERSIDADES

Proyecto MAHOD: 101106986
MSCA Information Science and Engineering

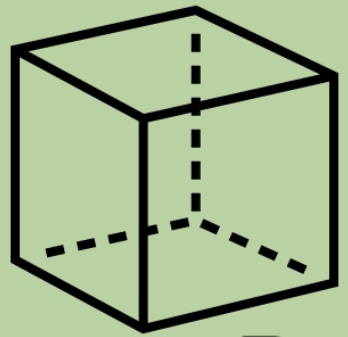
Recovering the retention scores

NOMINAL $\mathbf{X}_A$

$\mathbf{C}_{m/z}$  $\times$ $\mathbf{D}_{smpls}$

$\mathbf{A}_{1t_R}$ $\mathbf{B}_{2t_R}$

w/ ERROR $\mathbf{X}_A + \mathbf{E}$

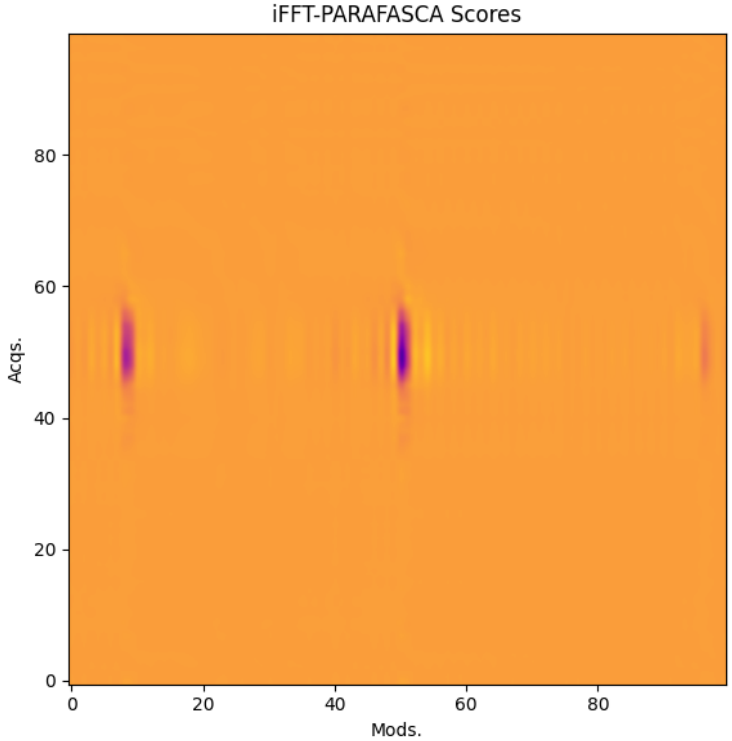
$\mathbf{C}_{m/z} + \mathbf{E}$  $\times$ $\mathbf{D}_{smpls} + \mathbf{E}$

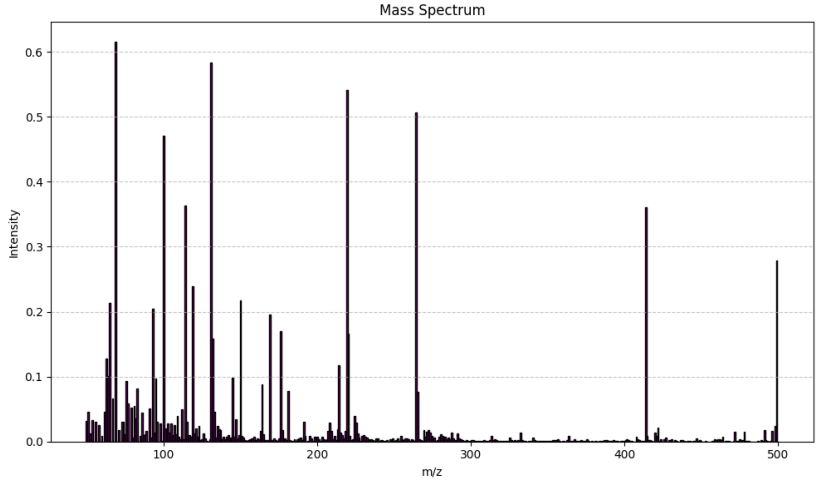
$\mathbf{A}_{1t_R} + \mathbf{E}$ $\mathbf{B}_{2t_R} + \mathbf{E}$

MINISTERIO DE CIENCIA, INNOVACIÓN Y UNIVERSIDADES

Proyecto MAHOD: 101106986
MSCA Information Science and Engineering

Results of 2DFFT-PARAFASCA





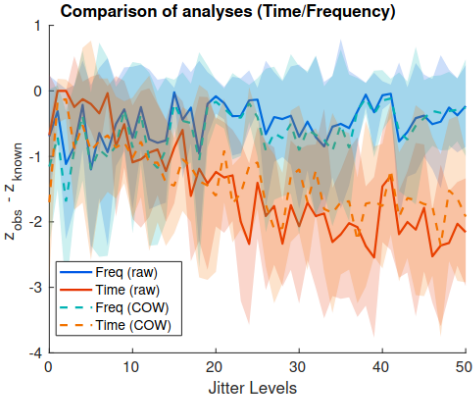
Results are localized by retention patterns, not individual peaks.

MINISTERIO DE CIENCIA, INNOVACIÓN Y UNIVERSIDADES

Proyecto MAHOD: 101106986
MSCA Information Science and Engineering

Further yet...

Heterogeneous data analysis - what if the peaks are in completely different locations?



Same colour := same factor

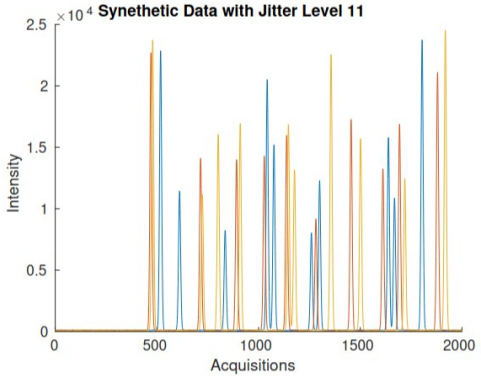
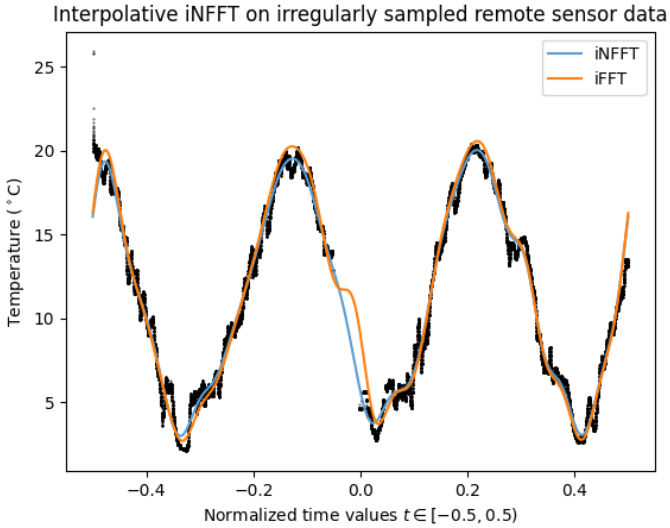


Figure A.16: 3 samples for jitter level 11

Proyecto MAHOD: 101106986
MSCA Information Science and Engineering

$$\hat{h}_k = Af(x_j)\left(W_k - W_k L\left(I + UW_k L\right)^{-1} UW_k\right)$$



A direct solution to the interpolative inverse non-uniform fast Fourier transform (i-iNFFT) problem for spectral analyses of non-equidistant time-series data

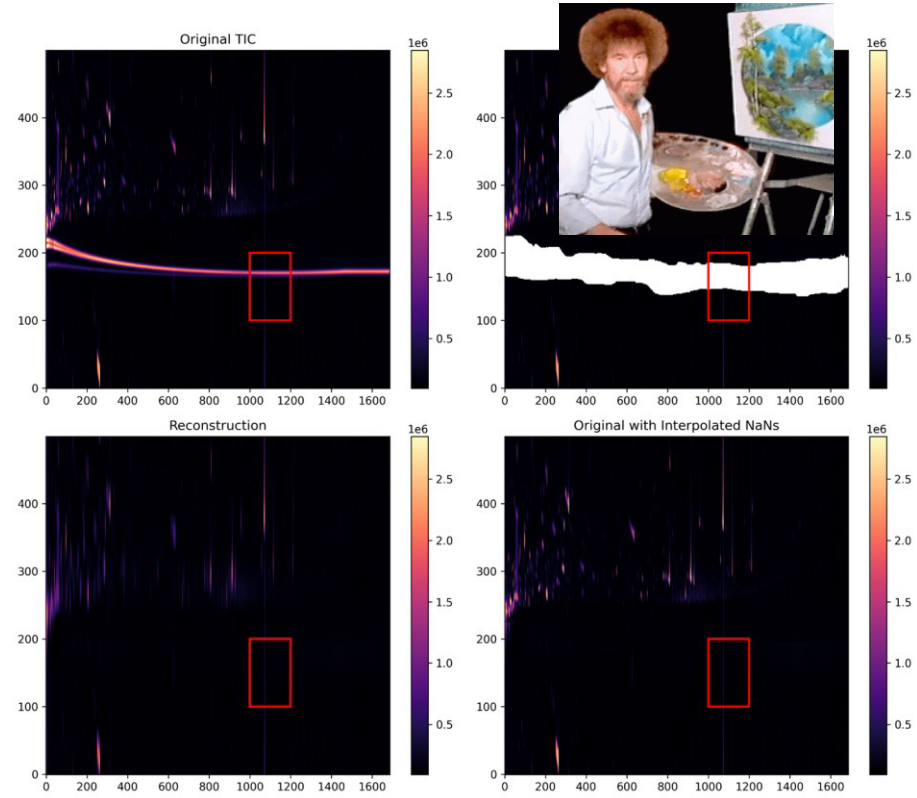
Michael Sorochan Armstrong, José Carlos Pérez-Girón, José Camacho *Member, IEEE*, Regino Zamora

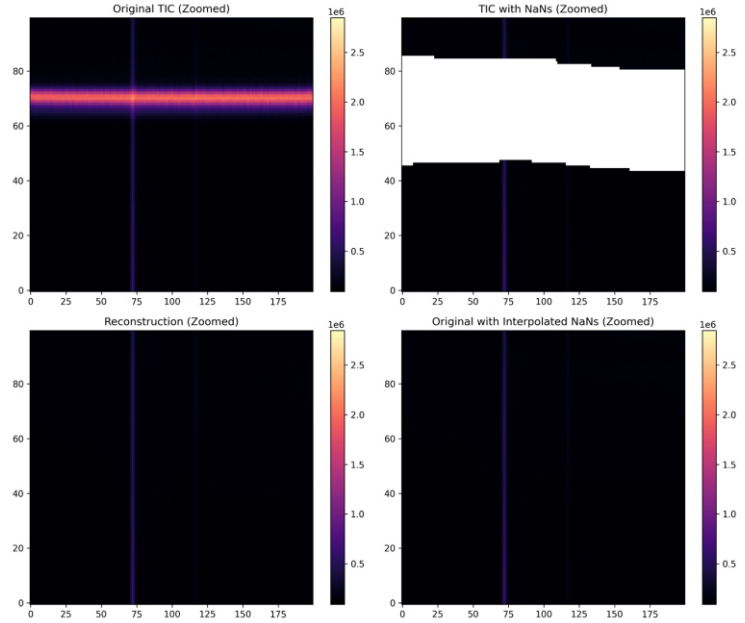
MINISTERIO DE CIENCIA, INNOVACIÓN Y UNIVERSIDADES

Proyecto MAHOD: 101106986
MSCA Information Science and Engineering

Source: Star Wars

Interpolative transforms





MINISTERIO DE CIENCIA, INNOVACIÓN Y UNIVERSIDADES

Proyecto MAHOD: 101106986
MSCA Information Science and Engineering

Applications of Frequency Analysis

Heterogeneous data analysis

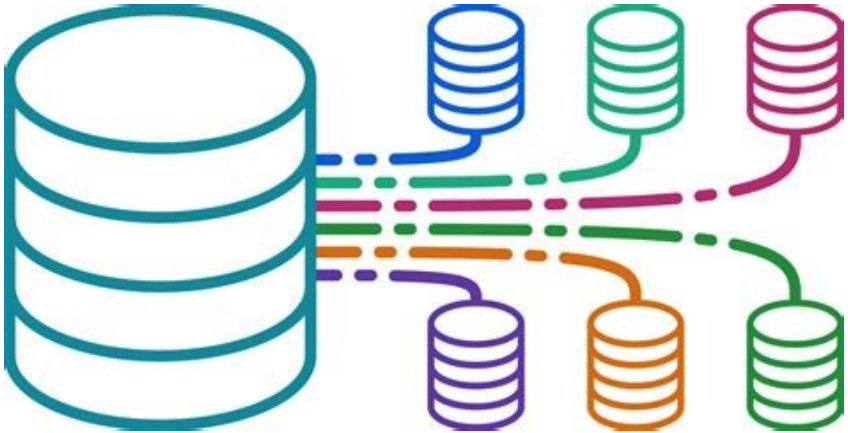
Process control

Interpolating irrelevant information

Extrapolation for saturated masses

Exploratory data analysis

Statistical analysis



MINISTERIO DE CIENCIA, INNOVACIÓN Y UNIVERSIDADES
Proyecto MAHOD: 101106986
MSCA Information Science and Engineering

Acknowledgements

Computational Data Science (CoDaS) Lab

UNIVERSIDAD DE GRANADA

GOBIERNO DE ESPAÑA
MINISTERIO DE CIENCIA E INNOVACIÓN

mustard
multi-scale spatio-temporal analysis of research data



MARIE CURIE ACTIONS

MAHOD
101106986

Dr. José Camacho
Dr. Regino Zamora

UNIVERSIDAD DE GRANADA

chemistry matters

Dr. Courtney Sandau
Dr. Ifeoluwa Grace Idowu

MINISTERIO DE CIENCIA, INNOVACIÓN Y UNIVERSIDADES

Proyecto MAHOD: 101106986
MSCA Information Science and Engineering

Acknowledgements

Centro de Investigación en Tecnologías de la Información y las Comunicaciones de la Universidad de

Universidad, Investigación e Innovación

MINISTERIO DE CIENCIA, INNOVACIÓN Y UNIVERSIDADES

Proyecto MAHOD: 101106986
MSCA Information Science and Engineering