Parameter estimation and forecasting

Cristiano Porciani

AIfA, Uni-Bonn

Questions?

Cosmological parameters

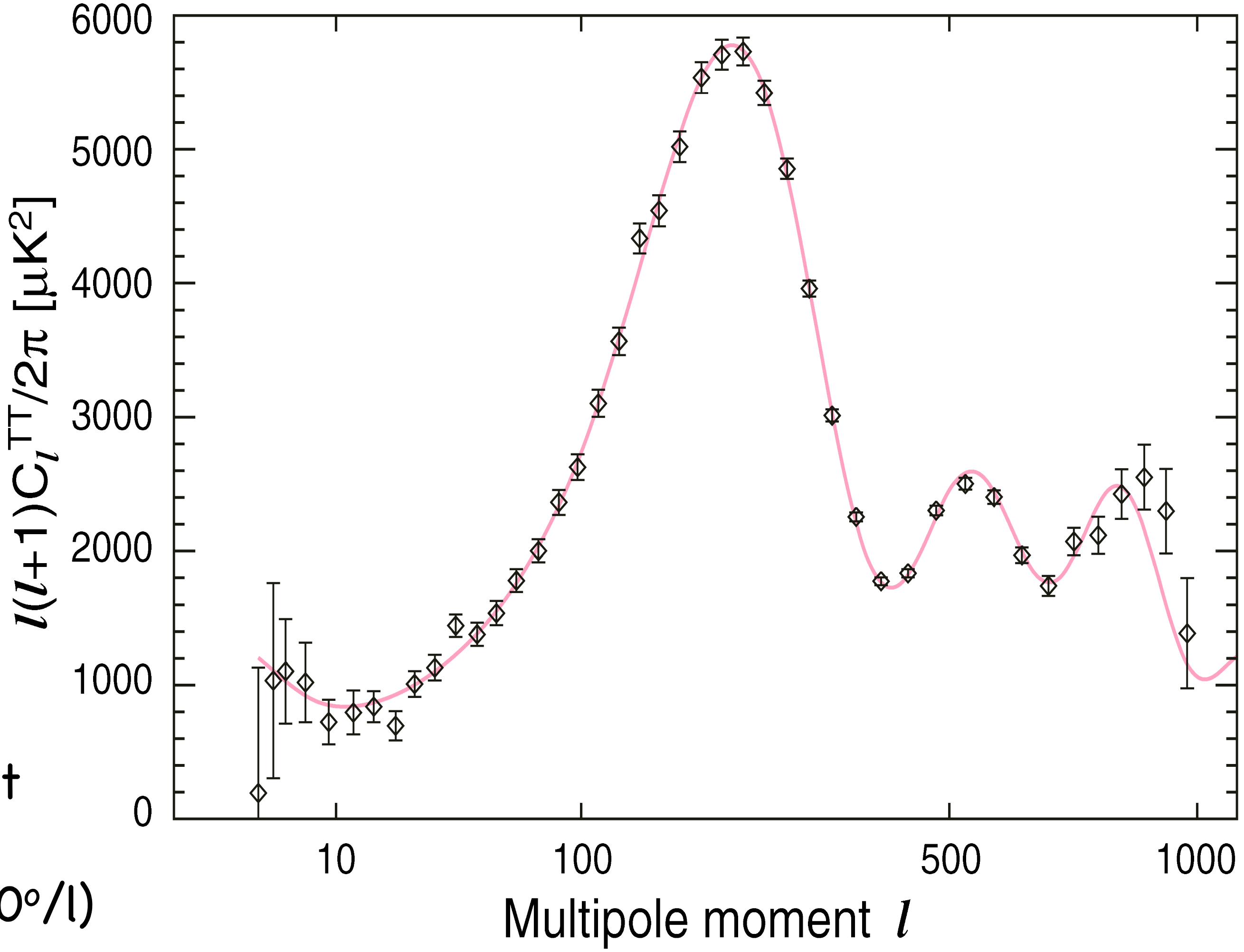
- A branch of modern cosmological research focuses on measuring cosmological parameters from observed data (e.g. the Hubble constant, the cosmic density of matter, etc.).

- In this class we will review the main techniques used for model fitting (i.e. extracting information on cosmological parameters from existing observational data) and forecasting (i.e. predicting the uncertainty on the parameters when future experiments will become available). The latter is a crucial ingredient for optimizing experimental design.

Key problems

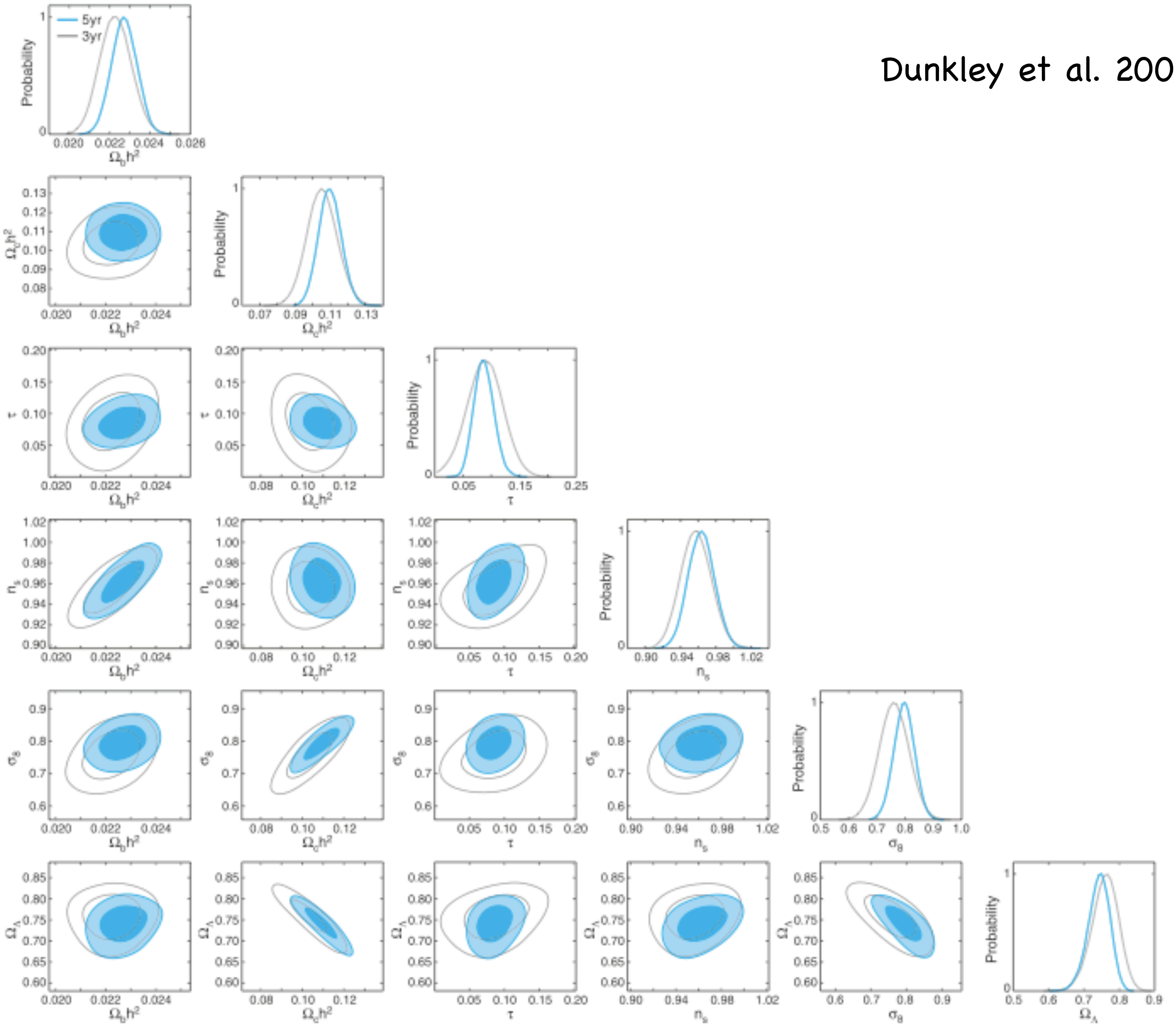
- How do you fit a model to data?
- How do you incorporate prior knowledge?
- How do you merge multiple sources of information?
- How do you treat uncertainties in model parameters?

Example: power spectrum of CMB temperature fluctuations

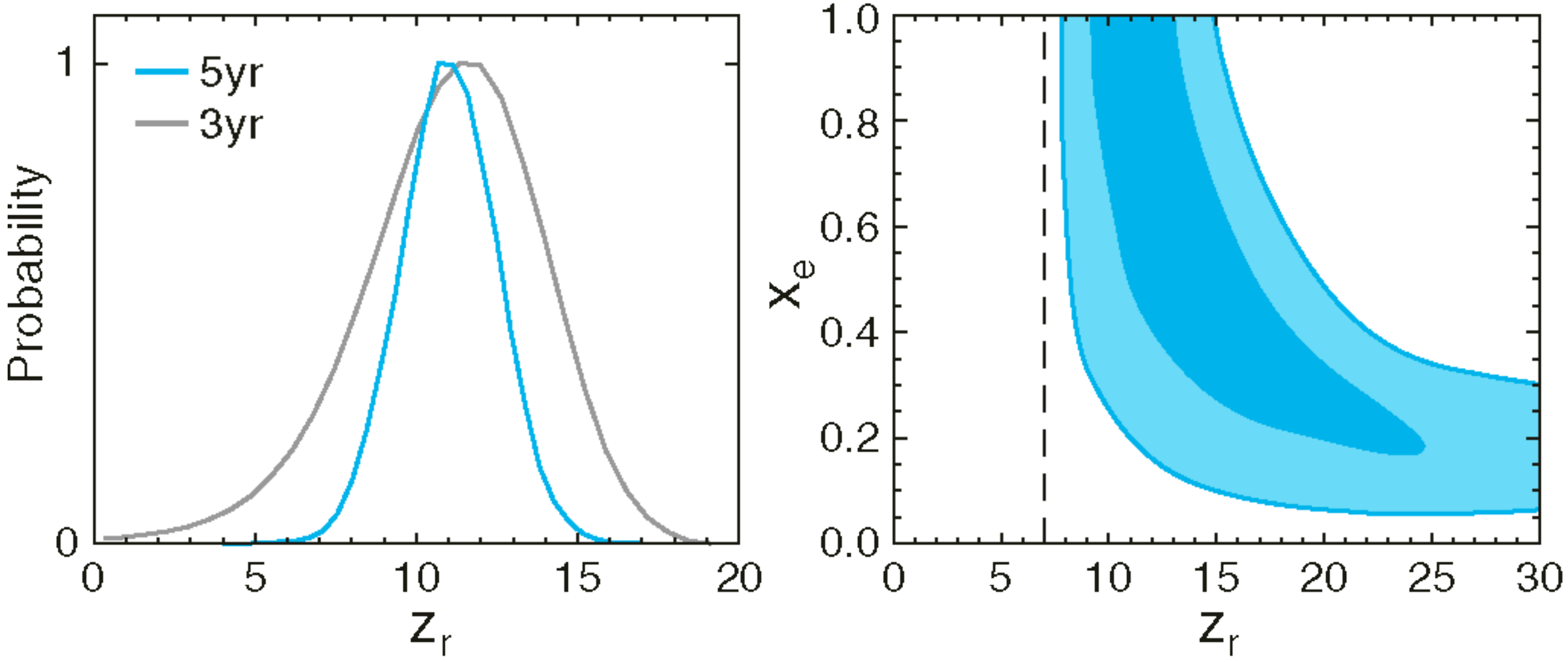


Variance at multipole l (angle ~180º/l)

Dunkley et al. 2009

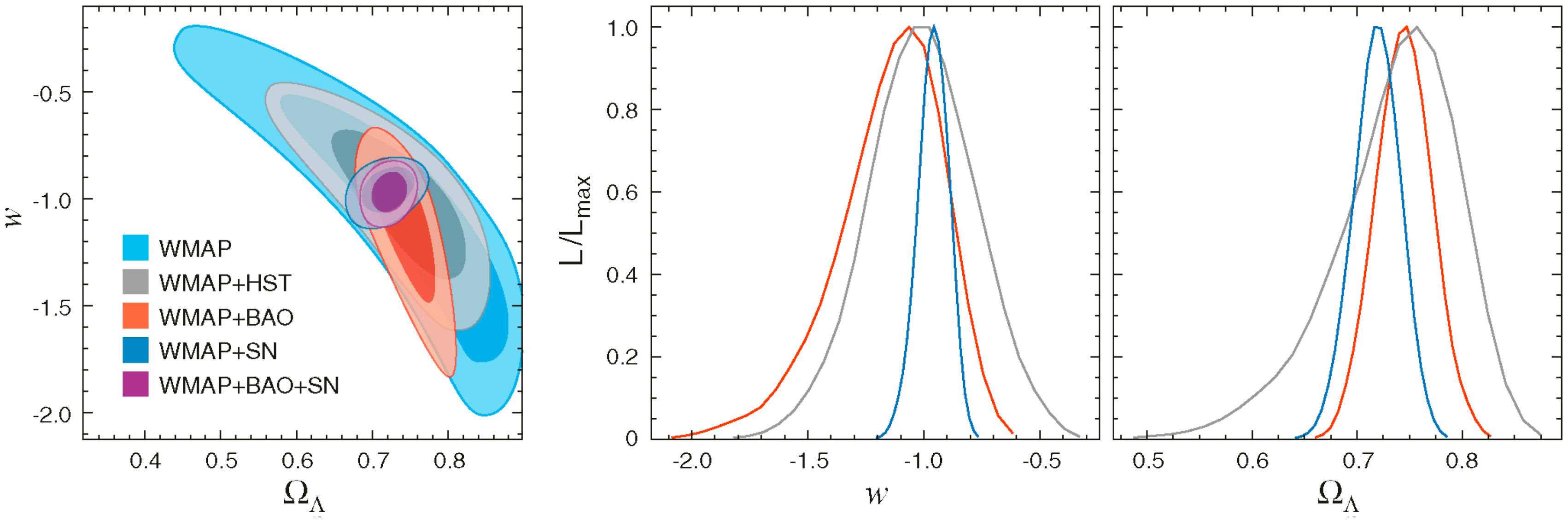


Dunkley et al. 2009



The current state of the art

Bennett 2006



What is the meaning of these plots?

- What's the difference between the 1D and the 2D plots?

- What is a confidence interval?

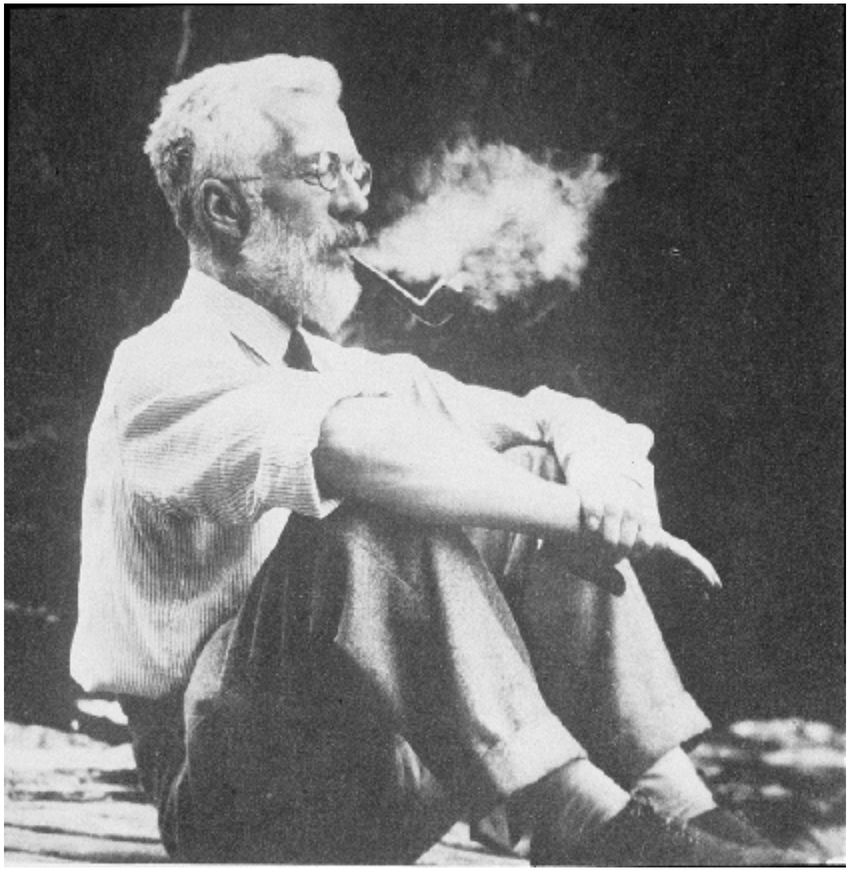
- What is a credibility interval?

- What does marginalisation mean?

- What's the difference between the frequentist and the Bayesian interpretation of statistics?

R.A. Fisher (1890-1962)

“Fisher was to statistics what Newton was to Physics” (R. Kass)



“Even scientists need their heroes, and R.A. Fisher was the hero of 20th century statistics” (B. Efron)

Fisher's concept of likelihood

- "Two radically distinct concepts have been confused under the name of 'probability' and only by sharply distinguishing between these can we state accurately what information a sample does give us respecting the population from which it was drawn." (Fisher 1921)

- "We may discuss the probability of occurrence of quantities which can be observed…in relation to any hypotheses which may be suggested to explain these observations. We can know nothing of the probability of the hypotheses…We may ascertain the likelihood of the hypotheses…by calculation from observations:…to speak of the likelihood…of an observable quantity has no meaning." (Fisher 1921)

- "The likelihood that any parameter (or set of parameters) should have any assigned value (or set of values) is proportional to the probability that if this were so, the totality of observations should be that observed." (Fisher 1922)

Probability of the data versus likelihood of the parameters

- Suppose you are counting how many cars pass in front of your window on Sundays between 9:00 and 9:02 am. Counting experiments are generally well described by the Poisson distribution. Therefore, if the mean counts are λ, the probability of counting n cars follows the distribution:

$$P(n\,|\,\lambda)=\frac{\lambda^n e^{-\lambda}}{n!}$$

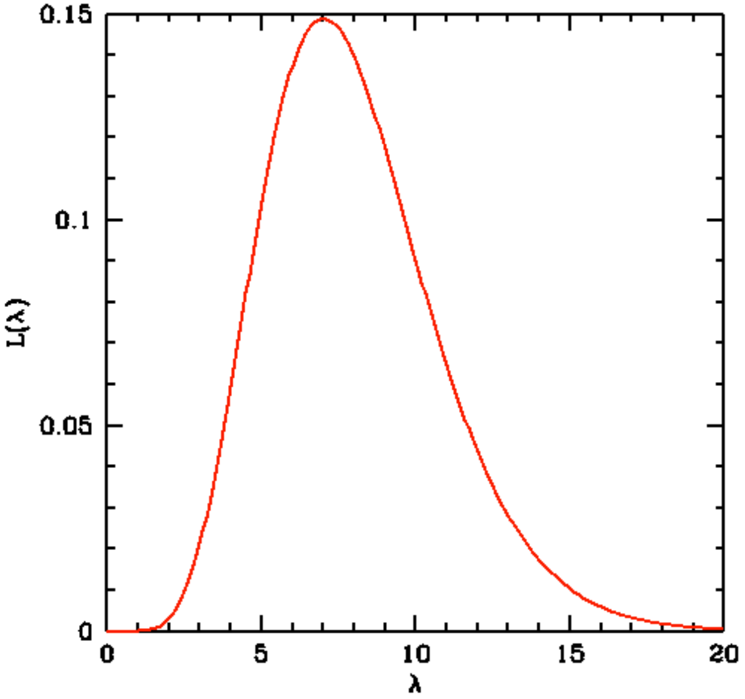
- This means that if you repeat the experiment many times, you will measure different values of n following the frequency P(n). Note that the sum over all possible n is unity.

- Now suppose that you actually perform the experiment once and you count 7. Then, the likelihood for the model parameter λ GIVEN the data is:

$$L(\lambda)=P(7\,|\,\lambda)=\frac{\lambda^7 e^{-\lambda}}{5040}$$

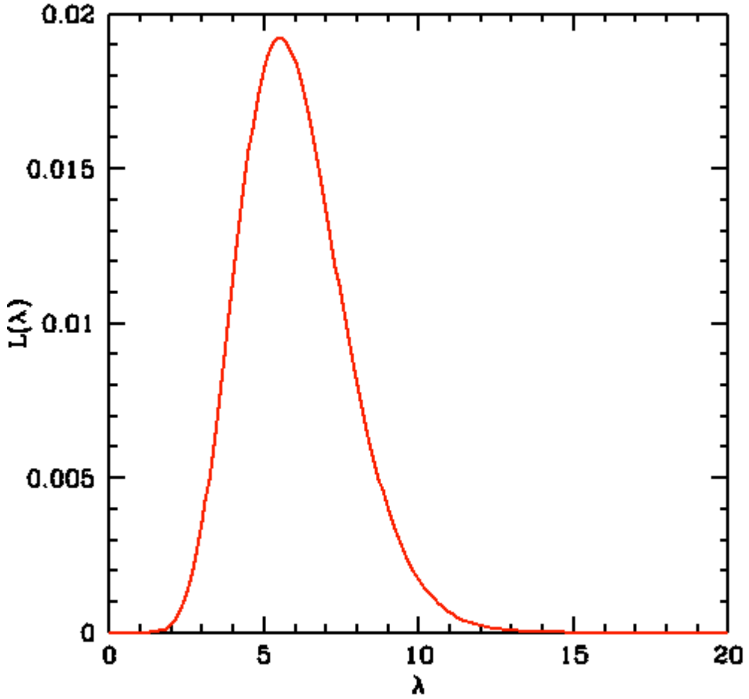
The likelihood function

- This is a function of λ only but it is NOT a probability distribution for λ! It simply says how likely it is that our measured value of n=7 is obtained by sampling a Poisson distribution of mean λ. It says something about the model parameter GIVEN the observed data.



The likelihood function

- Let us suppose that after some time you repeat the experiment and count 4 cars. Since the two experiments are independent, you can multiply the likelihoods and obtain the curve below. Note that now the most likely value is λ=5.5 and the likelihood function is narrower than before, meaning that we know more about λ.



Likelihood for Gaussian errors

- Often statistical measurement errors can be described by Gaussian distributions. If the errors σ_i of different measurements d_i are independent:

$$L(\theta) = P(d\,|\,\theta) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left[-\frac{\left(d_i - m_i(\theta)\right)^2}{2\sigma_i^2}\right]$$

$$-\ln L(\theta) = \sum_{i=1}^{N} \frac{(d_i - m_i(\theta))^2}{2\sigma_i^2} + \text{const.} = \frac{\chi^2(\theta,d)}{2} + \text{const.}$$

- Maximizing the likelihood corresponds to finding the values of the parameters $\theta = \{\theta_1,\ldots,\theta_n\}$ which minimize the χ^2 function (weighted least squares method).

The general Gaussian case

- In general, errors are correlated and

$$-\ln L(\theta) = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left[d_i - m_i(\theta)\right] C_{ij}^{-1} \left[d_j - m_j(\theta)\right] + \text{const.} = \frac{\chi^2(\theta,d)}{2} + \text{const.}$$

where $C_{ij}=<\varepsilon_i \ \varepsilon_j>$ is the covariance matrix of the errors.

- For uncorrelated errors the covariance matrix is diagonal and one reduces to the previous case.

- Note that the covariance matrix could also derive from a model and then depend on the model parameters. We will encounter some of these cases in the rest of the course.

The Likelihood function: a summary

- In simple words, the likelihood of a model given a dataset is proportional to the probability of the data given the model
- The likelihood function supplies an order of preference or plausibility of the values of the free parameters θ_i by how probable they make the observed dataset
- The likelihood ratio between two models can then be used to prefer one to the other
- Another convenient feature of the likelihood function is that it is functionally invariant. This means that any quantitative statement about the θ_i implies a corresponding statements about any one to one function of the θ_i by direct algebraic substitution

Maximum Likelihood

- The likelihood function is a statistic (i.e. a function of the data) which gives the probability of obtaining that particular set of data, given the chosen parameters θ_1, ... , θ_k of the model. It should be understood as a function of the unknown model parameters (but it is NOT a probability distribution for them)

- The values of these parameters that maximize the sample likelihood are known as the Maximum Likelihood Estimates or MLE's.

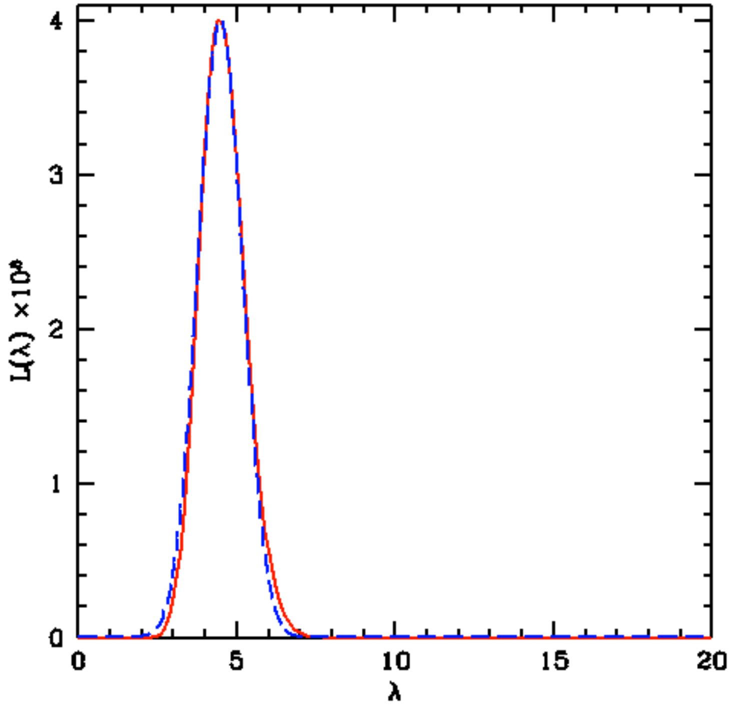
- Assuming that the likelihood function is differentiable, estimation is done by solving

$$\frac{\partial L(\theta_1,...,\theta_k)}{\partial \theta_i}=0 \qquad \text{or} \qquad \frac{\partial \ln L(\theta_1,...,\theta_k)}{\partial \theta_i}=0$$

- On the other hand, the maximum value may not exists at all.

Back to counting cars

- After 9 experiments we collected the following data: 7, 4, 2, 6, 4, 5, 3, 4, 5. The new likelihood function is plotted below, together with a Gaussian function (dashed line) which matches the position and the curvature of the likelihood peak (λ =4.44). Note that the 2 curves are very similar (especially close to the peak), and this is not by chance.



Score and information matrix

- The first derivative of the log-likelihood function with respect to the different parameters is called the Fisher score function:

$$S_i = \frac{\partial \ln L(\theta)}{\partial \theta_i}$$

- The Fisher score vanishes at the MLE.

- The negative of the Hessian matrix of the log-likelihood function with respect to the different parameters is called the observed information matrix:

$$O_{ij} = -\frac{\partial^2 \ln L(\theta)}{\partial \theta_i \, \partial \theta_j}$$

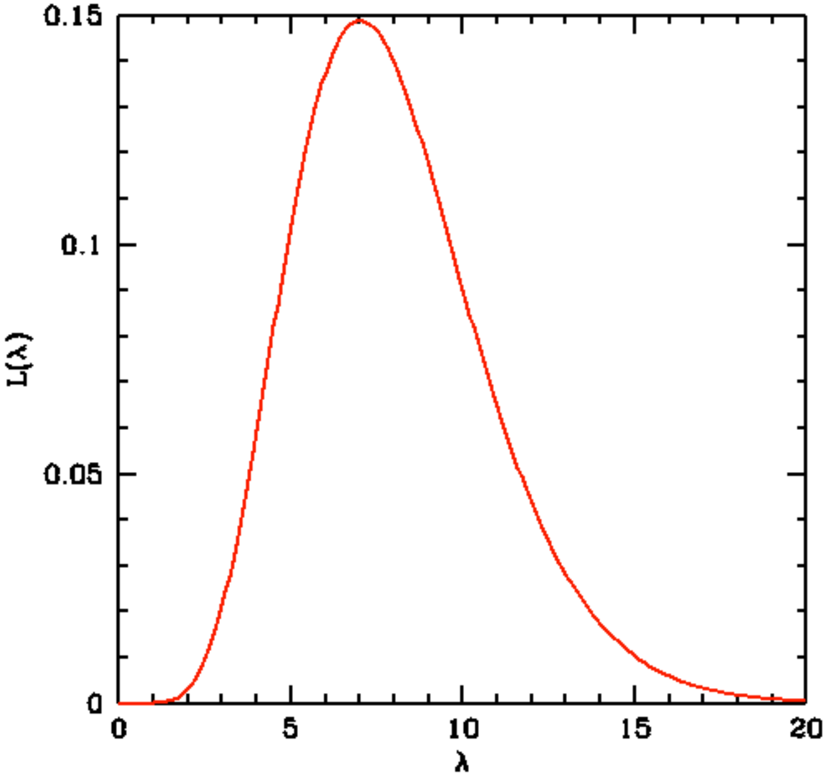
- The observed information matrix is definite positive at the MLE. Its elements tell us how broad is the likelihood function close to its peak and thus with what accuracy we determined the model parameters.

Example

1 datapoint

Low information

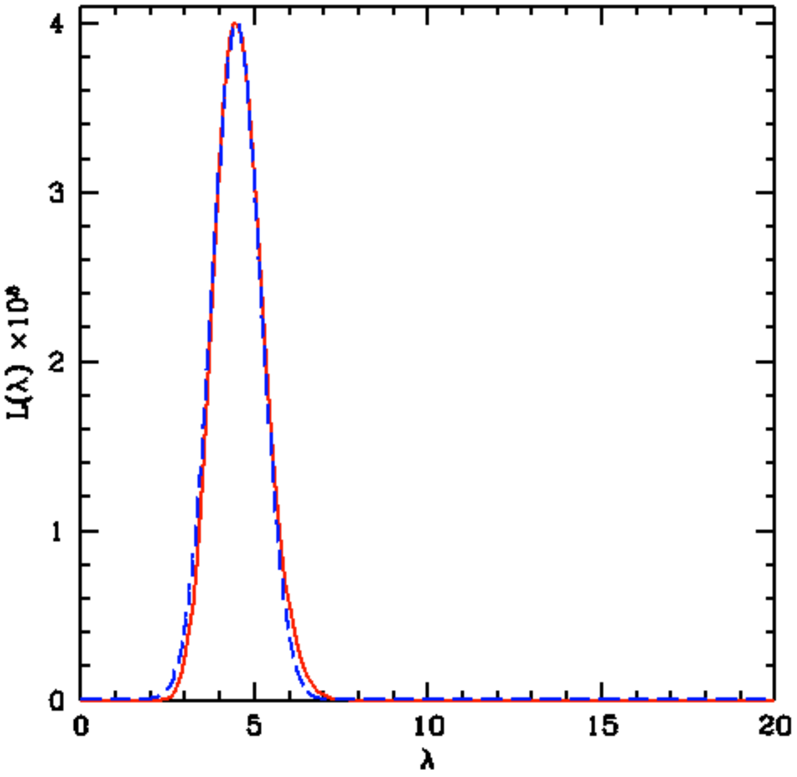
Large uncertainty in λ



9 datapoints

High information

Small uncertainty in λ



Fisher information matrix

- If we took different data, then the likelihood function for the parameters would have been a bit different and so its score function and the observed information matrix.

- Fisher introduced the concept of information matrix by taking the ideal ensemble average (over all possible datasets of a given size) of the observed information matrix (evaluated at the true value of the parameters).

$$F_{ij} = -\left\langle \frac{\partial^2 \ln L(\theta)}{\partial\theta_i \, \partial\theta_j} \right\rangle$$

- Under mild regularity conditions, it can been shown that the Fisher information matrix also corresponds to

$$F_{ij} = \left\langle \frac{\partial \;\ln L(\theta)}{\partial\theta_i} \frac{\partial \;\ln L(\theta)}{\partial\theta_j} \right\rangle$$

i.e. to the covariance matrix of the scores at the MLE's.

Cramér-Rao bound

- The Cramér-Rao bound states that, for ANY unbiased estimator of a model parameter θ_i, the measurement error (keeping the other parameters constant) satisfies

$$\Delta\theta_i \geq \frac{1}{\sqrt{F_{ii}}}$$

- For marginal errors that also account for the variability of the other parameters (see slide 35 for a precise definition), instead, it is the inverse of the Fisher information matrix that matters and

$$\Delta\theta_i \geq \sqrt{F^{-1}_{ii}}$$

Fisher matrix with Gaussian errors

- For data with Gaussian errors, the Fisher matrix assumes the form (the notation is the same as in slide 20)

$$F_{ij} = \frac{1}{2}\mathrm{Tr}\left[C^{-1}C_{,i}C^{-1}C_{,j} + C^{-1}M_{ij}\right]$$

where

Information from the noise

Information from the signal

$$M_{ij} = m_{,i}m_{,j}^{T} + m_{,j}m_{,i}^{T}$$

(note that commas indicate derivatives with respect to the parameters while data indices are understood)

Properties of MLE's

As the sample size increases to infinity (under weak regularity conditions):

- MLE's become asymptotically efficient and asymptotically unbiased
- MLE's asymptotically follow a normal distribution with covariance matrix (of the parameters) equal to the inverse of the Fisher's information matrix (that is determined by the covariance matrix of the data).

However, for small samples,

- MLE's can be heavily biased and the large-sample optimality does not apply

Maximizing likelihood functions

- For models with a few parameters, it is possible to evaluate the likelihood function on a finely spaced grid and search for its minimum (or use a numerical minimisation algorithm).

- For a number of parameters >>2 it is NOT feasible to have a grid (e.g. 10 point in each parameter direction, 12 parameters = 10^{12} likelihood evaluations!!!)

- Special statistical and numerical methods needs to be used to perform model fitting.

- Note that typical cosmological problems consider models with a number of parameters ranging between 6 and 20.

Forecasting

- Forecasting is the process of estimating the performance of future experiments for which data are not yet available

- It is a key step for the optimization of experimental design (e.g. how large must be my survey if I want to determine a particular parameter to 1% accuracy?)

- The basic formalism has been developed by Fisher in 1935

Figure of merit

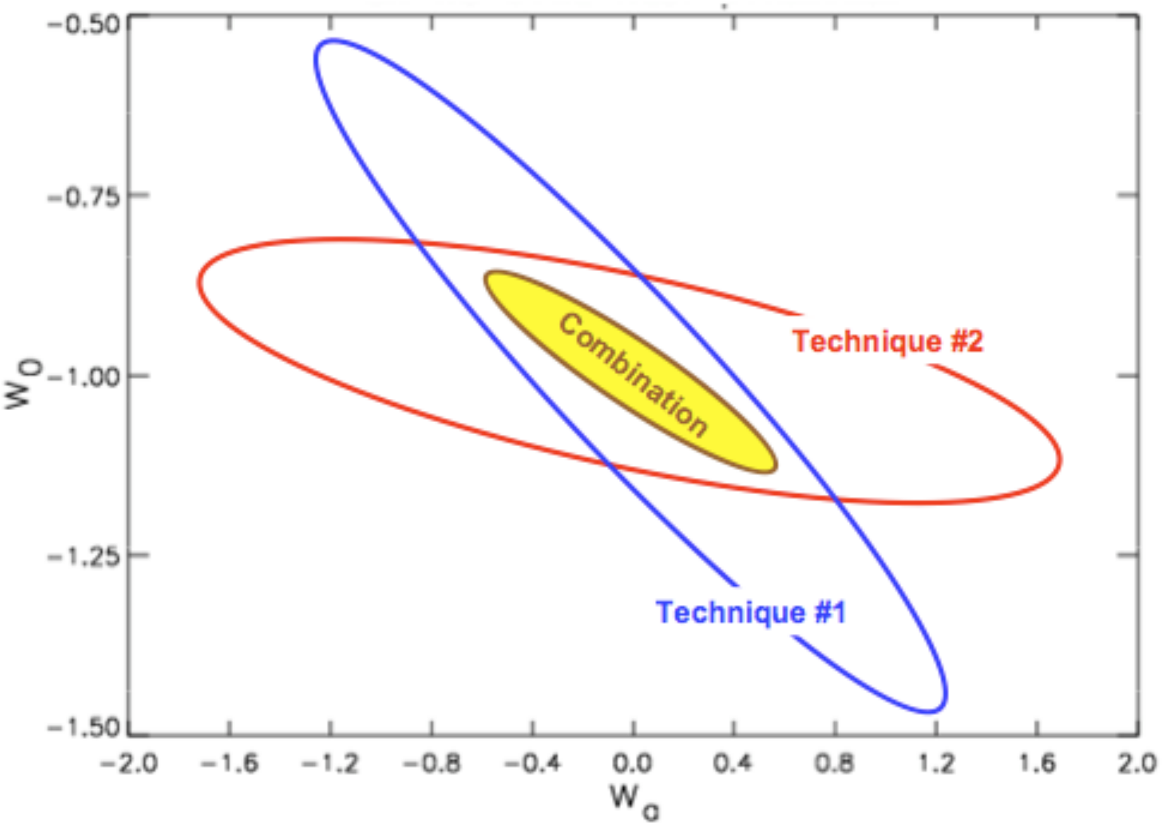
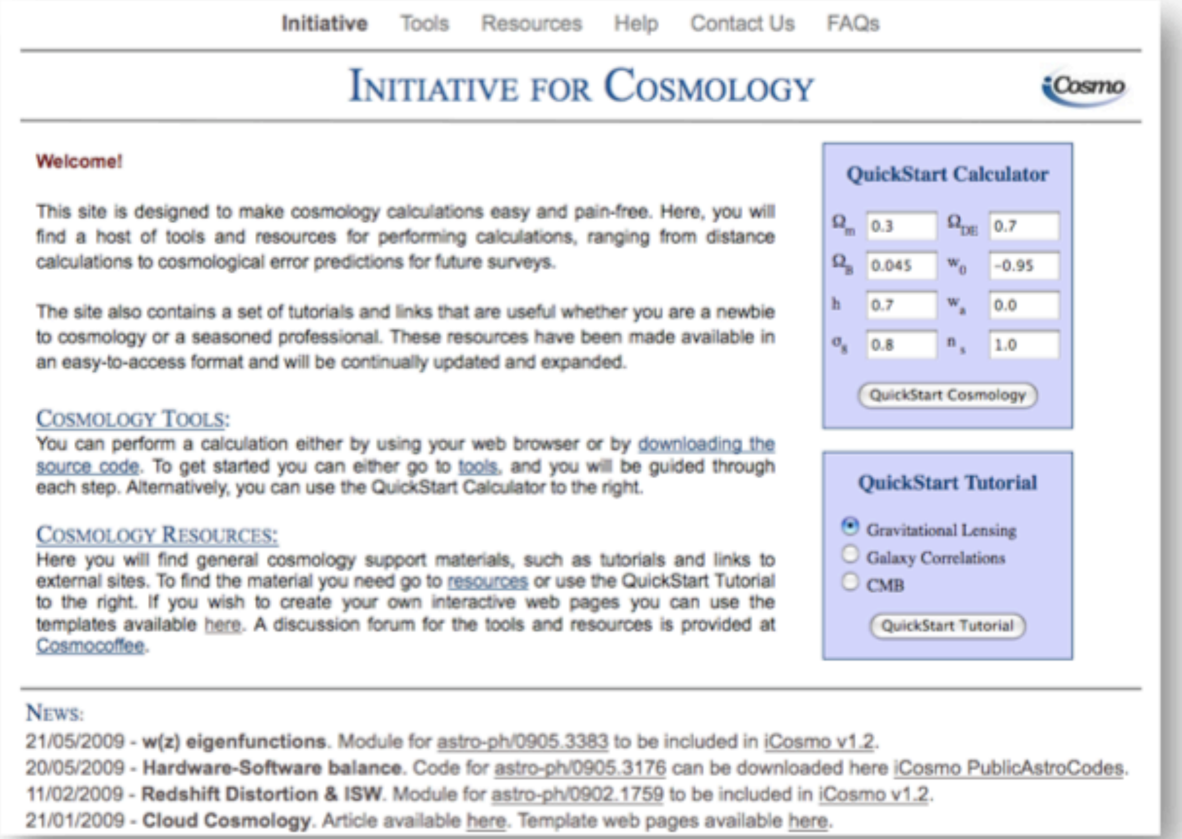


Figure of merit = 1 / (area of the ellipse)

iCOSMO.org



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 Cosmocoffee.

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

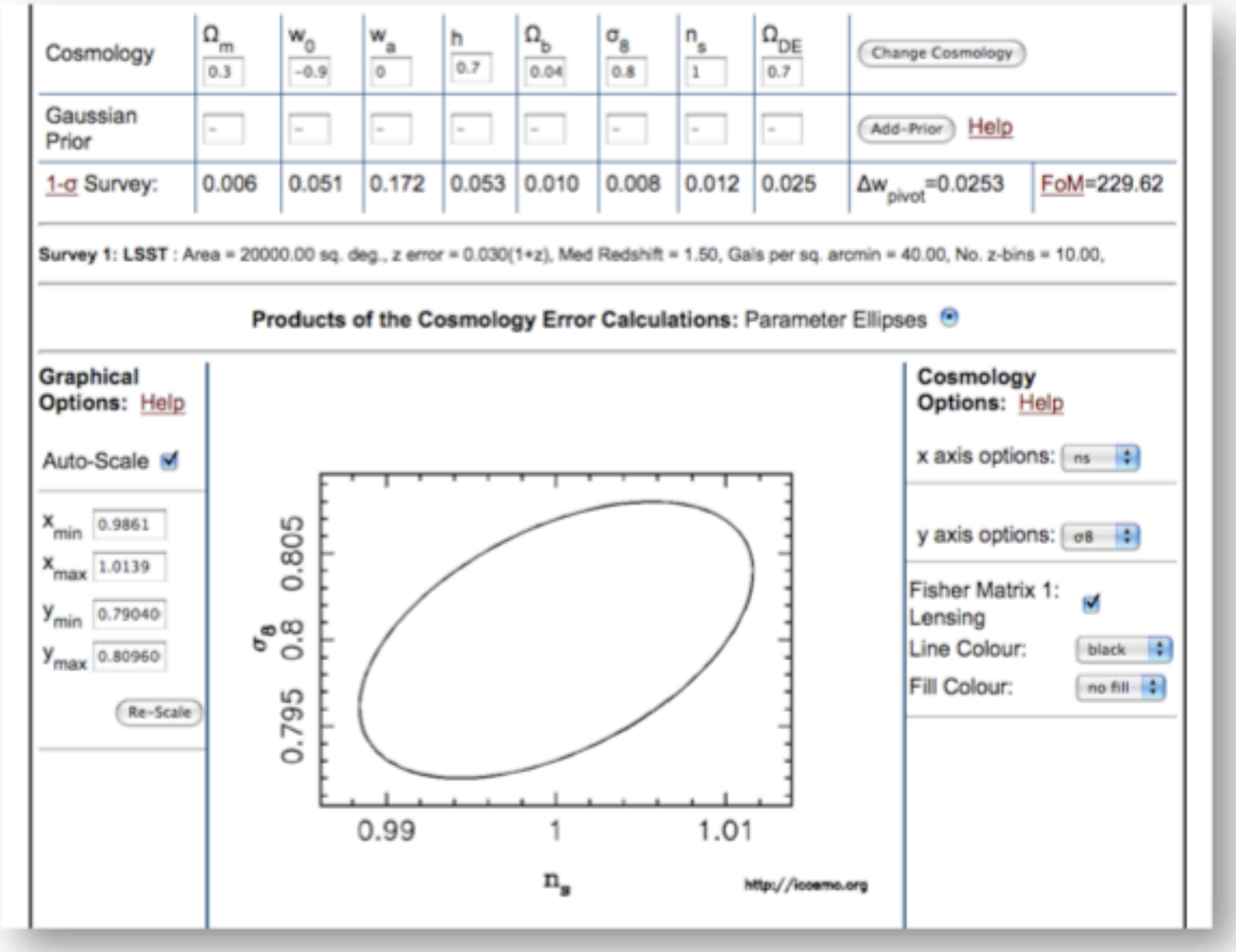
QuickStart Tutorial

● Gravitational Lensing
○ Galaxy Correlations
○ CMB

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.

Open source Fisher matrices



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	Δw_{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

Cosmology Options: Help

x axis options: ns

y axis options: σ8

Fisher Matrix 1: Lensing ☑

Line Colour: black

Fill Colour: no fill

Fisher 4cast (Matlab toolbox)

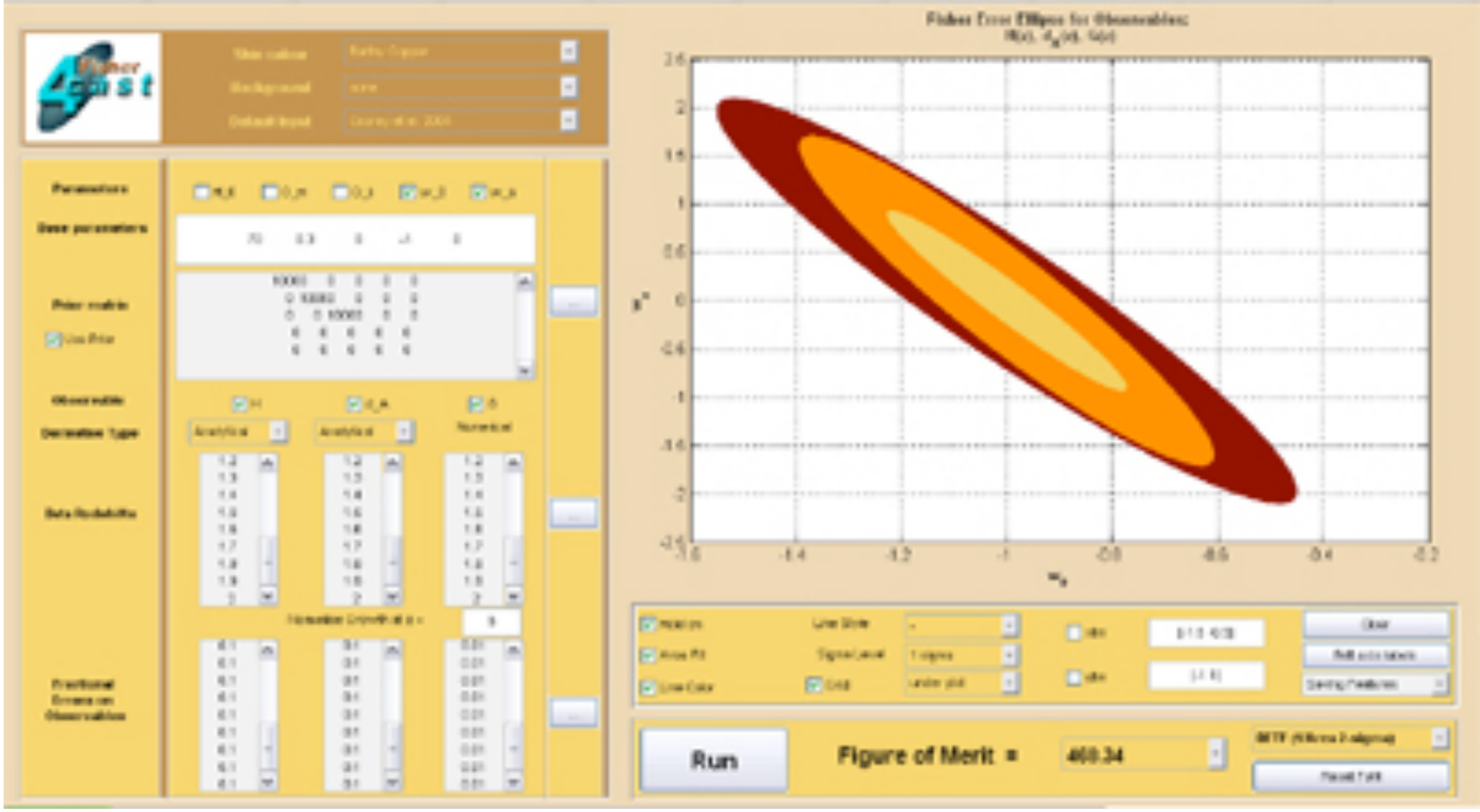


Figure of Merit = 468.34

Counting cars, again

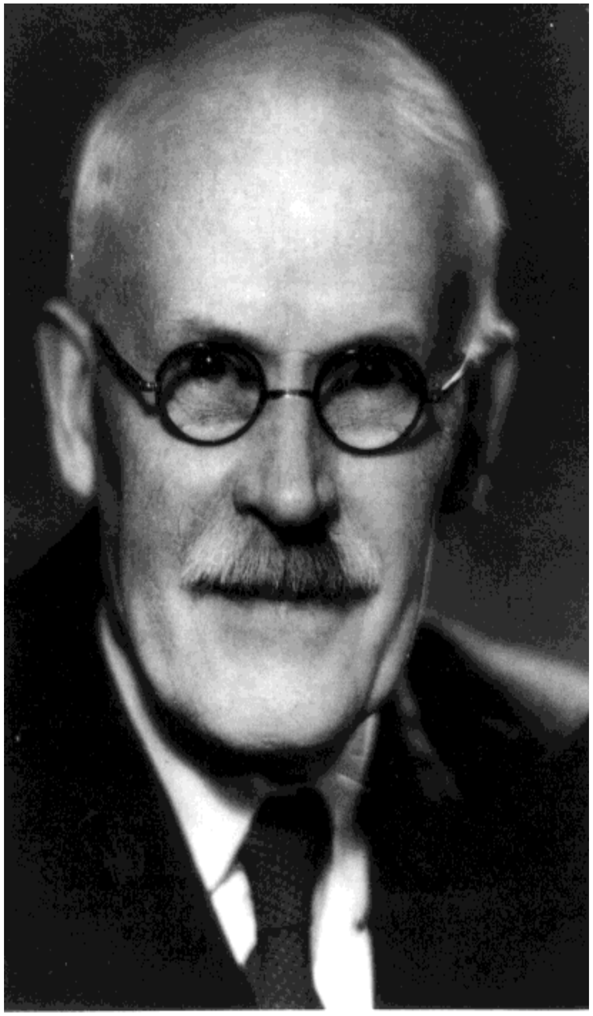
- In our study of the car counts we implicitly assumed that all the values of λ are equally likely a priori (i.e. before we started taking the data). However, we didn't consider that an automatic gate regulates the traffic in our street and does not allow more than 8 cars to enter every 10 minutes. Therefore λ cannot be larger than 8 and the likelihood derived from our counts should have been truncated at λ=8.

- Also, we live close to a church and whenever there is a wedding the traffic is more intense than usual. This means that on wedding days a higher value of λ is more likely than on non-wedding days.

- Moreover, a fellow that had been living in our flat before us did the same exercise and told us that he obtained λ=4.2±0.5.

- Is there a way to account for all this information in our study?

The Bayesian way



What is probability?

- Frequentist: the long-run expected frequency of occurrence of a random event

- Axiomatic: given a sample space Ω, a σ-algebra F of events E (a set of subsets of Ω), we call probability measure a real function on F such that P(E)≥0, P(Ω)=1, and for any countable series of pairwise disjoint events $P(E_1 \cup E_2 \cup \ldots \cup E_N)=P(E_1)+P(E_2)+\ldots+P(E_N)$. These are known as Kolmogorov axioms.

- Bayesian: a measure of the degree of belief (the plausibility of an event given incomplete knowledge)

Reasoning with beliefs

- There is 90% chance that today it will rain
- There is a 30% chance that my favourite football team will win the league this year
- There is a 10% chance that I will fail the observational cosmology examination
- There is a 0.1% chance that I will die before being 30
- There is 68.3% chance that H_0 lies between 67 and 73 km/s/Mpc

Rev. T. Bayes

Bayes theorem

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

Posterior probability for the parameters given the data

Prior probability for the parameters (what we know before performing the experiment)

Likelihood function

$$p(x \mid \theta) = L(x \mid \theta)$$

Evidence (normalization constant useful for Bayesian model selection)

$$p(x) = \int p(x \mid \theta)\, p(\theta)\, d\theta$$

Bayesian estimation

- In the Bayesian approach to statistics, population parameters are associated with a posterior probability which quantifies our DEGREE OF BELIEF in the different values
- Sometimes it is convenient to introduce estimators obtained by minimizing the posterior expected value of a loss function
- For instance one might want to minimize the mean square error, which leads to using the mean value of the posterior distribution as an estimator
- If, instead one prefers to keep functional invariance, the median of the posterior distribution has to be chosen
- Remember, however, that whatever choice you make is somewhat arbitrary as the relevant information is the entire posterior probability density.

Estimation: frequentist vs Bayesian

- Frequentist: there are TRUE population parameters that are unknown and can only be estimated by the data

- Bayesian: only data are real. The population parameters are an abstraction, and as such some values are more believable than others based on the data and on prior beliefs.

Confidence vs. credibility intervals

- Confidence intervals (Frequentist): measure the variability due to sampling from a fixed distribution with the TRUE parameter values. If I repeat the experiment many times, what is the range within which 95% of the results will contain the true values?

- Credibility interval (Bayesian): For a given significance level, what is the range I believe the parameters of a model can assume given the data we have measured?

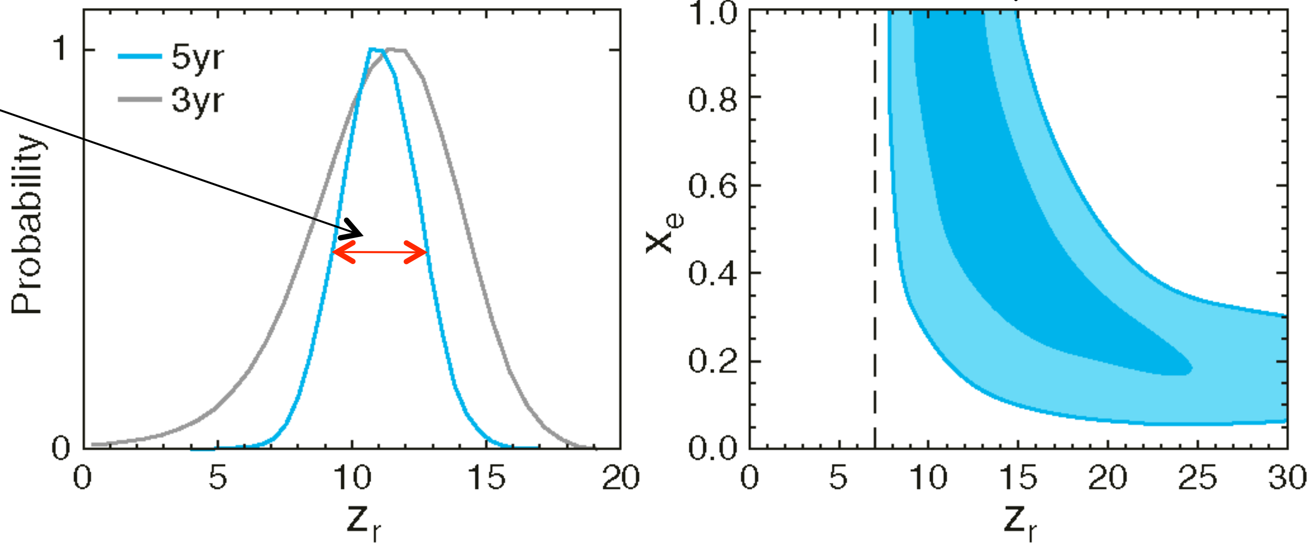
- They are profoundly DIFFERENT things even though they are often confused. Sometimes practitioners tend use the term “confidence intervals” in all cases and this is ok because they understand what they mean but this might be confusing for the less experienced readers of their papers. PAY ATTENTION!

Marginalisation

Marginal probability: posterior probability of a given parameter regardless of the value of the others. It is obtained by integrating the posterior over the parameters that are not of interest.

$$p(\vartheta_2 \,|\, x) = \int p(\theta \,|\, x)\; d\theta_1 d\theta_3 \ldots d\theta_n$$

Marginal errors characterise the width of the marginal posterior distributions.



Dunkley et al. 2009

How can we do this in practice?

Markov Chain Monte Carlo

Andrey Andreyevic Markov
(1856-1922)

Monte Carlo Casino
(1863-now)

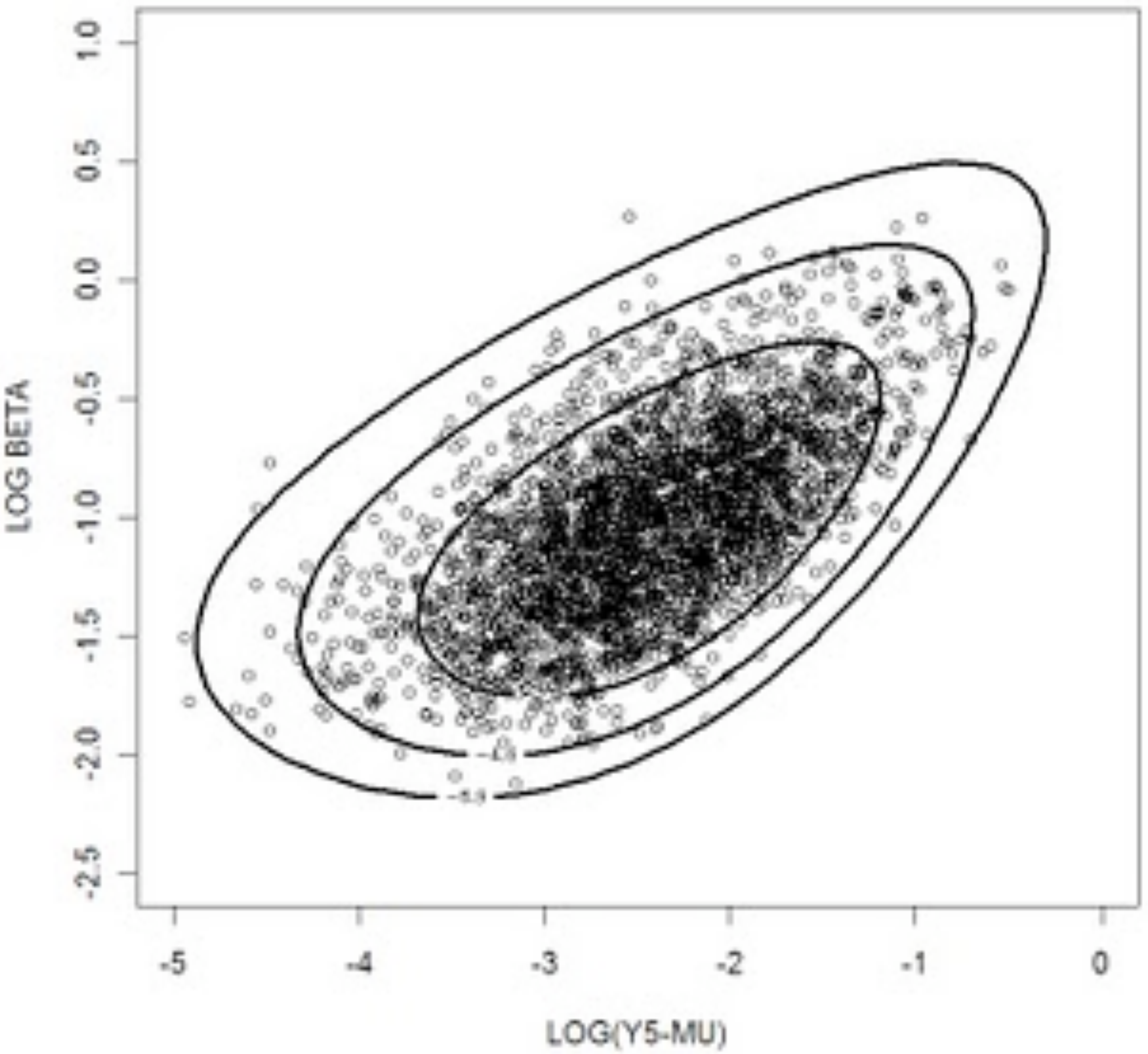
Markov Chain Monte Carlo

- WHAT? A numerical simulation method

- AIM: Sampling a given distribution function (known as the target density)

  i.e. generate a finite set of points in some parameter space that are drawn from a given distribution function.

- HOW? By building a Markov chain that has the desired distribution as its equilibrium distribution



1.0
0.5
0.0
-0.5
-1.0
-1.5
-2.0
-2.5
LOG BETA
-5
-4
-3
-2
-1
0
LOG(Y5-MU)

Markov chains

- A Markov chain is a sequence of random variables (or vectors) X_i (where i is an integer index: i=0,…,N) with the property that the transition probability

$$P(x_{N+1} \mid x_0,\ldots,x_N) = P(x_{N+1} \mid x_N)$$

This means that the future of the chain does not depend on the entire past but only on the present state of the process.

Monte Carlo

- The term Monte Carlo method refers, in a very general term, to any numerical simulation which uses a computer algorithm explicitly dependent on a series of (pseudo) random numbers

- The idea of Monte Carlo integration was first developed by Enrico Fermi in the 1930s and by Stanislaw Ulam in 1947

$$\int f(x)p(x)dx \approx \frac{1}{N}\sum_{i=1}^{N} f(x_i) \quad [\text{where the } x_i \text{ are samples from } p(x)]$$

- Ulam and von Neumann used it for classified work at Los Alamos and as a “code name” for the project chose “Monte Carlo” as a reference to the famous Casino in Monaco.

MCMC and Bayesian statistics

- The MCMC method has been very successful in modern Bayesian computing.
- In general (with very few exceptions) posterior densities are too complex to work with analytically.
- With the MCMC method, it is possible to generate samples from an arbitrary posterior density and to use these samples to approximate expectations of quantities of interest.
- Most importantly, the MCMC is guaranteed to converge to the target distribution under rather broad conditions, regardless of where the chain was initialized.
- Furthermore, if the chain is run for very long time (often required) you can recover the posterior density to any precision.
- The method is easily applicable to models with a large number of parameters (although the "curse of dimensionality" often causes problems in practice).

MCMC algorithm

- Choose a random initial starting point in parameter space, and compute the target density

- Repeat:

✓ Generate a step in parameter space from a proposal distribution, generating a new trial point for the chain.

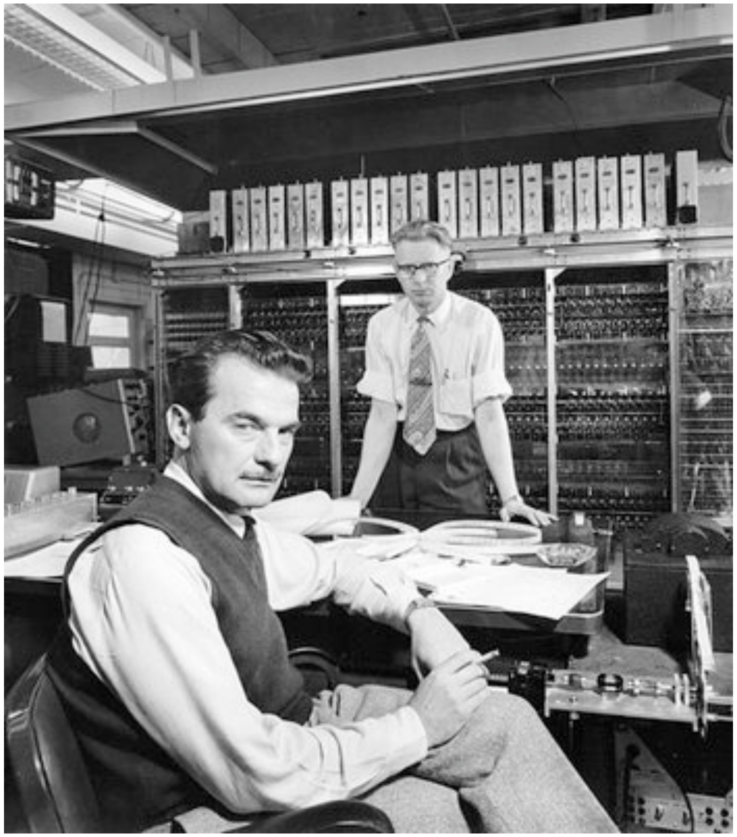
✓ Compute the target density at the new point, and accept it or not with the Metropolis-Hastings algorithm (see next slide).

✓ If the point is not accepted, the previous point is repeated in the chain.

- End Repeat

The Metropolis algorithm

Nicholas Constantine Metropolis (1915-1999)



"Equation of state calculation by fast computing machines"
Metropolis et al. (1953)

- After generating a new MCMC sample using the proposal distribution, calculate

$$r = \text{probability of acceptance} = \min\left(\frac{f(\theta_{new})}{f(\theta_{old})}, 1\right)$$

- Then sample u from the uniform distribution U(0,1)
- Set $\theta_{t+1} = \theta_{new}$ if u<r; otherwise set $\theta_{t+1} = \theta_t$
- Note that the number of iterations keeps increasing regardless of whether a proposed sample is accepted.

The Metropolis algorithm

- It can be demonstrated that the Metropolis algorithm works.

- The proof is beyond the scope of this course but, if you are curious, you can check standard statistics textbooks including Roberts (1996) and Liu (2001).

- You are not limited to a symmetric random-walk proposal distribution in establishing a valid sampling algorithm. A more general form, now known as the Metropolis-Hastings algorithm, was proposed by Hastings (1970). In this case:

$$r = \text{probability of acceptance} = \min\left(\frac{f(\theta_{new})q(\theta_t \mid \theta_{new})}{f(\theta_{old})q(\theta_{new} \mid \theta_t)}, 1\right)$$

The proposal distribution

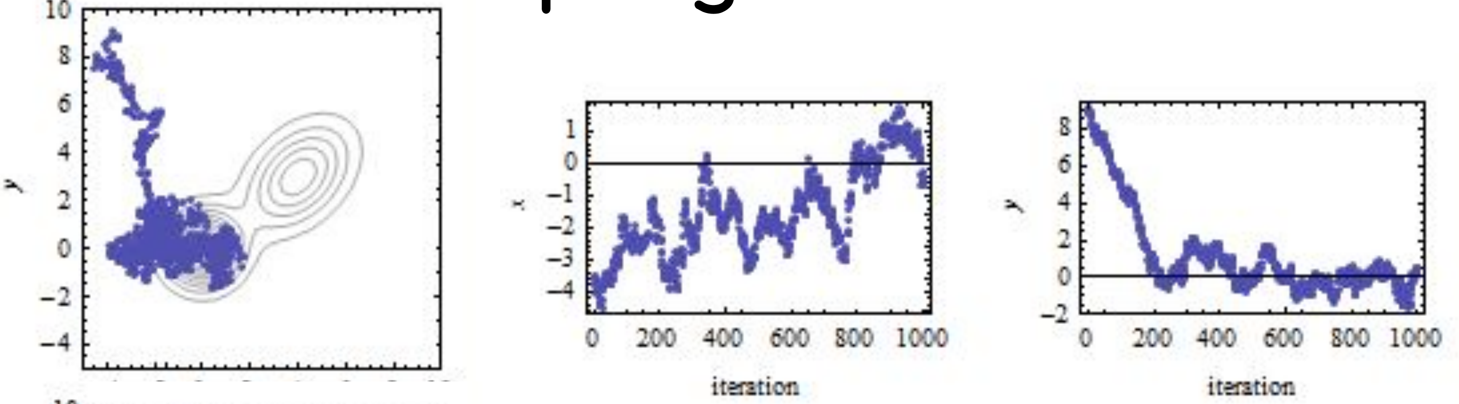
- If one takes too small steps, it takes long time to explore the target and the different entries of the chain are very correlated

- If one takes too large steps, almost all trials are rejected and the different entries of the chain are very correlated

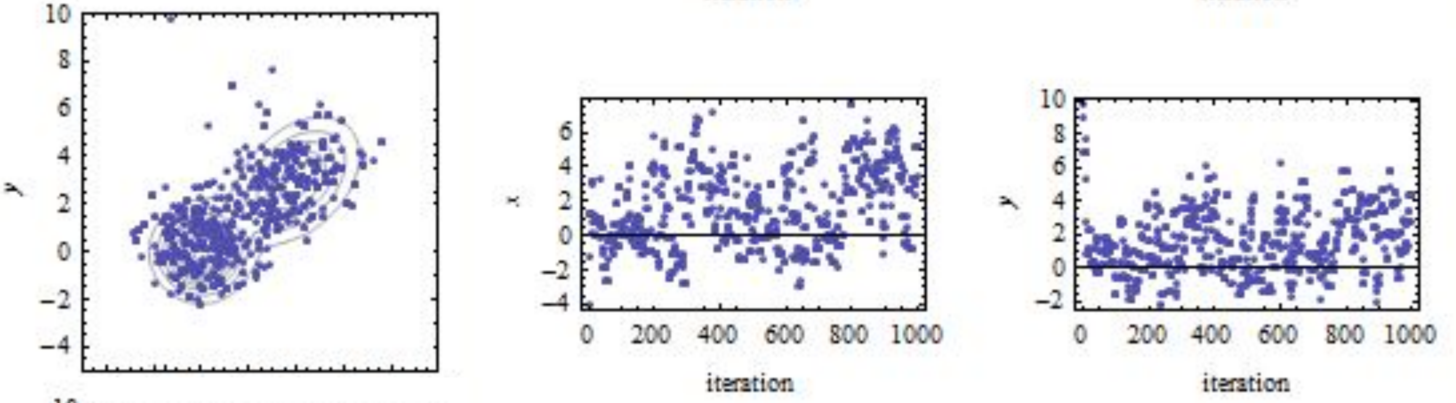
- There is an optimal proposal distribution (easy to identify if we knew already the target density)

Effect of the sampling distribution

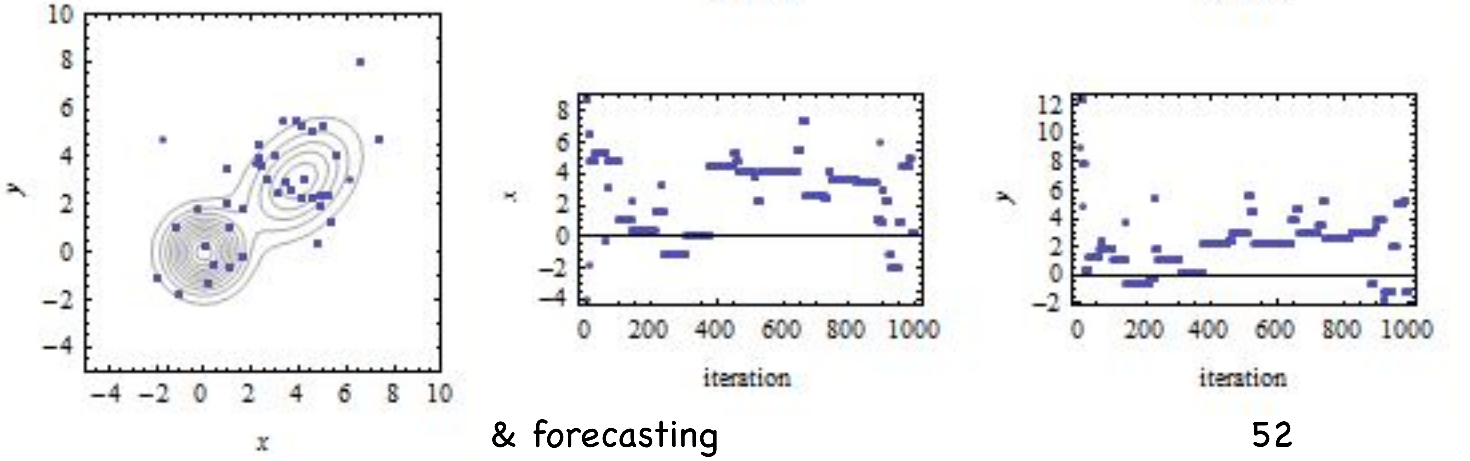
Gaussian proposal distrbution with $\sigma = 0.2$, acceptance rate =85.1%



Gaussian proposal distrbution with $\sigma = 2.2$, acceptance rate =37.9%



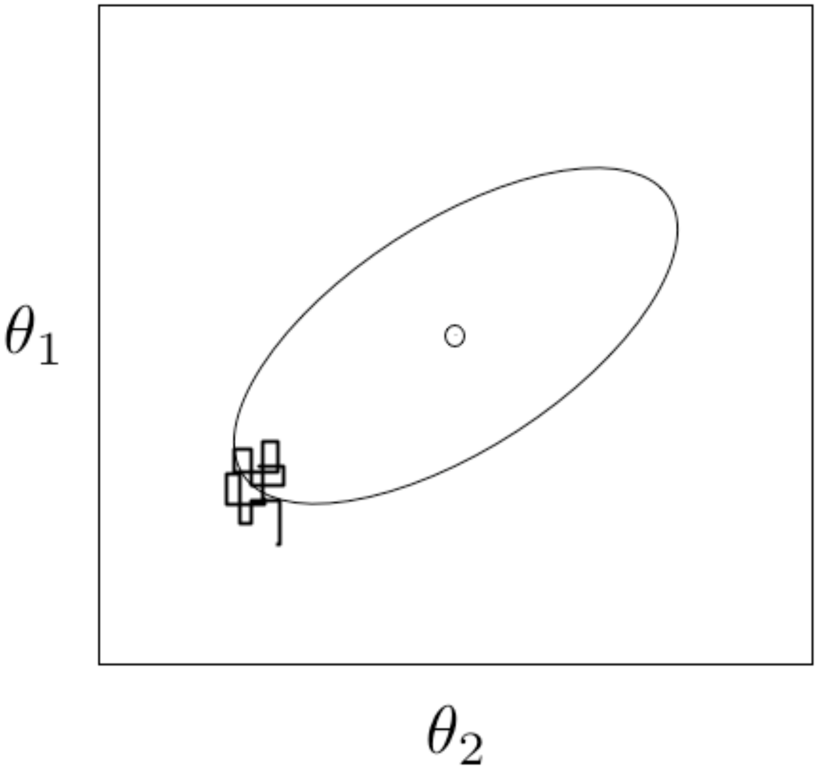
Gaussian proposal distrbution with $\sigma = 10.2$, acceptance rate =4.1%



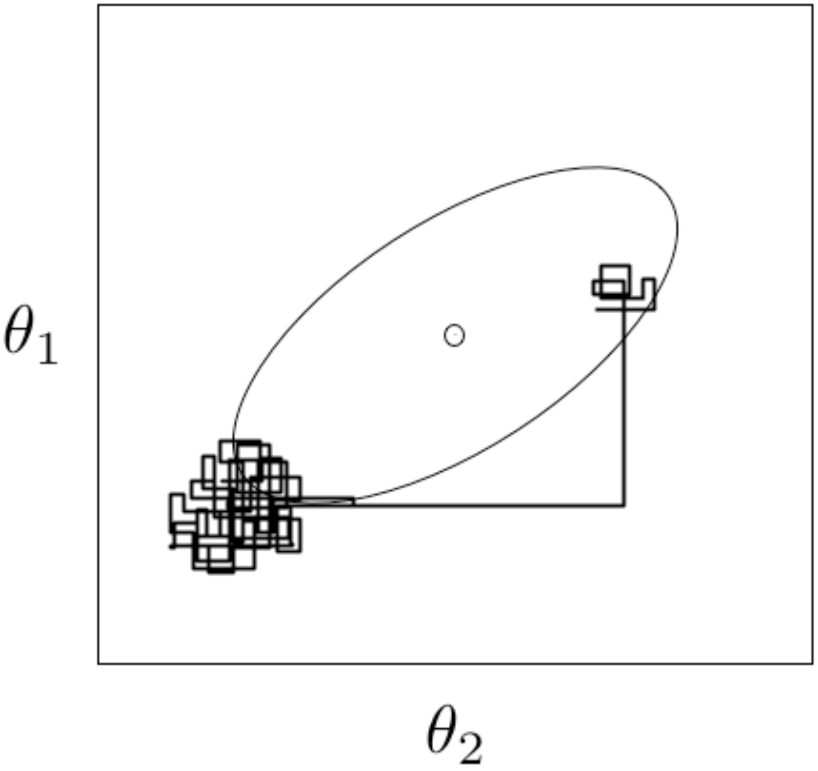
Mixing

Mixing refers to the degree to which the Markov chain explores the support of the posterior distribution. Poor mixing may stem from inappropriate proposals (if one is using the Metropolis-Hastings sampler) or from attempting to estimate models with highly correlated variables.

Bad mixing



Metastability



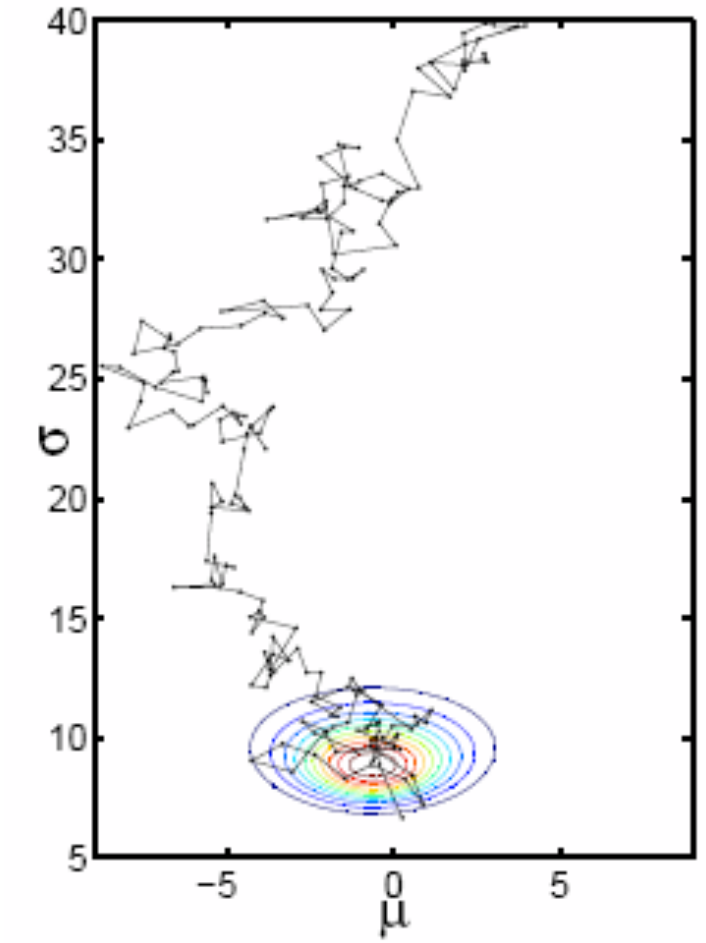
Burn-in

- Mathematical theorems guarantee that the Metropolis algorithm will asymptotically converge to the target distribution independently of its starting point.

- However, there will be an initial transient of unknown length during which the chain reaches its stationary state.

- In practice, you have to assume that after N_b iterations, the chain converged and started sampling from its target distribution.

- The value of N_b is called the burn-in number.



Issues with MCMC

- You have to decide whether the Markov Chain has reached its stationary distribution
- You have to decide the number of iterations to keep after the Markov Chain has reached stationarity
- Convergence diagnostics help to resolve these issues. Note, however, that most diagnostics are designed to verify a necessary but NOT sufficient condition for convergence.

Visual analysis via Trace Plots

- The simplest diagnostic is obtained by plotting the value of one model parameter versus the simulation index (i.e. the first point in the Markov chain has index 1, the second 2, and so on).

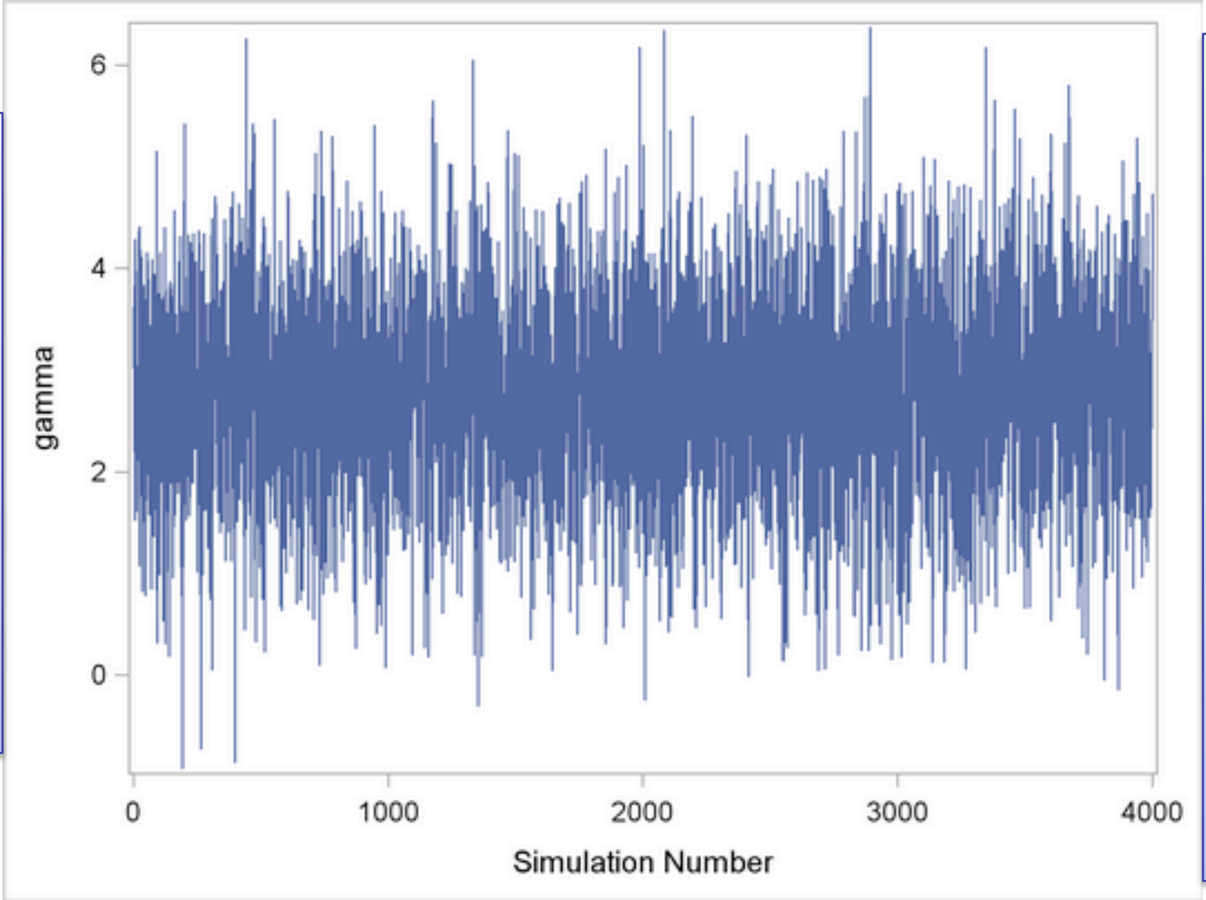
- This is called a Trace Plot.

- As we will see, a trace tells you if a longer burn-in period is needed, if a chain is mixing well, and gives you an idea about the stationary state of the chain.

- Trace plots must be produced for all the parameters, not only for those of interest! If some of parameters have bad mixing you cannot get accurate posterior inference for parameters that appear to have good mixing.

Example I

If the chain has reached stationarity the mean and the variance of the trace plot should keep relatively constant.

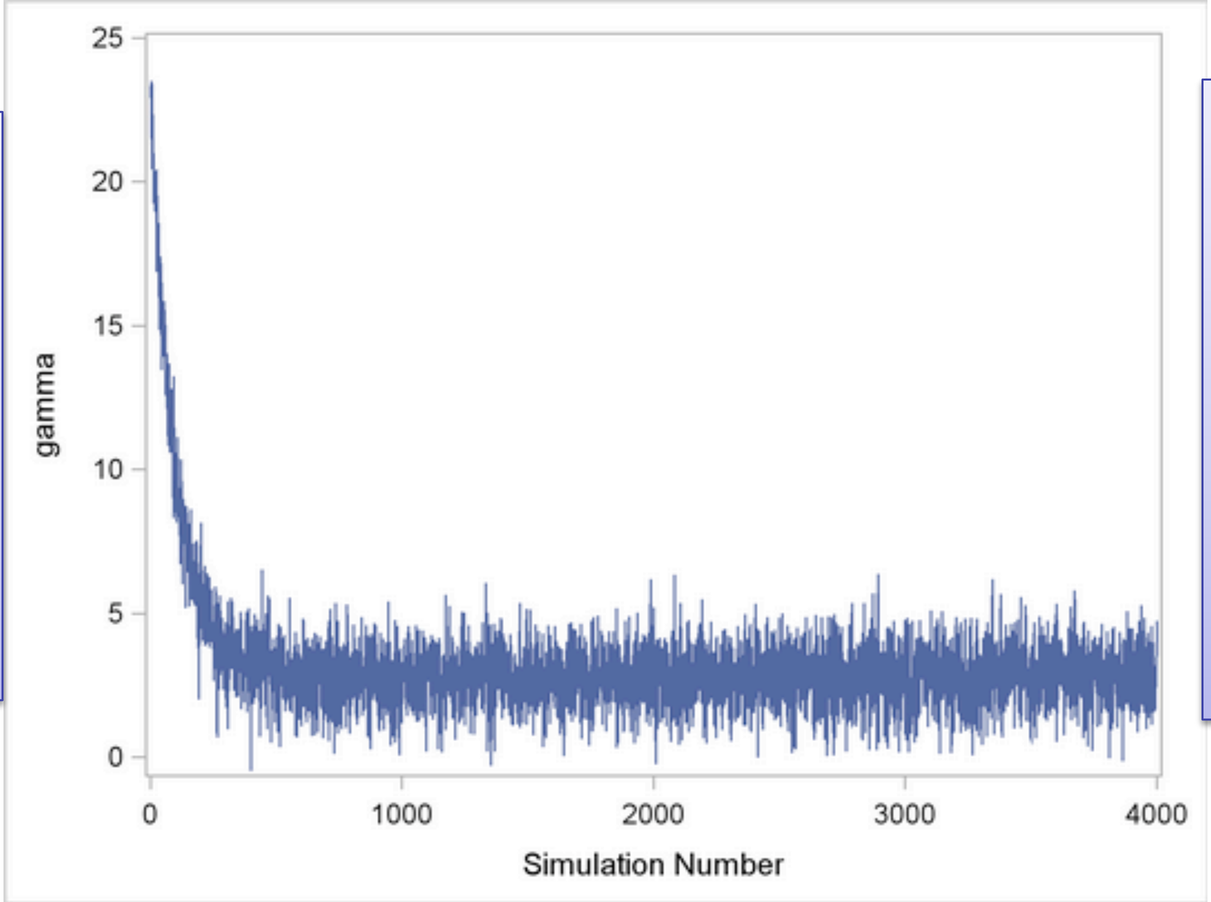


A chain that mixes well traverses the posterior space rapidly, and it can jump from a remote region of the posterior to another in relatively few steps.

The figure displays a “perfect” trace plot, not easy to achieve in high-dimensions

Example II

This chain starts at a very remote location and makes its way to the targeting distribution.

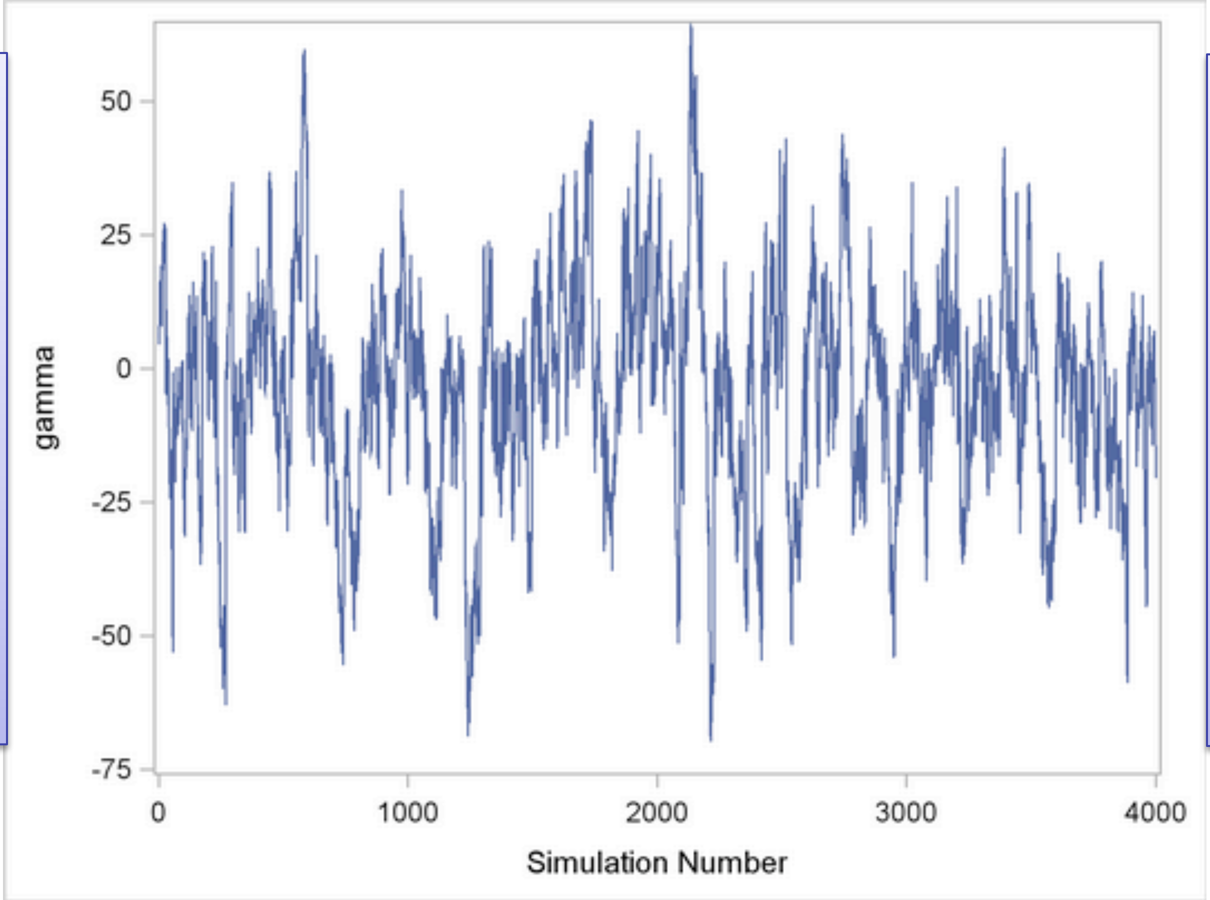


This chain mixes well locally and travels relatively quickly to the target distribution, reaching it in a few hundred iterations.

If you have a chain like this, increase the burn-in sample size.

Example III

This trace plot shows marginal mixing. The chain is taking small steps and does not traverse its distribution quickly.

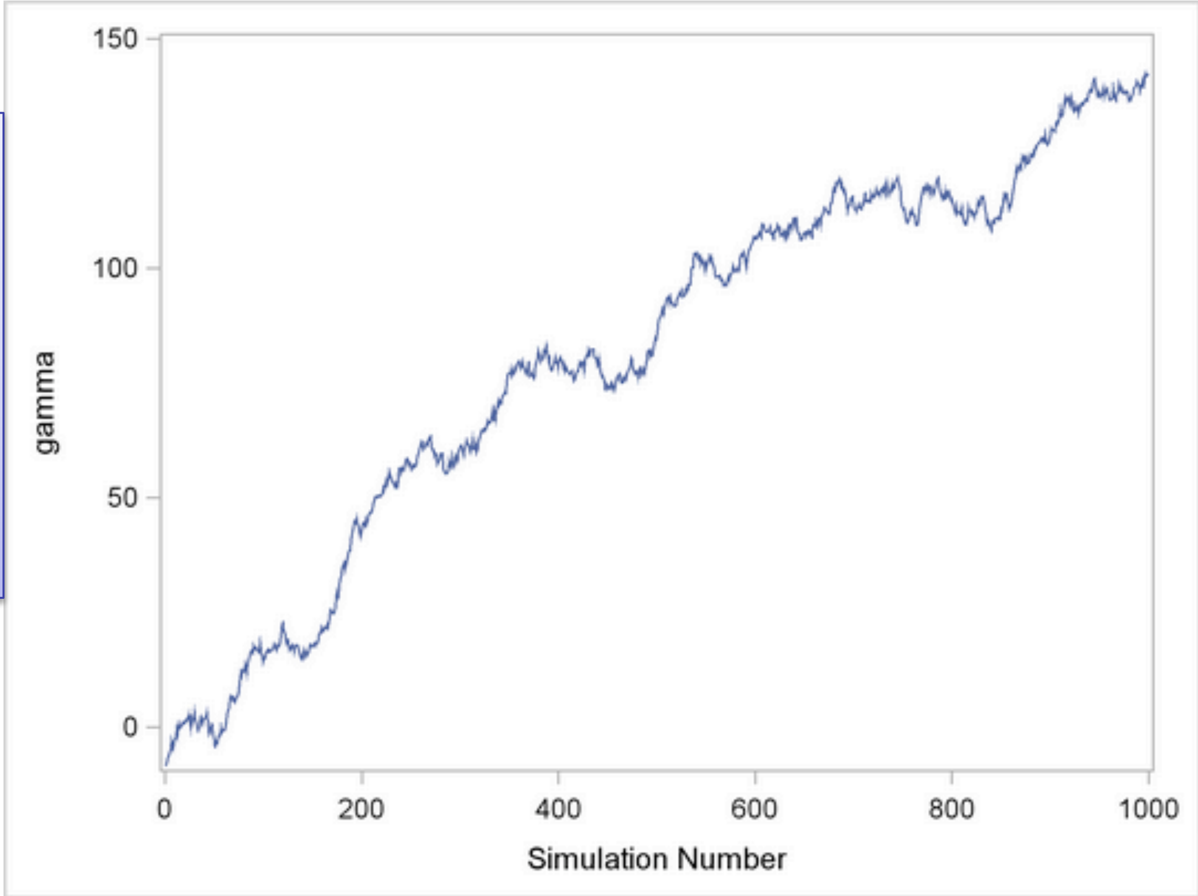


This type of trace plot is typically associated with high correlation among the samples. The chain takes too long to forget where it was before.

In order to obtain a given number of independent samples you need to run the chain for much longer.

Example IV

This chain has serious problems. It mixes very slowly, and it does not give any evidence of convergence.

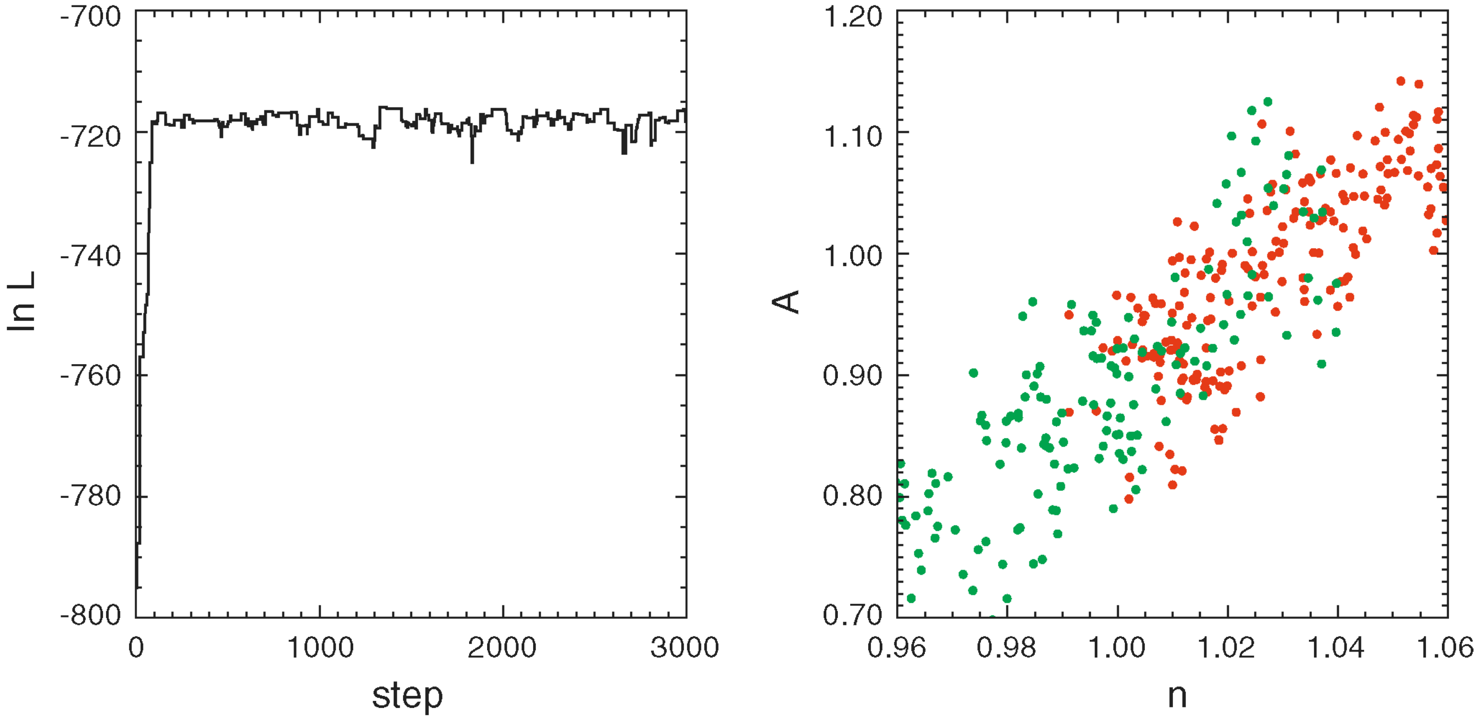


You would want to try to improve the mixing of this chain. For example, you might consider changing the proposal distribution or reparameterizing your model.

This type of chain is entirely unsuitable for making parameter inferences!

Convergence

Verde et al. 2003, first-year WMAP data



Although the trace plot on the left may appear to indicate that the chain has converged after a burn-in of a few hundred steps, in reality it has not fully explored the posterior surface.

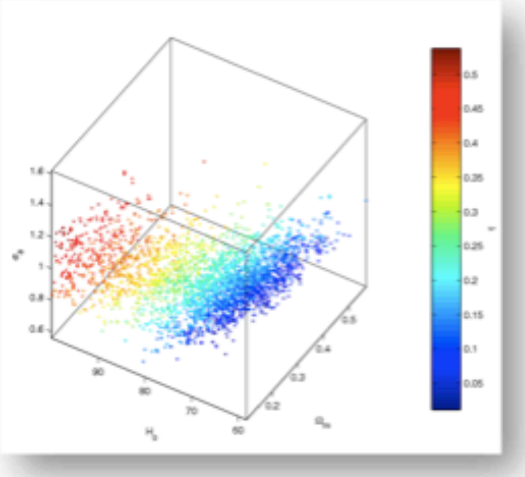
This is shown on the right where two chains of the same length are plotted. Using either of these two chains at this stage will give incorrect results for the best-fit cosmological parameters and their errors.

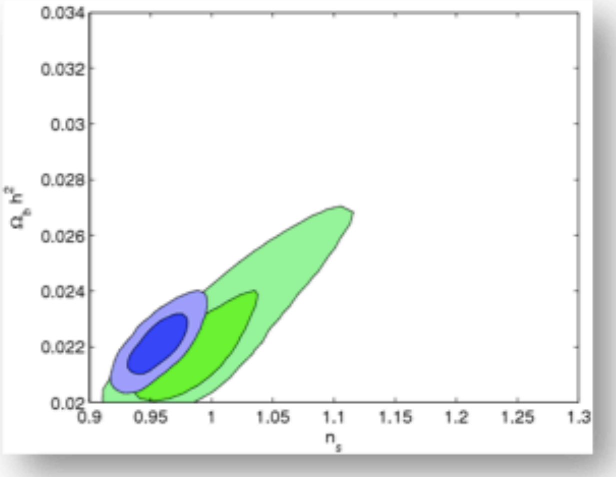
Statistical diagnostics

- Gelman-Rubin: uses parallel chains with dispersed initial values to test whether they all converge to the same target distribution.
- Geweke: tests whether the mean estimates of the parameters have converged by comparing means from the early and latter part of the Markov chain.
- Raftery-Lewis: Evaluates the accuracy of the estimated percentiles by reporting the number of samples needed to reach the desired accuracy.
- And many, many, more...

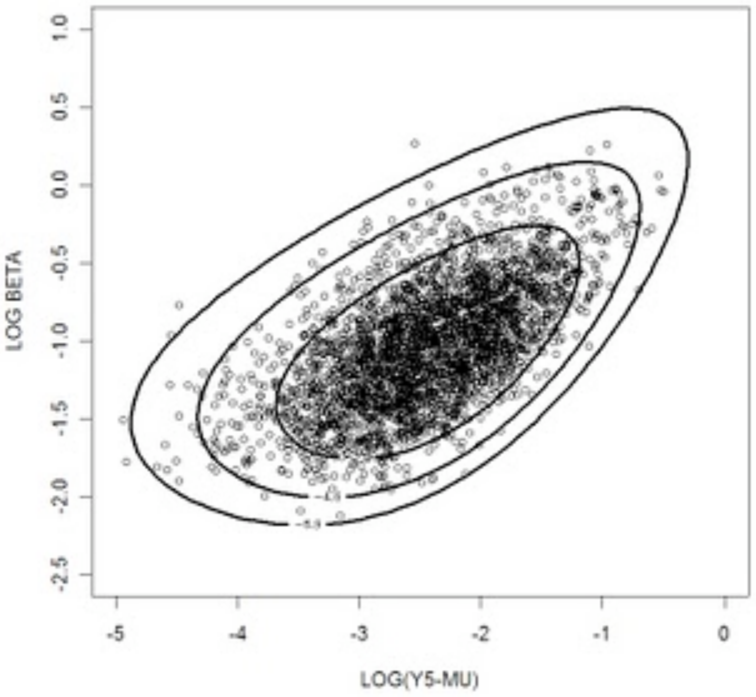
Marginalisation

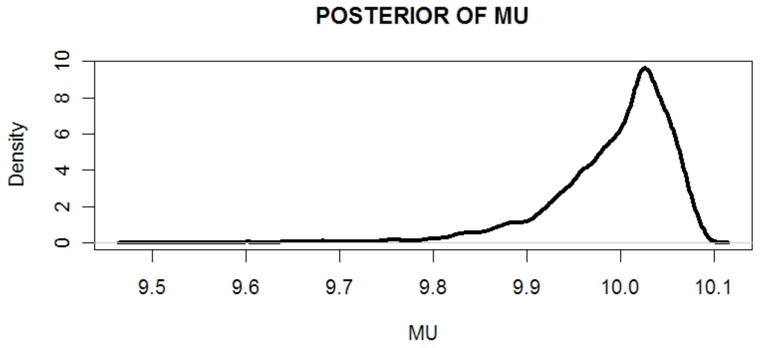
- Marginalisation is trivial
  - Each point in the chain is labelled by all the parameters
  - To marginalise, just ignore the labels you don't want

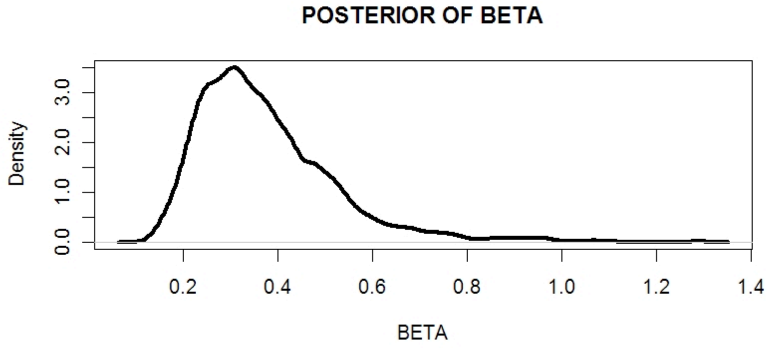




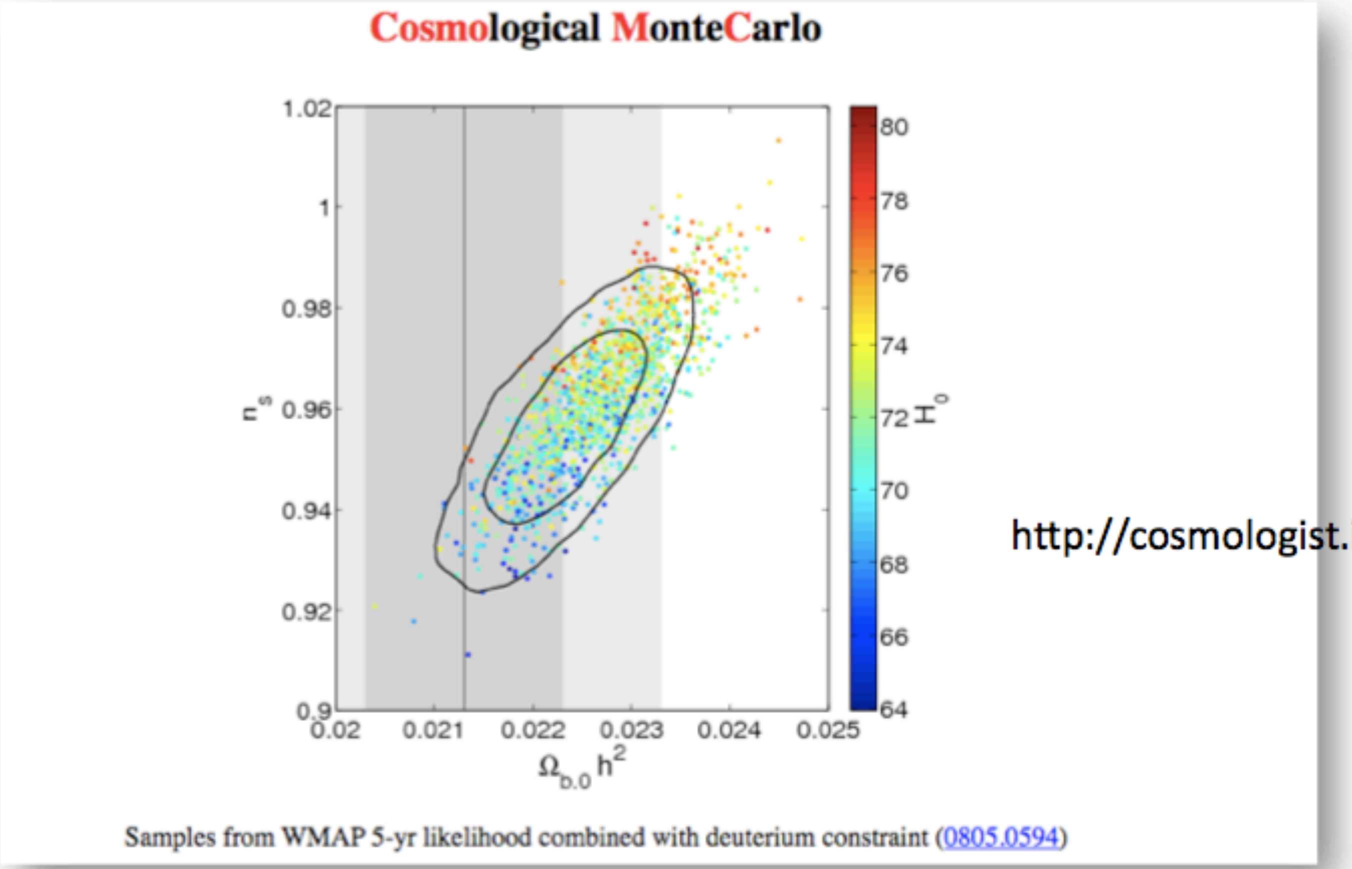
How to plot the results







CosmoMC



Cosmological MonteCarlo

Samples from WMAP 5-yr likelihood combined with deuterium constraint (0805.0594)

http://cosmologist.info/cosmomc/