Bayesian Analysis

Weeks 5+6

Introduction to Bayesian Analysis

- This is a huge subject with potential for year-long course.

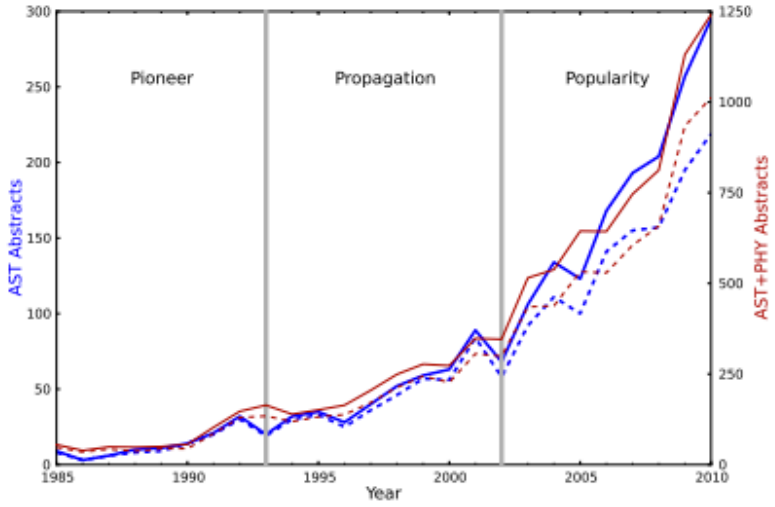


Fig. 1 Simple bibliometrics measuring the growing use of Bayesian methods in astronomy and physics, based on queries of the NASA ADS database in October 2011. Thick (blue) curves (against the left axis) are from queries of the astronomy database; thin (red) curves (against the right axis) are from joint queries of the astronomy and physics databases. For each case the dashed lower curve indicates the number of papers each year that include "Bayes" or "Bayesian" in the title or abstract. The upper curve is based on the same query, but also counting papers that use characteristically Bayesian terminology in the abstract (e.g., the phrase "posterior distribution" or the acronym "MCMC"); it is meant to capture Bayesian usage in areas where the methods are well-established, with the "Bayesian" appellation no longer deemed necessary or notable.

Tom Loredo arxiv:1208.3036

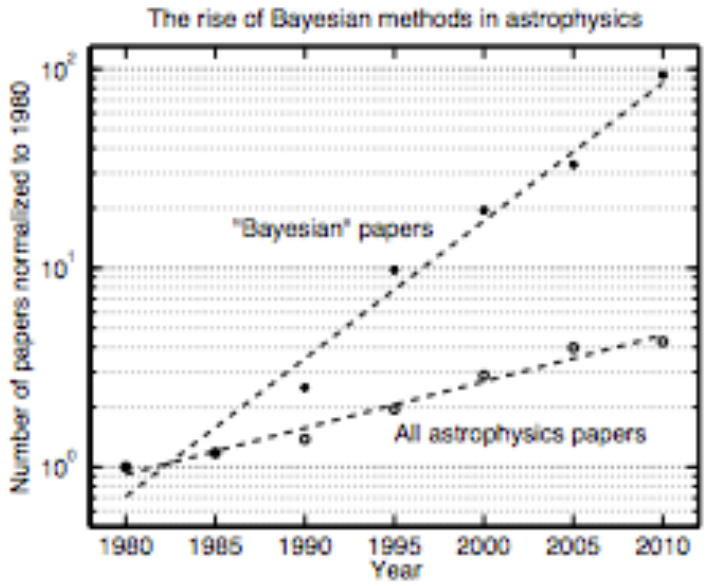
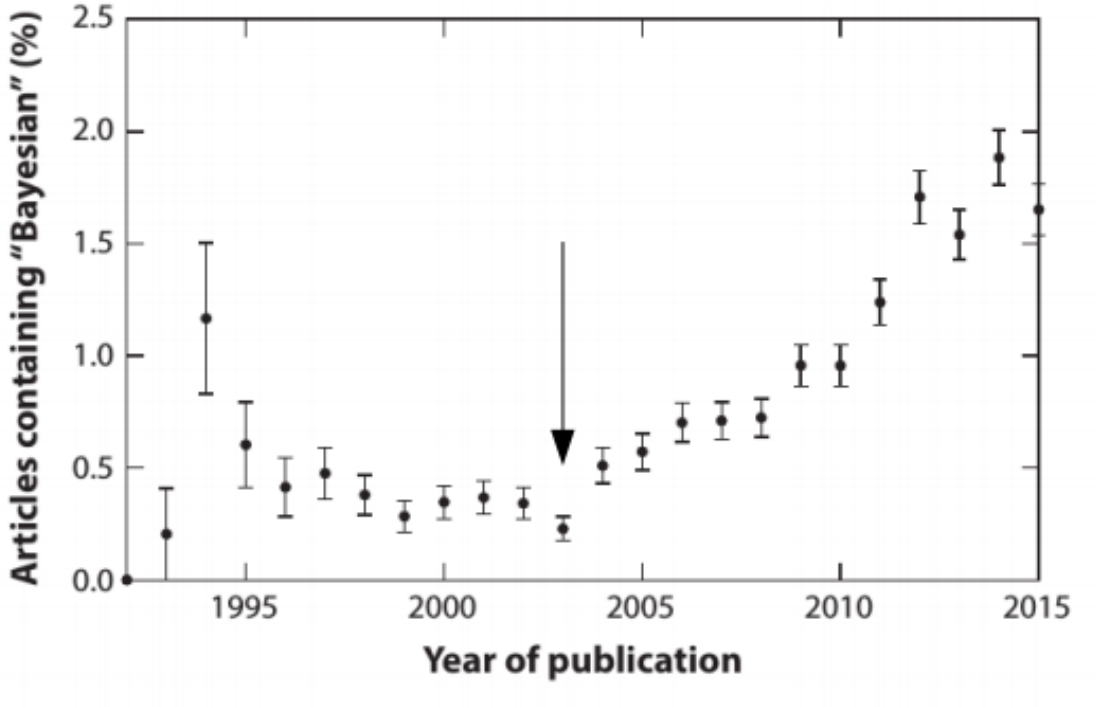


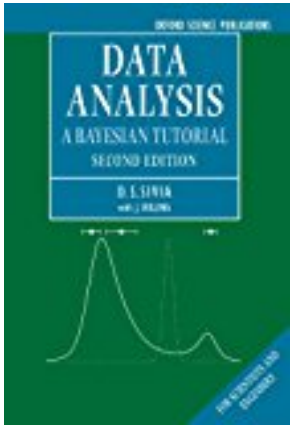
Fig. 3 Number of articles in astronomy and cosmology with "Bayesian" in the title, as a function of publication year (upper data points) and total number of articles (lower data points) as a function of publication year. Numbers are normalized to 1980 levels for each data series. The number of Bayesian papers doubles every 4.3 years, while the total number of papers doubles "only" every 12.6 years. At the present rate, by 2060 all papers on the archive will be Bayesian. (source: NASA/ADS).

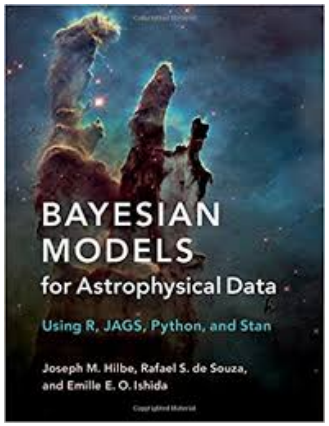
Bayesian Literature in Astronomy

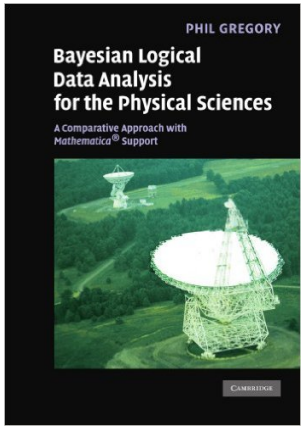


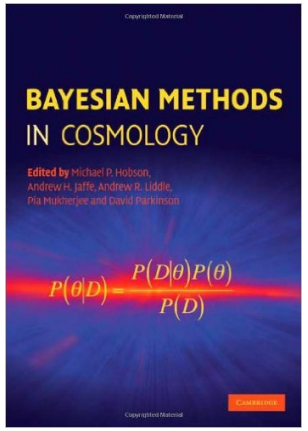
Books on Bayesian Analysis (Physicist Pov)

- Data Analysis – A Bayesian tutorial by D.S. Sivia and John Skilling
- Bayesian Logical Data Analysis for Physical Sciences
- Online book by E.T. Jaynes (Probability Theory, Logic of Science)
- Bayesian Methods in Cosmology by Michael Hobson et al









Arxiv resources on Bayesian Analysis

- https://arxiv.org/abs/0803.4089 Roberto Trotta
- https://arxiv.org/abs/1701.01467 Trotta (Bayesian methods in Cosmology)
- https://arxiv.org/abs/1301.1273 Louis Lyons (particle physics)
- https://arxiv.org/abs/1302.1721 Andrew Liddle
- https://arxiv.org/abs/0911.3105 Licia Verde (Statistical Methods in Cosmology)
- https://arxiv.org/abs/1208.3036 Tom Loredo (Bayesian Astrostatistics)
- https://arxiv.org/abs/1012.3589 Glenn Cowan (particle physics)
- https://arxiv.org/abs/1411.5018 J Vanderplas (Practical Introduction) (Easy to read)
- https://arxiv.org/abs/1809.02293 Eric Thrane & Colm Talbot(for GW astronomy)

Also lectures on Bayesian statistics in PSU summer school
Jake Van der Plas blog article

Introduction to Bayesian Statistics

- First introduced by Rev. Thomas Bayes (1702-1761) who wrote an article on how to combine an initial belief with new data to arrive at an improved belief

- Probability statements can be made for model parameters and models themselves.

- Both Classical (and frequentist) statistics use data likelihood. In frequentist statistics, likelihood function is used to find model parameters with the highest data likelihood. But likelihood function cannot be interpreted as PDF for the model parameters.

- Bayesian statistics extends the concept of data likelihood function by adding extra (or *prior*) information to the analysis and assigning pdfs to all model parameters and the models themselves.

Limitations of Max. Likelihood Estimates

Imagine you arrive at a bus stop (with no knowledge of the bus schedule)

and the next bus arrives *t* minutes later.

Q: What is the mean time τ between two successive buses , if the buses keep a regular schedule?

Ans: Wait time is distributed uniformly in the interval $0 \leq t \leq \tau$ and average wait time= $\tau/2$,

or $\tau = 2t$

Max Likelihood Estimates

Probability that you shall wait t minutes (likelihood of the data) is given by the uniform distribution

$$p(t|\tau) = \begin{cases} 1/\tau & \text{if } 0 \leq \mathrm{t} \leq \tau \\ 0 & \text{otherwise} \end{cases}$$

Because there is only a single data point, the data likelihood is simple equal to this probability. Maximum of p(tlτ) occurs for τ = t (or when τ is smallest)

Expectation value of τ or its median value diverges

Problem is solved when we introduce priors.

Essence of Bayesian idea

$$p(M|D) = \frac{p(D|M)p(M)}{p(D)}$$

Bayes theorem quantifies the rule for combining initial belief with new data to arrive at improved belief and "improved belief" is the product of "initial belief" and the probability that "initial belief" generated the observed data.

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

where model M includes k model parameters θ_p. (note that M and θ are used interchangeably)

- P(M, θ|D,I) is the *posterior* pdf for model M and some prior information *I*.
- P(DI M, θ) is the *likelihood* of the data *given* some model M and some fixed value of parameters θ
- P(M, θII) is the a priori joint probability for model M and its parameters θ in the absence of any of the data used to compute the likelihood and is often simply called the *prior*

Prior can be expanded as follows $p(M,\theta|I) = p(\theta|M,I)p(M|I)$

Usually for parameter estimation problems, we only need to specify P(θIM,I)

- P(D/I) is *probability of the data* or the prior predictive probability for D.

Provides normalization for the posterior pdf and is not explicitly computed when estimating model parameters.

Instead P(M, θ|D,I) for a given model is renormalized so that its integral over all model parameters is unity. Usually normalization is arbitrary.

One exception is model comparison where correct normalization of numerator in Bayes's formulae is important.

Why is Bayesian Interpretation Controversial

- P(M, θID,I) is not a probability in the strict sense (compared to the likelihood in frequentist statistics) . We accept that P(M, θID,I) corresponds to the state of our knowledge (or degree of belief) about a model and its parameters given data D and prior information *I* . This introduces the notion of probability for models and model parameters.

Bus-Stop Example

- Assume prior is proportion to 1/ τ. Posterior probability density function is given by :

$$p(\tau|t,I) = \begin{cases} t/\tau^2 & \text{if } \tau \geq \text{t} \\ 0 & \text{otherwise} \end{cases}$$

Note that the extra t is added so that integral of posterior is unity.

$$\int_t^\infty p(\tau|t,I)d\tau = \int_t^\infty \frac{C}{\tau^2}d\tau = 1 \longrightarrow \quad \text{C = t}$$

Median value of τ for $p(\tau|t, I)$ is given by 2t (in agreement with expectation)

Also p% quantiles are equal to (1- t/ τ_p)

Summary of Bayesian Statistical Inference

Notation : Replace M(θ) with M whenever the absence of explicit dependence on θ is not confusing.

- Construction of the Data likelihood P (D I M)
- Choice of the prior incorporating previous information that might exist, but is not used in computing the likelihood. P(θIM,I)
- Determination of posterior pdf using Bayes theorem. P (M I D)

This step is computationally intensive for multi-dimensional posterior problems. Also P(D I I) is not explicitly stated because P(M I D) can be properly normalized by renormalizing the product P(D I M) P(M)

- Best model parameters are found by maximizing P (M I D) which yield maximum *aposteriori estimate.* This *point estimate* is the natural analog to maximum likelihood estimation (MLE) from classical statistics.

Alternately calculate the Bayesian posterior means as follows :

$$\overline{\theta} = \int \theta p(\theta|D) d\theta$$

For a vector of parameters θ (i.e. one particular value) is obtained from $P(M, \theta|D, I)$ by the process of *marginalization* , i.e. by integrating $P(M, \theta|D, I)$ over all the other unknown parameters and then renormalization

$$\int p(\theta|D) d\theta = 1$$

Although there is no natural way to marginalize over nuisance parameters in frequentist statistic,
However , profile likelihood is used (particle physics literature)

$$\mathscr{L}(\theta_1) \equiv \max_{\theta_2,\ldots,\theta_N} \mathscr{L}(\theta),$$

- Uncertainty in parameter estimates is obtained via *credible regions* (Bayesian counterpart to frequentist confidence region). Obtained by integrating the posterior. Various numerical methods can be used to simulate samples from the posterior.

- Hypothesis testing as needed to make conclusions about model or parameter estimates.

Bayesian approach can be thought of as formalizing the process of continually refining our state of the knowledge about the world, beginning with no data and then updating that by multiplying by the likelihood once the data is observed to obtain the posterior.

Convergence of Posterior with More data

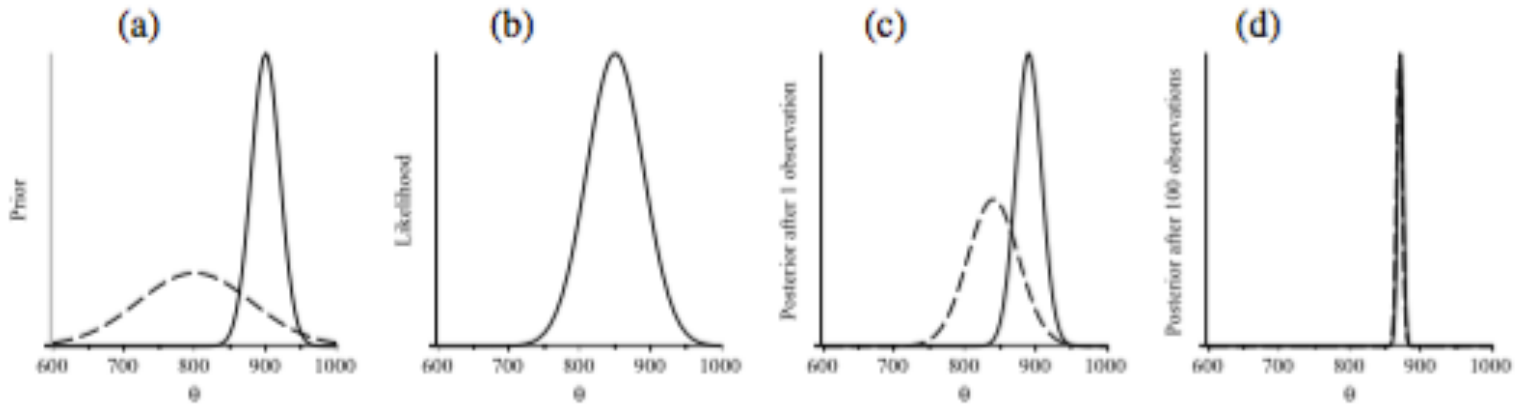


Fig. 4 Converging views in Bayesian inference. Two scientists having different prior believes $p(\theta)$ about the value of a quantity θ (panel (a), the two curves representing two different priors) observe one datum with likelihood $\mathscr{L}(\theta)$ (panel (b)), after which their posteriors $p(\theta|d)$ (panel (c), obtained via Bayes Theorem, Eq. (4)) represent their updated states of knowledge on the parameter. This posterior then becomes the prior for the next observation. After observing 100 data points, the two posteriors have become essentially indistinguishable (d).

Bayesian Priors

- Prior incorporates all other knowledge that might exist, but is used when computing the likelihood. Data may chronologically proceed the information in the prior

Examples of priors include knowledge from prior measurements of the same type as data at hand or different measurements that constrain the same quantity whose posterior pdf we are trying to constrain with new data. We may know the mass of an elementary particle from previous measurements with uncertainty.. Priors that incorporate information based on other measurements are called *informative priors*

Priors Assigned by Formal Rules

- When no information is available except for the data been analyzed then such priors are called non-informative priors.

  Consider a *flat prior* p(θ) $\propto$ constant since $\int p(\theta|I)d\theta = \infty$ -> this is not a pdf

  This is considered an improper prior. Not a problem as long as the resulting posterior is a well-defined pdf.

Principle of indifference : A set of mutually exclusive possibilities need to be assigned equal probabilities. For a fair die prior = 1/6

Principle of consistency: Prior for a location parameter should not change with translation of coordinate system and yields a flat prior. Prior for scale parameter should not depend on choice of units eg if scale parameter is **σ** and rescaling by a constant factor gives $p(\sigma|I)d\sigma = p(a\sigma|I)d(a\sigma)$

Solution to this is

$$P(\sigma|I) = p(a\sigma|I)d(a\sigma)$$

Principle of Maximum Entropy

- Entropy measures the information content of a pdf. Entropy for a pdf defined by N discrete values is given by :

$$S = -\sum_{i=1}^{N} p_i \ln(p_i)$$

This is also called Shannon's entropy. This can be generalized to a continuous distribution (Sivia 2006)

$$S = -\int_{-\infty}^{\infty} p(x) \ln\left(\frac{p(x)}{m(x)}\right) dx$$

where the ``Lebesgue measure" m(x) ensures that entropy is invariant under change of variables. (Sivia 06)

Principle of Maximum Entropy (contd)

- By maximizng entropy over a suitable set of pdfs, we find the distribution that is least informative (given the constraints)

- Power of the principle comes from a straightforward ability to add additional information about the prior distribution, such as the mean value and the variance.

Example : Consider a 6-faced die where we need to assign 6 prior probabilities. From principle of indifference prior probability is 1/6. But suppose we know the mean value (for a large number of rolls) then we need to adjust the prior probabilities to be consistent with this information.

$$\sum_{i=1}^{6} i p_i = \mu$$

$$\sum_{i=1}^{6} p_i = 1$$

Maximize

$$Q = -\sum_{i=1}^{6} p_i \ln\left(\frac{p_i}{m_i}\right) + \lambda_0(1 - \sum_{i=1}^{6} p_i) + \lambda_1(\mu - \sum_{i=1}^{6} i p_i)$$

- where m_i is the values of p_i when no additional information is known (in this case 1/6). Solution is given by

$$p_i = m_i \exp(-1 - \lambda_0) \exp(i\lambda_1)$$

Generalizes to Poisson distribution if no of discrete events is infinite with mean μ

For the continuous case maximum entropy solution for the prior is given by :

$$p(\theta|\mu) = \frac{1}{\mu} \exp\left(-\frac{\theta}{\mu}\right)$$

obtained by assuming a flat distribution on m(θ) (when no constraint on mean value) and assuming we know the expectation value of

$$\mu = \int \theta p(\theta) d\theta$$

Kullback-Leibler Divergence

$$KL = \sum_i p_i \ln\left(\frac{p_i}{m_i}\right)$$ and analogously for continuous case

KL divergence is sometimes called KL distance between the two pdfs.
But it is NOT the metric distance since value is not the same when p(x) and m(x) are switched.

In Bayesian statistics KL divergence can be used to measure the information gain When moving the prior distribution to posterior distribution. KL divergence also used in information theory.

Conjugate Priors.

- If posterior probability has the same functional form as prior probability, these priors are called conjugate priors. and represent a convenient way for generalizing computations.

Likelihood	Conjugate Prior
Gaussian	Gaussian
Binomial	Beta distribution
Poisson	Gamma Distribution
Multinomial	Dirichlet priors

Bayesian Parameter Uncertainty Quantification

To obtain Bayesian *credible region* estimate, we find a and b such that

$$\int_{-\infty}^{a} f(\theta)d\theta = \int_{b}^{\infty} f(\theta)d\theta = \alpha/2$$

The probability that the true values of θ is in the interval (a,b) is equal to 1-α
In analogy with classical confidence interval and (a,b) is called a 1-α Posterior interval

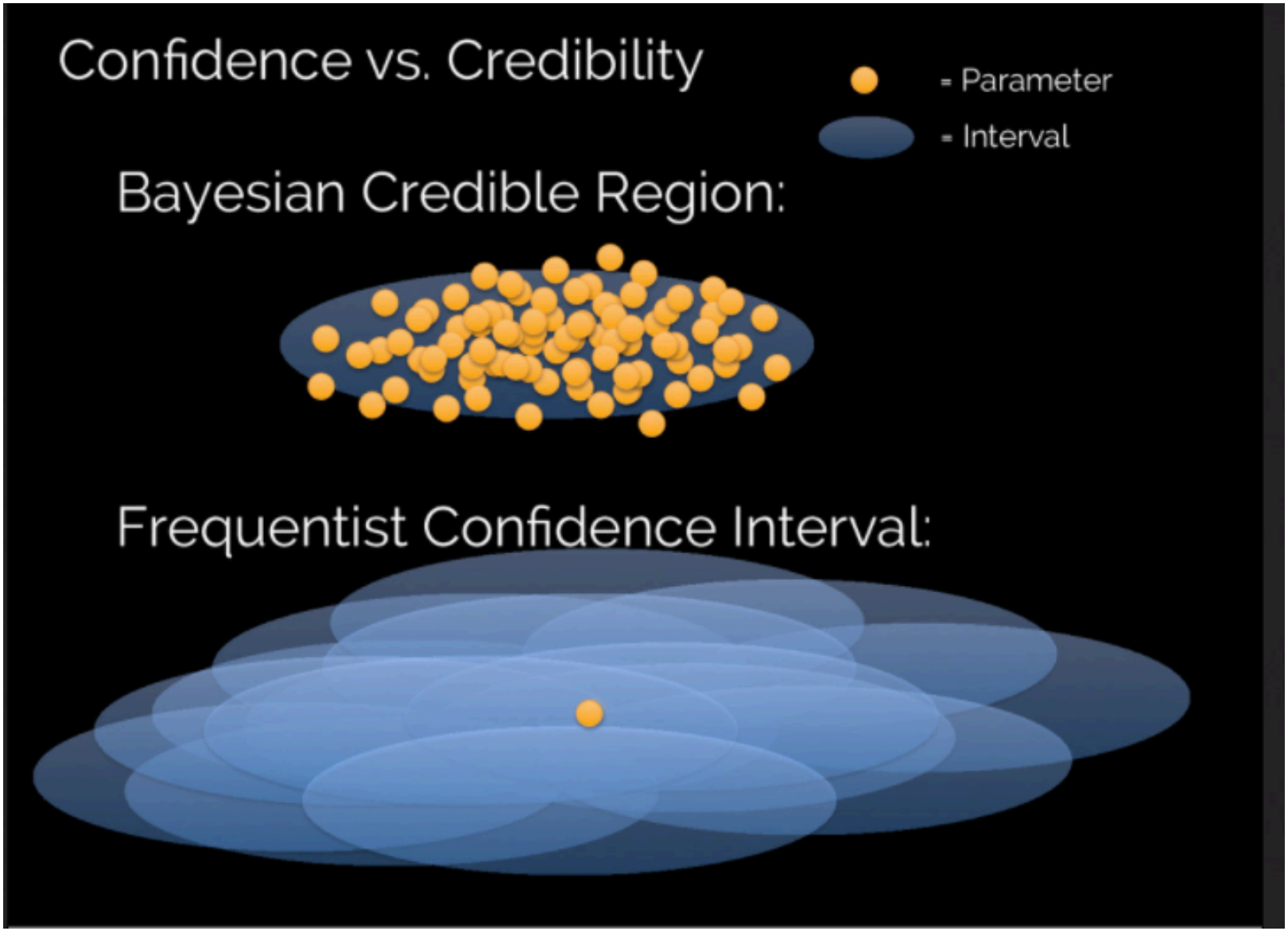
Difference between Confidence/Credible Regions

- Bayesianism : Probabilistic statement about *model parameters* given a *fixed credible region.*

Given our observed data, there is a 95% probability that the true value of θ lies within the credible region.

- Frequentism : Probabilistic statement about a recipe *for generating confidence intervals* given a *fixed model parameter.*

If our experiment is repeated many times, in 95% of these cases, the computed confidence interval will contain the true θ


Confidence vs. Credibility
= Parameter
= Interval
Bayesian Credible Region:
Frequentist Confidence Interval:

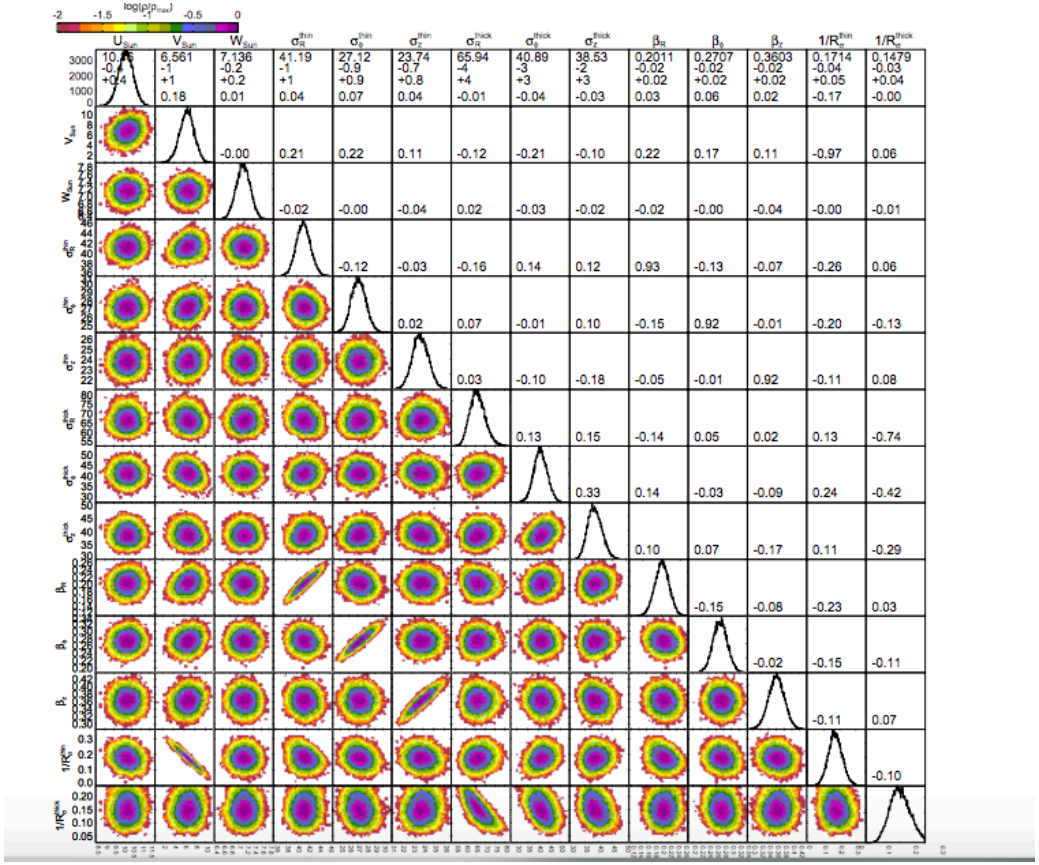
Marginalization of Parameters

Consider posterior pdf P (M ,θ_1, θ_2,..... θ_k) . Suppose we are only interested in θ_1 . In order to obtain the posterior pdf for θ_1 we integrate over all the uninteresting (or nuisance) parameters. This integration procedure is called the process of marginalization and the resulting pdf is called *marginal posterior pdf*

$$p(M, \theta_1 | D) = \int p(M, \theta_1, \theta_2, ...\theta_k | D) d\theta_2 ....d\theta_k$$

No analog of marginalization exists in frequentist statistics.

Examples of Marginalized Posteriors



arXiv:1405.7435

Bayesian Model Selection

- To find out which of two models say M_1 and M_2 are supported by data we compare their posterior probabilities via the *odds ratio* in favor of model M_2 over model M_1 (sometimes also called posterior odds)

$$O_{21} = \frac{P(M_2|D)}{P(M_1|D)}$$

Where the posterior probability that the model *M* is correct given data *D* is given by

$$P(M|D) = \frac{P(D|M)P(M)}{P(D)}$$

where

$$E(M) \equiv p(D|M) = \int p(D|M,\theta)p(\theta|M)d\theta$$

is called the evidence or marginal likelihood or Bayesian evidence. It quantifies the Probability that the data D would be observed *if* the model M were the correct model

Odds Ratio becomes :

$$O_{21} = \frac{E(M_2)p(M_2)}{E(M_1)p(M_1)} = B_{21}\frac{p(M_2)}{p(M_1)}$$

where

$$B_{21} = \frac{\int p(D|M_2,\theta_2)p(\theta_2|M_2)d\theta_2}{\int p(D|M_1,\theta_1)p(\theta_1|M_1)d\theta_1}$$

Bayes Factor

Interpretation of Odds Ratio

- Jeffreys Scale used as a qualitative guide to decide between model 2 and model 1

K	dHart	bits	Strength of evidence
$< 10^0$	< 0		negative (supports M_2)
10^0 to $10^{1/2}$	0 to 5	0 to 1.6	barely worth mentioning
$10^{1/2}$ to 10^1	5 to 10	1.6 to 3.3	substantial
10^1 to $10^{3/2}$	10 to 15	3.3 to 5.0	strong
$10^{3/2}$ to 10^2	15 to 20	5.0 to 6.6	very strong
$> 10^2$	> 20	> 6.6	decisive

K = Bayes Factor

wikipedia (note that model M_1 is in the numerator in wikipedia definition. of Bayes factor)

Open Questions/Limitations of Bayesian Analysis

- Criticism of Bayesian Model selection by Efsthathiou (for cosmology) (arXiv:0802.3185) and Bob Cousins (particle physics) (arXiv: 0807.1330) , where not enough care is given to selection of models and priors and posterior model probabilities maybe function of unjustified assumptions.
- How to deal with Lindley's paradox (where Bayesian and frequentist hypothesis testing lead to contradictory answers). (See arXiv:1310.3791)
- How to assess the completeness of a set of known models? Is there a principled way of constructing an *absolute* scale for model performance in a Bayesian context? (see arxiv:1511.02363 for some research on this)
- Is Bayesian model averaging (model independent estimation of parameters) useful ?
- Is there such a thing as a correct prior?)

- Many posteriors used in astronomical literature are improprer arXiv:1712.03549