

Astronomical Data Analysis: the Bayesics



Alan Heavens

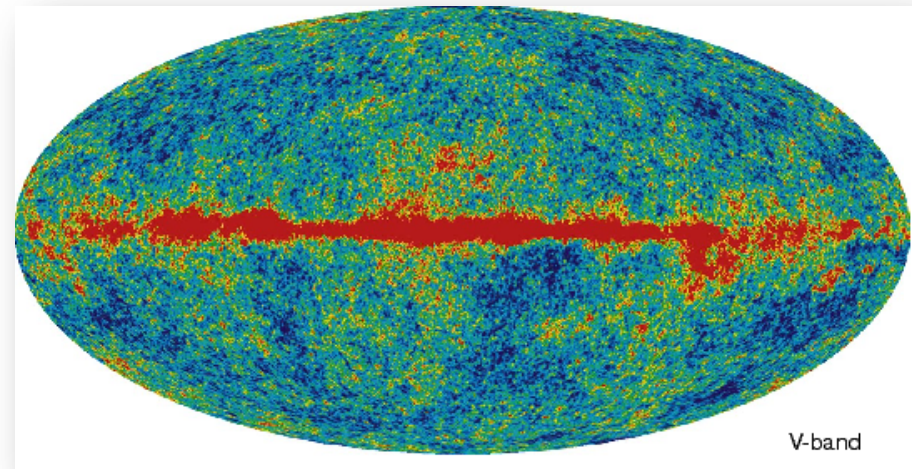
University of Edinburgh, UK

Lectures given at STFC Introductory School,
University of Glasgow, August 2011

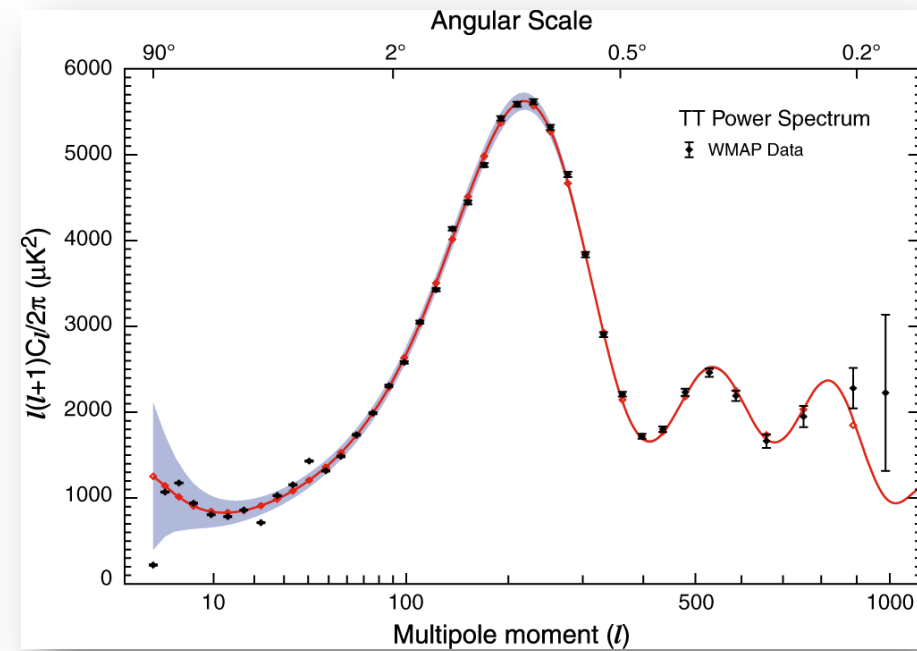
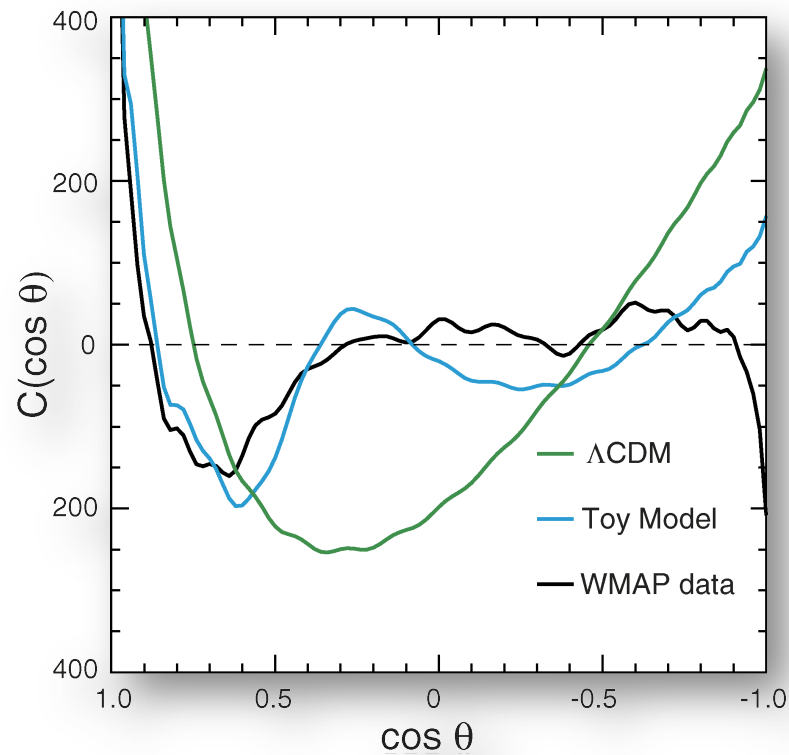
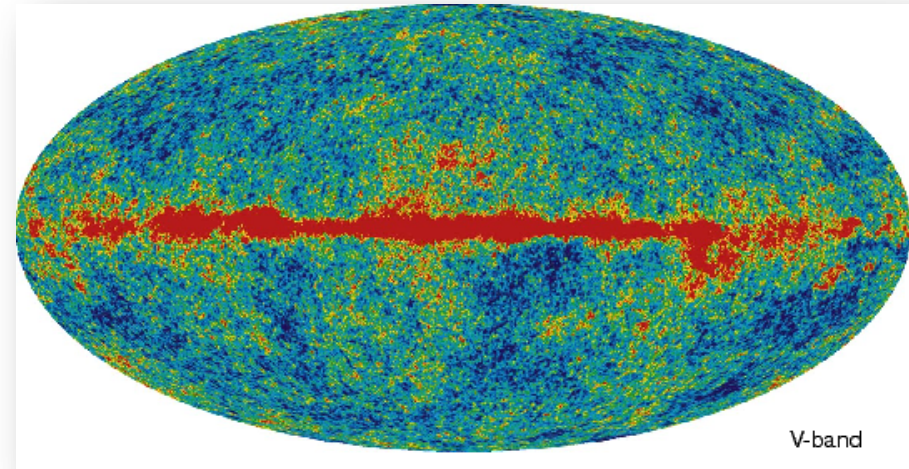


Outline

- Types of problem
- Bayes' theorem
- **Parameter Estimation**
 - Marginalisation
 - Errors
- Error prediction and experimental design:
Fisher Matrices
- **Model Selection**



LCDM fits the WMAP data well.



Inverse problems

- Most cosmological problems are *inverse problems*, where you have a set of data, and you want to infer something.
- Examples
 - Hypothesis testing
 - Parameter estimation
 - Model selection

Examples

- Hypothesis testing
 - Is the CMB radiation consistent with (initially) gaussian fluctuations?
- Parameter estimation
 - In the Big Bang model, what is the value of the matter density parameter?
- Model selection
 - Do cosmological data favour the Big Bang theory or the Steady State theory?
 - Is the gravity law General Relativity or higher-dimensional?

What is probability?

- **Frequentist view**: p describes the relative *frequency of outcomes* in infinitely long trials
- **Bayesian view**: p expresses our *degree of belief*
- **Bayesian view** is closer to what we seem to want from experiments: e.g. *given the WMAP data, what is the probability that the density parameter of the Universe is between 0.9 and 1.1?*
- Cosmology is in good shape for inference because we have decent model(s) with parameters – well-posed problem

Bayes' Theorem

- Rules of probability:
- $p(x)+p(\text{not } x) = 1$ sum rule
- $p(x,y) = p(x|y)p(y)$ product rule
- $p(x) = \sum_k p(x,y_k)$ marginalisation
- Sum \longrightarrow integral continuum limit (p =pdf)

- $p(x,y)=p(y,x)$ gives *Bayes' theorem*

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

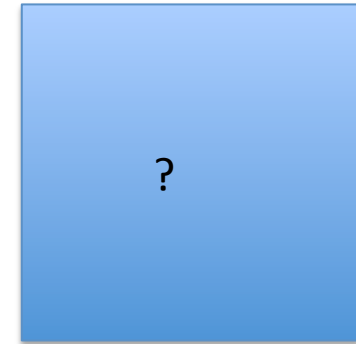
$p(x|y)$ is not the same as $p(y|x)$

- $x = \text{female}, y = \text{pregnant}$
- $p(y|x) = 0.03$
- $p(x|y) = 1$



An exercise in using Bayes' theorem

You choose
this one



Do you change your choice?

This is the Monty Hall problem



Bayes' Theorem and Inference

- If we accept p as a degree of belief, then what we often want to determine is*

$$p(\theta|x)$$

θ : model parameter(s), x : the data

To compute it, use Bayes' theorem

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

*This is RULE 1: start by writing down what it is you want to know

RULE 2: There is no RULE n , $n>1$

Posteriors, likelihoods, priors and evidence

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

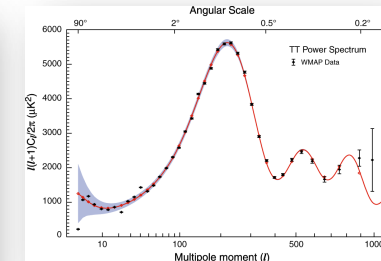
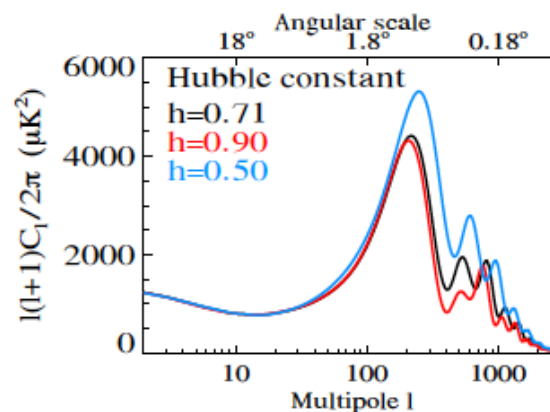
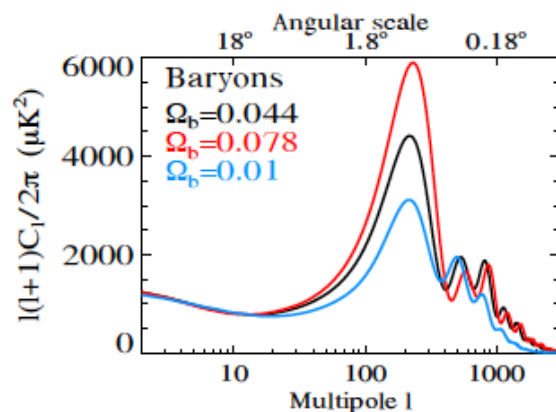
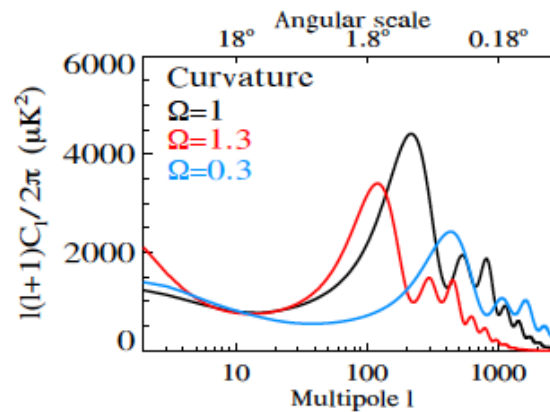
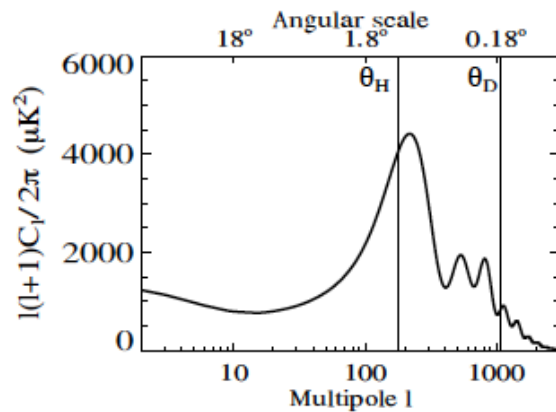
Posterior Likelihood L Evidence Prior

Note that we interpret these in the context of a model M , so all probabilities are really conditional on M (and indeed on any prior info I). E.g. $p(\theta) = p(\theta|M)$

The *evidence* looks rather odd – what is the probability of the data? For parameter estimation, we can ignore it – it simply normalises the posterior.

Noting that $p(x) = p(x|M)$ makes its role clearer. In *model selection* (from M and M'), $p(x|M) \neq p(x|M')$

Forward modelling $p(x|\theta)$



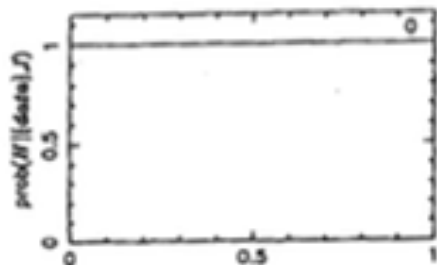
With noise properties we can predict the *Sampling Distribution* (the probability for a general set of data; the *Likelihood* is the probability for the specific data we have)

State your priors

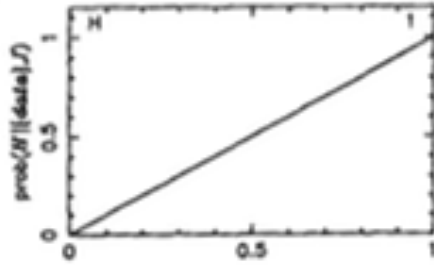
- In easy cases, the effect of the prior is simple
- As experiment gathers more data, the likelihood tends to get narrower, and the influence of the prior diminishes
- **Rule of thumb:** if changing your prior[†] to another reasonable one changes the answers significantly, you need more data
- **Reasonable priors?** Uninformative* – constant prior
- scale parameters in $[0, \infty)$; uniform in log of parameter (Jeffreys' prior*)
- **Beware:** in more complicated, multidimensional cases, your prior may have subtle effects...

[†] I mean the raw theoretical one, not modified by an experiment

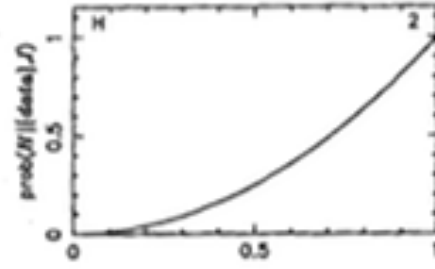
* Actually, it's better not to use these terms – other people use them to mean different things – just say what your prior is!



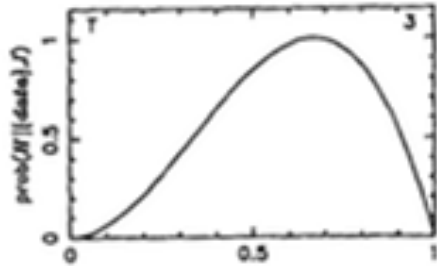
Bias-weighting for heads H



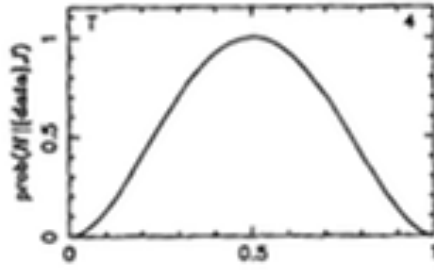
Bias-weighting for heads H



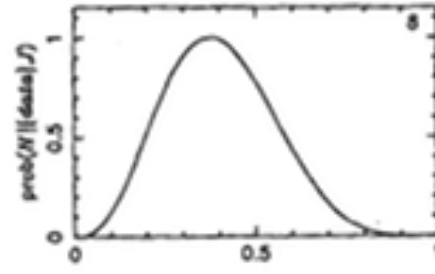
Bias-weighting for heads H



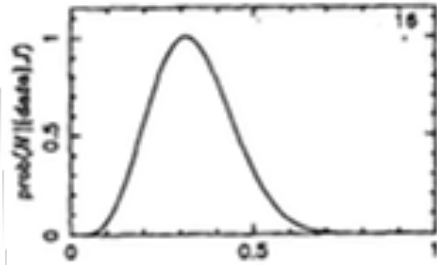
Bias-weighting for heads H



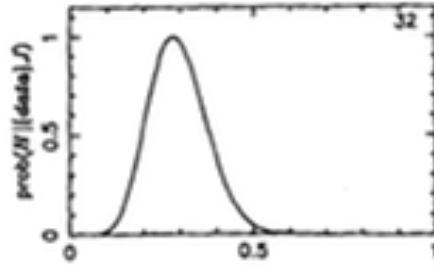
Bias-weighting for heads H



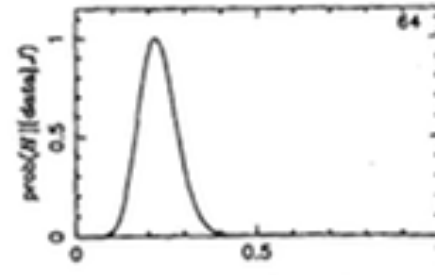
Bias-weighting for heads H



Bias-weighting for heads H



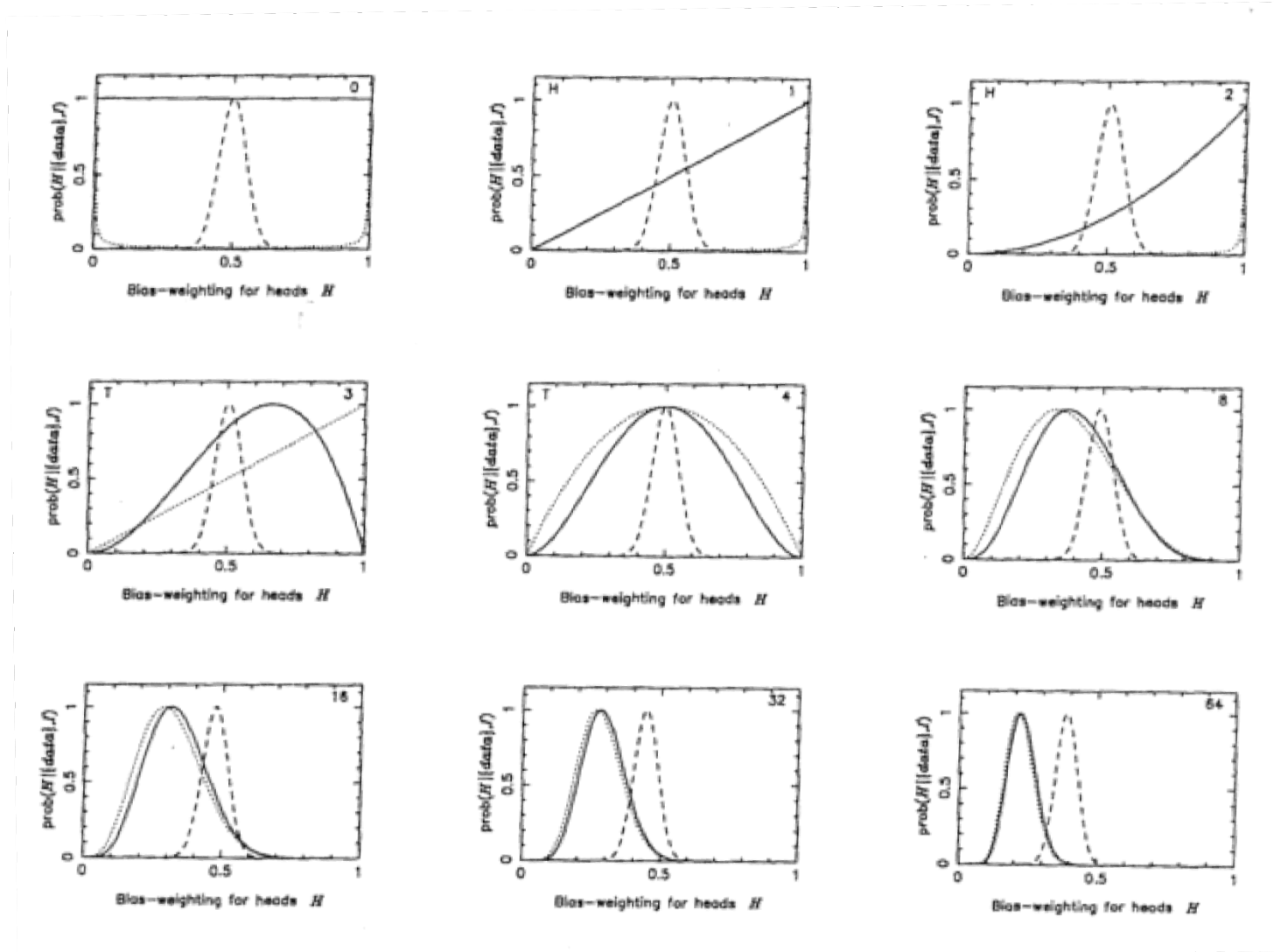
Bias-weighting for heads H



Bias-weighting for heads H

Sivia & Skilling. IS THE COIN FAIR?

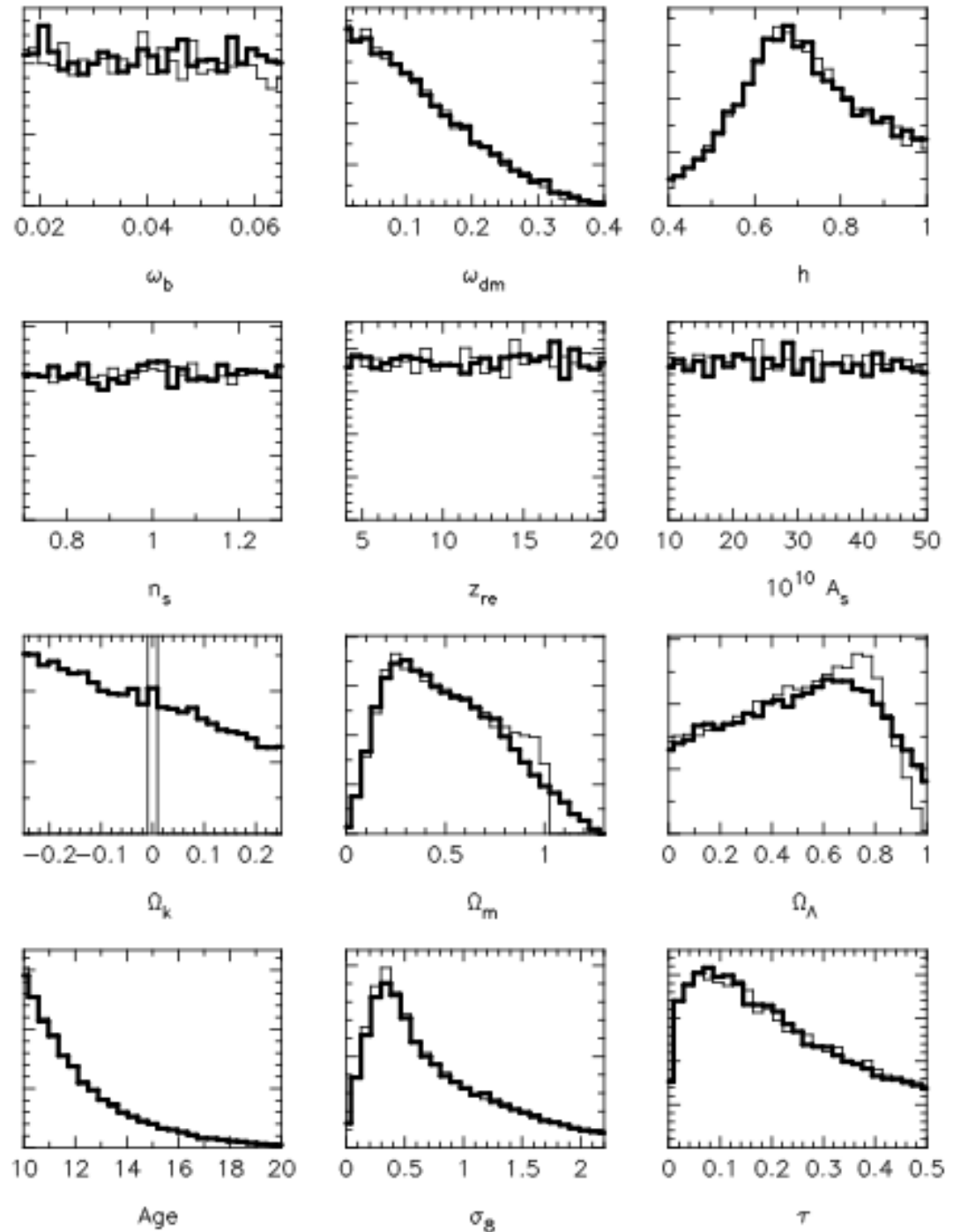
The effect of priors



Sivia & Skilling

- VSA CMB experiment

(Slosar et al 2003)

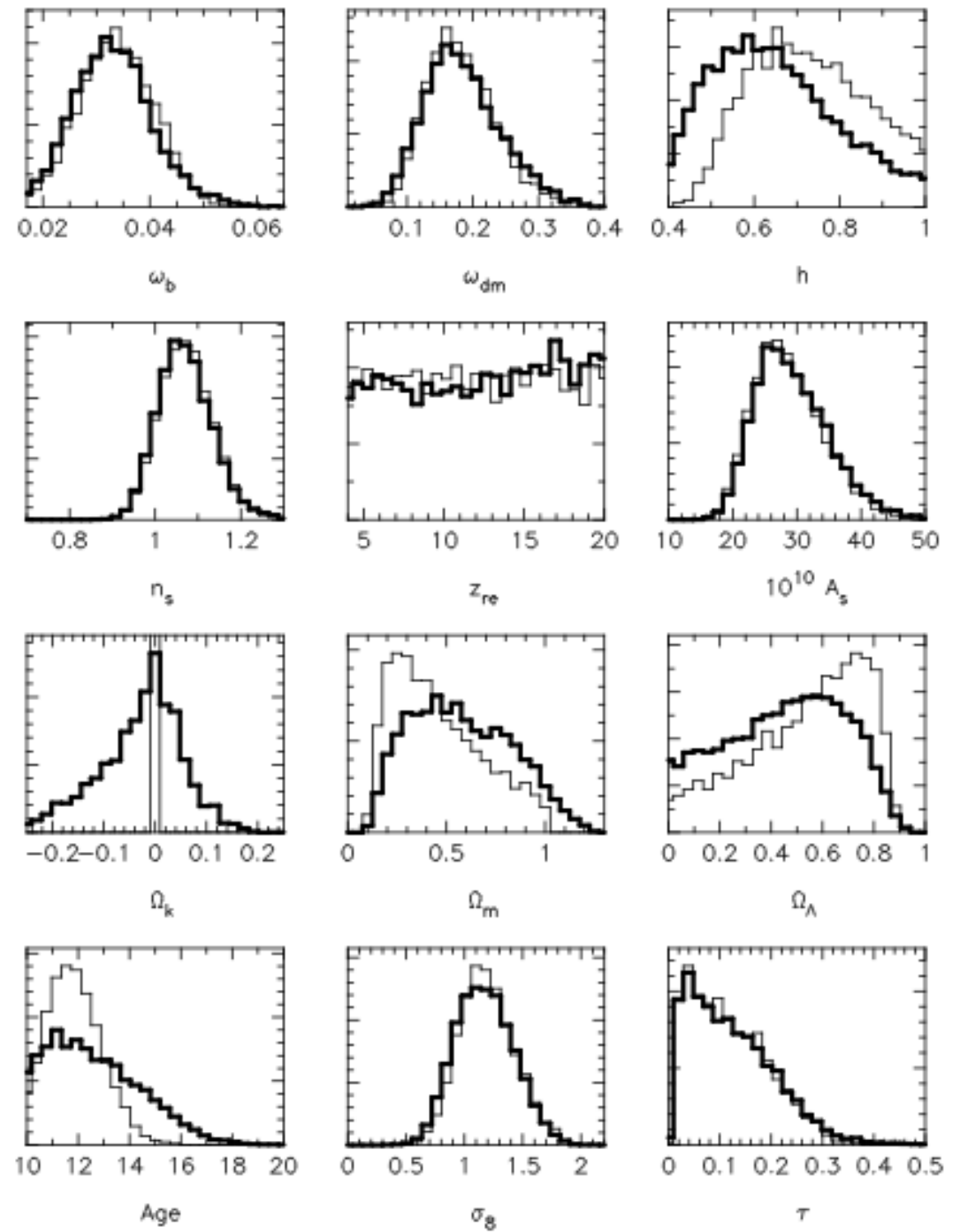


Priors: $\Lambda \geq 0$
 $10 \leq \text{age} \leq 20 \text{ Gyr}$

$h \approx 0.7 \pm 0.1$

There are no data in these plots – it is all coming from the prior!

VSA posterior



Estimating the parameter(s)

- Commonly the mode is used (the peak of the posterior)
- Mode = Maximum Likelihood Estimator, *if the priors are uniform*
- The *posterior mean* may also be quoted

$$\bar{\theta} = \int \theta p(\theta|x) d\theta$$

Errors

If we assume uniform priors, then the posterior is proportional to the likelihood.

If further, we assume that the likelihood is single-moded (one peak at θ_0), we can make a Taylor expansion of $\ln L$:

$$\ln L(x; \theta) = \ln L(x; \theta_0) + \frac{1}{2}(\theta_\alpha - \theta_{0\alpha}) \frac{\partial^2 \ln L}{\partial \theta_\alpha \partial \theta_\beta} (\theta_\beta - \theta_{0\beta}) + \dots$$

$$L(x; \theta) = L_0 \exp \left[-\frac{1}{2}(\theta_\alpha - \theta_{0\alpha}) H_{\alpha\beta} (\theta_\beta - \theta_{0\beta}) + \dots \right]$$

where the Hessian matrix is defined by these equations. Comparing this with a gaussian, the *conditional error* (keeping all other parameters fixed) is

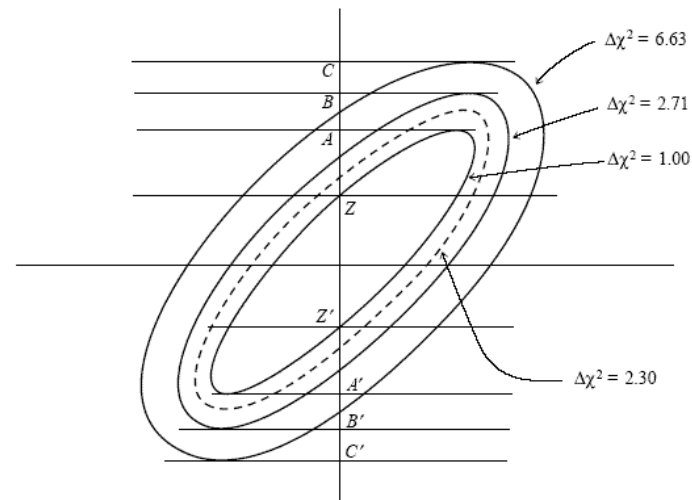
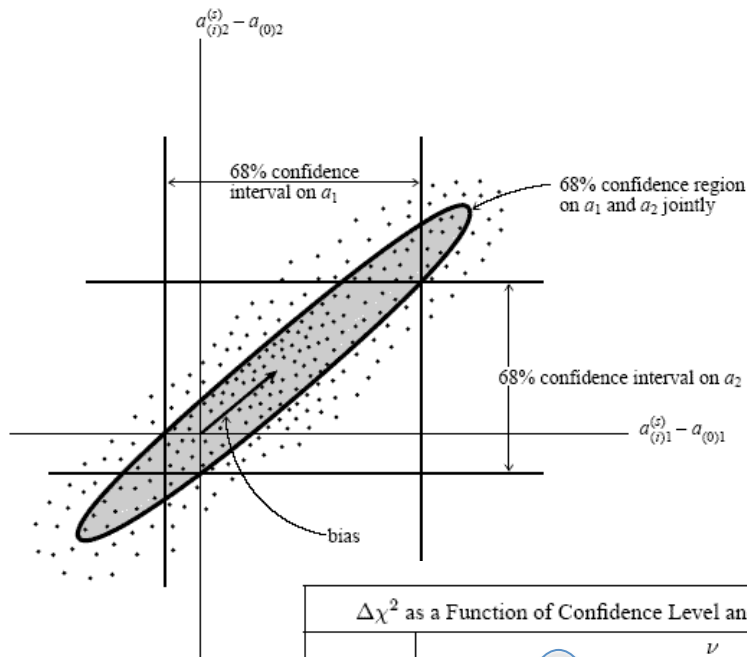
$$\sigma_\alpha = \frac{1}{\sqrt{H_{\alpha\alpha}}}$$

Marginalising over all other parameters gives the *marginal error*

$$\sigma_\alpha = \sqrt{(H^{-1})_{\alpha\alpha}}$$

How do I get error bars in several dimensions?

- Read Numerical Recipes, Chapter 15.6



$$L \propto e^{-\frac{1}{2}\chi^2}$$

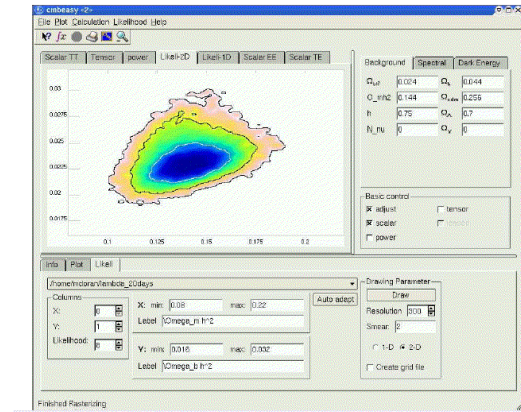
$\Delta\chi^2$ as a Function of Confidence Level and Degrees of Freedom						
p	ν					
	1	2	3	4	5	6
68.3%	1.00	2.30	3.53	4.72	5.89	7.04
90%	2.71	4.61	6.25	7.78	9.24	10.6
95.4%	4.00	6.17	8.02	9.70	11.3	12.8
99%	6.63	9.21	11.3	13.3	15.1	16.8
99.73%	9.00	11.8	14.2	16.3	18.2	20.1
99.99%	15.1	18.4	21.1	23.5	25.7	27.8

Beware! Assumes gaussian distribution

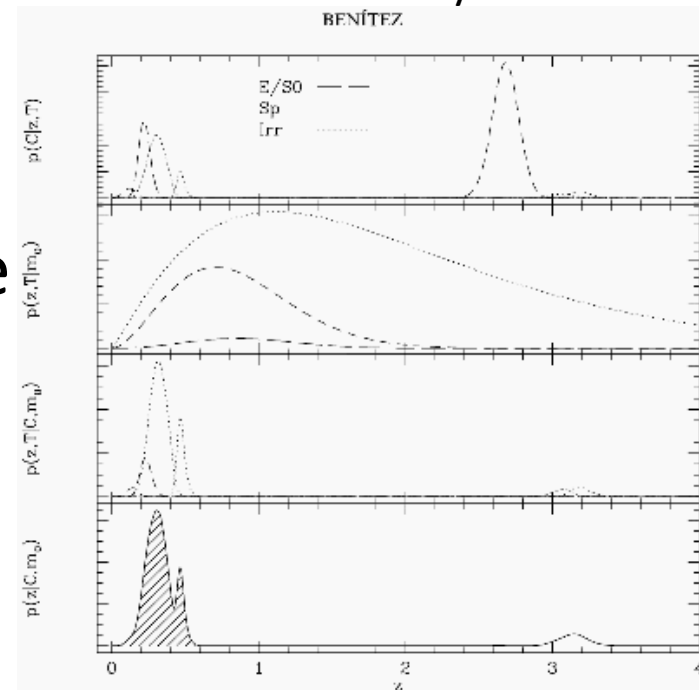
Say what your errors are!
e.g. 1σ , 2 parameter

Multimodal posteriors etc

- Peak may not be gaussian
- Multimodal? Characterising it by a mode and an error is probably inadequate. May have to present the full posterior.
- Mean posterior may not be useful in this case – it could be very unlikely, if it is a valley between 2 peaks.



From CMBEasy MCMC
BENÍTEZ



From BPZ

Fisher Matrices

- Useful for forecasting errors, and experimental design
- The likelihood depends on the data collected. Can we estimate the errors before we do the experiment?
- With some assumptions, yes, using the Fisher matrix

$$F_{\alpha\beta} \equiv - \left\langle \frac{\partial^2 \ln L}{\partial \theta_\alpha \partial \theta_\beta} \right\rangle$$

Gaussian errors

- If the data have gaussian errors (which may be correlated) then we can compute the Fisher matrix easily:

$$F_{\alpha\beta} = \frac{1}{2} \text{Tr} [C^{-1} C_{,\alpha} C^{-1} C_{,\beta} + C^{-1} M_{\alpha\beta}],$$

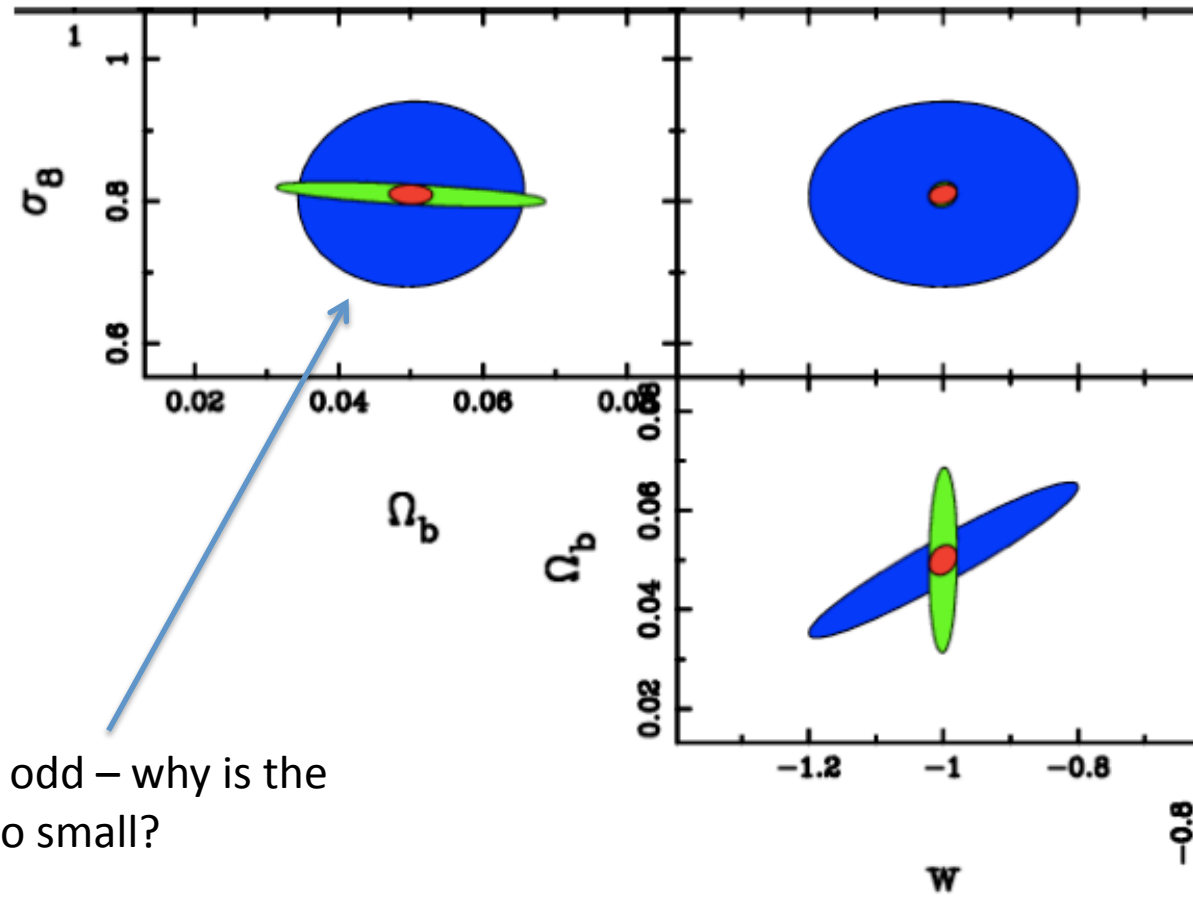
e.g. Tegmark, Taylor, Heavens 1997

Forecast
marginal error
on parameter α

$$\sigma_{\alpha} = \sqrt{(F^{-1})_{\alpha\alpha}}$$

$$\mu_{\alpha} = \langle x_{\alpha} \rangle \quad C_{\alpha\beta} = \langle (x - \mu)_{\alpha} (x - \mu)_{\beta} \rangle \quad M_{\alpha\beta} = \mu_{,\alpha} \mu_{,\beta}^T + \mu_{,\alpha}^T \mu_{,\beta}$$

Combining datasets



This looks odd – why is the red blob so small?

Open source Fisher matrices – icosmo.org

Initiative Tools Resources Help Contact Us FAQs

INITIATIVE FOR COSMOLOGY

Welcome!

This site is designed to make cosmology calculations easy and pain-free. Here, you will find a host of tools and resources for performing calculations, ranging from distance calculations to cosmological error predictions for future surveys.

The site also contains a set of tutorials and links that are useful whether you are a newbie to cosmology or a seasoned professional. These resources have been made available in an easy-to-access format and will be continually updated and expanded.

COSMOLOGY TOOLS:
You can perform a calculation either by using your web browser or by [downloading the source code](#). To get started you can either go to [tools](#), and you will be guided through each step. Alternatively, you can use the QuickStart Calculator to the right.

COSMOLOGY RESOURCES:
Here you will find general cosmology support materials, such as tutorials and links to external sites. To find the material you need go to [resources](#) or use the QuickStart Tutorial to the right. If you wish to create your own interactive web pages you can use the templates available [here](#). A discussion forum for the tools and resources is provided at [Cosmocooffee](#).

NEWS:
21/05/2009 - [w\(z\) eigenfunctions](#). Module for [astro-ph/0905.3383](#) to be included in [iCosmo v1.2](#).
20/05/2009 - [Hardware-Software balance](#). Code for [astro-ph/0905.3176](#) can be downloaded here [iCosmo PublicAstroCodes](#).
11/02/2009 - [Redshift Distortion & ISW](#). Module for [astro-ph/0902.1759](#) to be included in [iCosmo v1.2](#).
21/01/2009 - [Cloud Cosmology](#). Article available [here](#). Template web pages available [here](#).

QuickStart Calculator

Ω_m 0.3 Ω_{DE} 0.7
 Ω_B 0.045 w_0 -0.95
 h 0.7 w_a 0.0
 σ_8 0.8 n_s 1.0

[QuickStart Cosmology](#)

QuickStart Tutorial

Gravitational Lensing
 Galaxy Correlations
 CMB

[QuickStart Tutorial](#)

Cosmology	Ω_m 0.3	w_0 -0.9	w_a 0	h 0.7	Ω_B 0.04	σ_8 0.8	n_s 1	Ω_{DE} 0.7	Change Cosmology
Gaussian Prior	-	-	-	-	-	-	-	-	Add-Prior Help
1- σ Survey:	0.006	0.051	0.172	0.053	0.010	0.008	0.012	0.025	$\Delta w_{pivot}=0.0253$ $FoM=229.62$

Survey 1: LSST : Area = 20000.00 sq. deg., z error = 0.030(1+z), Med Redshift = 1.50, Gals per sq. arcmin = 40.00, No. z-bins = 10.00,

Products of the Cosmology Error Calculations: Parameter Ellipses

Graphical Options: [Help](#)

Auto-Scale

x_{min} 0.9861
 x_{max} 1.0139
 y_{min} 0.79040
 y_{max} 0.80960

[Re-Scale](#)

<http://icosmo.org>

Cosmology Options: [Help](#)

x axis options:

y axis options:

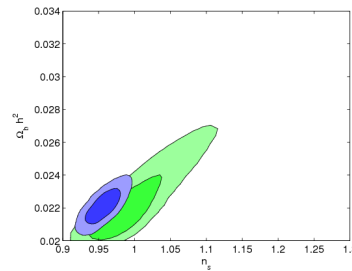
Fisher Matrix 1: Lensing

Line Colour:

Fill Colour:

Computing posteriors

- For 2 parameters, a grid is usually possible
 - Marginalise by numerically integrating along each axis of the grid



- For $\gg 2$ parameters it is not feasible to have a grid (e.g. 10 points in each parameter direction, 12 parameters = 10^{12} likelihood evaluations)
- Methods: Monte Carlo Markov Chain (MCMC) etc

Numerical Sampling methods:



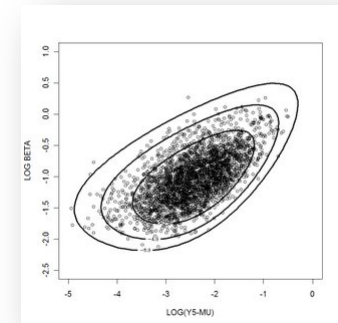
Markov Chain Monte Carlo



Aim of MCMC: generate a set of points in the parameter space whose *distribution function is the same as the target density*.

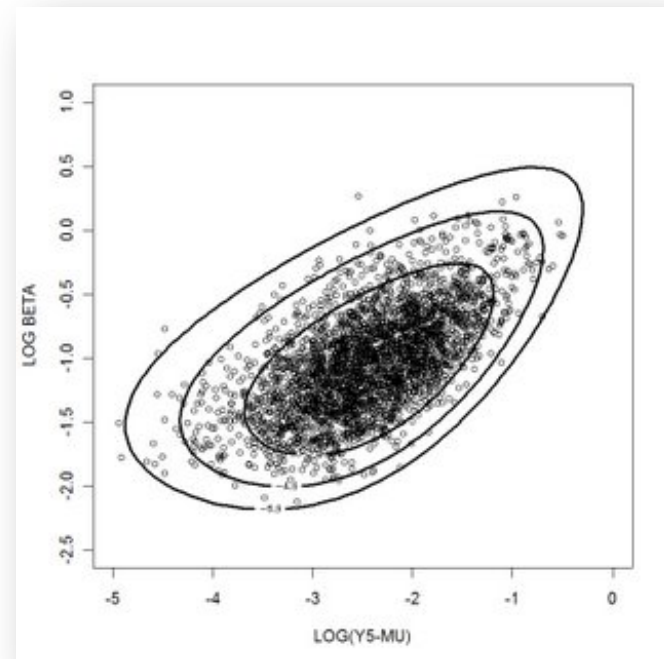
MCMC follows a Markov process - i.e. the next sample depends on the present one, but not on previous ones.

MCMC takes random steps and accepts or rejects the new point

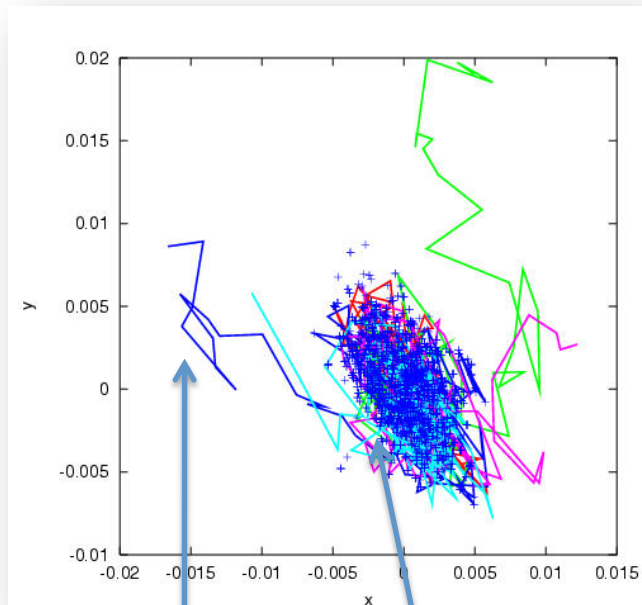


The proposal distribution

- Too small, and it takes a long time to explore the target
- Too large and almost all trials are rejected
- $q \sim$ 'Fisher size' is good.

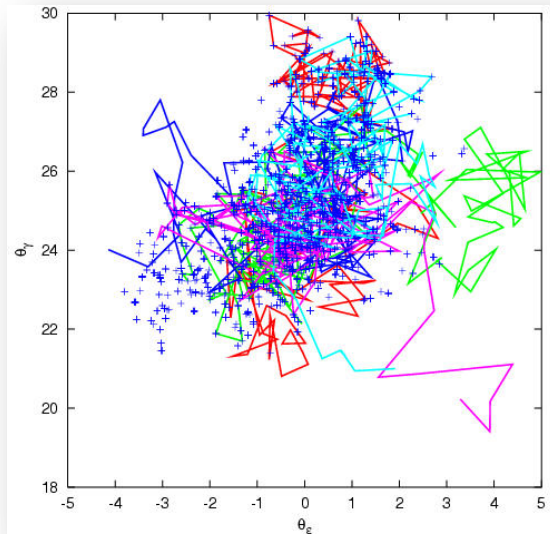


Burn-in and convergence



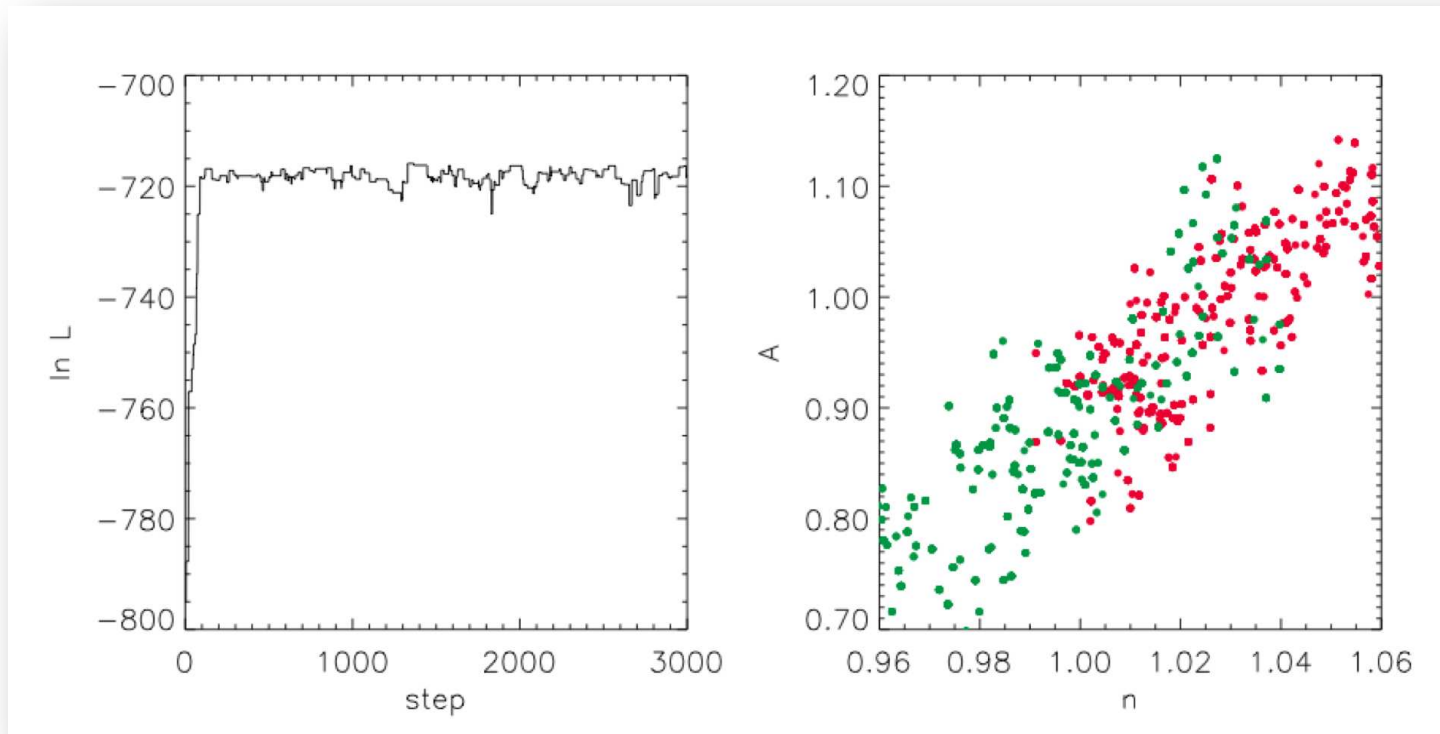
“Burn-in”

Points are correlated



You *must* use a convergence test.
Gelman-Rubin test is most common

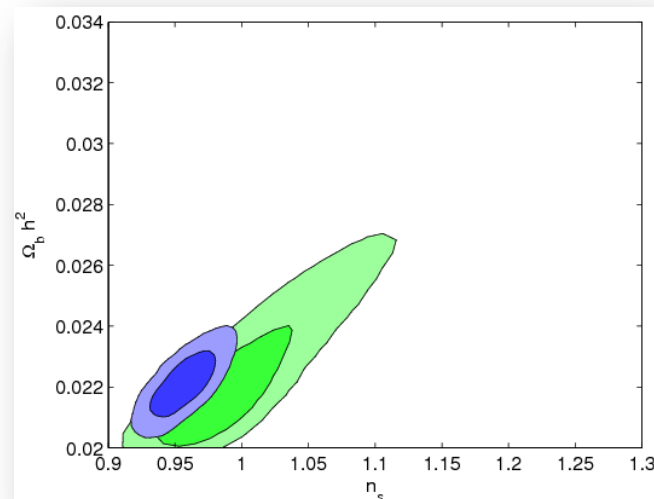
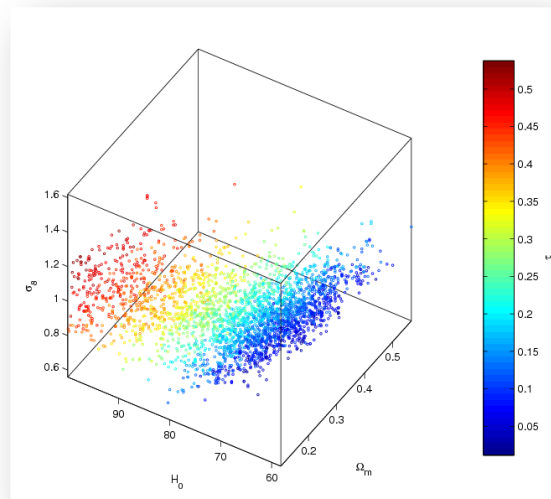
Unconverged chains



Verde et al 2003

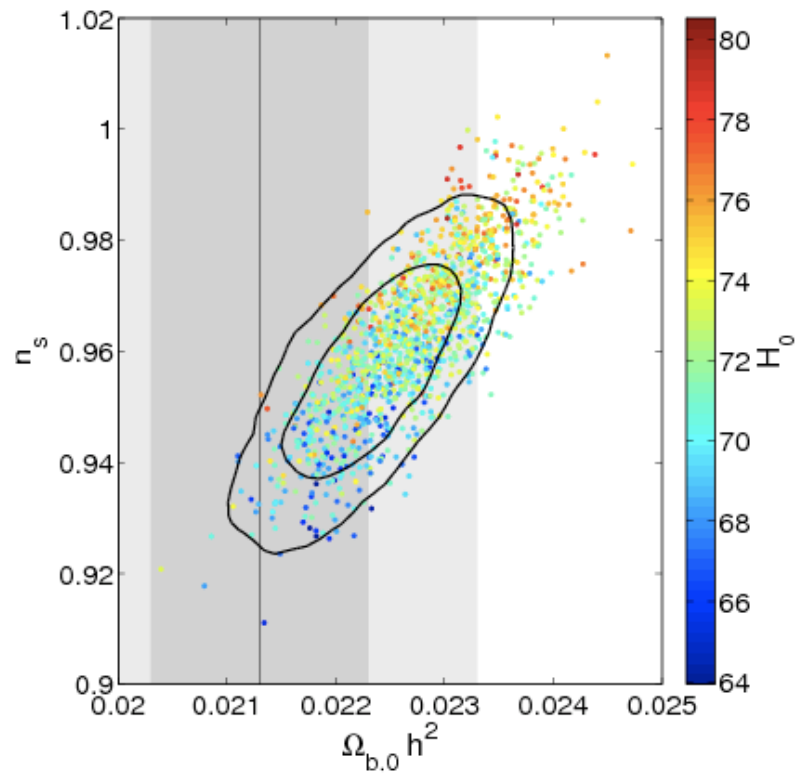
Marginalisation

- Marginalisation is trivial
 - Each point in the chain is labelled by all the parameters
 - To marginalise, just ignore the labels you don't want



CosmoMC

Cosmological MonteCarlo



<http://cosmologist.info/cosmomc/>

Samples from WMAP 5-yr likelihood combined with deuterium constraint ([0805.0594](https://arxiv.org/abs/0805.0594))



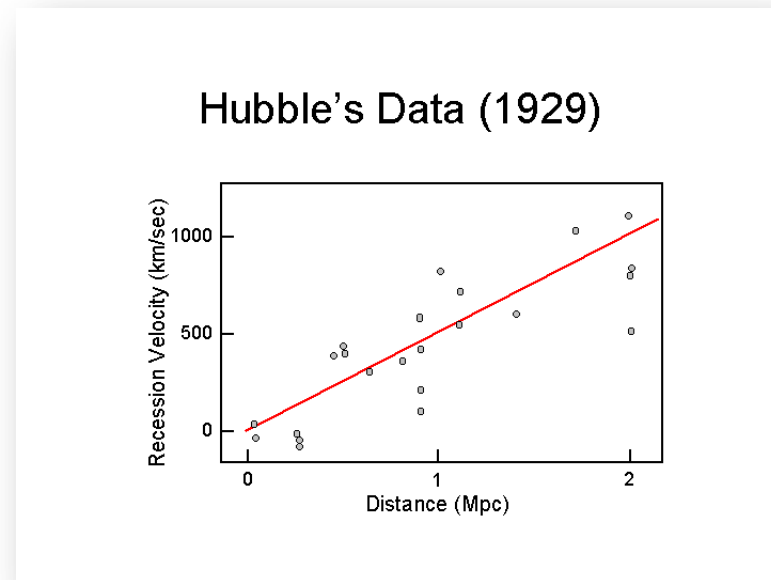
Model Selection



- Model selection: in a sense a higher-level question than parameter estimation
- Is the theoretical framework OK, or do we need to consider something else?
- We can compare widely different models, or may want to decide whether we need to introduce an additional parameter into our model (e.g. curvature)
- In the latter case, using likelihood alone is dangerous: the new model will always be at least as good a fit, and virtually always better, so naïve maximum likelihood won't work.

Hubble and Hendry

- E. Hubble has a theory that $v = Hr$ for all galaxies, where H is a free parameter.
- M. Hendry has a theory that $v = 0$ for all galaxies
- Who should we believe?



Bayesian approach

- Let models be M, M'
- Apply RULE 1: Write down what you want to know. Here it is $p(M | \mathbf{x})$ - the probability of the model, given the data.

More Bayes:

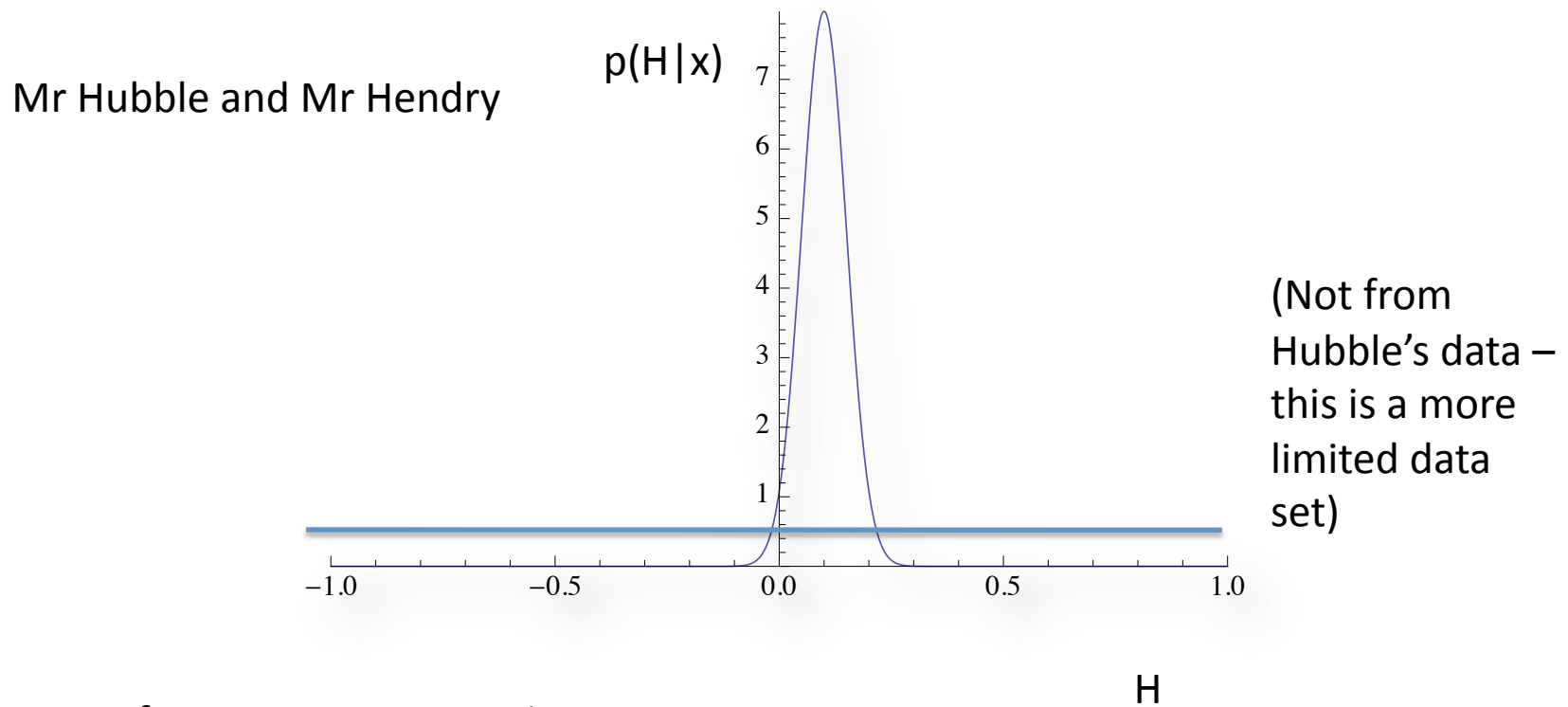
$$p(M|\mathbf{x}) = \frac{p(\mathbf{x}|M)p(M)}{p(\mathbf{x})}$$

$$\frac{p(M'|\mathbf{x})}{p(M|\mathbf{x})} = \frac{p(M') \int d\theta' p(\mathbf{x}|\theta', M')p(\theta'|M')}{p(M) \int d\theta p(\mathbf{x}|\theta, M)p(\theta|M)}$$

Define the Bayes factor as the ratio of [evidences](#):

$$B \equiv \frac{\int d\theta' p(\mathbf{x}|\theta', M')p(\theta'|M')}{\int d\theta p(\mathbf{x}|\theta, M)p(\theta|M)}$$

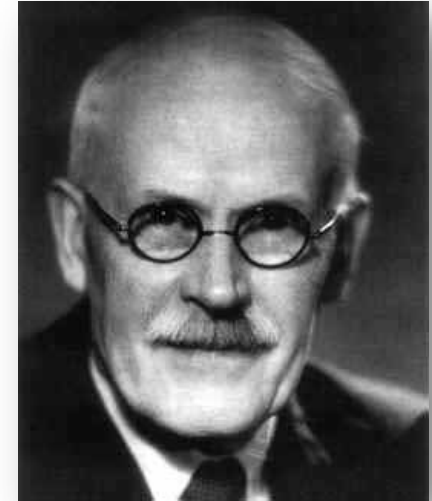
Which model is more likely?



Prior of extra parameter is $\frac{1}{2}$

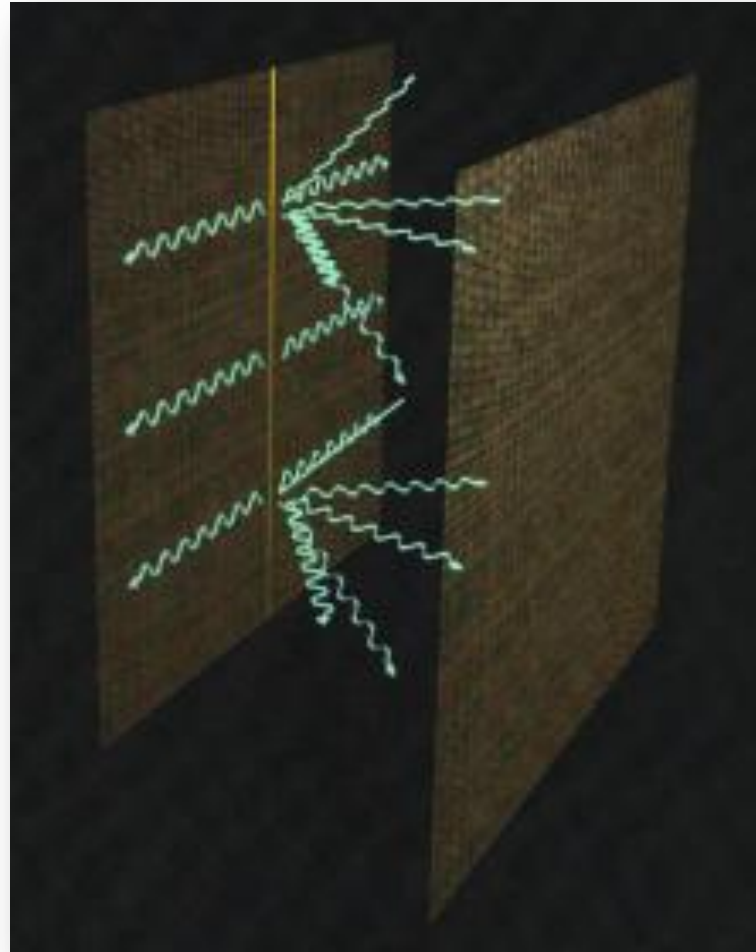
$$\frac{p(Hendry)}{p(Hubble)} = \frac{1.1}{0.5} = 2.2$$

Jeffreys' criteria



- Evidence:
- $1 < \ln B < 2.5$ 'substantial'
- $2.5 < \ln B < 5$ 'strong'
- $\ln B > 5$ 'decisive'
- These descriptions seem too aggressive:
 - $\ln B=1$ corresponds to a posterior probability for the less-favoured model which is 0.37 of the favoured model

Extra-dimensional gravity?



Evidence for beyond-Einstein gravity

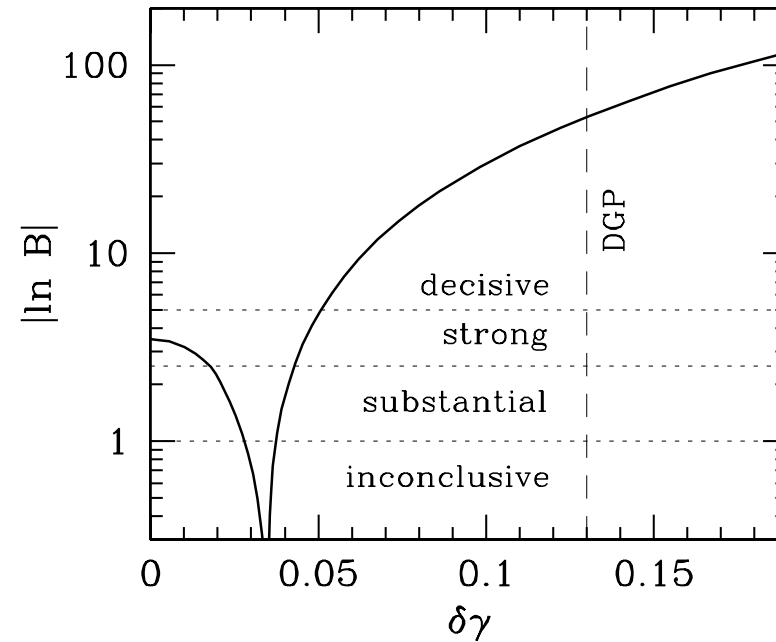
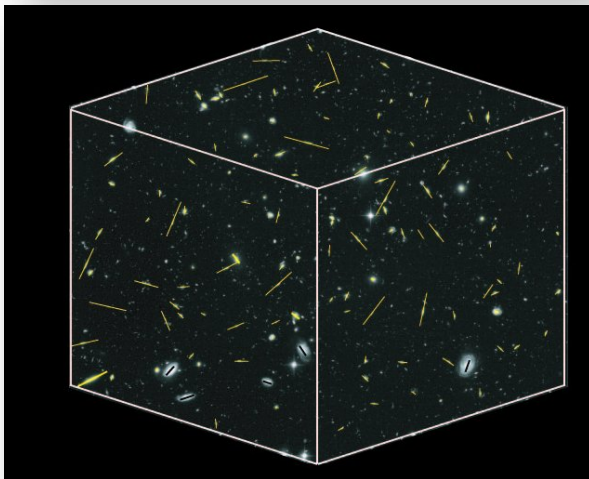
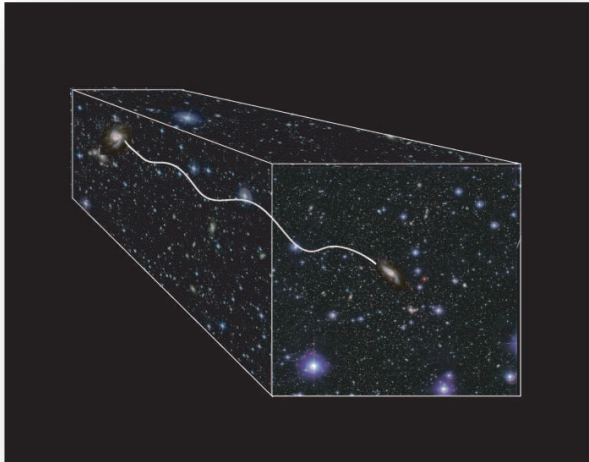
- How would we tell? Different growth rate

$$\frac{\delta_m}{a} \equiv g(a) = \exp \left\{ \int_0^a \frac{da'}{a'} [\Omega_m(a')^\gamma - 1] \right\}$$

$$\gamma = 0.55 (GR), 0.68 (Flat DGP)$$

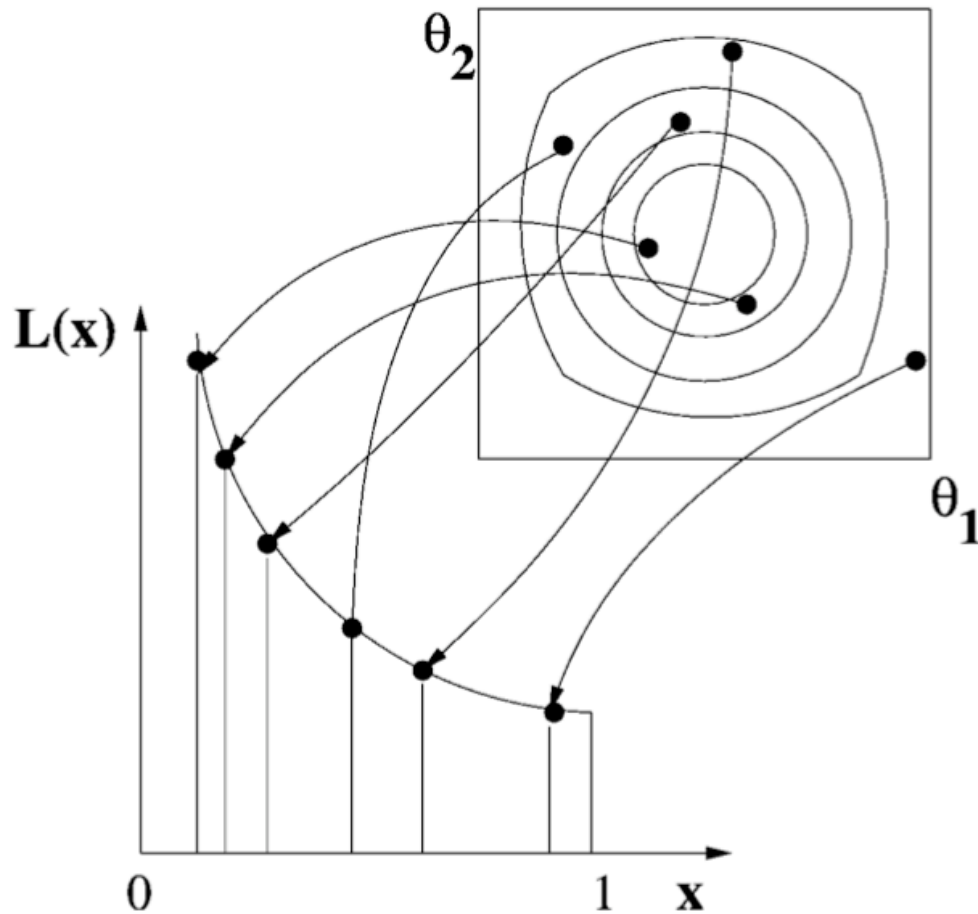
- Do the data demand an additional parameter?

Expected Evidence: braneworld gravity?



Heavens, Kitching & Verde 2007

Computing the Evidence: Nested Sampling



Skilling (2004)

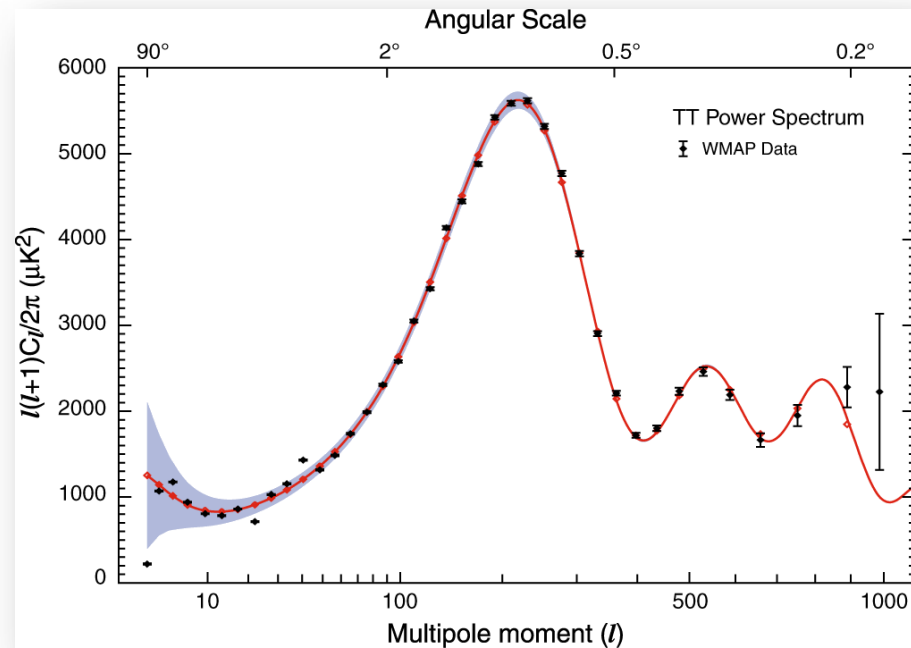
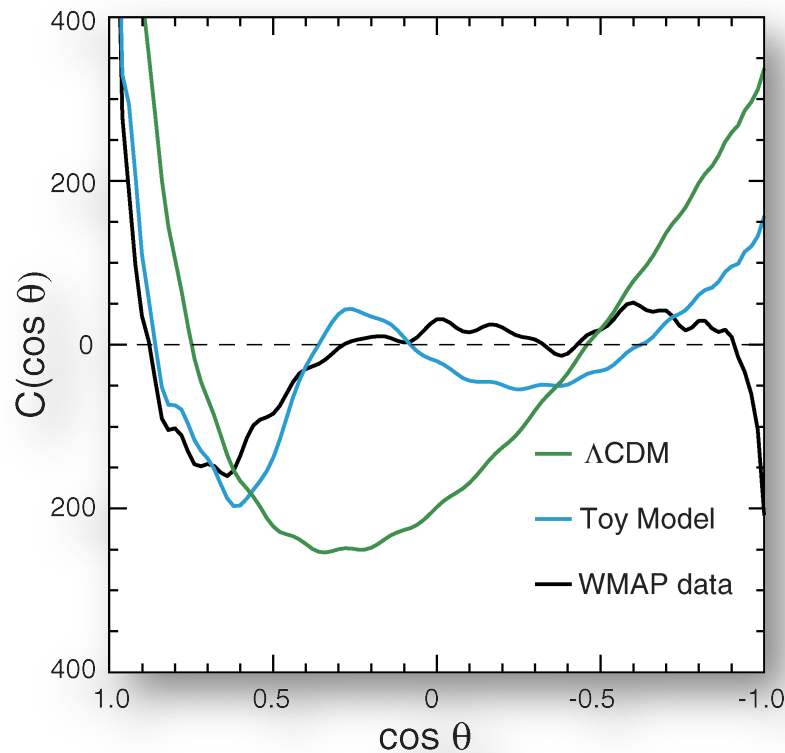
Sample from the prior volume, replacing the lowest point with one from a higher target density.

See: CosmoNEST (add-on for CosmoMC)

Multimodal? MultiNEST

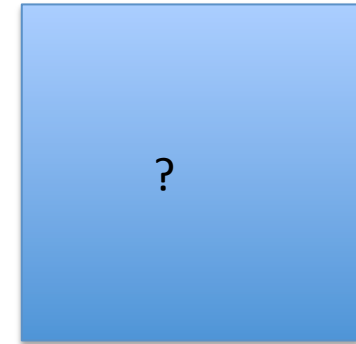
Back to WMAP

- Correlation function points are *highly correlated*; power spectrum points are not



An exercise in using Bayes' theorem

You choose
this one

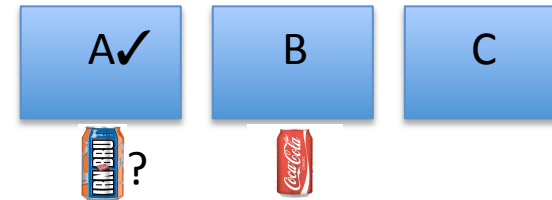


Do you change your choice?

This is the Monty Hall problem

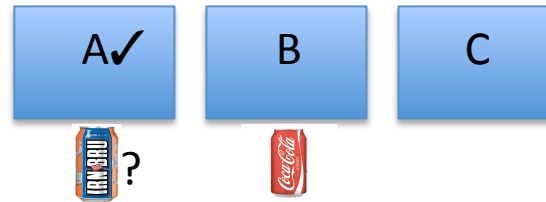


Monty Hall solution



- Rule 1: write down what it is you want
- Let a =Irn Bru is behind Door A (b, c similarly)
- Let B =Monty Hall opened Door B
- It is $p(a | B)$
- Now $p(a | B) = p(B | a)p(a)/p(B)$
- Evaluate $p(B) = p(B, a) + p(B, b) + p(B, c)$ (marginalisation)
 - $p(B) = p(B | a)p(a) + p(B | b)p(b) + p(B | c)p(c)$
 - $p(B) = (\frac{1}{2} \times \frac{1}{3}) + (0 \times \frac{1}{3}) + (1 \times \frac{1}{3}) = \frac{1}{2}$
- $p(a | B) = \frac{1}{2} \times \frac{1}{3} / \frac{1}{2} = \frac{1}{3}$ i.e. BETTER TO CHANGE

The one line reason (well, 3 lines)



- If you got it *right* first time, you'll get it *wrong* if you change
- If you got it *wrong* first time, you'll get it *right* if you change
- And *you are more likely to have got it wrong first time*