Fitting a model to data (part 2)

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

Jan Skowron, SJC, OA UW, 20.12.2022

Reminder: Generative model in a simple linear fit

- 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:

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

Reminder: Likelihood of the observed data

- 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

Outliers

- points deviate from the linear relation because of:
  - unmodeled experimental uncertainty
  - not-included rare sources of noise
  - or model doesn't apply to the data

Outliers

- approaches:

1) *objectively* **reject** outliers (or become insensitive to *bad* points)
2) **model** the data-point uncertainties in order to permit larger deviations

- both better than manual rejection, for reasons of:

1) objectivity
2) reproducibility

eg. “sigma clipping” is not the best, as it is a procedure but not a result of justifiable modeling

Outliers – parameters to descibe

- adding boolean vector, if a given point was good or bad
  - value 1 if point is good
  - value 0 if point is bad

  a vector $\{q_i\}_{i=1}^{N}$ of ones and zeros (N additional parameters!)
- adding a parameter: prior probability P_{bad} (chance) of how often an outlier appear in the observed data
- and for example: parameters describing the distribution of outliers, with (mean, variance) = (Y_{bad}, V_{bad})

  All this N+3 new parameters are used, in order to create a “generative model” for *all data* points observed in the sample

Parameters of outliers

- these additional parameters: $\{q_i\}_{i=1}^{N}$, P_{bad}, Y_{bad} and V_{bad} do not have to be know in advance
- in principal, one can *fit* for those
- and then, if not interested, we can *marginalize out* these from the posterior distribution of important parameters

Likelihood

L = p(data | parameters)

$$\mathscr{L} \equiv p(\{y_i\}_{i=1}^{N} | m, b, \{q_i\}_{i=1}^{N}, Y_{\rm b}, V_{\rm b}, I)$$

- probability of good point (Gaussian around the model)

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

- probability of bad point (Gaussian with sigma V_b and around some typical value Y_b)

$$\frac{1}{\sqrt{2\,\pi\,[V_{\rm b} + \sigma_{yi}^2]}}\,\exp\left(-\frac{[y_i - Y_{\rm b}]^2}{2\,[V_{\rm b} + \sigma_{yi}^2]}\right)$$

- Multiply probability for each point
  - good (*fg* – foreground)
  - bad (*bg* – background)

$$\mathscr{L} = \prod_{i=1}^{N} \left[p_{\rm fg}(\{y_i\}_{i=1}^{N} | m, b, I) \right]^{q_i} \left[p_{\rm bg}(\{y_i\}_{i=1}^{N} | Y_{\rm b}, V_{\rm b}, I) \right]^{[1-q_i]}$$

$$\mathscr{L} = \prod_{i=1}^{N} \left[\frac{1}{\sqrt{2\,\pi\,\sigma_{yi}^2}} \exp\left(-\frac{[y_i - m\,x_i - b]^2}{2\,\sigma_{yi}^2} \right) \right]^{q_i} \times \left[\frac{1}{\sqrt{2\,\pi\,[V_{\rm b} + \sigma_{yi}^2]}} \exp\left(-\frac{[y_i - Y_{\rm b}]^2}{2\,[V_{\rm b} + \sigma_{yi}^2]} \right) \right]^{[1-q_i]}$$

We have to penalize data rejection

We add prior on q_i:

total prior = prior on q_i x rest of the prior

$$p(m, b, \{q_i\}_{i=1}^N, P_\mathrm{b}, Y_\mathrm{b}, V_\mathrm{b}|I) = p(\{q_i\}_{i=1}^N|P_\mathrm{b}, I)\, p(m, b, P_\mathrm{b}, Y_\mathrm{b}, V_\mathrm{b}|I)$$

where:

$$p(\{q_i\}_{i=1}^N|P_\mathrm{b}, I) = \prod_{i=1}^{N} [1 - P_\mathrm{b}]^{q_i}\, P_\mathrm{b}^{[1-q_i]}$$

A binomial probability of a given sequence of good and bad points assuming given outlier probability P_b.

[Assumption: single probability for all points – but could be modified to weight / diferenciate probability of points being outliers]

Marginalization...

- denote all parameter as:

$$\boldsymbol{\theta} \equiv (m, b, \{q_i\}_{i=1}^{N}, P_{\rm b}, Y_{\rm b}, V_{\rm b})$$

- posterior (Bayes):

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

- marginalization of posterior:

$$p(m, b|\{y_i\}_{i=1}^{N}, I) = \int \mathrm{d}\{q_i\}_{i=1}^{N}\, \mathrm{d}P_{\rm b}\, \mathrm{d}Y_{\rm b}\, \mathrm{d}V_{\rm b}\;\; p(\boldsymbol{\theta}\,|\,\{y_i\}_{i=1}^{N}, I)$$

parameters of the line

nuisance parameters

posterior

$$p(m, b|\{y_i\}_{i=1}^N, I) = \int \mathrm{d}\{q_i\}_{i=1}^N \, \mathrm{d}P_\mathrm{b} \, \mathrm{d}Y_\mathrm{b} \, \mathrm{d}V_\mathrm{b} \quad p(\boldsymbol{\theta} \,|\, \{y_i\}_{i=1}^N, I)$$

- sum over all 2^N possible settings of the $\{q_i\}_{i=1}^N$
- this takes a long time!
- However, in this simple model, it can be done analytically:
- lets imagine that after marginalization, each *i*th point is drawn from a "***mixture***" of a straight-line and outlier population (with $[1-P_b]$ and $[P_b]$ probabilit

- instead of:

$$\mathscr{L} \;=\; \prod_{i=1}^{N} \left[p_{\rm fg}(\{y_i\}_{i=1}^{N}|m,b,I))\right]^{q_i} \;\left[p_{\rm bg}(\{y_i\}_{i=1}^{N}|Y_{\rm b},V_{\rm b},I)\right]^{[1-q_i]}$$

- we have:

$$\mathscr{L} \;\equiv\; \prod_{i=1}^{N} \left[(1-P_{\rm b})\,p_{\rm fg}(\{y_i\}_{i=1}^{N}|m,b,I)) + P_{\rm b}\,p_{\rm bg}(\{y_i\}_{i=1}^{N}|Y_{\rm b},V_{\rm b},I)\right]$$

- instead of:

$$\mathscr{L} = \prod_{i=1}^{N} \left[\frac{1}{\sqrt{2\,\pi\,\sigma_{yi}^2}} \exp\left(-\frac{[y_i - m\,x_i - b]^2}{2\,\sigma_{yi}^2} \right) \right]^{q_i} \times \left[\frac{1}{\sqrt{2\,\pi\,[V_\mathrm{b} + \sigma_{yi}^2]}} \exp\left(-\frac{[y_i - Y_\mathrm{b}]^2}{2\,[V_\mathrm{b} + \sigma_{yi}^2]} \right) \right]^{[1-q_i]}$$

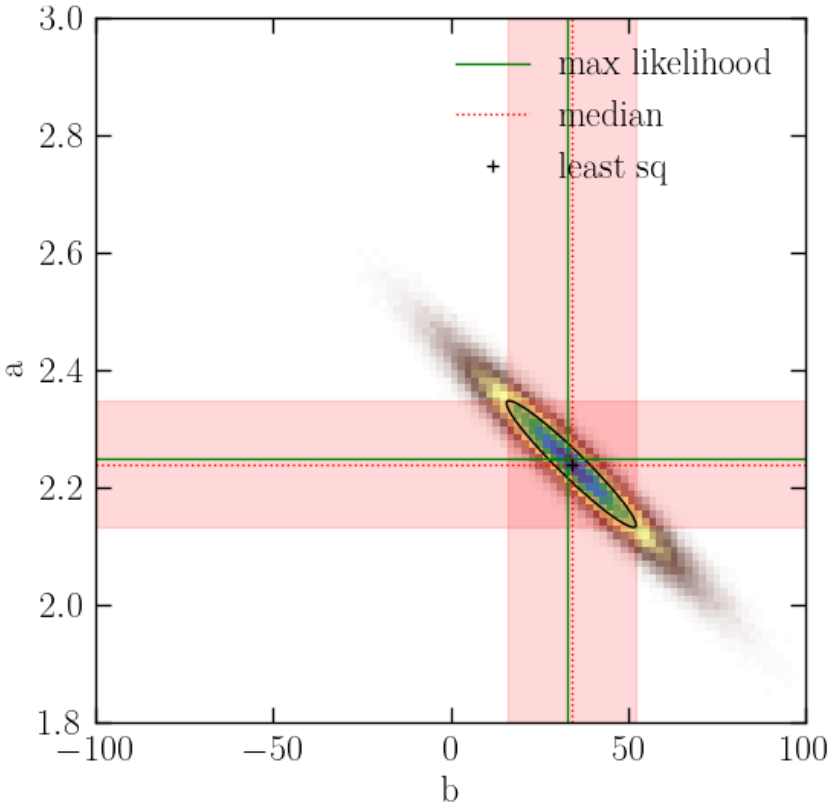
- we have:

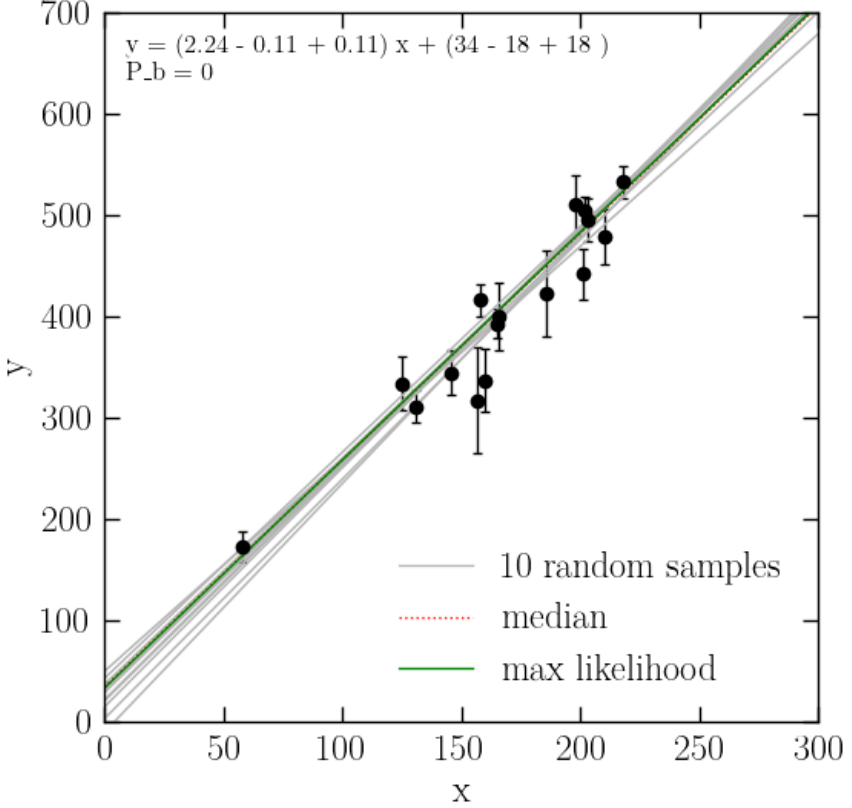
$$\mathscr{L} \propto \prod_{i=1}^{N} \left[\frac{1 - P_\mathrm{b}}{\sqrt{2\,\pi\,\sigma_{yi}^2}} \exp\left(-\frac{[y_i - m\,x_i - b]^2}{2\,\sigma_{yi}^2} \right) + \frac{P_\mathrm{b}}{\sqrt{2\,\pi\,[V_\mathrm{b} + \sigma_{yi}^2]}} \exp\left(-\frac{[y_i - Y_\mathrm{b}]^2}{2\,[V_\mathrm{b} + \sigma_{yi}^2]} \right) \right]$$

Exercises

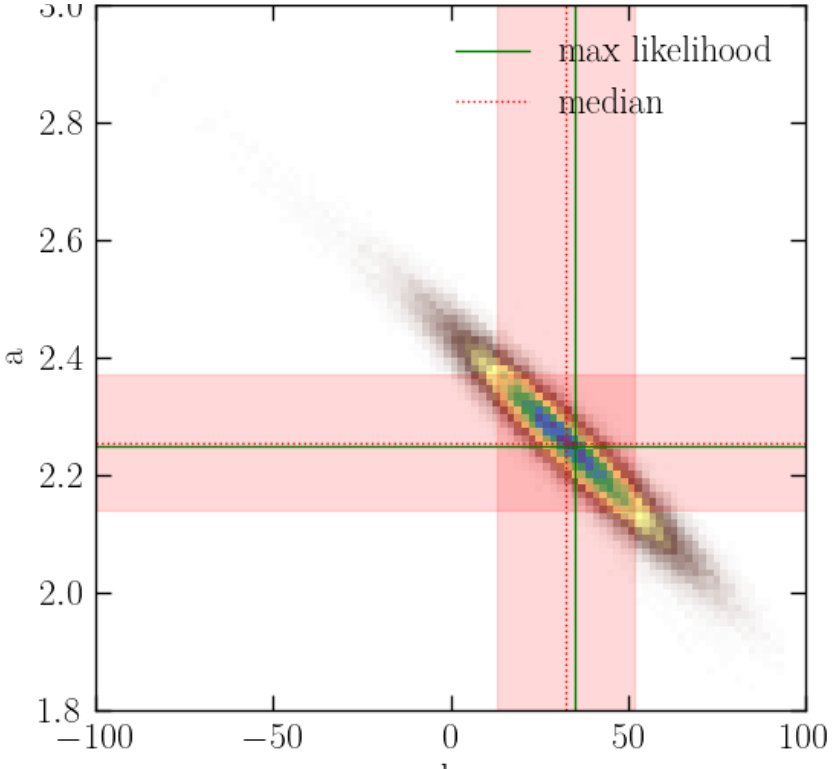
- Exercise 6 – fit model with outliers
- Exercise 7 – divide errorbars by 2 and fit again, see posterior for P_b
- Exercise 9 – compare parameters and their uncertainties between the fits with standard and shrunk errorbars

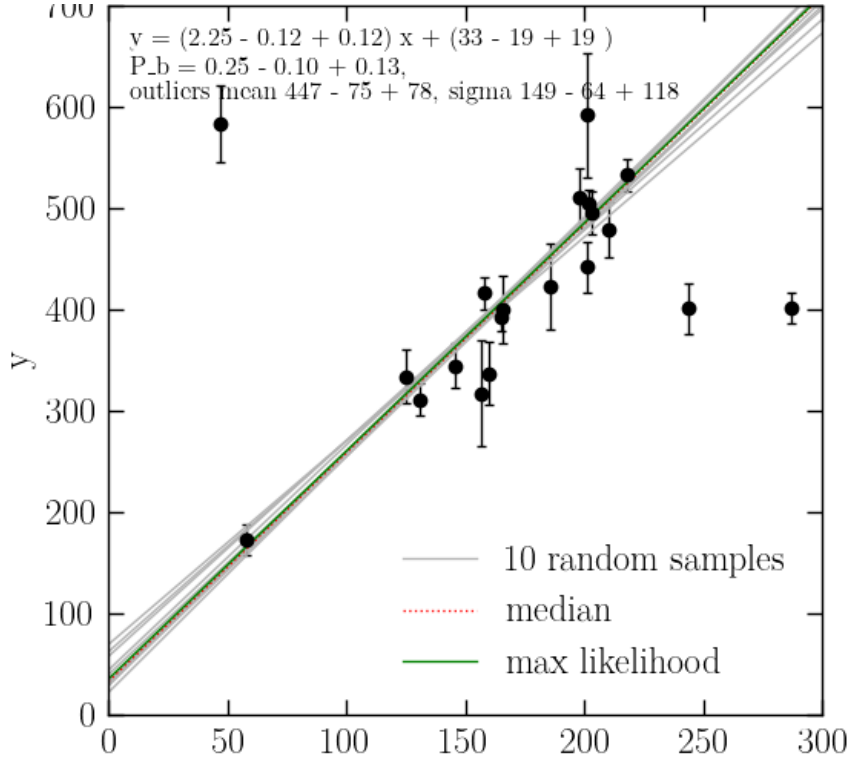
w/o outliers





Exercise 6





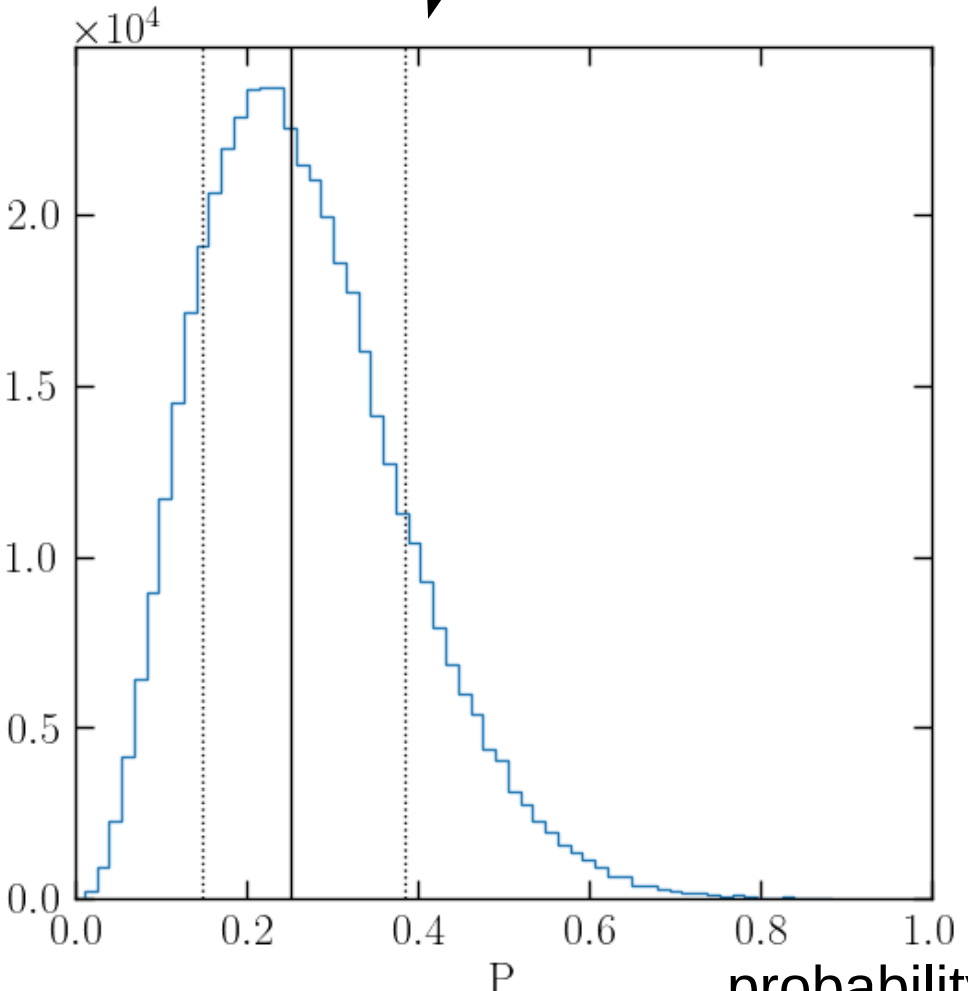
Exercise 6: fit model with outliers

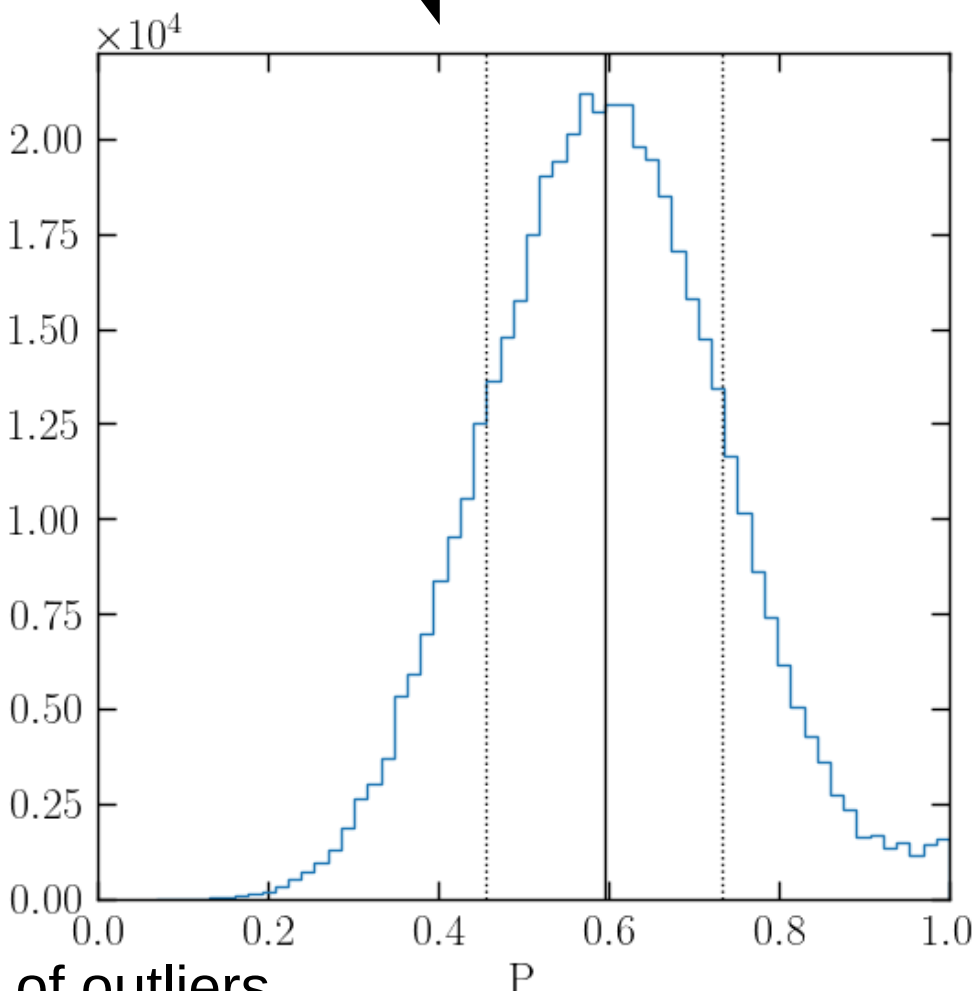
Exercise 7

(divide errorbars by 2 and fit again, see posterior for P_b)

unchanged errorbars

datapoint errorbars divided by 2





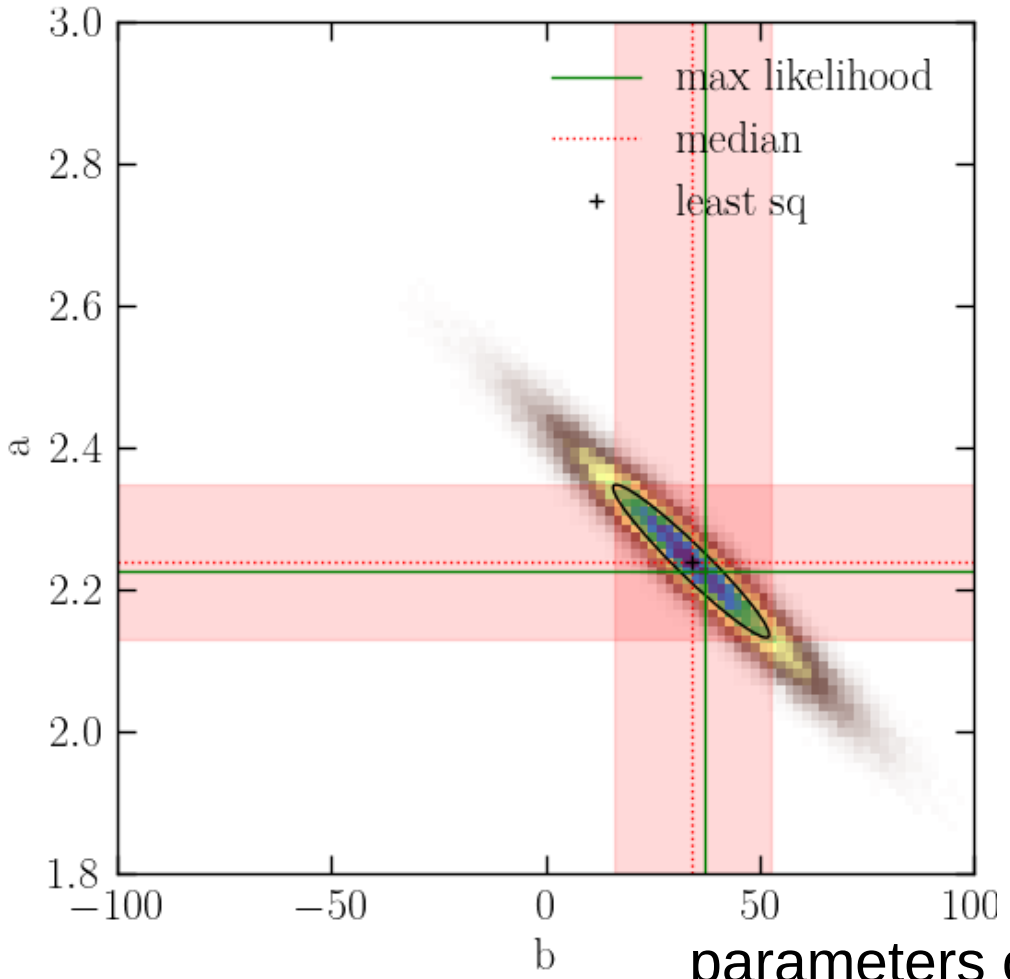
probability of outliers

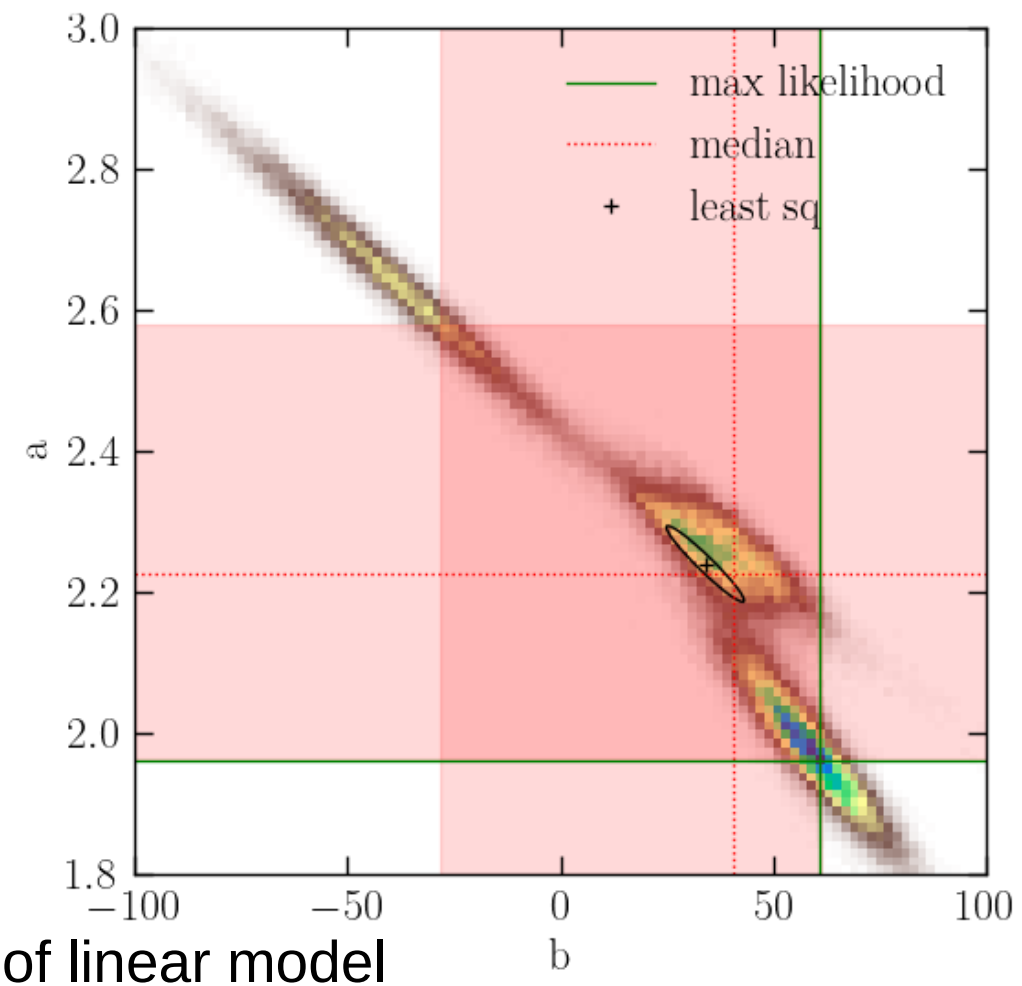
Exercise 9

compare parameters and their uncertainties between the fits with standard and shrunk errorbars

unchanged errorbars

datapoint errorbars divided by 2





parameters of linear model

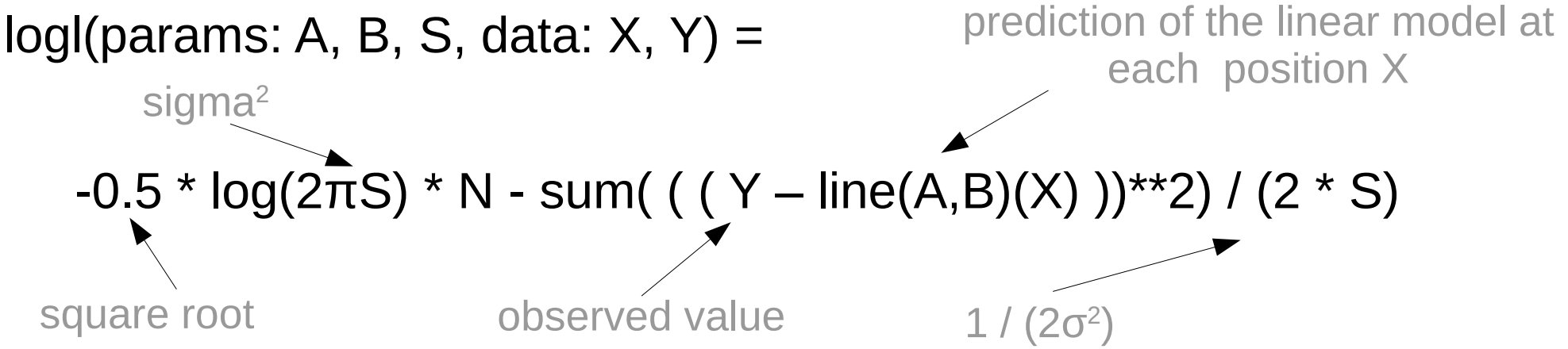
Also, the uncertainties can be fitted for!

- Exercise 12: assume that the uncertainties of the datapoints are not known
- let S be a new fit parameter, that gives the variance of the datapoints uncertainty
- we fit slope, intersection and variance

Exercise 12

- it is done with the same likelihood as in previous exercises
- (but do not forget the $1/\sqrt{(2\pi\sigma^2)}$ term!)

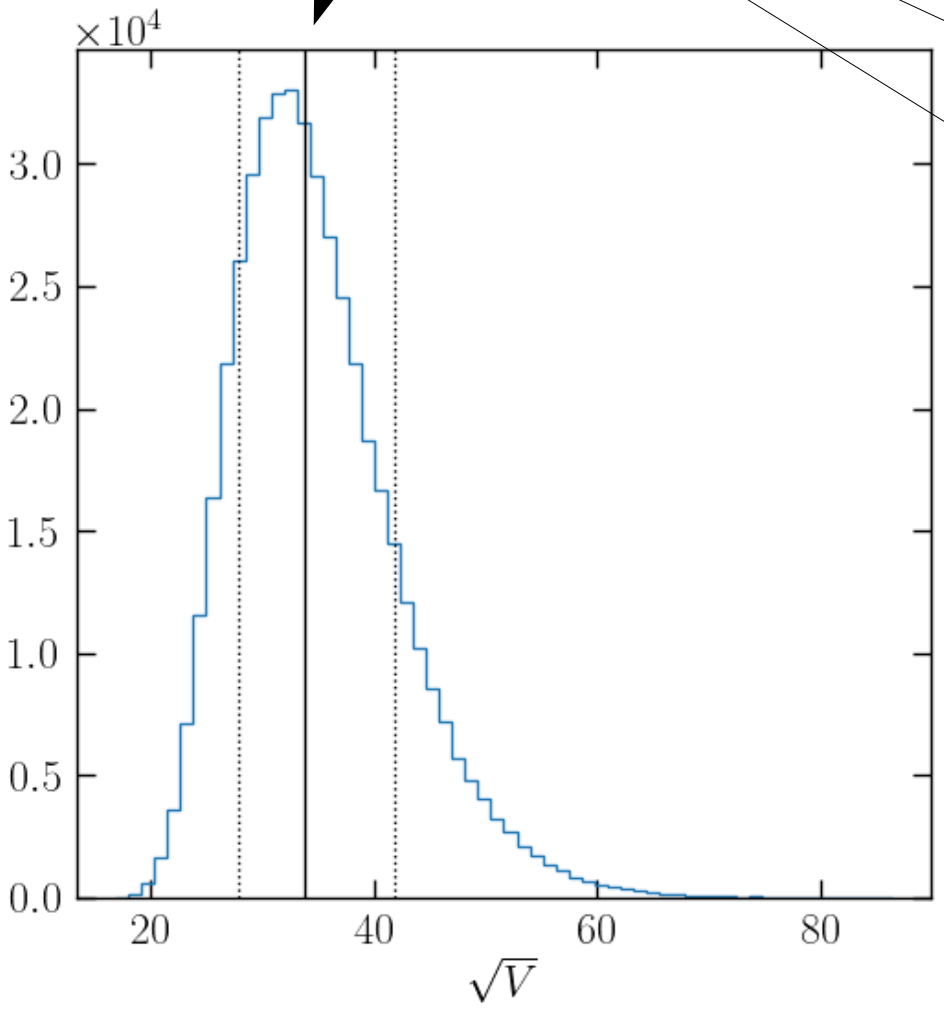
logl(params: A, B, S, data: X, Y) =

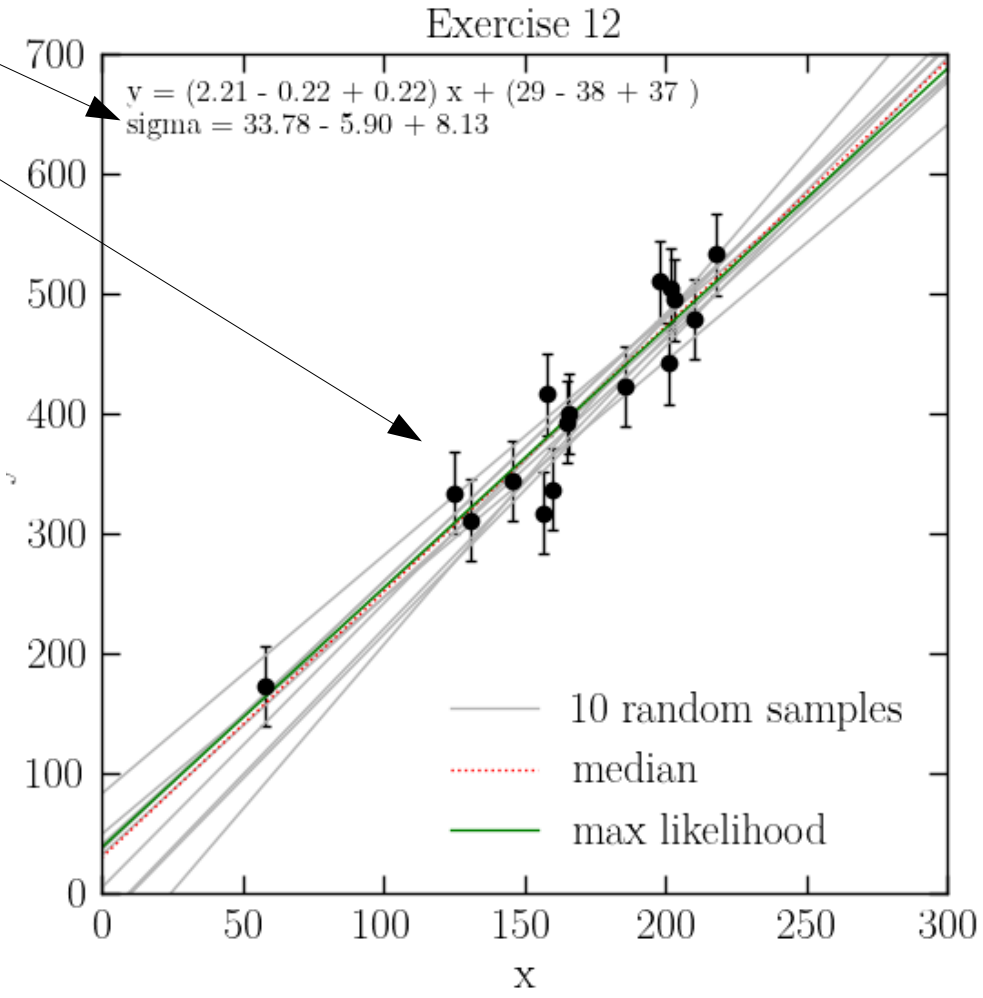


- where S is a variance of points ($\sigma_y^2 = S$)
- X and Y are coordinates of the datapoints
- A and B are slope and intersection of the linear model
- N is the count of the datapoints (eg. len(X))

Exercise 12

errorbars are fitted here





Non-Gaussian uncertainties

- easiest way to approach this is to simulate non-Gaussian errorbars with a sum of Gaussian distributions
- lets make a Gaussian mixture model:

some offset (typically 0)

$$p(y_i|x_i,\sigma_{yi},m,b) = \sum_{j=1}^{k} \frac{a_{ij}}{\sqrt{2\,\pi\,\sigma_{yij}^2}} \exp\left(-\frac{[y_i+\Delta y_{ij}-m\,x_i-b]^2}{2\,\sigma_{yij}^2}\right)$$

different variances of components

where:

$$\sum_{j=1}^{k} a_{ij} = 1$$

(similar to *[1-P]* and *P*)

Arbitrary 2d uncertainties

- Each datapoint has full 2d covariance matrix for errorbars:

$$(\sigma_{xi}^2, \sigma_{yi}^2) \quad + \quad \boldsymbol{S}_i \equiv \begin{bmatrix} \sigma_{xi}^2 & \sigma_{xyi} \\ \sigma_{xyi} & \sigma_{yi}^2 \end{bmatrix}$$

x_i y_i $sigma_{x,i}$ $sigma_{y,i}$ $cov_{xy,i}$

Probability of observing a point

at (x_i, y_i) when true value is (x, y):

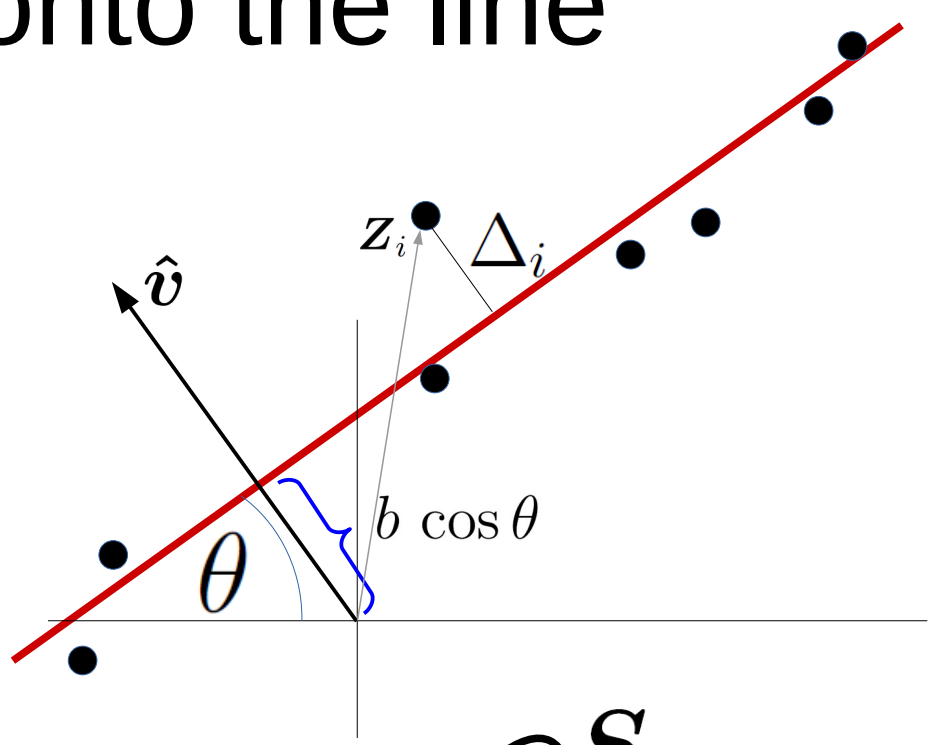
$$p(x_i, y_i | \boldsymbol{S}_i, x, y) = \frac{1}{2\,\pi\,\sqrt{\det(\boldsymbol{S}_i)}} \exp\left(-\frac{1}{2}\,[\boldsymbol{Z}_i - \boldsymbol{Z}]^\top\,\boldsymbol{S}_i{}^{-1}\,[\boldsymbol{Z}_i - \boldsymbol{Z}]\right)$$

where: $\boldsymbol{Z} = \begin{bmatrix} x \\ y \end{bmatrix}$; $\boldsymbol{Z}_i = \begin{bmatrix} x_i \\ y_i \end{bmatrix}$

We can project onto the line

line angle: $\theta = \arctan m$

$$\hat{\boldsymbol{v}} = \frac{1}{\sqrt{1+m^2}} \begin{bmatrix} -m \\ 1 \end{bmatrix} = \begin{bmatrix} -\sin\theta \\ \cos\theta \end{bmatrix}$$

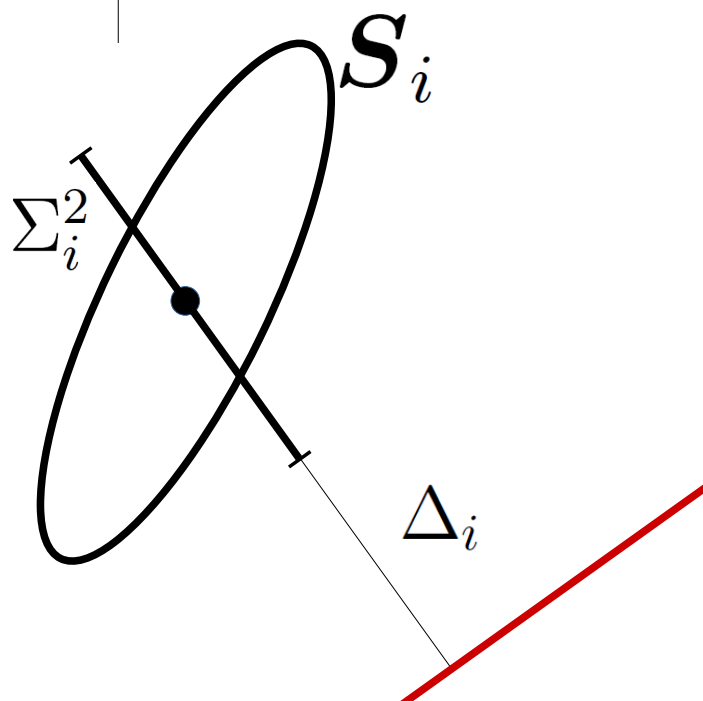


projected distance from the line:

$$\Delta_i = \hat{\boldsymbol{v}}^\top \boldsymbol{Z}_i - b\ \cos\theta$$

projected/orthogonal variance:

$$\Sigma_i^2 = \hat{\boldsymbol{v}}^\top \boldsymbol{S}_i\, \hat{\boldsymbol{v}}$$

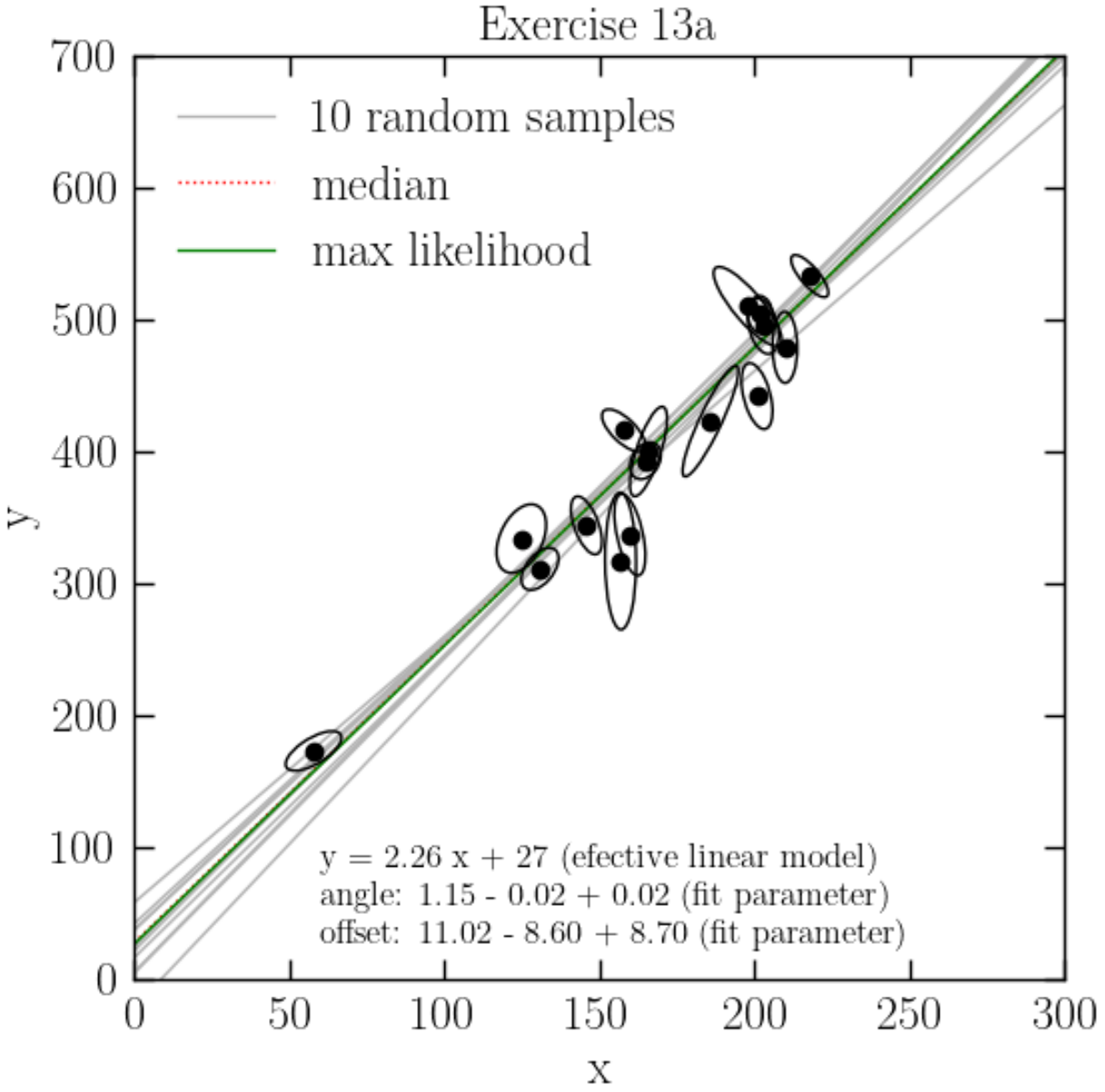


typical Gaussian likelihood:

$$\ln \mathscr{L} = K - \sum_{i=1}^{N} \frac{\Delta_i^2}{2\,\Sigma_i^2}$$

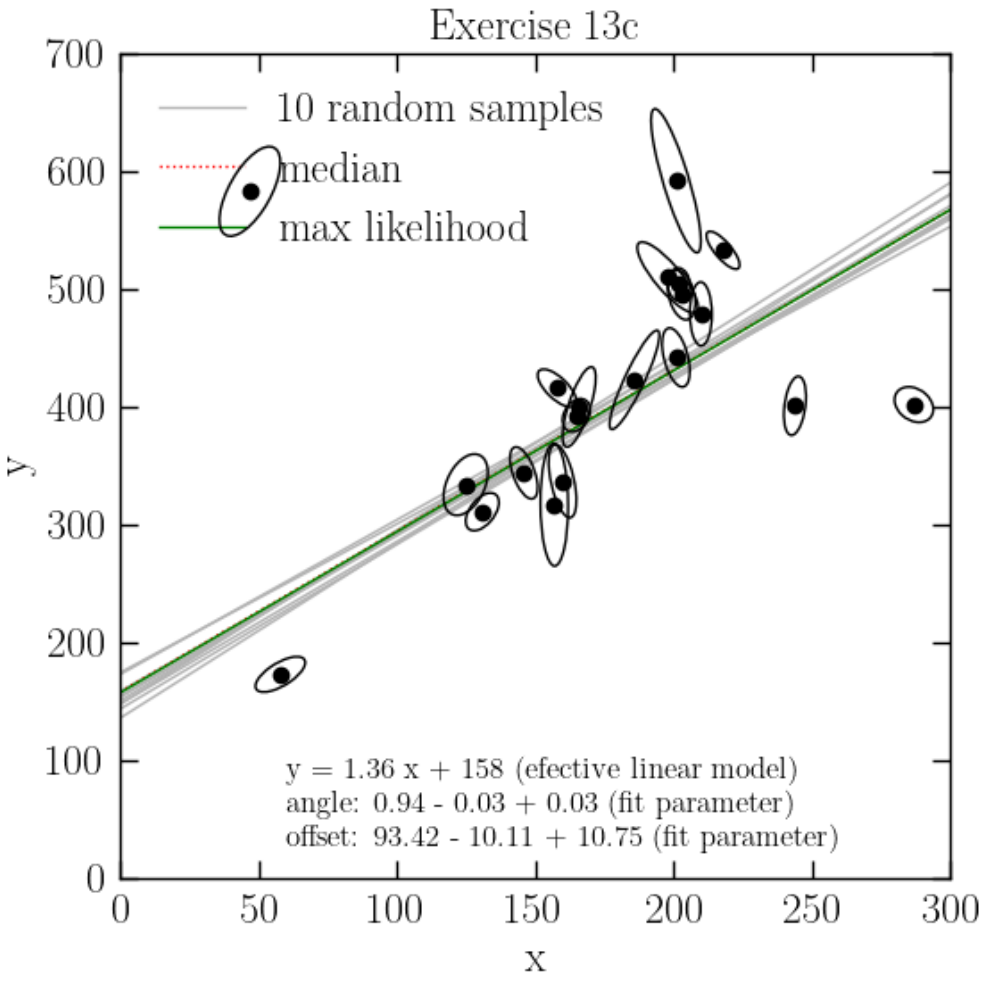
Exercise 13

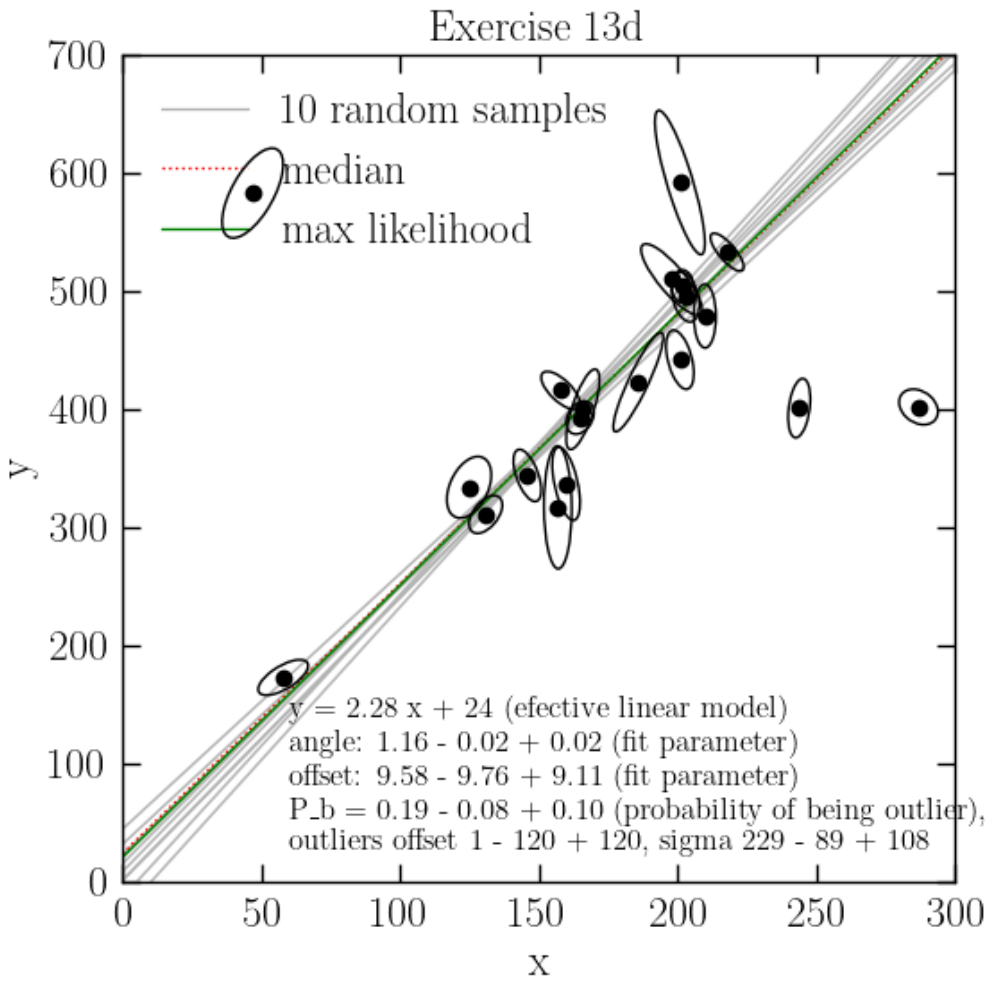
fit model to points with 2d errorbars)
(no outliers in the data)



Exercise 14

extend model to simultaneously fit population of outliers





standard linear model (outliers in the data)

mixture model with outliers accounted for (P_b = *19%*)

Intrinsic scatter of the linear model

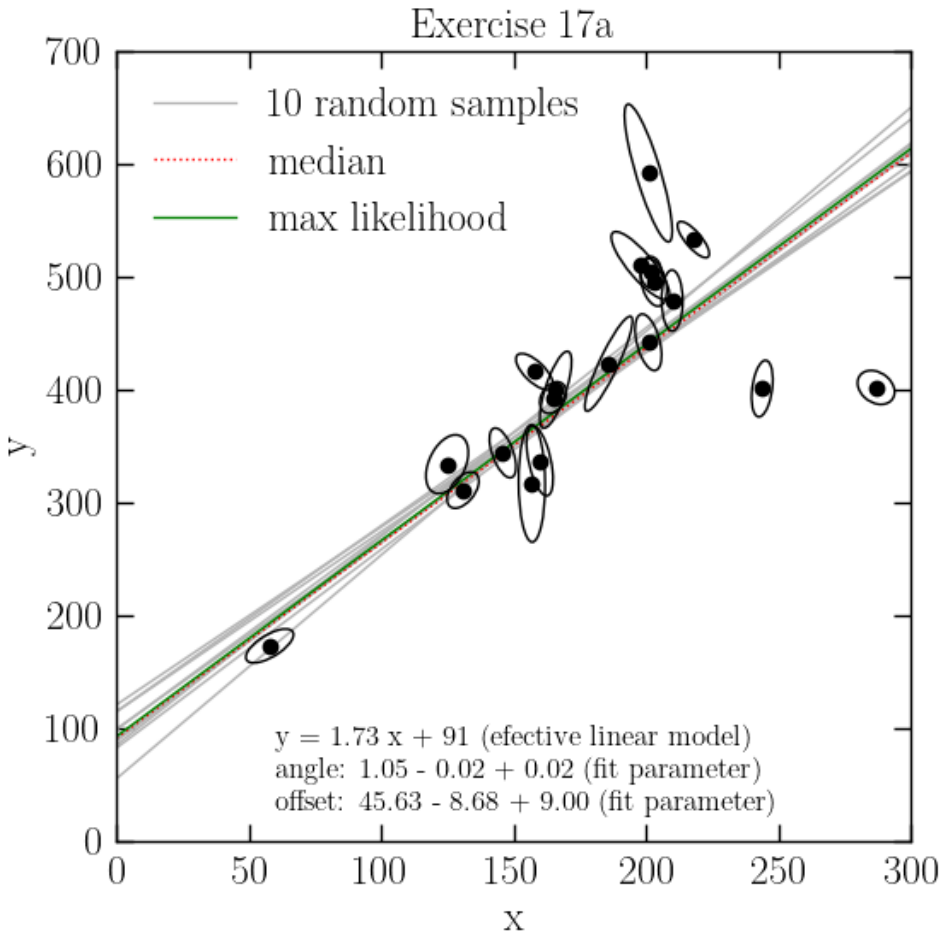
- what if we know, that the underlying model have some inherent scatter to it?
- we can model it by adding new parameters to describe this
- simplest case: we add variance (*V*)

$$\ln \mathscr{L} = K - \sum_{i=1}^{N} \frac{1}{2} \ln(\Sigma_i^2 + V) - \sum_{i=1}^{N} \frac{\Delta_i^2}{2\left[\Sigma_i^2 + V\right]}$$

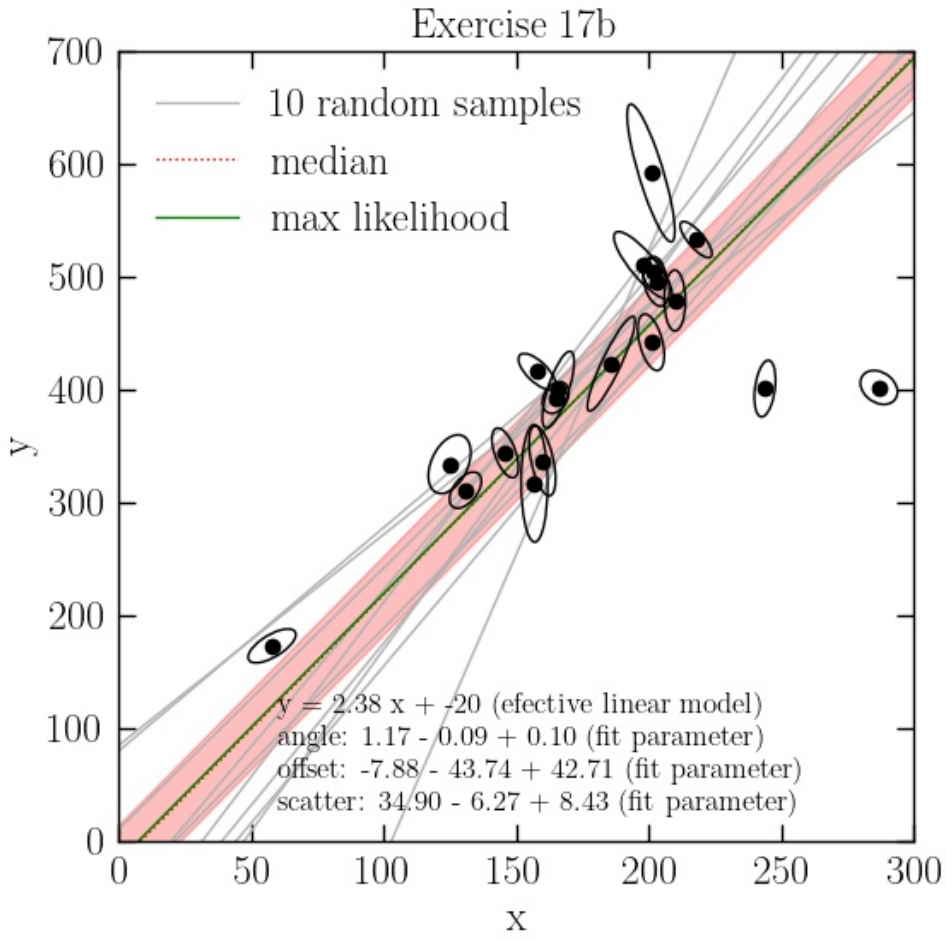
compared to previous: $\ln \mathscr{L} = K - \sum_{i=1}^{N} \frac{\Delta_i^2}{2\,\Sigma_i^2}$

Exercise 17 and 18

(add an intrinsic scatter to the linear model)

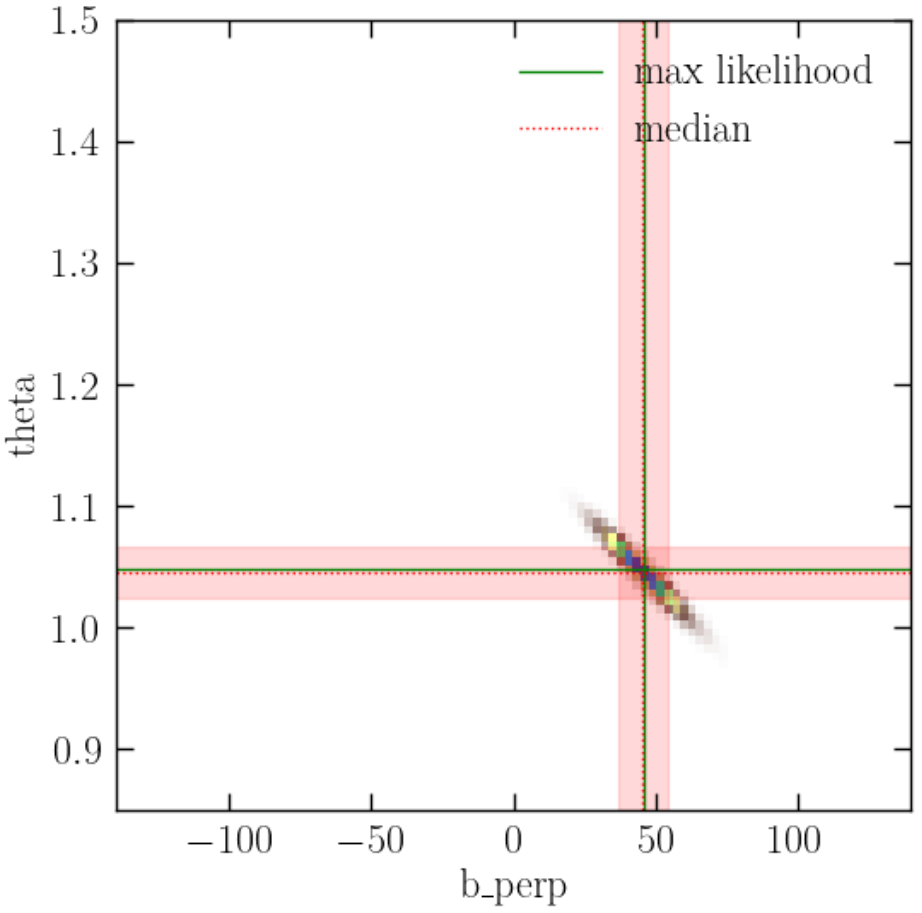


without scatter fitted

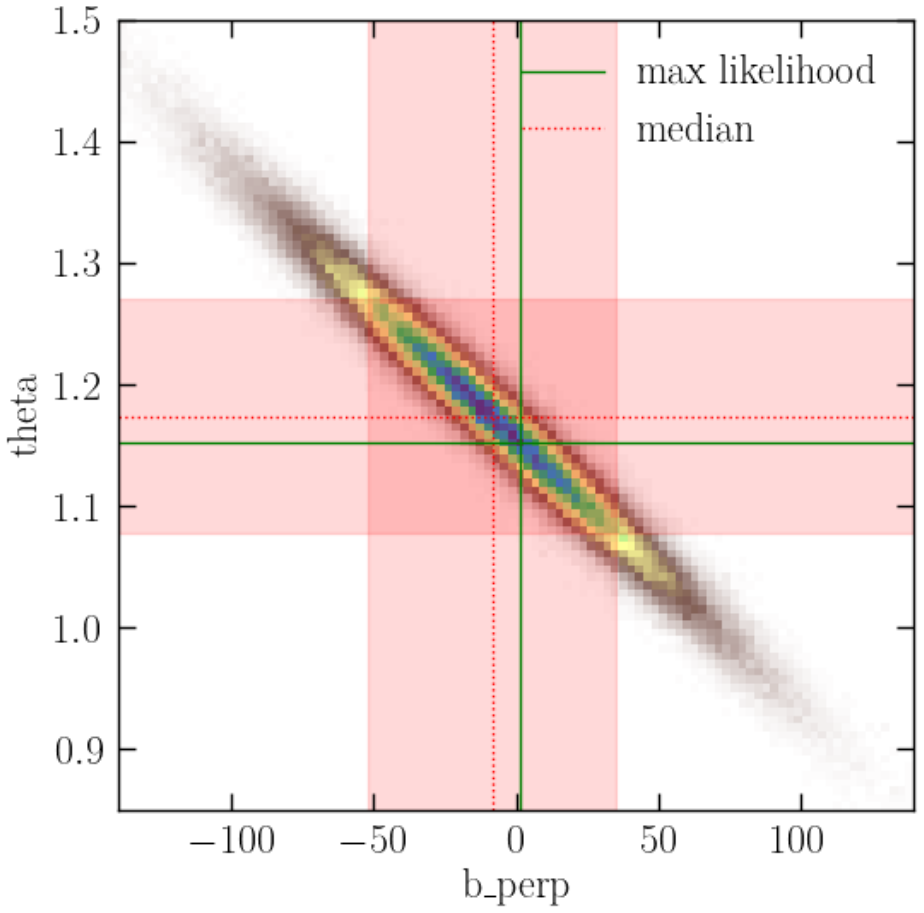


with intrinsic scatter as a fit parameter

Exercise 18

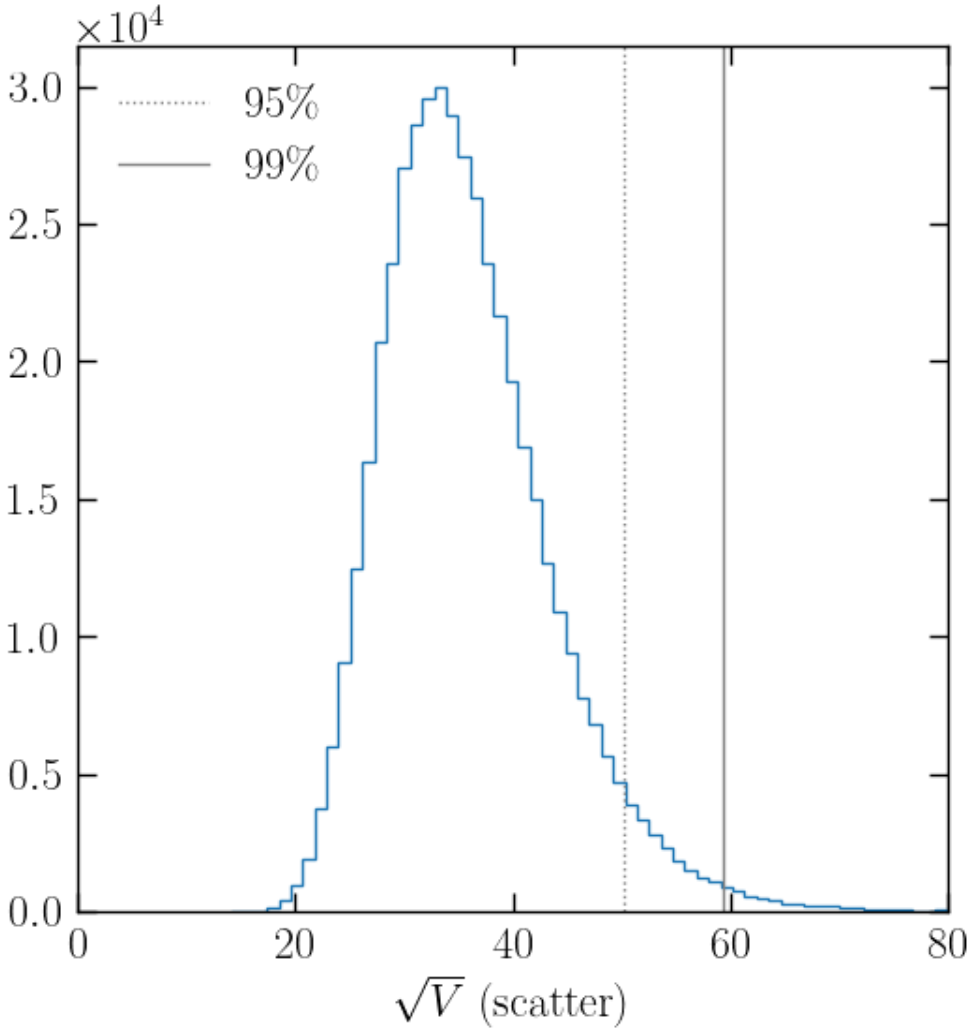


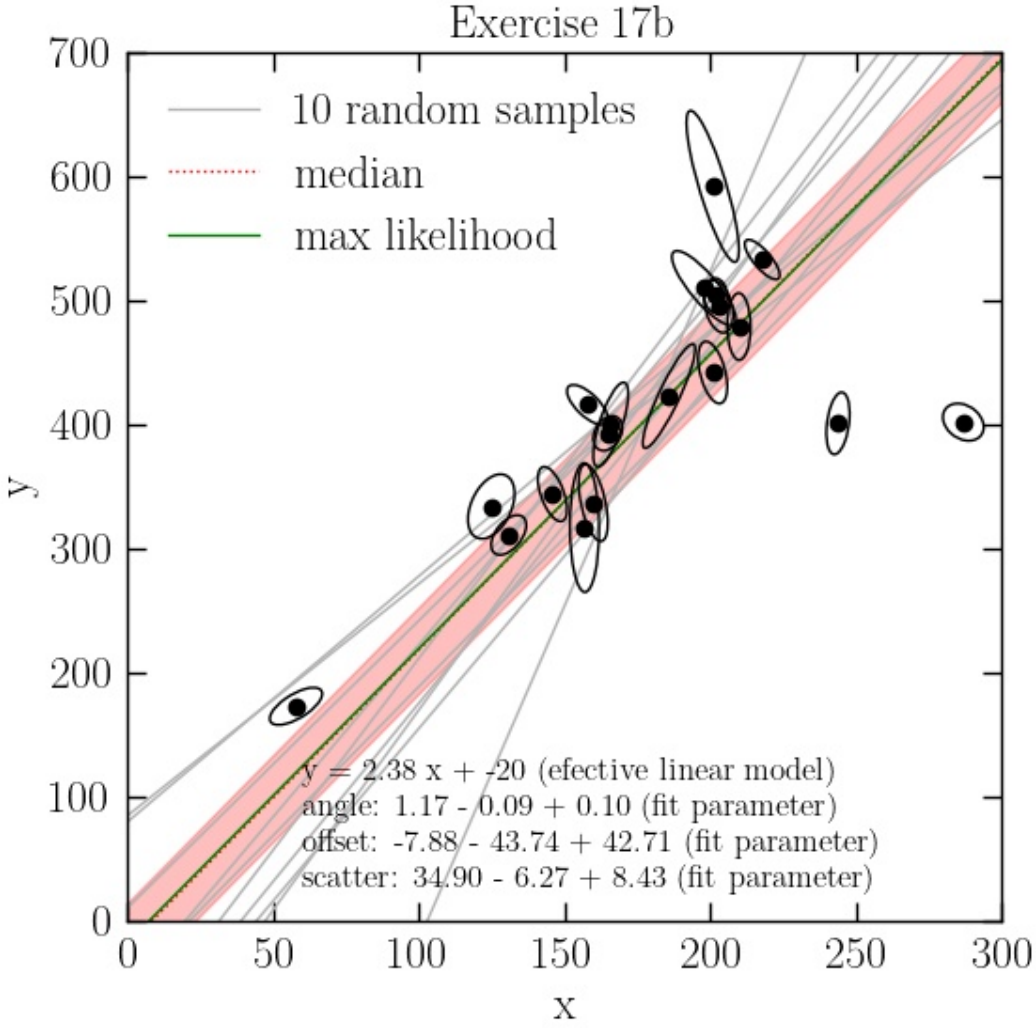
without scatter fitted



with scatter
(larger uncertainties of parameters)

Exercise 18
(generate maginalized posterior for the intrinsic scatter)





Fin

for details consult:
David W. Hogg, Jo Bovy, Dustin Lang (2010)
https://arxiv.org/abs/1008.4686