The evidence:

$$\mathcal{E} = \int L(d|\vec{\theta}(M)) \cdot p(\vec{\theta}(M)) \, d\vec{\theta}$$

⇝ completely integrates out all parameters : also called „marginal likelihood"
= the likelihood of obtaining the data at all

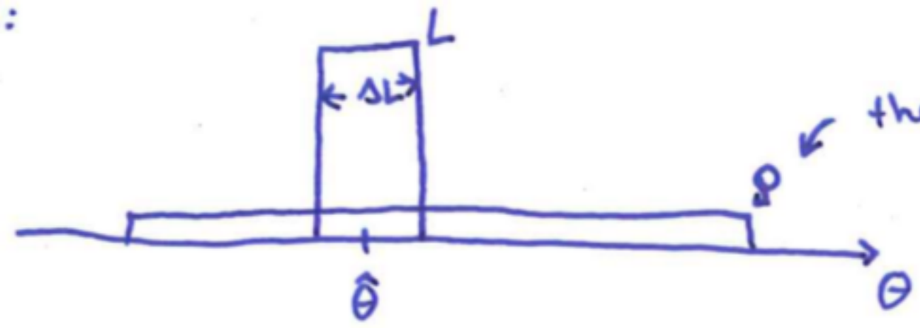
$\mathcal{E}$ is a numerical beast.
But it does the following / is needed for the following:

- Normalization of the Posterior
- penalize too complicated models (Def. "too": too complicated in the light of the available data)
  ⇝ „Occams Razor"
- Compare the likelihoods of different models
- test how many parameters the data want

Occams razor:

The evidence will sink with wide prior ranges ⇝ predictivity of a model is rewarded

Toy model:



the prior is normalized ⇒ $\int \rho \, d\theta = 1$

⇝ $\mathcal{E} = \int L(d|\theta) \, p(\theta) \, d\theta = L(\hat{\theta}) \, p(\hat{\theta}) \, \Delta L$

$p(\theta) \, \Delta p = 1$

$= L(\hat{\theta}) \, \frac{\Delta L}{\Delta p}$

will boost the evidence of a very good fit

wide prior ranges (i.e. an „unpredictive model") bring the evidence down

⇝ we'll apply the priors and the evidence for Model comparisons

Model comparisons

E.g.: • „standardmodel or „beyond standard model?"

- Model with 4 or 40 parameters?
- statistical fluke or new physics?

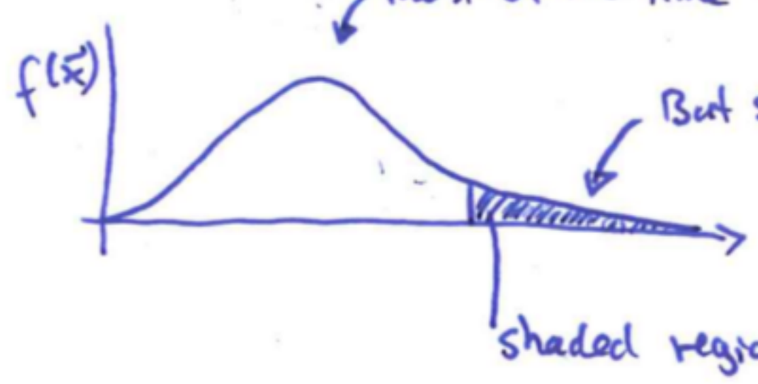
Modelcomparisons are rather easy to deal with in the Bayesian context, because the data are taken as fixed, and the credibility of the fitted model is investigated.

In contrast: Frequentist has $L(x|H)$, which is the likelihood that Hypothesis H has produced the data x.

But: $L(x|H) \neq L(H)$! From $L(x|H)$ we can't tell the probability of the Hypothesis.
↓
in the sense of „true" with a certain probability.

Let's investigate a Frequentists p-value:

the measured data set $\vec{x}$ stems from a probability distribution $f(\vec{x})$, which tells you how your measured data set would typically change if you measure another time



$p := L(T(\vec{x}) \geq T(x_{obs}))$ if T is some statistics on your data set

↓

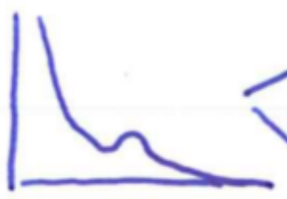
„the likelihood that the statistics can turn out as extreme, or more extreme, than in the data set which you analyse."

E.g. p-value = 0.05 : If the Hypothesis were true (which we don't know), then only in 5% of the times that you measure, it would produce you a data set as extreme, or more extreme, than the one you just calculated your p-value from.

{⇝ already while calculating the p-value, you assume the hypothesis to be true}

⇒ $L(x|H) \neq L(H)$ the likelihood that Hypothesis H has produced the data x is not the likelihood of H to be true.

(Not) Assuming the nullhypothesis to be true influences the way we speak:



Particle physicist: „It will show up as a peak: the Higgs will give itself away with a peak somewhere here, and its detection is significant with xyz-σ"

Astrophysicist/Cosmologist: „Hey, there is a peak! What has produced it?"

Nullhypothesis is assumed to be true: You can only detect something that exists

Why can't you use a p-value for model comparison?

- Because while calculating it, you already assume the Hypothesis to be true.
- Because it answers the Question „If the Hypothesis is true, then x% of the times..."

⇝ Example: one data set $\vec{x}$, two competing Hypothesis H_1 and H_2, only one is true.

Calculate $p_1 = L(T(\vec{x}|H_1) \geq T(\vec{x}_{obs}|H_1)$ ⇝ assumes H_1 is true

Calculate $p_2 = L(T(\vec{x}|H_2) \geq T(\vec{x}_{obs}|H_2)$ ⇝ assumes H_2 is true

} one of these calculations must base on wrong assumptions

↓

Calculating p-values doesn't allow to compare Models to each other

Bayesian Model comparison

we had the evidence

$$\mathcal{E} = p(d|M) = \int \underbrace{L(d|\vec{\Theta}(M))}_{\text{Likelihood (data go in here)}} \underbrace{p(\vec{\Theta}(M))}_{\text{prior (theory goes in here)}} d\vec{\Theta}$$

⇒ marginalizes out all parameters ⇝ evidence is the likelihood that the model M can produce the data at all.

Question of model comparison; e.g. „How likely is my model, in light of the available data?"

↓

Likelihoods are normalized, so this Question would only have an answer if you know „all possible Models".

⇝ But you can still compare multiple models with each other:

Evidence: $\mathcal{E} = L(d|M)$ ⇝ invert ⇝ $L(M|d)$ is the likelihood of your model, given the data

↓

Bayes theorem:

$$L(M|d) = L(d|M) \cdot \frac{p(M)}{p(d)}$$

$p(M)$: prior on model ??

$p(d)$: prior on data ??

→ take a ratio :)

2nd model: $L(M_*|d) = L(d|M_*) \frac{p(M_*)}{p(d)}$

⇝
$$\underbrace{\frac{L(M|d)}{L(M_*|d)}}_{} = \frac{p(M)\ L(d|M)}{p(M_*)\ L(d|M_*)} = \frac{p(M)}{p(M_*)} \frac{\mathcal{E}}{\mathcal{E}_*}$$

⇝ prior on data drops out, since in both cases, the data are the same

⇝ $\frac{p(M)}{p(M_*)}$ is usually set to 1.

Evidence ratio: $\frac{L(M|d)}{L(M_*|d)} = \frac{p(M)}{p(M_*)} \quad \frac{\mathcal{E}}{\mathcal{E}_*}$

The evidence ratio quantifies whether the data prefer M or M_*.

$\frac{\mathcal{E}}{\mathcal{E}_*}$ is what you care about.

The larger the evidence, the more likely your model.

„In the light of data d", is M or M_* more likely?"

Your beforhand belief of whether M or M_* are more likely.

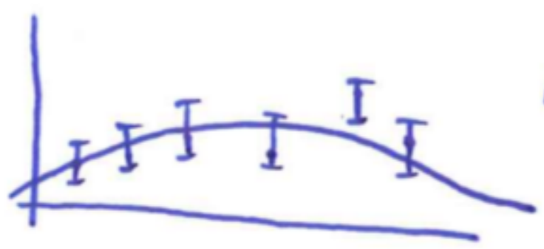
Example:

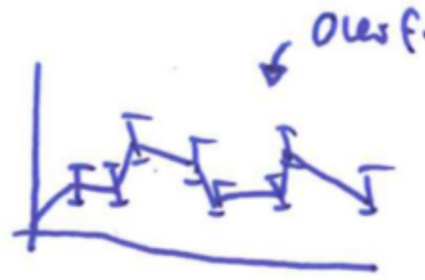
„With how many parameters should I fit my data?"

Problem: Increasing the number of parameters will always lead to an improvement I of the fit, with $I \geq 0$. ⇝ when do you stop?

(usually done with χ^2)

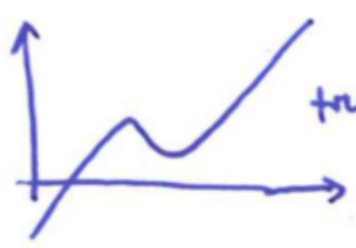
Classical problem: Overfitting / „Malen nach Zahlen".





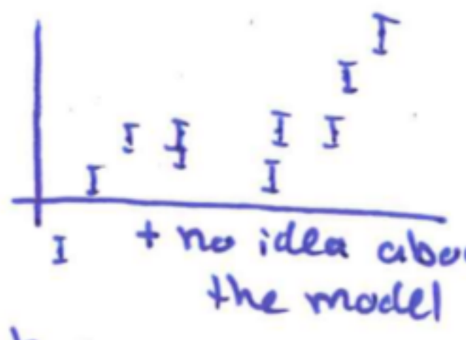
Completely misses the point of statistics: If you have estimated your error bars correctly then you do expect them to scatter around the best fit line.

E.g. true model: $ax + bx^2 + cx^3$, with 3 parameters a, b, c



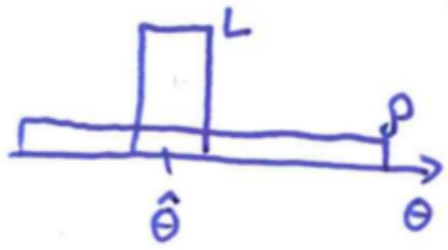
→

your data →



Let's find out with the evidence how many parameters the model has:

Reminder:



$\Rightarrow \mathcal{E} = L(\hat{\theta}) \cdot \frac{\Delta L}{\Delta \rho}$

goodness of fit — Occams Razor

Start with 1 parameter:

fit ax:

⇒ bad fit, bad χ^2 ⇒ low $L(\hat{\theta})$ ⇒ low $\mathcal{E}$

fit with 2 parameters:

$ax + bx^2$:

⇒ better fit, better χ^2 ⇒ larger $L(\hat{\theta})$ ⇒ larger $\mathcal{E}$

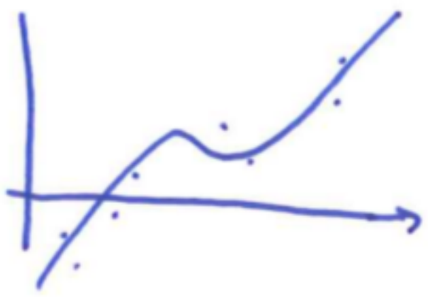
fit with 3 parameters:

$ax^4 + bx^2 + cx^3$ (the actual number of parameters of the true model)

$\Rightarrow \frac{\chi^2}{\text{dof}} \approx 1 \Rightarrow$ large $L(\hat{\theta}) \rightsquigarrow$ large $\mathcal{E}$

(over) fit with 4 parameters

$ax + bx^2 + cx^3 + dx^4$



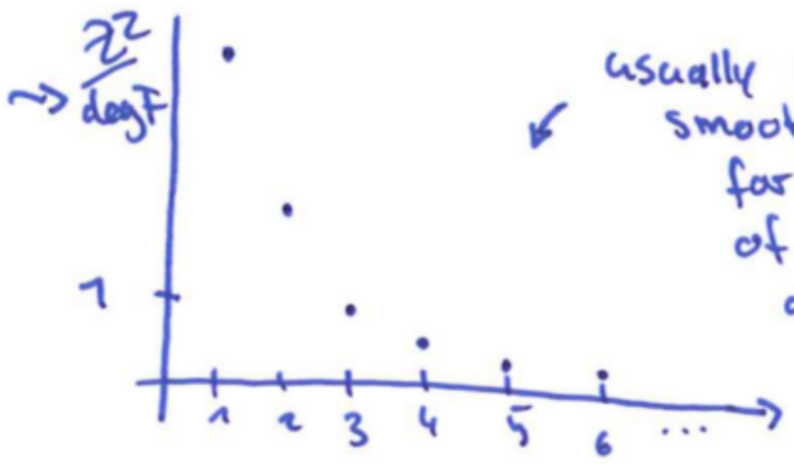
$\frac{\chi^2}{\text{dof}}$ keeps sinking. But the introduction of d does not make the fit so much better that the Occams razor factor for the new dimension is outweighted:

$\frac{\Delta L \text{ (d-direction)}}{\Delta p \text{ (d-parameter)}} \leftarrow$ large brings Evidence down

⋮

and so on, with ever more params

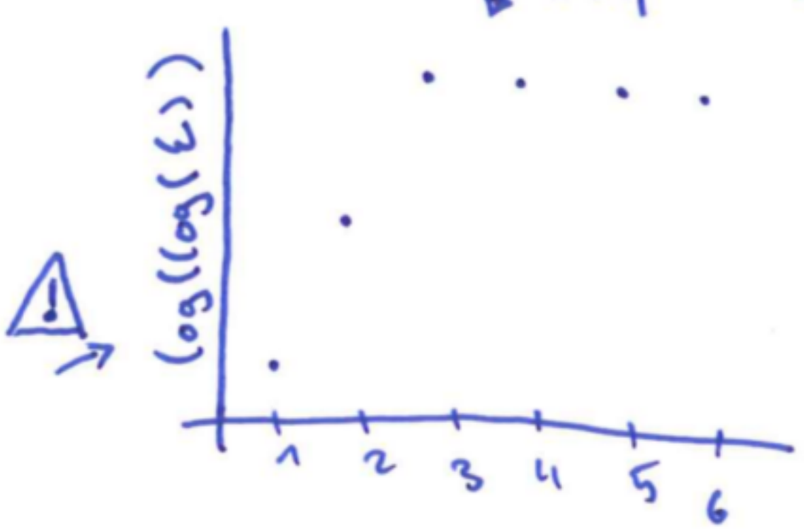
Results:

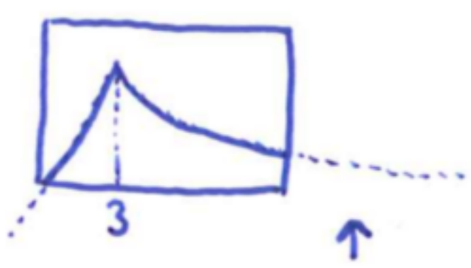


usually rather smooth, especially for huge number of data points and 20,40,60 parameters

→ it can be hard to judge how many parameter are needed

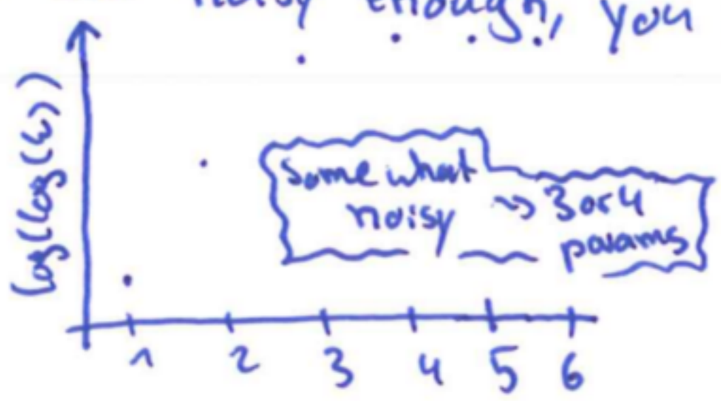
sharp breakoff ⇝ zoom in:

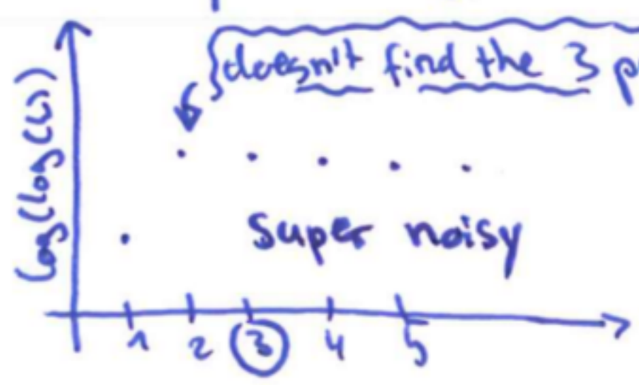




↑ dropping with Occams razor is invisible on the log(log(L)) scale

- Of course, if the data are noisy enough, you can miss parameters:





2nd Example: Standardmodel or Beyond standard model?

$$\frac{L(SM|d)}{L(BSM|d)} = \underbrace{\frac{p(SM)}{p(BSM)}}_{\text{set to unity}} \cdot \frac{\mathcal{E}(SM)}{\mathcal{E}(BSM)}$$

without specifying here what exactly your BSM does. (If you want to code it, of course you have to specify

$\rightsquigarrow \ln\left(\frac{\mathcal{E}(SM)}{\mathcal{E}(BSM)}\right) > 0 \Rightarrow$ Standardmodel is favoured

$< 0 \Rightarrow$ Beyond standard model is favoured

$\lesseqgtr$

But with which significance? $\rightsquigarrow$ Run tests and calibrate: {Jeffreys scale}

Jeffreys Scale:

$\left\|\ln\left(\frac{\mathcal{E}(SM)}{\mathcal{E}(BSM)}\right)\right\|$	odds	„conventionally called"	probability of the favoured model
≤ 1.0	3:1	„better data is needed"	≤ 0.75
≤ 2.5	12:1	weak evidence	0.923
≤ 5.0	$\leq 150:1$	moderate evidence	0.993
$\geqslant 5.0$	$> 150:1$	strong evidence	>0.993

$\rightsquigarrow$ measuring more precisely (= „improving parameter constraints") will ~~losing~~ push the evidence for good fitting model up

$\rightsquigarrow$ but only the evidence ratio tells whether this also happens with another model

if nature is indeed a BSM, then with the improvement of data we will observe how $\ln\left(\frac{\mathcal{E}(SM)}{\mathcal{E}(BSM)}\right)$ tipps from ~~<0~~ >0 to <0.

— Approximations of Likelihoods & Gaussianity —

We'll find Gaussians everywhere in statistics. Why?

Because:
- they can be handled analytically → fast!
- because many data are naturally Gauss-distributed
- Physicists like to count and Poisson-statistics has a Gaussian limit
- and if all else fails: **The Central Limit Theorem**

CTL: For large n, the sum of n random numbers that were drawn from a probability distribution with a finite variance σ^2, and higher moments, tends to be gaussian, with mean of the sums expetation value and variance $n\sigma^2$, (and no higher moments, because CTL is exactly that the higher moments decrease and asymptote to zero.) so only the first two moments survive ⇒ Gaussian.

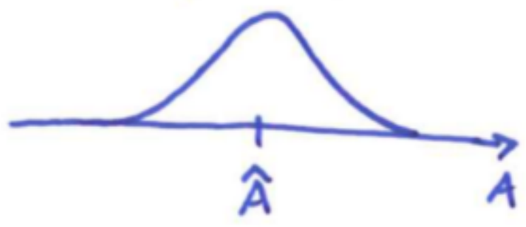
This works in the data space.

Does it mean that also a likelihood is usually Gaussian? ⇝ Gnuplot script

Other examples of non-Gaussian likelihoods:

- Parameter degeneracies: Imagine your data measure a function $f(A, x_1, x_2, x_3 \ldots)$, where the x_i are independent variables, and A a model parameter. Let's use $f = Ax_1$

⇒



Gaussian likelihood in A, if data are Gaussian

⇝ But maybe you know from theory that $A = a_1/a_2$, and you are interested in a_1 and a_2 ⇒ degeneracy:

⇝ non-Gaussian likelihood although data are Gaussian

⇝ So one can't always approximate a likelihood with a Gaussian
- ⇝ no analytical solutions
- ⇝ need witty numerical methods to keep computation times down (MCMC, Nested Sampling)
- ⇝ or need a well-working non-Gaussian approximation