Fitting a model to data
(part 1)

based on
David W. Hogg, Jo Bovy, Dustin Lang (2010)
https://arxiv.org/abs/1008.4686

Jan Skowron, SJC, OA UW, 06.12.2022

Least-square fitting

- good only if:
  - negligible uncertainties in one direction (eg. x)
  - Gaussian uncertainties in another direction (eg. y)

- rarely meet in practice

- goal: framework to consider outliers, arbitrarily covariant 2d uncertainties, various uncertainties distributions, etc.

Generative model

- fitting is non-arbitrary
- and model permits direct computation of the likelihoods and posterior distribution
  - this allows for subsequent marginalization (of posterior) over unimportant parameters

Straight line fits

- truly linear relations are rare in physics
- any transformation of coordinates often moves linear relation away from linearity
- even if looks linear – in the absence of theoretical reason for it – probably isn't

- fitting a straight line will introduce systematic errors, and can introduce overconfidence in the predicted values elsewhere

Simple straight-line fits are often useless

- providing solely a *slope and intercept* of the *best-fit* model rarely can be used by other researchers
- not for prediction of new data
- not for simulations

Weighted linear least-square fitting:

- set of points *(x_i, y_i)*
- together with Gaussian uncertainties in *y* direction σ_{yi}
- *perfect* knowledge in *x* direction
- and a model:

$$f(x) = m\,x + b$$

Matrices

$$\boldsymbol{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \cdots \\ y_N \end{bmatrix} ,$$

- vector

$$\boldsymbol{A} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \multicolumn{2}{c}{\cdots} \\ 1 & x_N \end{bmatrix} ,$$

- functions in linear model

  $f(x) = b\, f_0(x) + m\, f_1(x) + \ldots$

$$\boldsymbol{C} = \begin{bmatrix} \sigma_{y1}^2 & 0 & \cdots & 0 \\ 0 & \sigma_{y2}^2 & \cdots & 0 \\ & & \cdots & \\ 0 & 0 & \cdots & \sigma_{yN}^2 \end{bmatrix}$$

- covariance matrix (could be non-diagonal, if there are covariances among uncertainties of different points)

If not-over-constrained (set of linear equations)

$$\mathbf{Y} = \mathbf{A}\ \mathbf{X}$$

- where
  - Y – vector (of values)
  - X – parameters of model
  - A – model
- solution:

$$\mathbf{Y} = \mathbf{A}\ \mathbf{X}$$

$$\mathbf{A}^{-1}\ \mathbf{Y} = \mathbf{A}^{-1}\ \mathbf{A}\ \mathbf{X} \quad \text{- multiply by } \mathbf{A}^{-1}$$

$$\mathbf{A}^{-1}\ \mathbf{Y} = \mathbf{X}$$

$$\mathbf{X} = \mathbf{A}^{-1}\ \mathbf{Y}$$

“Best-fit values”

- given by **X**:

$$\begin{bmatrix} \mathrm{b} \\ \mathrm{m} \end{bmatrix} = \boldsymbol{X} = \left[\boldsymbol{A}^{\top} C^{-1} \boldsymbol{A}\right]^{-1} \left[\boldsymbol{A}^{\top} C^{-1} \boldsymbol{Y}\right]$$

if not-overconstained:

$$\mathbf{Y} = \mathbf{A}\ \mathbf{X}$$

$$\mathbf{A}^{-1}\ \mathbf{Y} = \mathbf{A}^{-1}\ \mathbf{A}\ \mathbf{X}$$

$$\mathbf{A}^{-1}\ \mathbf{Y} = \mathbf{X}$$

$$\mathbf{X} = \mathbf{A}^{-1}\ \mathbf{Y}$$

if over-constrained:

1) weight points with inverse covariance matrix:

$$\mathbf{C}^{-1}\ \mathbf{Y} = \mathbf{C}^{-1}\ \mathbf{A}\ \mathbf{X}$$

2) reduce dimensionality by multiplying with $\mathbf{A}^{\mathrm{T}}$:

$$\mathbf{A}^{\mathrm{T}}\ \mathbf{C}^{-1}\ \mathbf{Y} = \mathbf{A}^{\mathrm{T}}\ \mathbf{C}^{-1}\ \mathbf{A}\ \ \mathbf{X}$$

$$(\mathbf{A}^{\mathrm{T}}\ \mathbf{C}^{-1})\ \mathbf{Y} = (\mathbf{A}^{\mathbf{T}}\ \mathbf{C}^{\mathbf{-1}}\ \mathbf{A})\ \ \mathbf{X}$$

$$\mathbf{Y'} = \mathbf{A'}\ \mathbf{X}$$

This minimizes χ^2

- total squared error scaled by uncertainties:

$$\chi^2 = \sum_{i=1}^{N} \frac{[y_i - f(x_i)]^2}{\sigma_{yi}^2} \equiv [\boldsymbol{Y} - \boldsymbol{A}\,\boldsymbol{X}]^\top \boldsymbol{C}^{-1}\,[\boldsymbol{Y} - \boldsymbol{A}\,\boldsymbol{X}]$$

- if uncertainties are Gaussian and correctly scaled, the matrix:

$$\left[\boldsymbol{A}^\top \boldsymbol{C}^{-1}\,\boldsymbol{A}\right]^{-1}$$

- is the covariance matrix for parameters in **X**

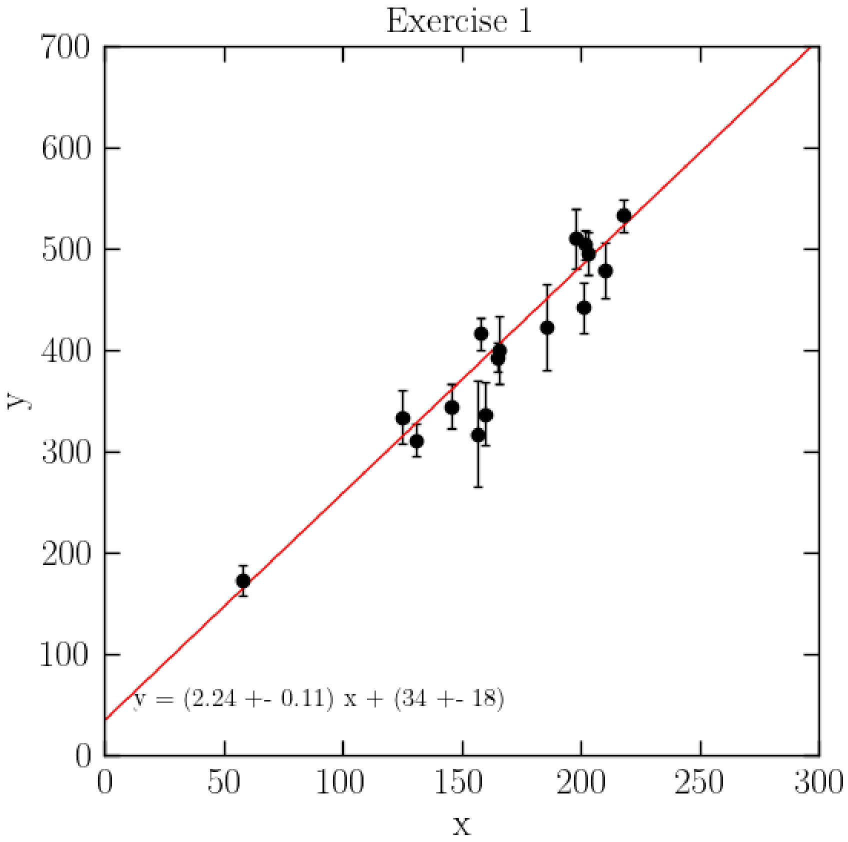
Exercises

- Exercise 1 – fit line to last 16 data points from the file data.txt (ignore uncertainties other than σ_y).
- Exercise 2 – fit line to all 20 data points
- Exercise 3 – extend model to fit parabola (to original points - last 16 points)

data.txt

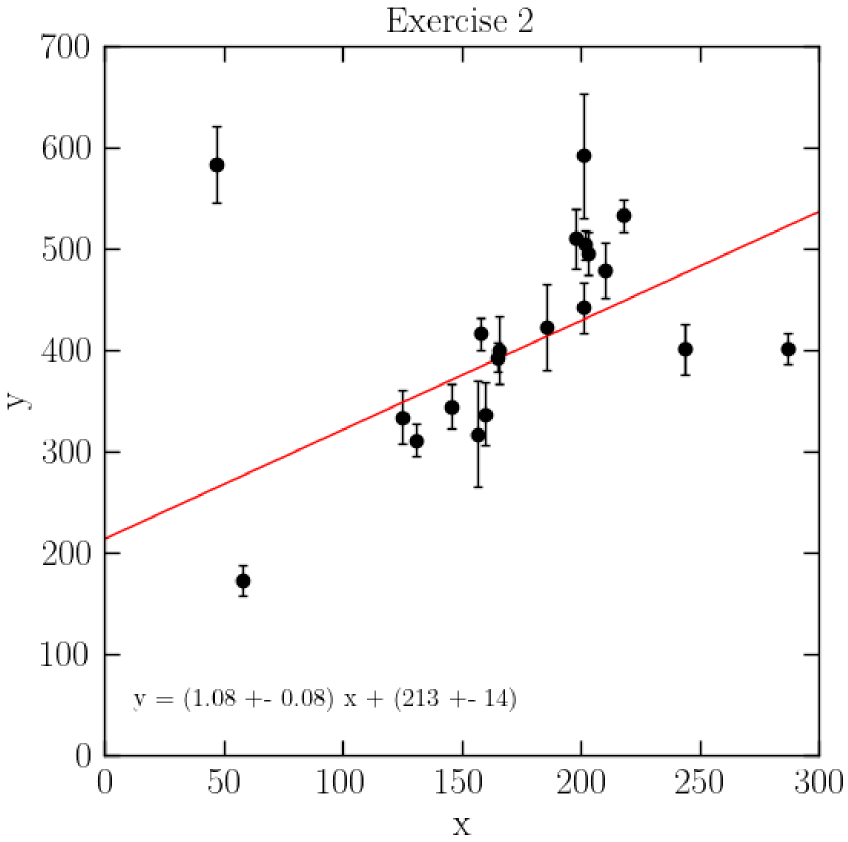
```
#  1    2      3      4        5        6
# No    x      y   sigma_y  sigma_x  rho_xy
   1   201    592    61         9     -0.84
   2   244    401    25         4      0.31
   3    47    583    38        11      0.64
   4   287    402    15         7     -0.27
   5   203    495    21         5     -0.33
   6    58    173    15         9      0.67
   7   210    479    27         4     -0.02
   8   202    504    14         4     -0.05
   9   198    510    30        11     -0.84
  10   158    416    16         7     -0.69
  11   165    393    14         5      0.30
  12   201    442    25         5     -0.46
  13   157    317    52         5     -0.03
  14   131    311    16         6      0.50
  15   166    400    34         6      0.73
  16   160    337    31         5     -0.52
  17   186    423    42         9      0.90
  18   125    334    26         8      0.40
  19   218    533    16         6     -0.78
  20   146    344    22         5     -0.56
# where the full uncertainty covariance matrix for each data point is given by:
#
#  |                                                 |
#  |       sigma_x^2          rho_xy*sigma_x*sigma_y |
#  |                                                 |
#  | rho_xy*sigma_x*sigma_y          sigma_y^2       |
```

Ex 1



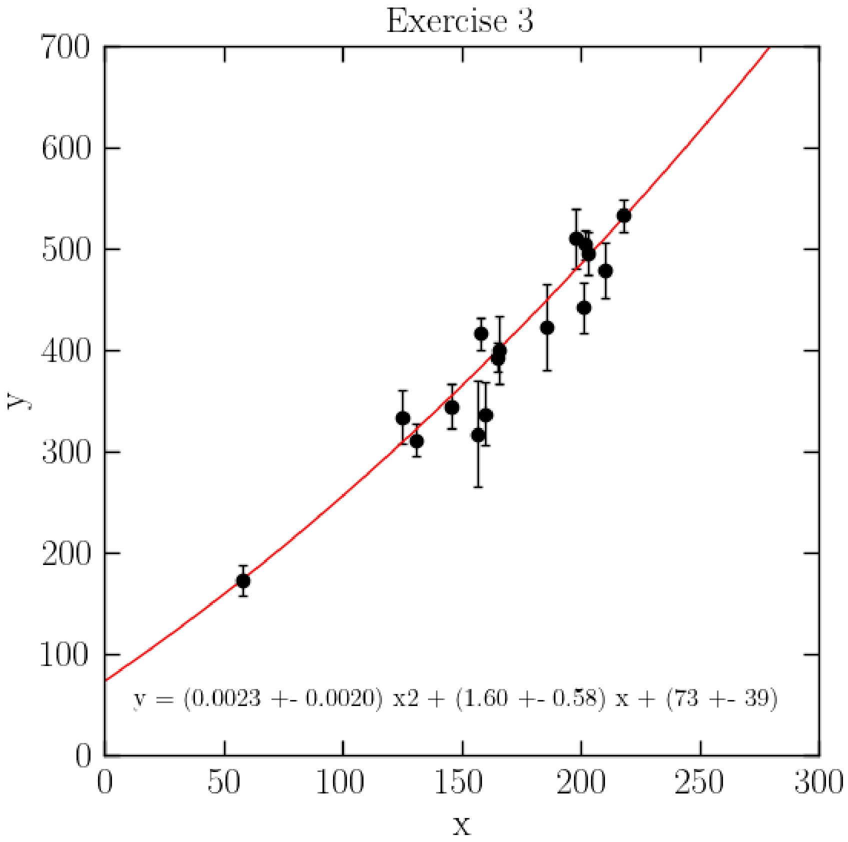
What is the standard uncertainty variance σ_m on the slope of the line?

Ex 2



What is the standard uncertainty variance σ_m on the slope of the line? Is there anything you don't like about the result? Is there anything different about the new points you have included beyond those used in Exercise 1?

Ex 3


Exercise 3
700
600
500
400
300
200
100
0
y
0
50
100
150
200
250
300
x
y = (0.0023 +- 0.0020) x2 + (1.60 +- 0.58) x + (73 +- 39)

Better way - Objective function

- all knowledge about the problem in one function
  - justified
  - scalar
  - monotonically represents the “quality of fit”

- and subsequently, procedures:
  - to find optimum
  - and to find posterior around optimum

Generative model for the data

  - parametrized
  - quantitative
- description of a statistical procedure that could reasonably have generated the data

Simple example

- data *really* do come from perfect model, this exact line: $$y = f(x) = mx + b$$
- only reason the data deviate from this narrow line is that the small offset was added
- this offset was drawn from a Gaussian distribution (with mean = 0, and known variance σ_y^2)

Simple example

- in this model, the probability of measuring given data point *y*$_i$ at given position *x*$_i$ is simply:

$$p(y_i|x_i,\sigma_{yi},m,b)=\frac{1}{\sqrt{2\,\pi\,\sigma_{yi}^2}}\,\exp\left(-\frac{[y_i-m\,x_i-b]^2}{2\,\sigma_{yi}^2}\right)$$

- in this case, the likelihood of observing the dataset we have observed is given by:

$$\mathscr{L}=\prod_{i=1}^{N}\,p(y_i|x_i,\sigma_{yi},m,b)$$

Simple example

- finding a line, is to find parameters *(m, b)* that maximize this likelihood

$$\mathscr{L} = \prod_{i=1}^{N} p(y_i | x_i, \sigma_{yi}, m, b)$$

- We can simplify this:

$$\ln \mathscr{L} \;=\; K - \sum_{i=1}^{N} \frac{[y_i - m\,x_i - b]^2}{2\,\sigma_{yi}^2} \;=\; K - \frac{1}{2}\,\chi^2$$

A justification! Minimizing χ^2, in fact, maximizes likelihood

Bayes theorem

- of course mind the prior, if important in the studied range of parameters:

likelihood

prior

$$p(m,b|\{y_i\}_{i=1}^N, I) = \frac{p(\{y_i\}_{i=1}^N|m,b,I)\,p(m,b|I)}{p(\{y_i\}_{i=1}^N|I)}$$

posterior

evidence

I - all information we have (like *x_i, σ_i*, etc.)

$\{y_i\}$ – all data we have

m, b – parameters of the model

p(m,b|I) – prior probability distribution of parameters without knowing data

Exercises

- Exercise 4 – calculate mean
- Exercise 5 – derivative of χ^2 in matrix form

The solution of **Exercise 4** by JS. t_i are the measurements and σ_i are the uncertainties. We assume there exists a true value of T that we tried to measure, but the measurements were scattered as a Gaussian process. We may then write, that the probability of observing a value t_i is:

$$p(t_i|T,\sigma_i) = \frac{1}{\sqrt{2\pi}\sigma_i}\exp\left(-\frac{(t_i-T)^2}{2\sigma_i^2}\right)$$

Then, the total likelihood of observing the whole sample is:

$$\mathcal{L} = \prod_{i=1}^{N} p(t_i|T,\sigma_i) = \prod_{i=1}^{N}\frac{1}{\sqrt{2\pi}\sigma_i}\exp\left(-\frac{(t_i-T)^2}{2\sigma_i^2}\right)$$

Lets take logarithm of both sides to get rid of exp:

$$\log\mathcal{L} = K - \sum_{i=1}^{N}\left(\frac{(t_i-T)^2}{2\sigma_i^2}\right)$$

To find the value of T that maximizes the likelihood, we take:

$$\frac{\partial\log\mathcal{L}}{\partial T} = \sum_{i=1}^{N}\frac{\partial}{\partial T}\left(\frac{(t_i-T)^2}{2\sigma_i^2}\right) = \sum_{i=1}^{N}\left(\frac{-2(t_i-T)}{2\sigma_i^2}\right) = \sum_{i=1}^{N}\frac{T-t_i}{\sigma_i^2}$$

The extremum is for such T that $\partial\log\mathcal{L}/\partial T = 0$, so:

$$\sum_{i=1}^{N}\frac{T-t_i}{\sigma_i^2} = 0$$

$$\sum_{i=1}^{N}\frac{T}{\sigma_i^2} = \sum_{i=1}^{N}\frac{t_i}{\sigma_i^2}$$

$$T = \sum_{i=1}^{N}\frac{t_i}{\sigma_i^2}\Big/\sum_{i=1}^{N}\frac{1}{\sigma_i^2}$$

which gave us the standard expression for a weighted mean.

The solution of **Exercise 5** by JS. The $\mathbf{X}$ is a vector of parameters (which multiply functions of the linear model), $\mathbf{A}$ is the matrix of model function values (in columns) evaluated for each coordinate (x_i, in rows), and $\mathbf{Y}$ is the vector of measurements (y_i) at each coordinate.

$$\begin{aligned}\chi^2 = \sum_{i=1}^{N} \frac{[y_i - f((x_i)]^2}{\sigma_{yi}^2} &= [\mathbf{Y} - \mathbf{AX}]^T \mathbf{C}^{-1} [\mathbf{Y} - \mathbf{AX}] = \\ &= \mathbf{Y}^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{Y} - \mathbf{Y}^{\mathrm{T}}\mathbf{C}^{-1}[\mathbf{AX}] - [\mathbf{AX}]^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{Y} + [\mathbf{AX}]^{\mathrm{T}}\mathbf{C}^{-1}[\mathbf{AX}] = \\ &= \mathbf{Y}^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{Y} - 2[\mathbf{AX}]^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{Y} + [\mathbf{AX}]^{\mathrm{T}}\mathbf{C}^{-1}[\mathbf{AX}]\end{aligned}$$

where we used fact that $\mathbf{Y}^{\mathrm{T}}\mathbf{C}^{-1}[\mathbf{AX}] \equiv [\mathbf{AX}]^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{Y}$ as both $\mathbf{Y}$ and $[\mathbf{AX}]$ are vectors and not matrices.

We try to find extremum of χ^2 in respect to $\mathbf{X}$, hence we calculate:

$$\frac{\partial \chi^2}{\partial \mathbf{X}} = \frac{\partial}{\partial \mathbf{X}}(\mathbf{Y}^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{Y} - 2[\mathbf{AX}]^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{Y} + [\mathbf{AX}]^{\mathrm{T}}\mathbf{C}^{-1}[\mathbf{AX}])$$

Term $\mathbf{Y}^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{Y}$ is constant in respect to $\mathbf{X}$ so can be omitted.

$$\begin{aligned}\frac{\partial \chi^2}{\partial \mathbf{X}} &= -2\frac{\partial}{\partial \mathbf{X}}([\mathbf{AX}]^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{Y}) + \frac{\partial}{\partial \mathbf{X}}([\mathbf{AX}]^{\mathrm{T}}\mathbf{C}^{-1}[\mathbf{AX}]) = \\ &= -2\mathbf{A}^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{Y} + 2\mathbf{A}^{\mathrm{T}}\mathbf{C}^{-1}[\mathbf{AX}]\end{aligned}$$

Equating $\frac{\partial \chi^2}{\partial \mathbf{X}}$ to 0 yields:

$$\begin{gathered}-2\mathbf{A}^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{Y} + 2\mathbf{A}^{\mathrm{T}}\mathbf{C}^{-1}[\mathbf{AX}] = 0 \\ \mathbf{A}^{\mathrm{T}}\mathbf{C}^{-1}[\mathbf{AX}] = \mathbf{A}^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{Y} \\ \mathbf{X} = [\mathbf{A}^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{A}]^{-1}\mathbf{A}^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{Y}\end{gathered}$$

Hence, the value of parameters vector $\mathbf{X}$ which gives minimal χ^2 is equal to $[\mathbf{A}^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{A}]^{-1}\mathbf{A}^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{Y}$.