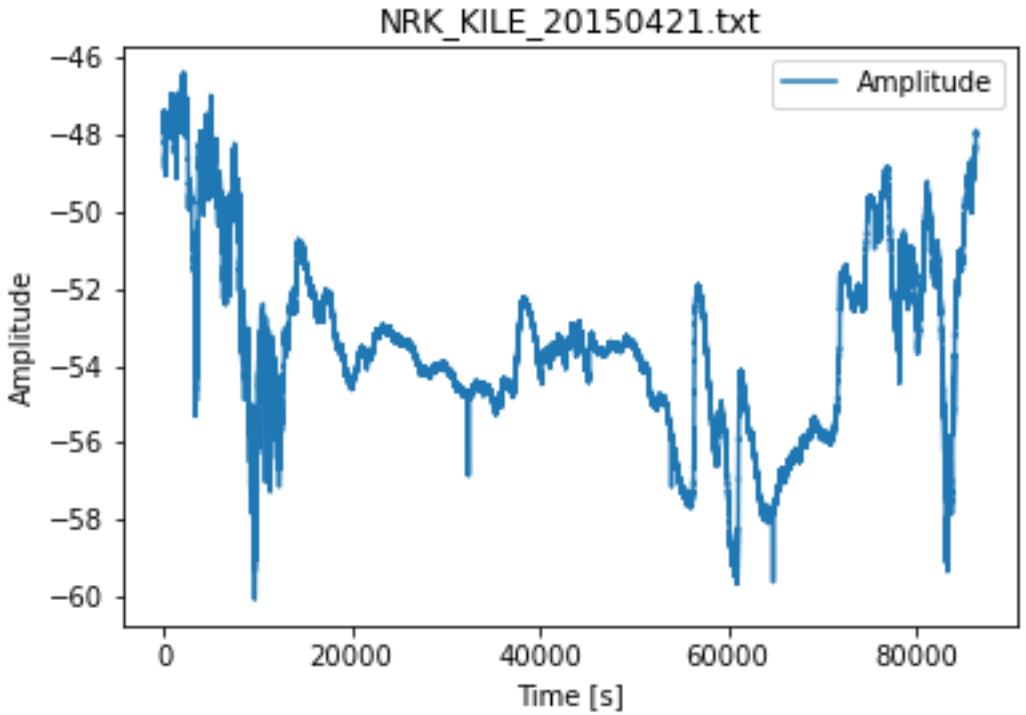
International Space Weather Summer Camp 2022: How to turn measurements into an analytical model – at the example of VLF data

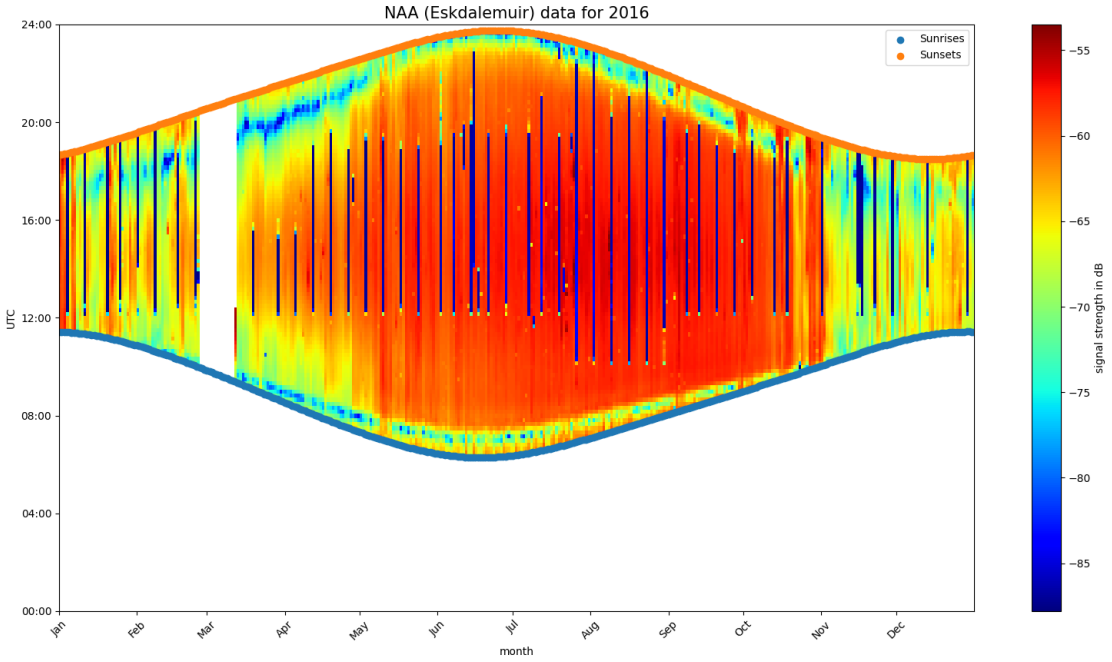
David Wenzel

German Aerospace Center (DLR)
Institute of Solar-Terrestrial Physics
Space Weather Observations

E-Mail: david.wenzel@dlr.de

Discussion


NRK_KILE_20150421.txt
Amplitude
−46
−48
−50
−52
−54
−56
−58
−60
0
20000
40000
60000
80000
Time [s]


NAA (Eskdalemuir) data for 2016
Sunrises
Sunsets
UTC
24:00
20:00
16:00
12:00
08:00
04:00
00:00
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
month
signal strength in dB
−55
−60
−65
−70
−75
−80
−85

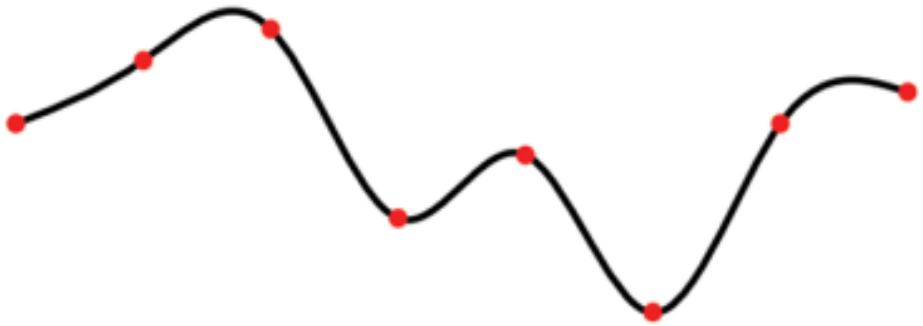
Curve approximation

Handling discrete data

- often measurements only at specific points
  → What is in between?

- interpolation with a polynome
- linear interpolation
- spline interpolation

- approximation
  → following a "trend"
  → eliminating oscillations



Handling discrete data

- often measurements only at specific points
  → What is in between?

- interpolation with a polynome
- linear interpolation
- spline interpolation

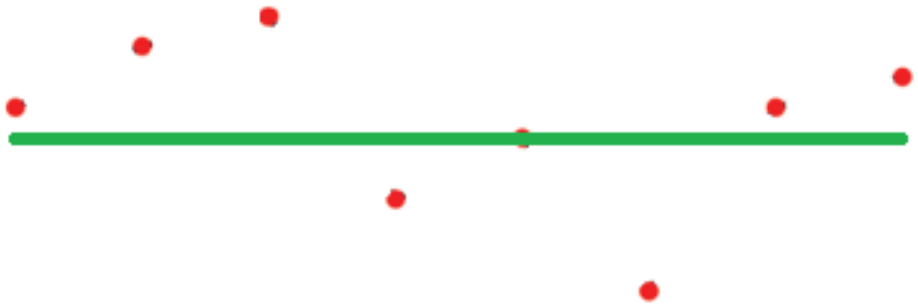
- approximation
  → following a "trend"
  → eliminating oscillations

Handling discrete data

- often measurements only at specific points
  → What is in between?

- interpolation with a polynome
- linear interpolation
- spline interpolation

- approximation
  → following a "trend"
  → eliminating oscillations



An algebraic approach

Estimating relationship between dependent and independent variables

(whiteboard)

→ attempt to solve overdetermined system

Why squaring the problem?

Pros:

- It turns things easier!
- absolute values not differentiable
- always unique fit line (unlike absolute distance)
- directly solvable
- small errors even reduced

Cons:

- large errors over-weighted
- more sensitive to outliers (than abs method)
- duplicates influence outcome

$$\sum_{i=0}^{N} \left(y_i - f(x_i)\right)^2 \rightarrow \min$$

Statistics' point of view: "regression"

- deterministic $y = f(x)$ → stochastic $y = f(x) + \varepsilon$

- standard deviation minimized
- minimum uncertainty
- maximal plausability

➔ only statistical, not causal relations

Dispute

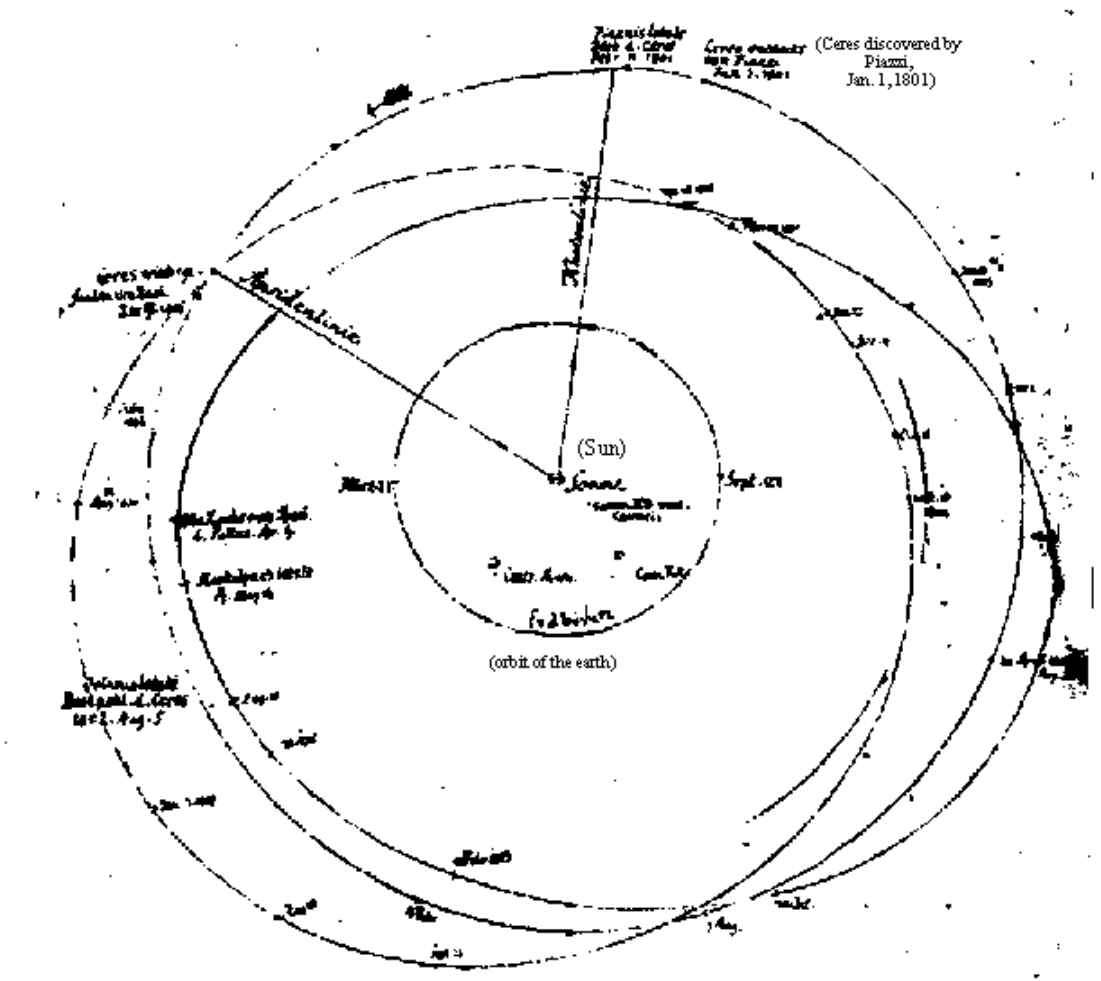
Adrien Marie Legendre, 1805
Nouvelles methodes pour la determination des orbites des come
→ first clear explanation with working example

Carl Friedrich Gauß, 1809
Theoria motus corporum coelestium in sectionibus conicus solem ambientium
→ much more refined and applicable
claimed use since 1795

Usage in astronomy and geodesy

- early 1801 Joseph Piazzi discovered a planetoid
- lost after 41 days of observation in Sun’s light
- "Where’s Ceres?" challenge
- Euler, Lagrange, Laplace, …
- late 1801 found very close to prediction

- 1793 republican France defines meter
- 10 mio. part of meridian quadrant
- angle measurements from Dunkirk to Barcelona
- calculated Earth‘s ellipticity
- notion of curvature



Sketch of the orbits of Ceres and Pallas (nachlaß Gauß, Handb. 4). Courtesy of Universitätsbibliothek Göttingen.

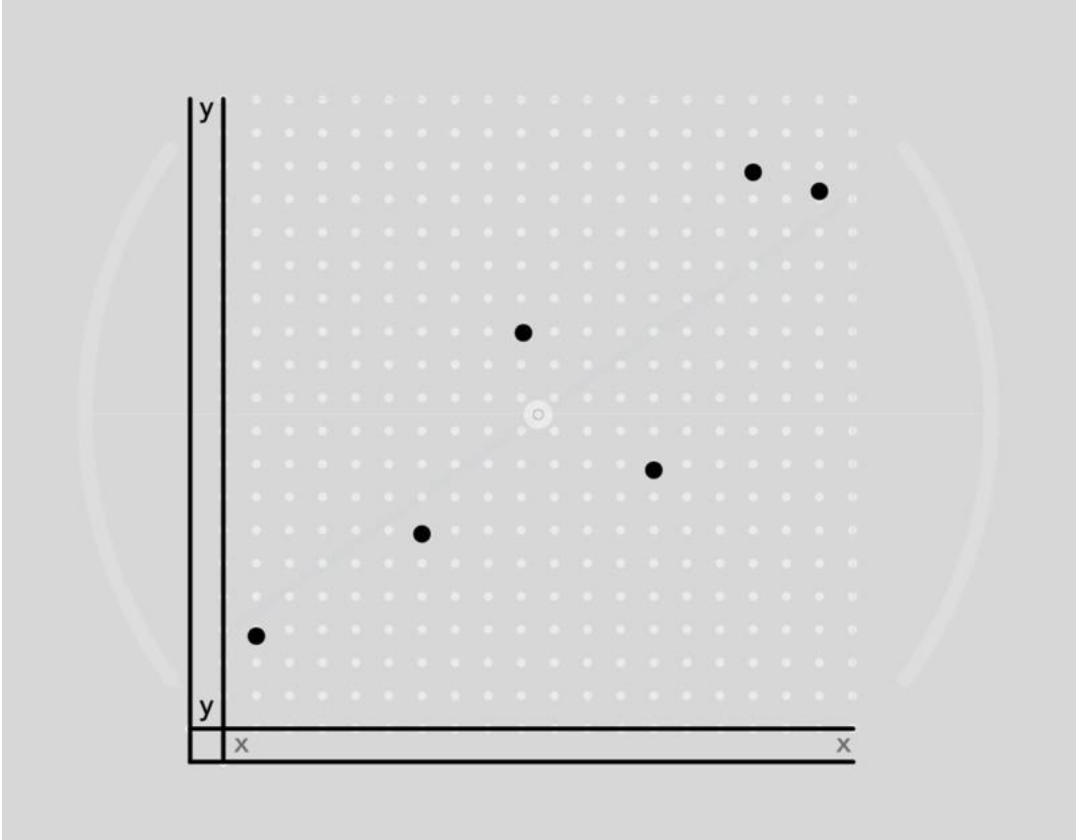
Physics' perspective

Balancing forces of springs

Really?

Hooke's law: $F \sim \Delta x$

→ Why is quadratic right and not absolute?



Variants

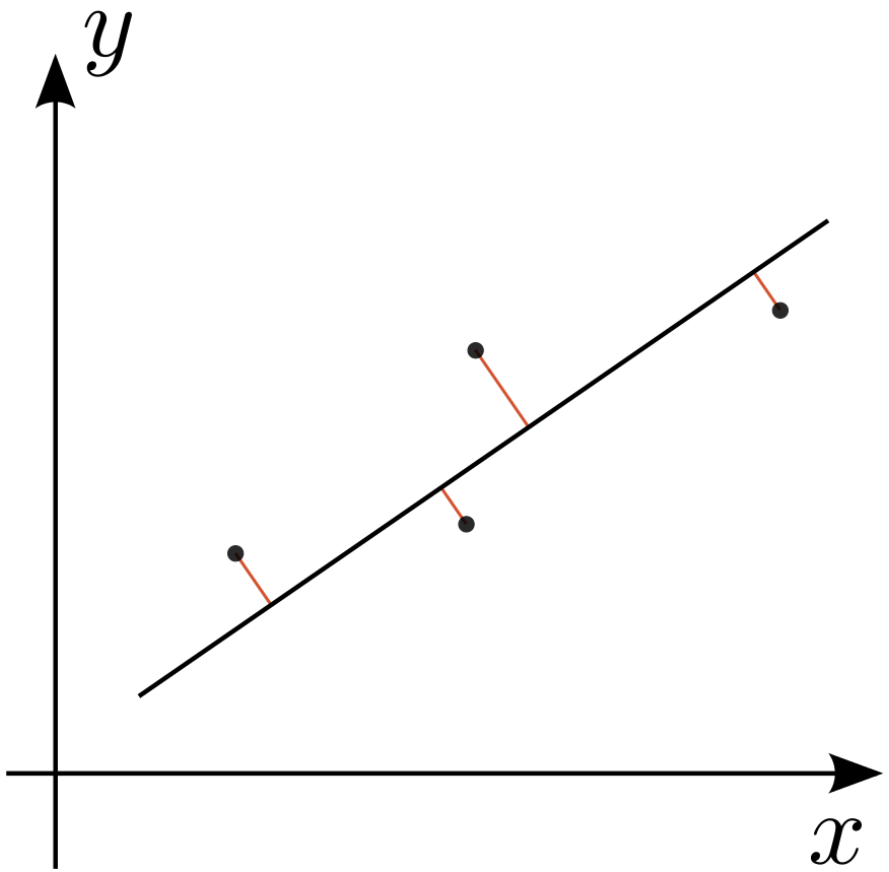
- "linear" does not mean "line"
  → linear combinations of arbitrary functions
  (remember what you get: simple ansatz is better)

- non-linear regression has unknown inside basis functions
  → iteratively solving adapted linear problems

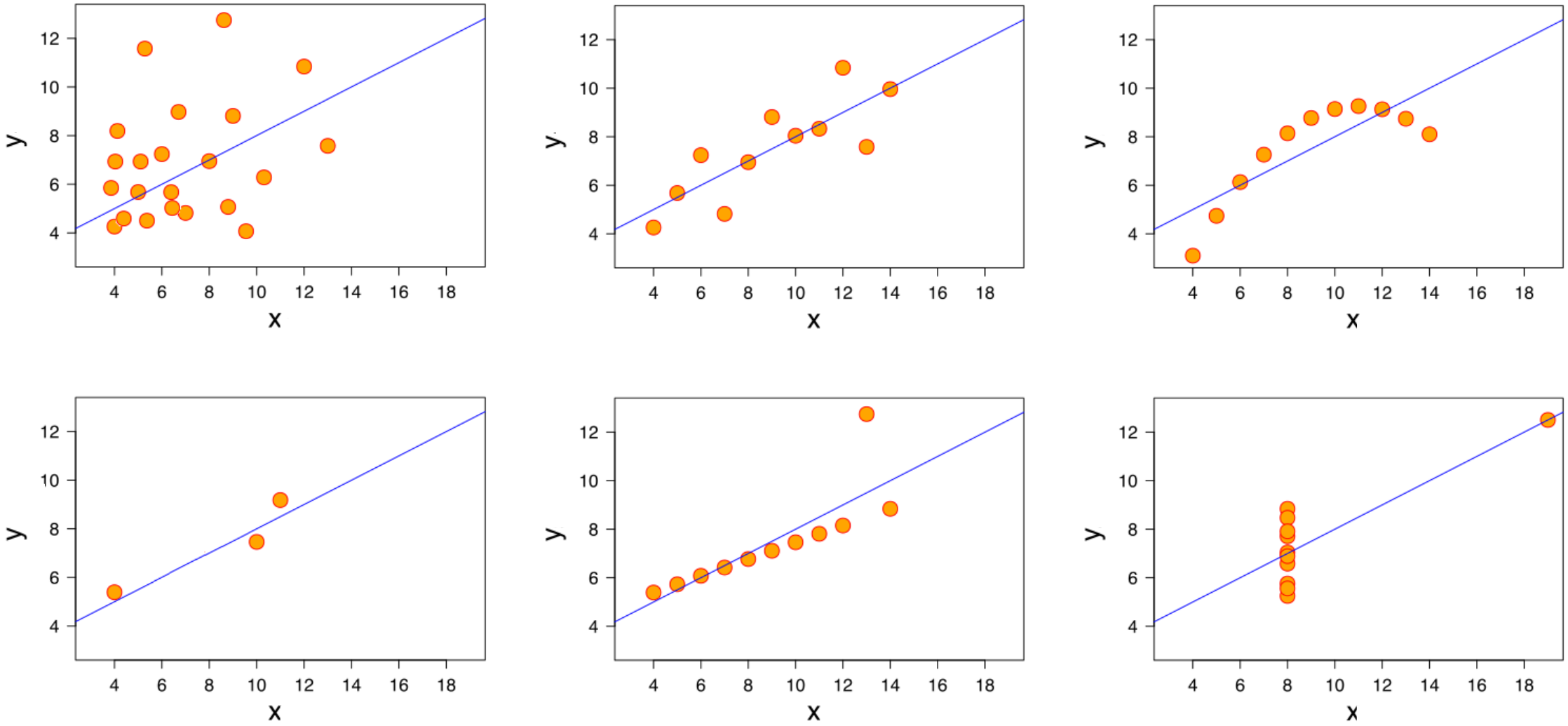
$$a \sin x + b \cos 2x + cx^3 \quad \leftrightarrow \quad ab,\ \sin ax,\ 2^b,\ x^c$$

- L_1 regression keeps distance proportions
  → but: possibly not unique

- Deming regression treats both variables equally
  → more "geometric"



Do not "make" it fit!



Tasks

- pick a day of your measurements
- get rid of the night according to your previously explored scheme
- calculate a fit
- plot the original data and the fit together

- repeat the computation for the other days
- arrange all fit curves in one 3D plot

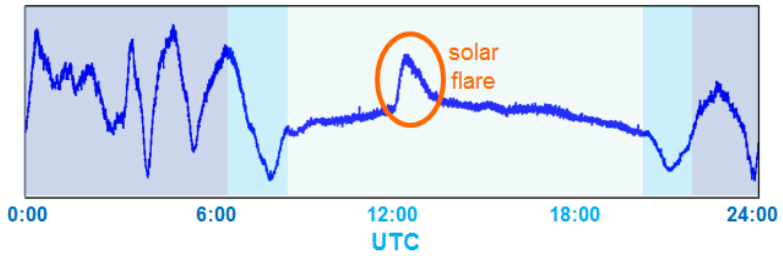
Useful:
- scipy.optimize
- SciPy is no SciFi – use it, but be aware of its power!

Improvements

Outliers

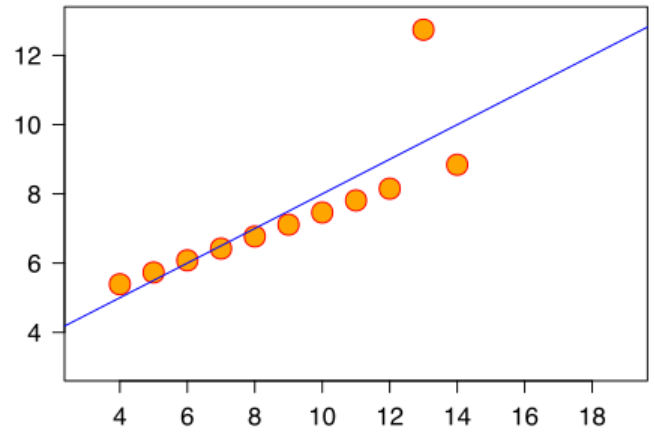
Remember sensitivity of LSM to outliers!

How can one detect and remove these…
…without manual selection?



Algorithmic approach:

- calculate the "errors" between measurement and fit
- determine *min* and *max* , as well as *mean*
- find a rule…
- e.g. delete all points with error greater *mean+K*(mean-min)*
- refit the reduced data set



→ actual deviation from the fit instead fixed value
→ no subjective choice such as eliminating flares or outages

Maximum

- ToD of different days are not "comparable"
  → fitting parameters may be distorted from one day to the other
  → the fitted solution may have an undesired shape (like maximum in the morning or even two maxima)

- one can normalise the time
  ➔ for each day, fit *amplitude* over *time-zenith* instead over *time*

- incorporates physically plausible information in the model
  → attach the maximum to the zenith
  → use *cos* directly instead of a shifted *sin*
  → simplify ansatz function by utilising approximation theorems

Symmetry

- polynomes or phase-shifted trigonometrics may be placed "wrongly"

- forcing the symmetry is possible
  → respect the axis in the model?

- Yes, if time is normalised!

- one can simply use ax^4+cx^2+e instead of $ax^4+bx^3+cx^2+dx+e$

  → reduce parameter number

- enforce known relations
  → match morning and afternoon lengths via deterministic formula directly inside the model