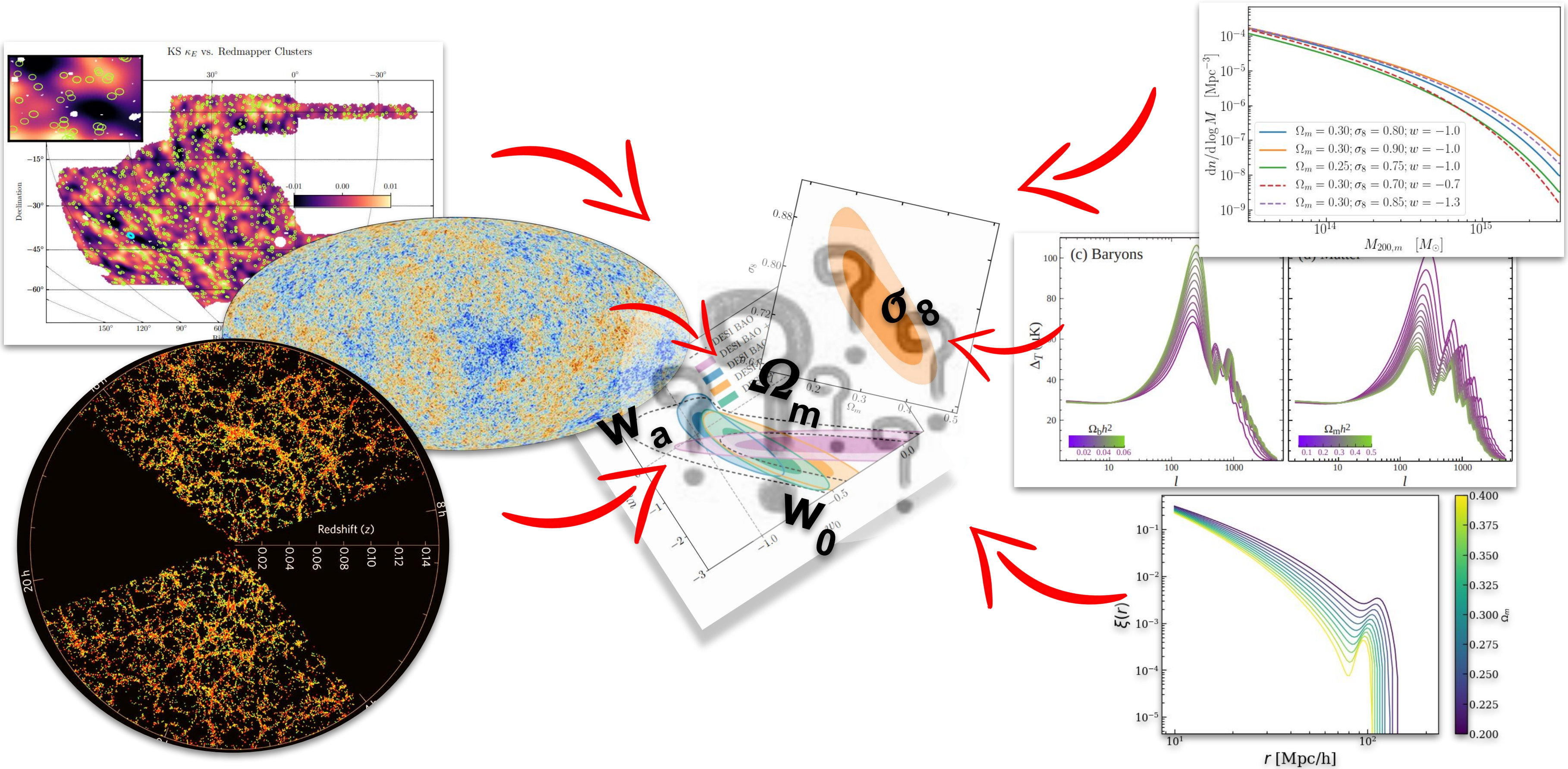
OBSERVATIONAL COSMOLOGY:

COSMOLOGICAL INFERENCE

A.A. 2025/2026 | Matteo Costanzi - University of Trieste / INAF

COSMOLOGICAL INFERENCE



BAYESIAN AND FREQUENTIST STATISTICS

- Probability as frequency:

The classical approach to statistics defines the probability of an event as "the number of times the event occurs over the total number of trials, in the limit of an infinite series of equiprobable repetitions."

- Probability as degree of belief:

The Bayesian viewpoint is based on the simple and intuitive tenet that: "probability is a measure of the degree of belief about a proposition".

BAYESIAN AND FREQUENTIST STATISTICS

- **Frequentist: model is fixed, data are repeatable. The model parameters are treated as fixed** (though unknown) **constants**, and probability quantifies the variability of data under repeated sampling.

- **Bayesian: data are fixed, model is repeatable. The unknown model parameters are modeled as random variables** with probability distributions, and probability statements refer to degree of belief given available information

This difference pervades all methodology: in Bayesian analysis even *single* events or non-repeatable parameters are assigned probabilities, whereas frequentists only assign probabilities to repeatable random events.

BAYESIAN AND FREQUENTIST STATISTICS

Say H_0 = (72 ± 8) km/s/Mpc.

- **Bayesian:** the posterior distribution for H_0 has 68% if its integral between 64 and 80 km/s/Mpc. The posterior can be used as a prior on a new application of Bayes' theorem.
- **Frequentist:** Performing the same procedure will cover the real value of H_0 within the limits 68% of the time. But how do I repeat the same procedure (generate a new H_0) if I only have one Universe?

Good references:

Bayesian: R. Trotta, “Bayes in the Sky”, https://arxiv.org/abs/0803.4089

Frequentist: Feldman & Cousins, “A Unified Approach to the Classical Statistical Analysis of Small Signals”, https://arxiv.org/abs/physics/9711021

Example of one cosmology inference done both Bayesian and frequentist way: G. Efstathiou, “The Statistical Significance of the Low CMB Multipoles”, https://arxiv.org/abs/astro-ph/0306431

Or “Robust constraints on tensor perturbations from cosmological data: a comparative analysis from Bayesian and frequentist perspectives” https://arxiv.org/pdf/2405.04455

BAYESIAN AND FREQUENTIST STATISTICS

- Bayesian:
  - can given probabilities for models
  - depends on both prior and likelihood (of data)
  - currently the dominant method in cosmology
- Frequentist:
  - doesn't give probabilities of models, only of hypotheses
  - doesn't depend on prior, just likelihood
  - currently the dominant method in particle physics

BAYES' THEOREM

Posterior

What you know after the experiment

$$P(p|dM) = \frac{P(d|pM)P(p|M)}{P(d|M)}$$

Evidence

Normalization constant

$$\int dpP(d|p,M)P(p|M)$$

Likelihood

What you learn from the experiment

$$\propto P(d|pM)P(p|M)$$

Prior

What you knew before the experiment

Observed data

Parameters

Model

Note: P(A|B) reads "the probability of A given B"

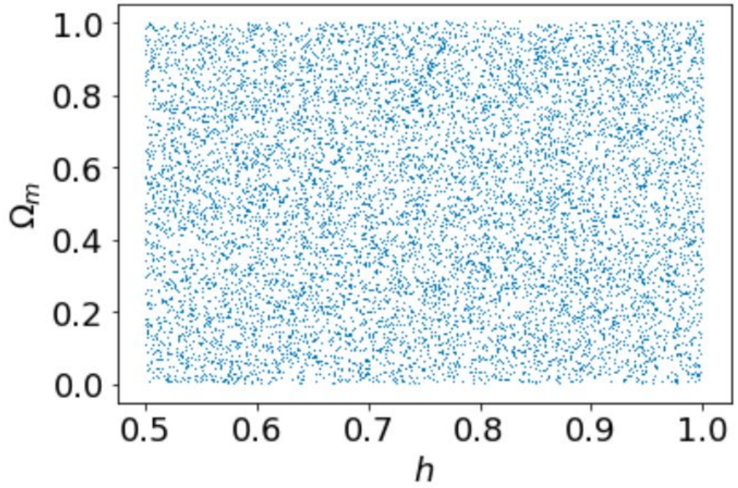
PRIORS

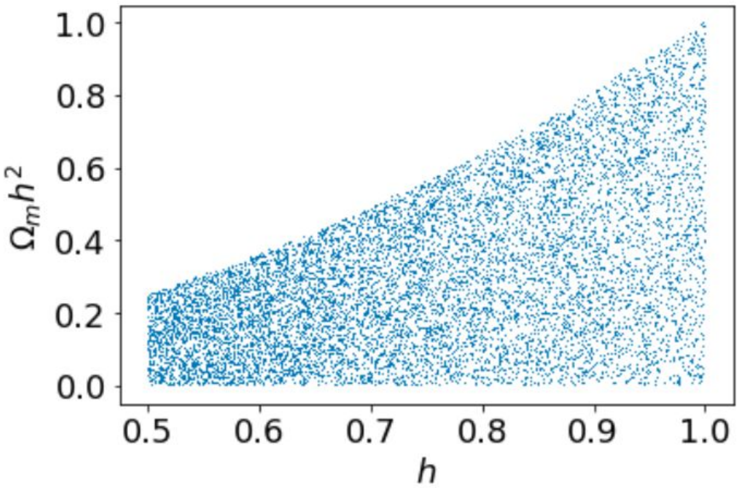
- Priors quantify what you knew about the parameters before you start
    - Theoretical limits, preferences, things that must be true from simpler data
- In regions where your likelihood is zero your prior doesn't matter for parameter estimation, but can for more advanced *model selection*
- It is common practice in cosmology to use uniform priors for most parameters

The rationale is that we should assign equal probability to equal states of knowledge. However, flat priors are not always as harmless as they appear. One reason is that a flat prior on a parameter θ does not correspond to a flat prior on a non–linear function of that parameter, $\psi(\theta)$. The two priors are related by:

$$p(\psi) = p(\theta) \left| \frac{\mathrm{d}\theta}{\mathrm{d}\psi} \right|$$

PRIORS





Jointly uniform priors on Ω_m - h

Implied priors on $\Omega_m h^2$ - h

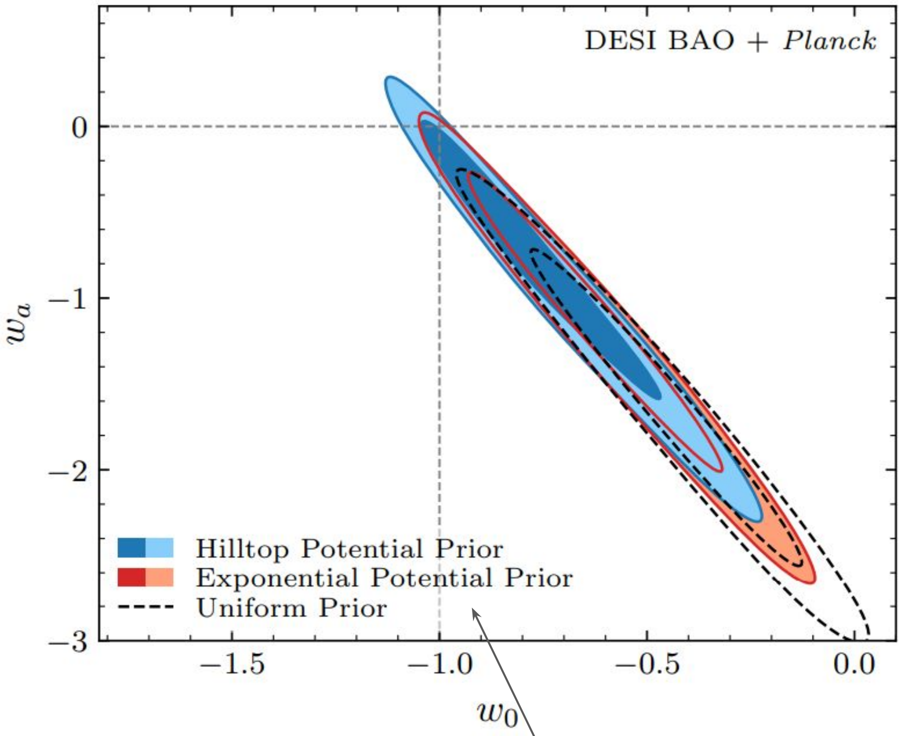
E.g. The parameter actually constrained by CMB data

PRIORS

How Theory-Informed Priors Affect DESI Evidence for Evolving Dark Energy (Toomey et al 2025):

Although uniform priors are sometimes adopted in an attempt to minimize subjective influence, this choice is far from neutral in a Bayesian framework. The use of broad uniform priors on the CPL parameters w_0 and w_a, often mischaracterized as “uninformative,” represents a specific choice that influences the posterior distributions, especially in the context of DESI measurements.

FIG. 1. Two-dimensional posteriors in w_0 and w_a at the 68% and 95% confidence levels from a joint fit to *Planck* PR4 CMB anisotropies with lensing and DESI DR2 BAO under different priors for evolving dark energy. The gray dashed lines mark the ΛCDM limit of the $w_0 w_a$CDM model, namely $(w_0, w_a) = (-1, 0)$. For two degrees of freedom, the results are consistent with ΛCDM at the $\sim 1.3\sigma$ (hilltop, blue) and $\sim 1.8\sigma$ (exponential, red) levels, in contrast with uniform priors (dashed black) that yield a $\sim 3.1\sigma$ deviation. The figure highlights the impact of prior choices, with theory-informed priors that shift the inference toward ΛCDM, significantly reducing the nominal preference for evolving dark energy.



two representative classes of thawing quintessence where DE is driven by a single scalar field ϕ minimally coupled to gravity

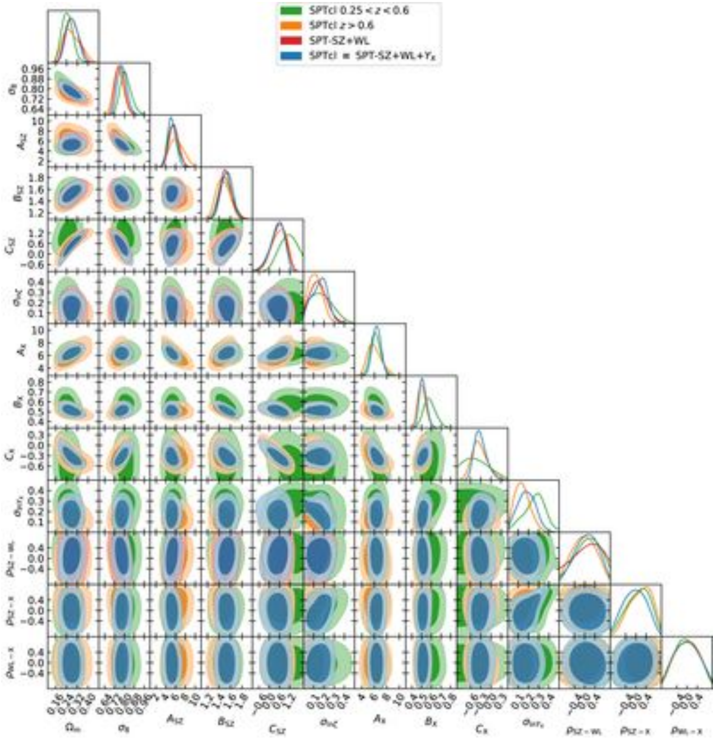
The challenge: map out a posterior in multi-dimensional parameter space.

Example: say there are just 10 parameters.
Lets say calculation takes just 1 second/model.
Say you want a grid with 20 values in each par.
Then

$$N = 20^{10} \simeq 10^{13}$$

$\Rightarrow$ it would take 300,000 years to do it!

$\Rightarrow$ Totally impossible, ever!!



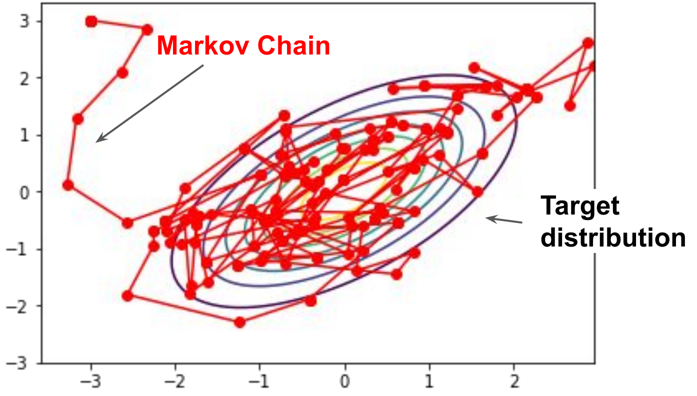
POSTERIOR ESTIMATION

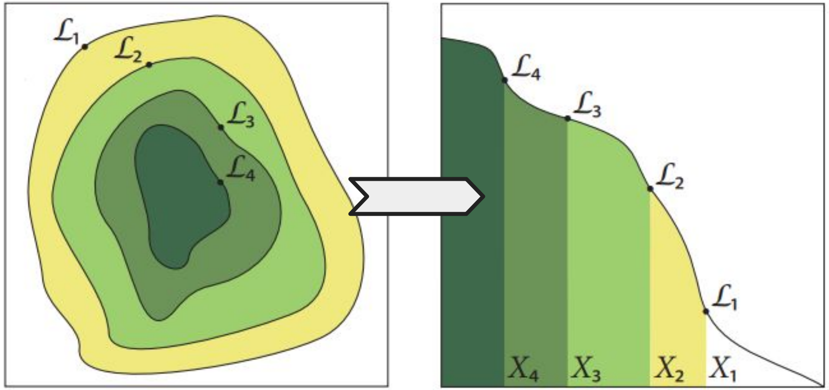
Over the past years many sampling techniques have been developed to overcome this issue (See this for a review). The general idea is to sample the parameter space in a clever way in order to map out the high-probability volumes. The methods can be divided in:

- **Monte Carlo Markov Chains methods: e.g.** `Metropolis-Hastings` **(Metropolis+1953),** `Emcee` **(Foreman-Mackey+2010)**

- **Nested sampling methods: e.g.** `Multinest` **(Feroz+2009,2013),** `Polychord` **(Handley+2015)**

In both case the density of the sampled points is proportional to the parameter posterior we seek to estimate





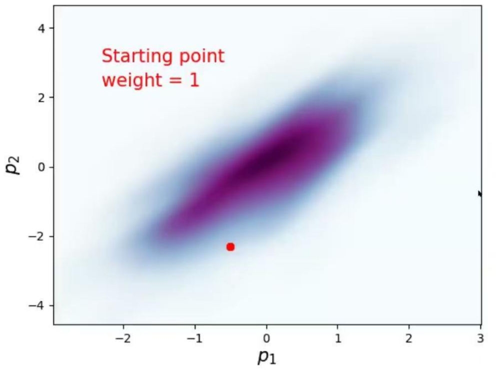
(a) (b)

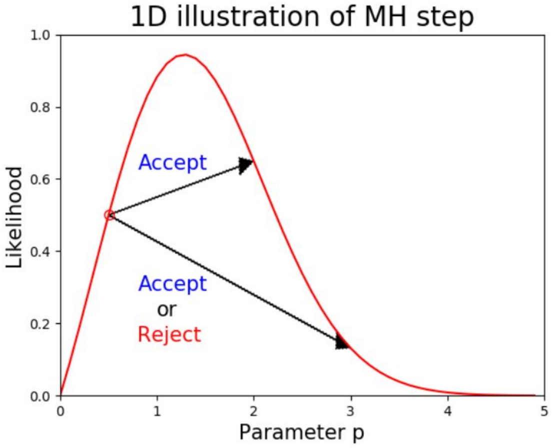
From Feroz+13

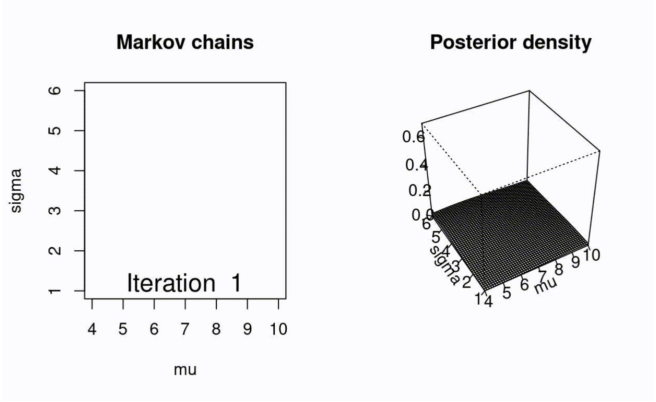
See Fowlie et al for a comparison of methods

MCMC: METROPOLIS-HASTING ALGORITHM

- at step t, at some parameters p_t
- propose move to $p_t' = p_t + \Delta p_t$ (randomly draw Δp_t)
- evaluate $r = L(p_t')/L(p_t)$
- MH step:
  - if r > 1 accept move
  - if r < 1 generate a random number $\alpha \in [0, 1]$
    - if α < r, accept move
    - if α > r, reject move
- t=t+1







One can prove that,
with this rule,
one asymptotically recovers the
true posterior

NESTED SAMPLING METHODS

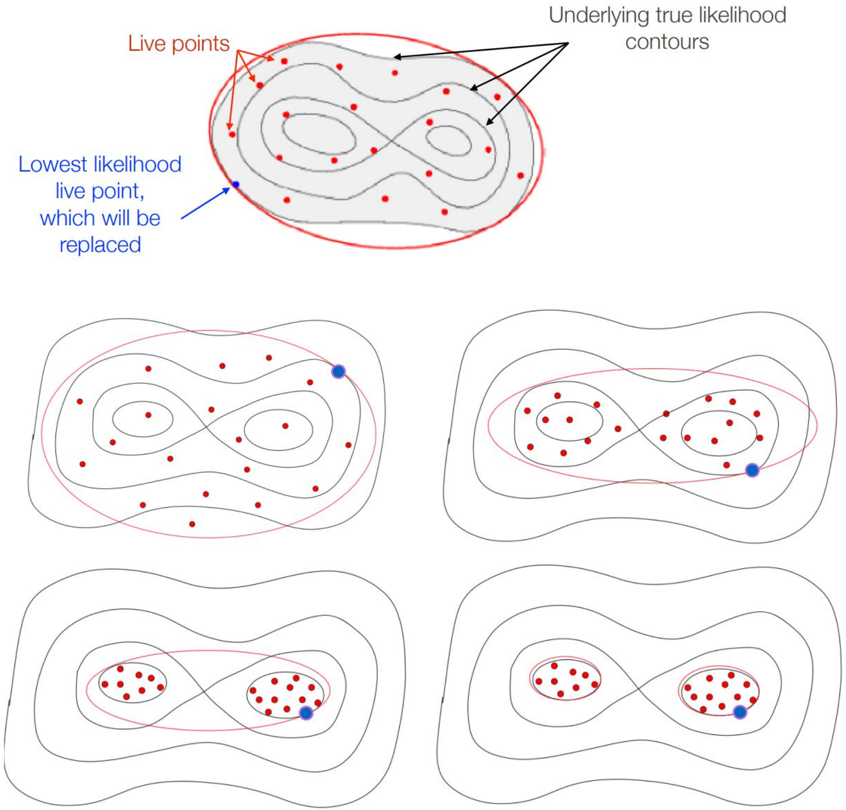
CORE IDEA: Instead of sampling the entire prior space (like MCMC), NS sequentially shrinks the sampling region by discarding low-likelihood points, concentrating on high-probability regions.

Algorithm:

1. Sample N point from the prior distribution
2. At each step *i-th:*
   a. Find the point with the lowest likelihood L_i
   b. Replace it with a new point having $L>L_i$
   c. Record the “dead point” and its likelihood
3. Stop when the remaining live point contribute negligibly to the evidence

Output:

1. Posterior samples
2. Evidence estimate

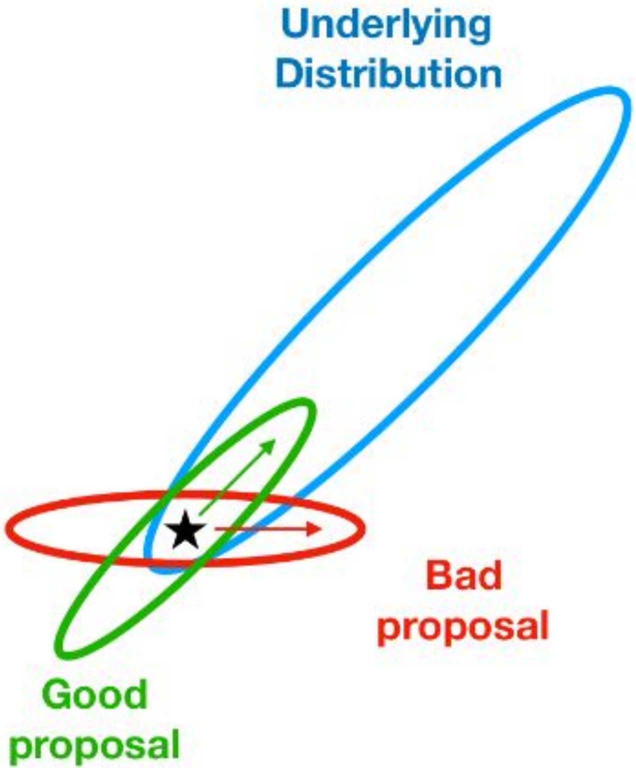


Lemos+22 (nice paper) to understand and compare sampler

Nice lecture on Nested methods

SAMPLING THE PARAMETER SPACE

- Efficiency of MH depends dramatically on how good the proposal is
- A bad proposal will not converge in any practical length of time
- The ideal proposal matches the shape of the underlying distribution
    - We don't know this, but can look for best approximation

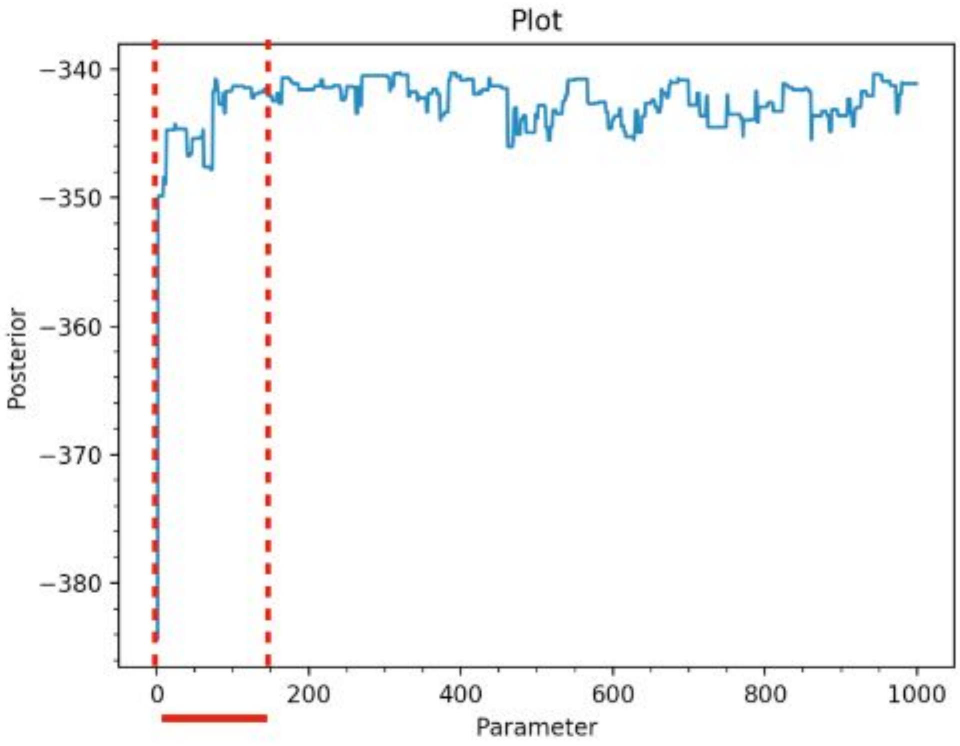


BURN-IN

- Unless you're doing a simulation where you know the truth, unlikely to start at the best-fit value

- Will take some iterations to get near this point

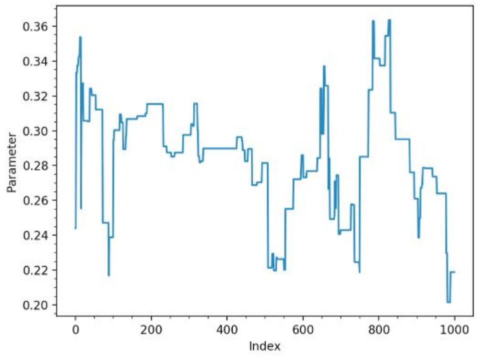
- Need to exclude these



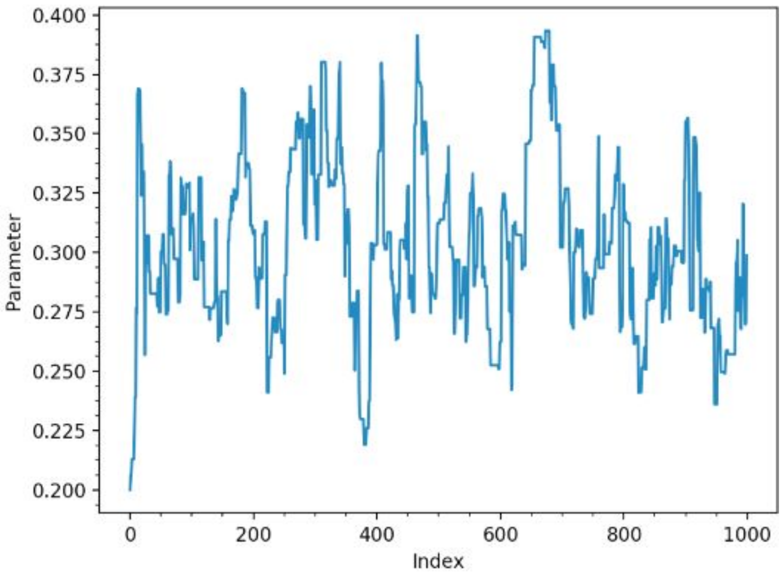
Burn-in - exclude from sampling

CHECKING CONVERGENCE

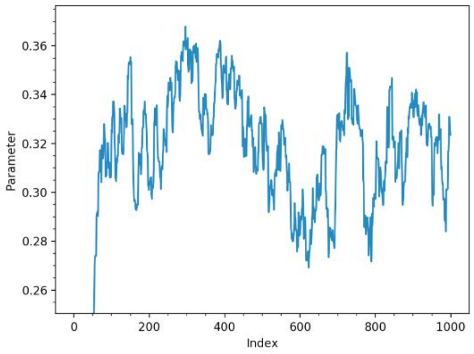
- Good MH chains look like white noise if you plot one parameters values throughout the chain



Bad - not long enough.
Chain getting stuck for long periods suggests covariance too large.
Acceptance rate too low.



Looks reasonable - could be a bit longer



Bad - not long enough.
Chain is random walking, taking long divergences from mean suggests covariance too small.
Acceptance rate too high.

INTERPRETING THE OUTPUT

WEIGHT	P_1	P_2	P_3		P_N
5	0.2	-0.3	0.15	...	2.8
1	-0.7	0.4	0.12	...	3.5
12	0.7	0.1	0.19	...	1.7
...	...	...	...	...	...
...	...	...	...	...	...

(~ MILLION ROWS)

To get the posterior probability,
simply histogram the parameter values vs weights - this is your posterior!

Want to look at posterior in p_3 marginalized over all other parameters?
Simply plot histogram of p_3 values vs weight (eaaasy!)

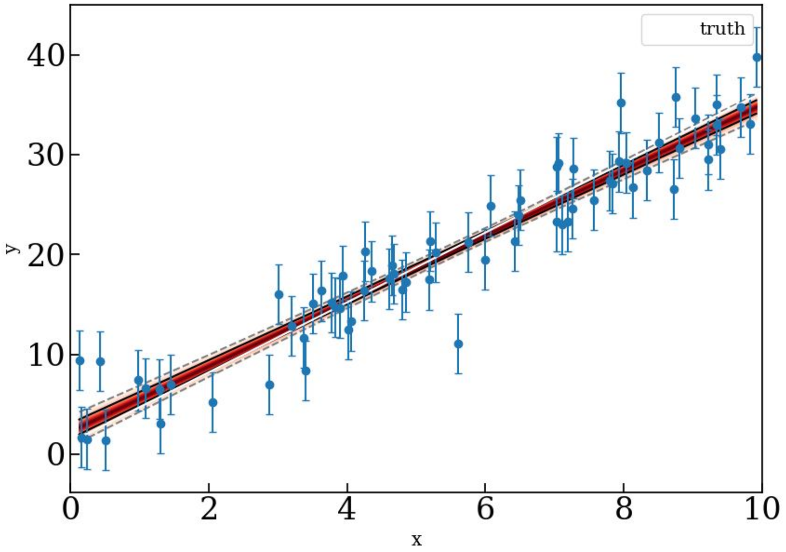
GOODNESS OF FIT

The goodness of fit is often estimated from the best-fit parameter values using a χ^2 statistic (which is formally correct only for Gaussian distributions):

$$\chi^2_{\mathrm{Best-fit}} = (\vec{d} - \vec{m}(\theta_{\mathrm{BF}}))^T C^{-1} (\vec{d} - \vec{m}(\theta_{\mathrm{BF}}))$$

where *C* is the data covariance matrix.

NOTE: This method does not account for the uncertainties on the estimated parameters ϑ.



GOODNESS OF FIT: p-value

- **To assess the goodness of fit from the $\chi^2_{best\text{-}fit}$ one computes p($\chi^2 > \chi^2_{BF} | \nu$) the probability to exceed the χ^2_{BF}, assuming a χ^2-distribution with ν degree of freedom: ν =N. Data points - N. effective parameters.**

- **But, the number of effective parameters, for correlated parameters and/or for a prior-informed analysis, is smaller than the total number of free parameters.**

Interpreting Reduced Chi-Square

- $(\chi^2_{red} \approx 1)$: The model is a good fit for the data. (Residuals are of the same order as the uncertainties.)
- $(\chi^2_{red} > 1)$: The model might not fit well.
- $(\chi^2_{red} \gg 1)$: Indicates systematic errors, underestimation of uncertainties, or a poor model.
- $(\chi^2_{red} < 1)$: The model may be overfitting or uncertainties might be overestimated.

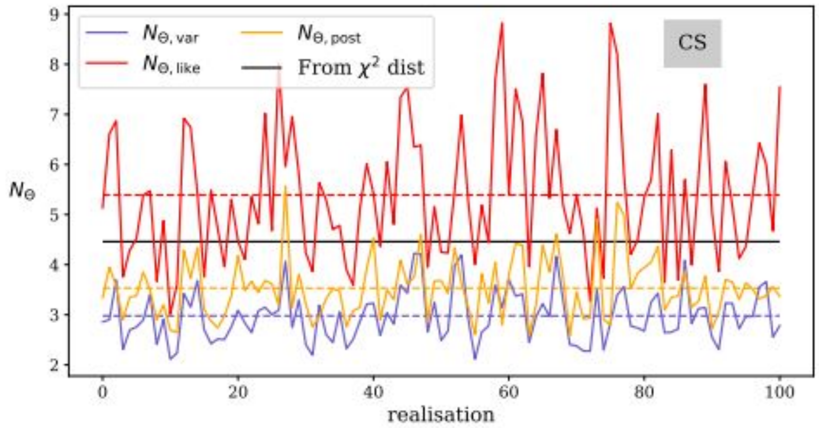
EFFECTIVE NUMBER OF D.O.F.

Effective number of constrained parameters:

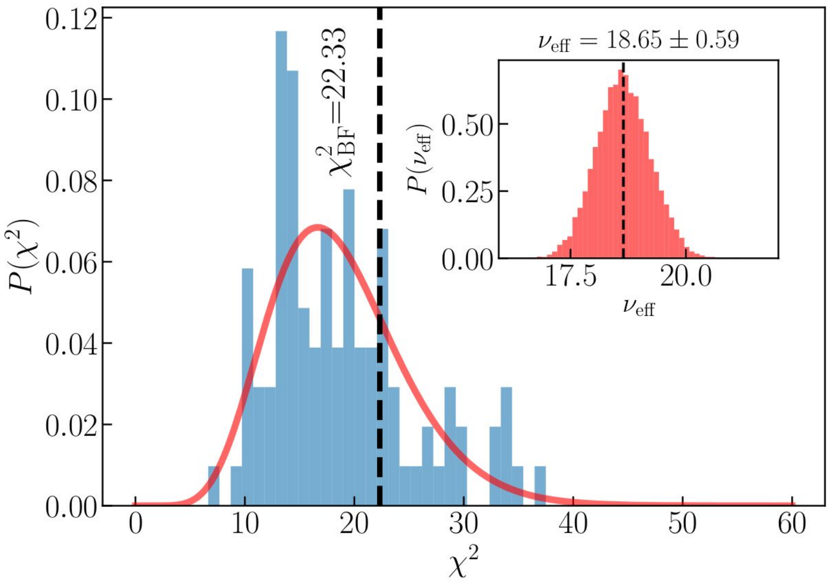
$$N_{\Theta,\mathrm{var}} = 2\left[\left\langle\left(\chi^2\right)^2\right\rangle_{\mathrm{Pr}(\boldsymbol{\Theta}|\boldsymbol{d})} - \left\langle\chi^2\right\rangle^2_{\mathrm{Pr}(\boldsymbol{\Theta}|\boldsymbol{d})}\right];$$

$$N_{\Theta,\mathrm{like}} = \left\langle\chi^2\right\rangle_{\mathrm{Pr}(\boldsymbol{\Theta}|\boldsymbol{d})} - \chi^2_{\mathrm{min}};$$

$$N_{\Theta,\mathrm{post}} = \left\langle\chi^2\right\rangle_{\mathrm{Pr}(\boldsymbol{\Theta}|\boldsymbol{d})} - \chi^2\left(\mathrm{argmax}\left[\mathrm{Pr}(\boldsymbol{\Theta}|\boldsymbol{d})\right]\right),$$



From Joachimi et al 2021



Distribution of the best-fit χ^2 values recovered from 100 mock data realizations generated from the best-fit model of the data. The red histogram in the inset plot shows the posterior distribution for the effective number of degrees of freedom obtained by fitting a χ^2 to the histogram. (DES Collaboration 20)

GOODNESS OF FIT: PPD

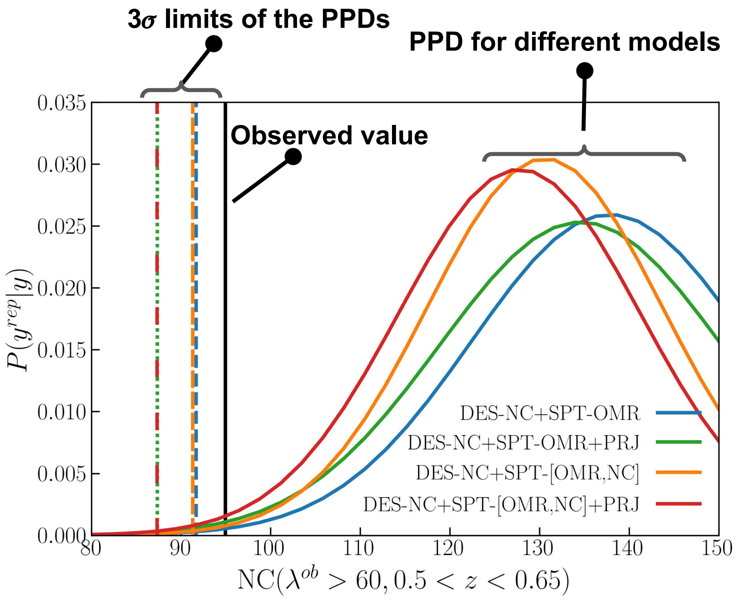
A more rigorous way to assess the goodness of fit which account for both the data and model uncertainty rely on the Posterior Predictive Distribution:

Parameter posteriors

$$P(y^{\mathrm{rep}}|y) = \int d\theta P(y^{\mathrm{rep}}|\theta) P(\theta|y)$$

Likelihood

The method consists of drawing simulated values from the posterior predictive distribution of replicated data and comparing these mock samples to the real data to assess their likelihood to be observed (see e.g. Doux+2021)



PPD for the observed cluster count in the highest λ/z bin of the DES Y1 data for 4 different model (Costanzi+21)

GOODNESS OF FIT: PTE

A common way to quantify the level of tension between y^{rep} and y^{obs} is to use a test statistic, $T(y, \vartheta)$, that can be computed for both y^{rep} and y^{obs}, and which may be a function of the parameters ϑ. A p-value – Probability To Exceed – can then be associated with the comparison between y^{rep} and y^{obs} via:

$$p = P(T(y^{rep}, \vartheta) > T(y^{obs}, \vartheta))$$

In other words, p is the probability of getting a higher test statistic for PPD realizations than $T(d\text{obs}, \Theta)$ by random chance. A very low p-value (say less that 0.01) would then be indicative that $d$obs was unlikely, while a high p-value (say greater than 0.99) could be indicative of a problem in the model, such as an overestimate of the noise covariance.

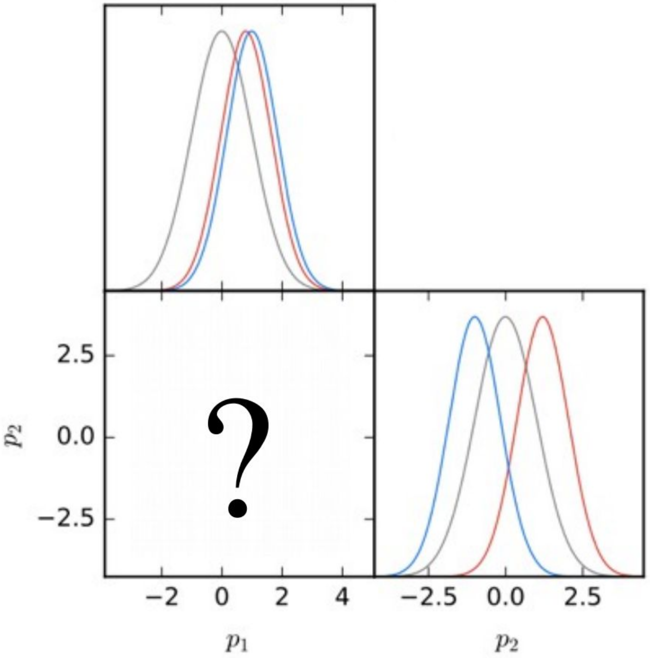
To assigning a p-value to the results of a PPD test requires a choice of test statistic. For high-dimensional data—in particular Gaussian data—a common choice is χ 2 , i.e.

$$T(d, \Theta) = (d - \mu(\Theta))^T C^{-1} (d - \mu(\Theta))$$

TENSION BETWEEN DATA SETS

Asses the level of tension (or agreement) between posteriors derived from different data sets might not be trivial in a multi-dimensional parameter space.

E.g. 1d marginalized posteriors which seem to be in agreement ...

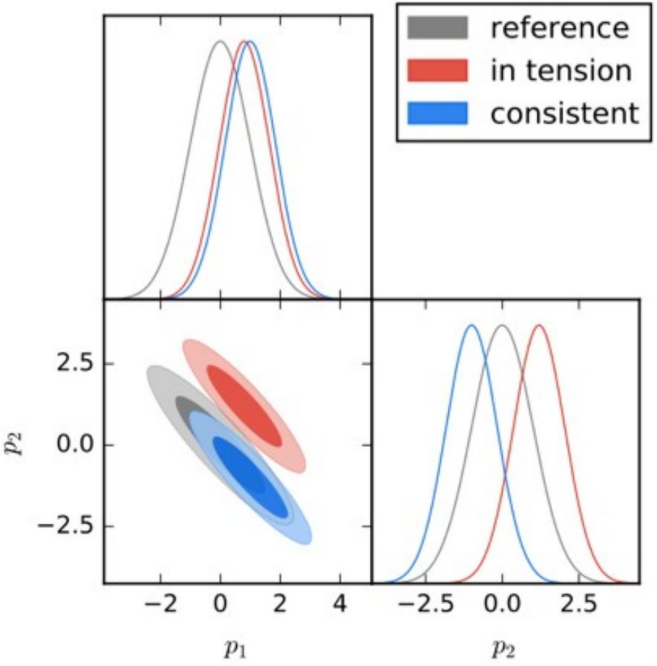


Credit A. Saro

TENSION BETWEEN DATA SETS

Asses the level of tension (or agreement) between posteriors derived from different data sets might not be trivial in a multi-dimensional parameter space.

E.g. 1d marginalized posteriors which seem to be in agreement ...



Credit A. Saro

… might hide tensions in higher dimension space due to “projection effects”

TENSION METRICS

There is no a unique "metric" to assess the level of tension/agreement between data sets, and there exist a number of technique which can be roughly splitted in:

- **Evidence-based methods** seek to answer the question:

Given hypothesis H_1: 'The assumed model is capable of generating the data observed by both experiments', and hypothesis H_2: 'The assumed model is not capable of generating the data observed by both experiments', which hypothesis is preferred by the data under the assumed model'?

- **Parameter-space methods** seek to answer the question:

What is the statistical significance of the differences between the posteriors for experiments A and B, within the parameter space analyzed by both experiments?

(Lemos+2020; see also e.g. Grandis+16, Charnock+17, Raveri+20)

Require the computation of the evidence:

$$P(\vec{d}) = \int \mathrm{d}\theta \mathcal{L}(\vec{d}|\vec{\theta})P(\vec{\theta})$$

In general can be computed directly from the parameter posteriors.
Require good sampling of the tails of the distributions

See also https://arxiv.org/pdf/2511.04661 for an application to data

TENSION METRICS

Bayes Ratio

The Bayes ratio R is defined for independent datasets A and B and for their combination AB as [210]:

$$R \equiv \frac{\mathcal{Z}_{AB}}{\mathcal{Z}_A \mathcal{Z}_B}, \qquad \text{(E1)}$$

where

$$\mathcal{Z}_D \equiv P(\mathbf{D}|\mathcal{M}) = \int d\mathbf{\Theta}\, \mathcal{L}(\mathbf{D}|\mathbf{\Theta}, \mathcal{M})\pi(\mathbf{\Theta}|\mathcal{M}). \qquad \text{(E2)}$$

In that expression, z_D is the Bayesian Evidence, L is the likelihood of observing the data given model M and parameter values Θ, and π is the prior probability of those parameters given the model. R can be viewed as a hypothesis test assessing the odds of both datasets being described with a single set of parameters (Z_{AB}) as opposed to two independent sets of parameters ($Z_A Z_B$)

Smaller values of R indicate stronger evidence of tension between measurements from datasets A and B

Jeffrey's scale	
ln R < -2.3	Strong Tension (10:1 odds)
-2.3< ln R <-1.2	Substantial tension (3:1 odds)
ln R>-1.2	Agreement

Caveat: the value of ln R depends strongly on the choice of parameter prior ranges

TENSION METRICS

Parameter difference technique:

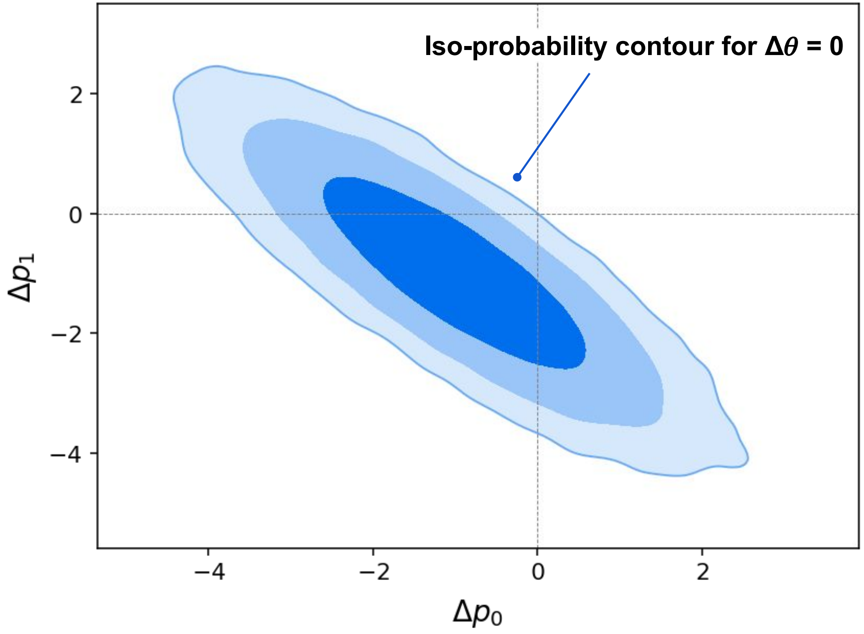
i) Compute the parameter difference probability distribution:

$$\mathcal{P}(\Delta\theta) = \int_{V_p} \mathcal{P}_A(\theta)\mathcal{P}_B(\theta - \Delta\theta)d\theta$$

ii) Determine posterior mass above the iso-probability contour for no shift, $\Delta\theta$ = 0

$$\Delta = \int_{\mathcal{P}(\Delta\theta) > \mathcal{P}(0)} \mathcal{P}(\Delta\theta)\, d\Delta\theta$$

The advantage of this technique is that it can be readily computed directly from the MCMC chains of experiment A and B



Toy model for a two parameter difference distribution. Credit M. Raveri

MODEL SELECTION

To determine which model is preferred by a given data set a simple comparison of χ^2s might not be sufficient (e.g. if the two models have a different number of parameters, or different priors)
Two widely used techniques for model selection are the evidence ratio and deviance information criterion:

Bayes Evidence Ratio:

$$\frac{P(M_1|\vec{d})}{P(M_2|\vec{d})} = \frac{P(\vec{d}|M_1)}{P(\vec{d}|M_2)} \frac{P(M_1)}{P(M_2)}$$

The evidence is larger for a model if more of its parameter space is likely and smaller for a model with large areas in its parameter space having low likelihood values, even if the likelihood function is sharply peaked.

Deviance Information Criterion:

$$\mathrm{DIC}(M) = 2\langle\chi^2\rangle_M - \chi^2_{\mathrm{MaxP}}(M)$$

The model with the lower DIC either fits better the data - lower $\langle\chi^2\rangle$- or has a lower level of complexity - lower$\langle\chi^2\rangle$ -χ^2_{MaxP}. It can be easily computed directly from the parameter posteriors

Check here for other methods

Comments on the prior dependence of the DESI results (https://arxiv.org/pdf/2407.06586)

uniform priors $\mathcal{U} = \frac{1}{b-a} = \frac{1}{d}$

$$B_{21} = \frac{\mathcal{P}_2 \int \mathcal{L}_2}{\mathcal{P}_1 \int \mathcal{L}_1} = \frac{\mathcal{P}_2}{\mathcal{P}_1}\Gamma = \frac{(d_{\nu_1} d_{\nu_2} d_{\nu_3} \ldots)_1}{(d_{\nu_1} d_{\nu_2} d_{\nu_3} \ldots)_2}\Gamma$$

We can change the prior range:

$$d^* = \alpha(b-a) = \alpha d$$

$$\log B^*_{21} = \log\left[\frac{(\alpha_{\nu_1}\alpha_{\nu_2}\alpha_{\nu_3}\ldots)_1}{(\alpha_{\nu_1}\alpha_{\nu_2}\alpha_{\nu_3}\ldots)_2}\right] + \log B_{21}$$

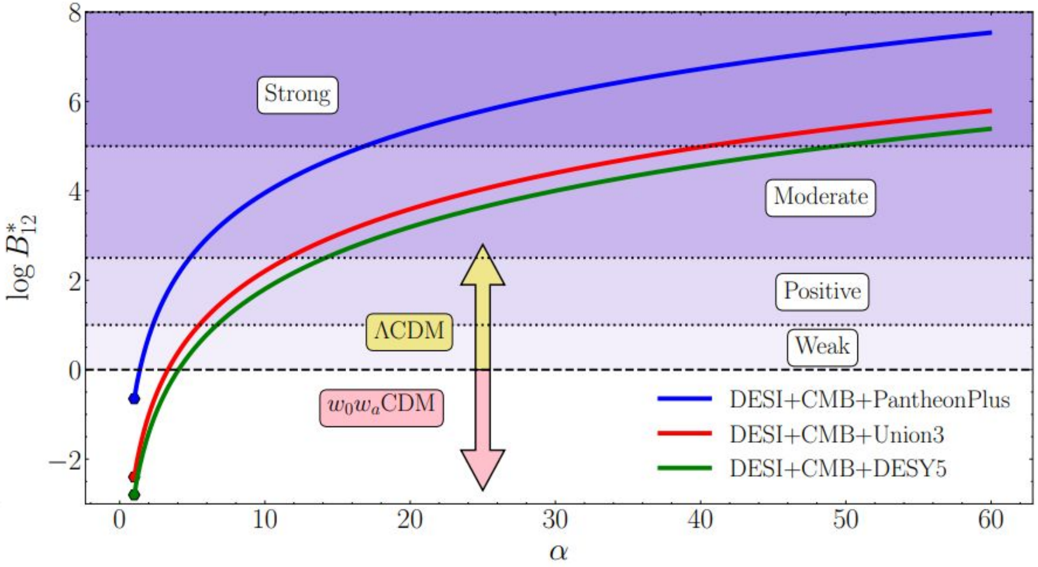
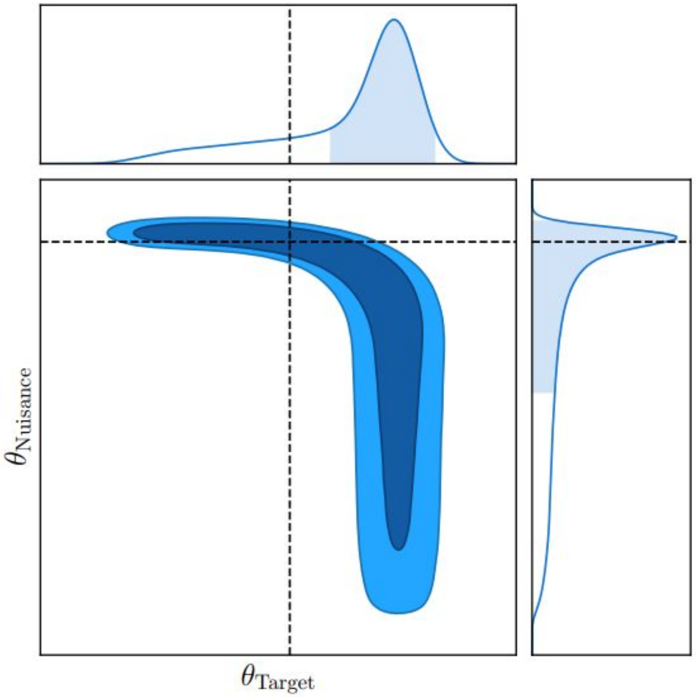


FIG. 1. Distribution of $\log B^*_{12}$ for α in the range [1, 60]. The colored regions indicate the strength of evidence for ΛCDM against w_0w_aCDM using Jeffreys' scale. The filled hexagon points at the start of the curves represent the original values from the DESI paper.

PRIOR VOLUME EFFECTS

The high-dimensionality of parameter spaces reduces the interpretability of posteriors to their one- and two-dimensional marginal distributions, when more information is available in the full dimensional distributions.



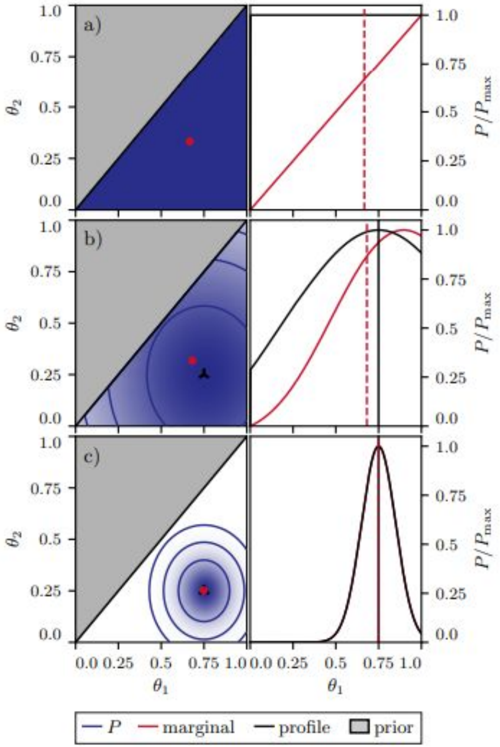
E.g.: Assume a highly-non Gaussian posterior. If we only have access to the 1D marginalized posterior of θ_{Target}, we would hardly estimate the fiducial value with our point estimators (e.g. using the BF point, or the mean/median of the distribution). In other words, even if the posterior in the entire parameter space is centered around the correct parameters, the posterior marginalized over the nuisance parameters can be off of the correct target parameters. This bias error is called the **prior volume or projection effect.**

See e.g.: https://arxiv.org/pdf/2405.00261

PRIOR VOLUME EFFECTS

The high-dimensionality of parameter spaces reduces the interpretability of posteriors to their one- and two-dimensional marginal distributions, when more information is available in the full dimensional distributions.

Difference between likelihood profile and marginal distribution for increasingly tighter posterior distributions



Another example is a poorly constrained, prior-limited parameters, which are projected over significant anisotropic volumes.

→ Posterior profiles do not suffer from projection effects as they are essentially insensitive to the volume of the parameter space. Using a simple metaphor, profiling can be thought of as observing the outline of the posterior landscape, whereas marginalization can be seen as measuring its column density.

https://arxiv.org/pdf/2409.09101