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 $\lambda = 8$.

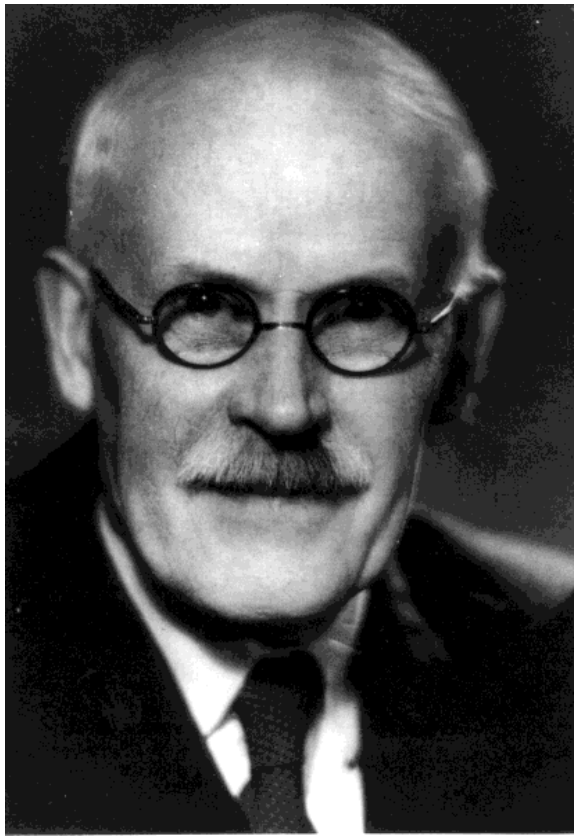
- Also, we live close to a church and whenever there is a wedding the traffic is more intense than usual. This means that on wedding days a higher value of λ is more likely than on non-wedding days.

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

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

The Bayesian way

Bruno de Finetti (1906-1985)



Harold Jeffreys (1891-1989)

What is probability?

- Probability is a modern concept first discussed in a correspondence between Blaise Pascal and Pierre de Fermat in 1654

- There is no unique definition of probability, statisticians are divided into different schools with contrasting views

- Classic definition: The probability of an event is the ratio of the number of cases favorable to it, to the number of all cases possible when nothing leads us to expect that any one of these cases should occur more than any other, which renders them, for us, equally possible. (Pierre-Simon de Laplace, *Théorie analytique des probabilités*, 1812)

- This is based on the "principle of insufficient reason" (or principle of indifference) which states that when cases are only distinguishable by their name they should be assigned the same probability

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 $P(E)\geq 0$, $P(\Omega)=1$, 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

De Finetti's game
Can you measure degree of belief?

Suppose we are on a trip and you say that you are "pretty sure" you locked the door of your flat. I want to determine how sure you are.

- I offer you to play a game: I propose you to draw a marble from a bag containing 95 red and 5 blue marbles. If you pick at random a red marble, I give you one million euros. Alternatively, I offer you to go back home and check the door. If you choose this option and the door is locked indeed, I give you one million euros.

- If you choose to pick a marble, it means that your degree of belief is lower than 95%

- I can then propose many other rounds of the game by progressively reducing the fraction of red marbles until you choose to go back. This would measure your degree of belief.

Posterior probability

$P(x|\theta)$: old name "direct probability"
It gives the probability of contingent events (i.e. observed data) for a given hypothesis (i.e. a model with known parameters θ)

$L(\theta)=P(x|\theta)$: modern name "likelihood function" or simply "likelihood"
It quantifies the likelihood that the observed data would have been observed as a function of the unknown model parameters (it can be used to rank the plausibility of model parameters but it is not a probability density for θ)

$P(\theta|x)$: old name "inverse probability"
modern name "posterior probability"
Starting from observed events and a model, it gives the probability of the hypotheses that may explain the observed data (i.e. of the unknown model parameters)

Rev. Thomas Bayes (1702-1761)

Bayes theorem

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

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

Posterior probability
for the parameters
given the data

Likelihood function

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

Evidence
(normalization
constant useful
for Bayesian
model selection)

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

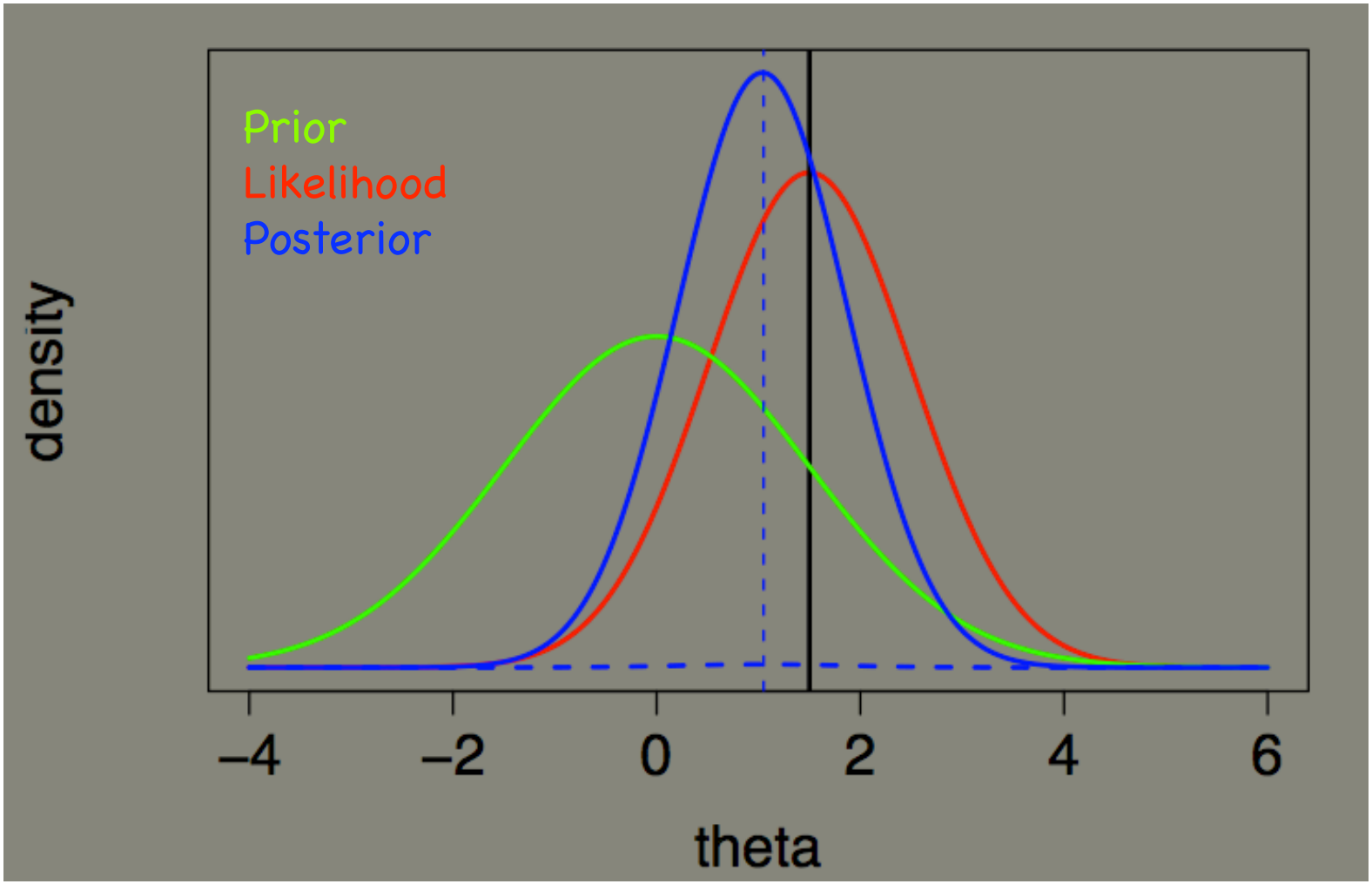
Pierre Simon (Marquis de) Laplace (1749-1827)

Bayes theorem

Let us re-write Bayes theorem emphasizing that all the probabilities that we generically called "p" are actually different functions:

$$P(\theta \mid x) = \frac{L(x \mid \theta)\ \pi(\theta)}{E(x)}$$

Bayes theorem visually



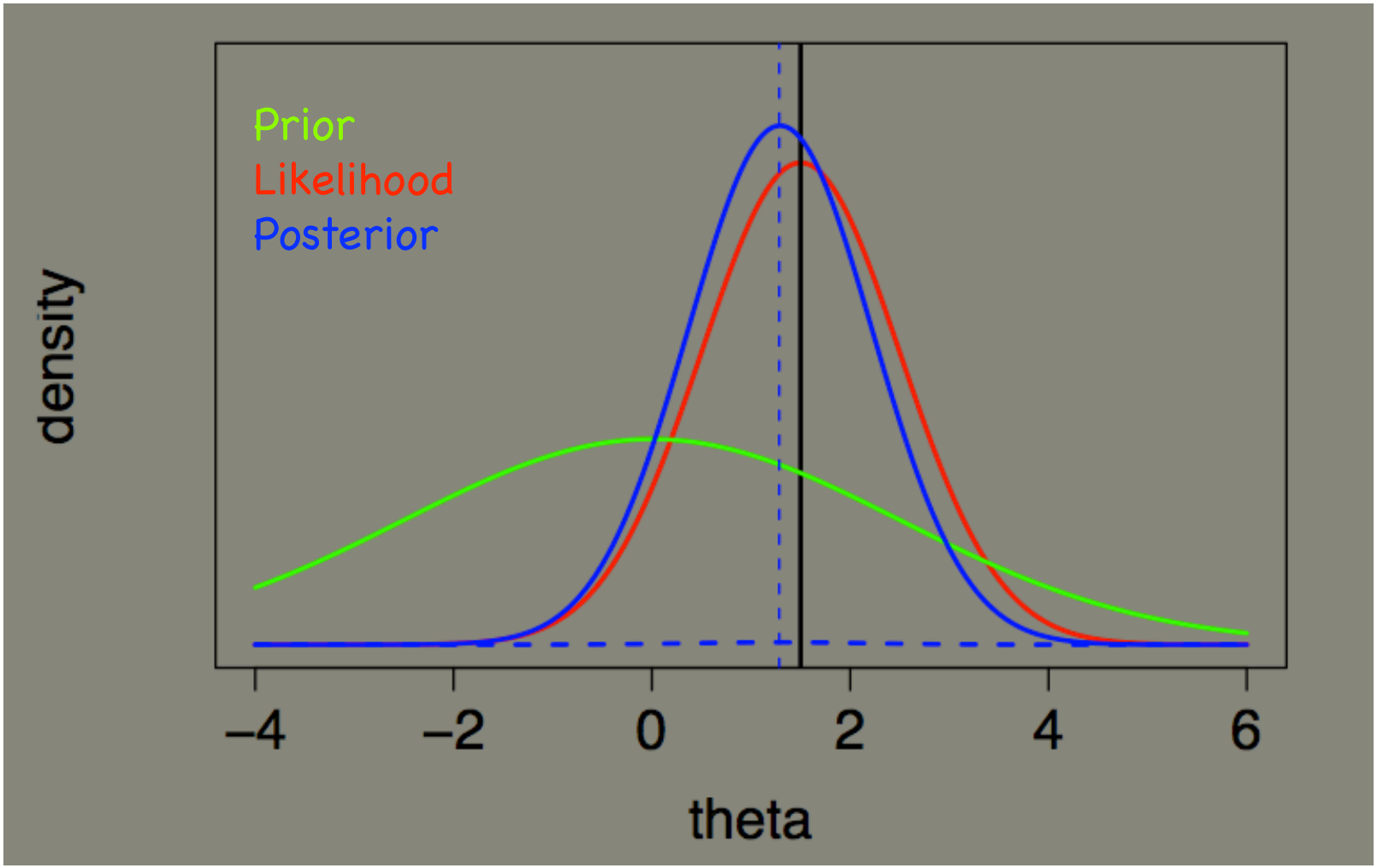
Priors

- Non-informative: if it has a minimal impact on the posterior distribution (i.e. it is "flat" with respect to the likelihood function). Non-informative priors are also called "vague", "diffuse" or "flat".

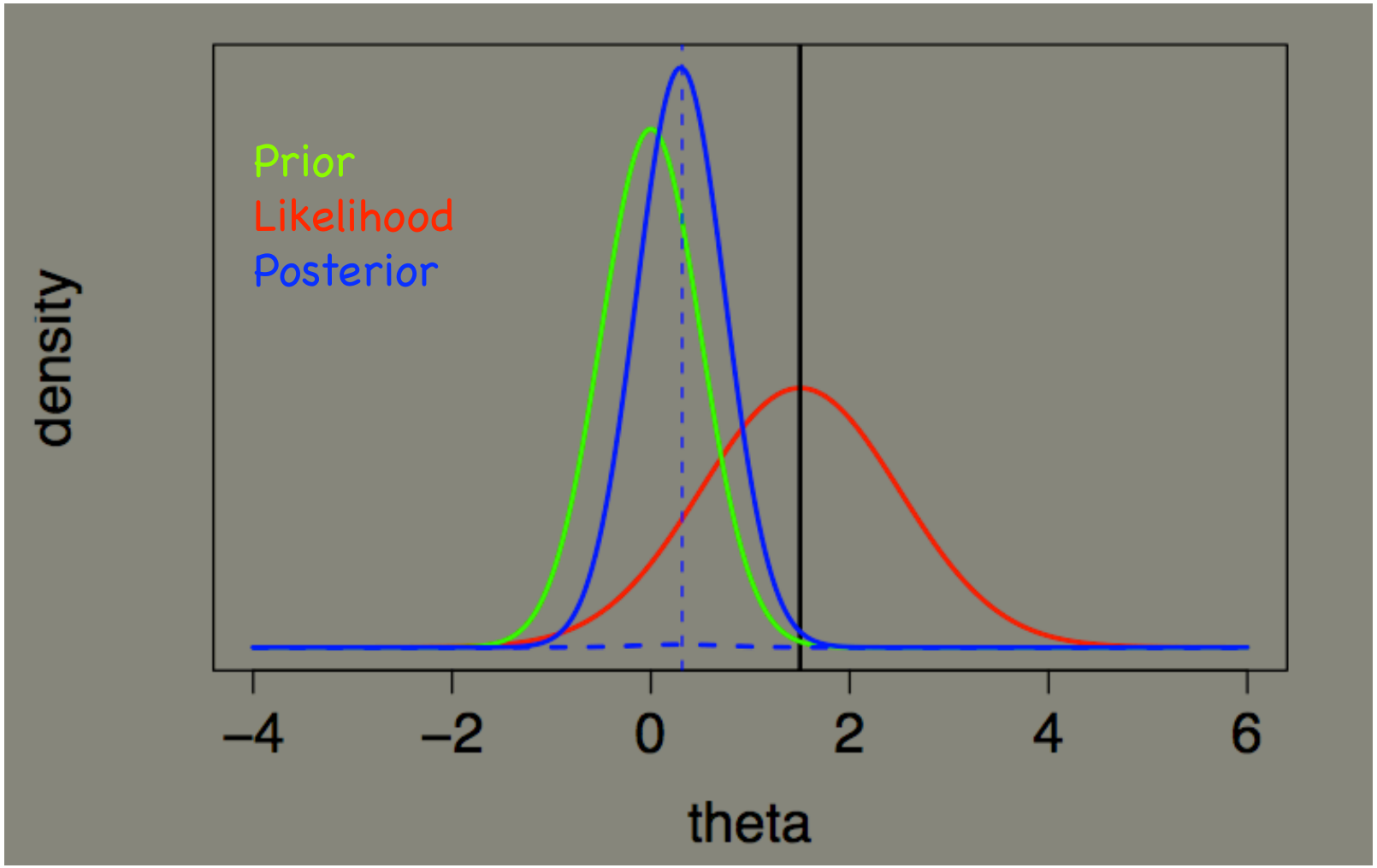
- Improper: if $\int \pi(\vartheta) d\theta = \infty$ e.g. uniform prior on the real line. Generally leads to a proper posterior but, if also the posterior is improper, inference is invalid.

- Informative: a prior which is not dominated by the likelihood. Must be handled with care in actual practice. On the other hand, illustrates the power of the Bayesian method: information gathered from previous study, past experience, or expert opinion can be combined with current information in a natural way

Not very informative prior



Very informative prior



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?

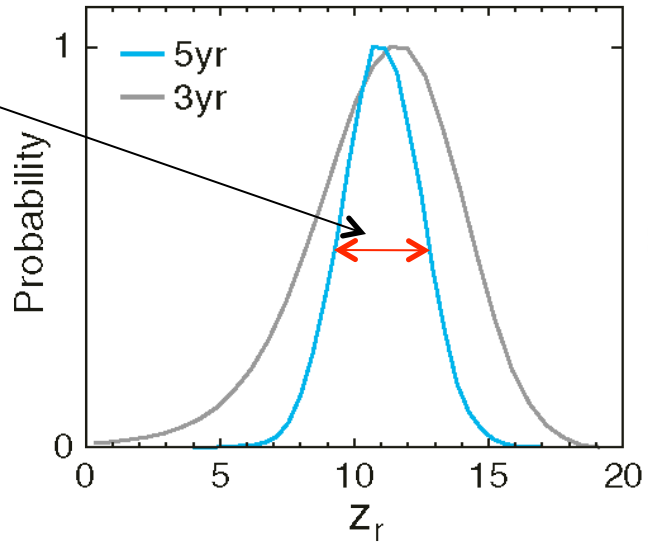
- They are profoundly DIFFERENT things even though they are often confused. Sometimes practitioners tend use the term "confidence intervals" in all cases and this is ok because they understand what they mean but this might be confusing for the less experienced readers of their papers. PAY ATTENTION!

Marginalisation

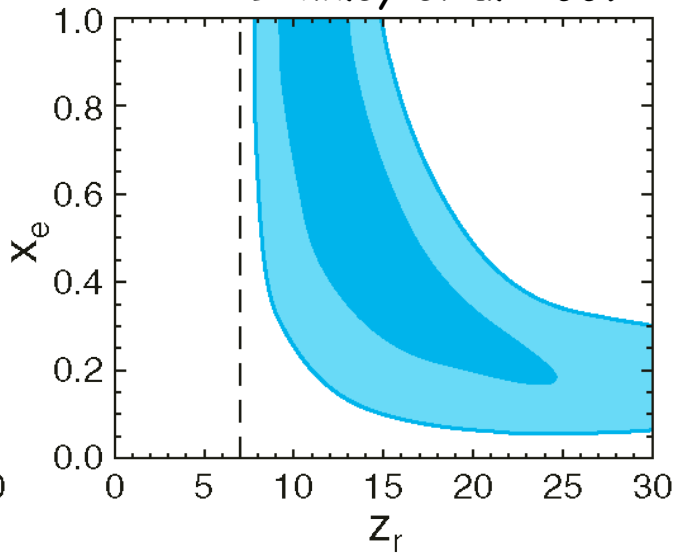
Marginal probability: posterior probability of a given parameter regardless of the value of the others. It is obtained by integrating the posterior over the parameters that are not of interest.

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

Marginal errors characterise the width of the marginal posterior distributions.



Dunkley et al. 2009



Pitfalls of Bayesian inference

- There is no "correct" way to choose a prior. Bayesian inference requires skills to translate subjective prior beliefs into a mathematically formulated prior.

- It can produce posterior distributions which are heavily influenced by priors.

- It often comes with high computational costs.

How can we do this in practice?

Markov Chain Monte Carlo

Andrey Andreyevic Markov
(1856-1922)

Monte Carlo Casino
(1863-now)

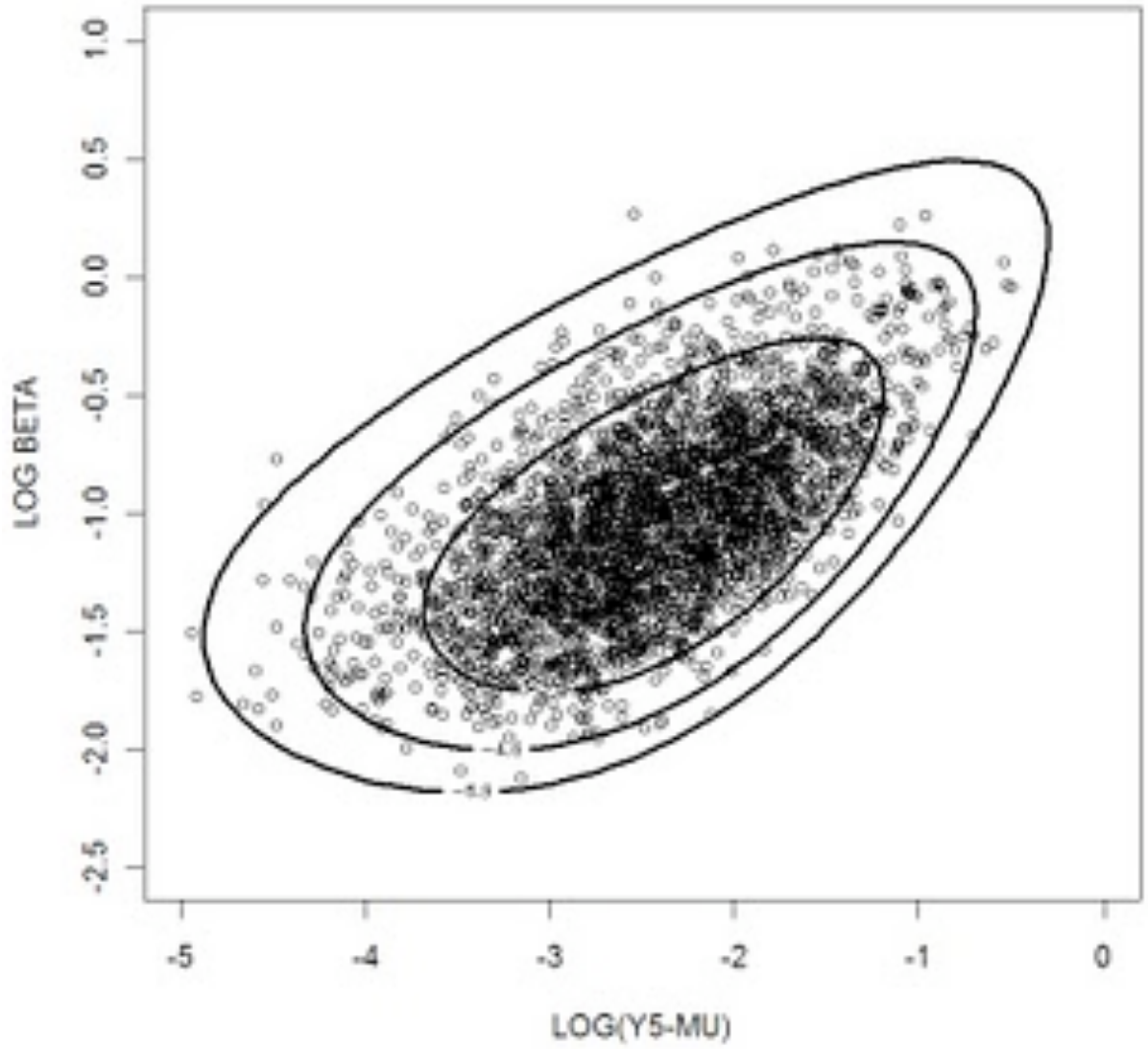
Markov Chain Monte Carlo

- 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


LOG BETA
1.0
0.5
0.0
-0.5
-1.0
-1.5
-2.0
-2.5
-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 meaning, 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).

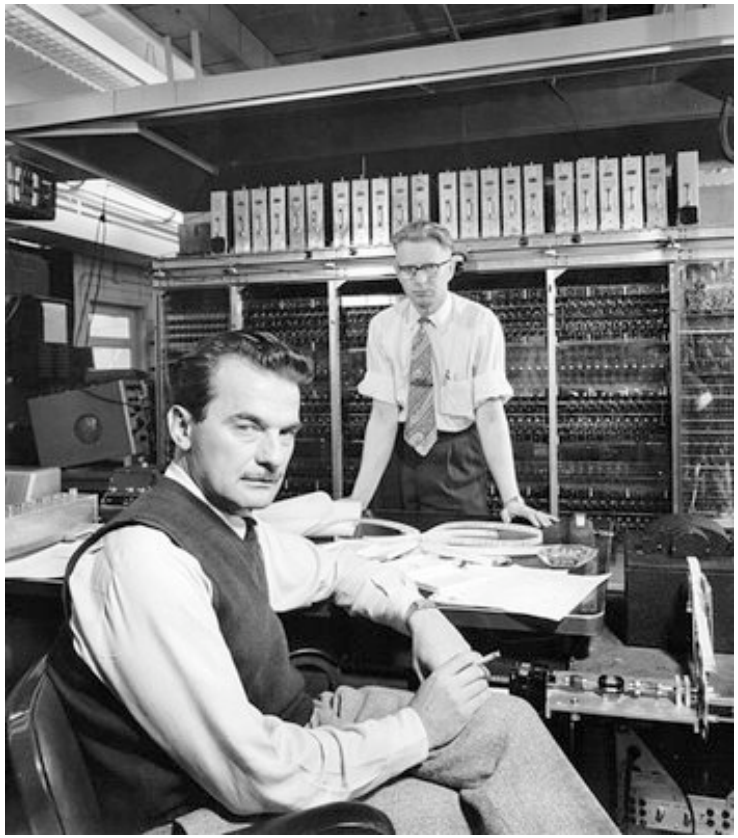
MCMC algorithm

- 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:

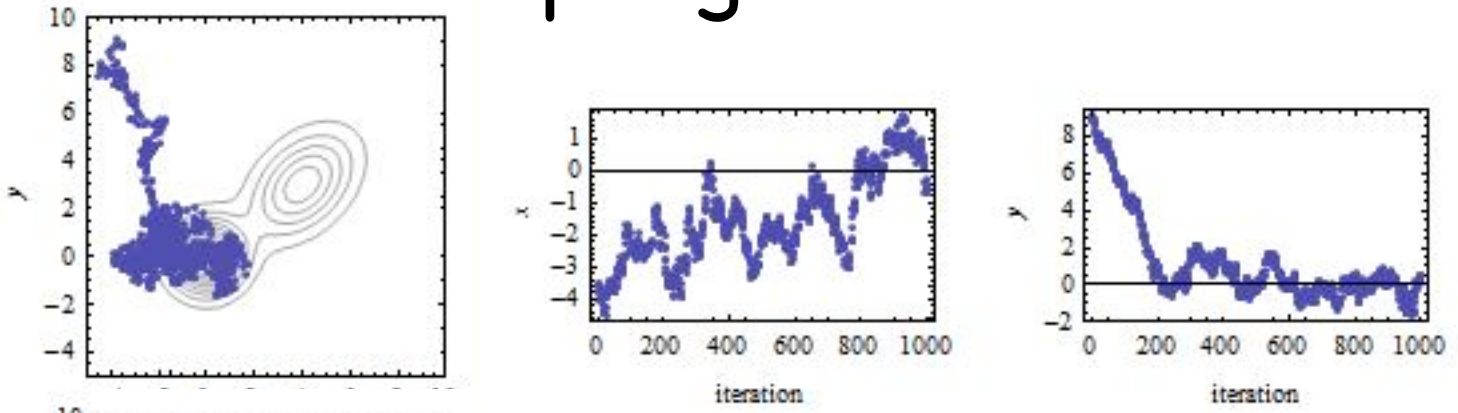
$$r = \text{probability of acceptance} = \min\left(\frac{f(\theta_{new})q(\theta_t \mid \theta_{new})}{f(\theta_{old})q(\theta_{new} \mid \theta_t)}, 1\right)$$

The proposal distribution

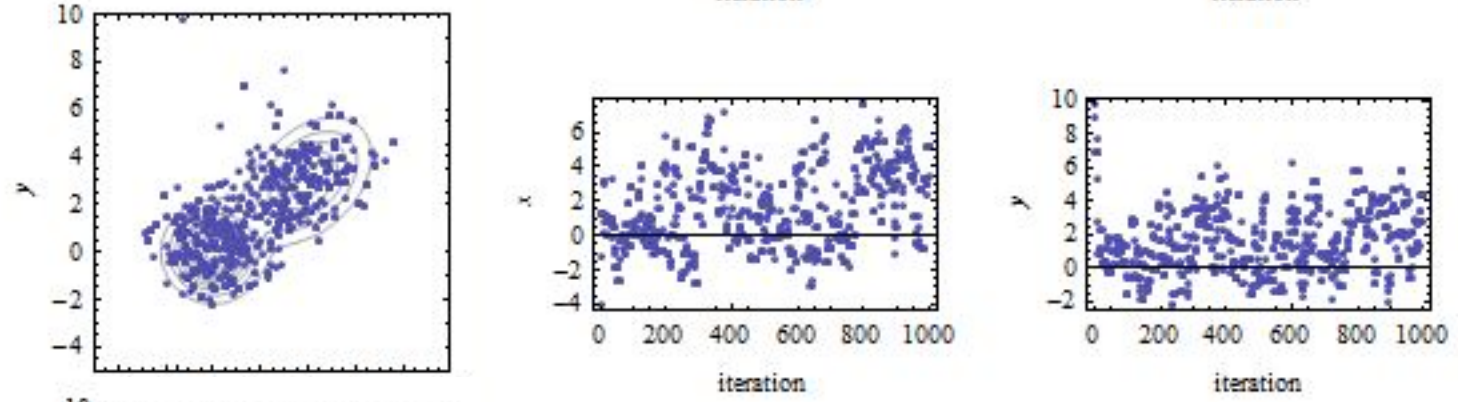
- If one takes too small steps, it takes long time to explore the target and the different entries of the chain are very correlated
- If one takes too large steps, almost all trials are rejected and the different entries of the chain are very correlated
- There is an optimal proposal distribution (easy to identify if we knew already the target density)

Effect of the sampling distribution

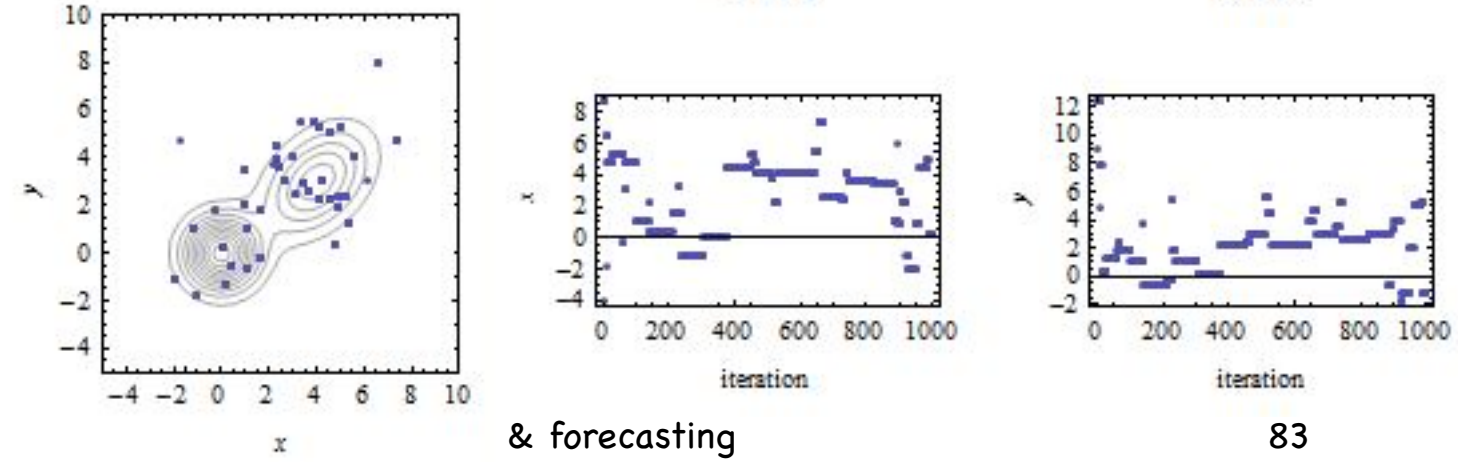
Gaussian proposal distribution with $\sigma = 0.2$, acceptance rate =85.1%



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



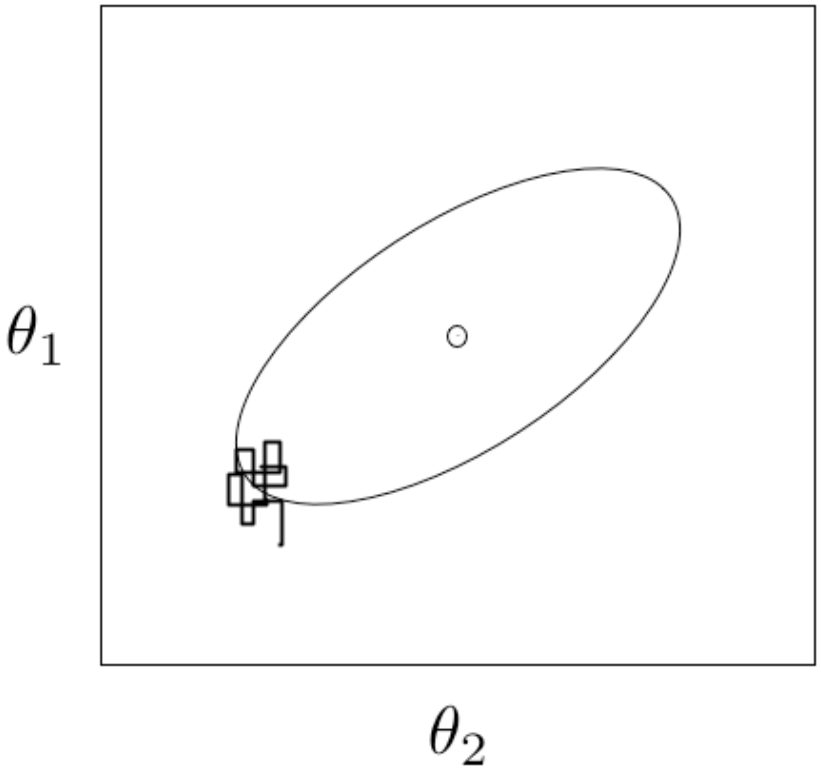
Gaussian proposal distribution with $\sigma = 10.2$, acceptance rate =4.1%



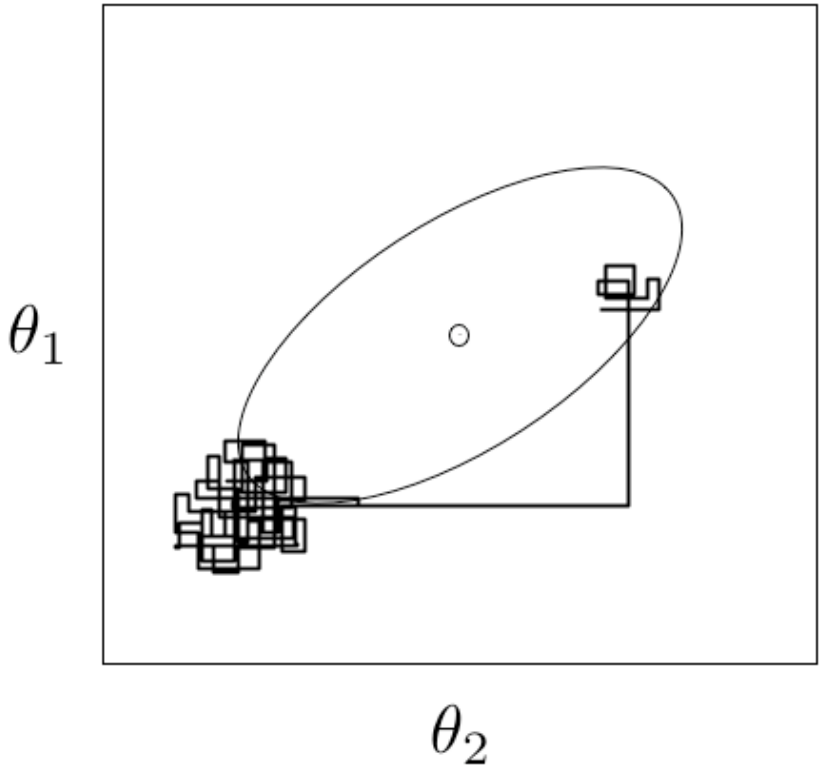
Mixing

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

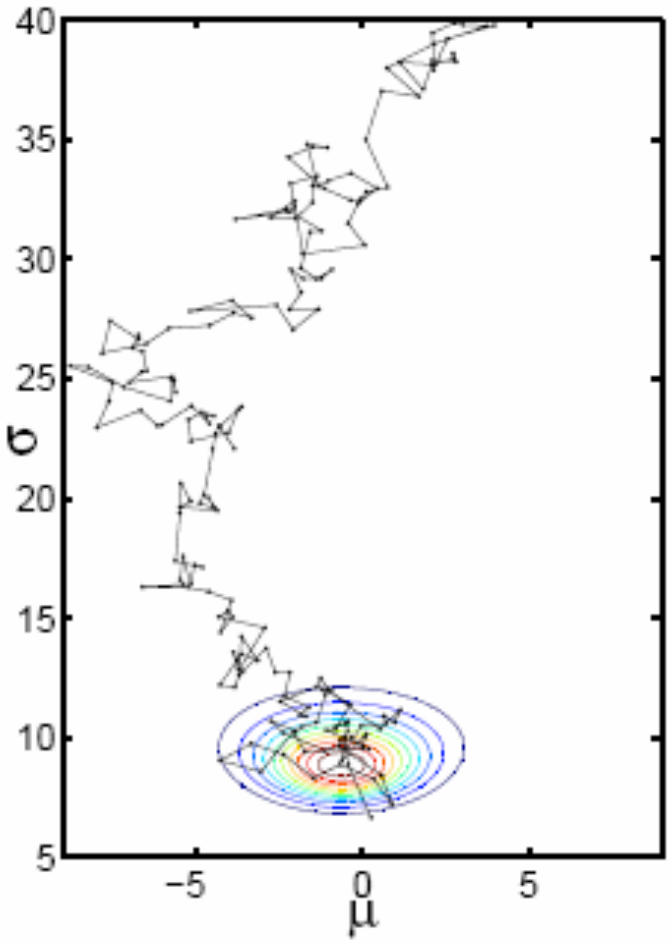


Metastability



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).

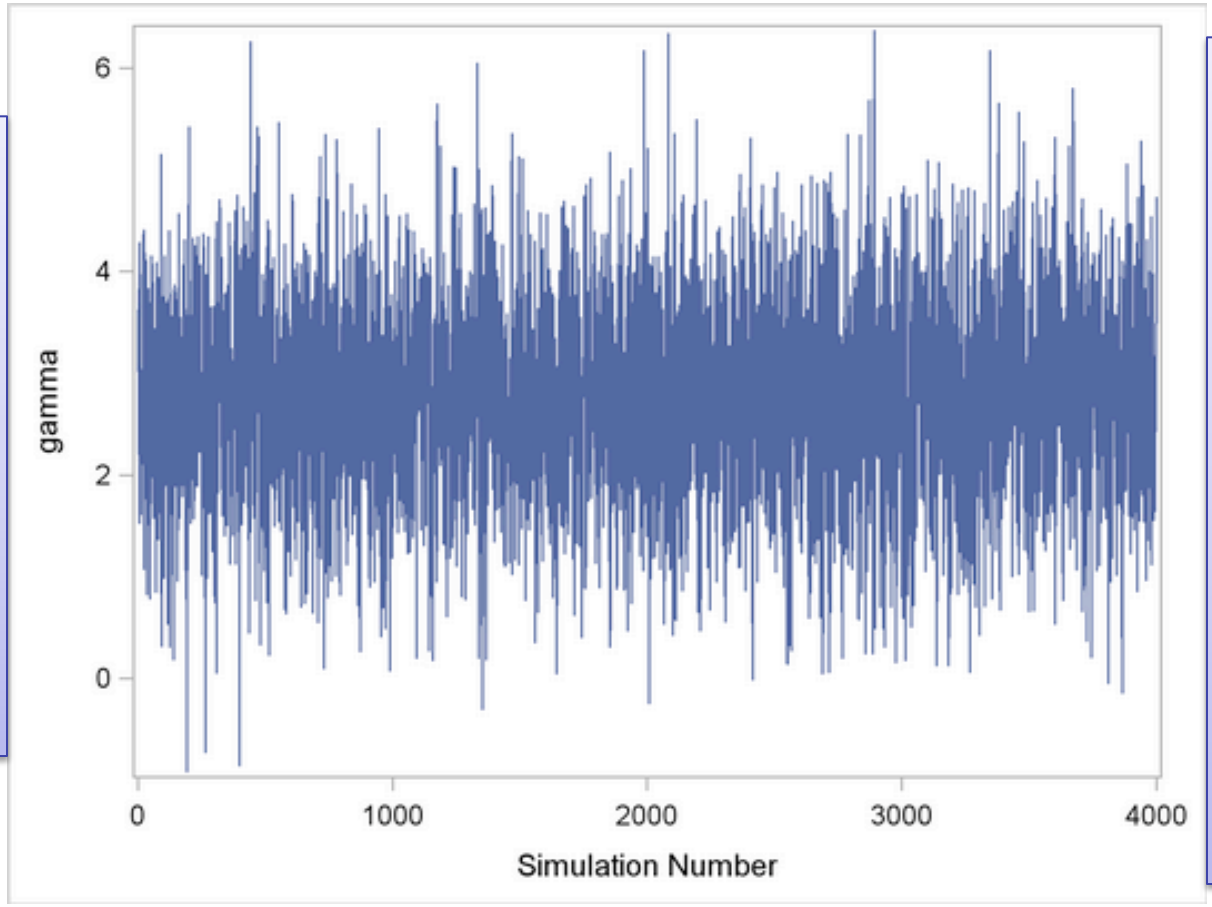
- This is called a Trace Plot.

- 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.

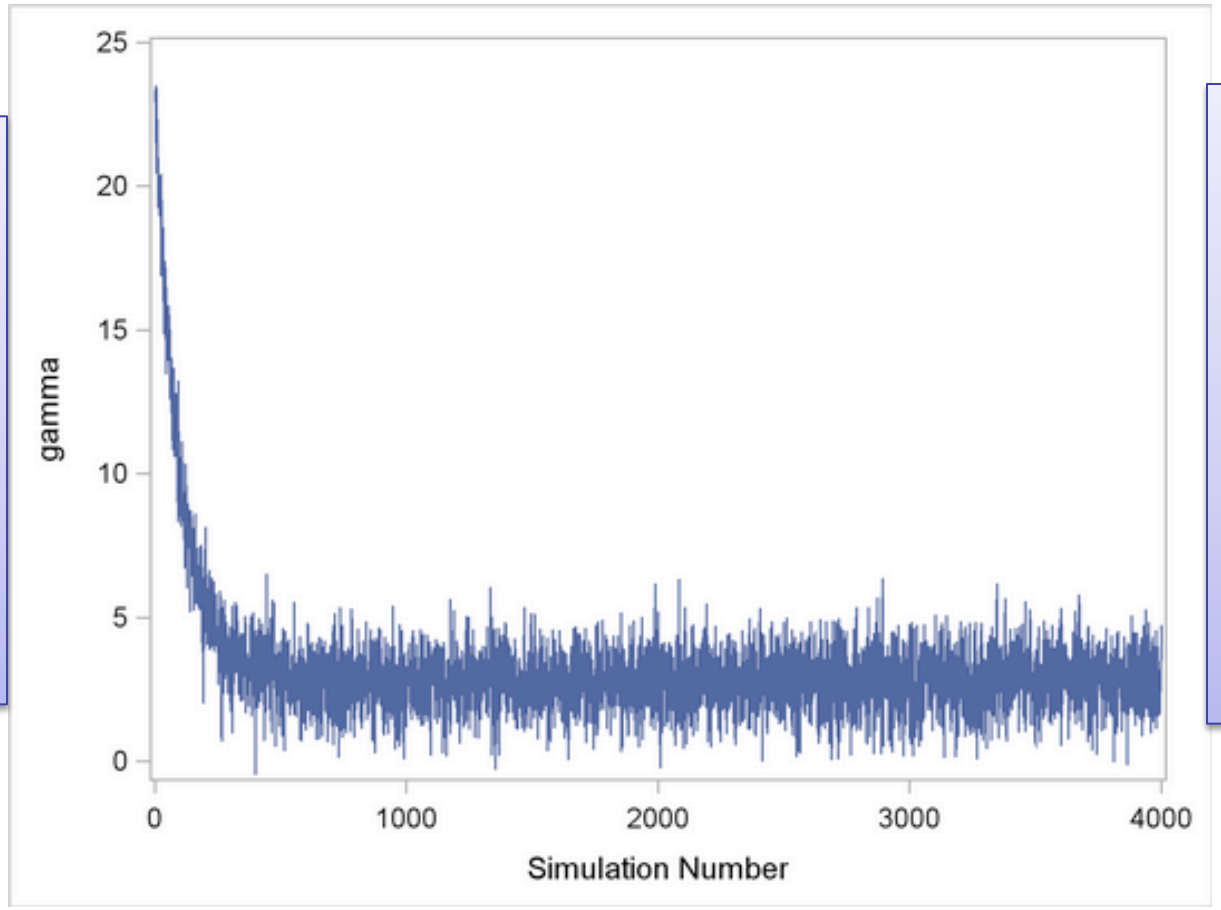


A chain that mixes well traverses the posterior space rapidly, and it can jump from a remote region of the posterior to another in relatively few steps.

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 .

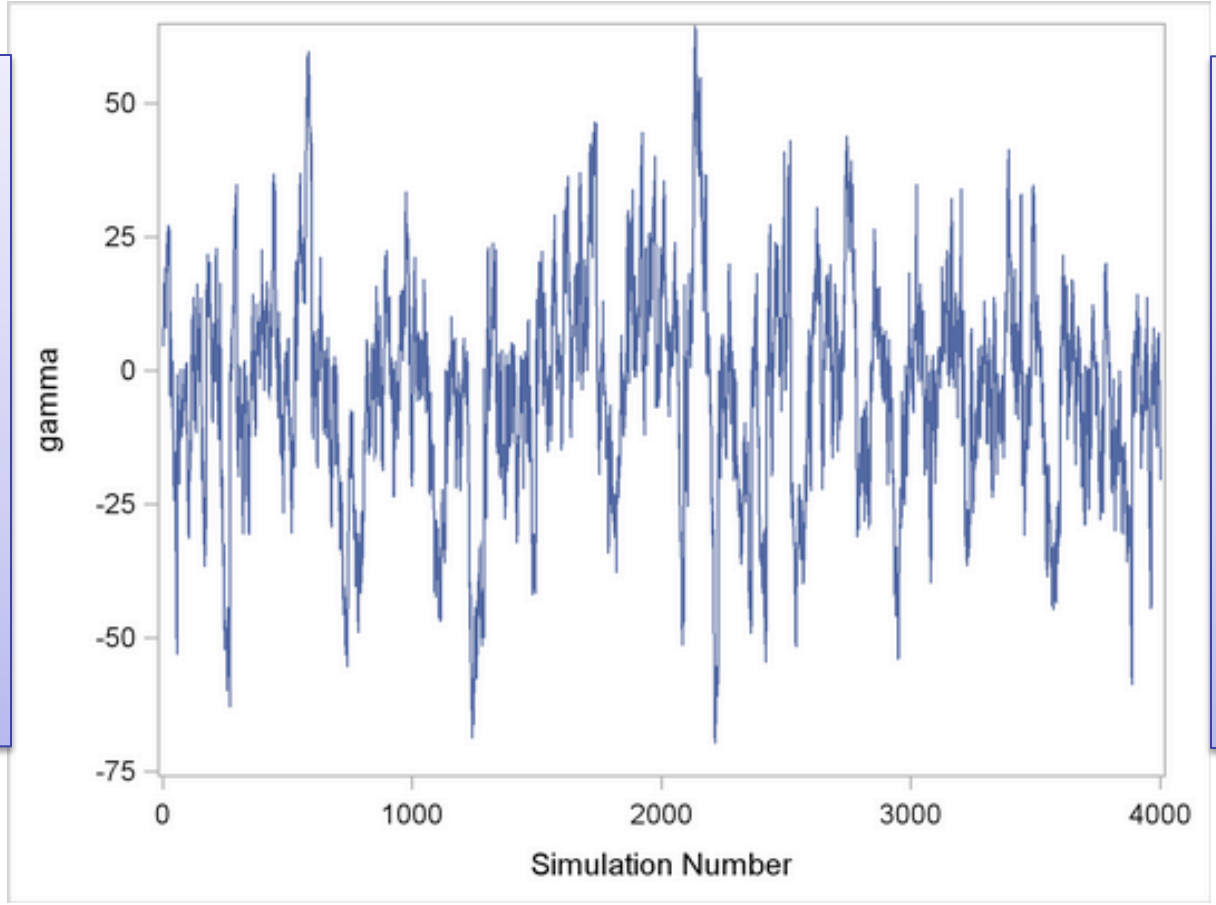


This chain mixes well locally and travels relatively quickly to the target distribution, reaching it in a few hundred iterations.

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.

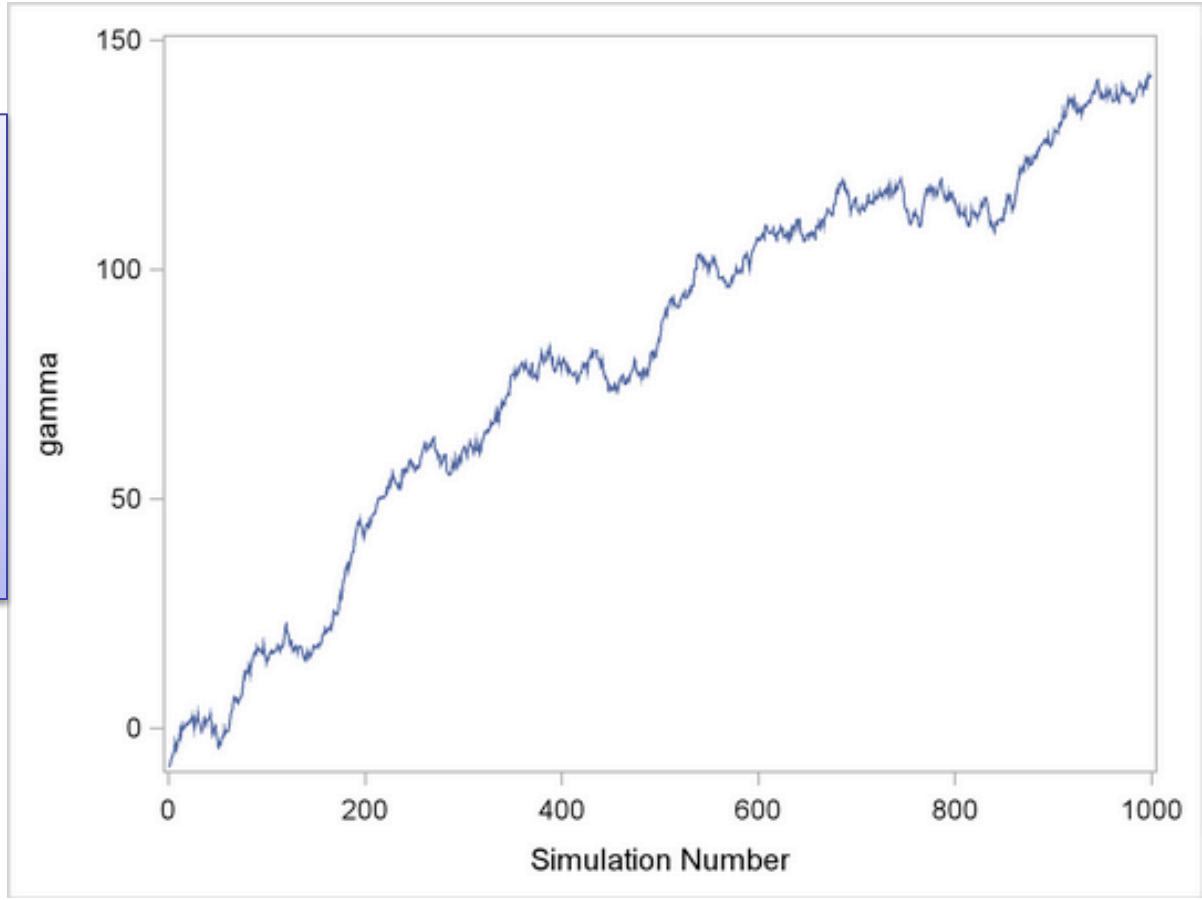


This type of trace plot is typically associated with high correlation among the samples. The chain takes too long to forget where it was before.

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.

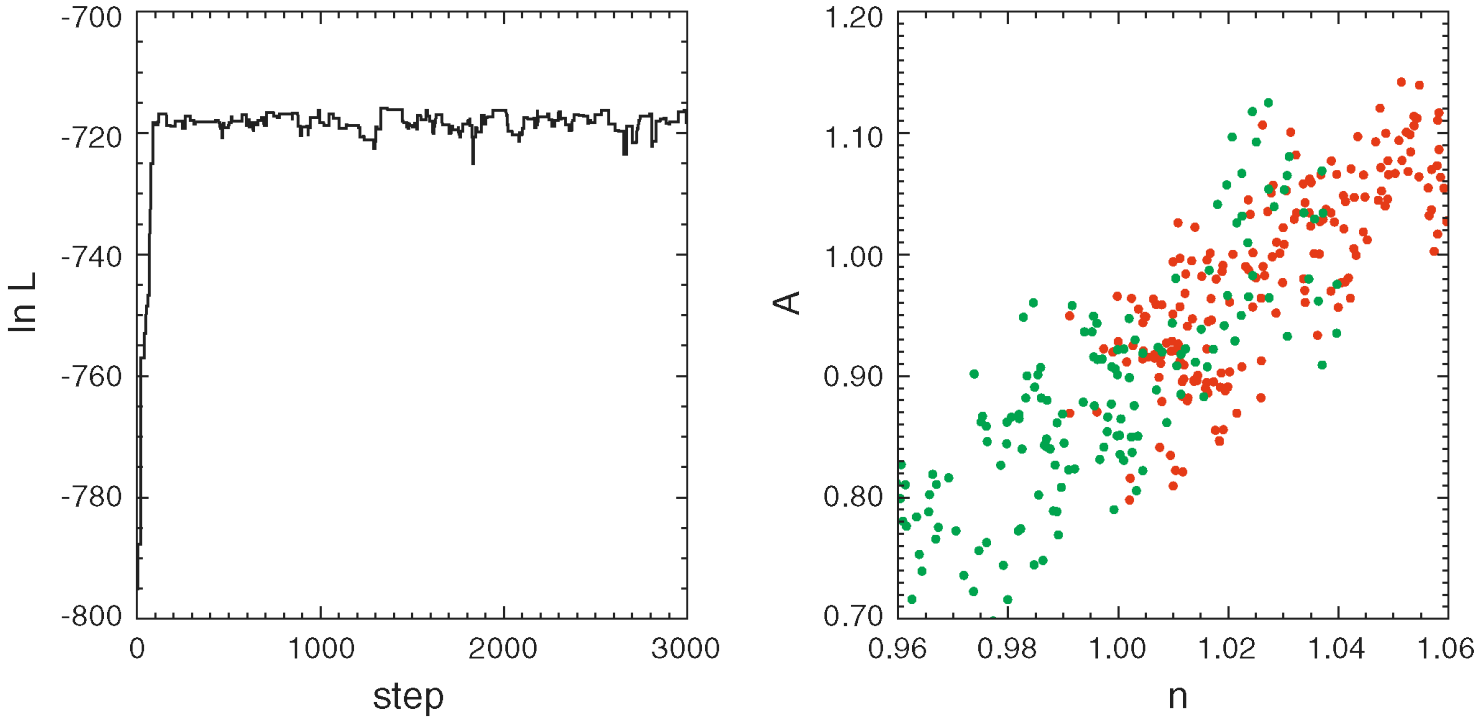


You would want to try to improve the mixing of this chain. For example, you might consider changing the proposal distribution or reparameterizing your model.

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.

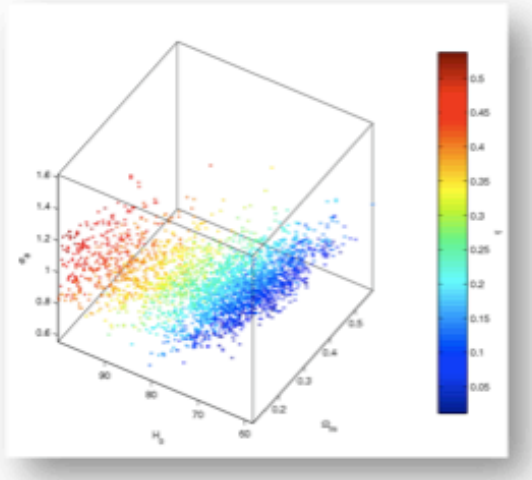
This is shown on the right where two chains of the same length are plotted. Using either of these two chains at this stage will give incorrect results for the best-fit cosmological parameters and their errors.

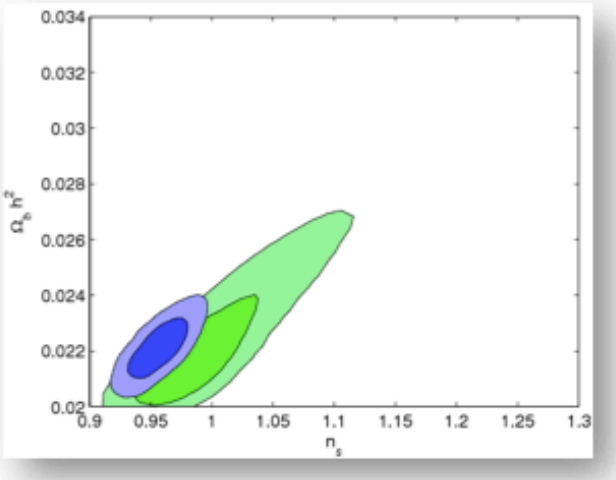
Statistical diagnostics

- 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...

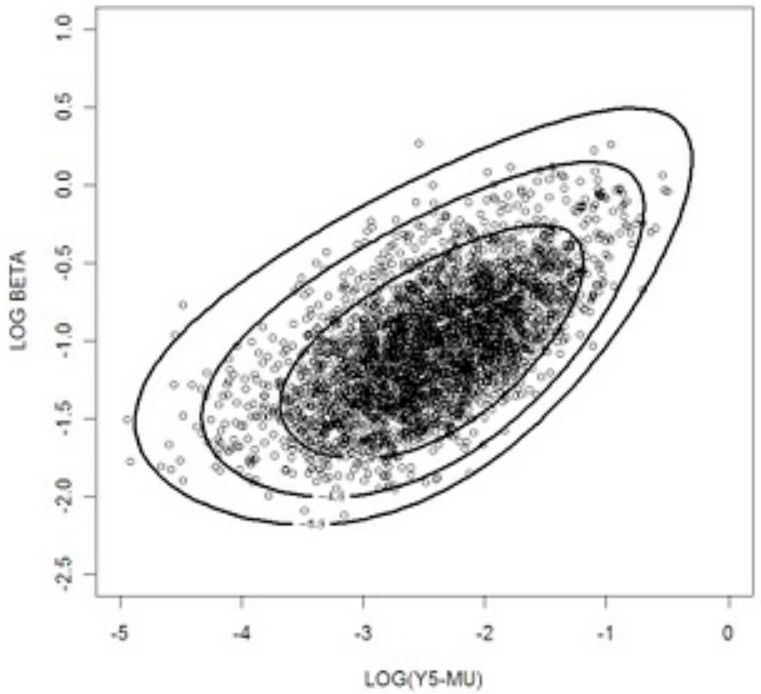
Marginalisation

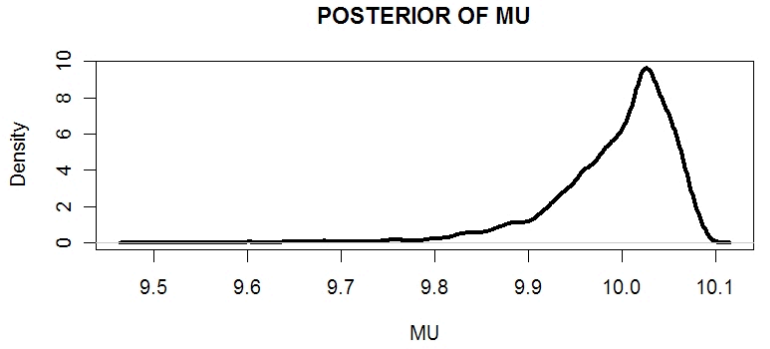
- Marginalisation is trivial
  - Each point in the chain is labelled by all the parameters
  - To marginalise, just ignore the labels you don't want

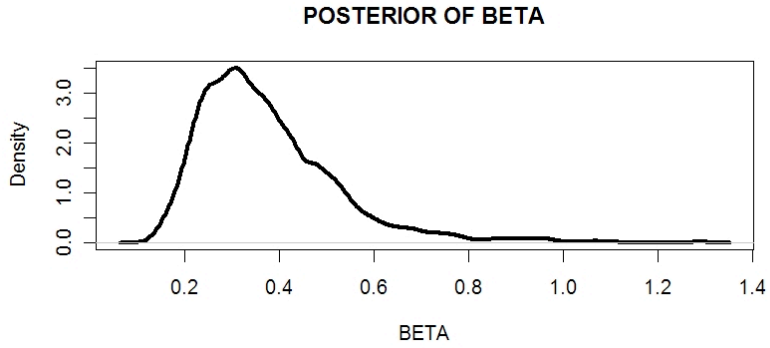




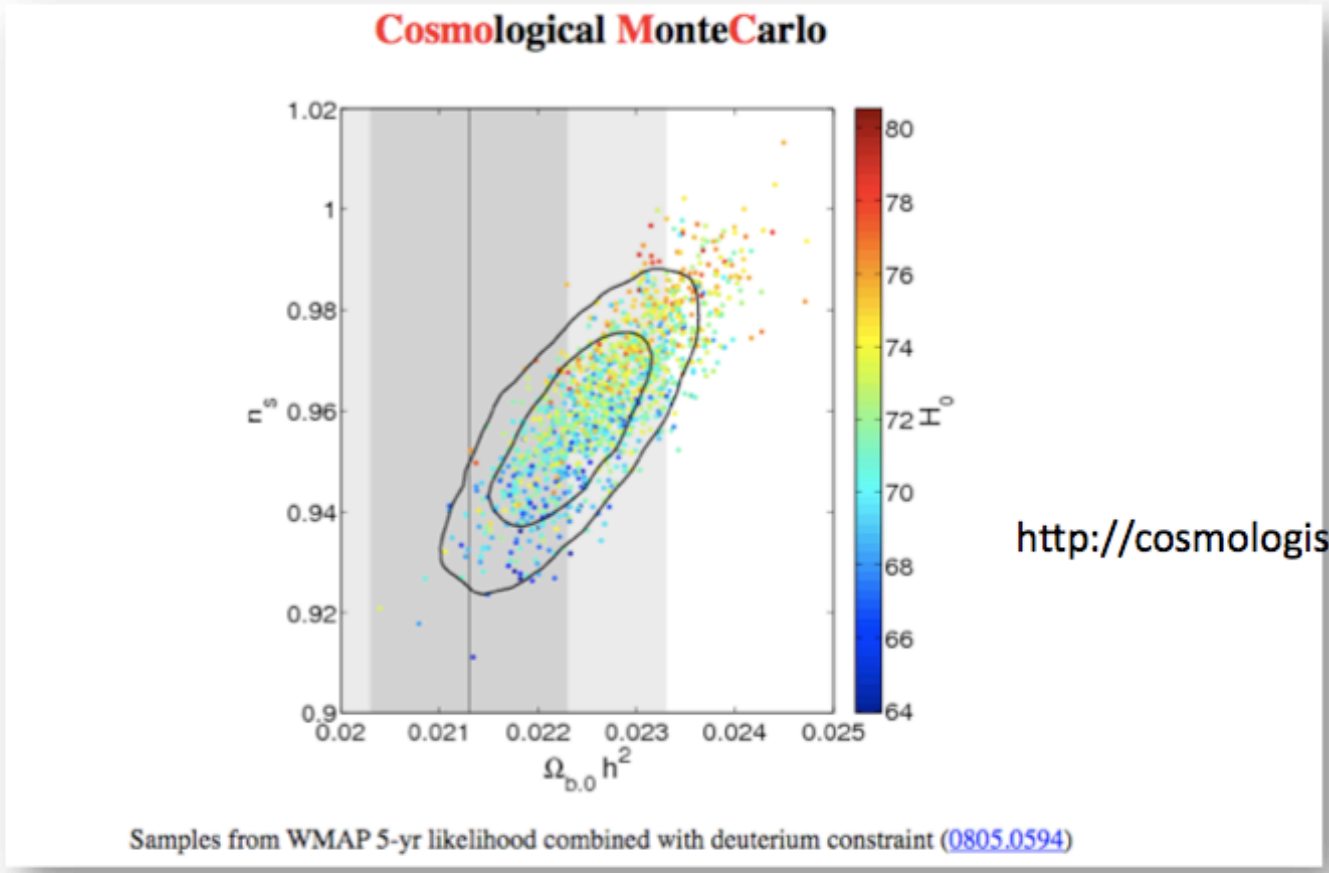
How to plot the results







CosmoMC



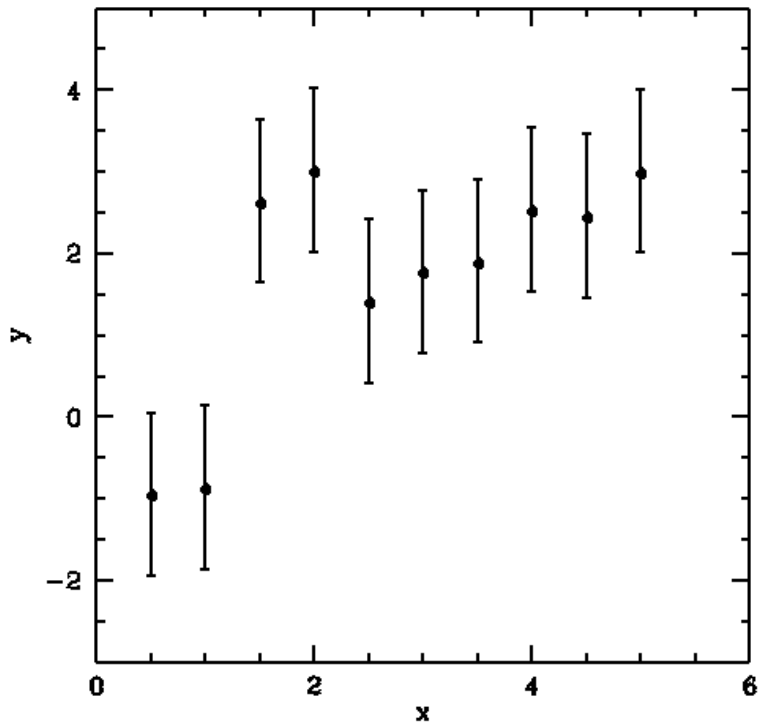
http://cosmologist.info/cosmomc/

Bayesian model comparison

- Suppose you have some data
- You want to fit them with some model
- More than a model is available with a different number of free parameters (e.g. vanilla ΛCDM, ΛCDM + massive neutrinos, ΛCDM + curvature, CDM + dynamic dark energy)
- How do you choose the "best model"?

- Central problem in learning: how to balance "goodness of fit" criteria against the complexity of models
- We want to avoid "overfitting"

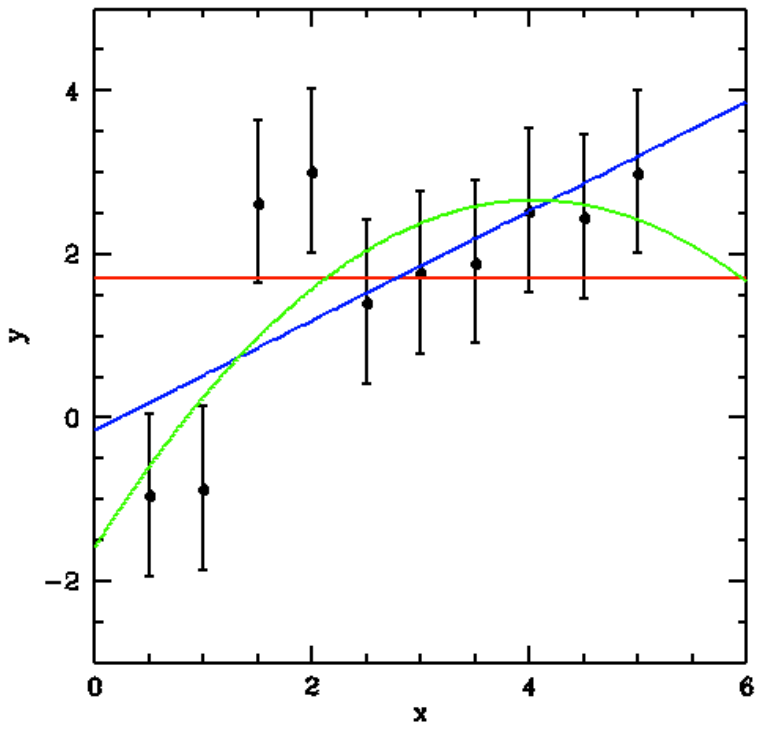
An example



- What function would you use to fit these data?

- Constant? y=c (1 parameter)
- Linear? y=bx+c (2 parameters)
- Quadratic? $y=ax^2+bx+c$ (3 parameters)

Maximum likelihood solution



- Constant fit: χ^2_{min}=19.33
- Linear fit: χ^2_{min}=10.02
- Quadratic fit: χ^2_{min}=7.79

- Which one should be preferred?

- I could get χ^2_{min}=0 by using a polynomial of degree 10

Bayes factor

Let's write Bayes theorem for the models in odds form:

$$K = \frac{P(M_1 \mid \vec{x})}{P(M_2 \mid \vec{x})} = BF \times \frac{\pi(M_1)}{\pi(M_2)}$$

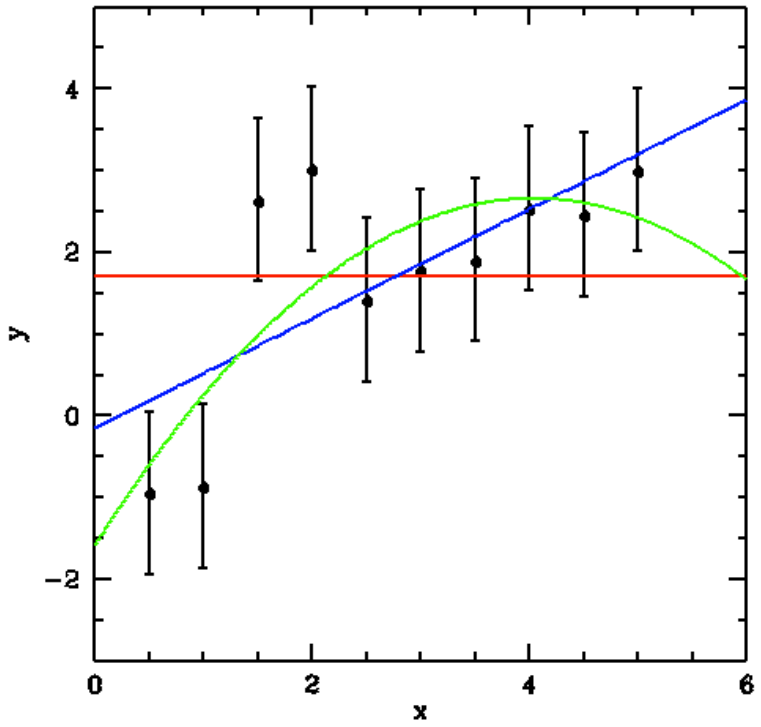
$$BF = \frac{P(\vec{x} \mid M_1)}{P(\vec{x} \mid M_2)} = \frac{\int L(\vec{x} \mid \vec{\vartheta}_1, M_1)\, \pi(\vec{\vartheta}_1 \mid M_1)\, d\vec{\vartheta}_1}{\int L(\vec{x} \mid \vec{\vartheta}_2, M_2)\, \pi(\vec{\vartheta}_2 \mid M_2)\, d\vec{\vartheta}_2}$$

In words: posterior odds = Bayes factor x prior odds
where the Bayes factor (BF) is given by the evidence ratio for the two models.
The prior ratio is often taken as unity. The evidence ratio penalizes for unnecessary complexity in the models: models are penalized if a small part of their prior parameter range matches the data (Occam factor).

Jeffrey's scale

Odds ratio	Strength of evidence
1<K<3	Barely worth mentioning
3<K<10	Substantial
10<K<30	Strong
30<K<100	Very strong
K>100	Decisive

Evidence and Bayes factors

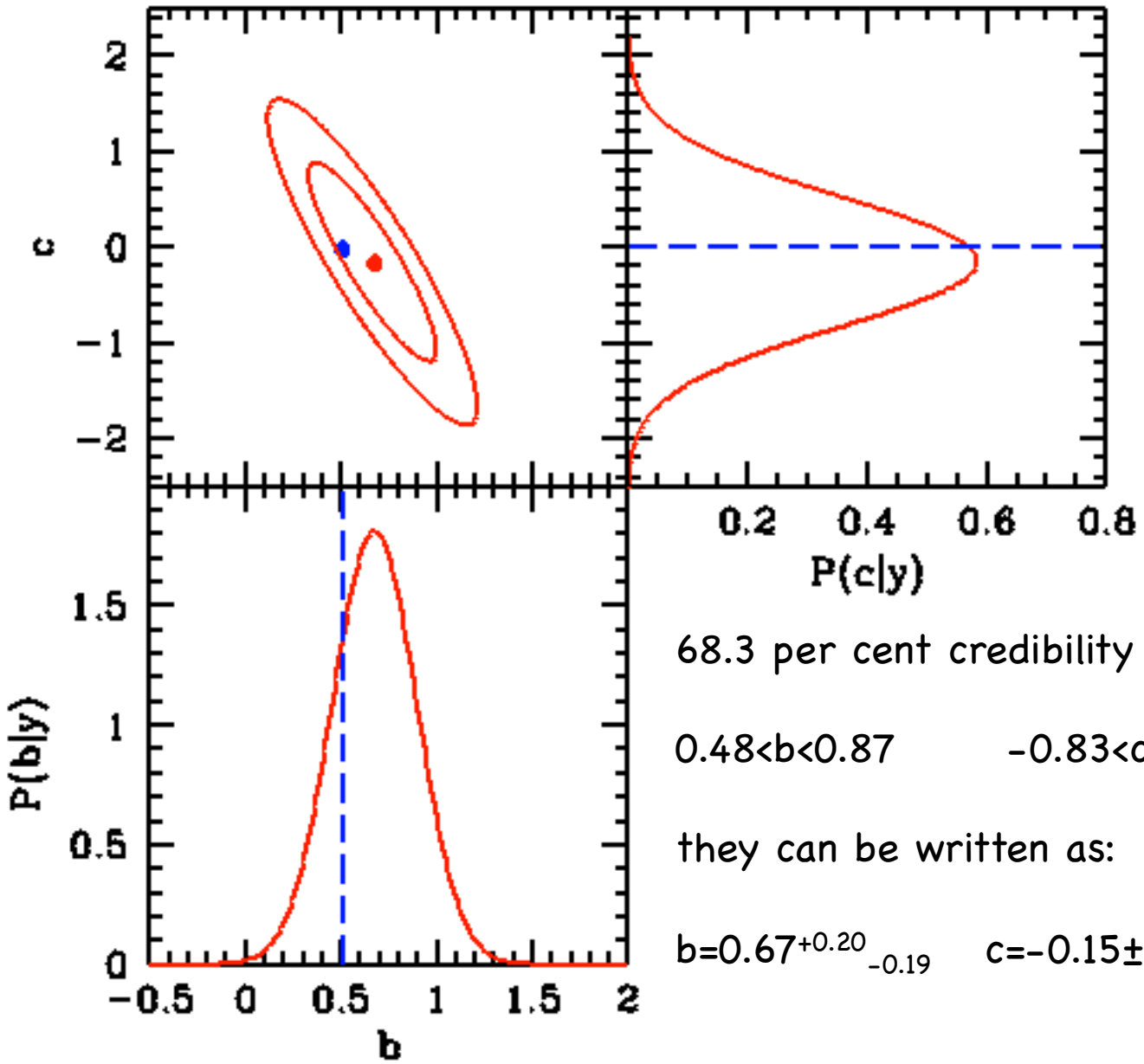


Evidence:

- Constant fit: 5.04×10^{-5}
- Linear fit: 2.92×10^{-3}
- Quadratic fit: 1.93×10^{-3}

- The linear fit is slightly preferred to the quadratic even though it has a worse χ^2_{min} (BF=1.5 while BF=57.9)

- I generated the data adding Gaussian noise with unit variance to the relation y=x/2 which was indeed linear



68.3 per cent credibility intervals

0.48<b<0.87 -0.83<c<0.53

they can be written as:

$b=0.67^{+0.20}_{-0.19}$ c=-0.15±0.68