1st IAA-CSIC Severo Ochoa School on Statistics, Data Mining, and Machine Learning

EXCELENCIA SEVERO OCHOA

Parametric modeling of time series: Gap-filling application

Javier Pascual Granado (IAA-CSIC, Spain)

PLATO IAA-CSIC Science Team

PLATO IAA-CSIC Science Team

Outline

- Introduction
- The interpolation problem
- Parametric modeling: ARMA
- MARA
- Conclusions

Software

- git pull
- Jupyter notebooks
- Python software:

  pip install numpy scipy astropy statsmodels matplotlib pandas ipython

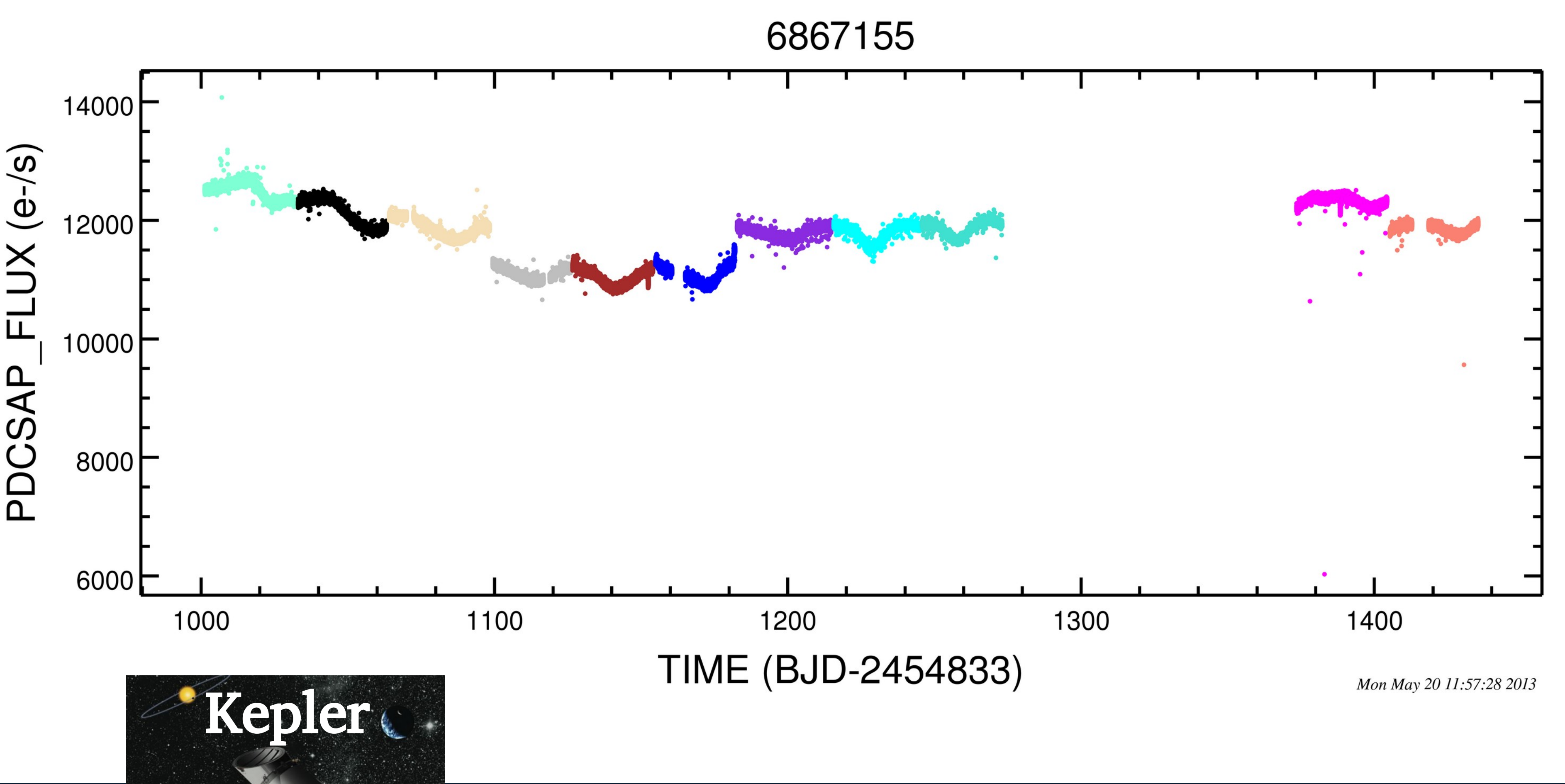
Introduction

Interpolation is a very usual problem in data analysis:

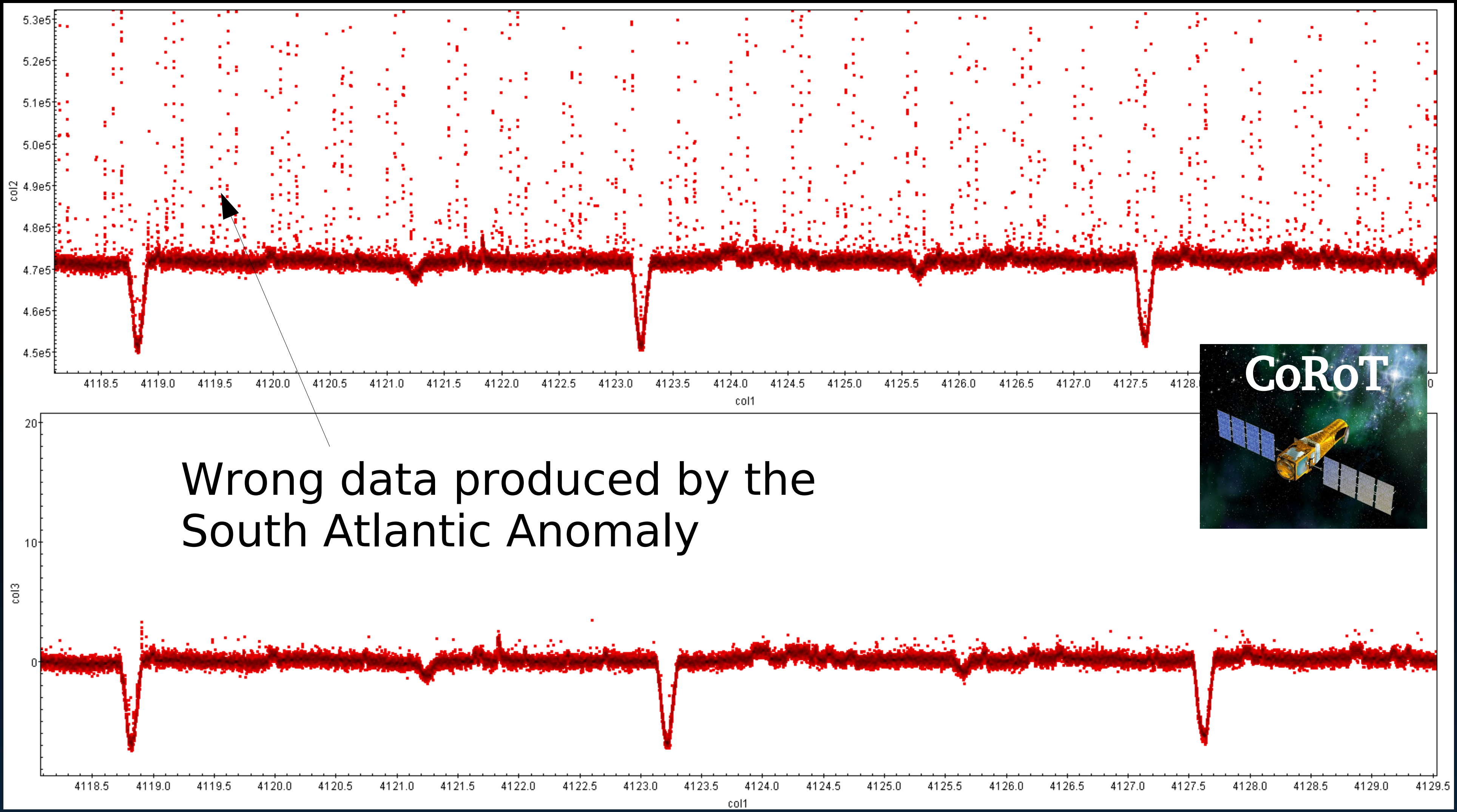
Gaps ↔ Irregular sampling ↔ Outliers

Signal Identification ↔ Data Modeling ↔ Interpolation ↔ Compression ↔ Noise Filtering ↔ …

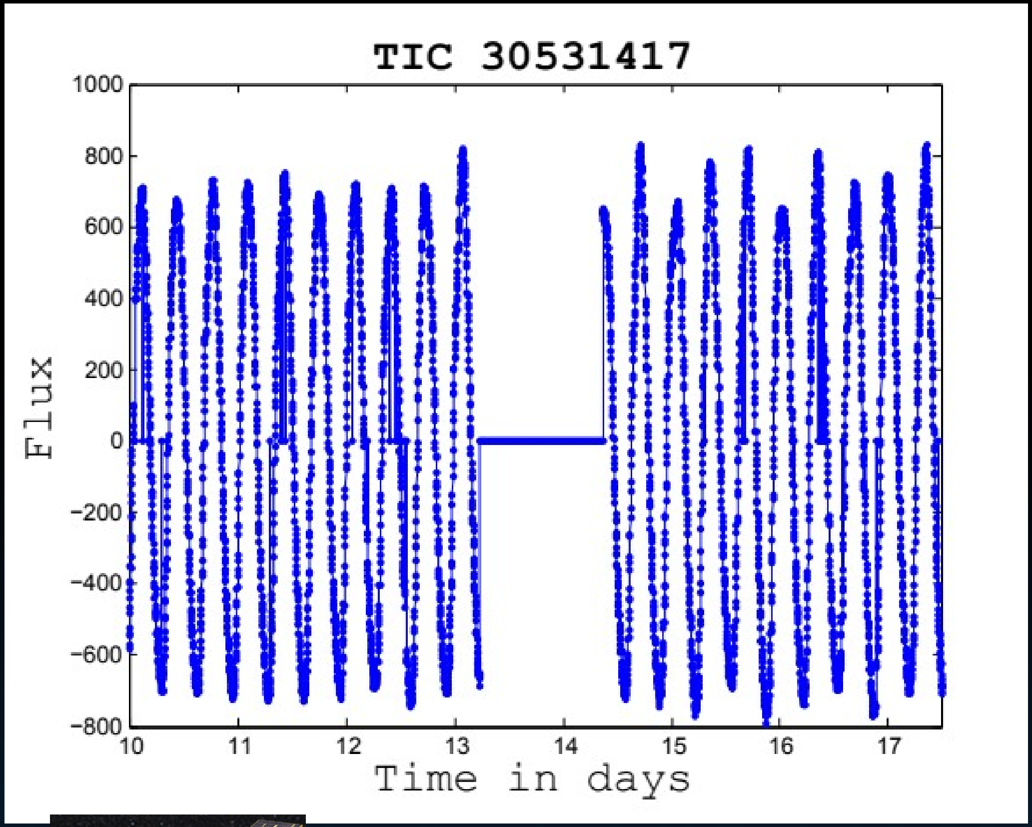
Why it is necessary to interpolate – part I

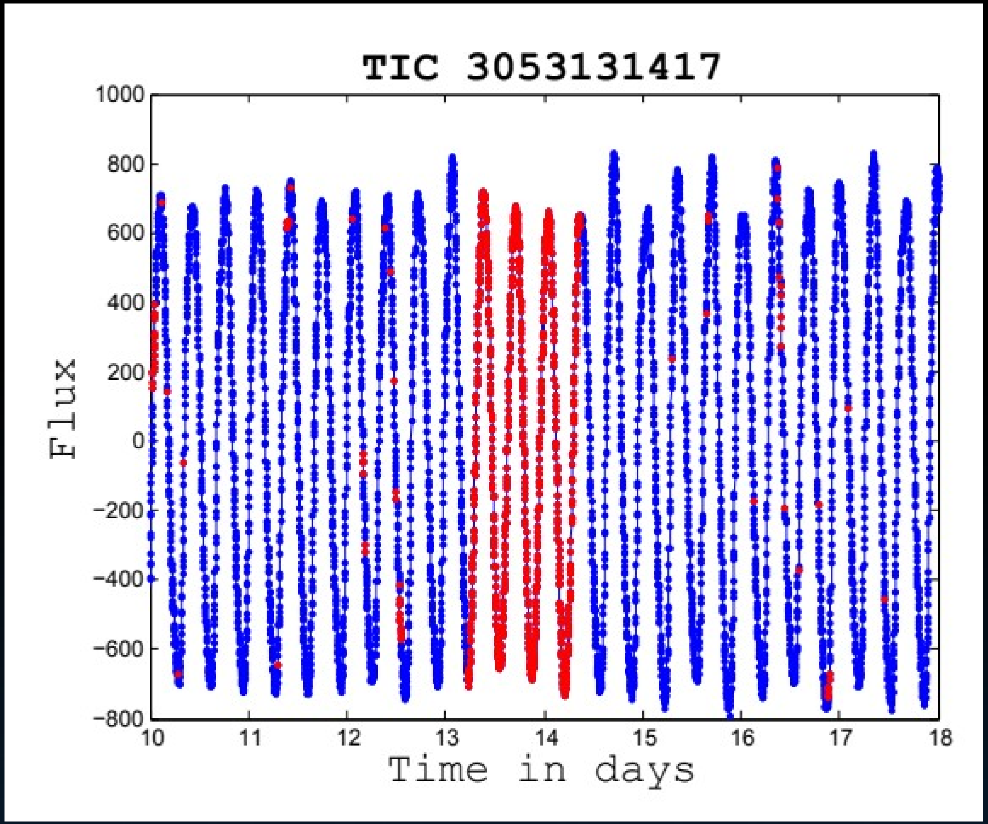


Why it is necessary to interpolate – part I



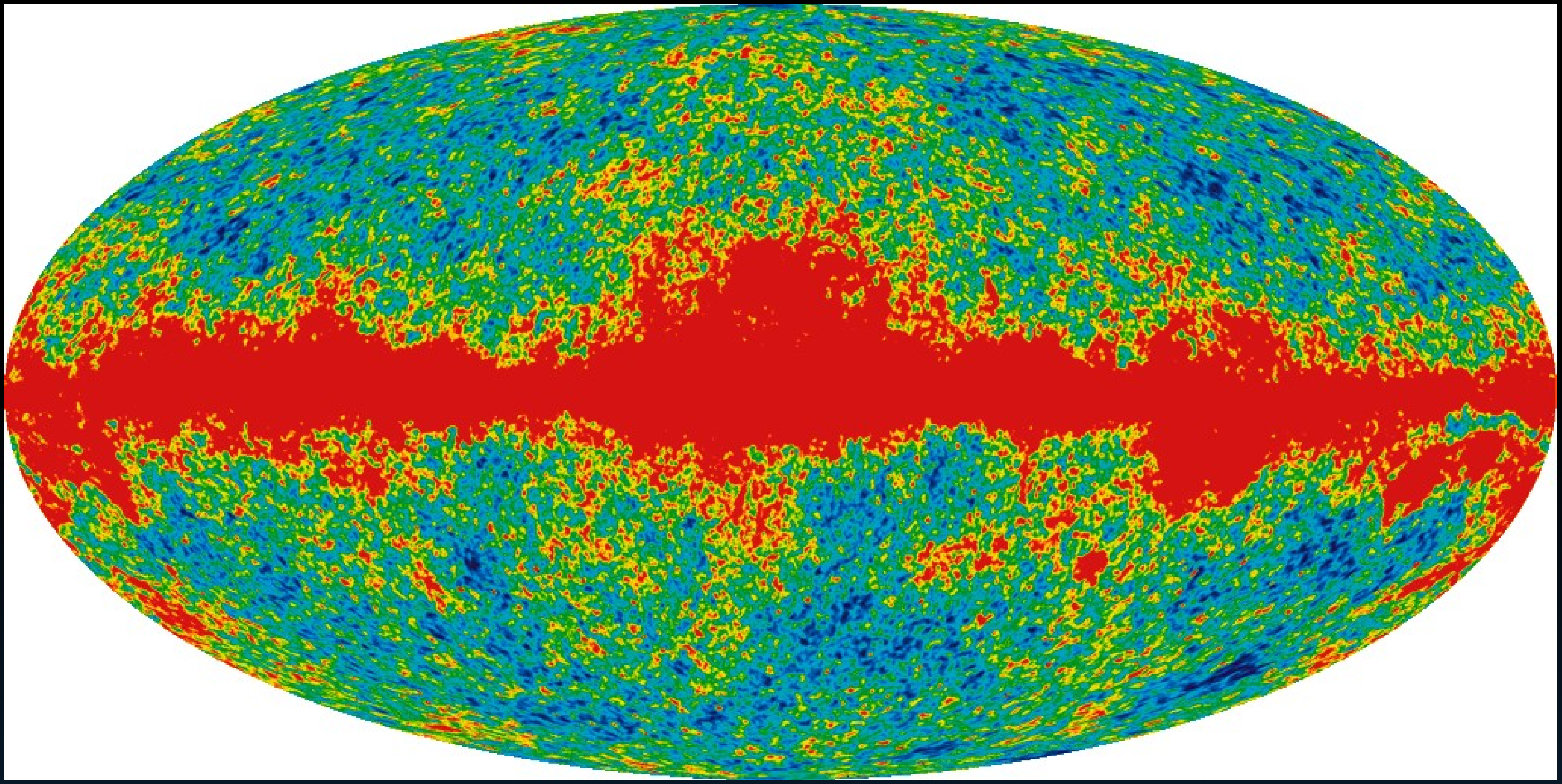
Why it is necessary to interpolate – part I





TESS 2-minute cadence data

Why it is necessary to interpolate – part I



WMAP full-sky map in Ka band. The red band is microwave emission from our Galaxy.

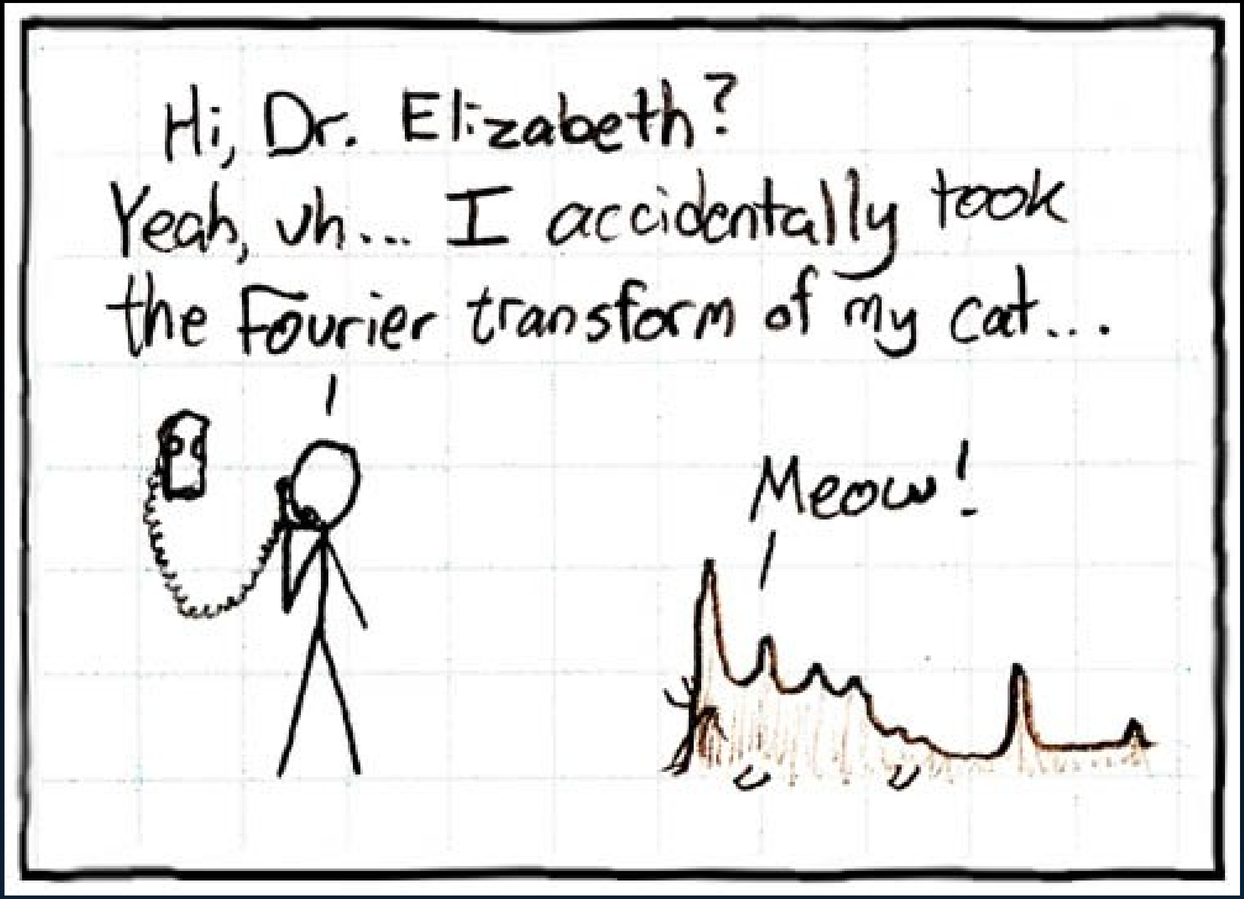
Why it is necessary to interpolate?

Why are gaps a problem?

Can't we just analyze the chunks of data that don't have gaps?

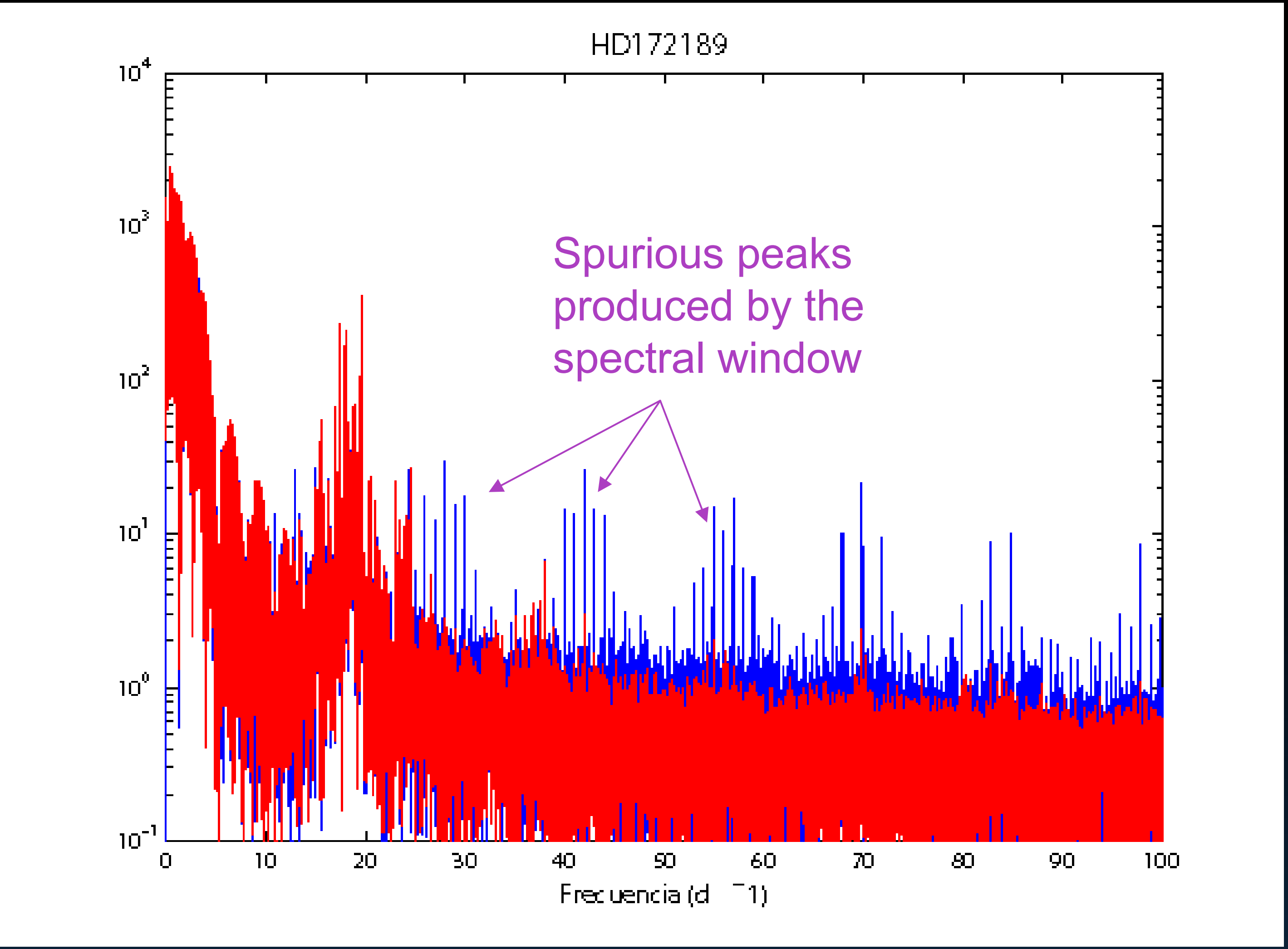
Why it is necessary to interpolate – part I

FOURIER



XKCD

Why it is necessary to interpolate – part I



Pascual-Granado et al., CUP 2012, IAUS285, pp.392-393

Spectral Window

FOURIER ANALYSIS WITH UNEQUALLY-SPACED DATA*

T. J. DEEMING
Dept. of Astronomy, The University of Texas at Austin, Tex., U.S.A.

(Received 22 March; in revised form 11 November, 1974)

$$w_T(t) = \begin{cases} 1; & (-T/2 \leqslant t \leqslant T/2) \\ 0; & \text{otherwise} \end{cases}$$

$$w_N(t) = \sum_{k=1}^{N} \delta(t - t_k).$$

Data windows

Discrete and Finite Fourier Transform

$$F_{T,N}(\nu) = \int_{-\infty}^{+\infty} w_{T,N}(t) f(t) e^{i2\pi\nu t}\,\mathrm{d}t, \longrightarrow F_{T,N}(\nu) = F(\nu) * \bar{W}_{T,N}(\nu),$$

Spectral Window

FOURIER ANALYSIS WITH UNEQUALLY-SPACED DATA*

T. J. DEEMING
Dept. of Astronomy, The University of Texas at Austin, Tex., U.S.A.

(Received 22 March; in revised form 11 November, 1974)

$$\left.\begin{aligned} w_T(t) &= \begin{cases} 1; & (-T/2 \leqslant t \leqslant T/2) \\ 0; & \text{otherwise} \end{cases} \\ w_N(t) &= \sum_{k=1}^{N} \delta(t - t_k). \end{aligned}\right\} \longleftarrow \text{Data windows}$$

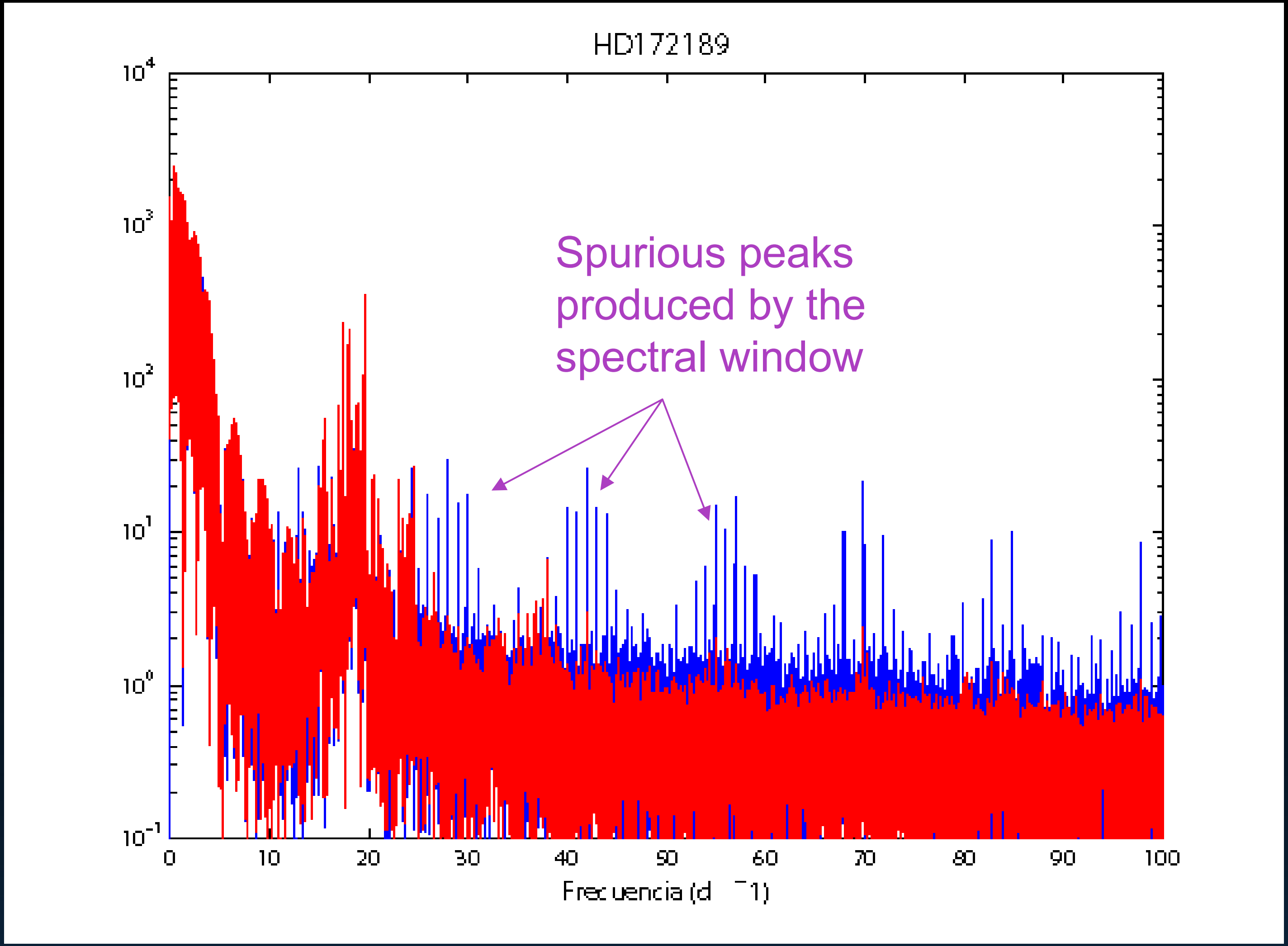
Discrete and Finite Fourier Transform

$$F_{T,N}(\nu) = \int_{-\infty}^{+\infty} w_{T,N}(t) f(t) e^{i2\pi\nu t}\, dt, \longrightarrow F_{T,N}(\nu) = F(\nu) * \bar{W}_{T,N}(\nu),$$

$$\boxed{W_N(\nu) = \sum_{k=1}^{N} e^{i2\pi\nu t_k} = \delta_N(\nu).}$$

Spectral Window Function

Why it is necessary to interpolate – part I



Pascual-Granado et al., CUP 2012, IAUS285, pp.392-393

So how do we interpolate in the gaps?

Interpolation - inpainting techniques

Ecce Homo, Sanctuary of Mercy church in Borja, Spain.

Interpolation - inpainting techniques

Left panel is the original from Elías García Martínez. Right panel shows restoration attempt from Cecilia Giménez.

A gap-filling method aimed to be <u>information preserving</u>

⬇

Unbiased

⬇

Non-closed form expression, fitting functions that can be analytic or not.

Interpolation

scipy.interpolate

Univariate interpolation

interp1d(x, y[, kind, axis, copy, ...])	Interpolate a 1-D function.
BarycentricInterpolator(xi[, yi, axis])	The interpolating polynomial for a set of points
KroghInterpolator(xi, yi[, axis])	Interpolating polynomial for a set of points.
barycentric_interpolate(xi, yi, x[, axis])	Convenience function for polynomial interpolation.
krogh_interpolate(xi, yi, x[, der, axis])	Convenience function for polynomial interpolation.
pchip_interpolate(xi, yi, x[, der, axis])	Convenience function for pchip interpolation.
CubicHermiteSpline(x, y, dydx[, axis, ...])	Piecewise-cubic interpolator matching values and first derivatives.
PchipInterpolator(x, y[, axis, extrapolate])	PCHIP 1-d monotonic cubic interpolation.
Akima1DInterpolator(x, y[, axis])	Akima interpolator
CubicSpline(x, y[, axis, bc_type, extrapolate])	Cubic spline data interpolator.
PPoly(c, x[, extrapolate, axis])	Piecewise polynomial in terms of coefficients and breakpoints
BPoly(c, x[, extrapolate, axis])	Piecewise polynomial in terms of coefficients and breakpoints.

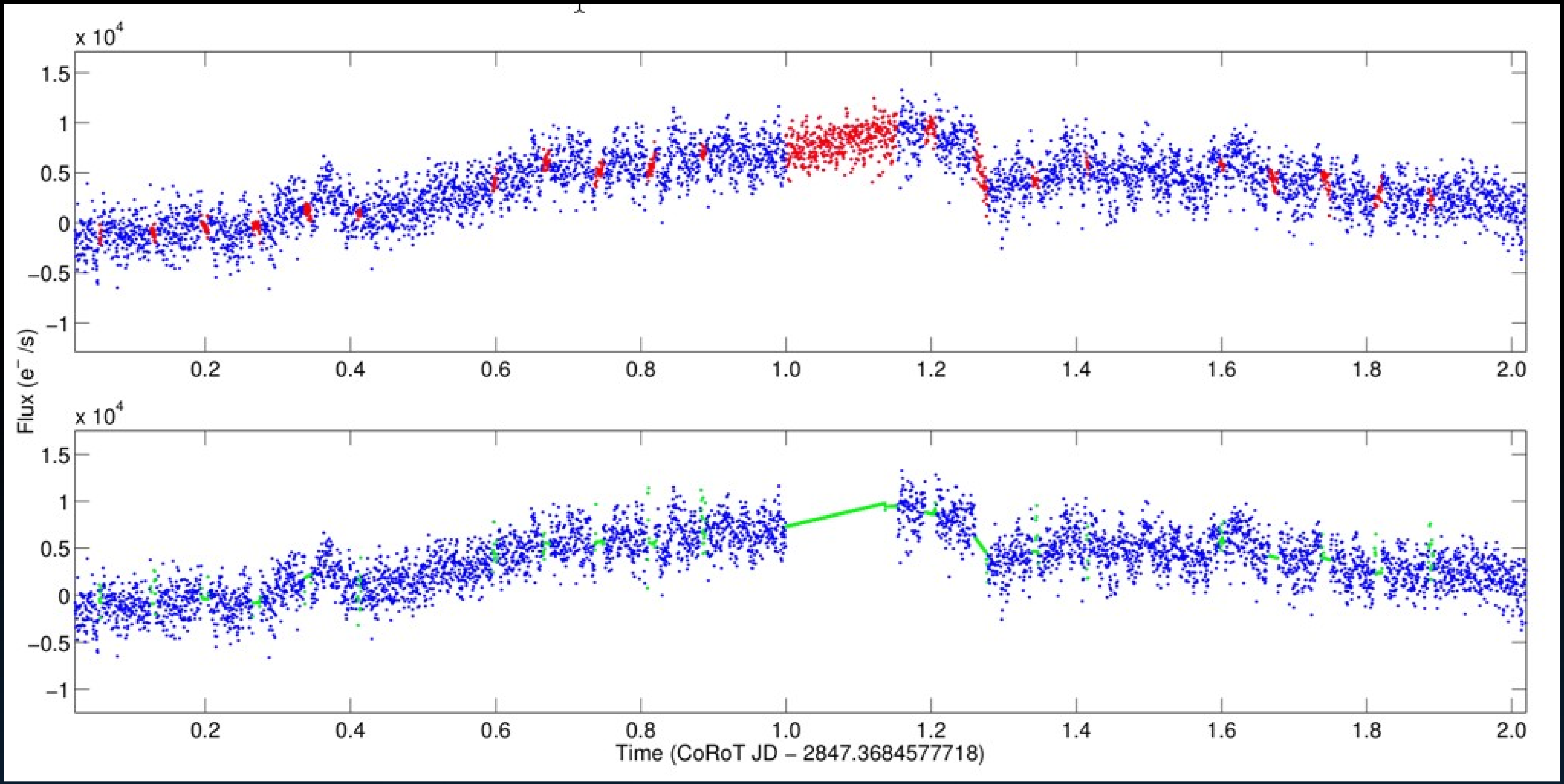
1-D Splines

BSpline(t, c, k[, extrapolate, axis])	Univariate spline in the B-spline basis.
make_interp_spline(x, y[, k, t, bc_type, ...])	Compute the (coefficients of) interpolating B-spline.
make_lsq_spline(x, y, t[, k, w, axis, ...])	Compute the (coefficients of) an LSQ B-spline.

Additional tools

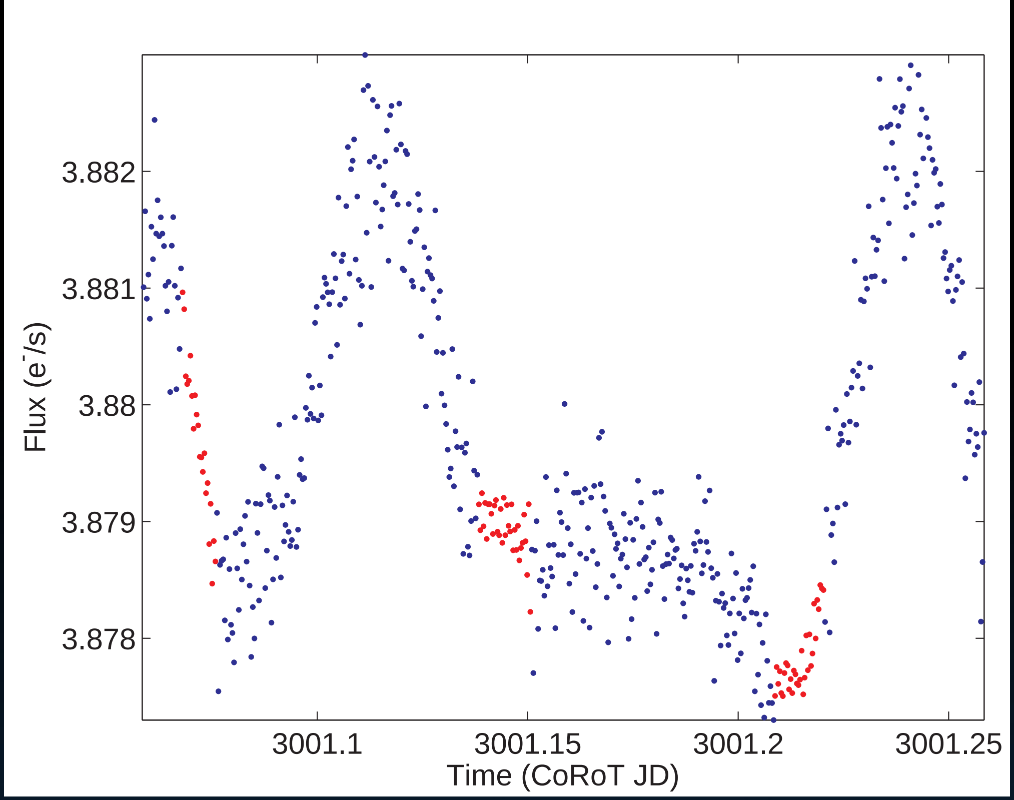
lagrange(x, w)	Return a Lagrange interpolating polynomial.
approximate_taylor_polynomial(f, x, degree, ...)	Estimate the Taylor polynomial of f at x by polynomial fitting.
pade(an, m[, n])	Return Pade approximation to a polynomial as the ratio of two polynomials.

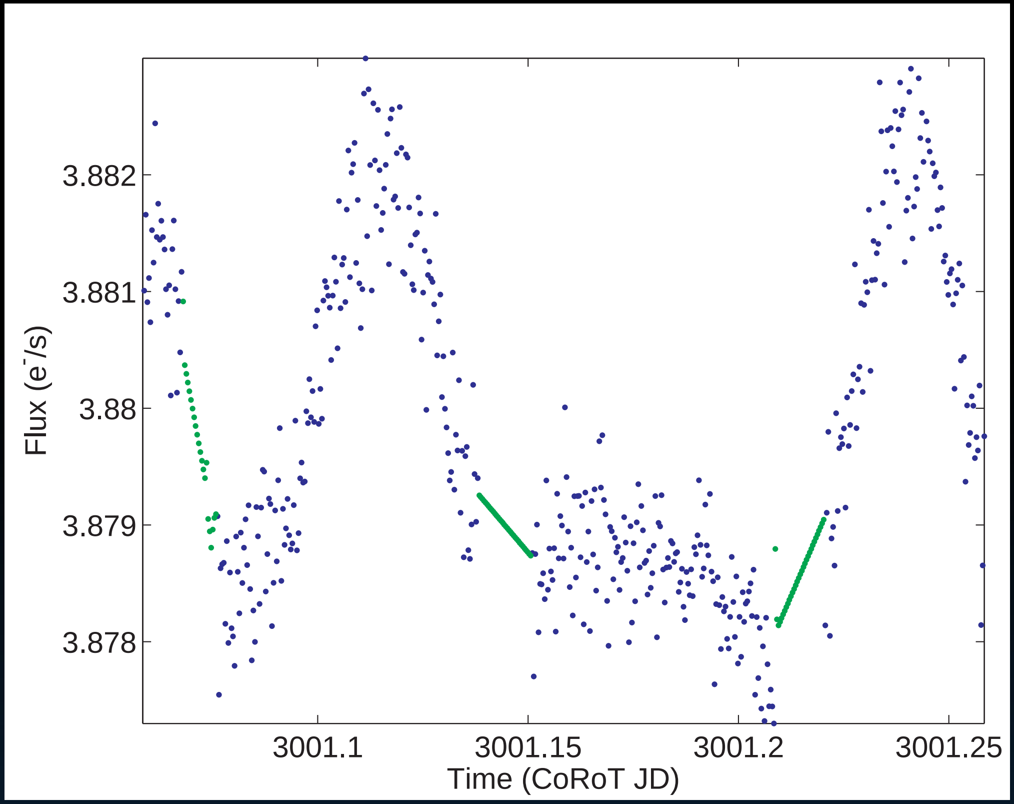
Interpolation: linear



HD49933 – solar-like star with periods ~ min

Interpolation: linear





HD 48784 – Delta Scuti star with periods ~ hr

Why it is necessary to interpolate – part II

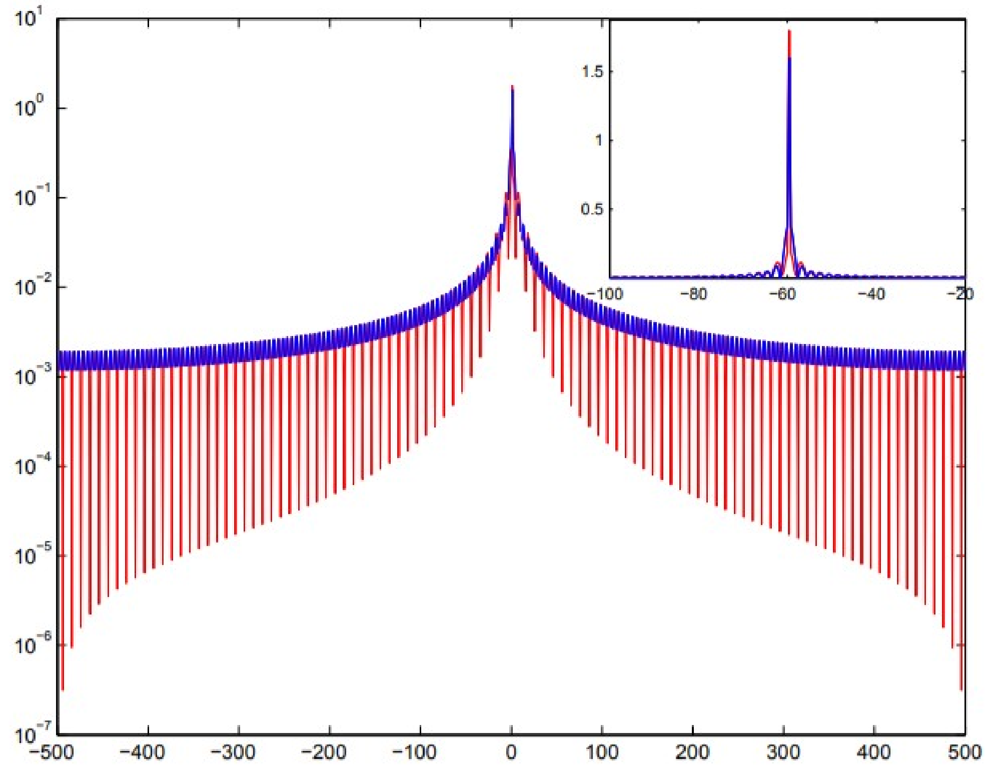


Fig. A.1. Spectral response function in log scale associated to gapped data (in blue) and linearly interpolated data (in red). See the inset for a zoom of the central peak in linear scale.

Pascual-Granado et al. 2018, A&A, 614, A40

Why it is necessary to interpolate – part II

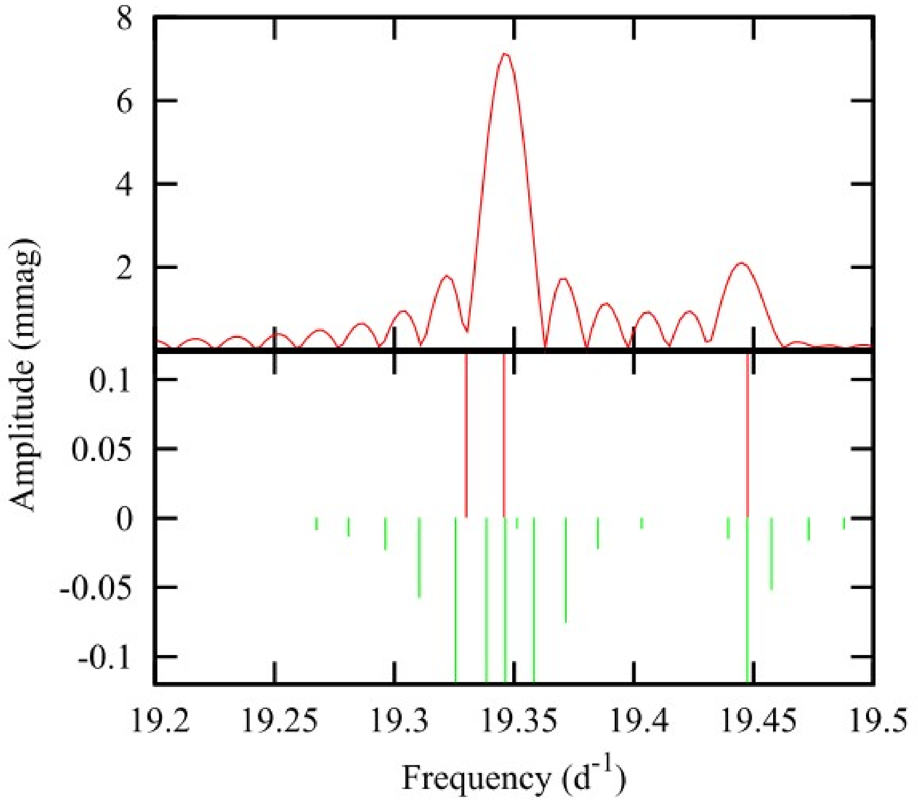
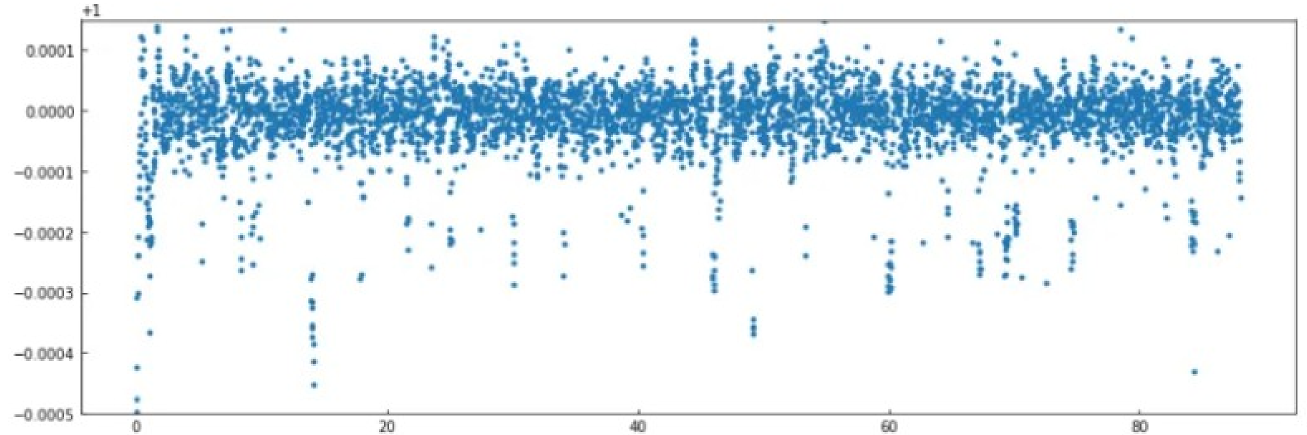


Figure 4. Schematic periodogram of known frequency components (with positive amplitudes) and extracted components (negative amplitudes) in a simulation.

Balona, L. 2014, MNRAS, 439,3453-3460

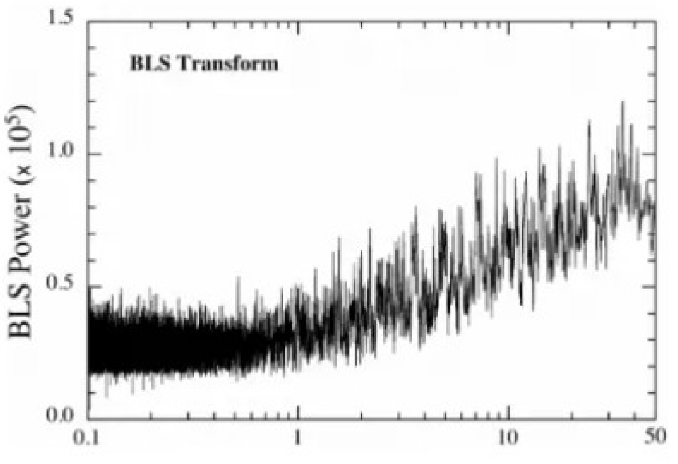
- Although the signal has just 3 frequencies, numerous frequencies of relatively high amplitudes are required by the non-linear least squares algorithm to fit the signal.
- Even though the difference between simulated and extracted f1 is only 0.0007 d^{-1} many fictitious components appear as a result of the insufficient quality of the fitting.

Ultra-precise data analysis



HD 139139 / EPIC 249706694

The Random Transiter Star



In the era of ultra-precise data from satellites we are going to need ultra-precise data analysis to understand what we observe

Information-preserving interpolation

A gap-filling information preserving method

Unbiased

Non-closed form expression, fitting functions that can be analytic or not.

ARMA interpolation (MIARMA)

Pascual-Granado, J., Garrido, R., and Suárez, J. C. 2015, A&A, 575, A78

Go to: www.menti.com

use the code: 97012

And answer this anonymous poll:

How familiar are you with ARIMA?


ARIMA DANTZA ESKOLA
segurma
902 108 088
CONECTADA
JAR EZAZU ARIMA DANTZAN!
DANZA CONTEMPORANEA
BALLET
URBAN
JAZZ FUNK
DANCE HALL
DANZA MODERNA
ORIENTAL
FLAMENCO
CAPOEIRA
GYM DANCE
BARRA-ESTIRAMIENTOS
YOGA TXIKI
YOGA/ PILATES
TEATRO MUSICAL

Autoregressive models

The class of autoregressive (AR) processes, and its extensions, autoregressive moving-average (ARMA) processes, are dense in the class of Gaussian linear processes.

Autoregressive models

The class of autoregressive (AR) processes, and its extensions, autoregressive moving-average (ARMA) processes, are dense in the class of Gaussian linear processes.

Wold Decomposition Theorem (Wold 1938)

"<u>Any stationary random process</u> can be decomposed into the sum of a purely random process and a linearly deterministic process, and further that the random part is a moving average, i.e. the convolution of a fixed, causal, invertible filter with an uncorrelated noise process."

Scargle, J. D. 1981, ApJS, 45

Autoregressive models

AR $$x_t = \sum_{k=1}^{p} \alpha_k x_{t-k} + a_t$$ **Purely Autoregressive**

MA $$x_t = -\sum_{k=1}^{q} b_k n_{t-k} \quad \rightarrow X = B*N$$ **Moving Average**

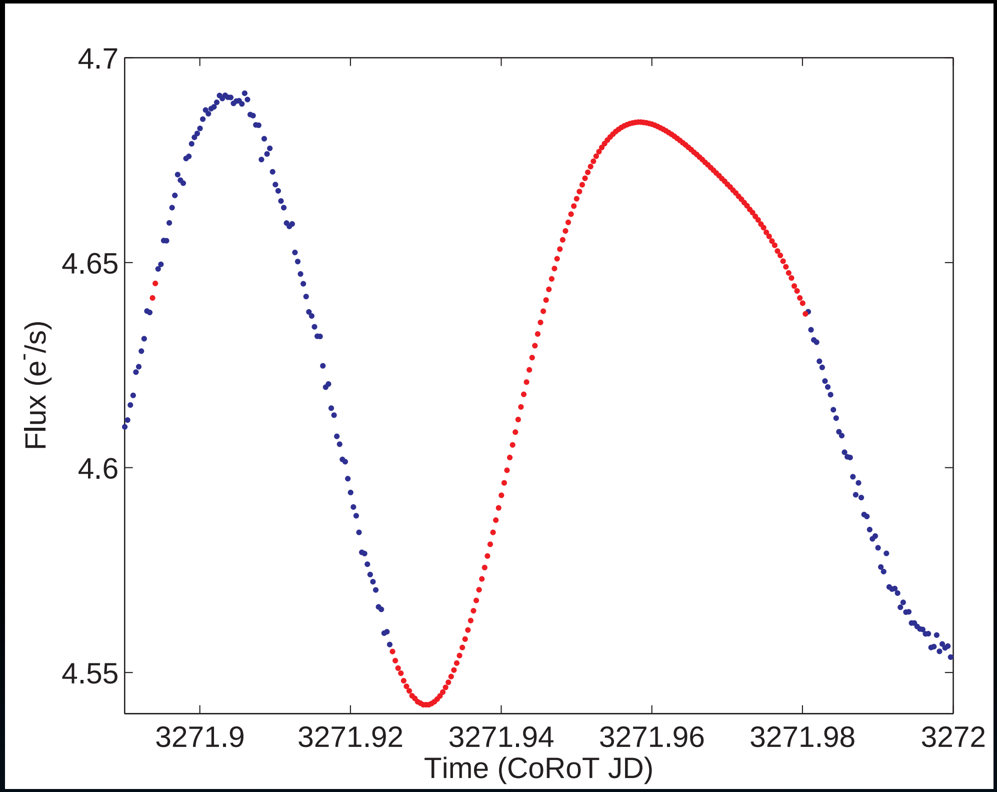
ARMA $$x_t = \sum_{k=1}^{p} \alpha_k x_{t-k} - \sum_{k=1}^{q} b_k n_{t-k} + a_t$$ **Mixed AR + MA**

More information: Feigelson, Eric D., Babu, Jogesh G., Caceres, Gabriel A., 2018, Frontiers in Physics, 6, 80

Go to Jupyter Notebook and load:

ARMA_pred1.ipynb

ARMA models fits deterministic signal too



e.g. a stochastically excited damped harmonic oscillation is described by an AR process of second order, i.e., p = 2 (Honerkamp, 2002)

$$x(t) = a_1\, x(t-1) + a_2\, x(t-2) + \eta(t)$$

This is the discretized version of the stochastic second order differential equation for the stochastically excited damped harmonic oscillation

$$\ddot{x}(t) = -\gamma\, \dot{x}(t) + \omega^2\, x(t) + \eta(t),$$

Roth, M., Zhugzhda, Yu D., 2010, Astronomy Letters, 36, 1

ARMA models fits deterministic signal too

Exercise 1

Make a simulated signal with just one harmonic component and try to fit an ARMA model to it. e.g.

```
nobs = 250
f1 = 0.1
t = np.arange(nobs)
yh = np.sin(2*np.pi*f1*t)
noise = np.random.normal(0,1,nobs)
```

Then plot it as in ARMA_pred1.ipynb

ARMA models fits deterministic signal too

<u>Hints:</u>

- Use some start params for the model fitting to converge:

```
model.fit(trend='nc', disp=-1, start_params=[1, 0])
```

- If you still have convergence problems and you want to force it to go through, you can try transparams=False

```
model.fit(trend='nc', disp=-1, start_params=[1, 0],
    transparams=False)
```

Autoregressive models

AR $x_t = \sum_{k=1}^{p} \alpha_k x_{t-k} + a_t$ **Purely Autoregressive**

MA $x_t = -\sum_{k=1}^{q} b_k n_{t-k}$ **Moving Average**

ARMA $x_t = \sum_{k=1}^{p} \alpha_k x_{t-k} - \sum_{k=1}^{q} b_k n_{t-k} + a_t$ **Mixed AR + MA**

How can we determine the orders p and q?

Go to Jupyter Notebook and load:

ARMA_pred2.ipynb

Exercise 2

Continue the previous example and try to fit other models in order to find the optimal one.

For more on this, check:

Time Series Analysis: Forecasting and Control (Wiley Series in Probability and Statistics) 5th Edition

by George E. P. Box, Gwilym M. Jenkins, Gregory C. Reinsel, Greta M. Ljung

CRITERION FOR SELECTION OF THE ORDER (P,Q)

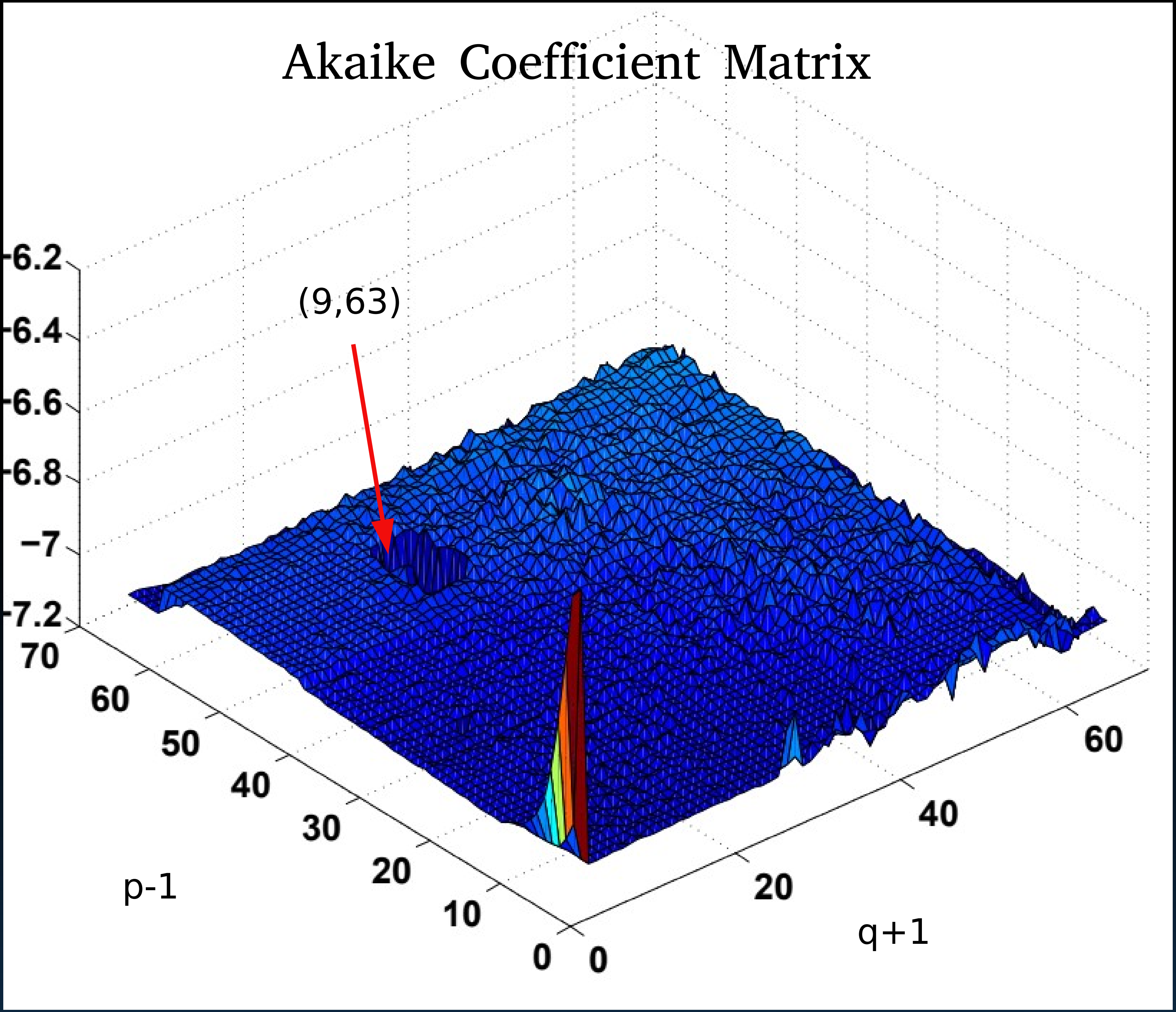
- An ungapped data segment is modelled. Iteration through p, q
- Given the k model, its Akaike coefficient is obtained (AIC_k)

$$AIC_k = N.\log(V) + 2(p + q)$$

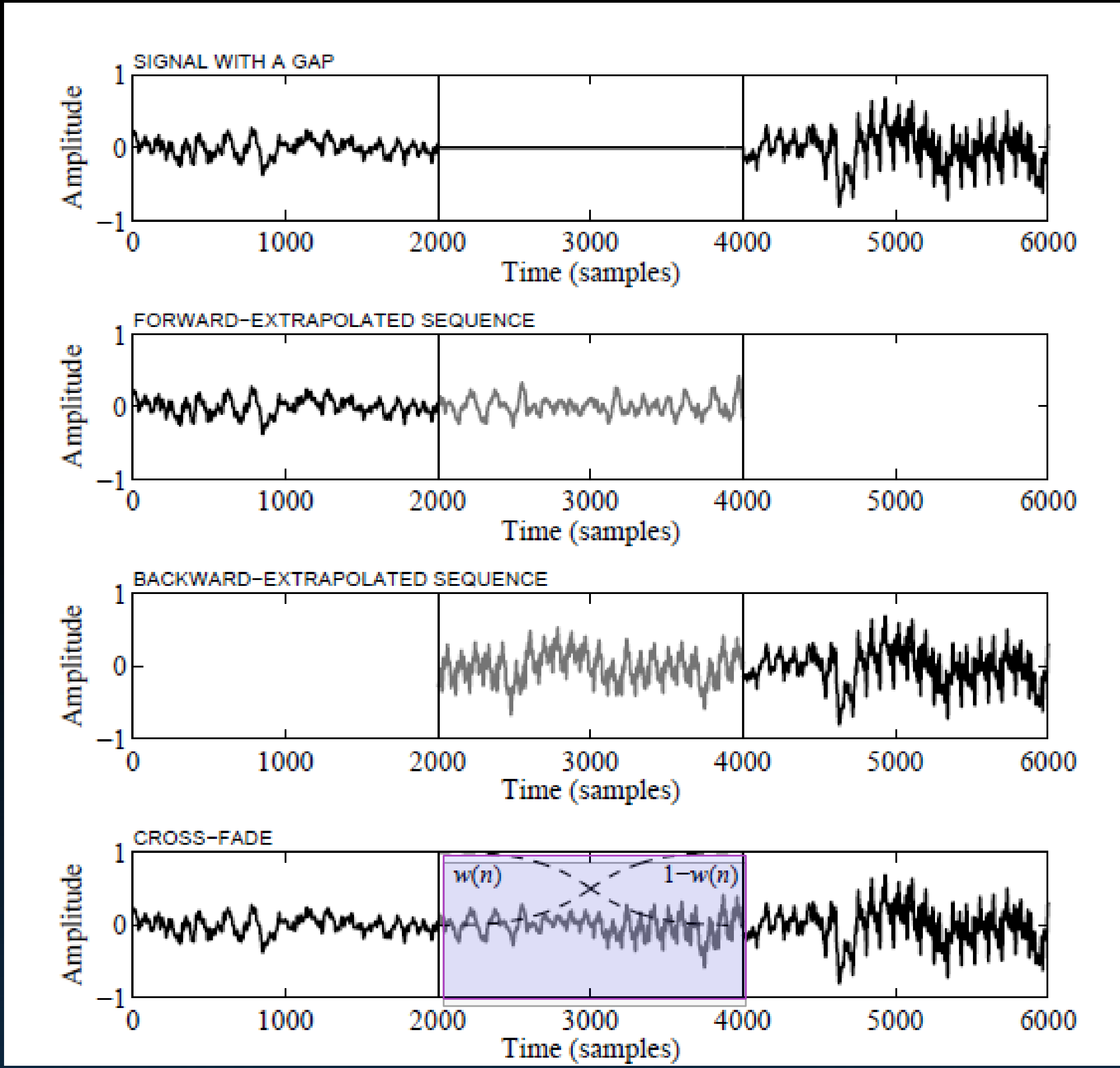
N = length of the data segment,

V = mean quadratic error of prediction

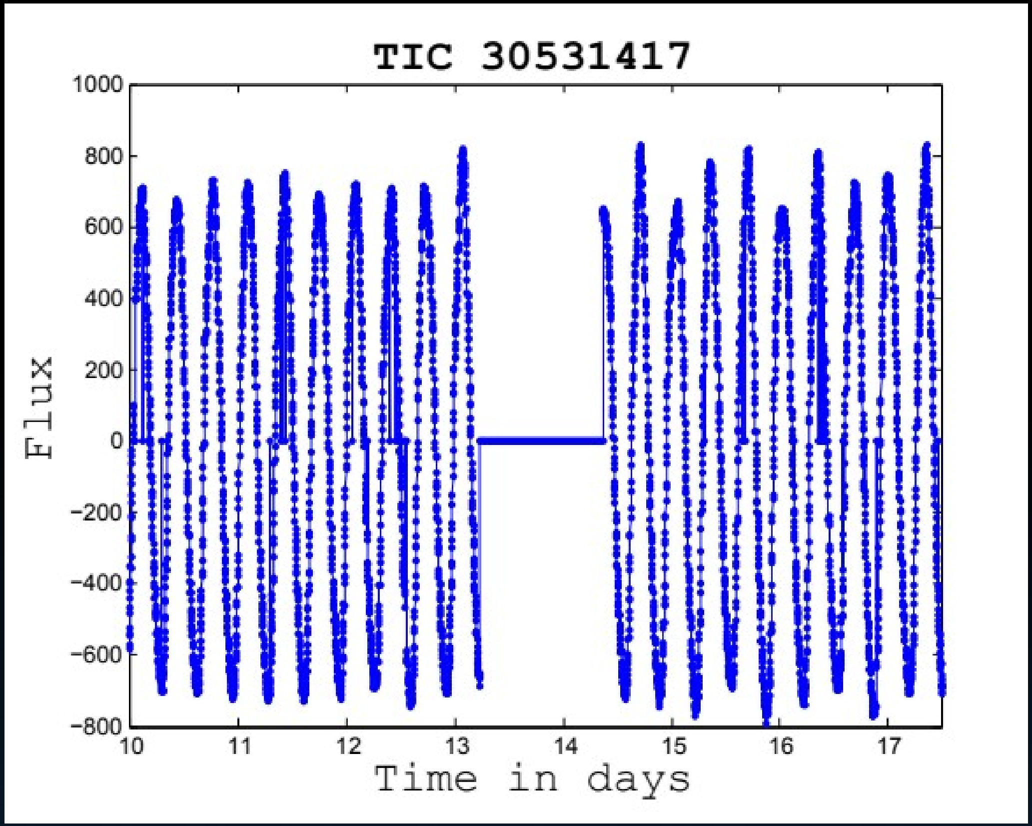
- Akaike criterion: the optimal model has min AIC_k
- Maximum Entropy Principle: guarantees that it is the best model that we can find with the information available.


Akaike Coefficient Matrix
(9,63)
-6.2
-6.4
-6.6
-6.8
−7
-7.2
70
60
50
40
30
20
10
0
p-1
0
20
40
60
q+1

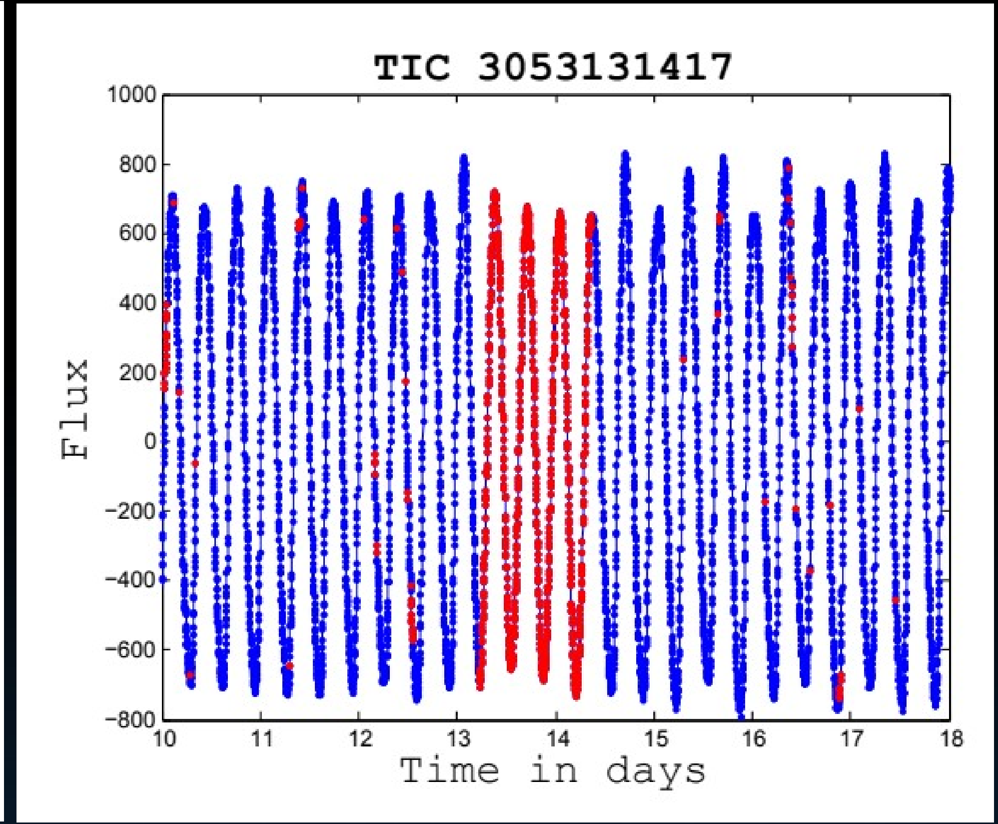
MIARMA


SIGNAL WITH A GAP
FORWARD−EXTRAPOLATED SEQUENCE
BACKWARD−EXTRAPOLATED SEQUENCE
CROSS−FADE
Amplitude
Time (samples)
w(n)
1−w(n)

MIARMA



TIC 30531417
Flux
Time in days



TIC 3053131417
Flux
Time in days

MIARMA **MARA**

MATLAB

Extension of the original algorithm:

* Nonstationary processes with ARIMA,
* Continuous time processes: CARMA, CARIMA
* Fractional integrated processes: ARFIMA, CARFIMA
* Fractal analysis with ARFIMA processes
* Multidimensional interpolation
* Parallelization of the computations

...

Module of AR Algorithms (MARA)

Lessons to take home

- Interpolation might be strictly necessary in order to perform ultra-precise data analysis and solve current challenges in astrophysics.
- Any data processing technique should be aimed to preserve the original information according to the scientific method.
- Use non-analytic models when you don't have any prior information about the signal.
- If you know that your data is stochastic or non-analytic don’t use analytic models for fitting/interpolating.
- Remember that ARMA can represent deterministic signals too.
- And finally, if you like the interpolations I’ve shown you here ask me about MARA.

“Music is the silence between the notes.”
– Claude Debussy

Thank you for your attention!