

A Hidden Gem: Unlocking the Power of Bayesian Inference in Particle Physics

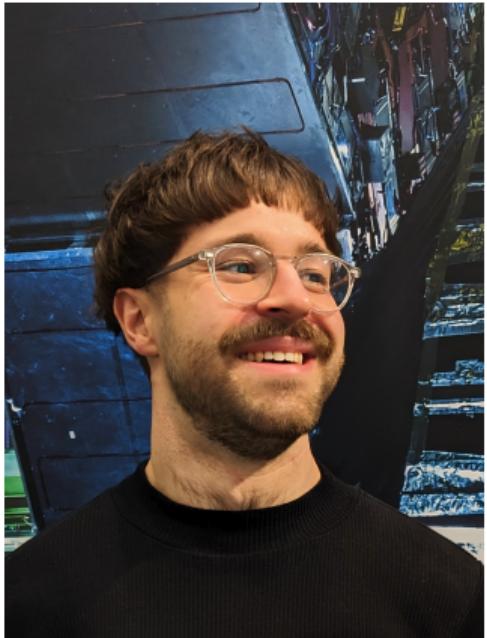
Lorenz Gärtner¹, Toni Sculac, Judith Katzy

¹LMU Munich

February 26, 2025

About me

- BSc @ University of Manchester
Physics with theoretical physics
Little stats
- MSc @ LMU Munich
Thesis on QFT in curved spacetime
Almost no stats
- Currently PhD @ LMU Munich
A lot of stats



What is a probability?

Kolmogorov probability axioms

1. $p(\Omega) = 1$, where Ω is the sample space.
2. $p(x) \geq 0$ for any event $x \subseteq \Omega$.
3. For any sequence of disjoint events x_1, x_2, \dots ,

$$p\left(\bigcup_{i=1}^{\infty} x_i\right) = \sum_{i=1}^{\infty} p(x_i)$$

Conditional probability

Is **defined** as the probability of an event x if we know that an event y is true $p(x|y)$.

$$p(x \cap y) = p(x|y)p(y) \quad \rightarrow \quad p(x|y) = \frac{p(x \cap y)}{p(y)}$$

Note

$$p(x \cap y) = p(y \cap x) \quad \text{but} \quad p(x|y) \neq p(x|y)$$

Probability interpretations

The probability axioms and rules allow you to calculate new probabilities from old ones.

To assign probabilities we need probability interpretations.

The interpretations **share the same mathematical framework**, but the meaning of $p(x)$ is different.

Frequentist interpretation

Assign a probability as relative frequency

$$p(x) = \lim_{N \rightarrow \infty} \frac{N_x}{N}$$

- Can only be assigned to repeatable experiments
- Not everything is repeatable...
- No probabilities of single events

Bayesian interpretation

Assign a probability $p(x)$ as *degree of belief*.

- Probability depends on the experimenters' knowledge.
- Inference results are subjective.
- Everything is a random variable.

Bayes' theorem

The **posterior** is

$$p(\text{theory}|\text{data}) = \frac{p(\text{data}|\text{theory})p(\text{theory})}{p(\text{data})}$$

- **Likelihood** $p(\text{data}|\text{theory})$
- **Prior** $p(\text{theory})$
- **Evidence** $p(\text{data}) = \int p(\text{data}|\text{theory})p(\text{theory})$

Can you derive it?

The common ground

Frequentist inference is based on

$$p(x|\theta)$$

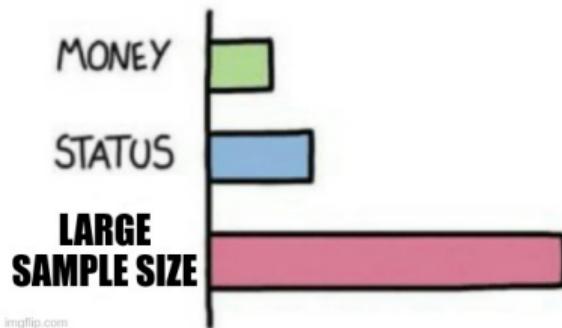
Bayesian inference is based on

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)}$$

→ You want the best possible $p(x|\theta)$.

The best possible model...

WHAT GIVES PEOPLE FEELINGS OF POWER



Physics does not care...

... about our interpretation

For many inference problems, the frequentist and Bayesian approaches give similar numerical values, even though they answer different questions.

BUT if results are different, you should understand why.

Parameter inference

Point estimates

Identify the most probable parameter point.

Interval estimation

Identify extended regions in parameter space based on compatibility with the data.

Frequentist point estimates: estimators



Estimator is a *statistic* $\hat{\theta}(x)$, i.e. a well chosen function of the data with the desited properties of

- *consistency*

$$\lim_{n_x \rightarrow \infty} E_x[\hat{\theta}] = \theta_{\text{true}}$$

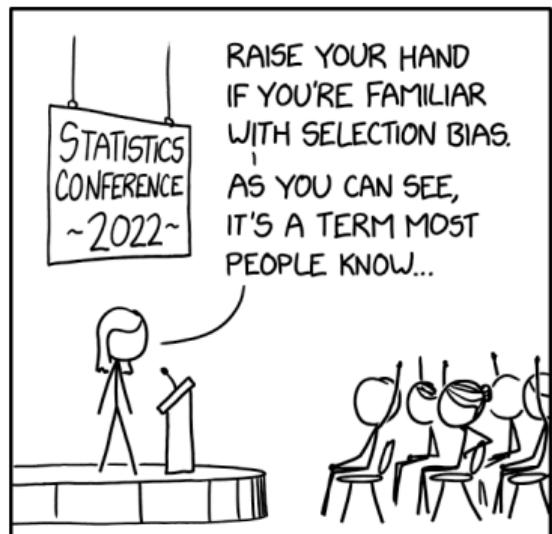
- *unbiasedness*

$$b = E_x[\hat{\theta}] - \theta_{\text{true}}$$

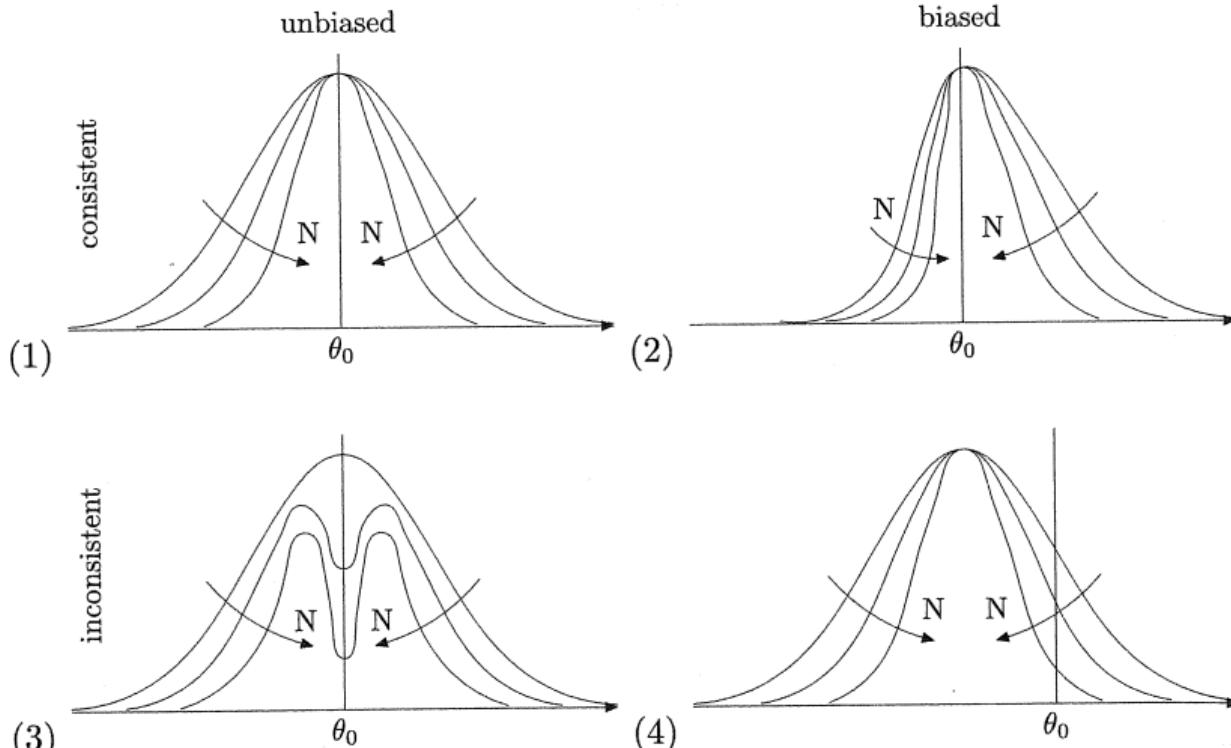
- *efficiency*

$$V(\hat{\theta}) = I(\theta)^{-1} = E_x \left[\left(\frac{\partial \ln p(x|\theta)}{\partial \theta} \right)^2 \right]^{-1}$$

- ...



A good data set is the key to success ...



Method of maximum likelihood

We find maximum likelihood estimators $\hat{\theta}$ by solving

$$\frac{\partial p(x|\theta)}{\partial \theta} \Big|_{\theta=\hat{\theta}} = 0$$

With the property

$$\lim_{N \rightarrow \infty} p\left(\sqrt{N}(\hat{\theta} - \theta_{true})\right) = \mathcal{N}\left(0, I^{-1}(\theta)\right)$$

This implies consistency, asymptotic unbiasedness and efficiency.

Bayesian point estimates

Mode

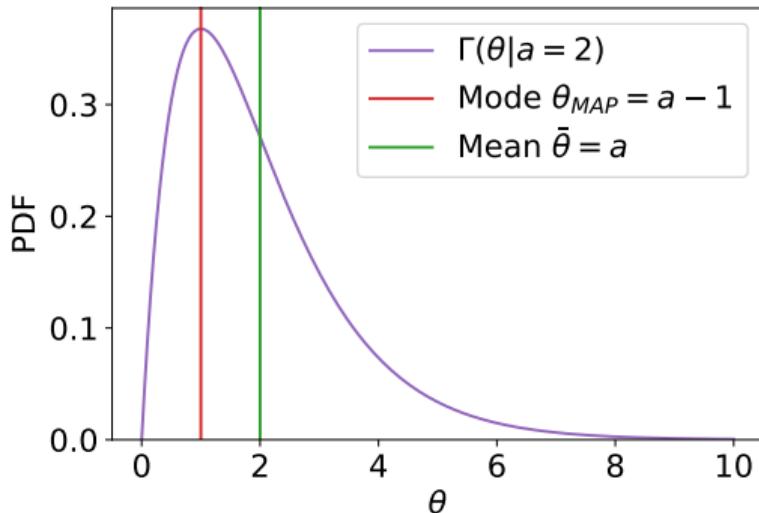
Value of θ with maximum a-posteriori probability

$$\theta^* = \operatorname{argmax}_{\theta} p(\theta|x)$$

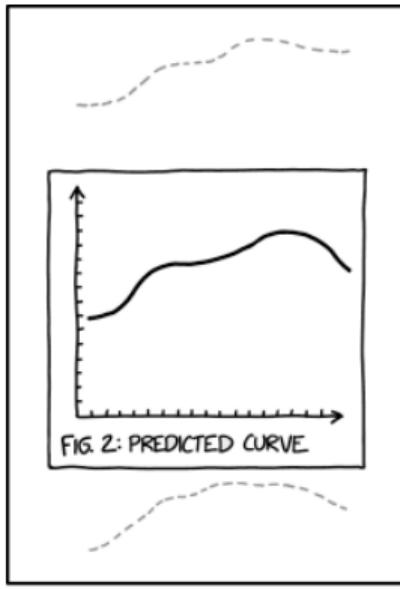
Mean

Expected value of θ under the posterior

$$\bar{\theta} = E_{p(\theta|x)}[\theta]$$



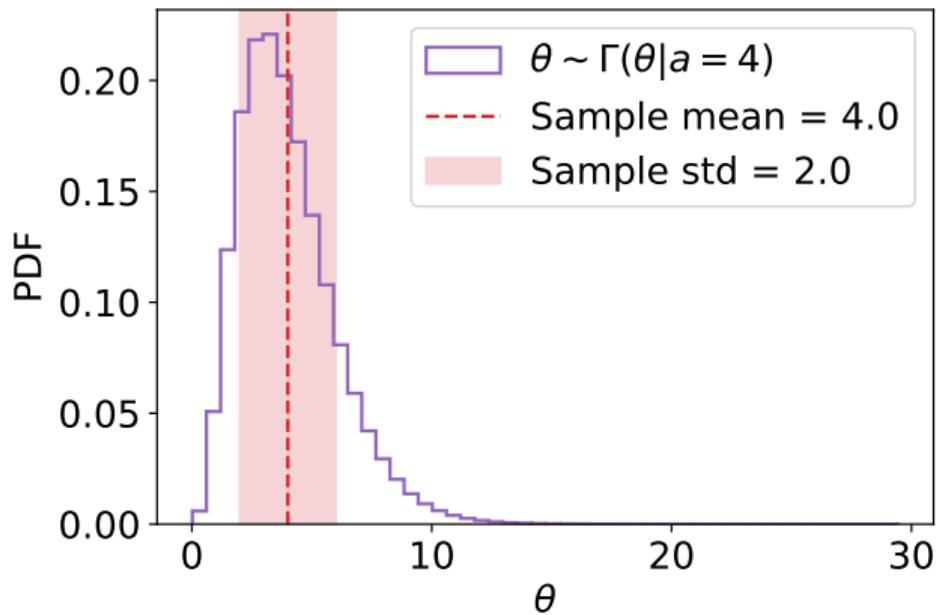
Intervals and limits



SCIENCE TIP: IF YOUR MODEL IS
BAD ENOUGH, THE CONFIDENCE
INTERVALS WILL FALL OUTSIDE
THE PRINTABLE AREA.

Intervals and limits

If an estimator PDF is not Normal, $\hat{\theta} \pm \sigma_{\hat{\theta}}$ is meaningless.



Frequentist intervals

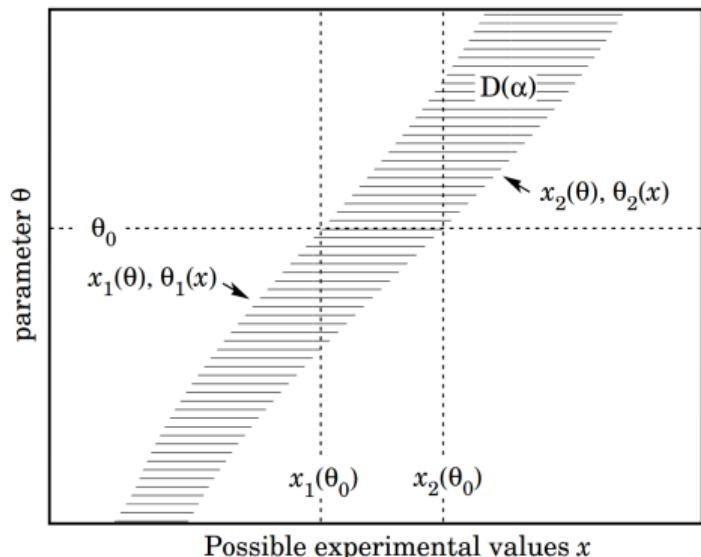
Neyman confidence belt

$$\int_{x_1}^{x_2} dx \, p(x|\theta) = 1 - \alpha$$

Not unique \rightarrow central interval

$$\int_{-\infty}^{x_1} dx \, p(x|\theta) = \int_{x_2}^{\infty} dx \, p(x|\theta) = \alpha/2$$

or upper/lower interval



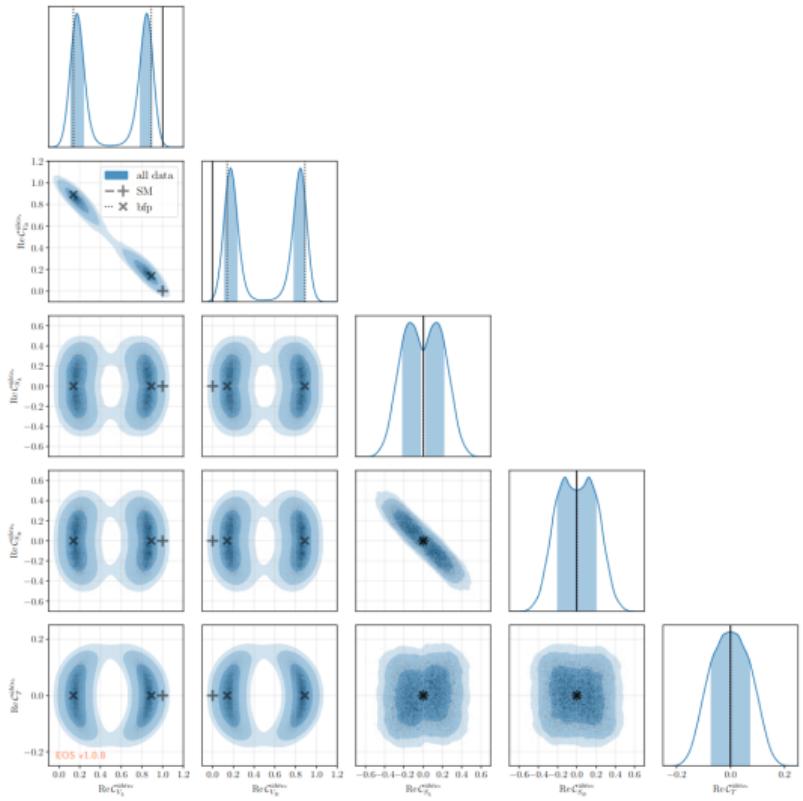
Bayesian intervals

Credible intervals $[\theta_1, \theta_2]$ cover $1 - \alpha$ of the posterior

$$\int_{\theta_1}^{\theta_2} d\theta p(\theta|x) = 1 - \alpha$$

- For upper/lower limits: set θ_1 or θ_2 to boundary
- *Smallest possible interval*
 - Highest (posterior) density intervals (HDI)

$b \rightarrow ul^- \bar{\nu}$ in the Weak Effective Theory



- Corner plots are great for visualization.
- Marginal posterior for Wilson coefficients
- Identify modes, credible intervals
- Best fit point \neq mode, why?

arXiv:2302.05268v2 [hep-ph]

Nuisance parameters and priors

Frequentist nuisance parameters

Models are not perfect → **systematic bias**

Solution: Include additional **nuisance** parameters ν ,

$$p(x|\theta, \nu).$$

Usually, we want to constrain nuisance parameters, but in the frequentist language **everything is data**.

Frequentist "priors"

Constrain nuisance parameters using *auxiliary data* a ,

$$p(x|\theta, \nu)p(a|\nu)$$

$p(a|\nu)$ represents our *degree of belief* in ν .

Often *auxiliary data* is *created* to match our desired constraint term.

Bayesian nuisance parameters

Priors can be defined using auxiliary data

$$p(\nu|a) \propto p(a|\nu)p_0(\nu)$$

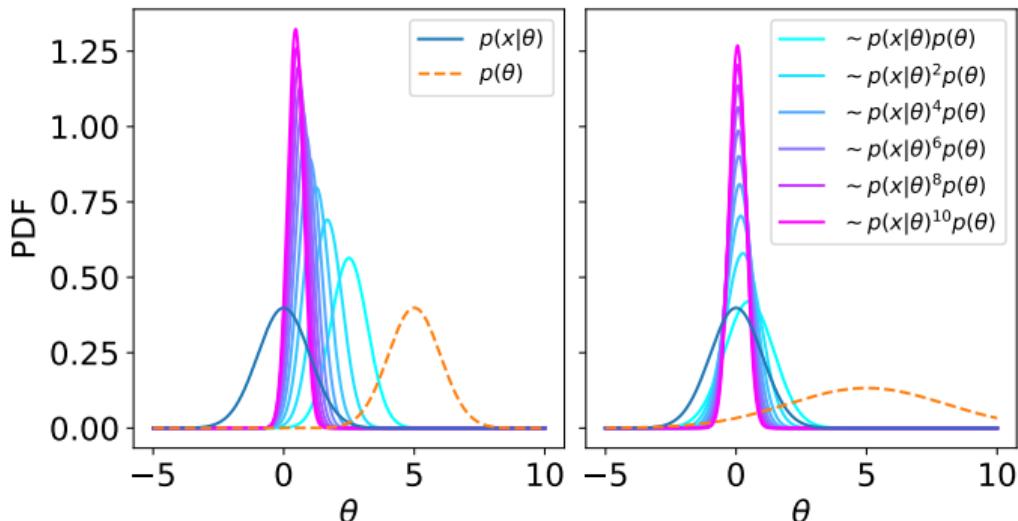
Only in the Bayesian case, other prior choices are also allowed, e.g.

$$p(\nu) = \mathcal{N}(\nu|\nu_0, \sigma_\nu^2)$$

Bayesian updating

Can generally use measurements to *update* our posterior

$$p(\theta|x_1, x_2) = \frac{p(x_2|\theta)}{p(x_2)} p(\theta|x_1) = \frac{p(x_2|\theta)}{p(x_2)} \frac{p(x_1|\theta)}{p(x_1)} p(\theta).$$

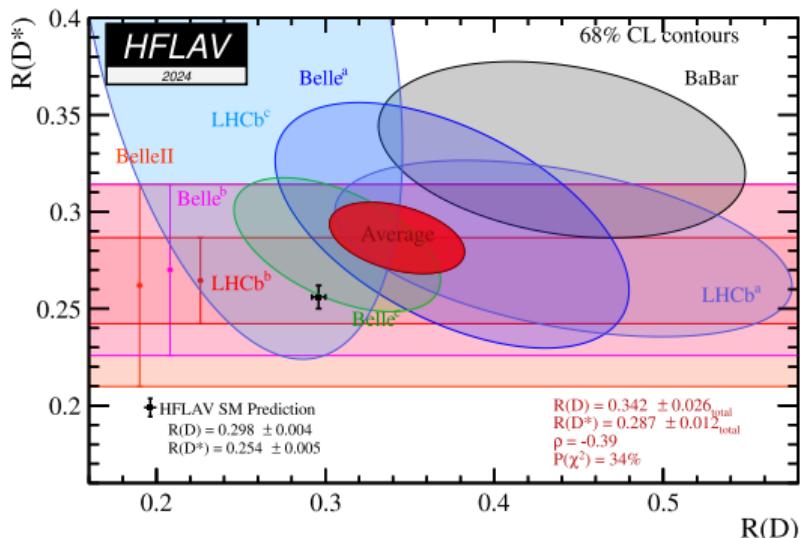


Bayesian updating in real life

PDFs for parameters are manifestly Bayesian.
 Combinations of $\mu_i \pm \sigma_i$ assume an underlying PDF for μ_i .

$$\mathcal{R}(D) = \frac{\mathcal{B}(\bar{B} \rightarrow D\tau^-\bar{\nu}_\tau)}{\mathcal{B}(\bar{B} \rightarrow DI^-\bar{\nu}_I)}$$

$$\mathcal{R}(D^*) = \frac{\mathcal{B}(\bar{B} \rightarrow D^*\tau^-\bar{\nu}_\tau)}{\mathcal{B}(\bar{B} \rightarrow D^*I^-\bar{\nu}_I)}$$

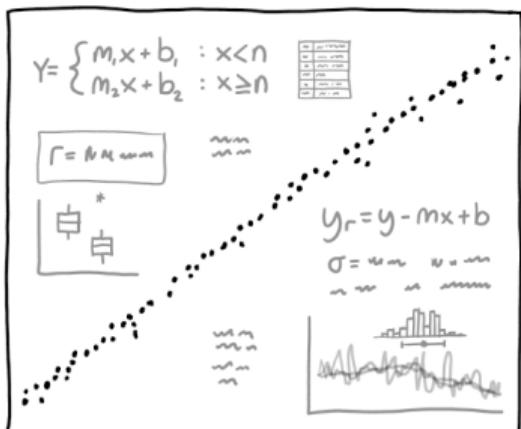


Actual combinations a bit more involved, see [HFLAV 2024](#).

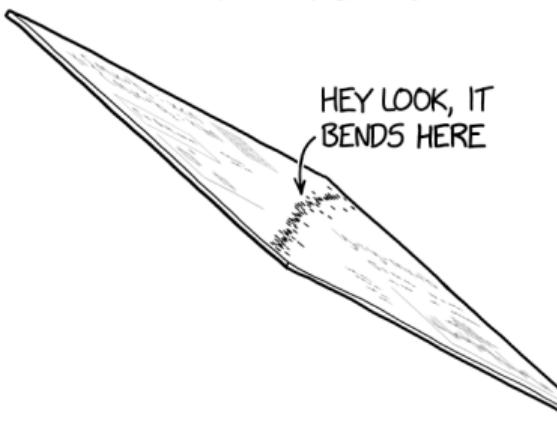
A simple linear model

HOW TO DETECT A CHANGE IN THE SLOPE OF YOUR DATA

NOVICE METHOD:

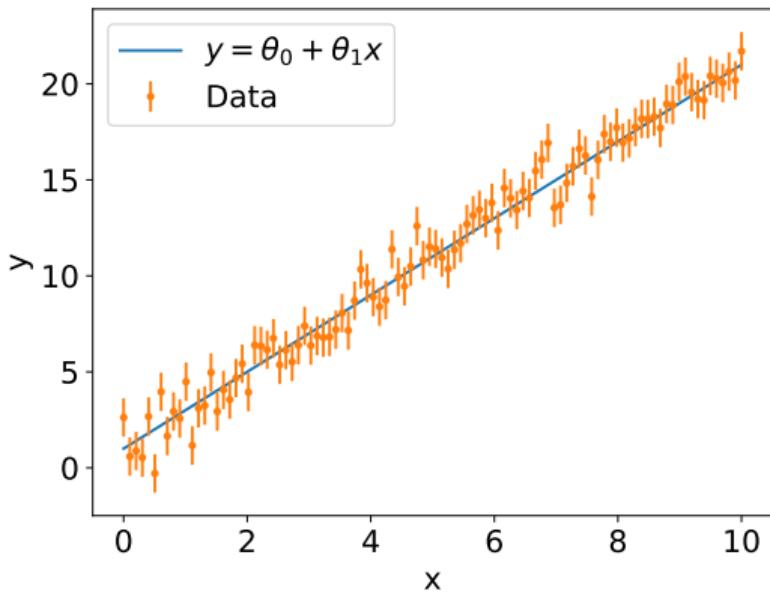


EXPERT METHOD:



A simple linear model

- Our independent data : $\mathbf{X} = (x_i, y_i, \sigma_i)$



A simple linear model

- Our independent data: $\mathbf{X} = (x_i, y_i, \sigma_i)$
- Our model:

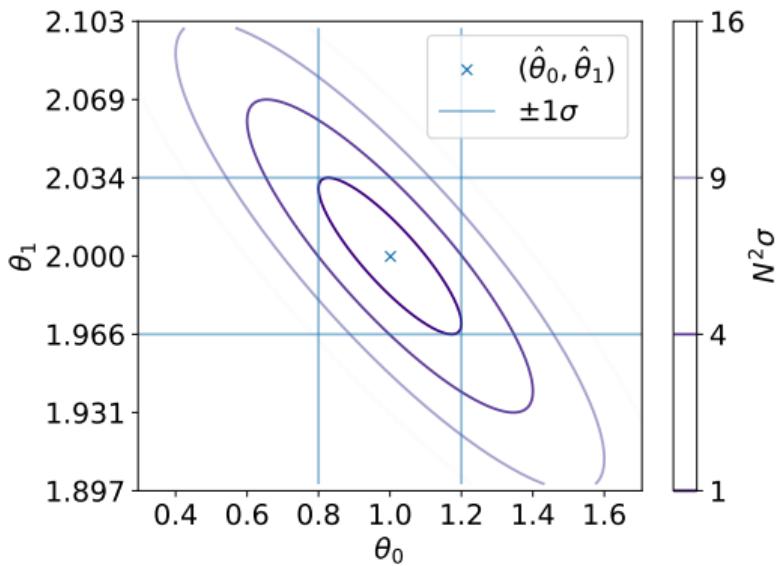
$$p(\mathbf{X} | \theta_0, \theta_1) = \prod_{x_i, y_i, \sigma_i \in \mathbf{X}} \mathcal{N}(y_i | \mu(x_i | \theta_0, \theta_1), \sigma_i^2)$$

$$\mu(x_i | \theta_0, \theta_1) = \theta_0 + \theta_1 x_i$$

→ We want to know about θ_0 , do not care about θ_1 .

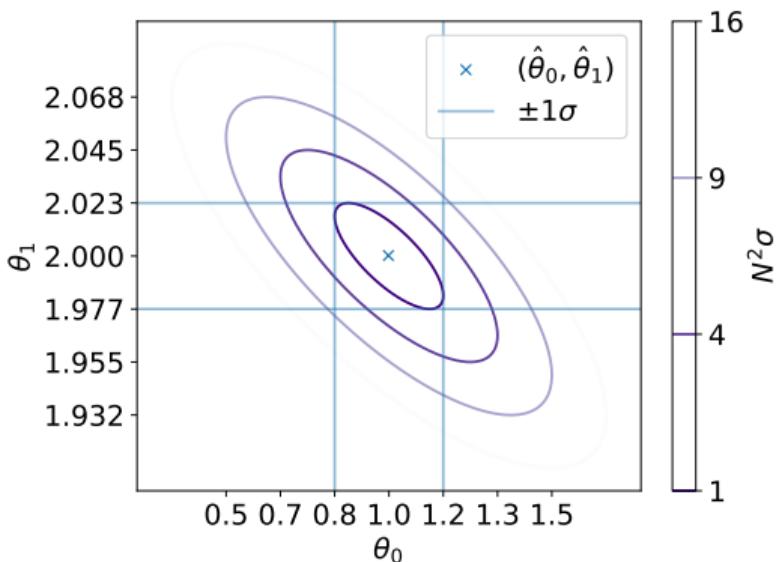
Unconstrained likelihood

$$-2 \log p(\mathbf{X} | \theta_0, \theta_1) = \sum_{x_i, y_i, \sigma_i \in \mathbf{X}} \frac{(y_i - \mu(x_i | \theta_0, \theta_1))^2}{\sigma_i^2}$$



Including a measurement of θ_1 : t_1, σ_{t_1}

$$-2 \log p(\mathbf{X} | \theta_0, \theta_1) = \sum_{x_i, y_i, \sigma_i \in \mathbf{X}} \frac{(y_i - \mu(x_i | \theta_0, \theta_1))^2}{\sigma_i^2} + \frac{(\theta_1 - t_1)^2}{\sigma_{t_1}^2}$$

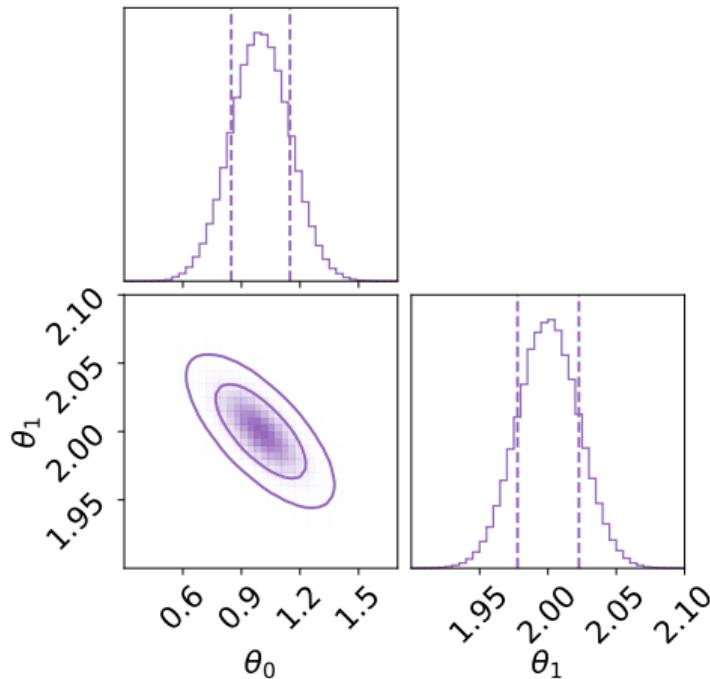


Posterior

$$p(\theta_0, \theta_1 | \mathbf{X}) \propto p(\mathbf{X} | \theta_0, \theta_1) p(\theta_0) p(\theta_1)$$

$$p(\theta_0) = \text{Uniform}(0, 2)$$

$$p(\theta_1) = \mathcal{N}(\theta_1 | t_1, \sigma_{t_1}^2)$$

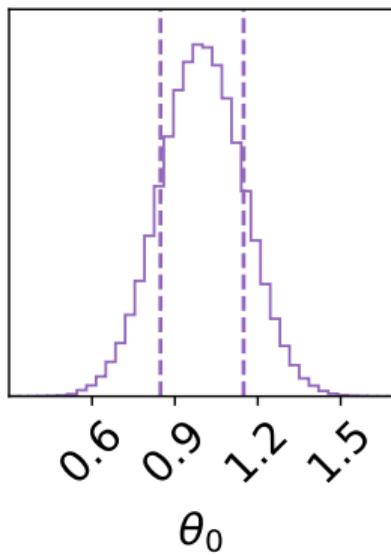


Marginal posterior

$$p(\theta_0 | \mathbf{X}) = \int d\theta_1 p(\theta_0, \theta_1 | \mathbf{X}) = \mathcal{N}(\theta_0 | \theta_0^*, \sigma_{\theta_0})$$

In this example, we get

- $\theta_0^* = \hat{\theta}_0$
- 68% HDI = $\hat{\theta}_0 \pm \sigma_{\theta_0}$



MCMC

The hard part ...

- We only want the posterior for θ alone.
 - Remove nuisance parameters by integrating over ν .
- The *marginal posterior* is

$$p(\theta|x) = \int d\nu p(\theta, \nu|x)$$

- Commonly a high dimensional integral
→ compute with Monte Carlo methods.

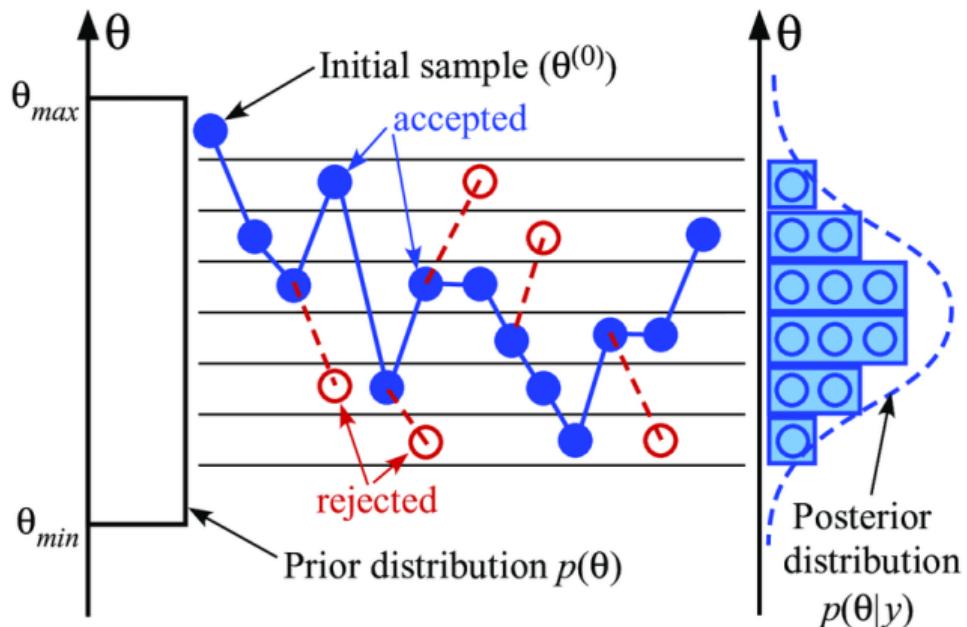
Markov chain

A sequence of events, where probability of the next state depends solely on the current state

$$\dots \rightarrow \theta_i \sim g(\theta_i | \theta_{i-1}) \rightarrow \theta_{i+1} \sim g(\theta_{i+1} | \theta_i) \rightarrow \dots$$

for some *proposal distribution* g .

Markov Chain Monte Carlo (MCMC)



Metropolis-Hastings

We loop

1. Generate $\theta \sim g(\theta|\theta_i)$
2. Update

$$\theta_{i+1} = \begin{cases} \theta & u \leq \min\left(1, \frac{p(\theta)g(\theta|\theta_i)}{p(\theta_i)g(\theta_i|\theta)}\right) \\ \theta_i & \text{otherwise} \end{cases}$$

where $u \sim \text{Uniform}(0, 1)$

Note: Need to define a proposal distribution $g(\theta|\theta_0)$.

Chains

In MCMC we generate a sequence

$$\theta_0 \rightarrow \theta_1 \rightarrow \theta_2 \rightarrow \dots$$

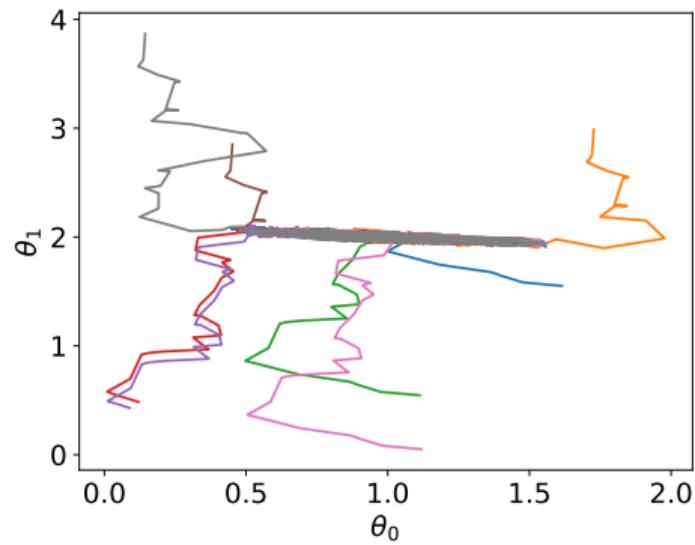
Only one start can land you
in local minima.

$$\theta_0^0 \rightarrow \theta_1^0 \rightarrow \theta_2^0 \rightarrow \dots$$

$$\theta_0^1 \rightarrow \theta_1^1 \rightarrow \theta_2^1 \rightarrow \dots$$

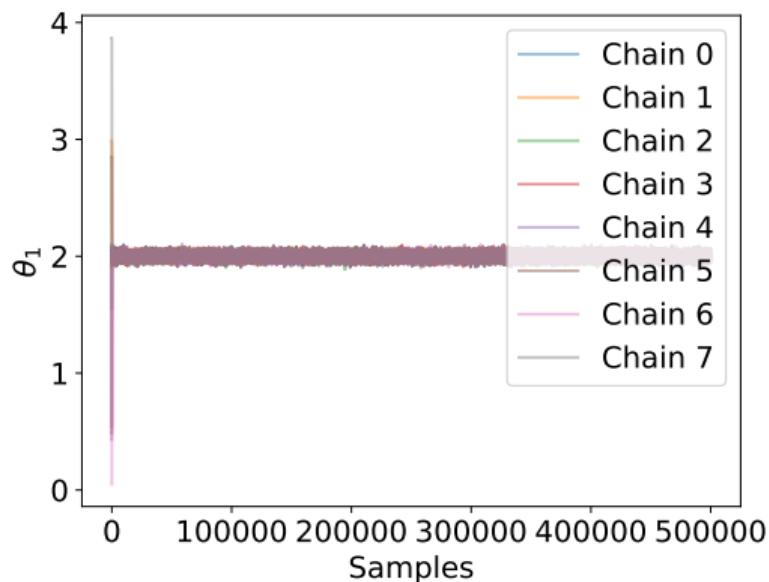
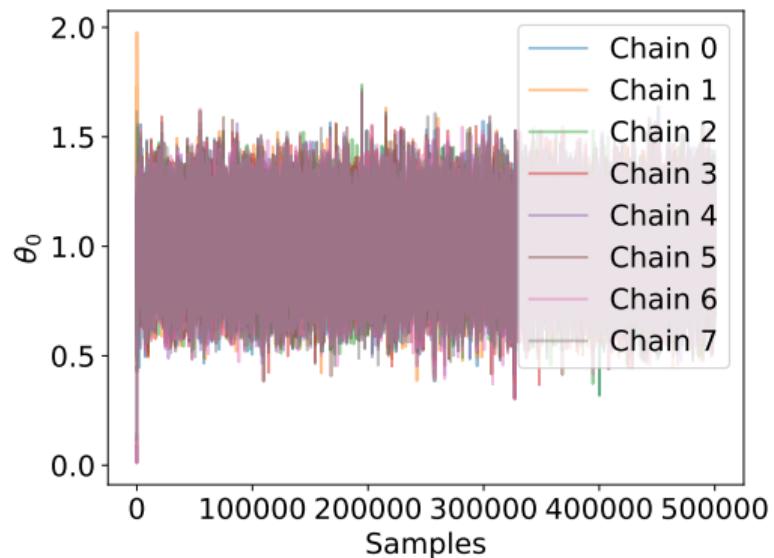
$$\theta_0^2 \rightarrow \theta_1^2 \rightarrow \theta_2^2 \rightarrow \dots$$

...



Convergence

Trace plots are a useful convergence diagnostic



... but one can become more fancy.

Model comparison

Point-wise: P-values

$$P = \int_{\lambda_{obs}}^{\infty} d\lambda p(\lambda) = 1 - CDF_p(\lambda_{obs}), \quad \lambda = -2 \ln \frac{L(\boldsymbol{\eta} = 0, \hat{\boldsymbol{\chi}})}{L(\hat{\boldsymbol{\eta}}, \hat{\boldsymbol{\chi}})}$$

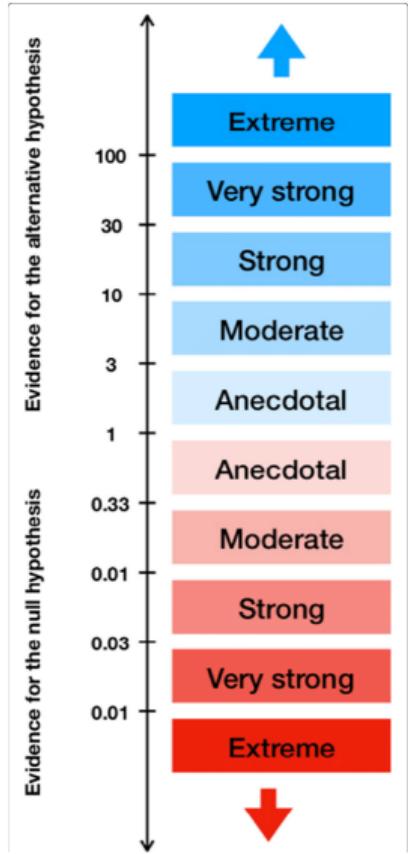
Averaged: Bayes factor

Compare the probabilities of the observed data being produced by a given model.

$$p(\theta|x, M) = \frac{p(x|\theta, M) p(\theta|M)}{p(x|M)}$$

$$p(x|M) = \int d^n\theta p(x|\theta, M) p(\theta|M)$$

$$B = \frac{p(x|M_1)}{p(x|M_0)}$$



$b \rightarrow ul^-\bar{\nu}$ in the Weak Effective Theory

fit model M	goodness of fit			
	χ^2	d.o.f.	p value [%]	$\ln Z(M)$
SM	44.18	48	63.03	372.5 ± 0.4
CKM	43.75	47	60.78	372.4 ± 0.4
WET	36.13	43	76.17	376.5 ± 0.4

Table 1. Goodness-of-fit values for the three main fits conducted as part of this analysis. We provide $\chi^2 = -2 \ln P(\text{data} | \vec{x}^*)$ at the best-fit point \vec{x}^* next to the p value and the natural logarithm of the evidence $\ln Z$. We find that the p values associated with each individual likelihood are larger than 42%.

$$B = \exp(\ln Z(WET) - \ln Z(SM)) = 54.6$$

arXiv:2302.05268v2 [hep-ph]

Tools to try

