Curso de Cosmologia Pós – 2023/2

Part VII:
Bayesian Statistics

[R. Trotta 1701.01467]

[Amendola & Tsujikawa cap. 13]

Miguel Quartin
Instituto de Física, UFRJ
Astrofísica, Relativ. e Cosmologia (ARCOS)

Topics

- Short review of probability theory
- Bayes' Theorem
- The Likelihood method
- Model Selection
- [optional] Fisher Matrix

Probabilities

- Classical interpretation of probability: infinite realization limit of relative frequencies
  - Probability: "the number of times the event occurs over the total number of trials, in the limit of an infinite series of equiprobable repetitions."
- Let's define 2 random (stochastic) variables x and y (e.g. numbers on a die roll).
  - $p(X)$ is the probability of getting the result $x = X$
  - $p(X, Y)$ or $p(X \cap Y)$→ prob of getting results $x = X$ AND $y = Y$
  - $p(X \mid Y)$ or $p(X ; Y)$ → prob of $x = X$ given the fact that $y = Y$
  - $p(X \cup Y)$ → prob of getting results $x = X$ OR $y = Y$

Probabilities (2)

- Some properties:
  - Joint probabilities are symmetric

$$p(A \cap B) = p(B \cap A) \qquad p(A, B) = p(B, A)$$

  - Joint prob of *independent* events

$$p(A, B) = p(B)p(A)$$

  - Joint prob of *dependent* events

$$p(A, B) = p(B)p(A|B)$$

  - Disjoint prob of *mutually exclusive* events

$$p(A, B) = p(A) + p(B)$$

    - In particular

$$p(A, \overline{A}) = 1 = p(A) + p(\overline{A})$$

Probabilities (3)

- Let's discuss the conditional probability property:

$$p(A, B) = p(B)p(A|B)$$

  - Suppose A refers to “person that studies physics” and B to “person that plays piano”
  - Suppose also that we know that: $p(B) = 1/100$ and $p(A, B) = 1/1000$
    - In other words, out of 1000 random people, 10 will play the piano and 1 will play the piano AND be a physicist
    - So, if someone plays piano, he has 1/10 chance of being also a physicist

$$p(A, B) = \frac{1}{1000} = p(B)p(A|B) = \frac{1}{100}\frac{1}{10}$$

Bayes' Theorem

- Note that

$$p(A|B) \neq p(B|A)$$

  - The probability of A given B is **not** the prob of B given A.
  - E.g.: The probability of winning the lottery given that you played twice in your life is **not** the same as the probability that you played twice in your life given that you won the lottery!
- From the symmetry of the joint probabilities we get

$$p(A,B) = p(B,A) \quad \rightarrow \quad p(B)p(A|B) = p(A)p(B|A)$$

- This is the Bayes Theorem of conditional probabilities

$$p(A|B) = \frac{p(A)p(B|A)}{p(B)}$$

Interpretation

- Classical "frequentist" interpretation of probability (infinite realization limit of relative frequencies) has limitations
  - It is *circular* → assumes that the repeated trials have same probability of outcomes
  - Cannot deal with unrepeatable situations [e.g. *(i)* probability I will die in a car accident; *(ii)* prob the Big-Bang happened the way it did]
    - "what is the probability that it rained in Manaus during D. Pedro II 43rd birthday?"
  - How to correct for finite realizations? How many realizations are needed for the frequencies to be approx. the probabilities? This approximation is to which % accuracy?

Interpretation (2)

- The Bayesian interpretation is based on Bayes' Theorem
  - Re-interpret the theorem not in terms of regular random variables but in terms of data (D) and theory (T)
  - Inverse statistical problem: what is the probability that theory T is correct given we measured the data D?

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

  - The "theory" might be a model (such as ΛCDM or DGP) of just the parameter values of an assumed model (such as Ω_{m0} and $\Omega_{\Lambda 0}$, assuming ΛCDM).

Interpretation (3)

- Bayesian analysis has some philosophical implications
  - The best theory will be the most probable theory
  - Bayesian analysis carry a mathematically precise formulation of Occam's Razor: “if 2 hypotheses are equally likely, the hypothesis with the fewest assumptions should be selected”.
    - Only strong reason before the 18th century to choose Copernicus' model over Ptolomy's
    - Karl Popper: we prefer simpler theories to more complex ones “because their empirical content is greater; and because they are better testable” → simple theories are more easily falsifiable

The Likelihood Method

- Let's define the data as a vector x and the parameters as a vector θ
- We write Bayes' theorem as
  - $P \rightarrow$ posterior probability
  - $p(\theta) \rightarrow$ prior probability
  - $L \rightarrow$ likelihood function
  - $g(x) \rightarrow$ the evidence $\rightarrow$ just a normalization factor for parameter estimation

$$P(\theta_j|x_i) = \frac{L(x_i|\theta_j)p(\theta_j)}{g(x_i)}$$

- We are often interested in the posterior

- In the literature the posterior is sometimes referred also as simply the "likelihood", but this is not technically correct
  - Although they may be the same function of the parameters!

The Likelihood Method (2)

- The posterior is a probability, so it has to be normalized to unity

$$\int P(\theta_j|x_i)\mathrm{d}^n\theta_j = 1 = \frac{\int L(x_i|\theta_j)p(\theta_j)\mathrm{d}^n\theta_j}{g(x_i)}$$

$$g(x_i) = \int L(x_i|\theta_j)p(\theta_j)\mathrm{d}^n\theta_j$$

- This integral is called the evidence
  - $g(x)$ does not depend on the parameters, so it is useless for parameter determination
  - But very useful to choose which is the best model

Prior and Prejudice

- Priors are inevitable in the likelihood (posterior) method
  - Frequentist don't like it → subjective prior knowledge
  - Bayesianists have to learn to love it → after all, we always know something before the analysis
    - E.g.: we can use $p(\Omega_{m0} < 0) = 0$ as it does not make sense to have negative matter density
    - We can add information from previous experiment. E.g. experiment A measured $h = 0.72 \pm 0.08$, so we can use, say, the Gaussian prior

$$p(h) = \frac{1}{2\pi(0.08)} \exp\left[-\frac{1}{2}\frac{(h-0.72)^2}{(0.08)^2}\right]$$

Prior and Prejudice

- You are free to choose your prior → but choice must be explicit
- You HAVE to choose a prior → $p(\theta) = 1$ *is* a particular prior, which under a parameter change will no longer be constant.
  - E.g.: $p(t) = 1 \neq p(z) = 1 \neq p(\log t) = 1$ …
  - E.g. 2: a measurement of $\Omega_{\Lambda 0}$ assumes the strong prior that the model is ΛCDM
- Priors may be subjective, but analysis is objective
- Priors are an advantage of Bayes → no inference can be made without assumptions
- Data can show that the priors were "wrong"

The Likelihood Method (3)

- If we are not interested in model selection we can neglect the function $g(x)$
  - The posterior P must then be normalized
  - The best-fit parameter values are the ones that maximize P

$$P(\theta_j^{\text{best}}|x_i) = P_{\max}$$

  - The $n\%$ confidence region $\mathbb{R}$ of the parameters is the region around the best fit for which

$$\int_R P(\theta_j|x_i)\mathrm{d}^n\theta_j = n/100$$

    - The confidence region in general is not symmetric

The Likelihood Method (4)

- If the likelihood (i.e. the posterior) has many parameters, it is interesting to know what information it has in each parameter (or each pair) independently of the others
  - We must do a weighted sum on the other parameters
    - This is referred to as marginalization over a parameter

$$P(\Omega_{m0}, \Omega_{\Lambda 0}) \equiv \int P(\Omega_{m0}, \Omega_{\Lambda 0}, C) \mathrm{d}C$$

$$P(\Omega_{m0}) \equiv \int P(\Omega_{m0}, \Omega_{\Lambda 0}) \mathrm{d}\Omega_{\Lambda 0}$$

The Likelihood Method (5)

- It is customary to use the following confidence regions: 68.3%, 95.4% and 99.73%. The reason is that for Gaussian posteriors, these are the 1, 2 and 3 standard deviations.
  - We therefore often refer to these regions, for simplicity, as just the 1σ, 2σ and 3σ regions
- "Highest Density Intervals" (HDI) are preferred over percentiles
- Note: HDI is the standard in cosmology, but not in e.g. medicine
  - Say, if for Ω_{m0} the best fit is 0.3 and the 68% confidence region (1σ) is [0.1, 0.4]

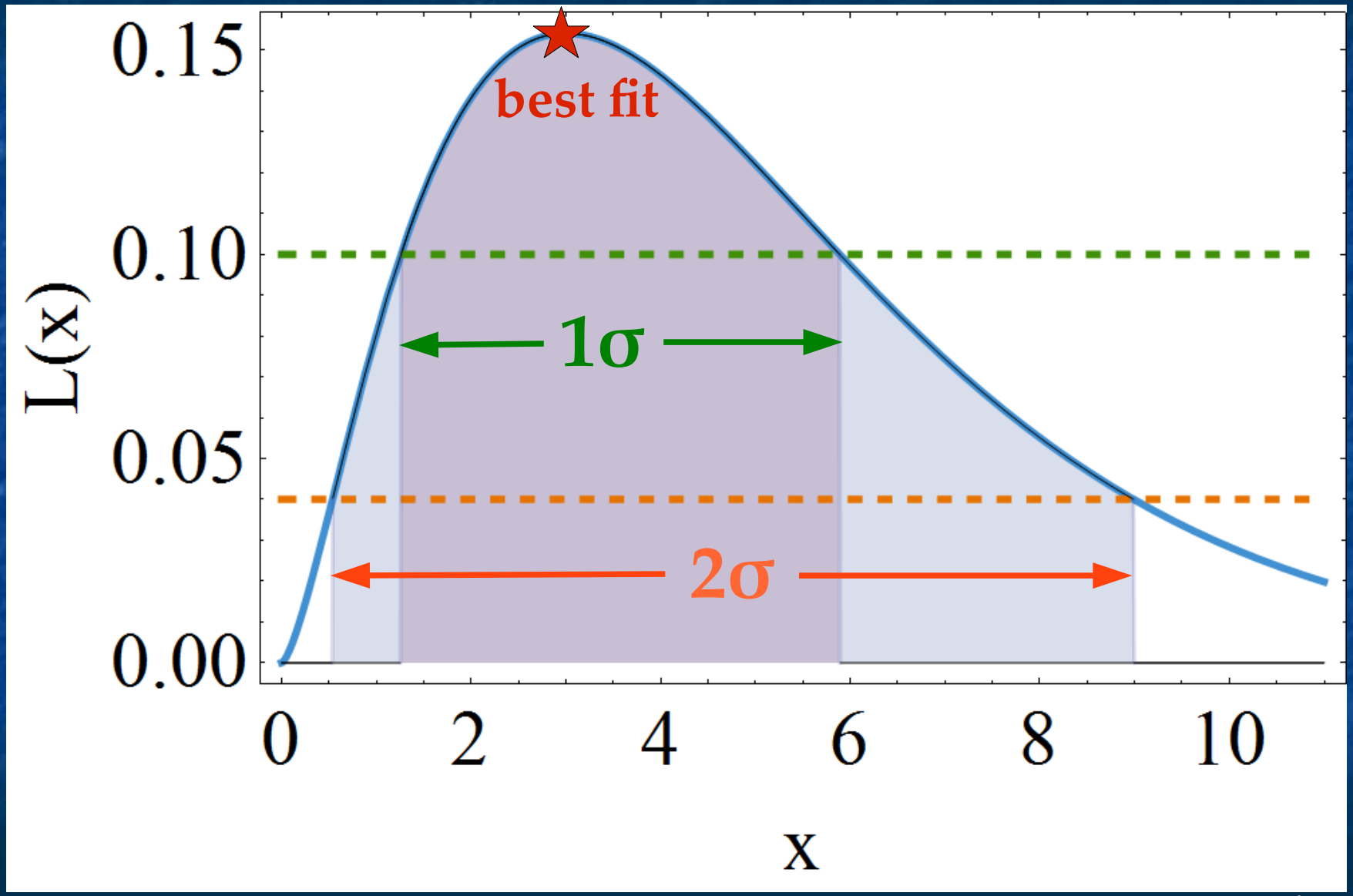
$$P_{\max}(\Omega_{m0}) = P(0.3) \qquad P(0.1) = P(0.4)$$

$$\int_{0.1}^{0.4} P(\Omega_{m0}) \mathrm{d}\Omega_{m0} = 0.683 \quad \rightarrow \quad \Omega_{m0} = 0.3^{+0.1}_{-0.2}$$

  - Note that here the 2σ region will in general not be

$$\Omega_{m0} \neq 0.3^{+0.2}_{-0.4} \ (95.4\%\mathrm{CI})$$

Highest Density Intervals



How to Build the Likelihood?

- The likelihood is a function of the data → functional form depends on the instrument used to collect the data
  - Usually instruments have (approximately) either Gaussian or Poisson noise.
  - It is common to assume by default a Gaussian noise
- If the likelihood is Gaussian in the data and the data are independent (uncorrelated errors) we have

$$L(\boldsymbol{x}|\boldsymbol{\theta}) = L_0 \prod_i \exp\left[-\frac{1}{2}\left(\frac{x_i^{\rm obs} - x_i^{\rm teo}(\boldsymbol{\theta})}{\sigma_i}\right)^2\right]$$

$$L(\boldsymbol{x}|\boldsymbol{\theta}) = L_0 \exp\left[-\frac{1}{2}\sum_i \left(\frac{x_i^{\rm obs} - x_i^{\rm teo}(\boldsymbol{\theta})}{\sigma_i}\right)^2\right]$$

Fisher Matrix

- In a nutshell → the Fisher Matrix method is an approximation for the computation of the posterior under the assumption that it is Gaussian in the parameters
  - Advantages:
    - very fast to compute (either analytically or numerically)
    - gives directly the (elliptical) confidence-level contours
  - Disadvantages:
    - gives wrong results when non-Gaussianity is strong
    - no intrinsic flags to warn you when non-Gaussianity is strong
    - numerical derivatives can be noisy
- For a 4-page quick-start guide, see: arXiv:0906.4123
- For more detail, see Amendola & Tsujikawa, Sect. 13.3

Fisher Matrix (2)

- We write the posterior as a multivariate gaussian

$$L \approx N \exp\left[-\frac{1}{2}\left(\theta_i - \hat{\theta}_i\right) F_{ij} \left(\theta_j - \hat{\theta}_j\right)\right]$$

  The matrix **F** is called the Fisher (or information) matrix

- To compute **F**, we Taylor expand the posterior near its peak – the maximum likelihood (ML) point
  - We need to compute first this point but this is simple:
    - When doing forecasts for future experiments, we know the ML beforehand (it is our fiducial model)
    - For real data → multi-dim. minimization algorithms are fast

$$\ln L\left(\theta_i\right) \approx \ln L\left(\hat{\theta}_i\right) + \frac{1}{2} \left.\frac{\partial^2 \ln L\left(\theta_i\right)}{\partial\theta_i \partial\theta_j}\right|_{\mathrm{ML}} \left(\theta_i - \hat{\theta}_i\right)\left(\theta_j - \hat{\theta}_j\right)$$

$$F_{ij} \equiv - \left.\frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial\theta_i \partial\theta_j}\right|_{\mathrm{ML}}$$

Properties of the Fisher Matrix

- Once we have **F**, the covariance matrix is simply its inverse
  - For 2 parameters:

$$[F]^{-1} = [C] = \begin{bmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{bmatrix}$$

- The ellipses axis lengths (α a and α b) and rotation angle are given by the eigenvalues and eigenvectors of **C**:

$$a^2 = \frac{\sigma_x^2 + \sigma_y^2}{2} + \sqrt{\frac{(\sigma_x^2 - \sigma_y^2)^2}{4} + \sigma_{xy}^2}$$

$$b^2 = \frac{\sigma_x^2 + \sigma_y^2}{2} - \sqrt{\frac{(\sigma_x^2 - \sigma_y^2)^2}{4} + \sigma_{xy}^2}$$

$$\tan 2\theta = \frac{2\sigma_{xy}}{\sigma_x^2 - \sigma_y^2}$$

σ	CL	$\Delta\chi^2$	α
1	68.3%	2.3	1.52
2	95.4%	6.17	2.48
3	99.73%	11.8	3.44

Properties of the Fisher Matrix (2)

- Marginalization over a parameter → simply remove the line & column of that parameter from $\mathbf{C} = \mathbf{F}^{-1}$ and invert the new, reduced $\mathbf{C}$
- Fixing a parameter to its best fit → simply remove the line & column of that parameter from $\mathbf{F}$
- Adding datasets → simply add $\mathbf{F}_{tot} = \mathbf{F}_1 + \mathbf{F}_2$
- Changing variables → simple jacobian transformation

$$[\mathrm{M}] = \begin{bmatrix} \frac{\partial x}{\partial a} & \frac{\partial x}{\partial b} & \frac{\partial x}{\partial c} \\ \frac{\partial y}{\partial a} & \frac{\partial y}{\partial b} & \frac{\partial y}{\partial c} \\ \frac{\partial z}{\partial a} & \frac{\partial z}{\partial b} & \frac{\partial z}{\partial c} \end{bmatrix}$$

$$[F'] = [M]^T [F][M]$$

Covariance Matrix

- When data is correlated, we need to compute the covariance matrix Σ

$$\Sigma_{ij} = cov(X_i, X_j) = E[(X_i - \mu_i)(X_j - \mu_j)]$$

$$\Sigma = \begin{bmatrix} \mathrm{E}[(X_1 - \mu_1)(X_1 - \mu_1)] & \cdots & \mathrm{E}[(X_1 - \mu_1)(X_n - \mu_n)] \\ \mathrm{E}[(X_2 - \mu_2)(X_1 - \mu_1)] & \cdots & \mathrm{E}[(X_2 - \mu_2)(X_n - \mu_n)] \\ \vdots & \ddots & \vdots \\ \mathrm{E}[(X_n - \mu_n)(X_1 - \mu_1)] & \cdots & \mathrm{E}[(X_n - \mu_n)(X_n - \mu_n)] \end{bmatrix}$$

Covariance Matrix (2)

- The cov matrix is related to the correlation matrix:

$$corr(\mathbf{X}) = \left(\operatorname{diag}\Sigma\right)^{-1/2} \Sigma \left(\operatorname{diag}\Sigma\right)^{-1/2}$$

  - It corresponds to the cov matrix of the standardized random variable set

$$X_i/\sigma(X_i)$$

- We can also define the cross-covariance between 2 vectors

$$cov(\mathbf{X}, \mathbf{Y}) = E\left[(\mathbf{X} - E(\mathbf{X})(\mathbf{Y} - E(\mathbf{Y})^T\right]$$

- Some properties of Σ:
  - It is positive-semidefinite and symmetric

$$cov(\mathbf{X_1} + \mathbf{X_2}, \mathbf{Y}) = cov(\mathbf{X_1}, (Y)) + cov(\mathbf{X_2}, (Y))$$

$$cov(\mathbf{A}\,\mathbf{X} + \mathbf{a}) = \mathbf{A}\, cov(\mathbf{X})\, \mathbf{A}^T$$

Covariance Matrix (3)

- To compute the cov matrix, we need to compute expected values (means) over many realizations
- Sometimes it can be computed analytically
- But more often it cannot, and one has to rely on simulations of mock data
  - Mock data, or mock catalogs, are collections of random realizations of data according to some distribution
  - Many mock (Monte Carlo) catalogs have to be generated in order to estimate the cov matrix with good precision
    - Never forget the golden rule: statistical errors decrease as sqrt(N)

Comparison of Different Methods

- When we have a posterior with non-linear dependence in the parameters (i.e., it is not Gaussian in the parameters, even if it is Gaussian in the data), the Fisher Matrix approach might yield incorrect results
- We have then several options to compute it. Let's assume we have *N* parameters. The most common are:
  - Fisher Matrix analysis anyway, for a first estimate
  - DALI method (higher order Fisher Matrix – arXiv:1401.6892)
    - First mature codes now becoming available
  - Grid analysis → compute numerically the posterior for a *N*-dimensional tensor, which is the exterior product of the different vectors of values for each parameter
    - Must guess the parameter ranges, or apply trial & error
      - Run first a very coarse-grained tensor, the refine
    - Very simple to code and implement

Comparison of Different Methods

- MCMC analysis → usually based on the Metropolis-Hastings algorithm
  - Goal: probe the N-dimensional space in a ”non-retangular” way → concentrate in the high-posterior space → more efficient search
  - We will study it later
- Nested Sampling analysis → see 1306.2144 & 1506.00171
- Comparison of techniques:
  - Grid → algorithm complexity grows as exp(N) [only OK for up to ~6 params]
    - Simple to code and good to learn
  - Fisher Matrix → fast and simple approx., maybe very wrong
  - DALI → more accurate, but lacks simplifying properties of FM
  - MCMC → complexity is $N \log(N)$, but code may require tuning
    - Lots of good MCMC codes in python out there
  - Nested Sampling → can be faster than MCMC

Model Selection

- We now want to address the more general problem: how to tell which of 2 competitive theories are statistically better given some data?
- Frequentist approach: compare the χ^2 and the number of degrees of freedom (d.o.f.) of the data in the 2 theories
  - The reduced χ^2 is the χ^2 / d.o.f.
    - Should be close to 1
  - The χ^2–distribution with k degrees of freedom is the distribution of a sum of the squares of k independent standard normal (i.e. gaussian) random variables.
  - The p.d.f. is given by

$$f_k(x) = \frac{x^{(k/2)-1}\exp^{-x/2}}{2^{k/2}\Gamma(k/2)}, \quad x \geq 0$$

Model Selection (2)

- This is the distribution if the likelihood of the data was exactly given by

$$L(\boldsymbol{x}|\boldsymbol{\theta}) = f_0 \exp\left[-\frac{1}{2}\sum_i \left(\frac{x_i^{\rm obs} - x_i^{\rm teo}(\boldsymbol{\theta})}{\sigma_i}\right)^2\right]$$

  - In a nutshell, it is the sum of squares of the "distance, in units of standard deviations, between data points and theoretical curve"
    - We refer to the total χ^2 as the sum $\displaystyle\sum_i \left(\frac{x_i^{\rm obs} - x_i^{\rm teo}(\boldsymbol{\theta})}{\sigma_i}\right)^2$
    - Frequentist mantra: good models have $\chi^2_{\rm red} = \dfrac{\chi^2}{k} \simeq 1$
      - Compute the p-value

$$\text{p-value} = \int_{\chi^2}^{\infty} f_k(x)\mathrm{d}x$$

Model Selection (3)

- The bayesian equivalent to χ^2 comparison is the Bayes ratio → ratio of evidences of models "1" and "2"
  - For a model "M" the evidence is

$$E(\mathbf{x}, M) = \int L(\mathbf{x}|\boldsymbol{\theta}^M) p(\boldsymbol{\theta}^M) \mathrm{d}^n \boldsymbol{\theta}^M$$

  - The Bayes factor between 2 models is just

$$B_{12} = \frac{\int L(x|\theta_i^{M_1}) p(\theta_i^{M_1}) \mathrm{d}^n \theta_i^{M_1}}{\int L(x|\theta_i^{M_2}) p(\theta_i^{M_2}) \mathrm{d}^n \theta_i^{M_2}}$$

  - $B_{12} > 1$ → model 1 is favored by the data (and vice-versa)
  - If you have an a priori reason to favor a model → generalize the above to include model priors $P(M|\mathbf{x}) \propto E(\mathbf{x}|M) p(M)$

Model Selection (4)

- The Bayes factor has several advantages over simple χ^2
  - If the data is poor and a particular parameter of one model is unconstrained by it, the model is not penalized
    - E.g.: a given dark energy model has a parameter related to, say, cluster abundance at $z = 2$, for which data is poor. This is good, because poor data ≠ poor model!
    - Mathematically → the posterior is approx. flat on this parameter → assuming (as usual) that the priors are independent we have that:

$$p(\boldsymbol{\theta}) = \prod_i p(\theta_i) \qquad E(\mathbf{x}, M) = \int L(\mathbf{x}|\boldsymbol{\theta}^M) p(\boldsymbol{\theta}^M) \mathrm{d}^n \boldsymbol{\theta}^M$$

$$\int p(\theta_p) \mathrm{d}\theta_p = 1 \qquad = \int L(\mathbf{x}|\boldsymbol{\theta}^{M-1}) p(\boldsymbol{\theta}^{M-1}) \mathrm{d}^n \boldsymbol{\theta}^{M-1}$$

Model Selection (5)

- To get a better intuition, we can study the simple case of Gaussian likelihoods + Gaussian priors → analytical $E(x)$
  - Assuming uncorrelated parameters, the posterior is then (integrating over the data):

$$P(\theta) = \prod_i L\left(x;\theta_i\right) p\left(\theta_i\right)$$

$$= \prod_i L_{\max,i} \left(2\pi\sigma_{P,i}^2\right)^{-\frac{1}{2}} \exp\left[-\frac{\left(\theta_i - \theta_i^{(B)}\right)^2}{2\sigma_{B,i}^2} - \frac{\left(\theta_i - \theta_i^{(P)}\right)^2}{2\sigma_{P,i}^2}\right]$$

$$= \prod_i L_{\max,i} \left(2\pi\sigma_{P,i}^2\right)^{-\frac{1}{2}} \exp\left[-\frac{1}{2}\frac{\left(\theta_i - \theta_i^{*}\right)^2}{\sigma_{i*}^2}\right] \exp\left[-\frac{1}{2}\frac{\left(\theta_i^{(B)} - \theta_i^{(P)}\right)^2}{\sigma_{B,i}^2 + \sigma_{P,i}^2}\right]$$

$$\theta_i^* = \frac{\sigma_{B,i}^2\theta_i^{(P)} + \sigma_{P,i}^2\theta_i^{(B)}}{\sigma_{B,i}^2 + \sigma_{P,i}^2} \qquad \sigma_{i*}^2 = \frac{\sigma_{P,i}^2\sigma_{B,i}^2}{\sigma_{B,i}^2 + \sigma_{P,i}^2}$$

Model Selection (6)

- The evidence is then given by

$$E = \int L\left(x;\theta_i\right) p\left(\theta_i\right) \mathrm{d}\theta_i$$

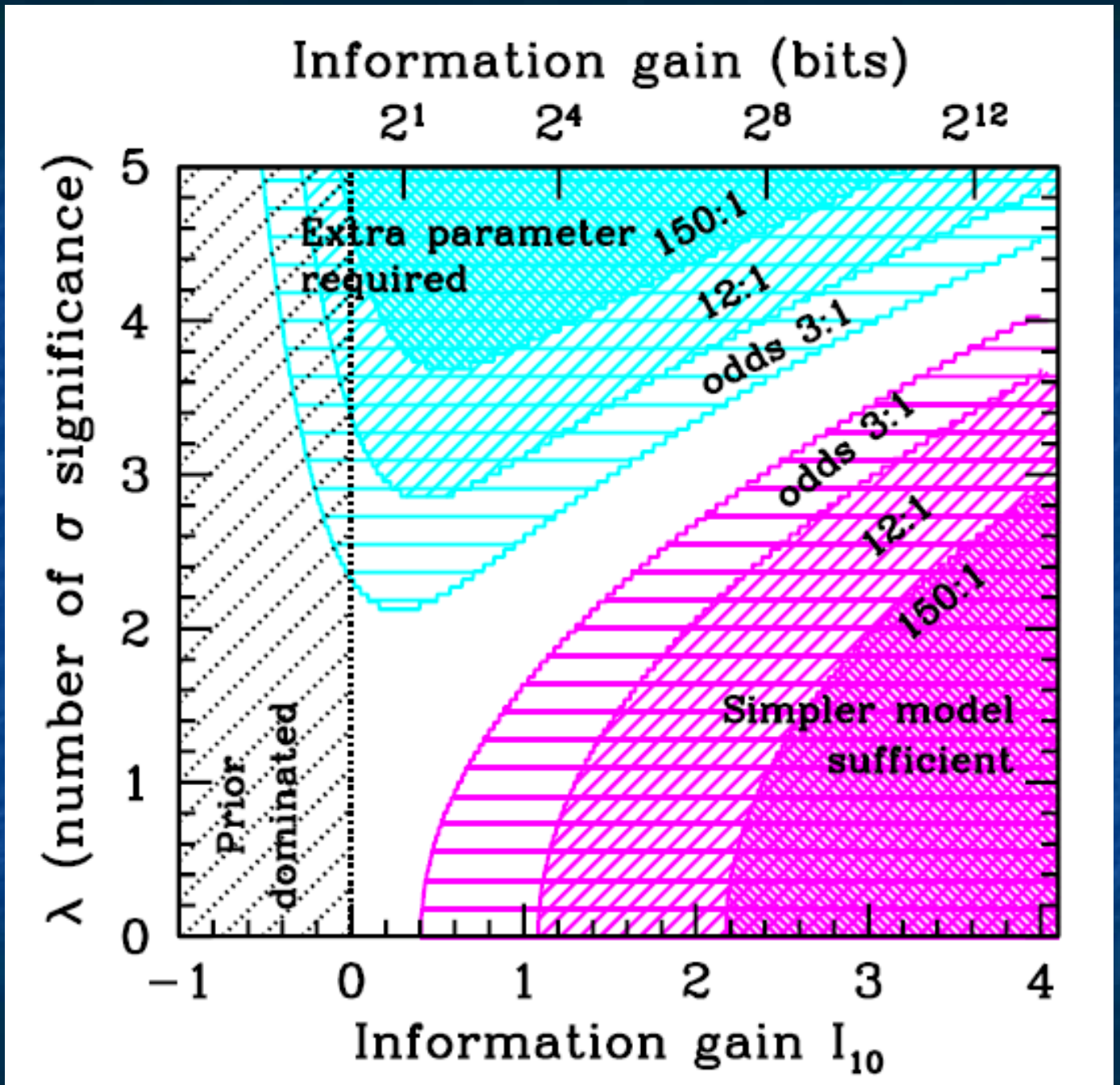
$$= \prod_i L_{\max,i} \frac{\sigma_{i*}}{\sigma_{P,i}} \exp\left\{-\frac{1}{2}\left[\left(\frac{\theta_i^{(B)}}{\sigma_{B,i}}\right)^2 + \left(\frac{\theta_i^{(P)}}{\sigma_{P,i}}\right)^2 - \left(\frac{\theta_i^*}{\sigma_{i*}}\right)^2\right]\right\}$$

  - Let's analyze the 3 distinct terms above
    - $f_{\max}$ is the max likelihood → how well the model fits the data
    - $\sigma_{i*}/\sigma_{P.i}$ is always < 1 → penalizes extra parameters constrained by the data → Ockham's Razor factor
    - exp[…] → penalizes cases where prior best fit is very different than posterior best fit

Jeffrey's Scale

- As we have seen: $B_{12} > 1 \rightarrow$ model 1 is favored by the data (and vice-versa)
  - There is no absolute rule of how big must B_{12} be to conclude whether one model must be replaced by another
  - A simple rule-of-thumb though is just to use a simple scale to guide the discussion. Jeffrey's scale is often used
    - Note: this scale has no fundamental grounds!!!

$\lvert \ln B_{01} \rvert$	Odds	Probability	Strength of evidence
< 1.0	$\lesssim 3:1$	< 0.750	Inconclusive
1.0	$\sim 3:1$	0.750	Weak evidence
2.5	$\sim 12:1$	0.923	Moderate evidence
5.0	$\sim 150:1$	0.993	Strong evidence


Information gain (bits)
2^1
2^4
2^8
2^{12}
λ (number of σ significance)
5
4
3
2
1
0
Extra parameter required
150:1
12:1
odds 3:1
odds 3:1
12:1
150:1
Simpler model sufficient
Prior dominated
−1
0
1
2
3
4
Information gain I_{10}

Extra Slides

Example: 2-pt correlation function

- Let's study one particular example involving the 2-point correlation function in astronomy
- We want to study how a given class of objects are distributed in the sky
  - Let's focus on galaxies, for instance
  - Given a random galaxy in a location, the 2-point correlation function $\xi(r)$ describes the *excess probability* that another galaxy will be found within a given (scalar) distance r, compared to an *uniform distribution*
  - Because gravity attracts objects they tend to cluster together, so we expect $\xi(r)$ to decrease as r increases

Example: 2-pt correlation function

- In principle, ξ depends on the vector **r**, but if our data is assumed to be statistically homogeneous (same statistics everywhere), then x only depends on r = |**r**|.
  - Let's assume this
- If we define the number of galaxies dN at a small volume dV, located at distance r from a given galaxy, and the average density as ρ_0,we have

$$\mathrm{d}N = \rho_0 \mathrm{d}V \left[1 + \xi(r)\right]$$

  - It's an excess probability → we have an integral constraint

$$\xi(r) = \frac{\mathrm{d}N(r)}{\rho_0 \mathrm{d}V} - 1 = \frac{E(\rho_r)}{\rho_0} - 1$$

$$\int \xi(r)\mathrm{d}V = \frac{1}{\rho_0}\int \frac{\mathrm{d}N}{\mathrm{d}V}\mathrm{d}V - V = \frac{N}{\rho_0} - V = 0$$

Example: 2-pt correlation function

- If the correlation ξ is positive (*negative*), there are more (*less*) particles than an uniform distribution
- For a given catalog, unless the volume has a very simple geometry (say, a perfect sphere), it is impossible to compute the correlation function or its cov matrix analytically
- We can estimate ξ in a given catalog with the following estimator, where DD means number of galaxies at a distance (r, r + Δr) in the data, and RR the same in a random, uniform catalog with the same volume:

$$\xi = \frac{DD}{RR} - 1$$

Example: 2-pt correlation function

$$\xi = \frac{DD}{RR} - 1$$

- We compute the numbers DD(r) and RR(r) for all pairs of objects (here, galaxies)

- We do this for a number of distance bins
  - In each bin, we need to estimate the error bars
  - Since the same objects enter different bins, the data is highly correlated → we need to compute the cov matrix
    - We need to generate many random data catalogs!

Example: 2-pt correlation function