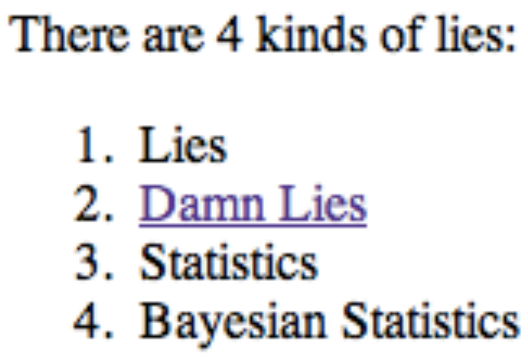
Statistics

arXiv:0712.3028; arXiv:0911.3105
Numerical recipes (the “bible”)

Licia Verde
ICREA & ICC UB-IEEC
OiU, Oslo
http://icc.ub.edu/~liciaverde



There are 4 kinds of lies:

1. Lies
2. Damn Lies
3. Statistics
4. Bayesian Statistics

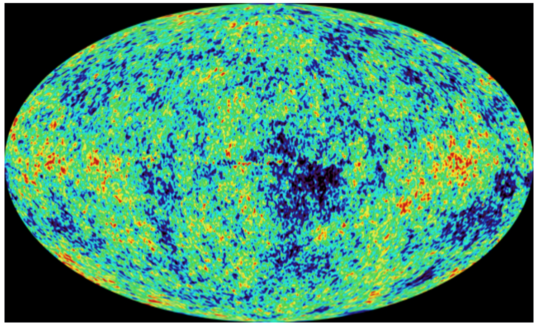
Ned Wright

outline

- Introduction Bayes vs Frequentists
- priors
- modeling and statistical inference
- Monte Carlo methods.
- Different type of errors.
- Conclusions.

What’ s is all about

DATA



?

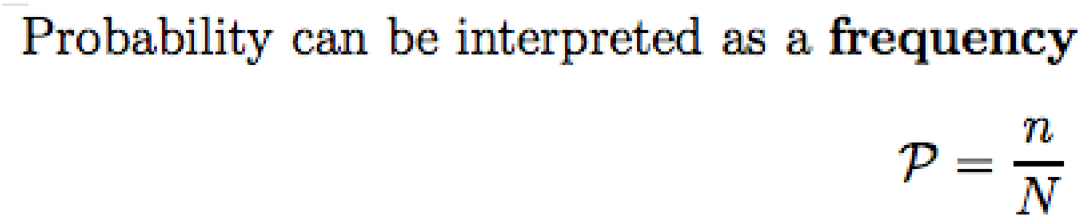
Measurement errors

Cosmic Variance

Models,
models parameters

LCDM? w? etc…

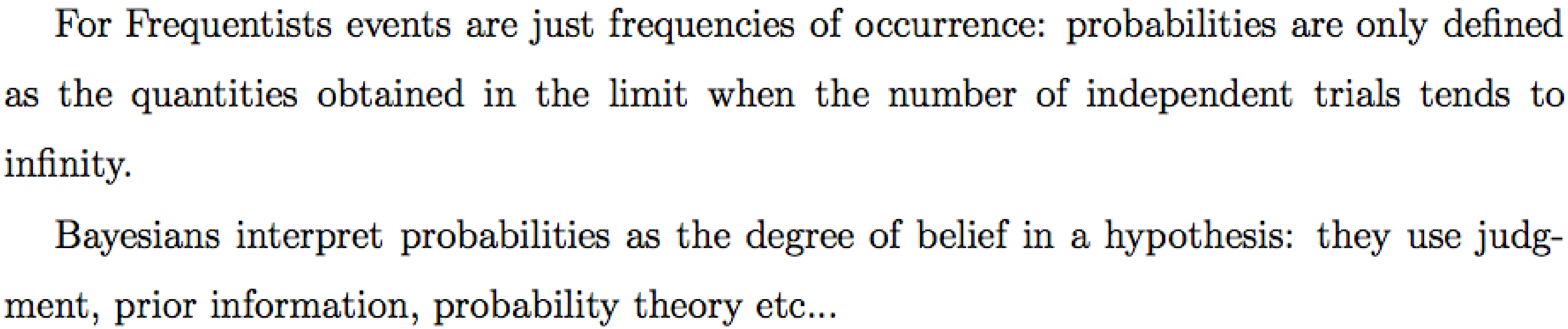
Probabilities



Probability can be interpreted as a **frequency**

$$\mathcal{P} = \frac{n}{N}$$

Frequentists vs Bayesian



For Frequentists events are just frequencies of occurrence: probabilities are only defined as the quantities obtained in the limit when the number of independent trials tends to infinity.

Bayesians interpret probabilities as the degree of belief in a hypothesis: they use judgment, prior information, probability theory etc...

Bayesians and Frequentists often criticize each other; many physicists take a more pragmatic approach about what method to use.

Probabilities (background)

Concept of Random variable x

Probability distribution $\mathcal{P}(x)$

Properties of probability distribution:

1. $\mathcal{P}(x)$ is a non negative, real number for all real values of x.

2. $\mathcal{P}(x)$ is normalized so that [1] $\int dx \mathcal{P}(x) = 1$

3. For mutually exclusive events x_1 and x_2, $\mathcal{P}(x_1 + x_2) = \mathcal{P}(x_1) + \mathcal{P}(x_2)$ the probability of x_1 or x_2 to happen is the sum of the individual probabilities. $\mathcal{P}(x_1 + x_2)$ is also written as $\mathcal{P}(x_1 U x_2)$ or $\mathcal{P}(x_1 .OR. x_2)$.

4. In general:

$$\mathcal{P}(a,b) = \mathcal{P}(a)\mathcal{P}(b|a) \quad ; \qquad \mathcal{P}(b,a) = \mathcal{P}(b)\mathcal{P}(a|b) \qquad\qquad \mathcal{P}(a,b) = \mathcal{P}(b,a).$$

For independent events then $\mathcal{P}(a,b) = \mathcal{P}(a)\mathcal{P}(b)$.

Ex. Produce examples of this last case

We might want to add:

$$P(a) = \sum_{b} P(a, b)$$

Useful later when talking about marginalization

Bayes theorem

prior Likelihood

$$\mathcal{P}(H|D) = \frac{\mathcal{P}(H)\mathcal{P}(D|H)}{\mathcal{P}(D)}$$

Posterior

From

$$\mathcal{P}(a,b) = \mathcal{P}(a)\mathcal{P}(b|a) \quad ; \qquad \mathcal{P}(b,a) = \mathcal{P}(b)\mathcal{P}(a|b)$$

Fundamental difference here; “statistical INFERENCE”

Prior: how do you chose P(H)? Back to this later.

Drawbacks: Examples, discussion

r log r

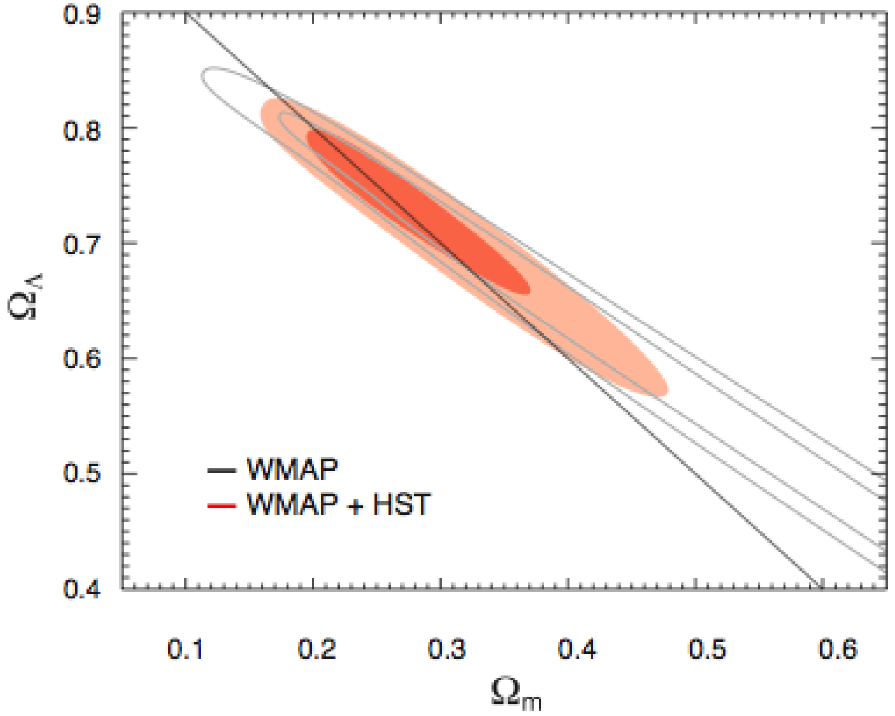
τ log τ exp(-2 τ)

comparing $\mathcal{P}(x)$ with $\mathcal{P}(f(x))$;

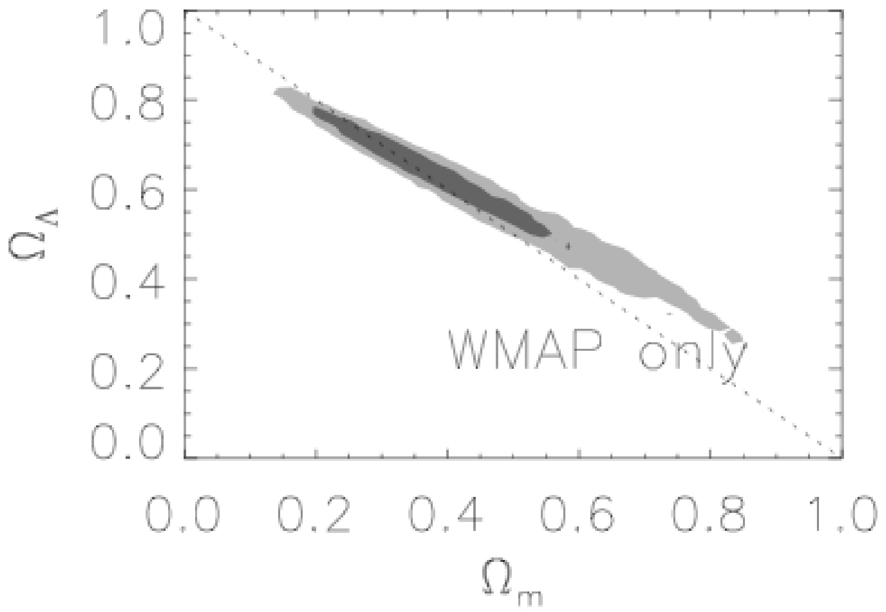
$$\mathcal{P}(f) = \mathcal{P}(x(f)) \left|\frac{df}{dx}\right|^{-1}$$

?

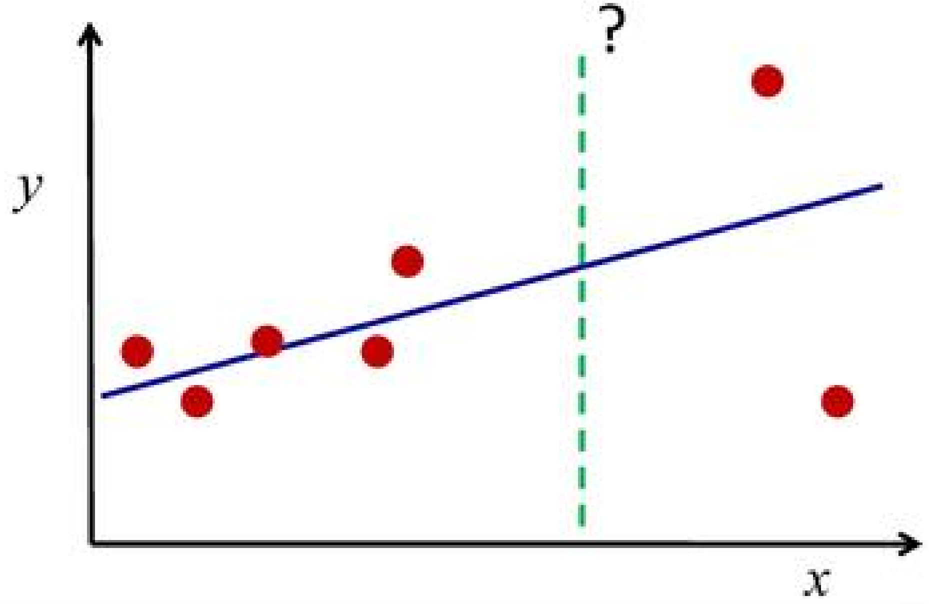
Spergel et al 2007



Spergel et al 2003



The importance of the prior



Priors are not generally bad!

Random fields, probabilities and Cosmology

Average statistical properties

Particulary important: $\delta(\vec{x}) = \delta\rho(\vec{x})/\rho$

Ensamble: all the possible realizations of the true underlying Universe

Inference: examples

The Cosmological principle: models of the universe are homogeneous on average; in widely separated regions of the Universe the density field has the same statistical properties

A crucial assumption: we see a fair sample of the Universe

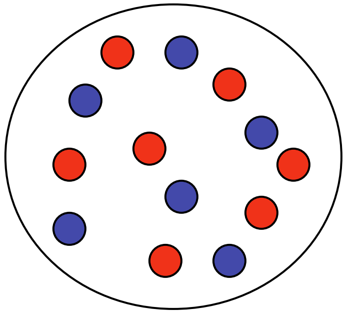
Ergodicity then follows: averaging over many realizations is equivalent to averaging over a large(enough) volume

Tools… statistics! Correlation functions etc…

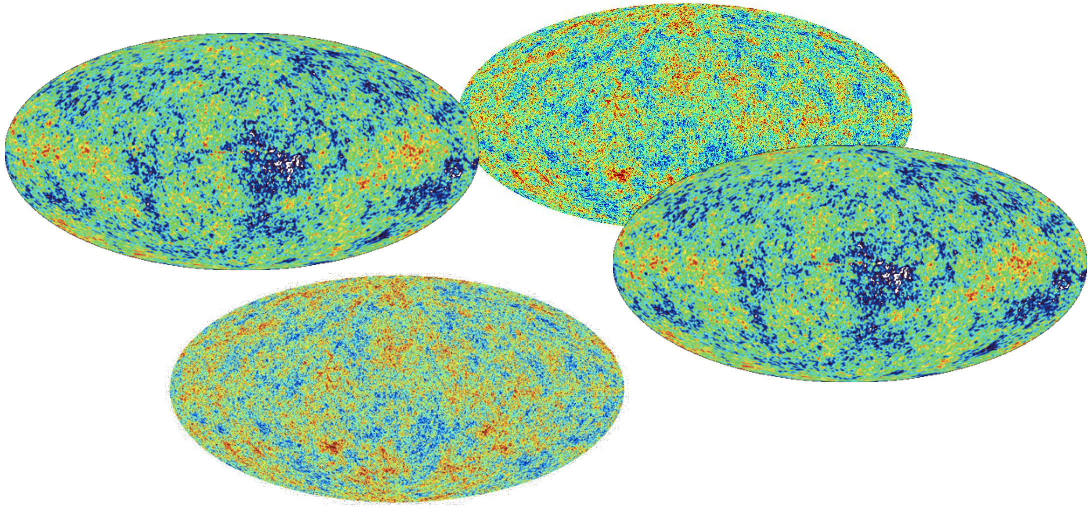
Big advantage of being Bayesian

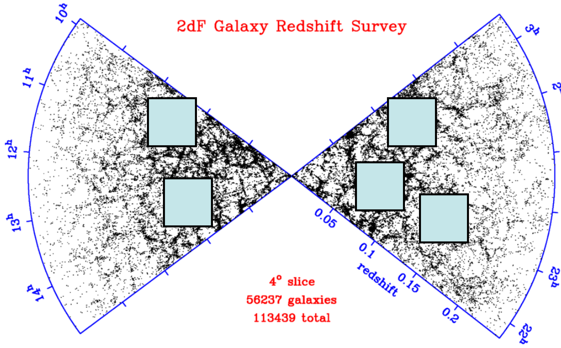
- Urn example

(in reality NOT transparent)



Cosmic variance



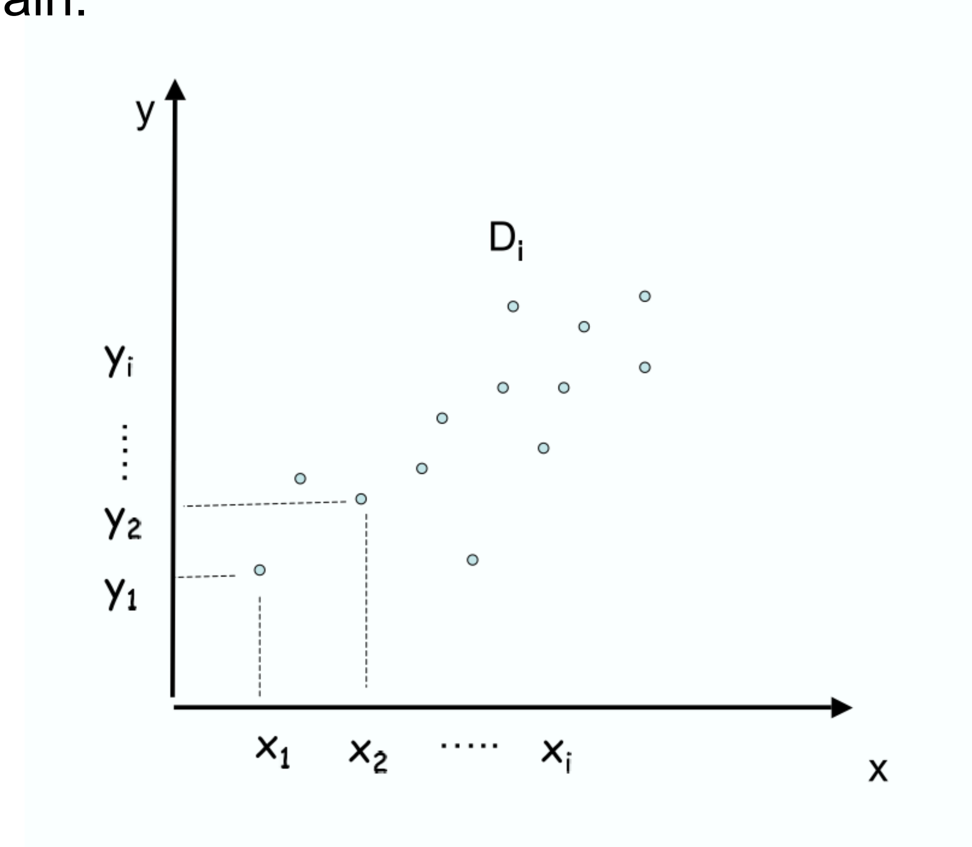


Modeling of data and Statistical inference

Read numerical recipes chapter 15, read it again, then when you have to apply all this, read it again.

example

Fit this with a line



Need a "figure of merit"

Least squares….

What you want:

- Best fit parameters
- Error estimates on the parameters
- A statistical measure of the goodness of fit (possibly)

Bayesian: “what is the probability that a particular set of parameters is correct?”

Figure of merit: “given a set of parameters this is the probability of occurrence of the data”

Least squares fit….

$$\chi^2 = \sum_i w_i [D_i - y(x_i|\vec{\alpha})]^2$$

you can show that the minimum variance weights are $w_i = 1/\sigma_1^2$.

And what if data are correlated?

$$\chi^2 = \sum_{ij} (D_i - y_i) F_{ij} (D_j - y_j) = (\vec{D} - \vec{y}) C^{-1} (\vec{D} - \vec{y})$$

In general: chi-squared

Goodness of fit?

If all is Gaussian, the probability of χ^2 at the minimum follows a χ^2 distribution, with ν=n-m degrees of freedom

# data points

#parameters

$$\mathcal{P}(\chi^2 < \hat{\chi}^2, \nu) = \mathcal{P}(\nu/2, \hat{\chi}^2/2) = \Gamma(\nu/2, \hat{\chi}^2/2)$$

Incomplete gamma function

$$Q = 1 - \mathcal{P}(\nu/2, \hat{\chi}^2/2)$$

Goodness of fit if evaluated at the best fit

Too small Q?

a) Model is wrong! Try again…

b) Real errors are larger

c) non-Gaussian

In general Monte-Carlo simulate….

Too large Q?

a) Errors overestimated

b) Neglected covariance?

c) Non-Gaussian (almost never..)

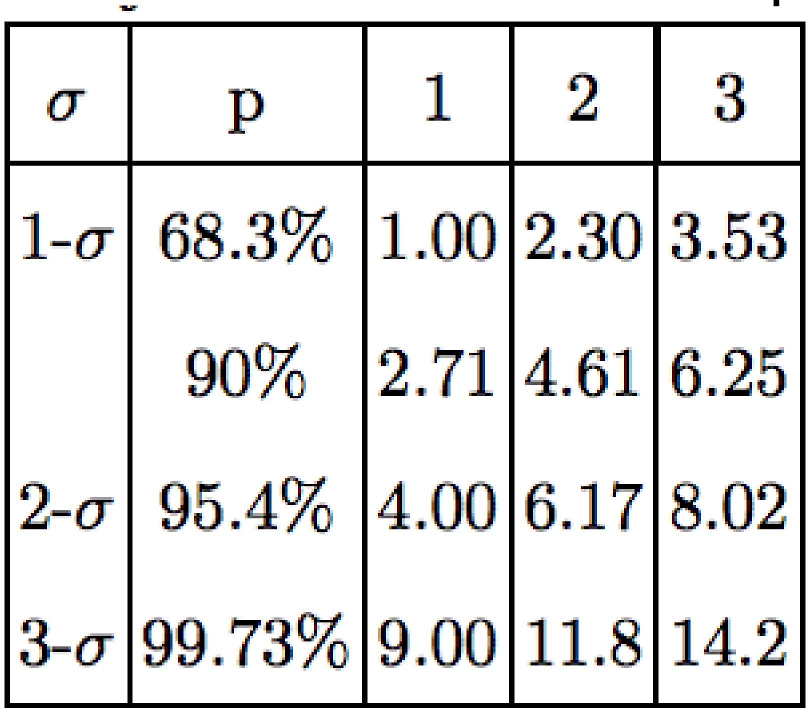
P.S chi-by-eye?

Confidence regions

If m is the number of fitted parameters for which you want to plot the joint confidence region and p is the confidence limit desired, find the $\Delta\chi^2$ such that the probability of a chi-Square variable with m degrees of freedom being less than $\Delta\chi^2$ is p. Use the Q function above.

Confidence regions

Number of parameters



σ	p	1	2	3
1-σ	68.3%	1.00	2.30	3.53
	90%	2.71	4.61	6.25
2-σ	95.4%	4.00	6.17	8.02
3-σ	99.73%	9.00	11.8	14.2

$\Delta\chi^2$

Joint confidence levels

Likelihoods

Remember Bayes …

$$\mathcal{P}(H|D) = \frac{\mathcal{P}(H)\mathcal{P}(D|H)}{\mathcal{P}(D)}$$

set $\mathcal{P}(D) = 1$ Back to this later

In many cases, can invoke the central limit theorem

a multi-variate Gaussian:

$$\mathcal{L} = \frac{1}{(2\pi)^{n/2}|detC|^{1/2}} \exp\left[-\frac{1}{2}\sum_{ij}(D-y)_i C_{ij}^{-1}(D-y)_j\right]$$

where $C_{ij} = \langle (D_i - y_i)(D_j - y_j)\rangle$ is the covariance matrix.

Confidence levels

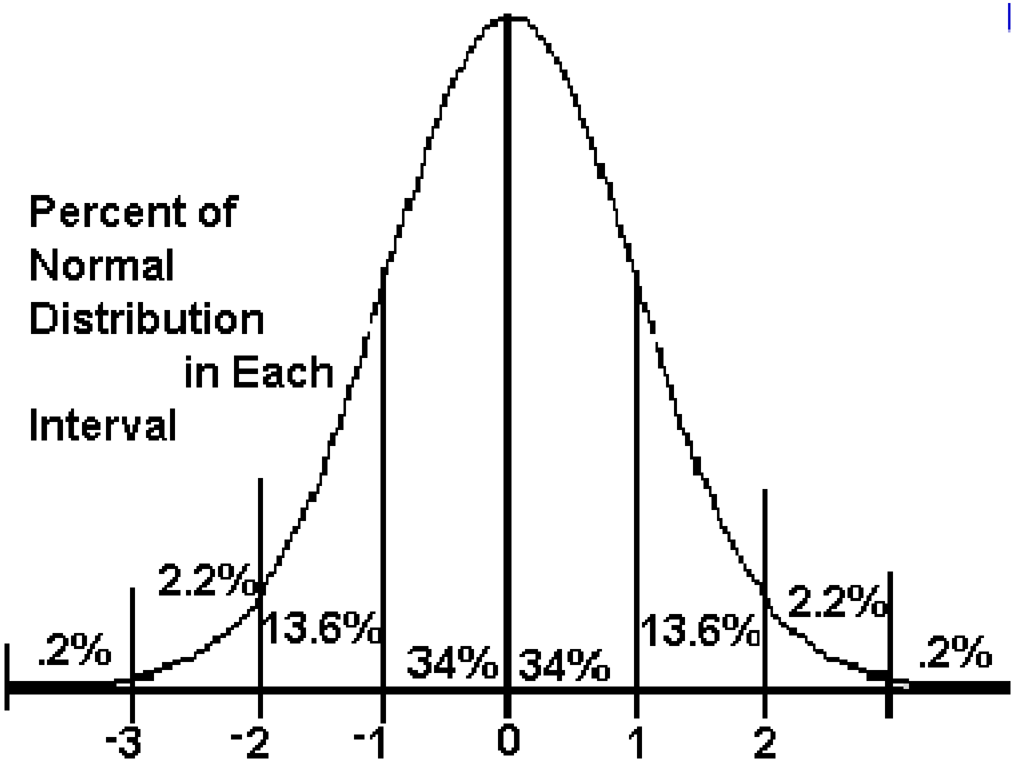
Bayesians $\int_R \mathcal{P}(\vec{\alpha}|D)d\vec{\alpha}$ =0.683.. or 0.95… or…

Integrating over the hypothesis

Classical: likelihood ratio

$$-2\ln\left[\frac{\mathcal{L}(\vec{\alpha})}{\mathcal{L}_{max}}\right] \leq \mathrm{threshold}$$

visually



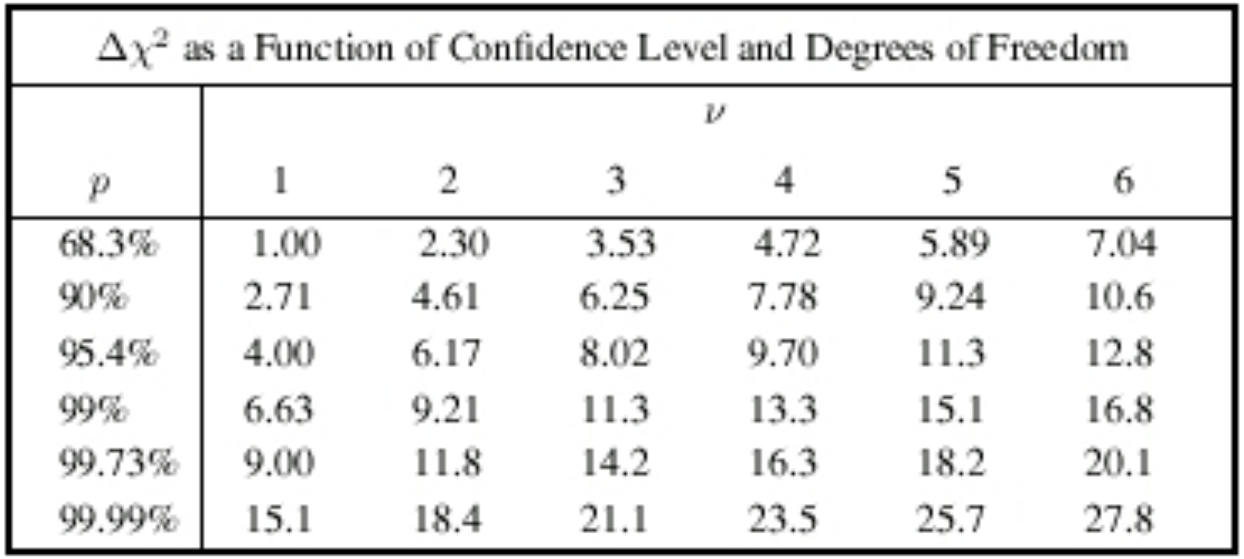
In higher dimensions….

Questions for you

- in what simple case can you make an easy identification of the likelihood ratio with the chi-square?

- In what case can you make an easy identification between the two approaches?

There is a BIG difference between χ^2 & reduced χ^2



$\Delta\chi^2$ as a Function of Confidence Level and Degrees of Freedom

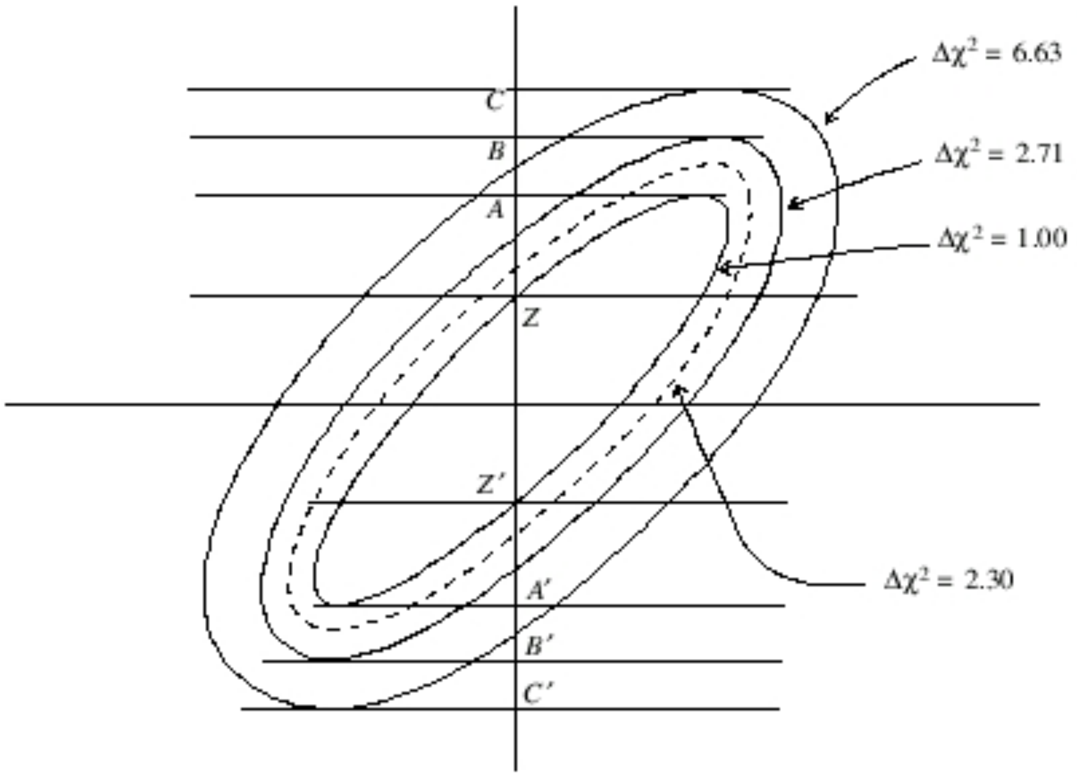
	ν					
p	1	2	3	4	5	6
68.3%	1.00	2.30	3.53	4.72	5.89	7.04
90%	2.71	4.61	6.25	7.78	9.24	10.6
95.4%	4.00	6.17	8.02	9.70	11.3	12.8
99%	6.63	9.21	11.3	13.3	15.1	16.8
99.73%	9.00	11.8	14.2	16.3	18.2	20.1
99.99%	15.1	18.4	21.1	23.5	25.7	27.8

Only for multivariate Gaussian with constant covariance

$$-2\ln \mathcal{L} = \chi^2$$

From: “Numerical recipes” Ch. 15

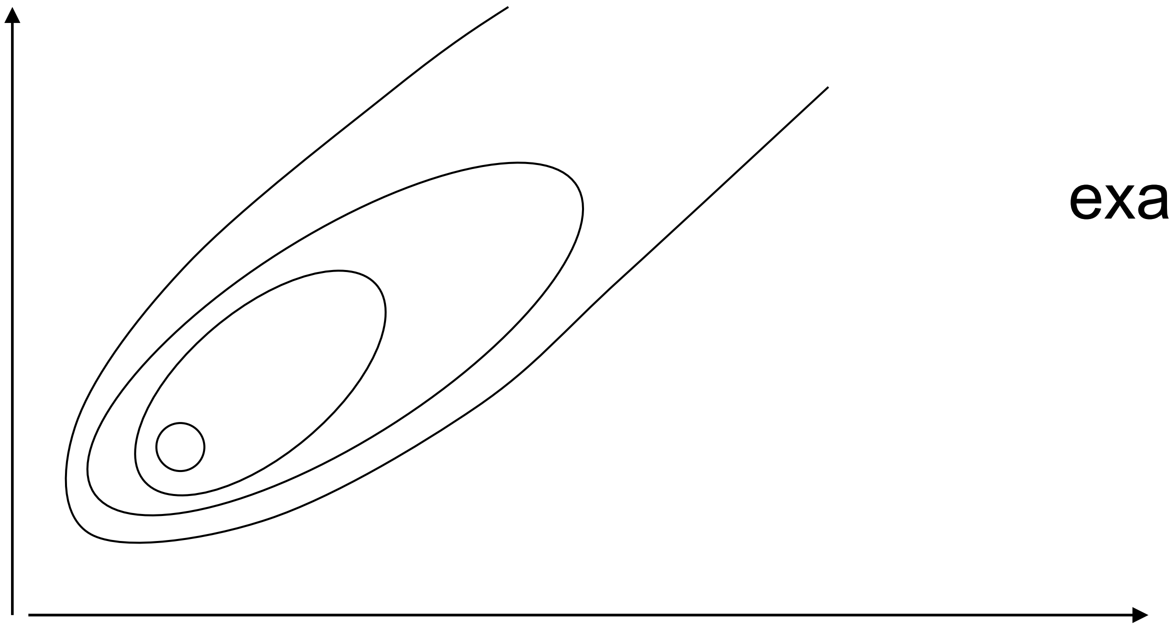
If likelihood is Gaussian and Covariance is constant



Example: for multi-variate Gaussian

Marginalization

$$P(\alpha_1..\alpha_j|D) = \int d\alpha_{j+1}, ...d\alpha_m P(\vec{\alpha}|D)$$



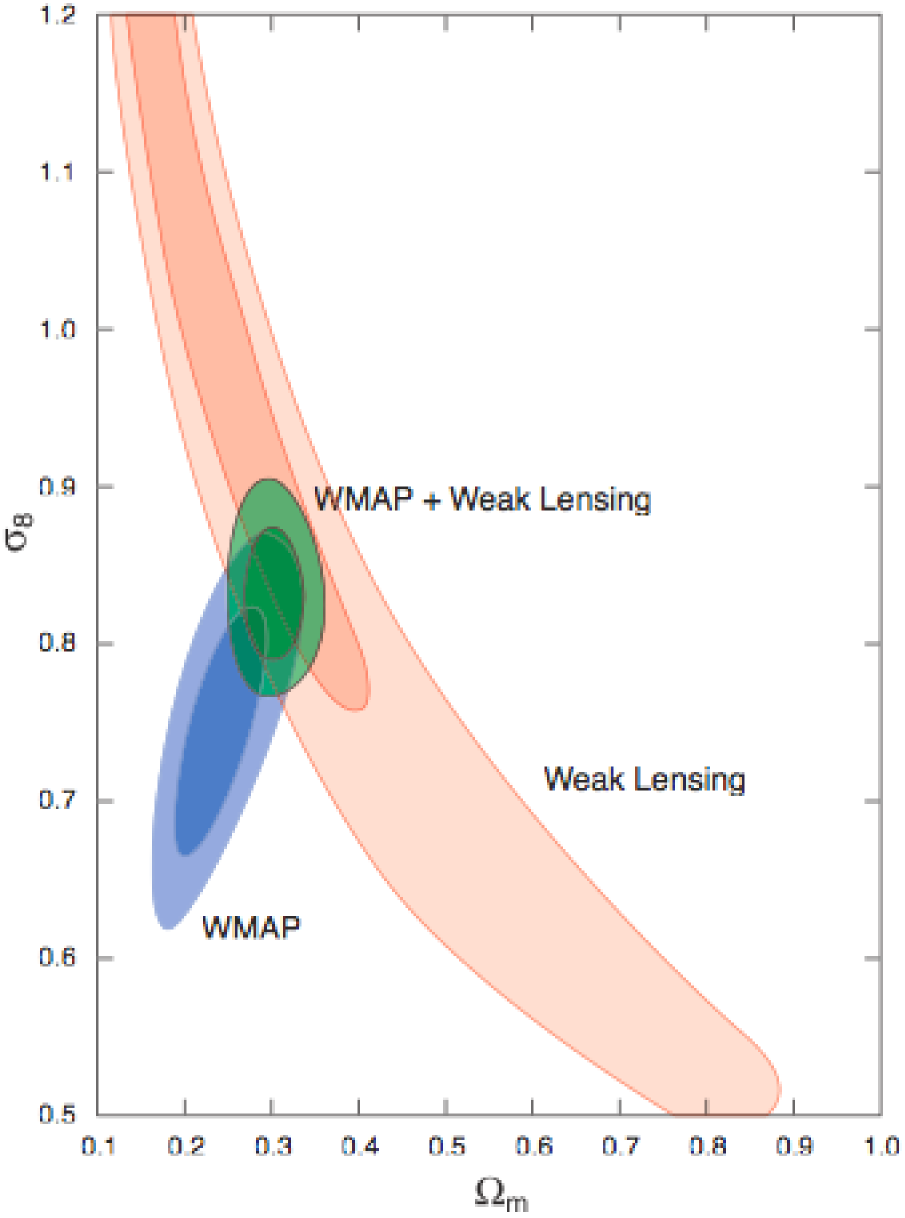
example

Other data sets

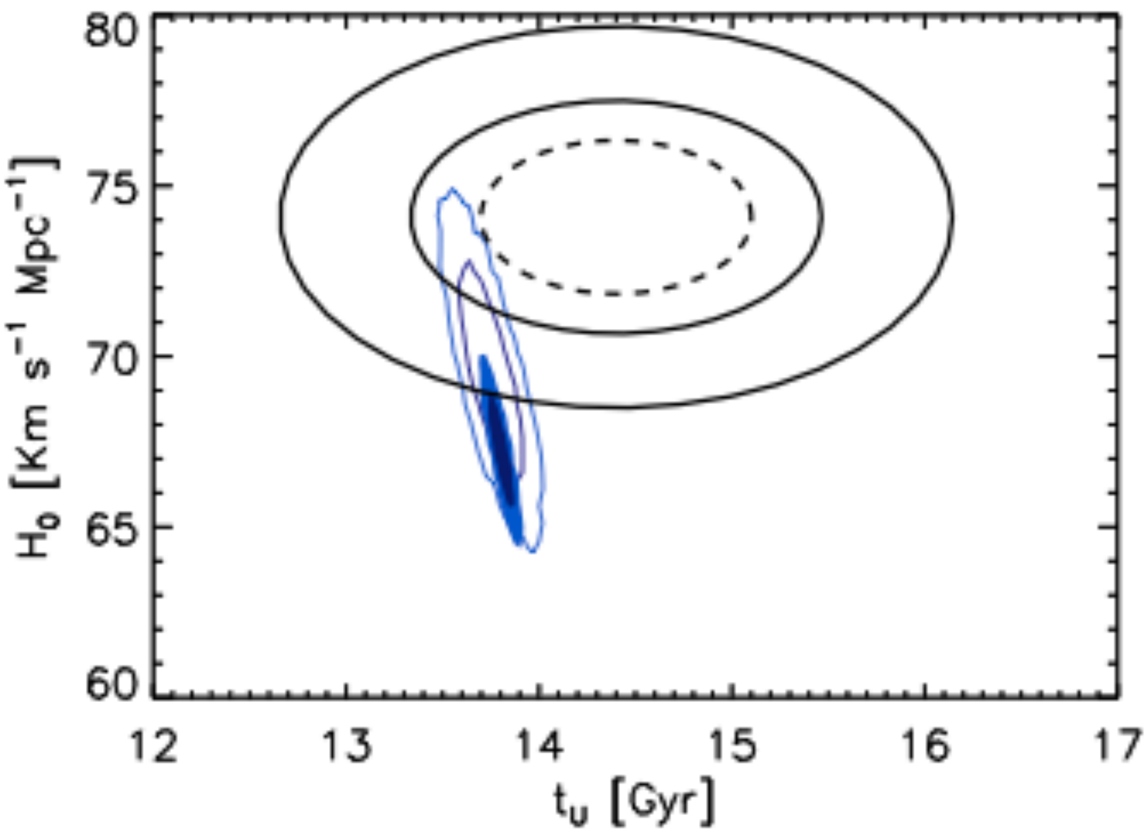
If independent, multiply the two likelihoods

(can use some of them as “priors”)

Beware of inconsistent experiments!



Spergel 2007



Lv, Protopapas, Jimenez, 2013

Useful trick for Gaussian likelihoods

e.g. marginalizing over point source amplitude

$$P(\alpha_1..\alpha_{m-1}|D) = \int \frac{dA}{(2\pi)^{\frac{m}{2}}||C||^{\frac{1}{2}}} e^{\left[-\frac{1}{2}(C_i-(\hat{C}_i+AP_i))\Sigma_{ij}^{-1}(C_j-(\hat{C}_j+AP_j))\right]}$$

$$\times \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2}\frac{(A-\hat{A})^2}{\sigma^2}\right]$$

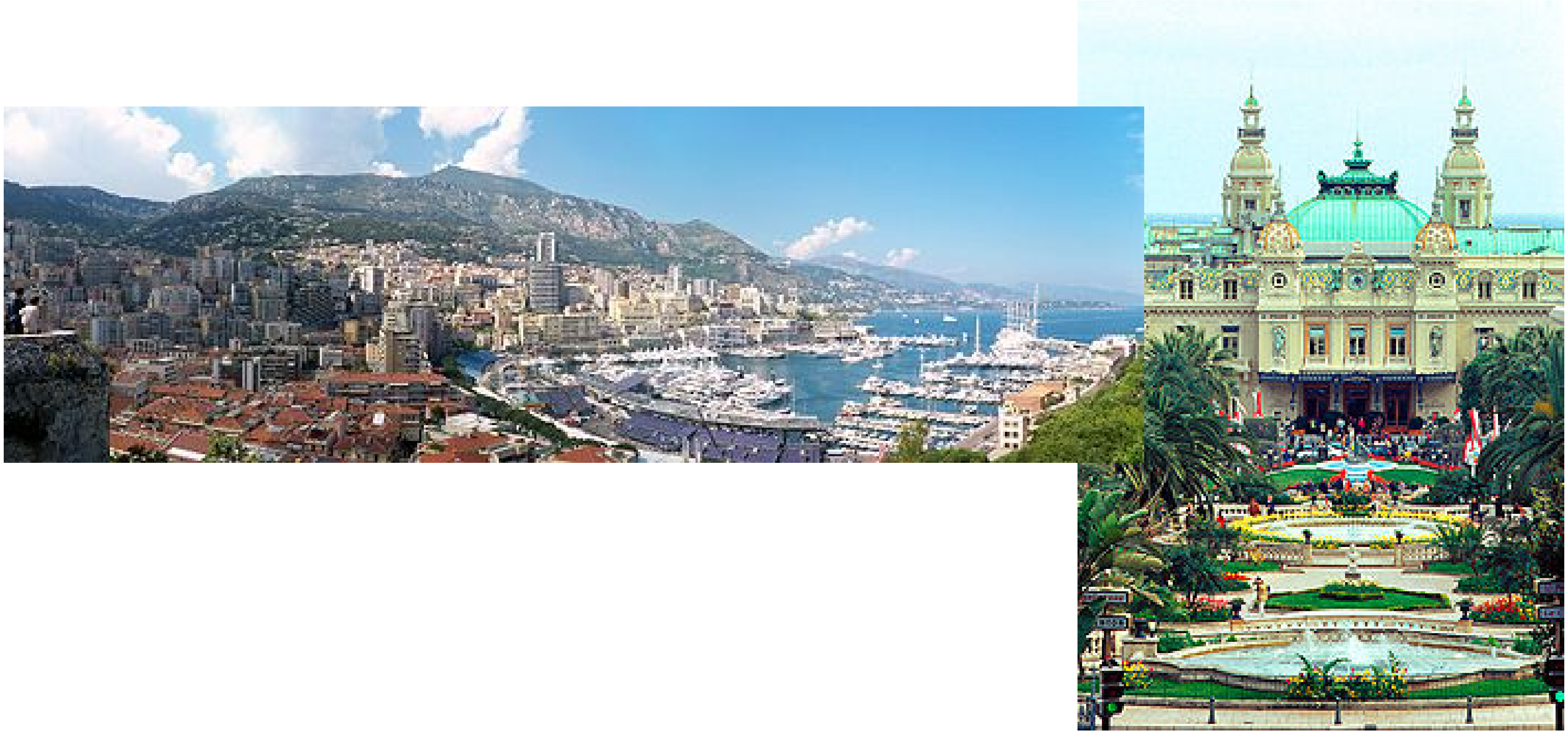
The trick is to recognize that this integral can be written as:

$$P(\alpha_1..\alpha_{m-1}|D) = C_0 \exp\left[-\frac{1}{2}C_1 - 2C_2A + C_3A^2\right] dA$$

$$\text{substitution } A \longrightarrow A - C_2/C_3$$

result $\propto \exp[-1/2(C_1 - C_2^2/C_3)]$

Monte Carlo methods

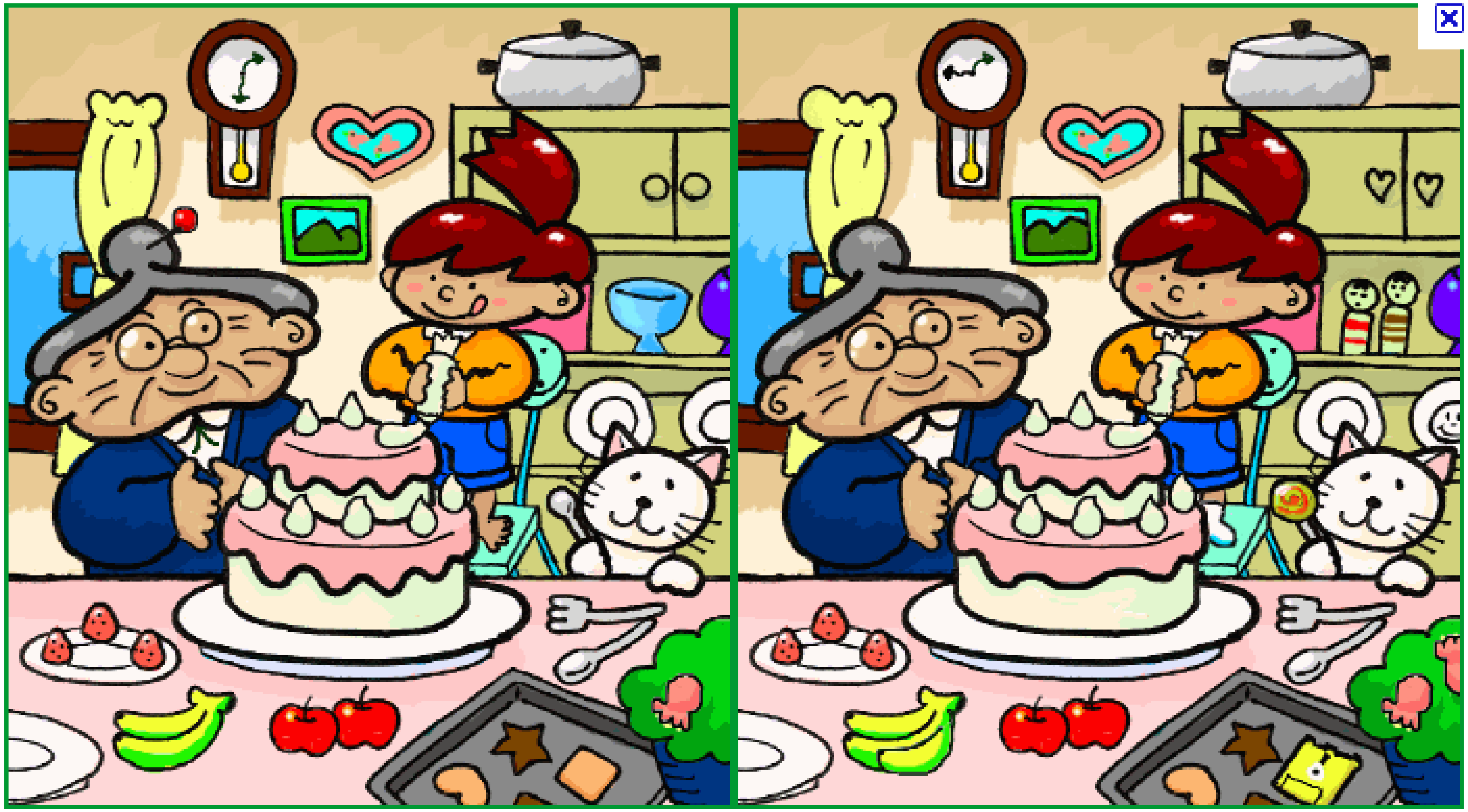


Monte Carlo methods

a) Monte Carlo error estimation

b) Monte Carlo Markov Chains

Your brain does it!



Spot the differences…

Intro to:
Monte Carlo

Simple problem: what’s the mean of a large number of objects?

What’s the mean height of people in Madrid?

$$\sum_{i=1}^{N} \frac{h_i}{N}$$

If N is very large this is untractable soo…

If n<<N but still a fair sample, great!

$$\sim \sum_{i=1}^{n} \frac{h_i}{n}$$

In probability: $\int f(x)P(x)dx \sim \frac{1}{S}\sum^{S} f(x^s)$ if $x^s \sim P(x)$

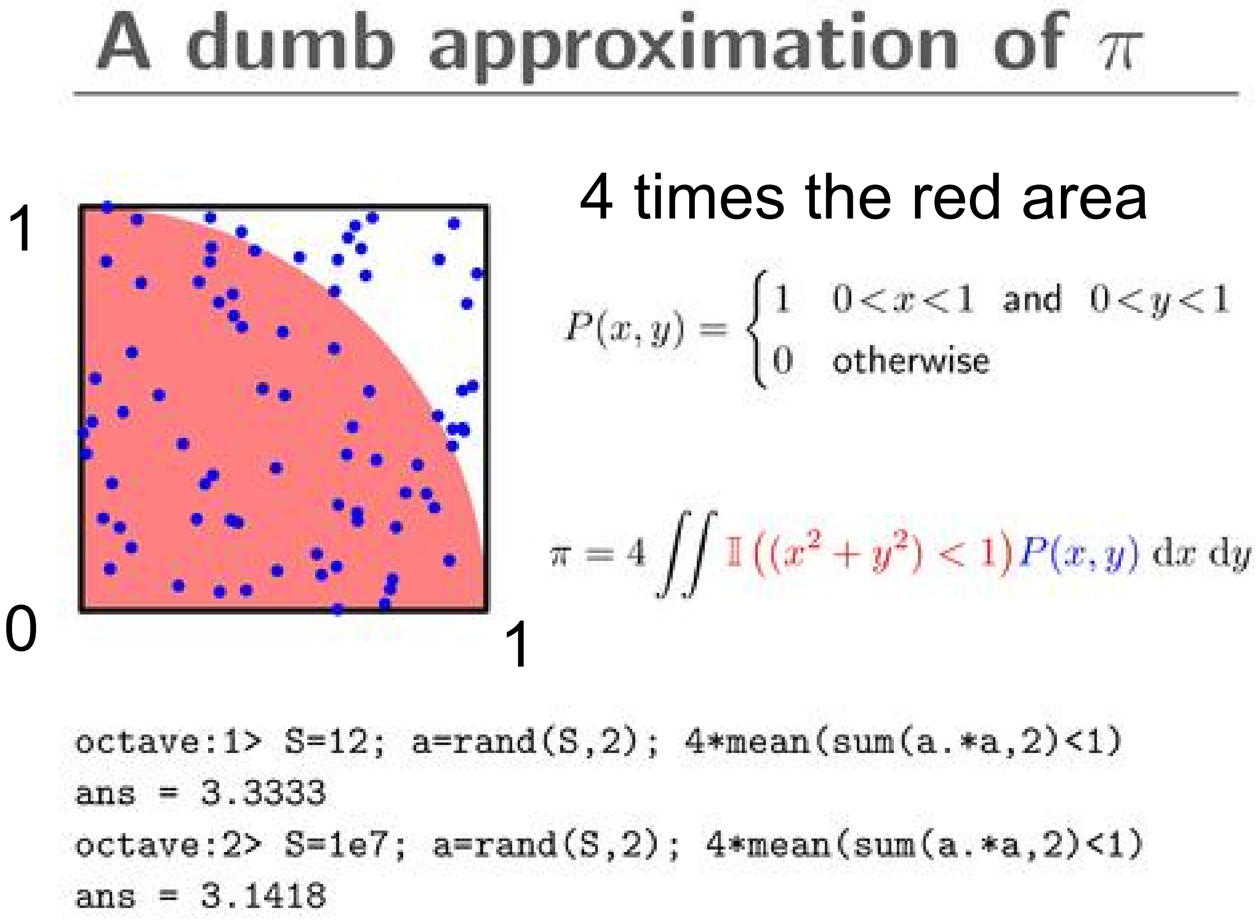
In Bayesian inference:

$$p(x|D) = \int P(x|\theta, D)P(\theta|D)d\theta \sim \frac{1}{S}\sum P(x|\theta^s, D) \quad \text{if } \theta^s \sim P(\theta|D)$$

You can show that:

The estimator is unbiased
 and you can quantify the variance of the estimator:
The error shrinks like $S^{1/2}$

Very simple example:



A dumb approximation of π

4 times the red area

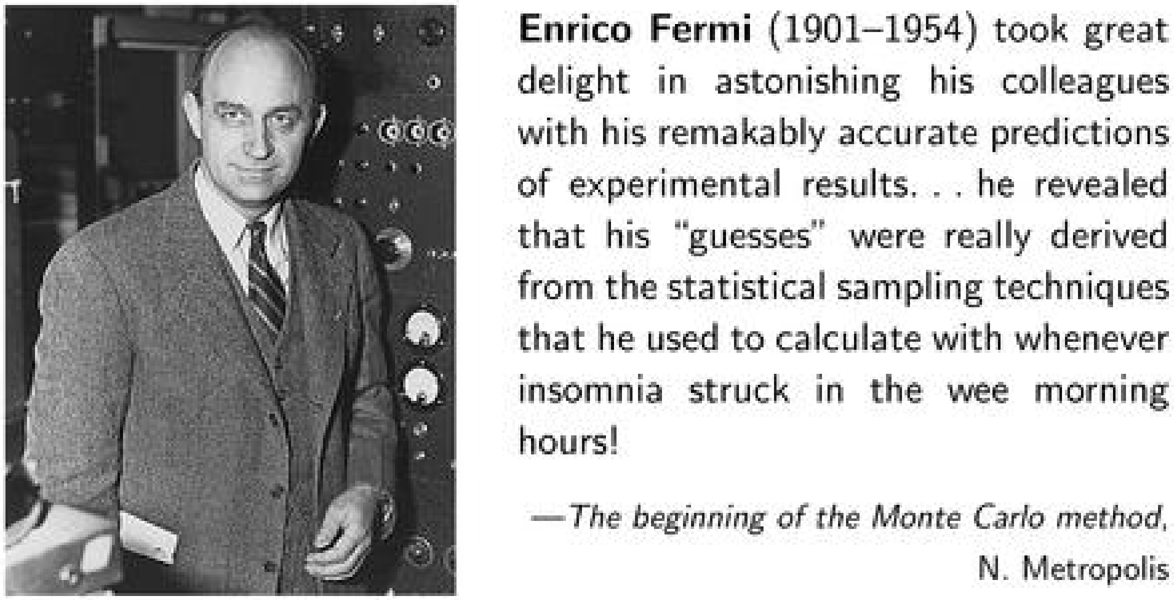
$$P(x,y) = \begin{cases} 1 & 0<x<1 \text{ and } 0<y<1 \\ 0 & \text{otherwise} \end{cases}$$

$$\pi = 4 \iint \mathbb{I}\left((x^2+y^2)<1\right) P(x,y)\,\mathrm{d}x\,\mathrm{d}y$$

```
octave:1> S=12; a=rand(S,2); 4*mean(sum(a.*a,2)<1)
ans = 3.3333
octave:2> S=1e7; a=rand(S,2); 4*mean(sum(a.*a,2)<1)
ans = 3.1418
```

There are better ways to compute π, so use mcmc only when right to use…

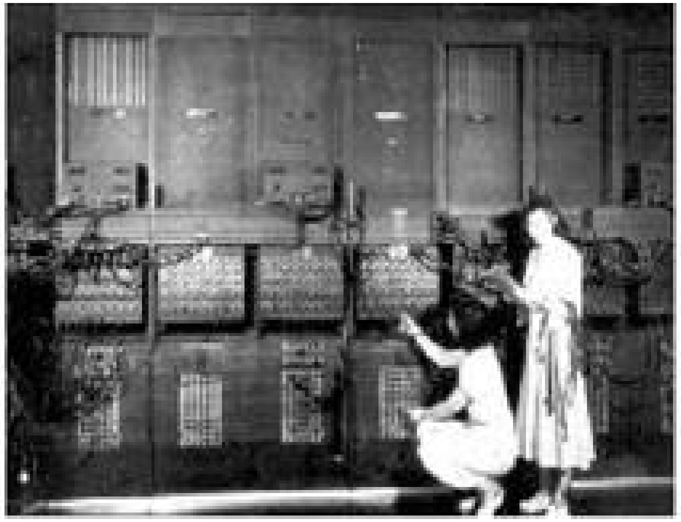
Historical note



Enrico Fermi (1901–1954) took great delight in astonishing his colleagues with his remakably accurate predictions of experimental results. . . he revealed that his "guesses" were really derived from the statistical sampling techniques that he used to calculate with whenever insomnia struck in the wee morning hours!

—*The beginning of the Monte Carlo method,*
N. Metropolis

history

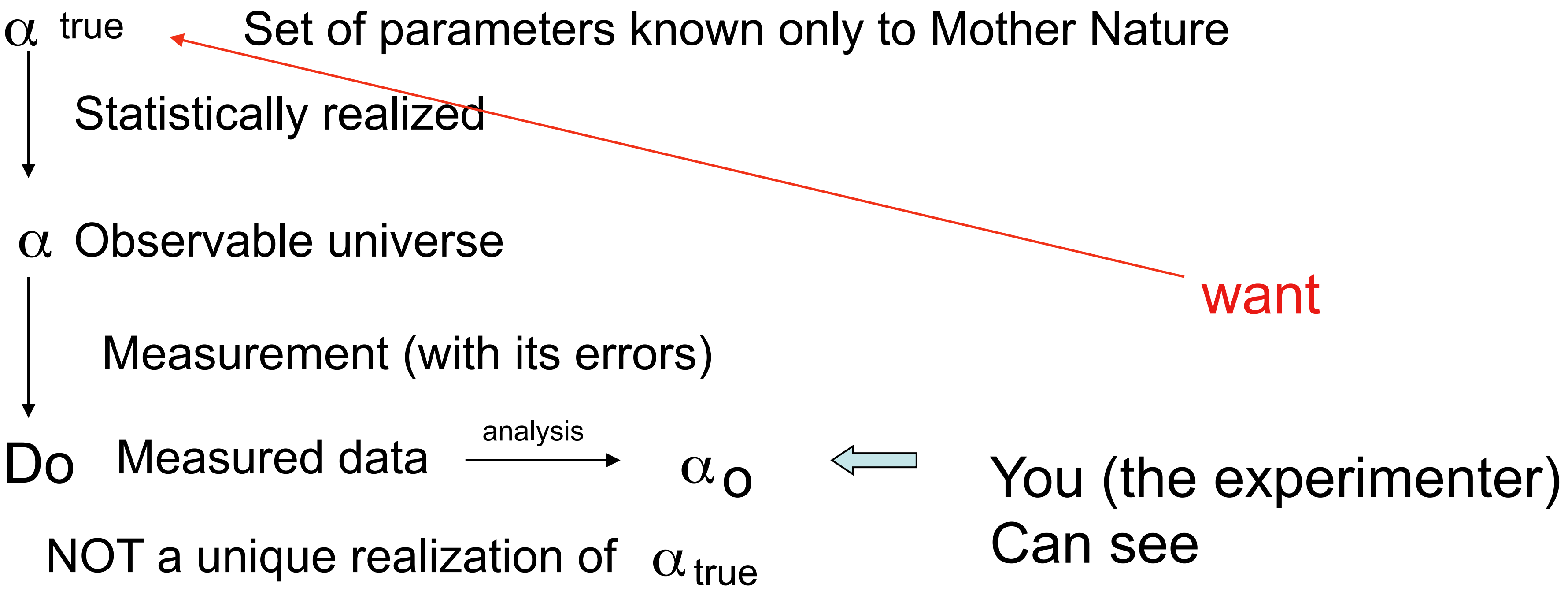


Monte Carlo methods

a) Monte Carlo error estimation

Back to parameter estimation and confidence regions

Conceptual interpretation in cosmology



α_{true} Set of parameters known only to Mother Nature

Statistically realized

α Observable universe

want

Measurement (with its errors)

D_o Measured data $\xrightarrow{\text{analysis}}$ α_o ⇐ You (the experimenter) Can see

NOT a unique realization of α_{true}

There could be infinitely many realizations (hypothetical data sets) $D_1, D_2, ...$

Each one with best fit parameters $\alpha_1, \alpha_2, ...$

Expect: $< \alpha_i >= \alpha_{\rm true}$

If I knew the distribution of $\alpha_i - \alpha_{\rm true}$ That' d be all I need

Trick: say that (hope) $\alpha_0 \sim \alpha_{\rm true}$

In many cases we can simulate the distribution of $\alpha_i - \alpha_0$

Make many synthetic realizations of universes where α_0 is the truth; mimic the observational process in all these mock universes, estimate the best fit parameters from each; map $\alpha_S - \alpha_0$

Very important tool

How to sample from the probability distribution?

- For some well known univariate probability distributions there are numerical routines http://cg.scs.carleton.ca/~luc/rnbookindex.html
- In other cases there may be numerical techniques to sample P(x) [more later]
- Importance sampling: (if you know how to sample from Q but not from P)

$$\int f(x)P(x)dx = \int f(x)\frac{P(x)}{Q(x)}Q(x)dx \sim \frac{1}{S}\sum_{s=1}^{S} f(x^s)\frac{P(x^s)}{Q(x^s)} \text{ if } x^s \sim Q(x)$$

Some Q are more suitable for P than others….

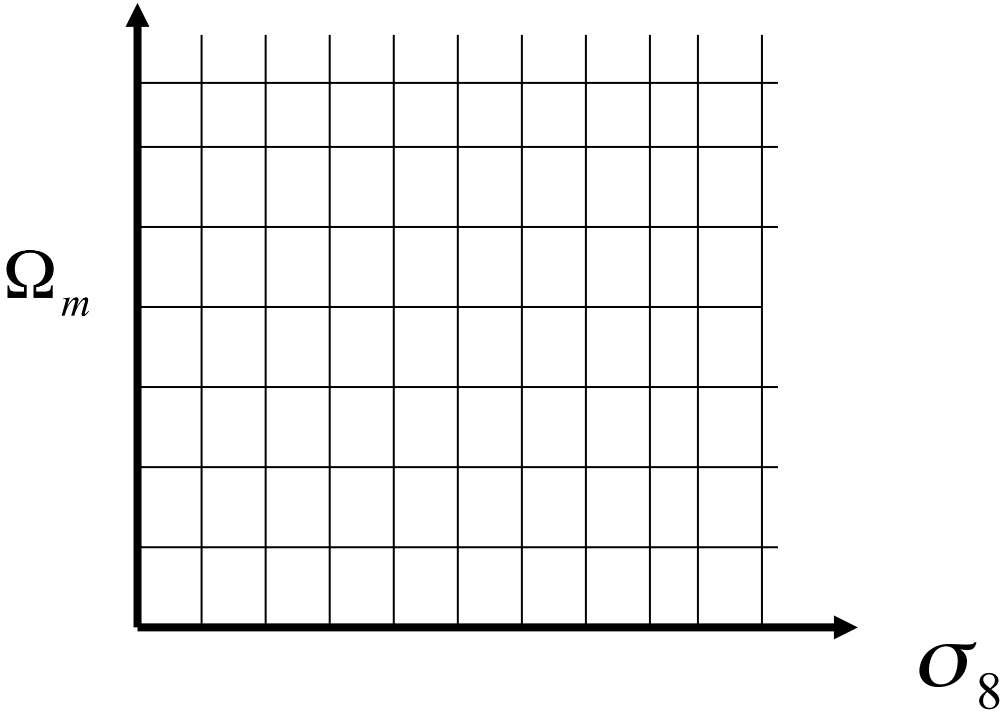
b) Monte Carlo Markov Chains

Explore likelihood surface

Grid-based approach

Operationally:

e.g., 2 params: 10 x 10



What if you have (say) 6 parameters?

6 params. 20 pixels/dim
= 6.7×10^7 evals

say 1.6 s/eval

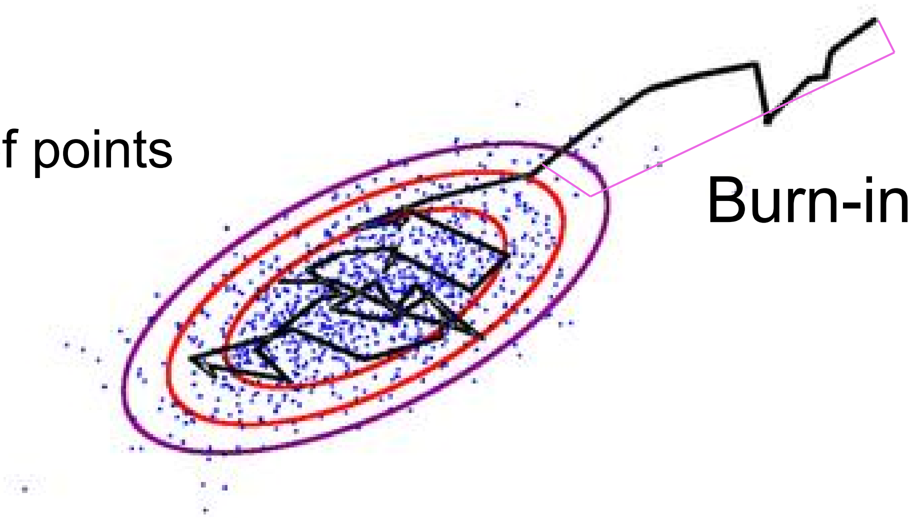
~1200 days!

You’ve got a problem !

Monte Carlo Markov Chains

So you have a higher-dimensional probability distribution,
you want to sample in a way proportional to it ,
with a random walk

Goal: density of points
proportional to
the probability



Start at an
arbitrary point

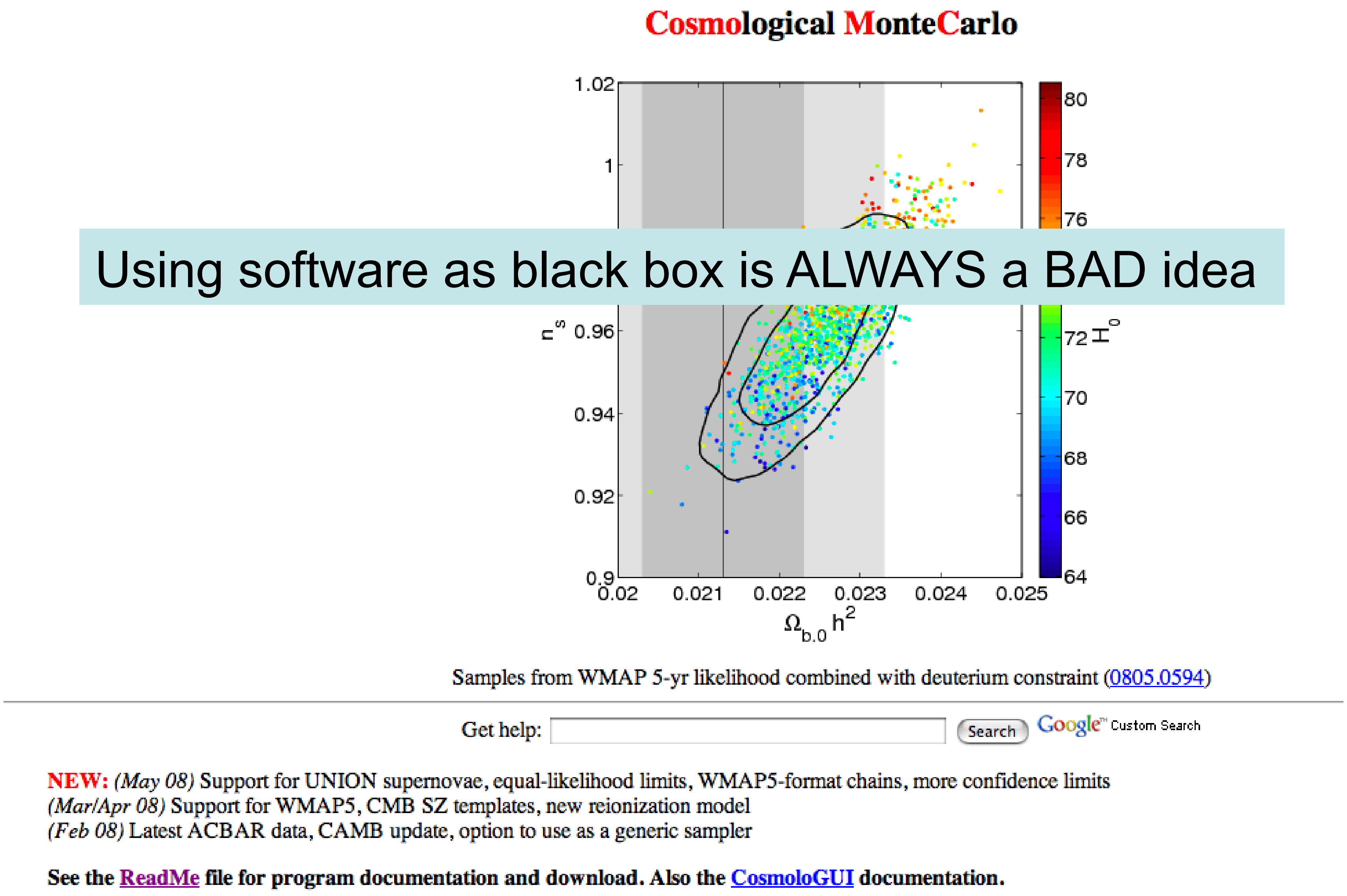
Take Markov steps

Burn-in

MCMC gives approximated, correlated samples from the target distribution

Monte Carlo Markov Chains

http://cosmologist.info/cosmomc/



Cosmological MonteCarlo

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

Get help: [ ] Search Google™ Custom Search

NEW: *(May 08)* Support for UNION supernovae, equal-likelihood limits, WMAP5-format chains, more confidence limits
(Mar/Apr 08) Support for WMAP5, CMB SZ templates, new reionization model
(Feb 08) Latest ACBAR data, CAMB update, option to use as a generic sampler

See the ReadMe file for program documentation and download. Also the CosmoloGUI documentation.

Using software as black box is ALWAYS a BAD idea

Markov Chain Monte Carlo (MCMC)

Standard in CMB analyses (publicly available COSMOMC)

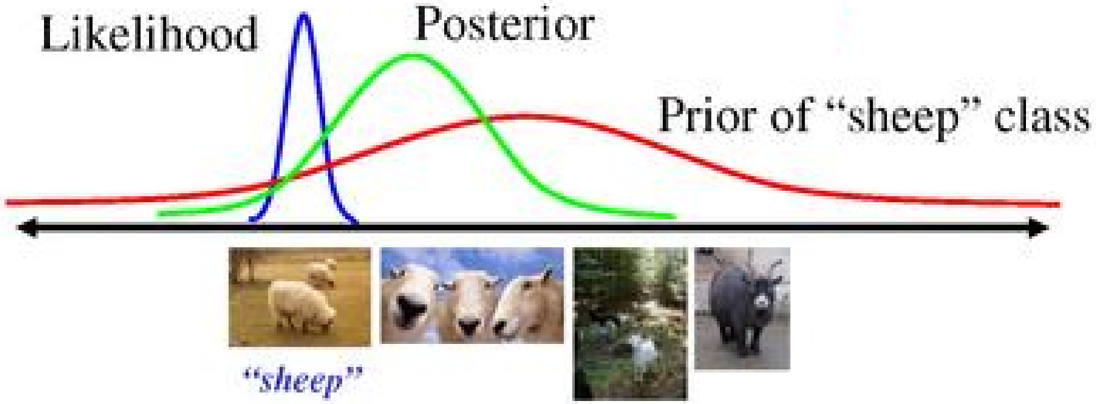
Simulate

$$p(h|d) = \frac{p(d|h)\,p(h)}{\sum_{h'\in H} p(d|h')\,p(h')}$$

Set of cosm. Params

$$C_\ell^{\rm th}$$

Bayes



Generate random draws from the posterior that are a fair sample of the probability(Likelihood) surface

Markov Chain Monte Carlo (MCMC)

Random walk in parameter space

At each step, sample one point in parameter space

The density of sampled points $\propto$ posterior distribution

(grid example)
FAST: before 10^7 likelihood evaluations, now< 10^5

marginalization is easy:
just project points and recompute their density

Adding external data sets is often very easy

Operationally (Metropolis-Hastings):

1. Start at a random location in parameter space: $\alpha_i^{\text{old}}\ \mathcal{L}^{\text{old}}$

2. Try to take a random step in parameter space: $\alpha_i^{\text{new}}\mathcal{L}^{\text{new}}$

3a. If $\mathcal{L}^{\text{new}} \geq \mathcal{L}^{\text{old}}$ Accept (take and save) the step, "new"--> "old" and go to 2.

3b. If $\mathcal{L}^{\text{new}} < \mathcal{L}^{\text{old}}$ Draw a random number x uniform in 0,1

If $x \geq \dfrac{\mathcal{L}^{\text{new}}}{\mathcal{L}^{\text{old}}}$ do not take the step (i.e. save "old") and go to 2.

If $x < \dfrac{\mathcal{L}^{\text{new}}}{\mathcal{L}^{\text{old}}}$ do as in 3a.

KEEP GOING….

"Take a random step"

The probability distribution of the step is the "**proposal distribution**", which you should not change once the chain has started.

The proposal distribution (the step-size) is crucial to the MCMC efficiency.

Steps too small step poor mixing

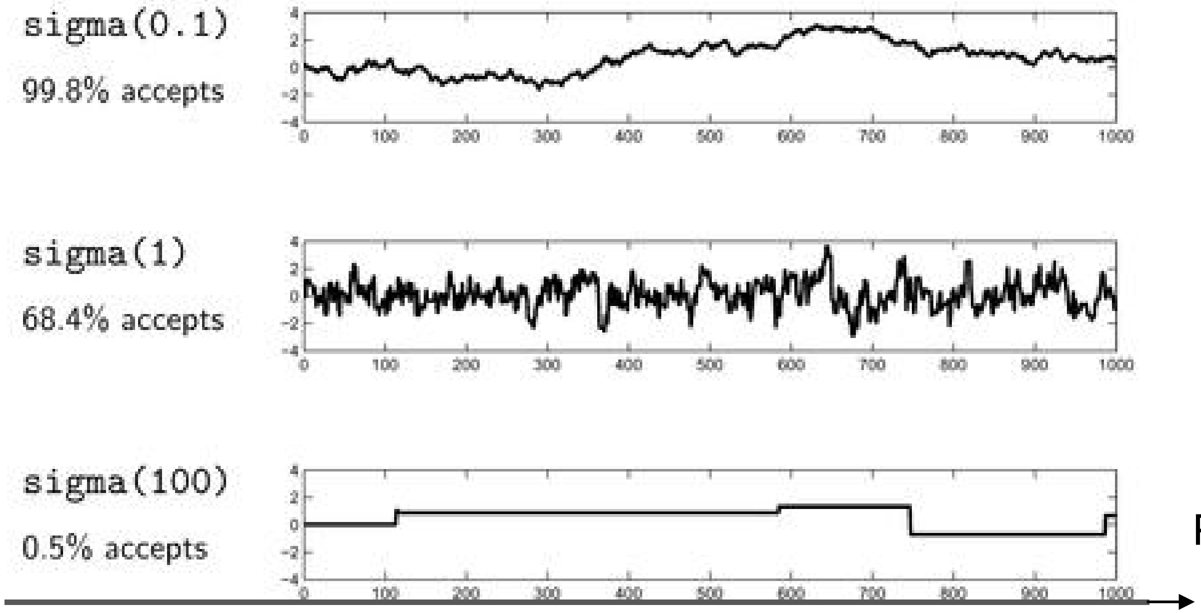
Steps too big step poor acceptance rate

"fair sample of the likelihood surface", remember?

The importance of stepsize

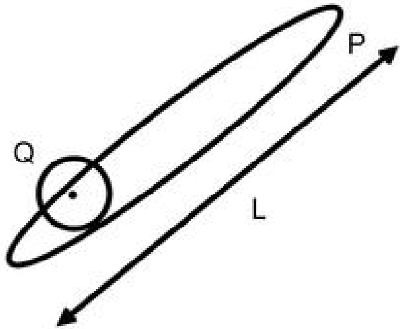
Likelihood

Poor exploration

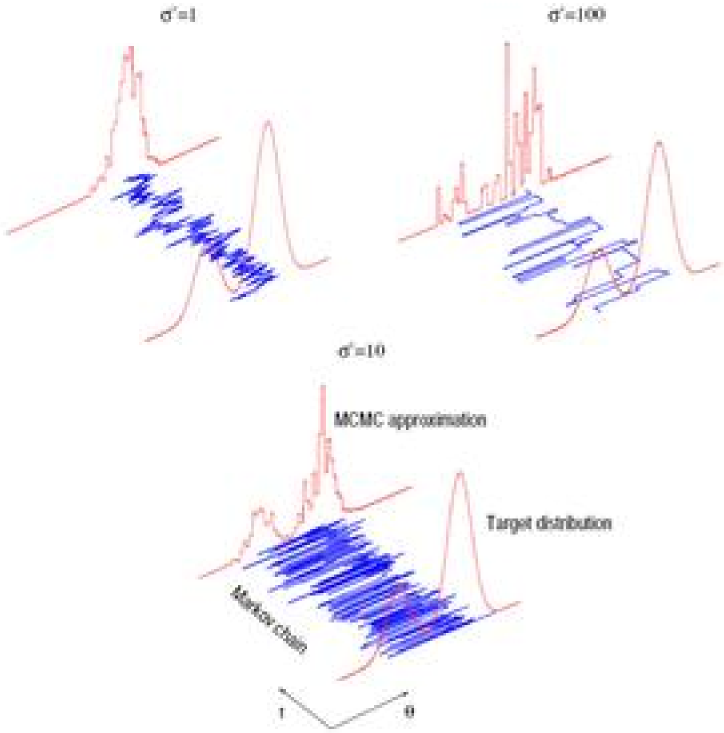


Poor exploration

Step number



The importance of stepsize

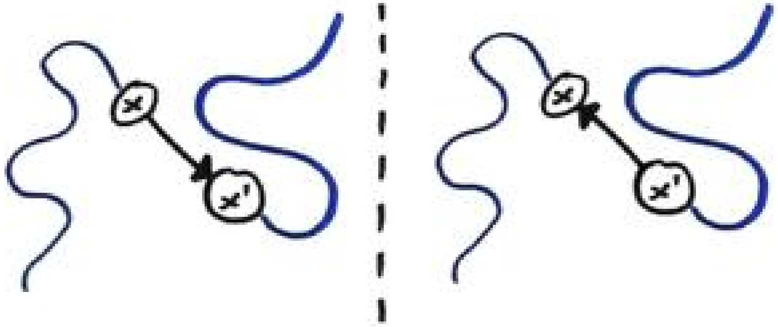


Take a random step

For statisticians: transition operators

Detailed balance: (beware of boundaries….)

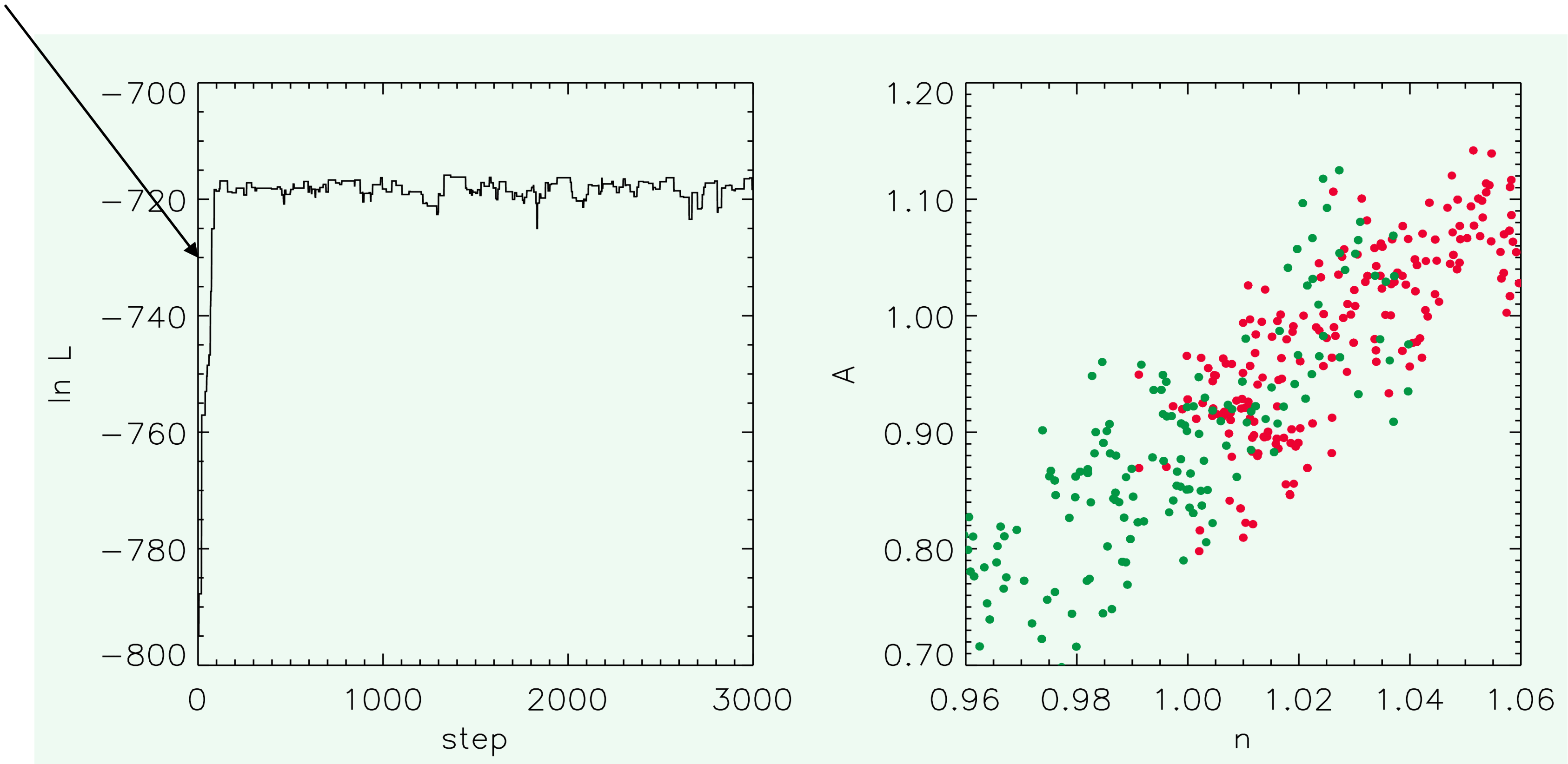
Detailed balance means $\rightarrow x \rightarrow x'$ and $\rightarrow x' \rightarrow x$ are equally probable:



When the MCMC has forgotten about the starting location and has well explored the parameter space you’re ready to do parameter estimation.

USE a MIXING and CONVERGENCE criterion!!!

Burn-in



Gelmans and Rubin convergence

Recommended: start 4 to 8 chains at well separated points

M chains, N elements

Chain mean $$\bar{y}^j = \frac{1}{N}\sum_{i=1}^{N} y_i^j,$$

Vector with parameters value

Mean of distrib. $$\bar{y} = \frac{1}{NM}\sum_{ij=1}^{NM} y_i^j .$$

Variance between chains $$B_n = \frac{1}{M-1}\sum_{j=1}^{M}(\bar{y}^j - \bar{y})^2$$

And within $$W = \frac{1}{M(N-1)}\sum_{ij}(y_i^j - \bar{y}^j)^2$$

$$\boxed{\hat{R} = \frac{\frac{N-1}{N}W + B_n\left(1+\frac{1}{M}\right)}{W}}$$

Always >1 by construction

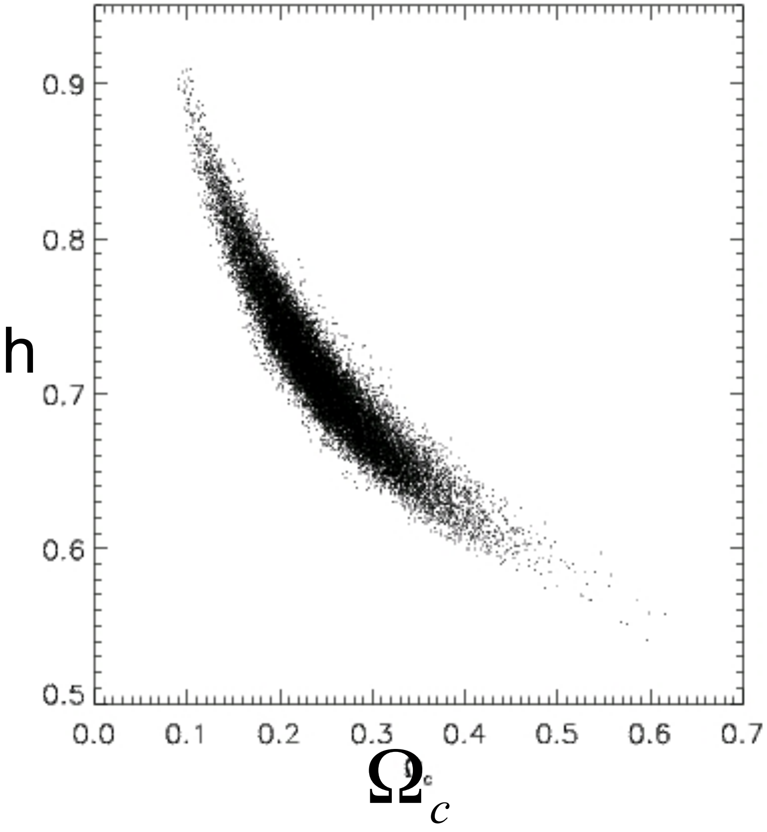
Require <1.03

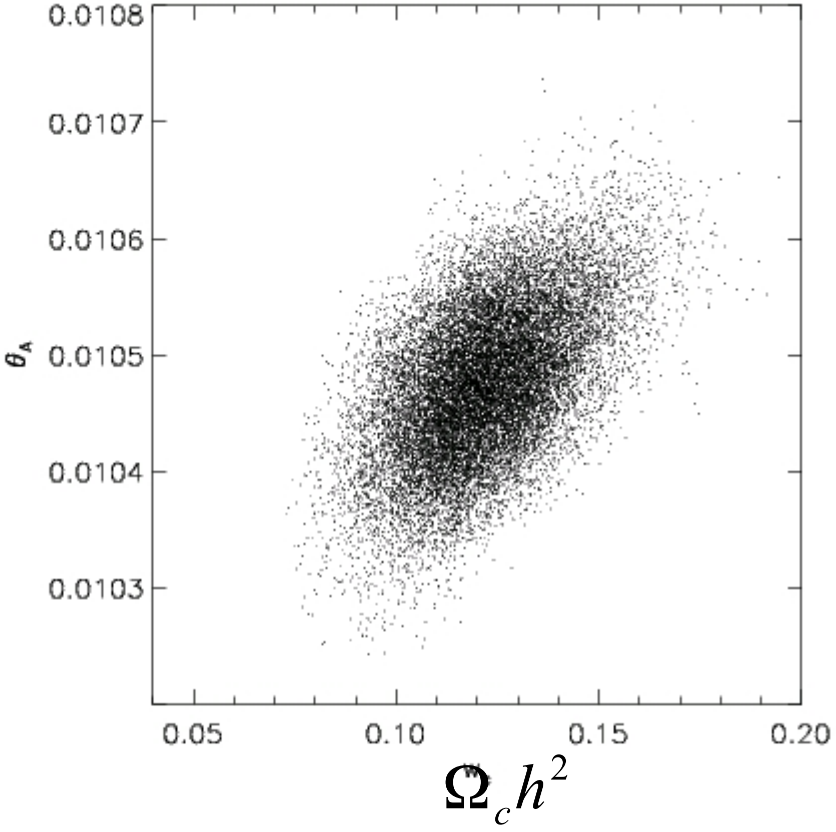
Unconverged chains are just nonsense

Metropolis-Hastings is NOT the only implementation,

Other options are:
Gibbs Sampler
Rejection method
Hamiltonian Monte-Carlo
Simulated annealing (though you do not get an MCMC)

Beware of DEGENERACIES





Reparameterization. e.g., Kososwsky et al. 2002

$$\theta_A = \frac{r_s(a_{dec})}{D_A(a_{dec})}$$

Even “better”:

Cosmomc has the option of computing the covariance for the parameters
Find the axis of the multi dim. degeneracies
perform a rotation and re-scaling to obtain azimutally symmetric contours

An improve MCMC efficiency by factor of up to 10

It is still a linear operation

Where’ s the prior ?

Once you have the MCMC output:

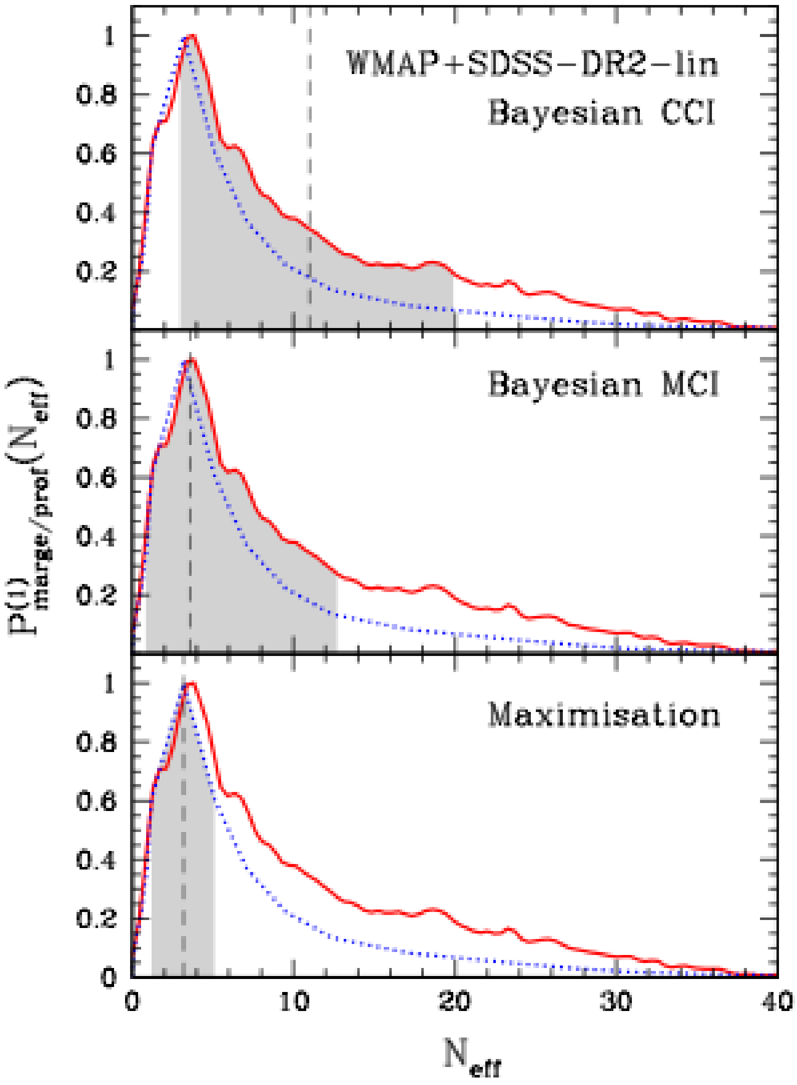
- The density of points in parameter space gives you the posterior distribution
- To obtain the marginalized distribution, just project the points
- To obtain confidence intervals, - integrate the "likelihood" surface

  -compute where e.g. 68.3% of points lie
- To each point in parameter space sampled by the MCMC give a weight proportional to the number of times it was saved in the chain
- To add to the analysis another dataset (that does not require extra parameters) renormalize the weight by the "likelihood" of the new data set.

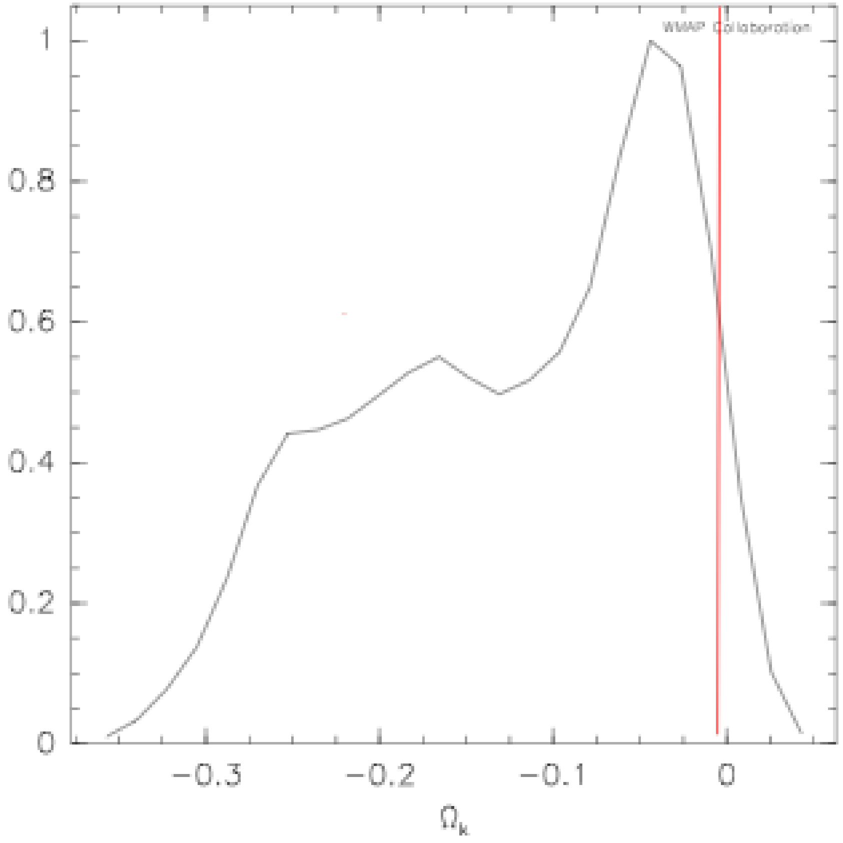
  No need to re-run!

  warning: if new data set is not consistent with the old one--> nonsense

Errors, what errors?



Hamann et al. arXiv:0705.0440


WMAP Collaboration
1
0.8
0.6
0.4
0.2
0
−0.3
−0.2
−0.1
0
Ω_k

Statistical vs systematic errors

Statistics can tell you how to deal with statistical errors

As a data set grows, the statistical errors shrink;
systematic errors do not shrink

You’ ve got a problem.

Rumsfeld can help:

There are known knowns. These are things we know that we know. There are known unknowns. That is to say, there are things that we know we don't know. But there are also unknown unknowns. There are things we don't know we don't know.

Donald Rumsfeld

Jokes aside: some interesting literature has appeared

Key concepts today

- Probability
- Bayes theorem
- Modeling of data and statistical inference
- Likelihoods and chisquare
- Confidence levels; confidence regions
- Monte Carlo methods
- Monte-Carlo errors
- MCMC
- Errors, what errors?

Exercise(s)

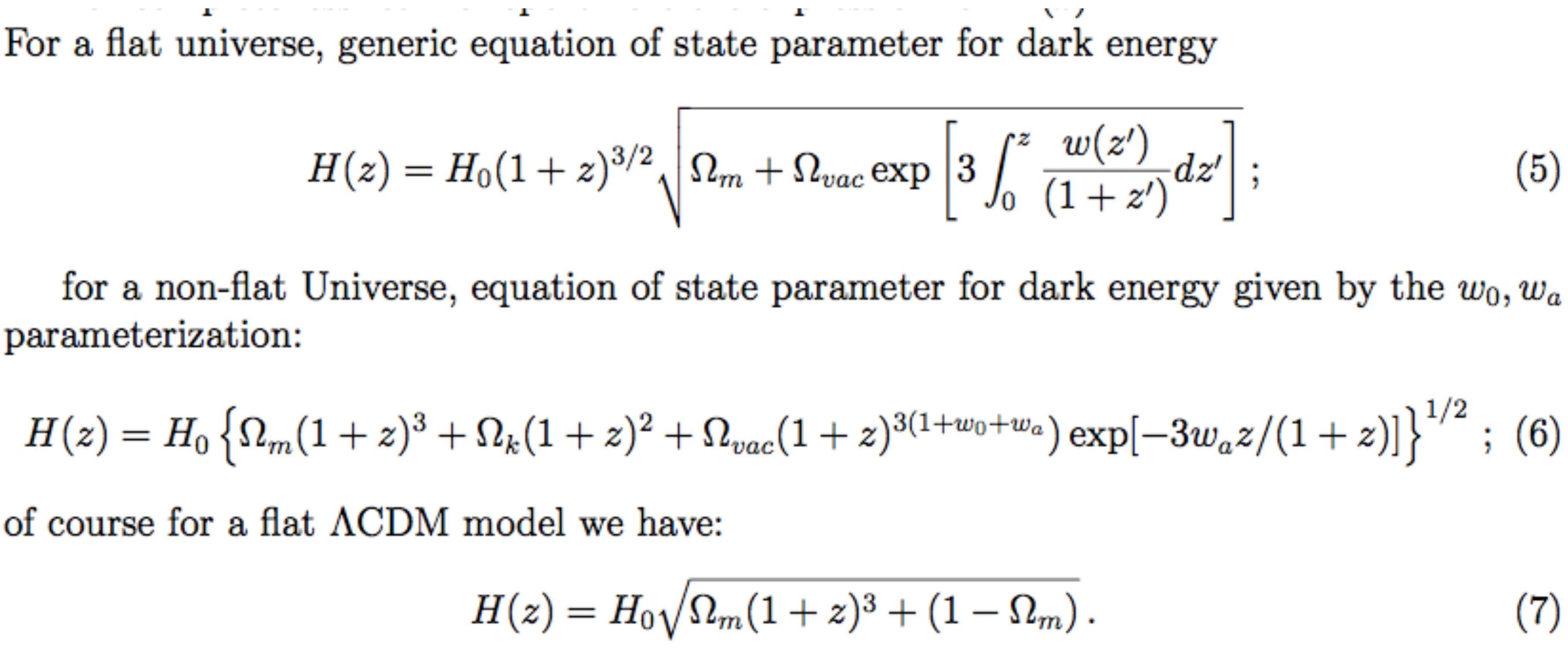
Monte Carlo integration

π

A multi-dimensional Gaussian (compare to analytics)

Write your own MCMC

Simple example: H(z)



For a flat universe, generic equation of state parameter for dark energy

$$H(z) = H_0(1+z)^{3/2}\sqrt{\Omega_m + \Omega_{vac}\exp\left[3\int_0^z \frac{w(z')}{(1+z')}dz'\right]}\;; \qquad (5)$$

for a non-flat Universe, equation of state parameter for dark energy given by the w_0, w_a parameterization:

$$H(z) = H_0\left\{\Omega_m(1+z)^3 + \Omega_k(1+z)^2 + \Omega_{vac}(1+z)^{3(1+w_0+w_a)}\exp[-3w_a z/(1+z)]\right\}^{1/2}\;; \quad (6)$$

of course for a flat ΛCDM model we have:

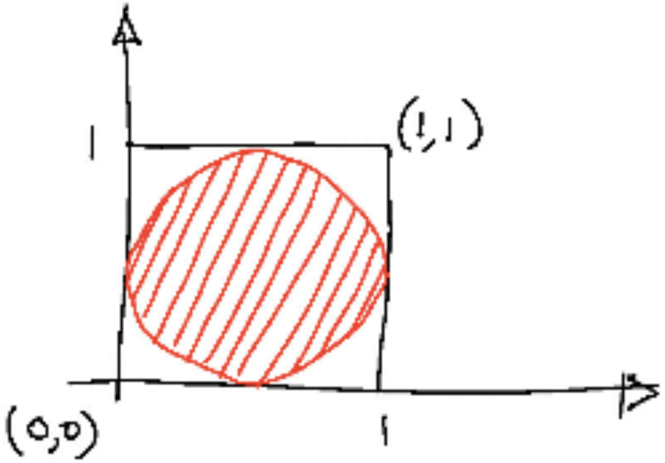
$$H(z) = H_0\sqrt{\Omega_m(1+z)^3 + (1-\Omega_m)}\,. \qquad (7)$$

H(z) data from:

http://www.physics-astronomy.unibo.it/en/research/areas/astrophysics/
cosmology-with-cosmic-chronometers

The idea

What is the probability that a dart thrown uniformly at random will hit the red area?



$$P(\text{area}) = \pi\left(1/2\right)^2 = \frac{\pi}{4}$$