LUDWIG-
MAXIMILIANS-
UNIVERSITÄT
MÜNCHEN

Illuminating the dark side of statistics: Bayesian inference in particle physics

Lorenz Gärtner

LMU Munich

26.03.2025

ORIGINS Excellence Cluster

Belle II

PUNCH 4NFDI

DFG

About me

- BSc @ University of Manchester
  Physics with theoretical physics

- MSc @ LMU Munich
  More theory . . .

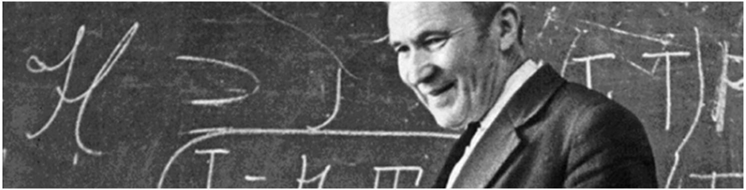
  – – – almost no stats – – –

- Currenly PhD @ LMU Munich
  A lot of stats

➜ Never too late to start

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 $x_1, x_2, \ldots,$

$$p\left(\bigcup_i x_i\right) = \sum_i p(x_i)$$

Probability interpretations

Axioms tell you how to calculate with probabilities.

How do we assign probabilities in the first place?

Probability interpretations

Axioms tell you how to calculate with probabilities.

How do we assign probabilities in the first place?
➜ Need probability interpretations.

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

Frequentist interpretation

Assign a probability as relative frequency

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

- For repeatable experiments only.
- **Data** is random.

Bayesian interpretation

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

- Inference results are subjective.
- **Parameters** are random.

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** $p(\text{theory})$
- **Marginal likelihood**
  $p(\text{data}) = \int p(\text{data}|\text{theory})p(\text{theory})$

[...] *nearly all physicists tend to misinterpret frequentist results as statements about the theory given with the data.*

A. L. Read

Bayesian beliefs

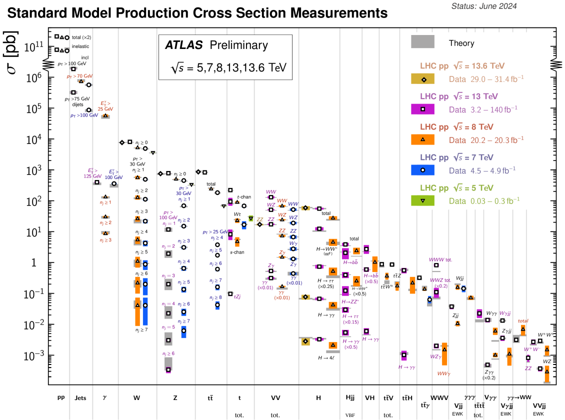
$$p(SM|data)$$

$$p(SUSY|data)$$

Bayesian beliefs

$$p(SM|data)$$

$$p(SUSY|data)$$



ATLAS

Bayesian updating

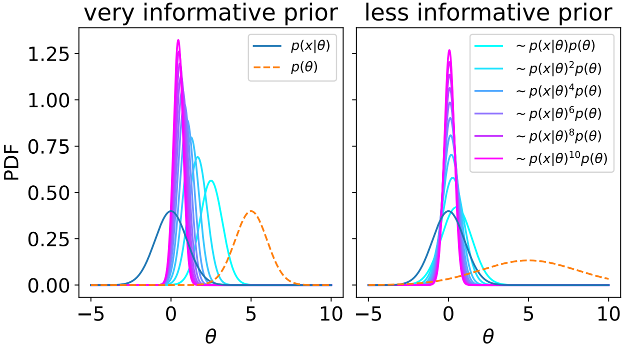
Can generally use measurements to *update* our posterior

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

Bayesian updating

Can generally use measurements to *update* our posterior

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



The common ground

Frequentist inference is based on

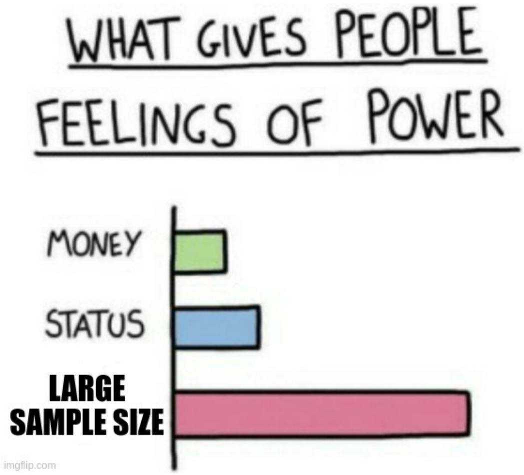
$$\boxed{p(x|\theta)}$$

Bayesian inference is based on

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

You want the best possible $\boxed{\boldsymbol{p(x|\theta)}}$!

The best possible model...



How do we make data useful?

Parameter inference

Point estimates

Identify the most probable parameter point.

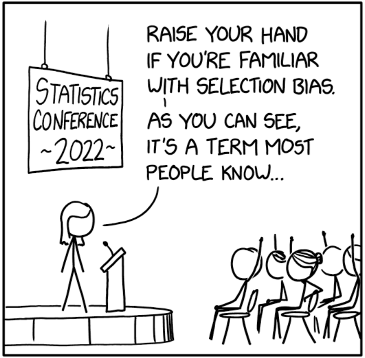
Interval estimation

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

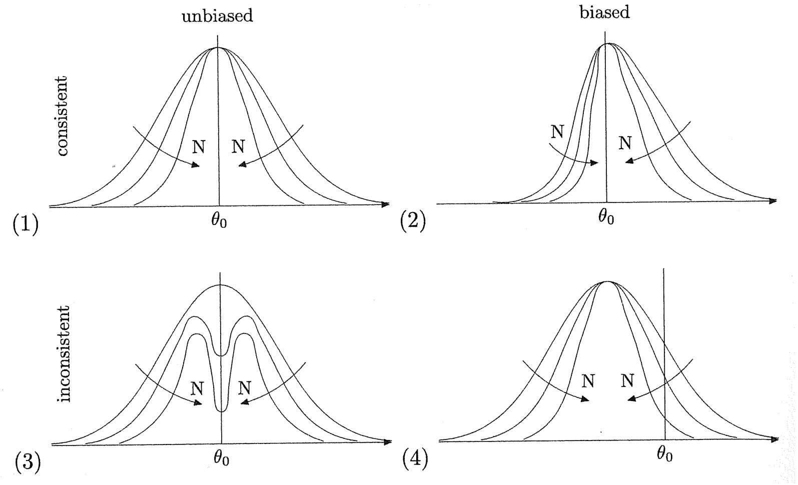
Frequentist point estimates: estimators

Estimator is a *statistic* $\hat{\theta}(x)$, with desired properties

- **consistency**
  $\lim_{N_x \to \infty} E_x[\hat{\theta}] = \theta_{true}$
- **unbiasedness**
  $b = E_x[\hat{\theta}] - \theta_{true}$
- **efficiency**
  $$V(\hat{\theta}) = I(\theta)^{-1} = E_x\left[\left(\frac{\partial \ln p(x|\theta)}{\partial \theta}\right)^2\right]^{-1}$$
- ...



Estimator properties



Statistical Methods in Experimental Physics, F. James

Method of maximum likelihood

Maximum likelihood estimators $\hat{\theta}$ by solving

$$\hat{\theta} = \mathrm{argmax}_{\theta}\, p(x|\theta)$$

Method of maximum likelihood

Maximum likelihood estimators $\hat{\theta}$ by solving

$$\hat{\theta} = \text{argmax}_\theta \, p(x|\theta)$$

Asymptotically,

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

➜ consistency, asymptotic unbiasedness and efficiency

Bayesian point estimates

Mode

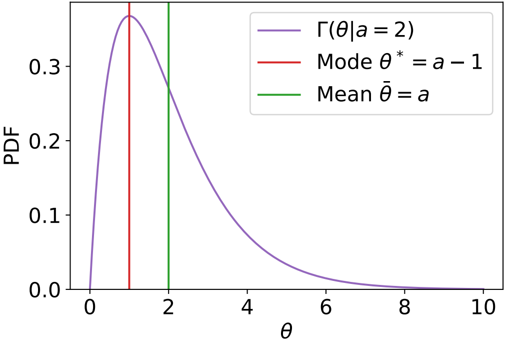
Value of θ with maximum posterior probability

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

Mean

Expected value of θ under the posterior

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



Mode vs. ML estimator

Stationary point of the posterior at θ^*:

$$0 = \frac{\partial p(\theta|x)}{\partial \theta}\bigg|_{\theta=\theta^*} \propto \left(\frac{\partial p(x|\theta)}{\partial \theta} p(\theta) + p(x|\theta)\frac{\partial p(\theta)}{\partial \theta}\right)\bigg|_{\theta=\theta^*}$$

Mode vs. ML estimator

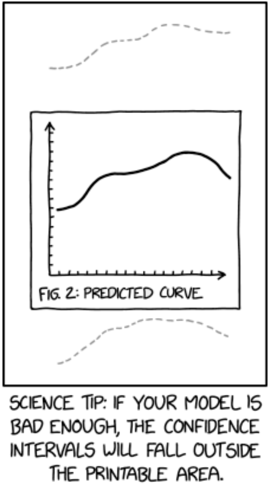
Stationary point of the posterior at θ^*:

$$0 = \frac{\partial p(\theta|x)}{\partial\theta}\bigg|_{\theta=\theta^*} \propto \left(\frac{\partial p(x|\theta)}{\partial\theta}p(\theta) + p(x|\theta)\frac{\partial p(\theta)}{\partial\theta}\right)\bigg|_{\theta=\theta^*}$$

$$\boxed{\implies \theta^* = \hat{\theta} \quad \text{if} \quad \frac{\partial p(\theta)}{\partial\theta}\bigg|_{\theta=\theta^*} = 0}$$

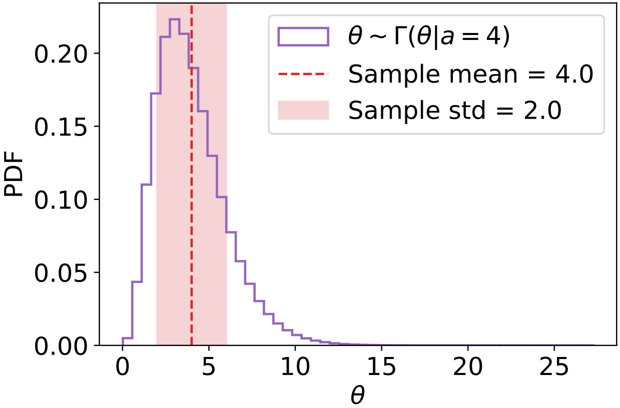
Posterior mode and ML estimate agree for *flat* priors.

Intervals and limits



Non-Normal PDFs

For non-normal estimator PDFs, $\hat{\theta} \pm \sigma_\theta$ can be misleading.



Frequentist *confidence* intervals

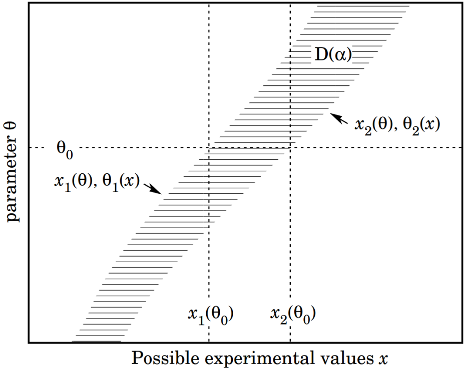
Neyman confidence belt

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

Not unique ➜ *central interval*

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

or *upper/lower interval*

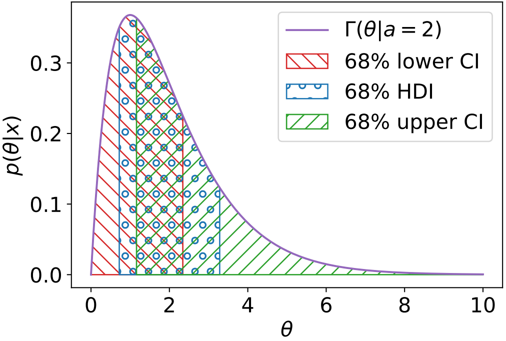


Bayesian *credible* intervals

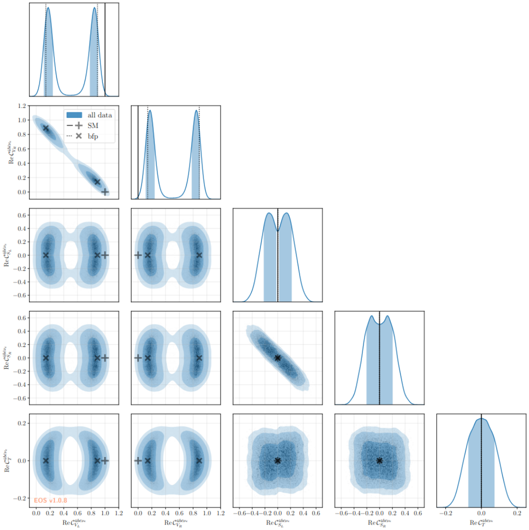
Credible intervals (CI) $[\theta_1, \theta_2]$ cover $1 - \alpha$ of the posterior

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

- Upper/lower CI
- Highest (posterior) density intervals (HDI)



$b \to u l^- \bar{\nu}$ in the Weak Effective Theory



- Corner plots are great for visualization
- Posterior for Wilson coefficients

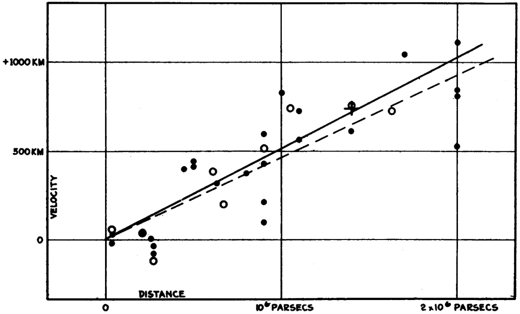
➜ Modes, credible intervals, . . .

arXiv:2302.05268v2 [hep-ph]

Nuisance parameters and priors

Nuisance parameters

Models are not perfect
➜ **systematic bias**



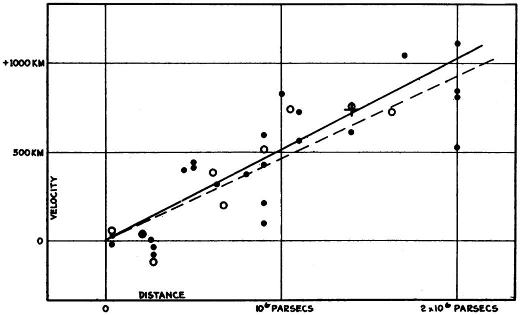
Hubble 1929

Nuisance parameters

Models are not perfect
➜ **systematic bias**

Solution:
Nuisance parameters ν,

$$p(x|\psi, \nu)$$



Hubble 1929

Generally, want to **constrain** nuisance parameters.

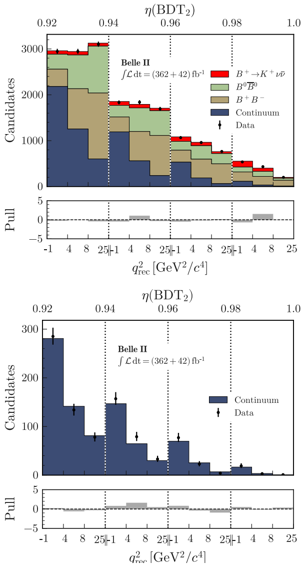
Frequentist "priors"

Frequentist: **everything is data**

Constrain ν with *auxiliary data a*,

$$p(x|\psi,\nu)p(a|\nu).$$

Belle II 2024



Frequentist "priors"

Frequentist: **everything is data**

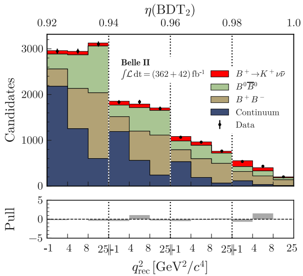
Constrain ν with *auxiliary data a*,

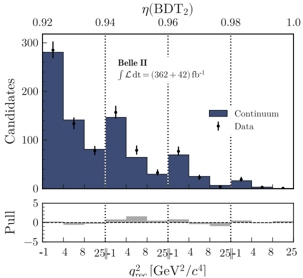
$$p(x|\psi,\nu)p(a|\nu).$$

Often: **create** *auxiliary data* to match our desired constraint term.

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

Belle II 2024





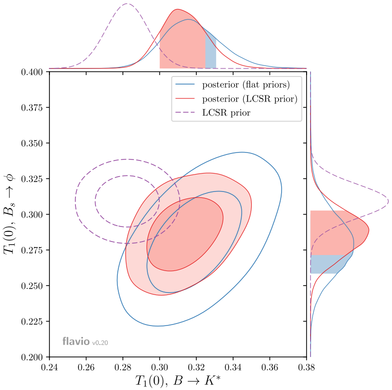
Bayesian nuisance parameters

Priors from auxiliary data

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

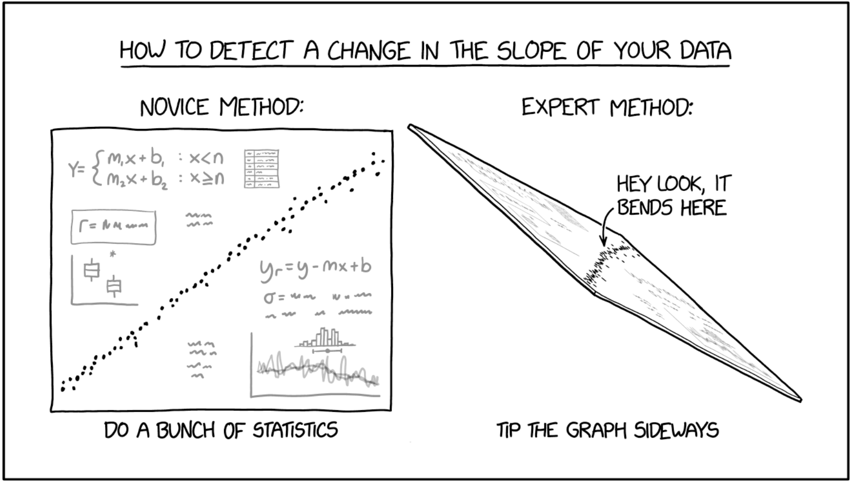
Only Bayesian allows other prior choices

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



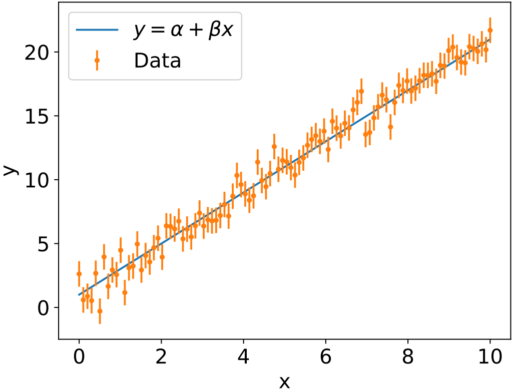
Paul 2017

A simple linear model



A simple linear model

- Independent data : $\boldsymbol{X} = (x_i, y_i, \sigma_i)$



A simple linear model

- Independent data: $\boldsymbol{X} = (x_i, y_i, \sigma_i)$

➜ Model = product of normal distributions:

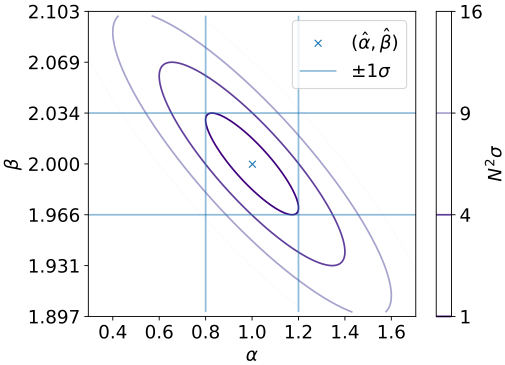
$$p(\boldsymbol{X}|\alpha,\beta) = \prod_{x_i, y_i, \sigma_i \in \boldsymbol{X}} \mathcal{N}(y_i|\mu(x_i|\alpha,\beta), \sigma_i^2)$$

$$\mu(x_i|\alpha,\beta) = \alpha + \beta x_i$$

- Care about $\psi = \alpha$, not $\nu = \beta$.

Frequentist analysis

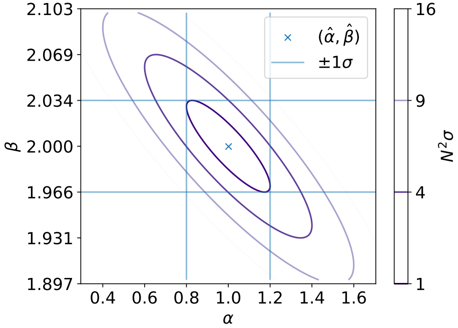
$$-2\log p(\boldsymbol{X}|\alpha,\beta) = \sum_{x_i, y_i, \sigma_i \in \boldsymbol{X}} \frac{(y_i - \mu(x_i|\alpha,\beta))^2}{\sigma_i^2}$$



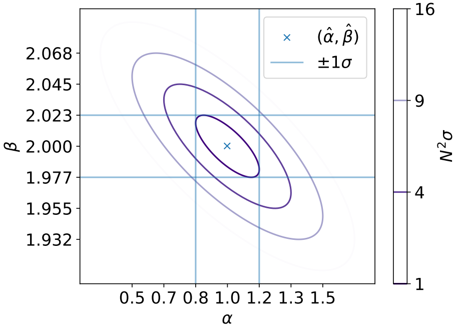
Including a measurement of β: b, σ_b

$$-2\log p(\boldsymbol{X}|\alpha,\beta) = \sum_{x_i,y_i,\sigma_i \in \boldsymbol{X}} \frac{(y_i - \mu(x_i|\alpha,\beta))^2}{\sigma_i^2} + \frac{(\beta - b)^2}{\sigma_b^2}$$

unconstrained



constrained

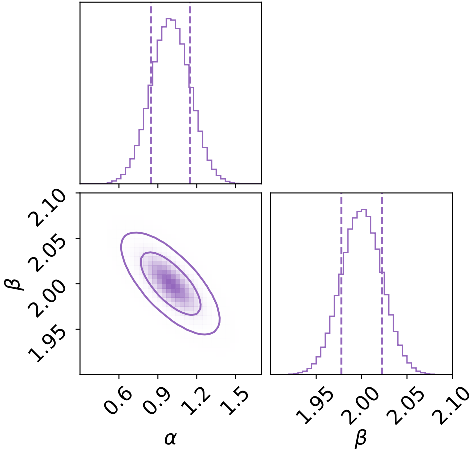


Posterior

$$p(\alpha, \beta|\boldsymbol{X}) \propto p(\boldsymbol{X}|\alpha, \beta)p(\alpha)p(\beta)$$

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

$$p(\beta) = \mathcal{N}(\beta|b, \sigma_b^2)$$

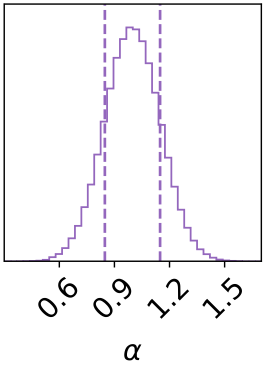


Marginal posterior

$$p(\alpha|\boldsymbol{X}) = \int d\beta \, p(\alpha, \beta|\boldsymbol{X}) = \mathcal{N}(\alpha|\alpha^*, \sigma_\alpha)$$

In this example, we get

- $\alpha^* = \hat{\alpha}$
- 68% HDI = $\hat{\alpha} \pm \sigma_\alpha$



How do we marginalize?

Interested in $p(\psi|x)$ and not $p(\psi, \nu|x)$.

➜ **Marginal posterior**

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

Commonly a high dimensional integral

➜ **Monte Carlo integration**

Markov Chain Monte Carlo (MCMC)

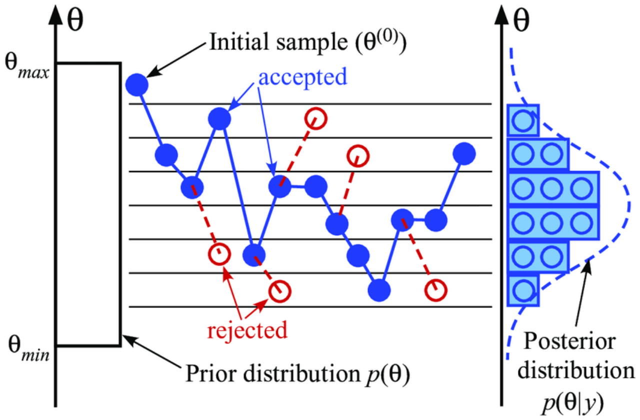
Markov chain

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

$$\ldots \to \theta_i \sim g(\theta_i|\theta_{i-1}) \to \theta_{i+1} \sim g(\theta_{i+1}|\theta_i) \to \ldots$$

for some *proposal distribution* g.

MCMC integration



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: for example $g(\theta|\theta_0) = \mathcal{N}(\theta|\theta_0, \sigma)$.

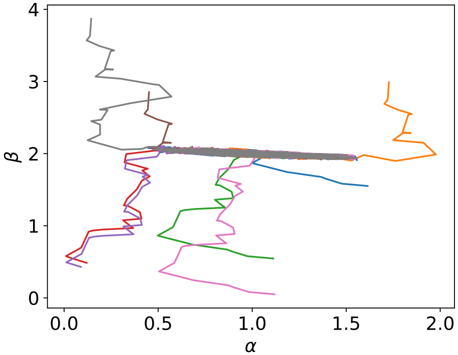
Chains

In MCMC we generate a sequence

$$\theta_0 \to \theta_1 \to \theta_2 \to \dots$$

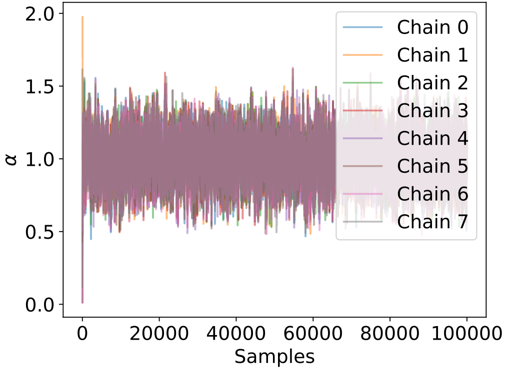
Only one start can land you a in local minima.

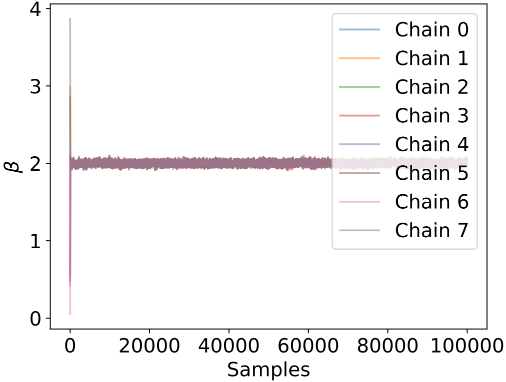
$$\theta_0^0 \to \theta_1^0 \to \theta_2^0 \to \dots$$
$$\theta_0^1 \to \theta_1^1 \to \theta_2^1 \to \dots$$
$$\theta_0^2 \to \theta_1^2 \to \theta_2^2 \to \dots$$
$$\dots$$



Convergence

Trace plots are a useful convergence diagnostic





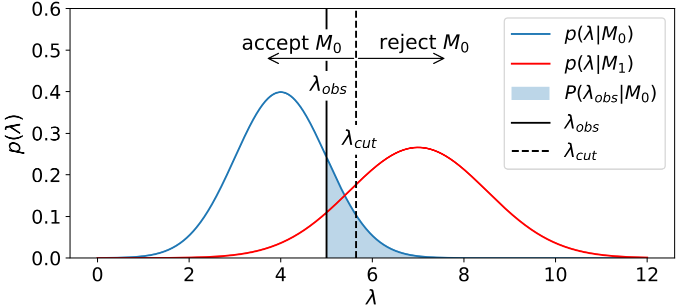
... but one can become more fancy.

Tools to try

How do we compare model performance?

Frequentist: P-values

$$P(\lambda_{obs}|M_0) = \int_{\lambda_{obs}}^{\infty} d\lambda\, p(\lambda|M_0), \quad \lambda = -2\ln\frac{p(x|\hat{\theta}_0, M_0)}{p(x|\hat{\theta}_1, M_1)}\,^{\dagger}$$



† Likelihood ratio = optimal test statistic ➜ Newman-Pearson lemma

Averaged: Bayes factor

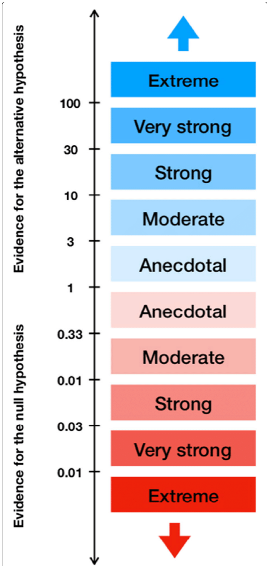
Compare the probabilities of the observed data being produced by a given model.

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

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

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

Do you see a potential hazard?



$b \to u l^- \bar{\nu}$ in the Weak Effective Theory

	goodness of fit			
fit model M	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]