Today: - handling outliers
- Markov Chain Monte Carlo

Note: outlier material follows arxiv: 1008.4686 Section 3 (pg 11). More explanation & exercises there too.

Outliers

Experimental science always encounters glitches. How do we deal with them?

- by hand?
- by some heuristic? (sigma-clipping)
- with statistics!

↳ So far, we have written a generative model for our data: eg, $y_i = b + m x_i + e_i$

Now, we need to model how data can be bad!

We'll say: data can come from our good ("foreground") or bad ("background") distribution. For the background, we'll use a Gaussian (whose parameters we'll fit).

- two ways of looking at this setup. equivalent

① For N data points, add N new parameters, q_i, $q_i = \begin{cases} 0 & \text{if data point } i \text{ is bad} \\ 1 & \text{if data point } i \text{ is good} \end{cases}$

Our "good" distribution is

P_{fg} $\quad y_i \sim N(mx_i + b, \sigma_i^2)$

out "bad" dist. is

P_{bg} $\quad y_i \sim N(Y, V)$

So we can write the combo as:

$$P(y_i \mid m, b, q_i, Y, V) = P_{fg}(y_i \mid m, b)^{q_i} \times P_{bg}(y_i \mid Y, V)^{(1-q_i)}$$

Now, we're going to <u>marginalize out</u> the parameter q_i:

$$\int P(y_i \mid m, b, q_i, Y, V)\, p(q_i)\, dq_i$$

q_i only takes value 0, 1, so

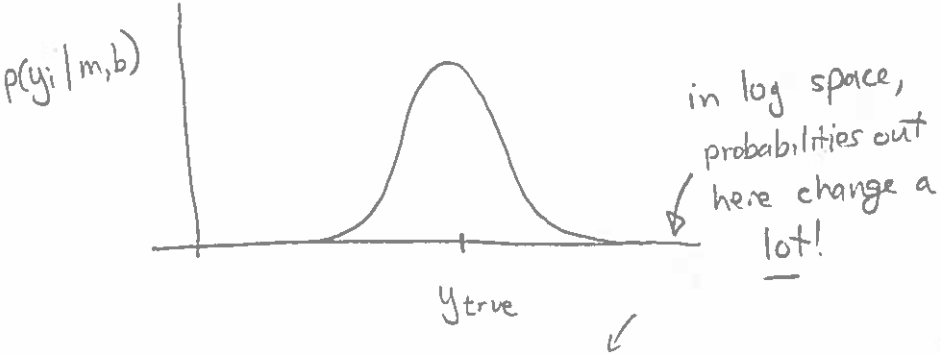
$$= \sum p(y_i \mid \cdots q_i = 0 \cdots)\, p(q_i = 0) + p(y_i \mid \cdots q_i = 1 \cdots)\, p(q_i = 1)$$

Call $p(q_i=0)$ "P_{bad}" - a priori probability that any given data point is bad.

Then $p(q_i=1) = 1 - P_{bad}$, and

$$p(y_i \mid m, b, Y, V, P_{bad}) = P_{bad} \times P_{bg}(y_i \mid Y, V) + (1 - P_{bad}) \times P_{fg}(y_i \mid m, b)$$

This is the "mixture model" or "fg-bg" model version. We got rid of the N params q_i.



Ratio between 10σ and 9σ is way bigger than between 2σ and 1σ

→ Fit is dominated by the outliers.