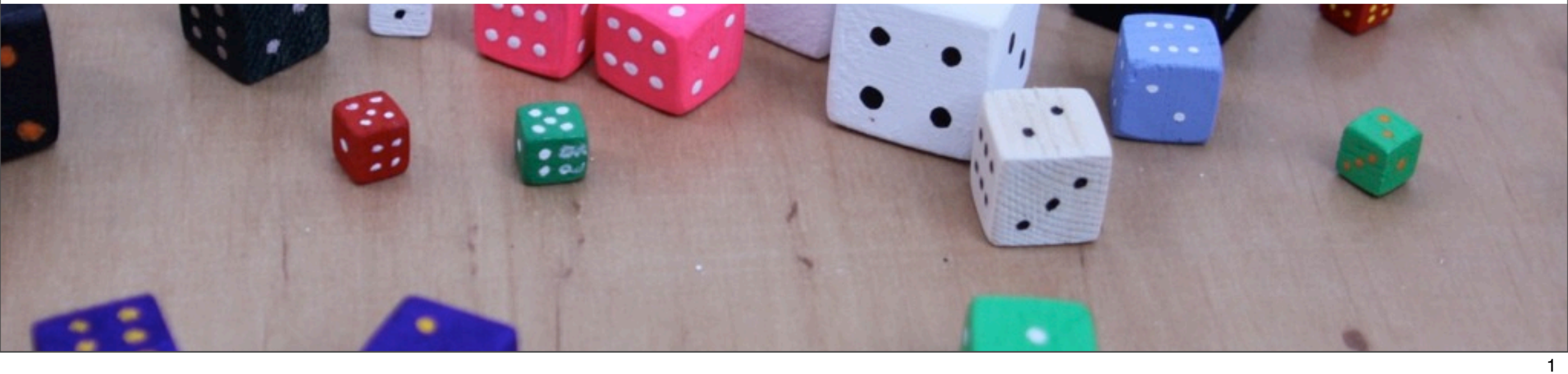
Advanced statistical methods for cosmology and astroparticle physics

Nov 16-18 2011

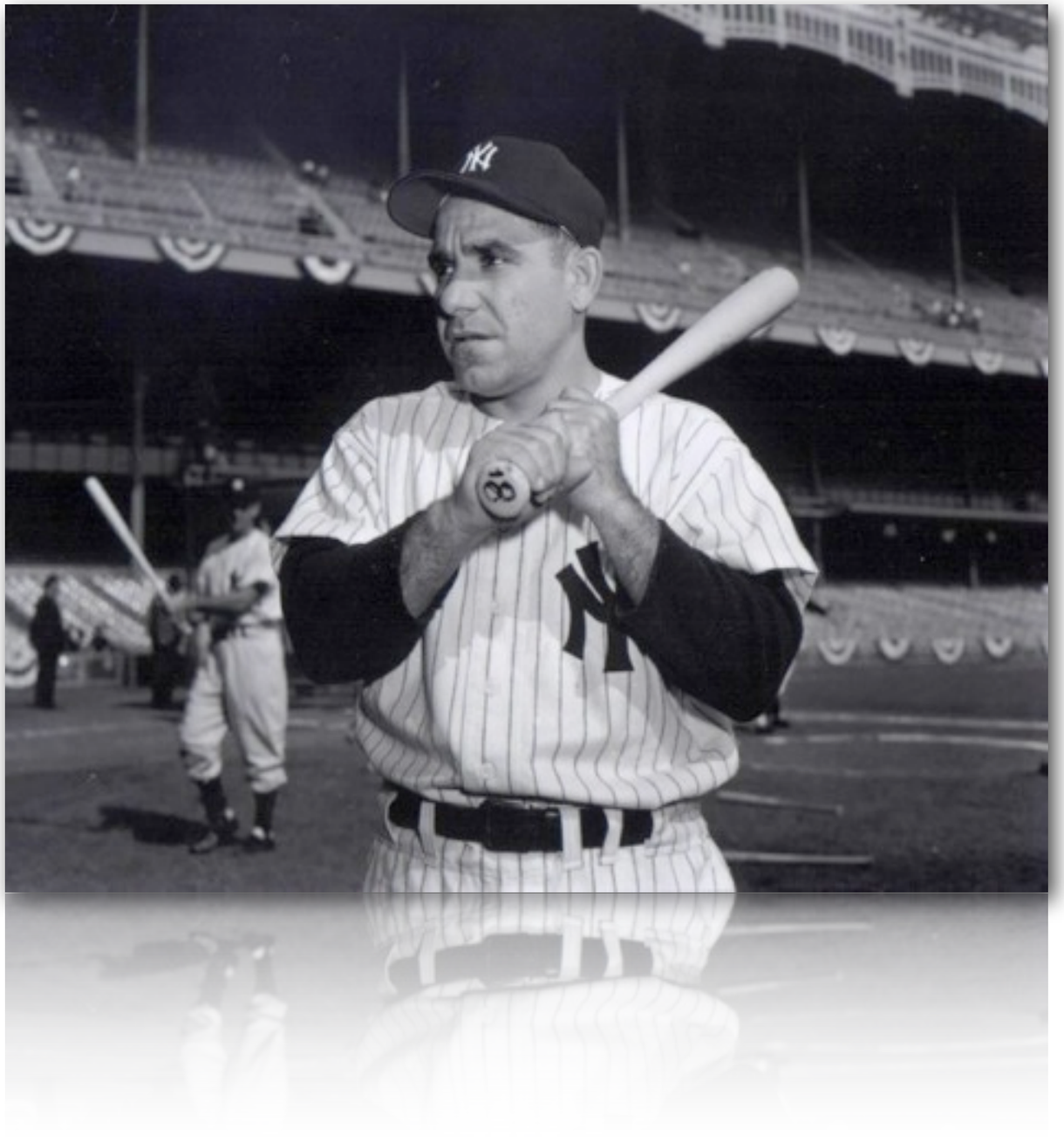
Roberto Trotta, Imperial College London



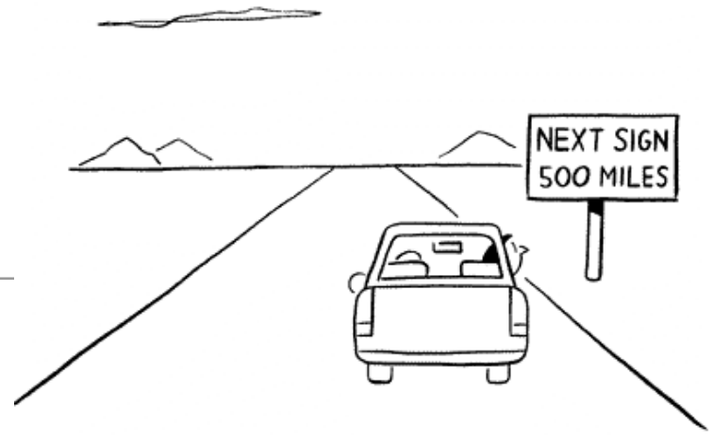
A word about statistics:

90% of the game is half mental.

Yogi Berra



Contents



- **1. Foundational aspects: what is probability?**
  Probability as frequency; Probability as degree of knowledge; Bayes Theorem; Priors; Building the likelihood function; Combination of multiple observations; Nuisance parameters

- **2. Learning from experience: Bayesian parameter inference**
  Markov Chain Monte Carlo methods; Importance sampling; Nested sampling; Reporting inferences; Credible regions vs confidence regions; The meaning of sigma

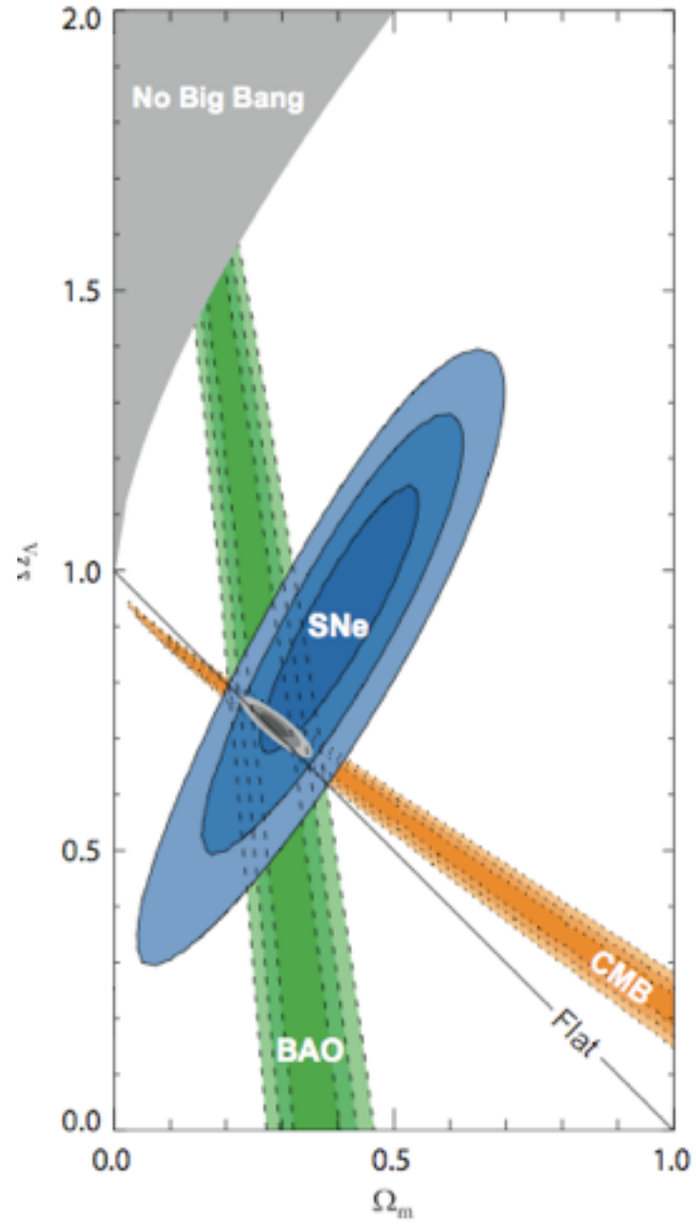
- **3. Bayesian model selection and cosmological applications**
  The different levels of inference; The Bayesian evidence and the Bayes factor; Computing Bayes factors; Information criteria for approximate model selection; The meaning of significance; Comparison with classical hypothesis testing; Model complexity; Bayesian model averaging

- **4. Experiment optimization and prediction**
  Fisher matrix formalism; Figures of merit; Expected usefulness of an experiment ; Survey optimization; Experimental utility; Bayesian adaptive exploration.

What you will learn

Imperial College London

- What does it mean to say that Ω_m = 0.28 ± 0.02 ?
- How do you get plots like this and what do they mean?
- How can you quantitatively compare different models for your observations?



Astier et al (2006)

Recommended reading

Imperial College London

- R. Trotta, “Bayes in the sky: Bayesian inference and model selection in cosmology” *Contemporary Physics*, 49, 2 (2008), 71-104 (arXiv: 0803.4089)
- *Bayesian methods in cosmology*, Hobson et al (eds), CUP (2010)
- Tom Loredo’s Bayesian papers: http://www.astro.cornell.edu/staff/loredo/bayes/tjl.html

- G. D’Agostini, *Probability and Measurement Uncertainty in Physics - a Bayesian Primer* (1995), hep-ph/9512295
- E.T. Jaynes, *Probability Theory: The Logic of Science*, CUP (2003)
- D. MacKay, *Information theory, Inference & Learning Algorithms*, CUP (2003) (available for free on the web for onscreen viewing)
- P. Gregory, *Bayesian logical data analysis for the physical sciences*, CUP (2003)

“Why should i bother?”

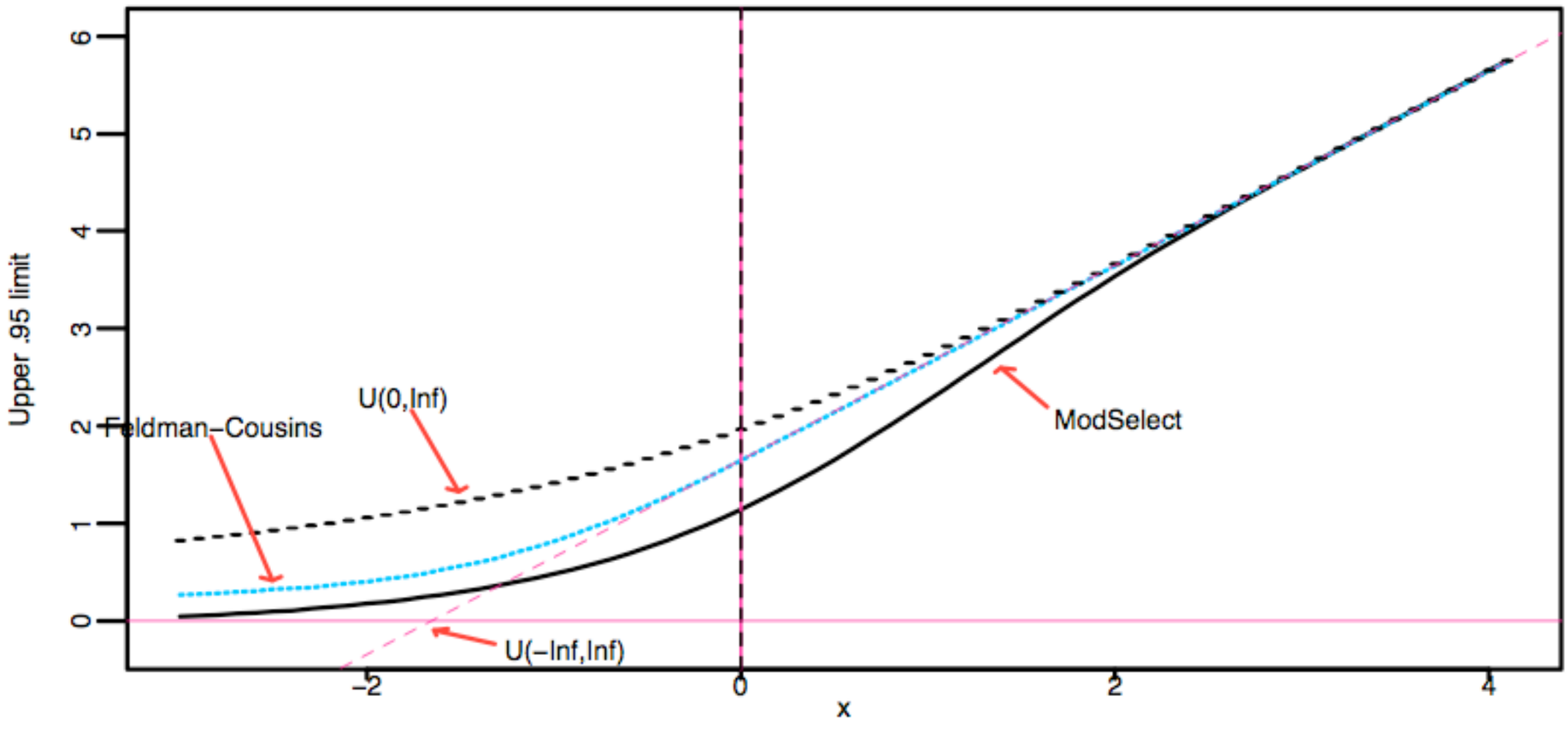
Imperial College London

- Increasingly complex models and data: "chi-square by eye" simply not enough
- "If it's real, better data will show it":
  but all the action is in the "discovery zone" around 3-4 sigma significance. This is a moving target.
- Don't waste time explaining effects which are not there
- Plan for the future: which is the best strategy? (survey design & optimization)
- In some cases, there will be no better data! (cosmic variance)

The neutrino example

Imperial College London

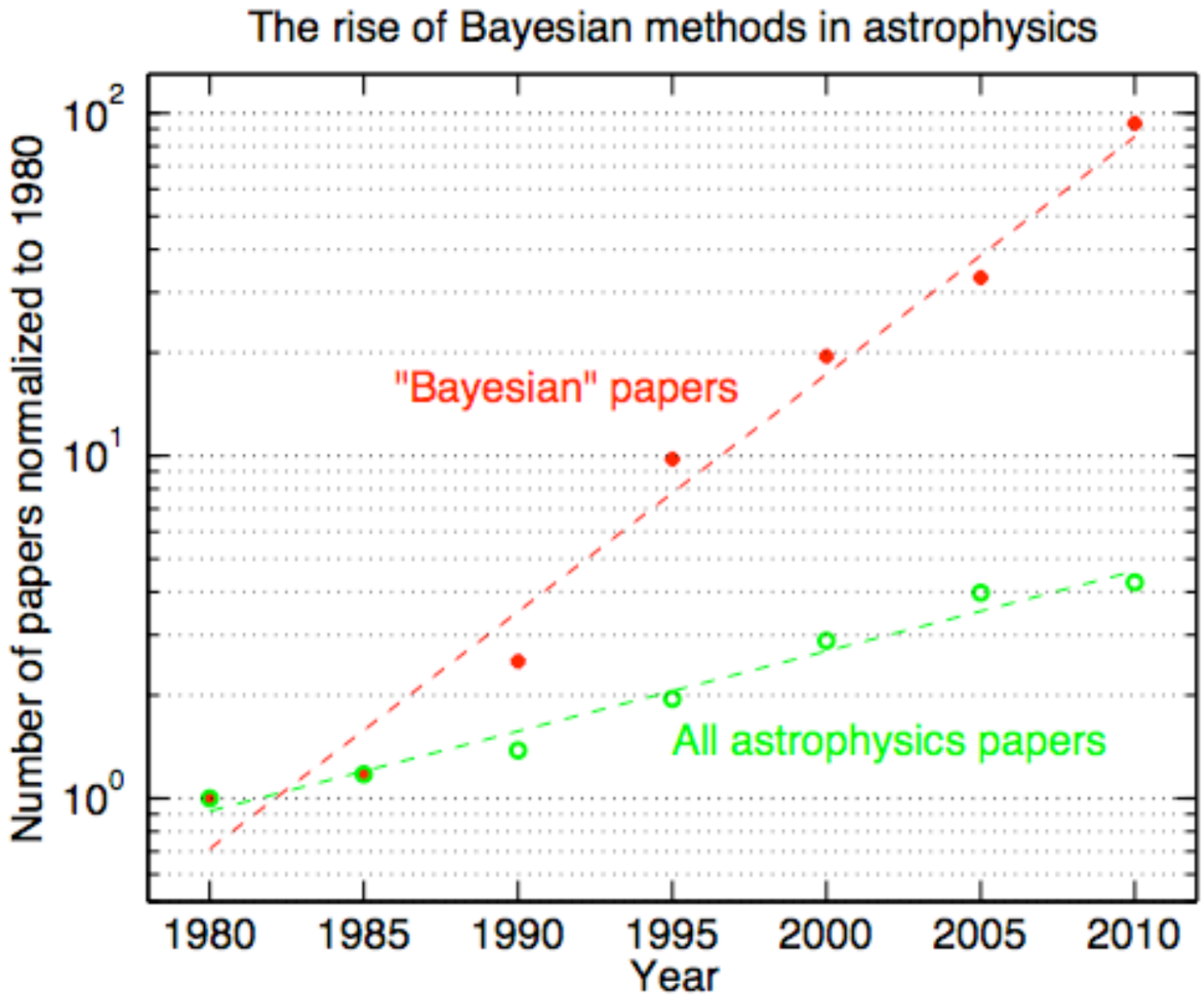
Upper 95% limit on neutrino mass as a function of observed value for different statistical methods



Brad Efron (PHYSTAT 2003)

Bayesian methods on the rise

Imperial College London



Bayes' theorem

Imperial College London

posterior

likelihood

prior

$$P(\theta|d,I) = \frac{P(d|\theta,I)P(\theta|I)}{P(d|I)}$$

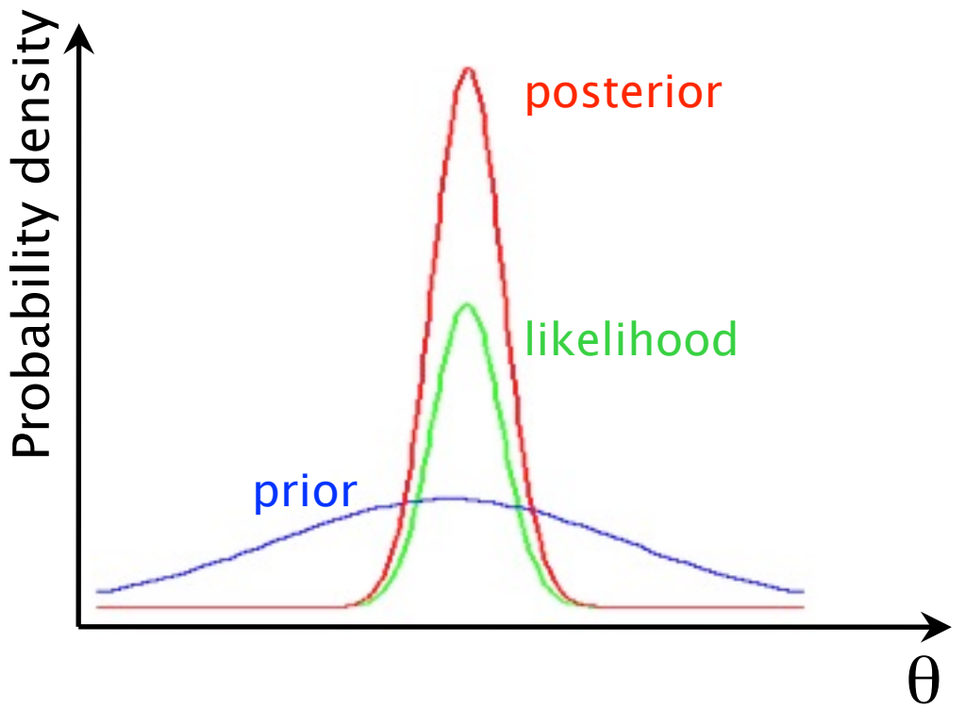
evidence

θ: parameters
d: data
I: any other external information, or the assumed model

For parameter inference it is sufficient to consider

$$P(\theta|d,I) \propto P(d|\theta,I)P(\theta|I)$$

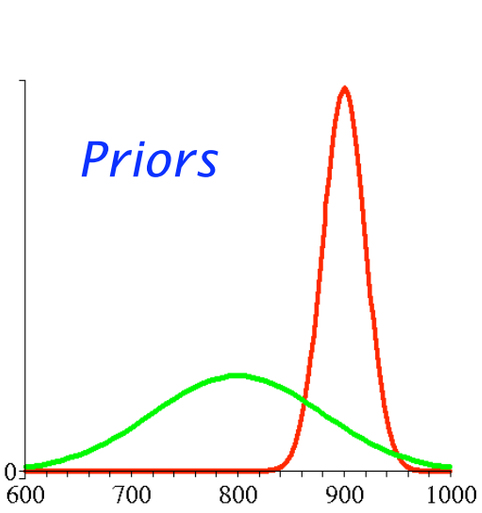
$$\text{posterior} \propto \text{likelihood} \times \text{prior}$$

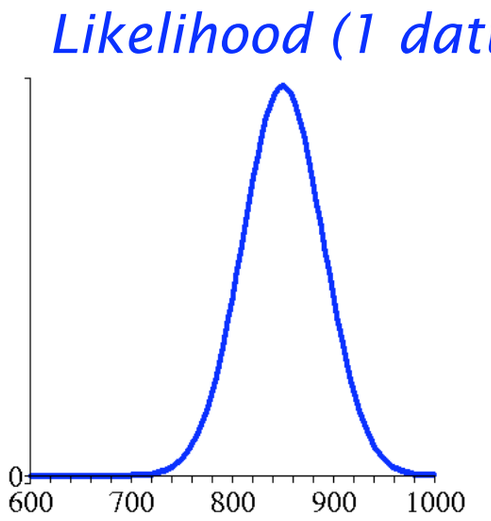


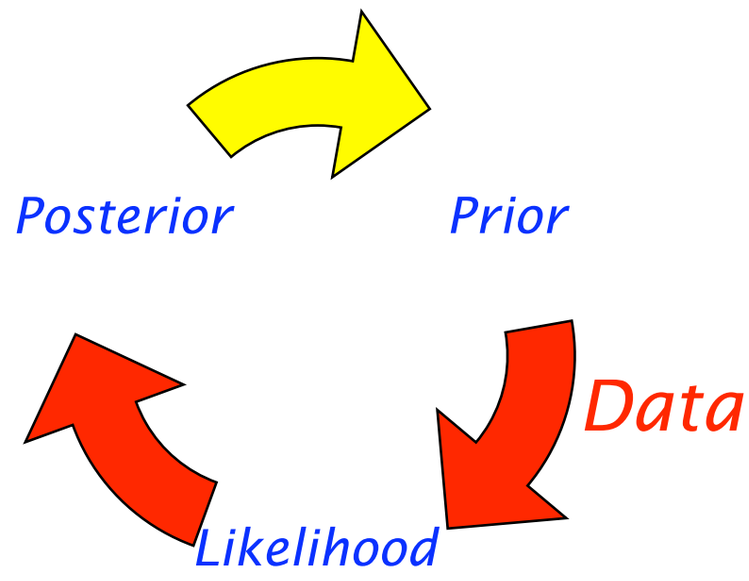
The matter with priors

Imperial College London

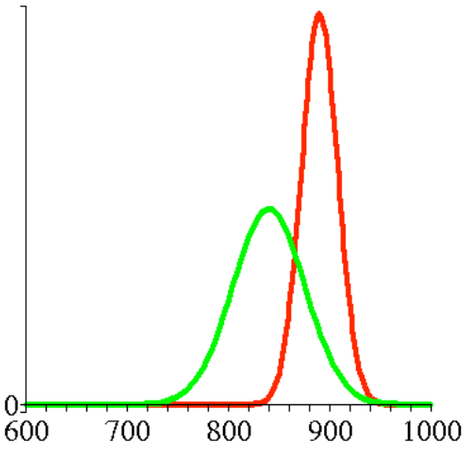
- In parameter inference, prior dependence will **in principle** vanish for strongly constraining data.
  A sensitivity analysis is mandatory for all Bayesian methods!



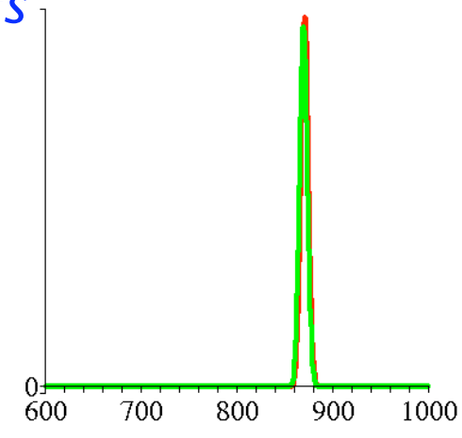




Posterior after 1 datum



Posterior after 100 data points



Inference in many dimensions

Imperial College London

Usually our parameter space is multi-dimensional: how should we report inferences for one parameter at the time?

BAYESIAN

Marginal posterior:

$$P(\theta_1|D) = \int L(\theta_1, \theta_2) p(\theta_1, \theta_2) d\theta_2$$

FREQUENTIST

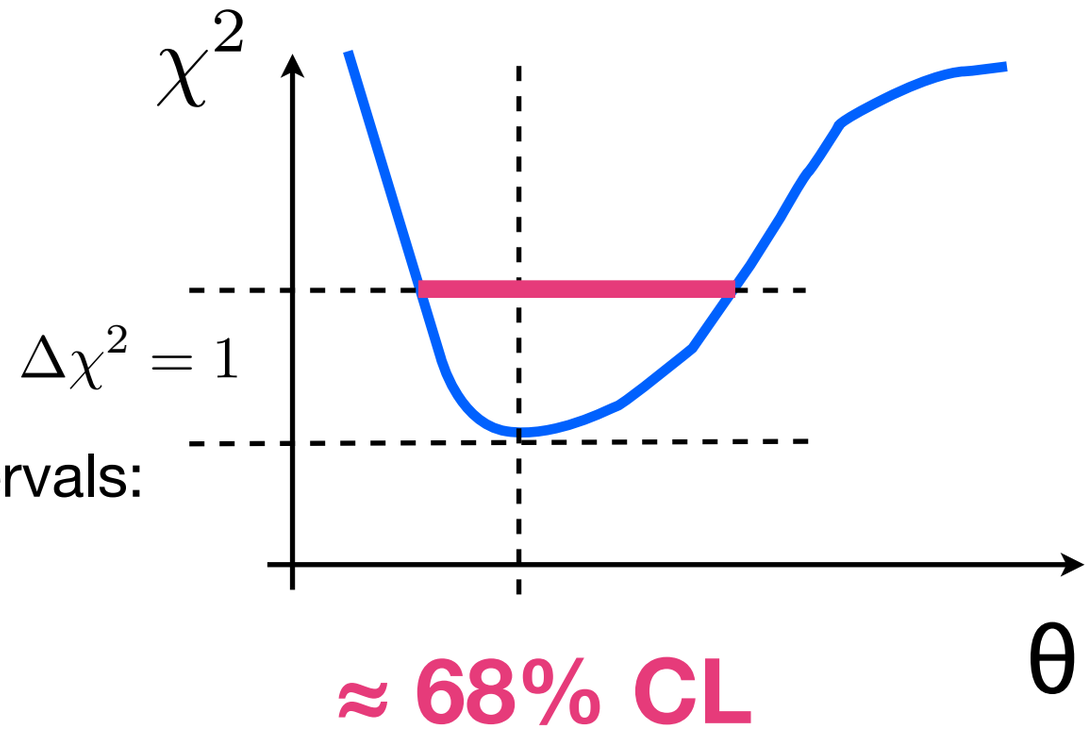
Profile likelihood:

$$L(\theta_1) = max_{\theta_2} L(\theta_1, \theta_2)$$

Confidence intervals: Frequentist approach

Imperial College London

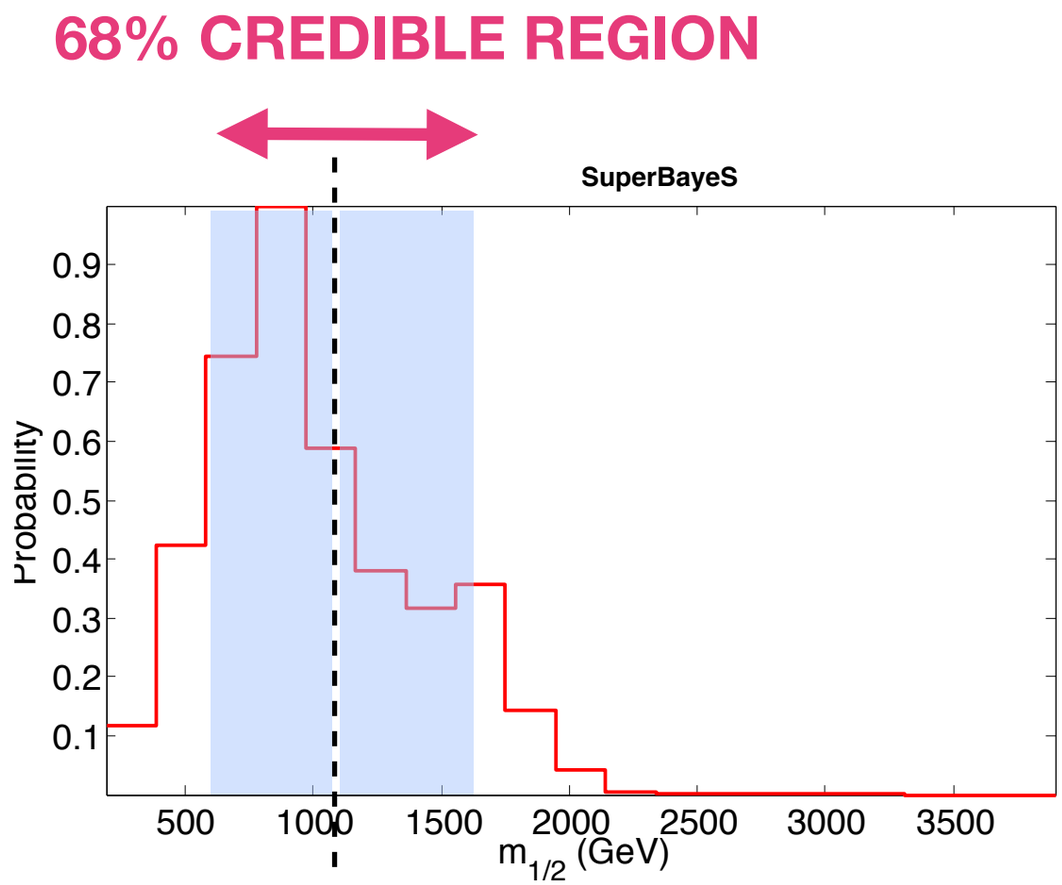
- **Likelihood-based methods:** determine the best fit parameters by finding the minimum of -2Log(Likelihood) = chi-squared
  - Analytical for Gaussian likelihoods
  - Generally numerical
  - Steepest descent, MCMC, ...
- Determine approximate confidence intervals: Local Δ(chi-squared) method



Credible regions: Bayesian approach

Imperial College London

- Use the prior to define a metric on parameter space.
- **Bayesian methods:** the best-fit has no special status. Focus on region of large posterior probability mass instead.
  - Markov Chain Monte Carlo (MCMC)
  - Nested sampling
  - Hamiltonian MC
- Determine posterior credible regions: e.g. symmetric interval around the mean containing 68% of samples



The good news

Imperial College London

- Marginalisation and profiling give exactly identical results for the linear Gaussian case.

- This is not suprising, as we already saw that the answer for the Gaussian case is numerically identical for both approaches

- And now the bad news: THIS IS NOT GENERICALLY TRUE!

- A good example is the **Neyman-Scott problem:**

  - We want to measure the signal amplitude μ_i of N sources with an uncalibrated instrument, whose Gaussian noise level σ is constant but unknown.

  - Ideally, measure the amplitude of calibration sources or measure one source many times, and infer the value of σ

Neyman-Scott problem

Imperial College London

- In the Neyman-Scott problem, no calibration source is available and we can only get 2 measurements per source. So for N sources, we have N+1 parameters and 2N data points.

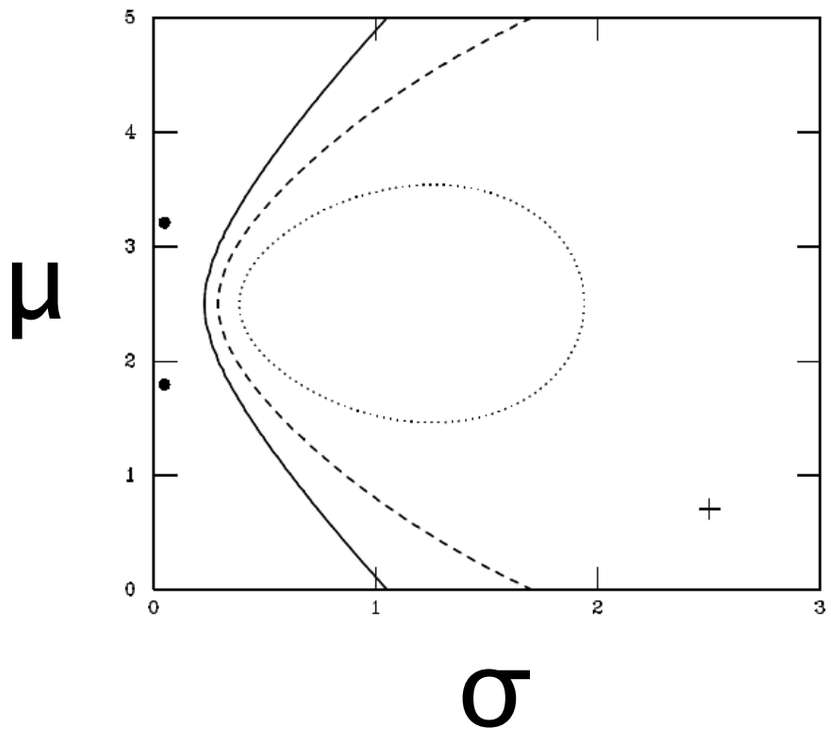
- The profile likelihood estime of σ converges to a biased value σ/sqrt(2) for N → ∞

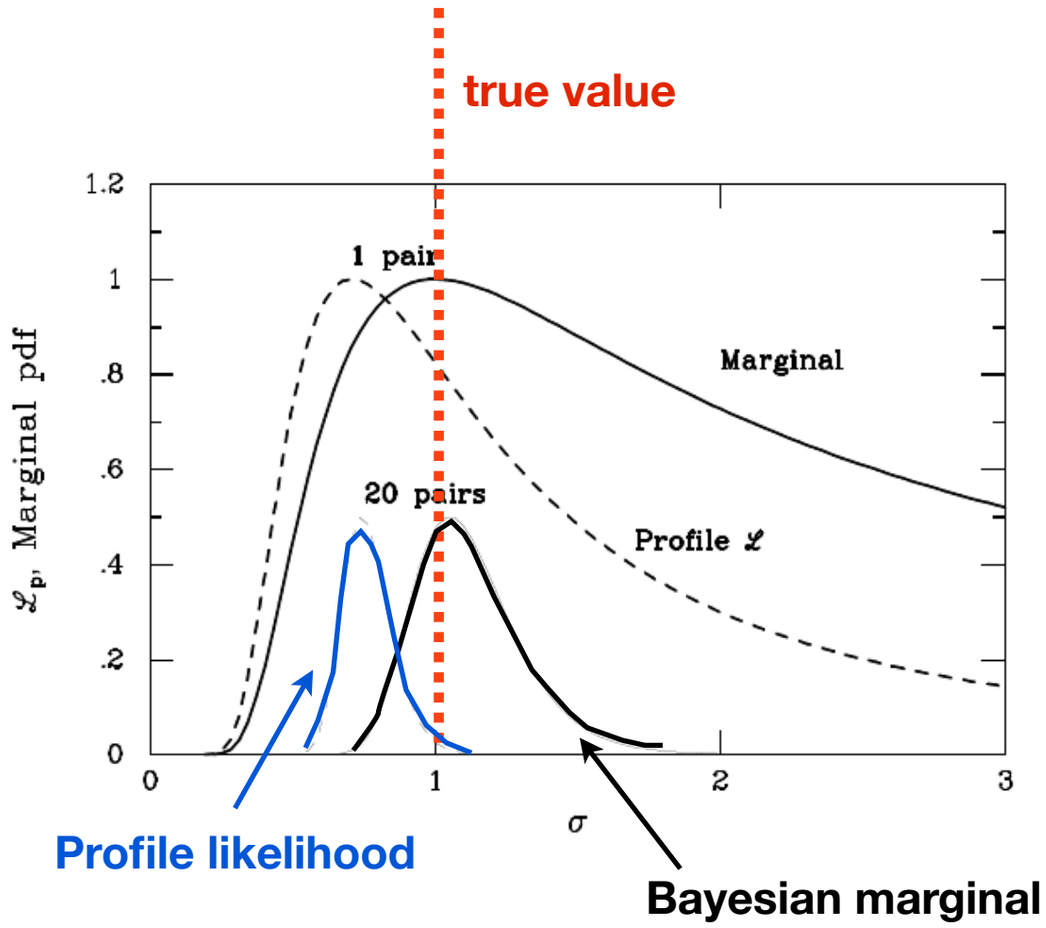
- The Bayesian answer has larger variance but is unbiased

Neyman-Scott problem

Imperial College London

Tom Loredo, talk at Banff 2010 workshop:





Marginalization vs profiling (maximising)

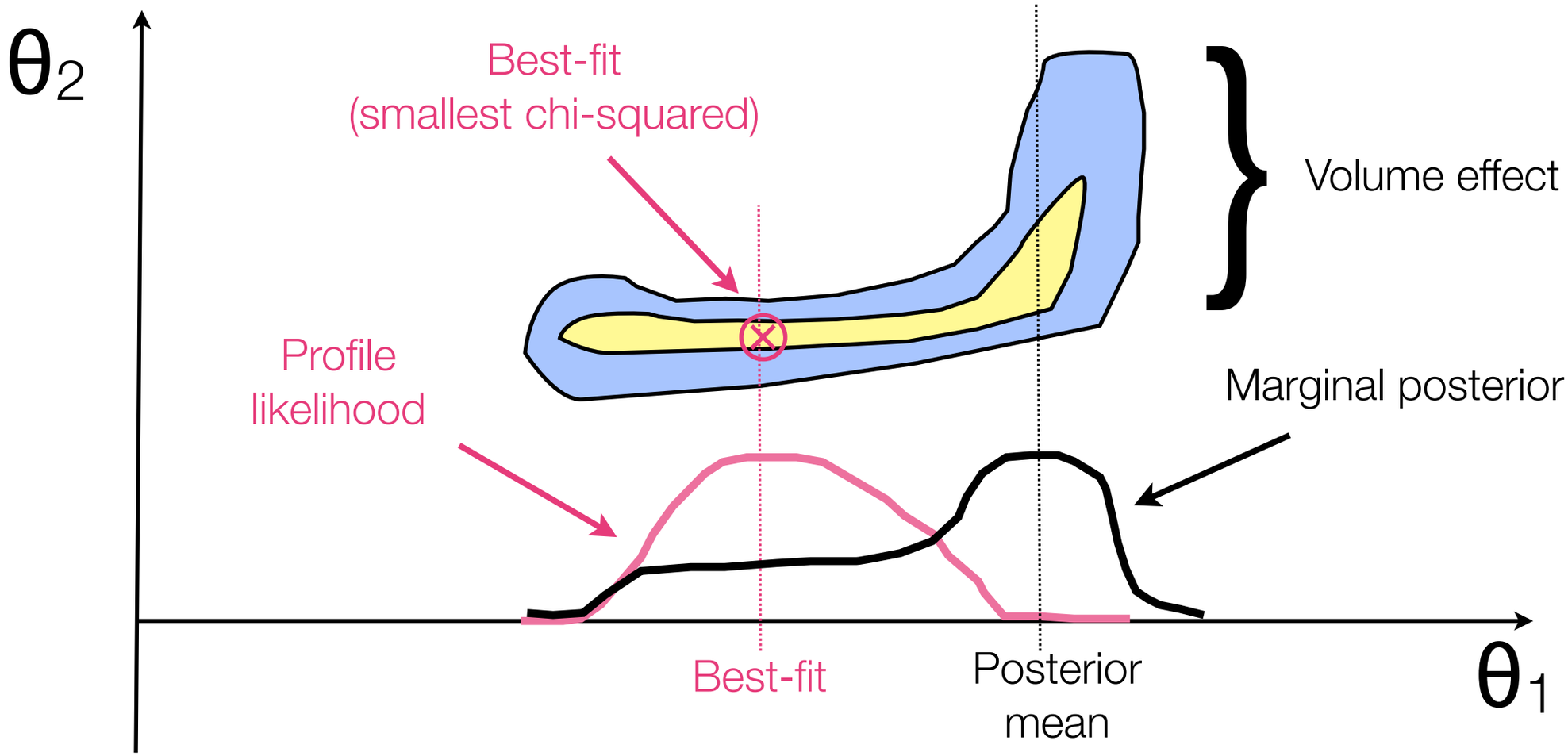
Imperial College London

Marginal posterior:

$$P(\theta_1|D) = \int L(\theta_1,\theta_2)p(\theta_1,\theta_2)d\theta_2$$

Profile likelihood:

$$L(\theta_1) = max_{\theta_2} L(\theta_1,\theta_2)$$



(2D plot depicts likelihood contours - prior assumed flat over wide range)

Marginalization vs profiling (maximising)

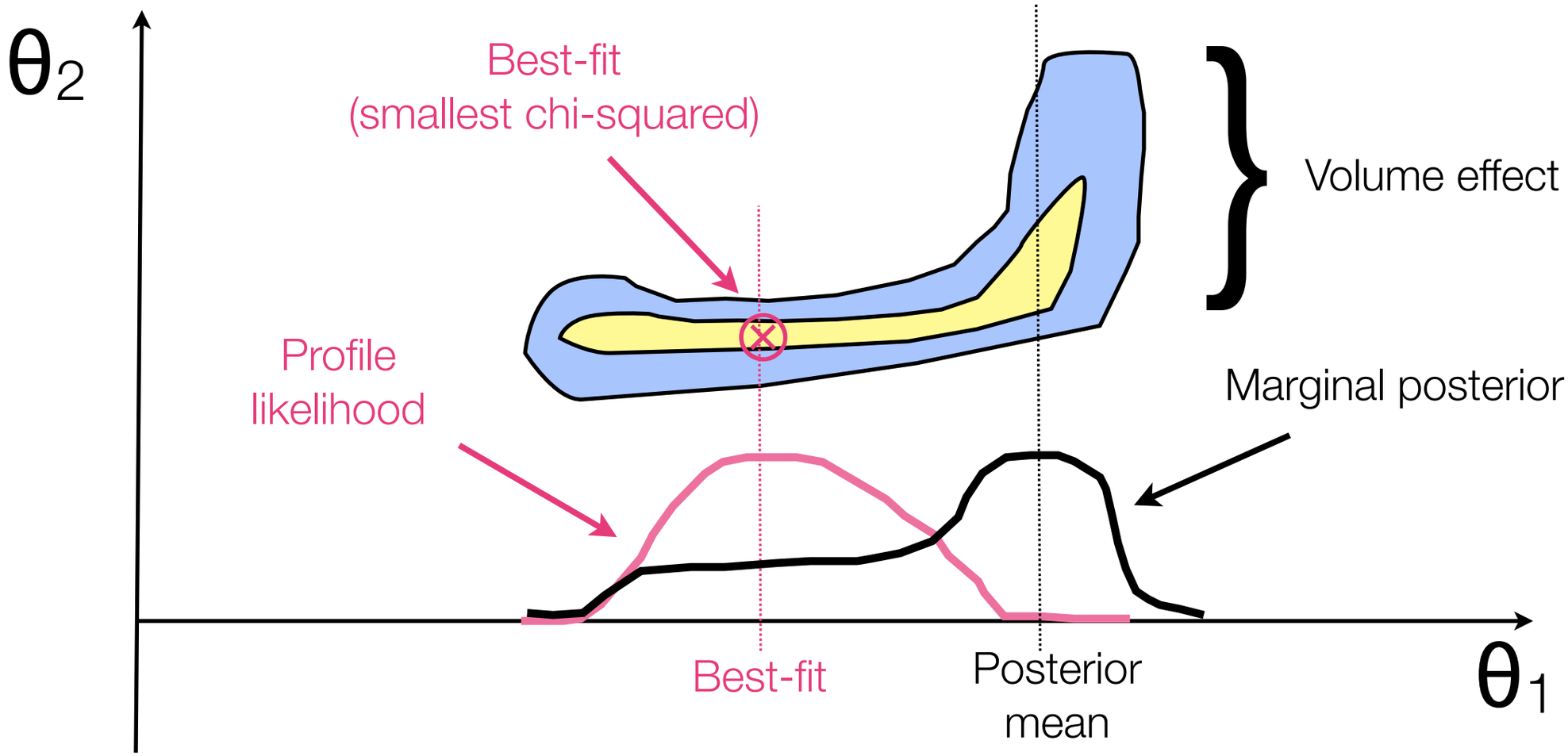
Imperial College London

Physical analogy: (thanks to Tom Loredo)

Likelihood = hottest hypothesis
Posterior = hypothesis with most heat

Heat: $$Q = \int c_V(x) T(x) dV$$

Posterior: $$P \propto \int p(\theta) L(\theta) d\theta$$



(2D plot depicts likelihood contours - prior assumed flat over wide range)

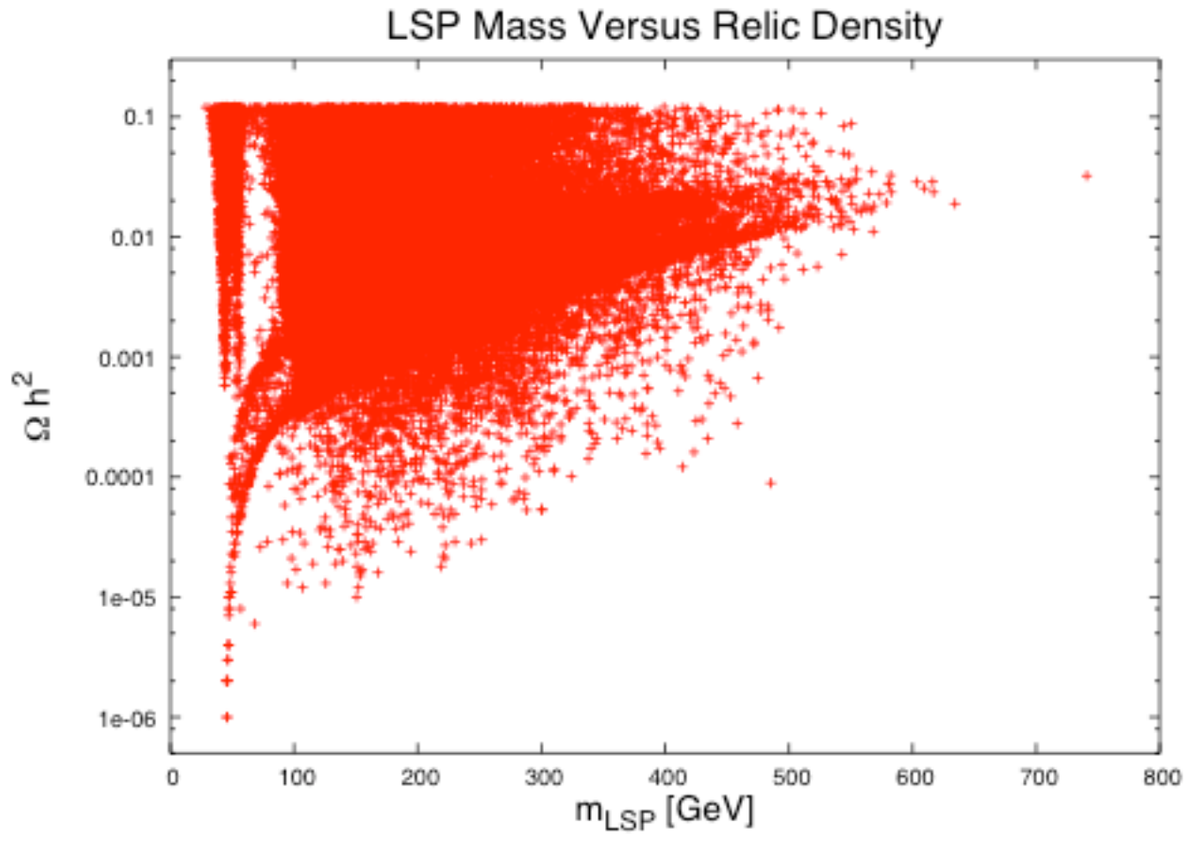
Markov Chain Monte Carlo

Exploration with “random scans”

Imperial College London

- Points accepted/rejected in a in/out fashion (e.g., 2-sigma cuts)
- No statistical measure attached to density of points: no probabilistic interpretation of results possible, although the temptation cannot be resisted...
- Inefficient in high dimensional parameters spaces (D>5)
- **HIDDEN PROBLEM:** Random scan explore only a very limited portion of the parameter space!

One recent example:
Berger et al (0812.0980)
pMSSM scans
(20 dimensions)



C. F. B
Hewett
Supers
Prejudi
[arXiv

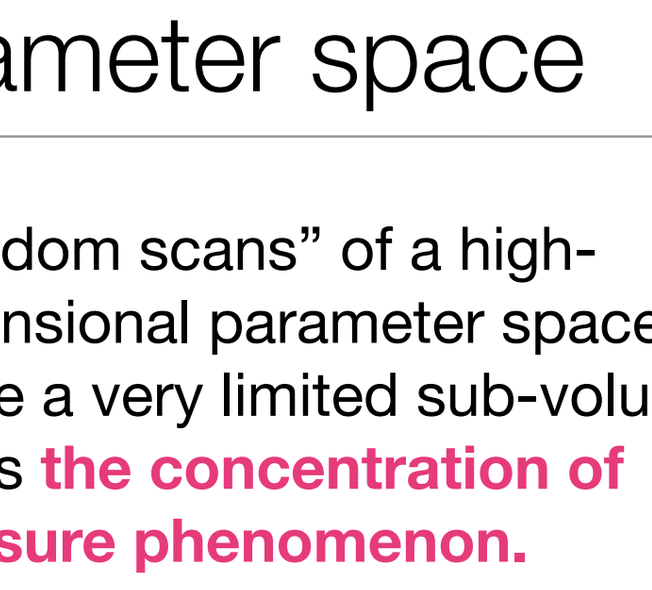
check t
pMSSM

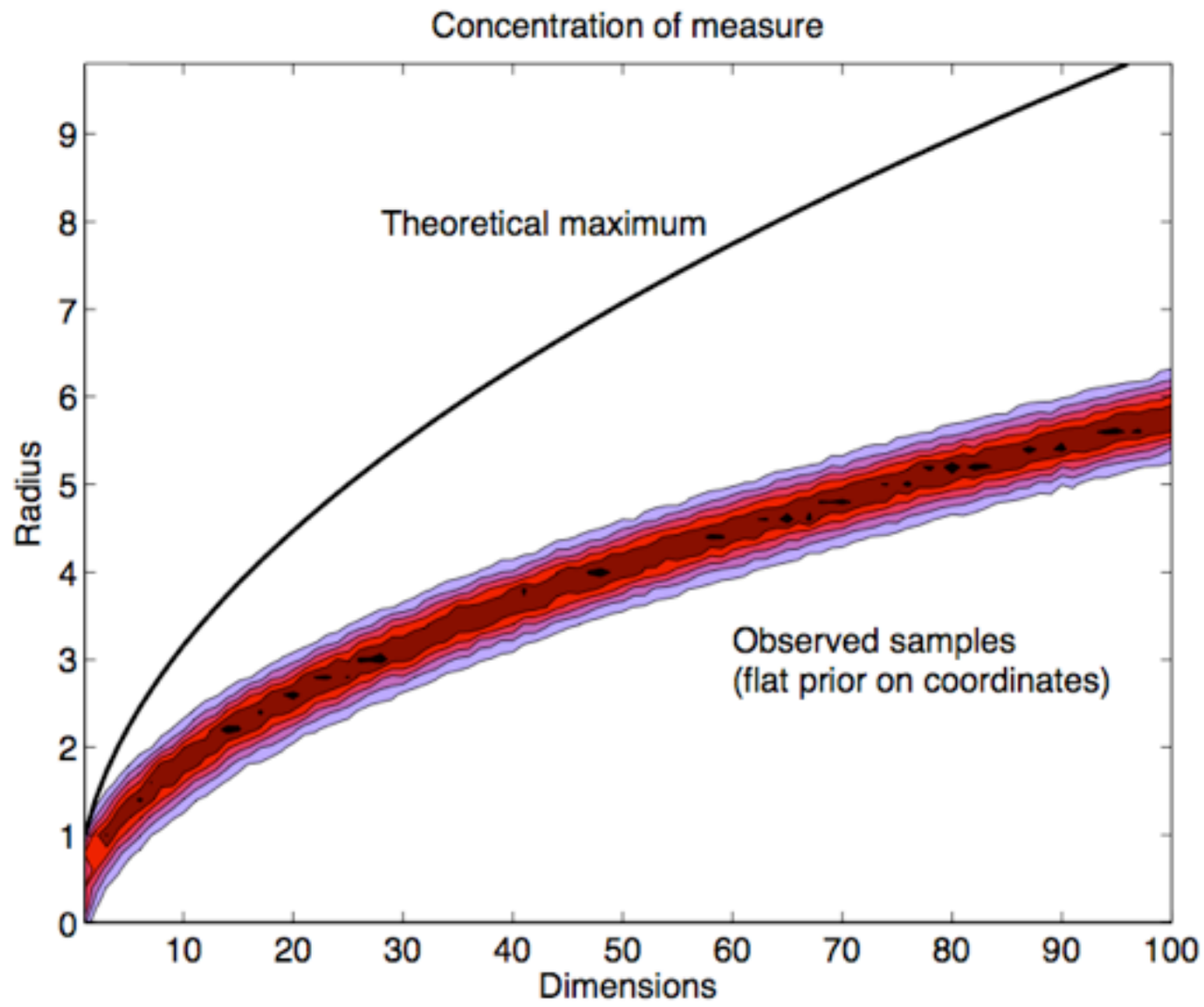
Random scans explore only a small fraction of the parameter space

Imperial College London

- "Random scans" of a high-dimensional parameter space only probe a very limited sub-volume: this is **the concentration of measure phenomenon.**

- **Statistical fact:** the norm of *D* draws from U[0,1] concentrates around $(D/3)^{1/2}$ with constant variance



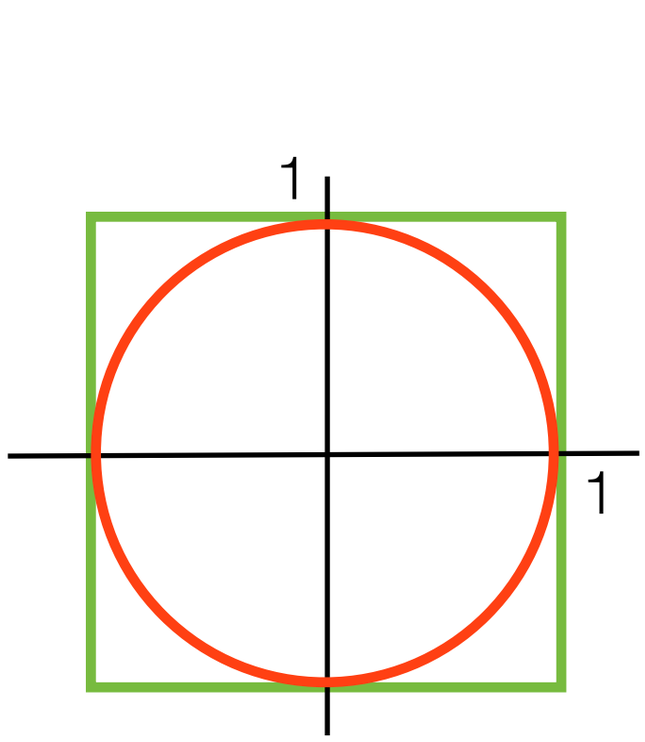


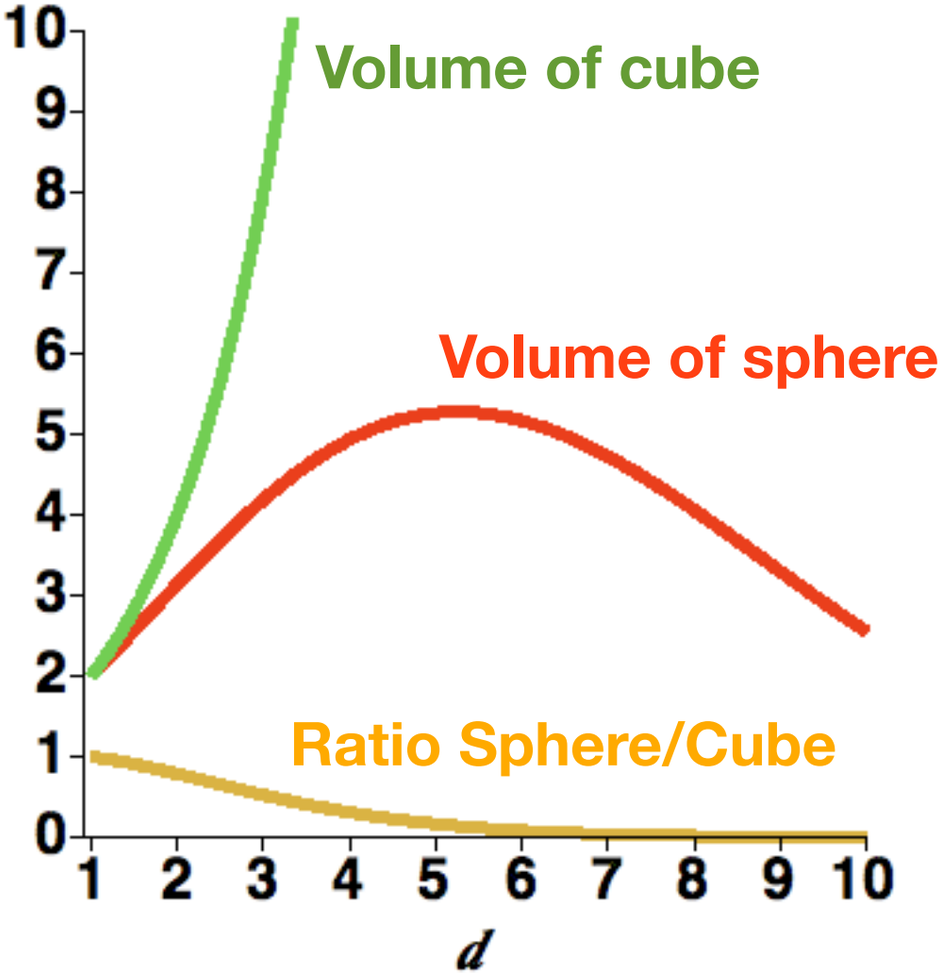
Geometry in high-D spaces

Imperial College London

- **Geometrical fact:** in *D* dimensions, most of the volume is near the boundary. The volume inside the spherical core of *D*-dimensional cube is negligible.

Together, these two facts mean that random scan only explore a very small fraction of the available parameter space in high-dimesional models.





Key advantages of the Bayesian approach

Imperial College London

- **Efficiency:** computational effort scales ~ N rather than k^N as in grid-scanning methods. Orders of magnitude improvement over grid-scanning.
- **Marginalisation:** integration over hidden dimensions comes for free.
- **Inclusion of nuisance parameters:** simply include them in the scan and marginalise over them.
- **Pdf's for derived quantities**: probabilities distributions can be derived for any function of the input variables

The general solution

Imperial College London

$$P(\theta|d, I) \propto P(d|\theta, I)P(\theta|I)$$

- Once the RHS is defined, how do we evaluate the LHS?
- Analytical solutions exist only for the simplest cases (e.g. Gaussian linear model)
- Cheap computing power means that numerical solutions are often just a few clicks away!
- Workhorse of Bayesian inference: Markov Chain Monte Carlo (MCMC) methods. A procedure to generate a list of samples from the posterior.

MCMC estimation

Imperial College London

$$P(\theta|d,I) \propto P(d|\theta,I)P(\theta|I)$$

- A Markov Chain is a list of samples θ_1, θ_2, θ_3,... whose density reflects the (unnormalized) value of the posterior

- A MC is a sequence of random variables whose *(n+1)*-th elements only depends on the value of the *n*-th element

- **Crucial property:** a Markov Chain converges to a stationary distribution, i.e. one that does not change with time. In our case, the posterior.

- From the chain, expectation values wrt the posterior are obtained very simply:

$$\langle\theta\rangle = \int d\theta P(\theta|d)\theta \approx \frac{1}{N}\sum_i \theta_i$$

$$\langle f(\theta)\rangle = \int d\theta P(\theta|d) f(\theta) \approx \frac{1}{N}\sum_i f(\theta_i)$$

Reporting inferences

Imperial College London

- **Once P(θ|d, I) found, we can report inference by:**
  - Summary statistics (best fit point, average, mode)
  - Credible regions (e.g. shortest interval containing 68% of the posterior probability for θ). Warning: this has **not** the same meaning as a frequentist confidence interval! (Although the 2 might be formally identical)
  - Plots of the marginalised distribution, integrating out nuisance parameters (i.e. parameters we are not interested in). This generalizes the propagation of errors:

$$P(\theta|d, I) = \int d\phi P(\theta, \phi|d, I)$$

What does x=1.00±0.01 mean?

Imperial College London

$$P(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right)$$

$$\text{Notation}: \quad x \sim N(\mu, \sigma^2)$$

- **Frequentist statistics (Fisher, Neymann, Pearson):**
  E.g., estimation of the mean μ of a Gaussian distribution from a list of observed samples x_1, x_2, x_3**...**
  The sample mean is the Maximum Likelihood estimator for μ:

  $\mu_{ML} = X_{av} = (x_1 + x_2 + x_3 + \ldots x_N)/N$

- **Key point:**
  in $P(X_{av})$, X_{av} is a random variable, i.e. one that takes on different values across an ensemble of infinite (imaginary) identical experiments. X_{av} is distributed according to $X_{av} \sim N(\mu, \sigma^2/N)$ **for a fixed true μ**
  The distribution applies to imaginary replications of data.

What does x=1.00±0.01 mean?

Imperial College London

- **Frequentist statistics (Fisher, Neymann, Pearson):**
  The final result for the confidence interval for the mean

$$P(\mu_{ML} - \sigma/N^{1/2} < \mu < \mu_{ML} + \sigma/N^{1/2}) = 0.683$$

- This means:
  If we were to repeat this measurements many times, and obtain a 1-sigma distribution for the mean, the true value μ would lie inside the so-obtained intervals 68.3% of the time

- This is not the same as saying: “The probability of μ to lie within a given interval is 68.3%”. This statement only follows from using Bayes theorem.

What does x=1.00±0.01 mean?

Imperial College London

- **Bayesian statistics (Laplace, Gauss, Bayes, Bernouilli, Jaynes):**

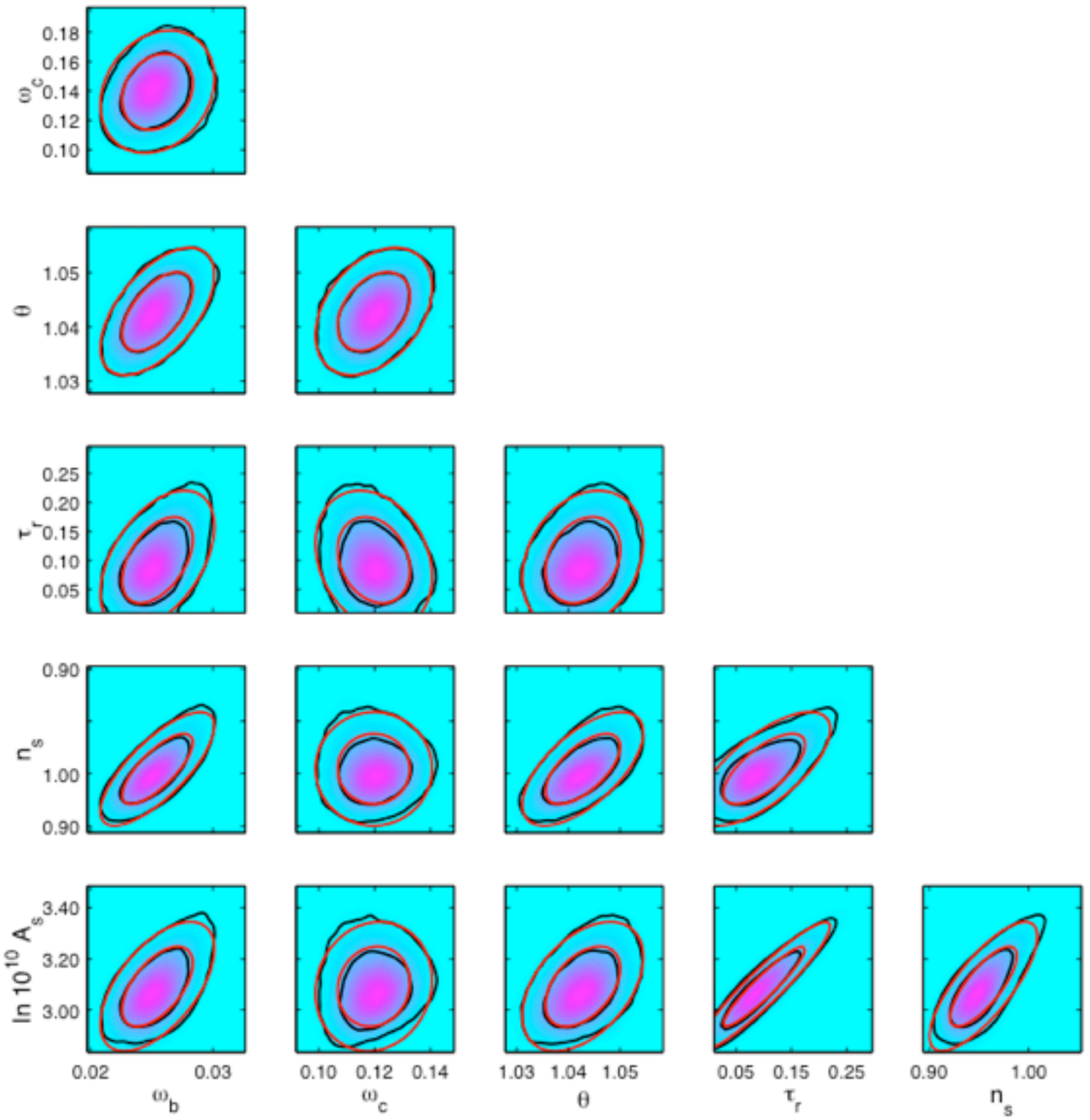
  After applying Bayes therorem $P(\mu | X_{av})$ describes the distribution of our degree of belief about the value of μ given the information at hand, i.e. the observed data.

- Inference is conditional only on the observed values of the data.

- There is no concept of repetition of the experiment.

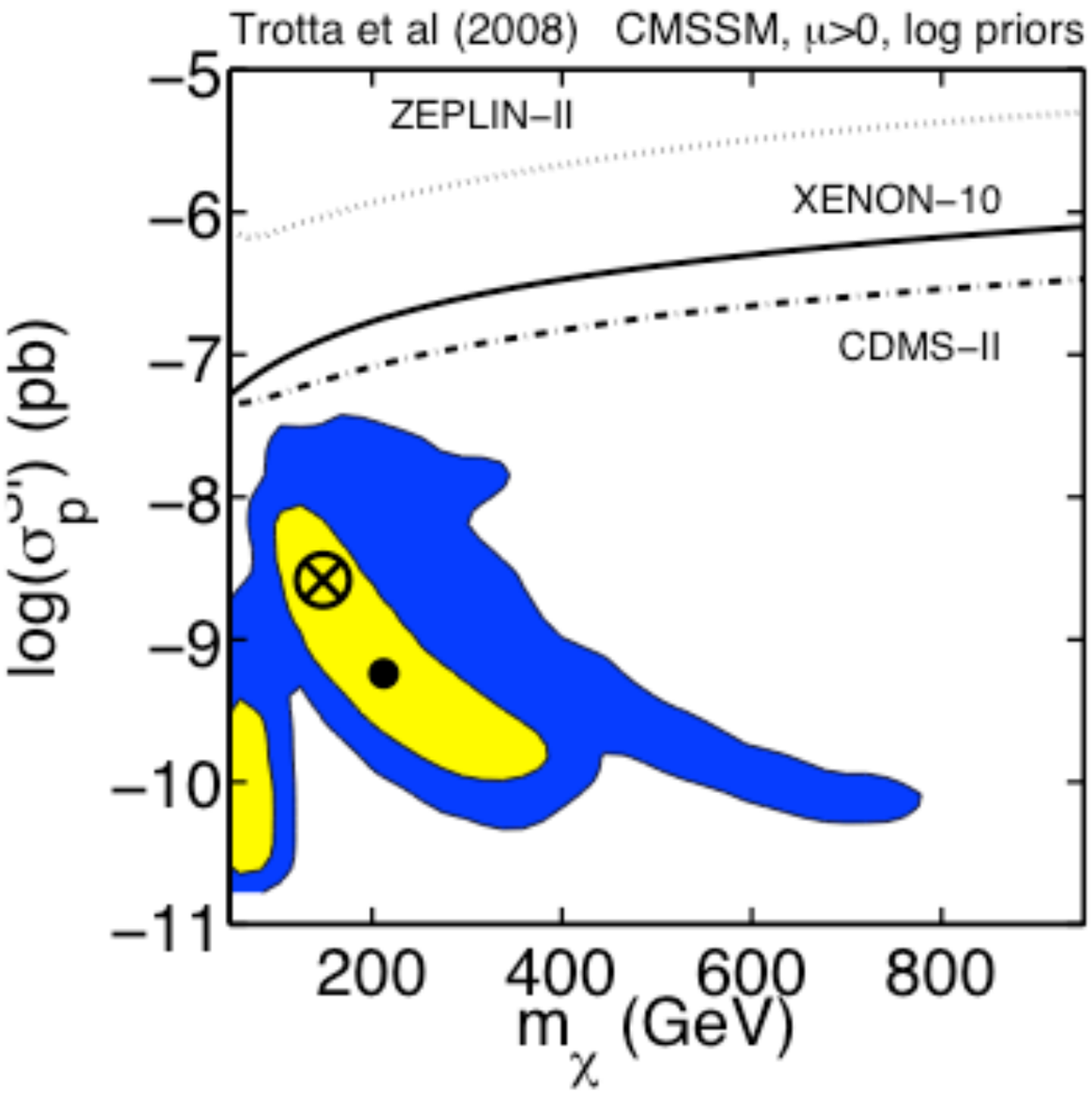
Gaussian case

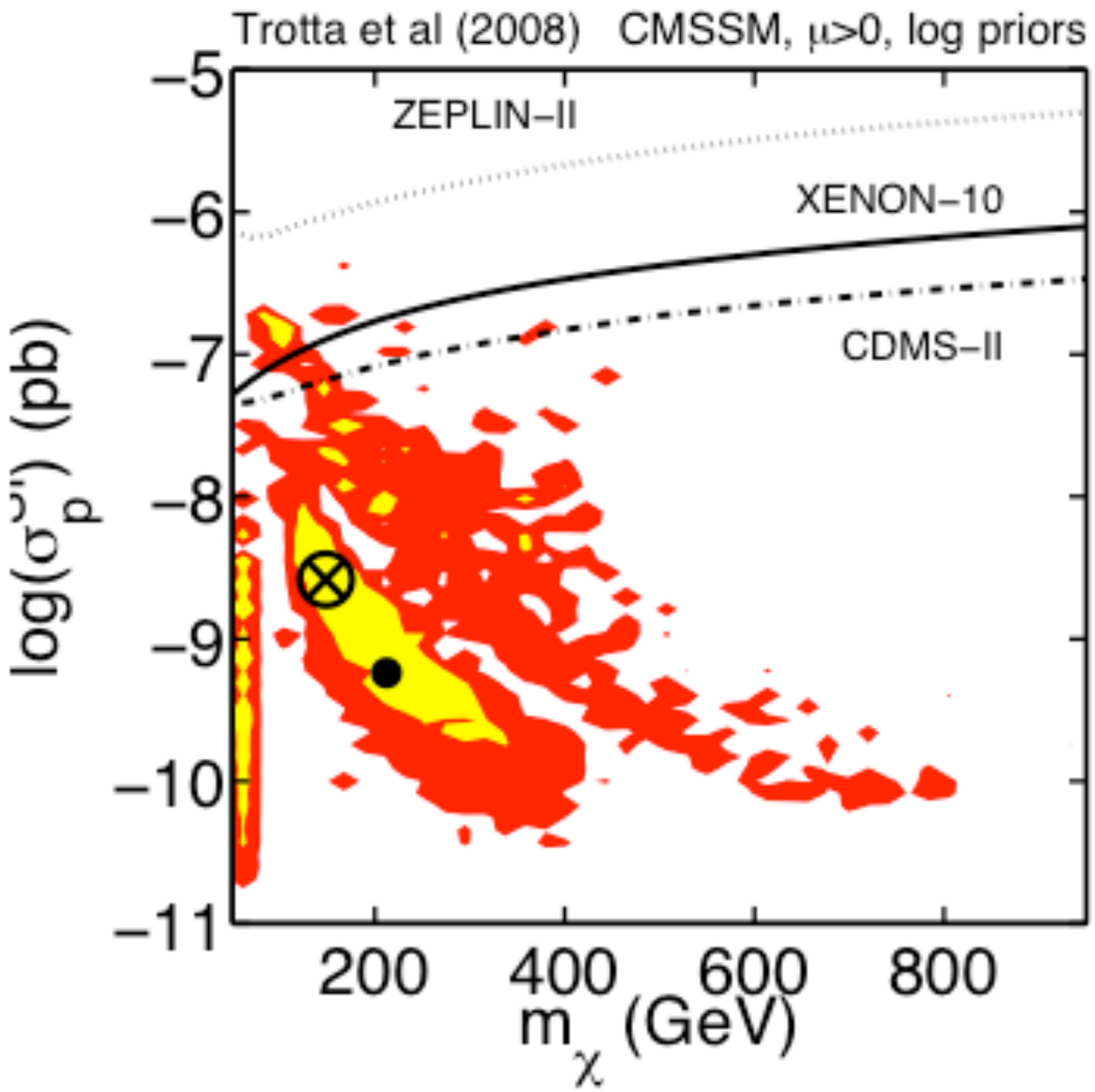
Imperial College London



Non-Gaussian posteriors

Imperial College London





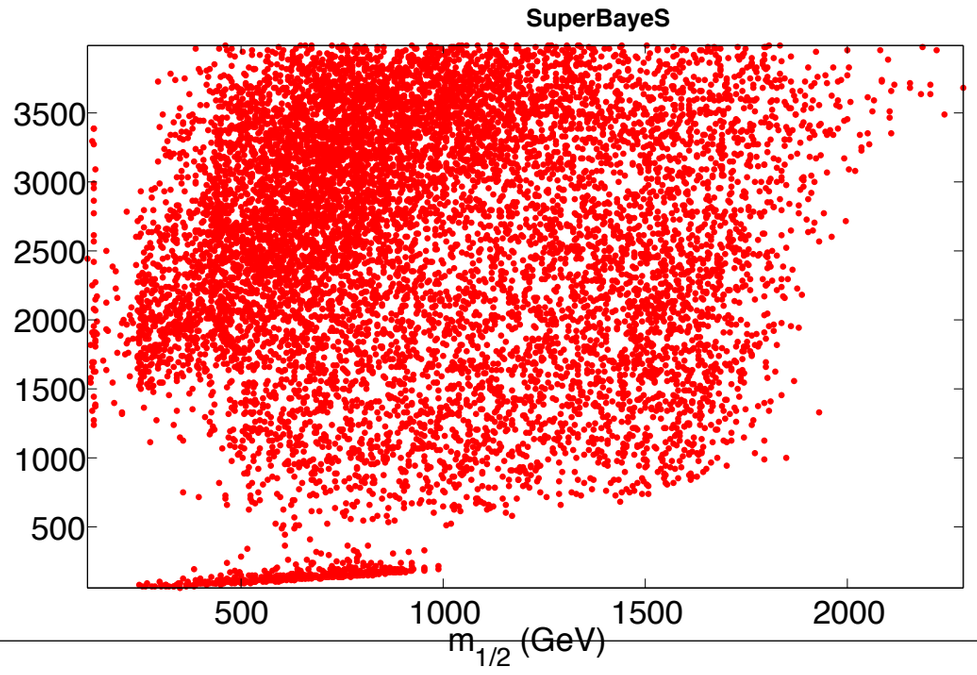
MCMC estimation

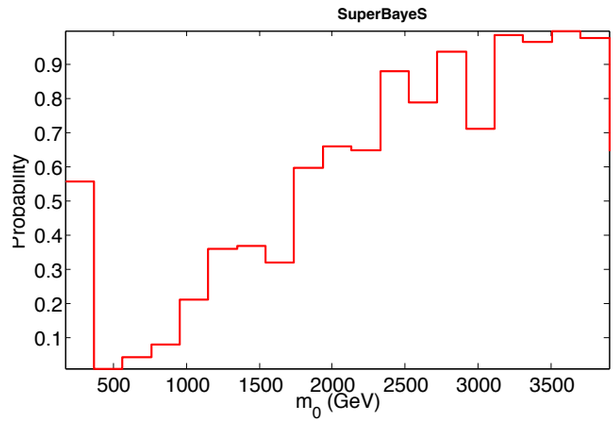
Imperial College London

- **Marginalisation becomes trivial:** create bins along the dimension of interest and simply count samples falling within each bins ignoring all other coordinates

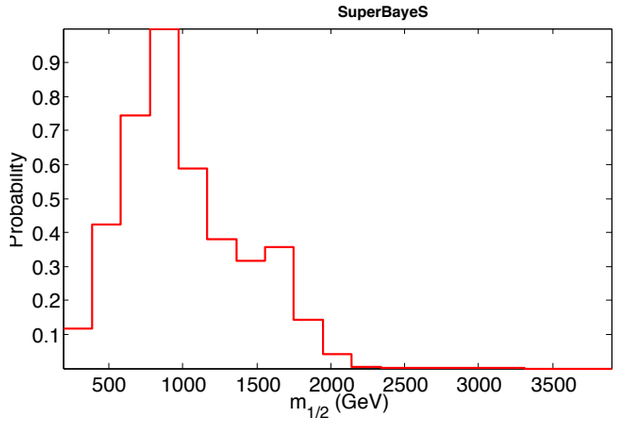
- Examples (from **superbayes.org**) :

2D distribution of samples from joint posterior





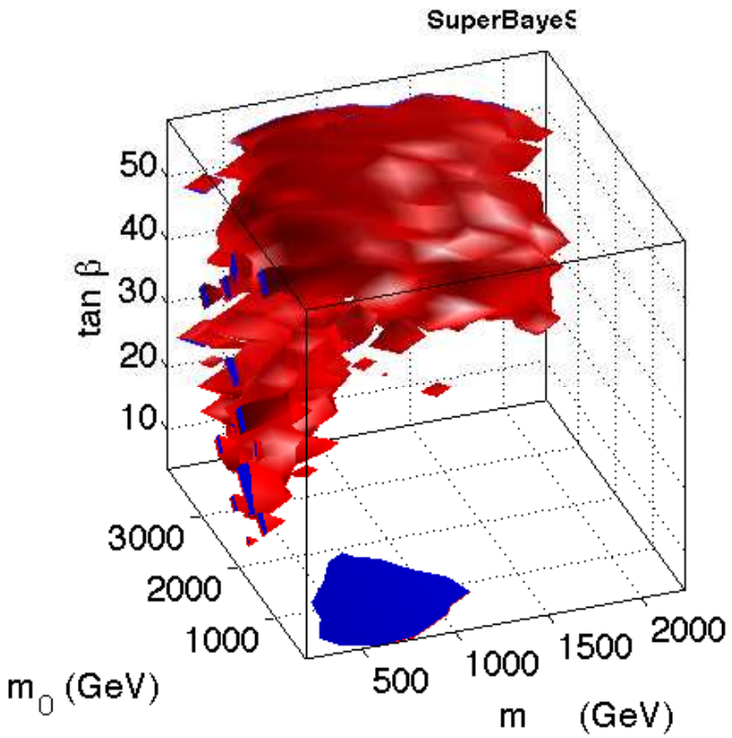
1D marginalised posterior (along y)

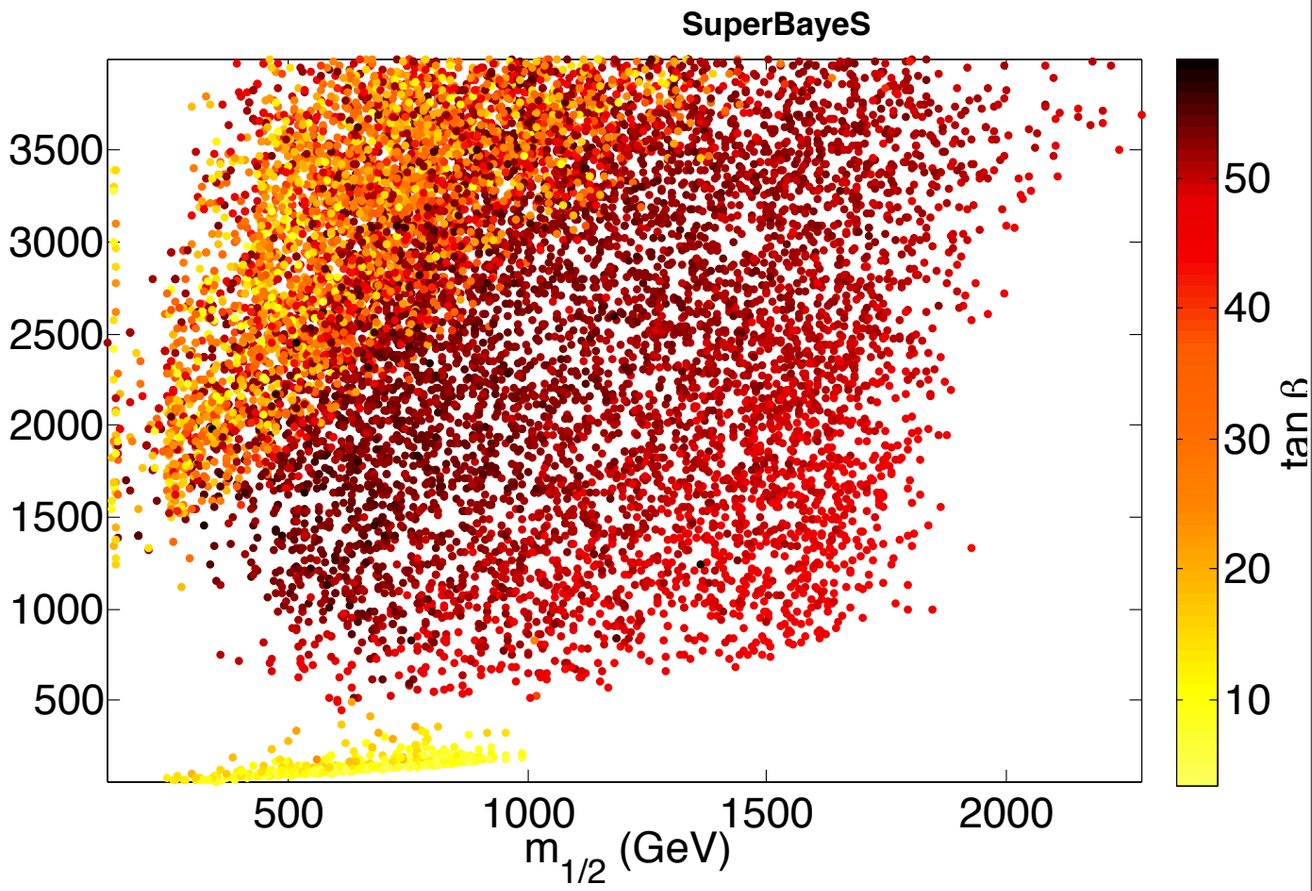


1D marginalised posterior (along x)

Fancier stuff

Imperial College London





The simplest MCMC algorithm

Imperial College London

- Several (sophisticated) algorithms to build a MC are available: e.g. Metropolis-Hastings, Hamiltonian sampling, Gibbs sampling, rejection sampling, mixture sampling, slice sampling and more...

- Arguably the simplest algorithm is the **Metropolis (1954) algorithm:**

  - pick a starting location θ_0 in parameter space, compute $P_0 = p(\theta_0|d)$

  - pick a candidate new location θ_c according to a proposal density $q(\theta_0, \theta_1)$
  - evaluate $P_c = p(\theta_c|d)$ and accept θ_c with probability $\alpha = \min\left(\frac{P_c}{P_0}, 1\right)$
  - if the candidate is accepted, add it to the chain and move there; otherwise stay at θ_0 and count this point once more.

Practicalities

Imperial College London

- Except for simple problems, achieving good MCMC **convergence** (i.e., sampling from the target) and **mixing** (i.e., all chains are seeing the whole of parameter space) can be tricky

- There are several diagnostics criteria around but none is fail-safe. Successful MCMC remains a bit of a black art!

- Things to watch out for:

  - Burn in time

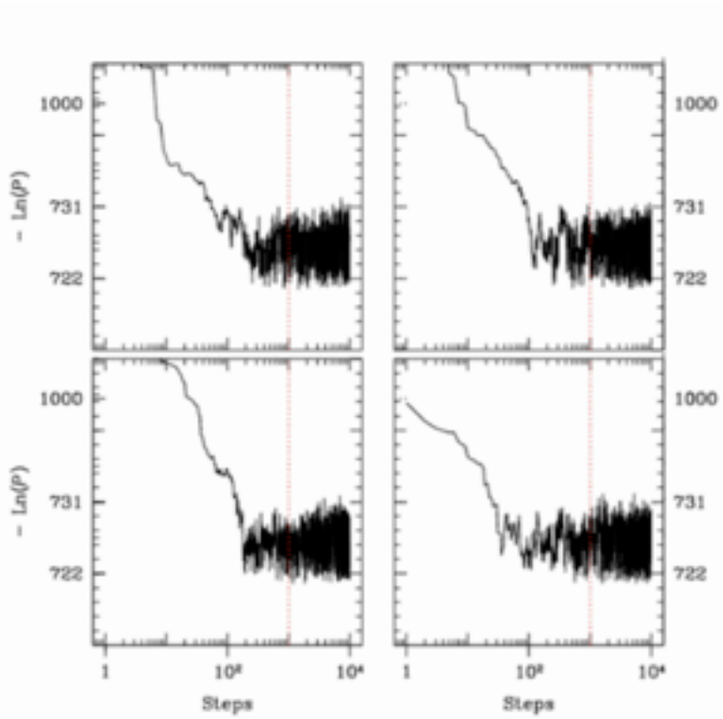
  - Mixing

  - Samples auto-correlation

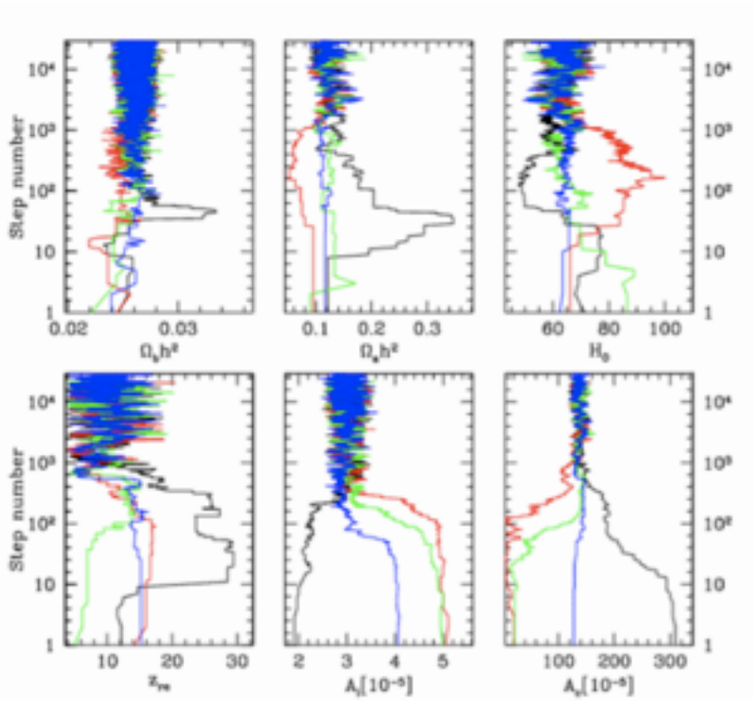
MCMC diagnostics

Imperial College London

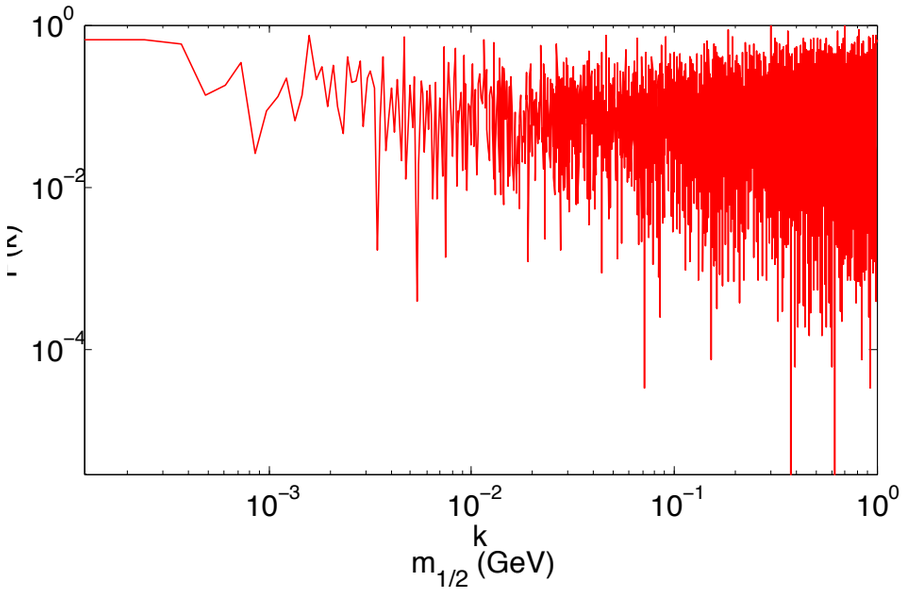
Burn in



Mixing



Power spectrum



(see astro-ph/0405462 for details)

Bayesian model comparison

Bayesian inference chain

Imperial College London

- Select a model (parameters + priors)
- Compute observable quantities as a function of parameters
- Compare with available data
    - derive parameters constraints: **PARAMETER INFERENCE**
    - compute relative model probability: **MODEL COMPARISON**
- Go back and start again

The 3 levels of inference

Imperial College London

LEVEL 1

I have selected a model M and prior $P(\theta|M)$

$$P(\theta|d, M) = \frac{P(d|\theta, M)P(\theta|M)}{P(d|M)}$$

Parameter inference
(assumes M is the true model)

LEVEL 2

Actually, there are several possible models: M_0, M_1,...

$$\text{odds} = \frac{\mathrm{P(M_0|d)}}{\mathrm{P(M_1|d)}}$$

Model comparison
What is the relative plausibility of M_0, M_1,... in light of the data?

LEVEL 3

None of the models is clearly the best

$$P(\theta|d) = \sum_i P(M_i|d)P(\theta|d, M_i)$$

Model averaging
What is the inference on the parameters accounting for model uncertainty?

Level 2: model comparison

Imperial College London

$$P(\theta|d,M) = \frac{P(d|\theta,M)P(\theta|M)}{P(d|M)}$$

Bayesian evidence or model likelihood

The evidence is the integral of the likelihood over the prior:

$$P(d|M) = \int_\Omega d\theta P(d|\theta,M)P(\theta|M)$$

Bayes' Theorem delivers the model's posterior:

$$P(M|d) = \frac{P(d|M)P(M)}{P(d)}$$

When we are comparing two models:

$$\frac{P(M_0|d)}{P(M_1|d)} = \frac{P(d|M_0)}{P(d|M_1)}\frac{P(M_0)}{P(M_1)}$$

The Bayes factor:

$$B_{01} \equiv \frac{P(d|M_0)}{P(d|M_1)}$$

Posterior odds = Bayes factor × prior odds

Scale for the strength of evidence

Imperial College London

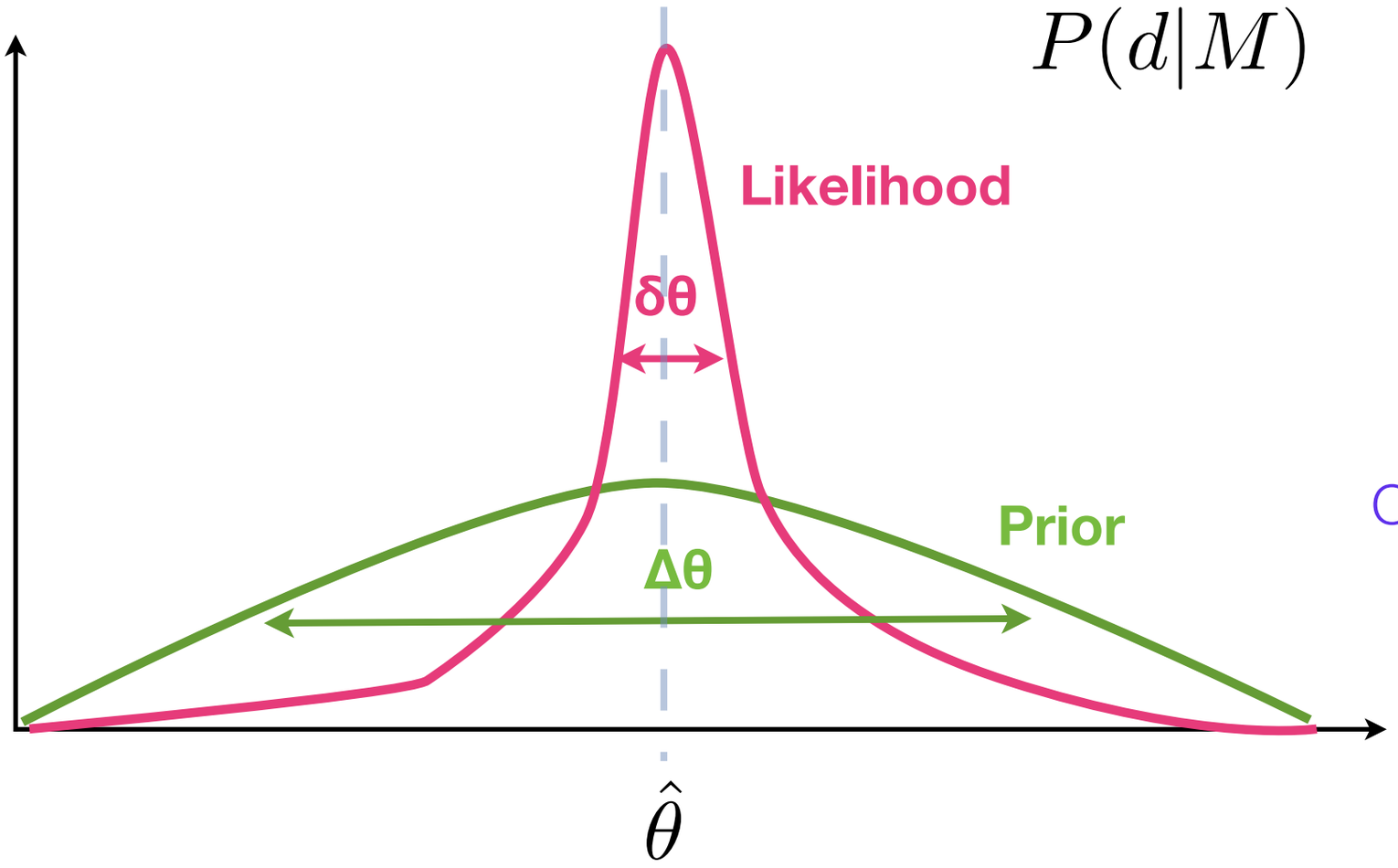
- A (slightly modified) Jeffreys' scale to assess the strength of evidence (**Notice:** this is empirically calibrated!)

|lnB|	relative odds	favoured model's probability	Interpretation
< 1.0	< 3:1	< 0.750	not worth mentioning
< 2.5	< 12:1	0.923	weak
< 5.0	< 150:1	0.993	moderate
> 5.0	> 150:1	> 0.993	strong

An automatic Occam's razor

Imperial College London

- Bayes factor balances quality of fit vs extra model complexity.
- It rewards highly predictive models, penalizing "wasted" parameter space



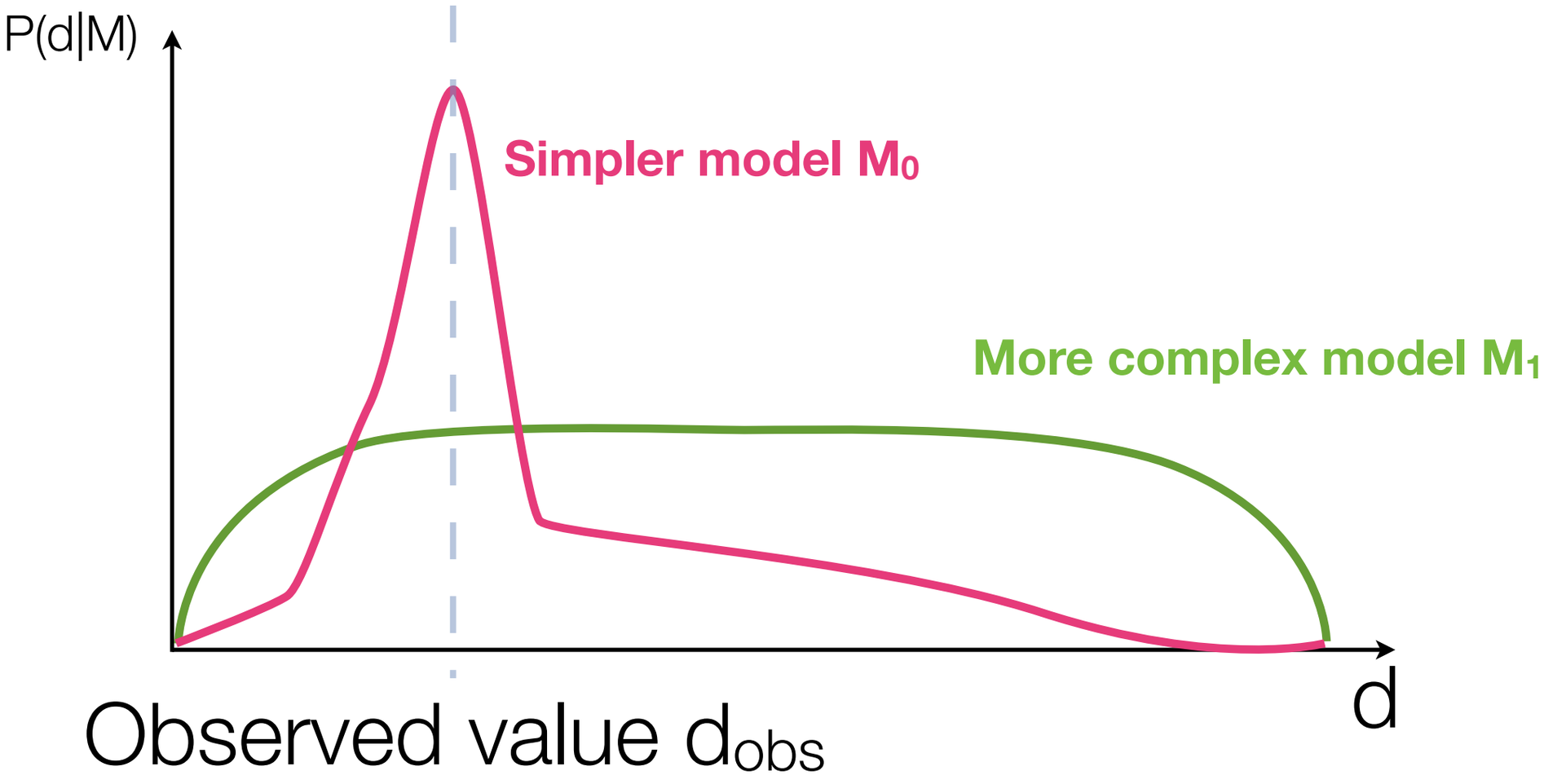
$$P(d|M) = \int d\theta L(\theta) P(\theta|M)$$
$$\approx P(\hat{\theta}) \delta\theta L(\hat{\theta})$$
$$\approx \frac{\delta\theta}{\Delta\theta} L(\hat{\theta})\hat{\theta}$$

Occam's factor

The evidence as predictive probability

Imperial College London

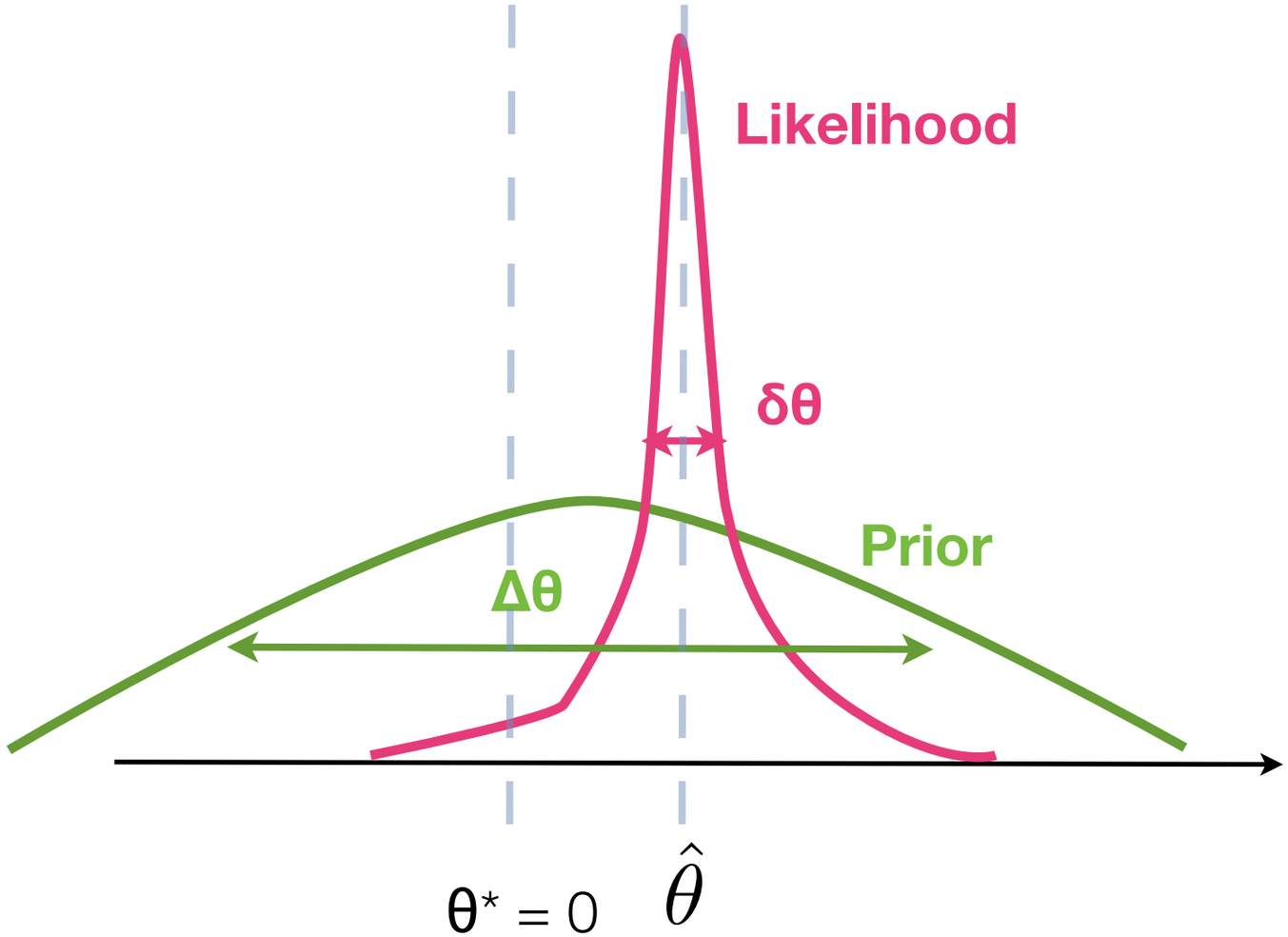
- The evidence can be understood as a function of d to give the predictive probability under the model M:



Simple example: nested models

- This happens often in practice: we have a more complex model, M_1 with prior $P(\theta|M_1)$, which reduces to a simpler model (M_0) for a certain value of the parameter, e.g. $\theta = \theta^* = 0$ (**nested models**)

- Is the extra complexity of M_1 warranted by the data?



Simple example: nested models

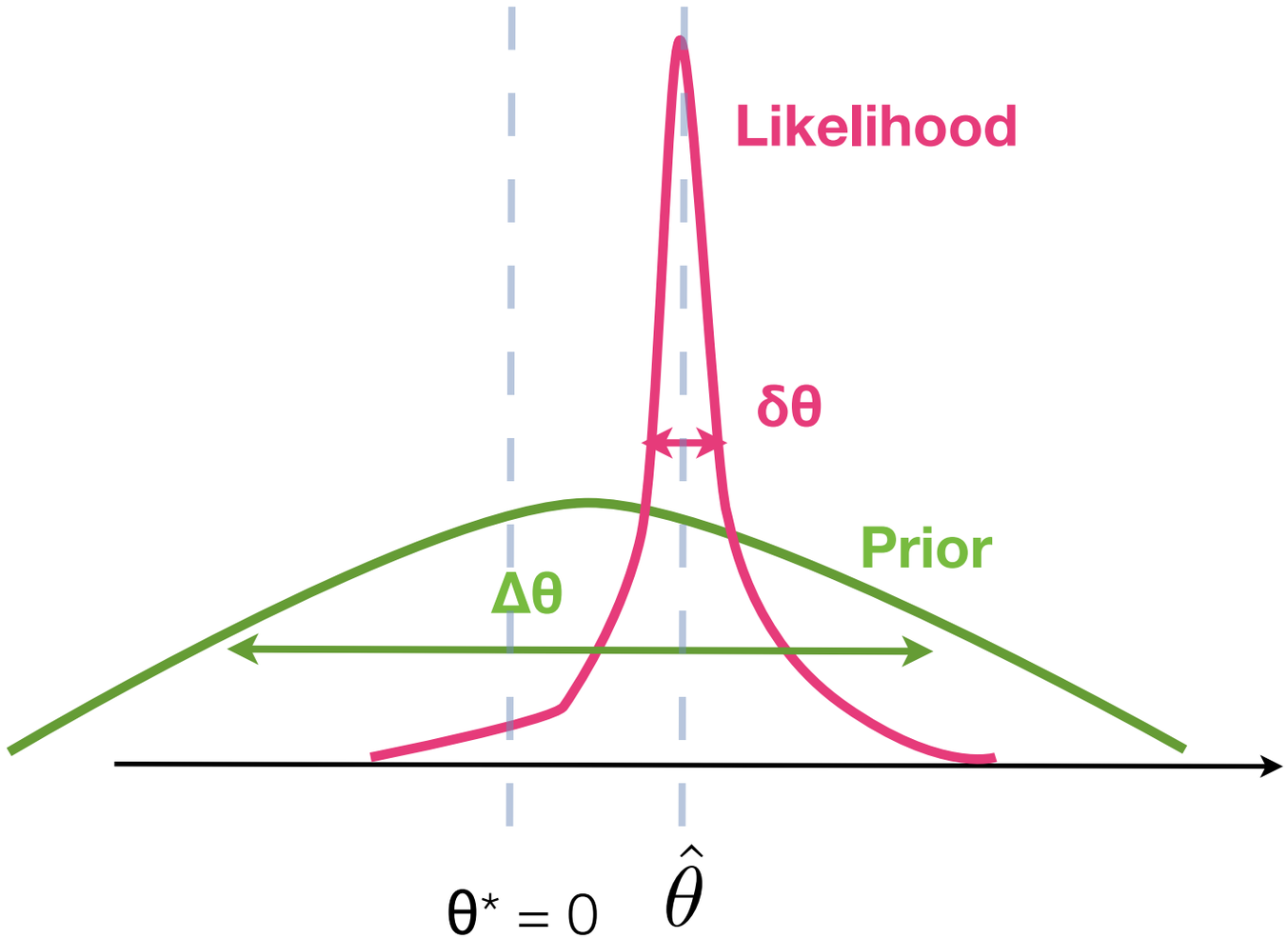
Define: $\lambda \equiv \frac{\hat{\theta} - \theta^\star}{\delta\theta}$

For “informative” data:

$$\ln B_{01} \approx \ln \frac{\Delta\theta}{\delta\theta} - \frac{\lambda^2}{2}$$

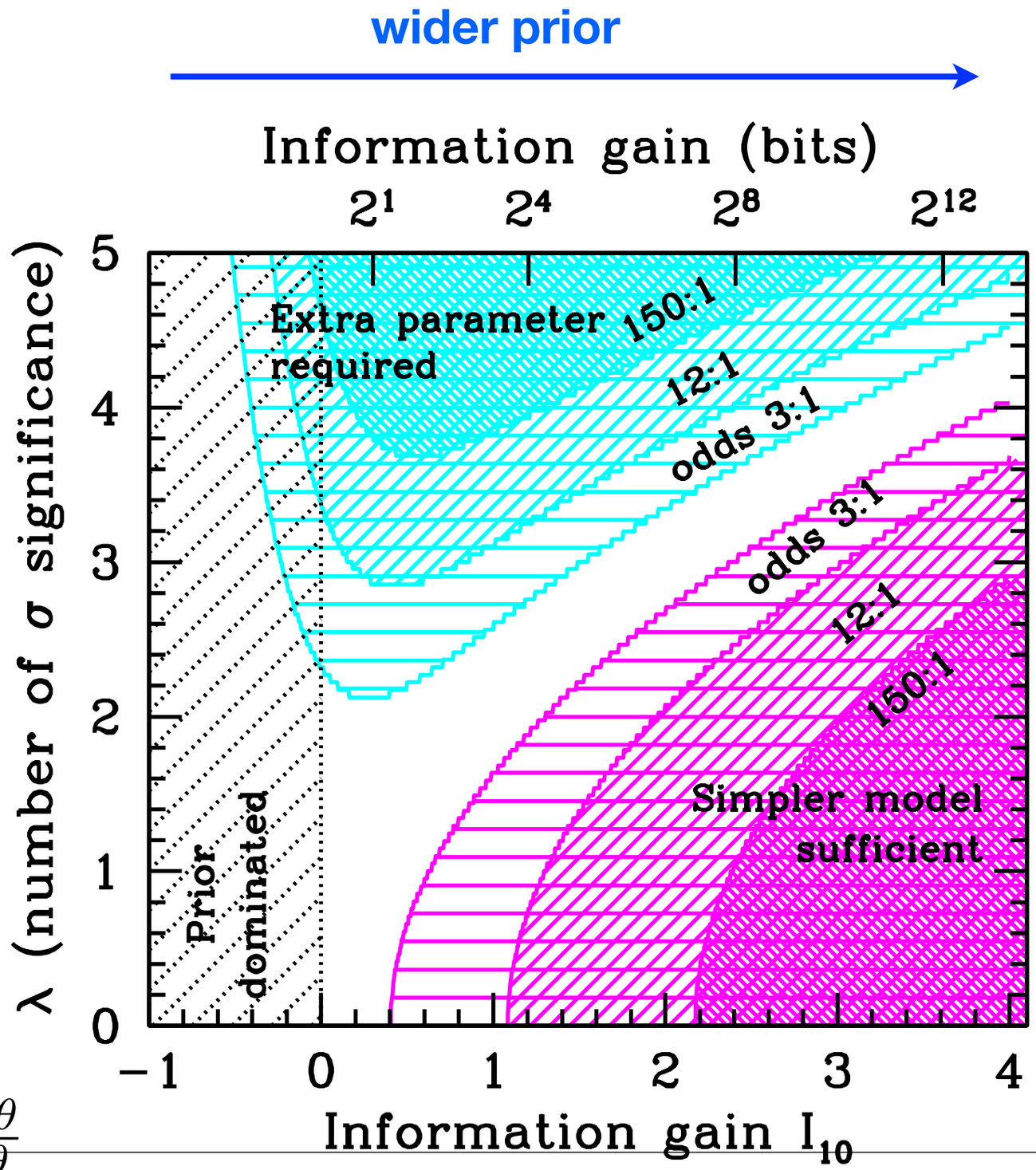
wasted parameter space (favours simpler model)

mismatch of prediction with observed data (favours more complex model)



The rough guide to model comparison

Imperial College London



$$I_{10} \equiv \log_{10} \frac{\Delta\theta}{\delta\theta}$$

Computing the evidence

Imperial College London

evidence: $P(d|M) = \int_\Omega d\theta P(d|\theta, M)P(\theta|M)$

Bayes factor: $B_{01} \equiv \frac{P(d|M_0)}{P(d|M_1)}$

- Usually computational demanding: multi-dimensional integral!
- Several techniques available:
  - Brute force: **thermodynamic integration**
  - **Laplace approximation:** approximate the likelihood to second order around maximum gives Gaussian integrals (for normal prior). Can be inaccurate.
  - **Savage-Dickey density ratio:** good for nested models, gives the Bayes factor
  - **Nested sampling:** clever & efficient, can be used generally

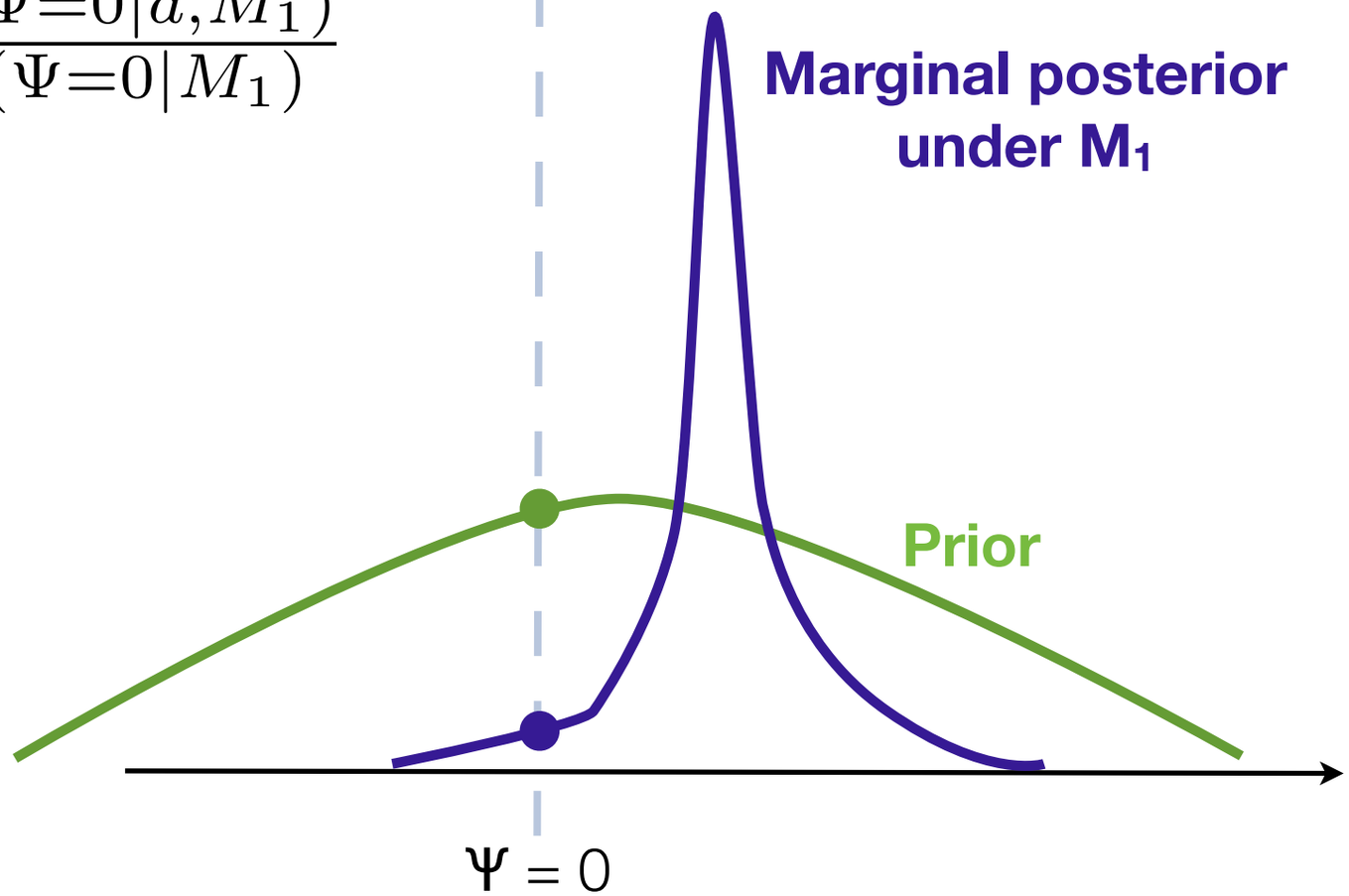
The Savage-Dickey density ratio

Imperial College London

- This methods works for nested models and gives the Bayes factor analytically.
- **Assumptions:** nested models (M_1 with parameters θ,Ψ reduces to M_0 for e.g. $\Psi=0$) and separable priors (i.e. the prior $P(\theta,\Psi|M_1)$ is uncorrelated with $P(\theta|M_0)$)
- Result:

$$B_{01} = \frac{P(\Psi=0|d,M_1)}{P(\Psi=0|M_1)}$$

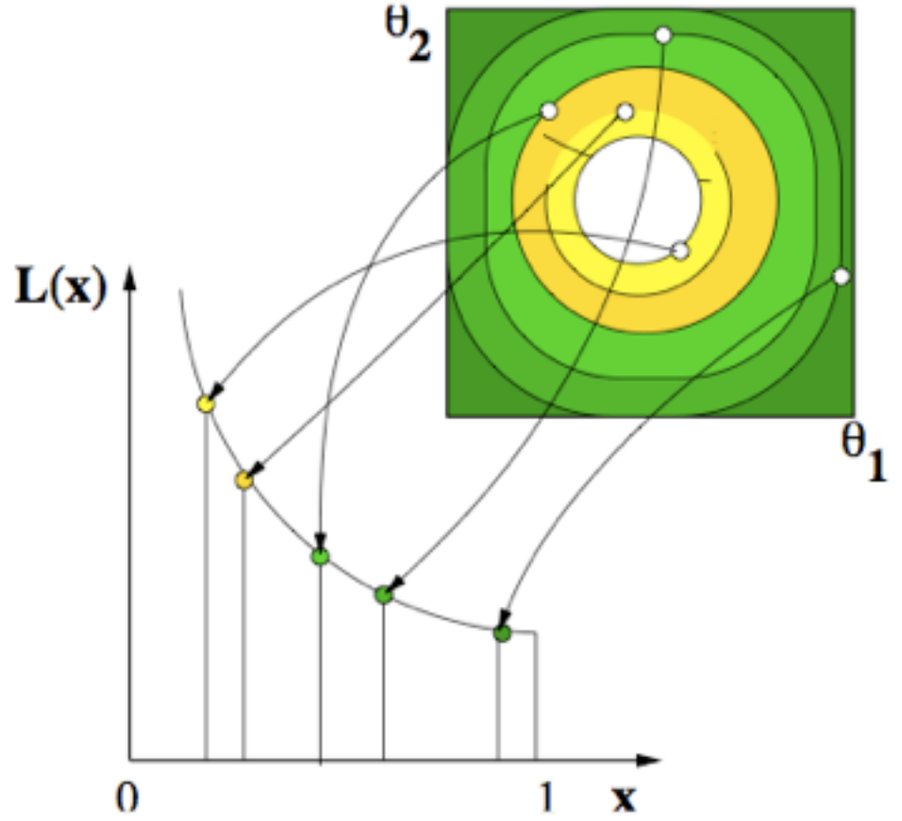
- **Advantages**:
  - analytical
  - often accurate
  - clarifies the role of prior
  - does not rely on Gaussianity



Nested sampling

Imperial College London

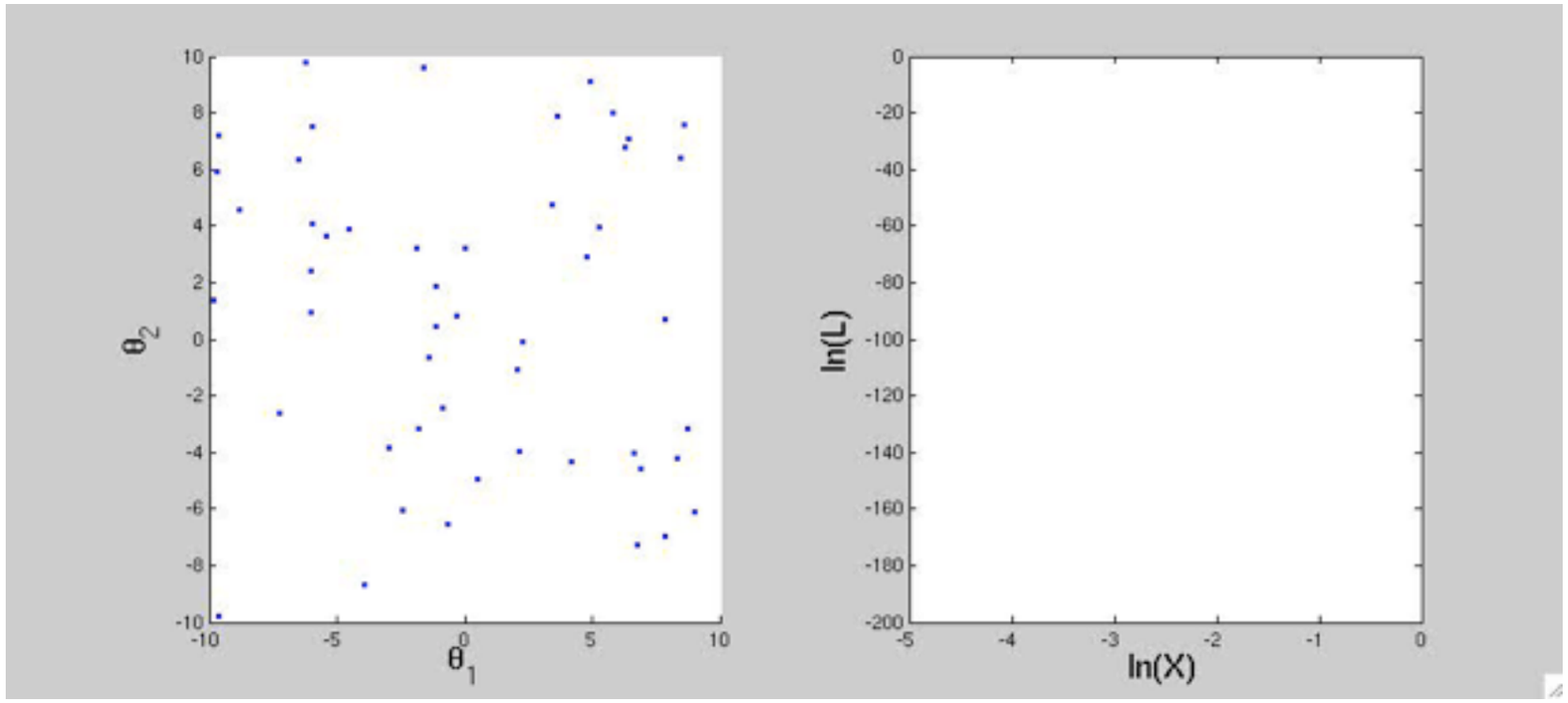
- Perhaps **the** method to compute the evidence
- At the same time, it delivers samples from the posterior: it is a highly efficient sampler! (much better than MCMC in tricky situations)
- Invented by John Skilling in 2005: the gist is to convert a *n*-dimensional integral in a 1D integral that can be done easily.

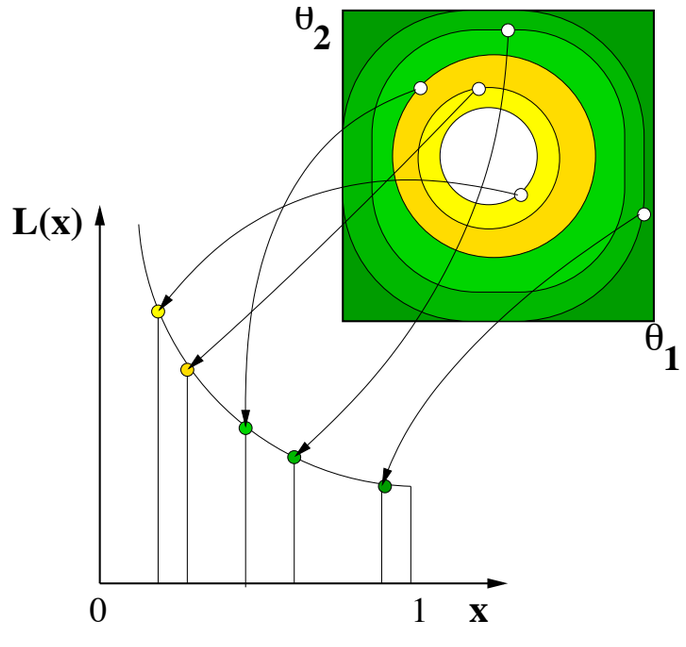


Liddle et al (2006)

Nested sampling

Imperial College London





(animation courtesy of David Parkinson)

An algorithm originally aimed primarily at the Bayesian evidence computation (Skilling, 2006):

$$X(\lambda) = \int_{\mathcal{L}(\theta)>\lambda} P(\theta) d\theta$$

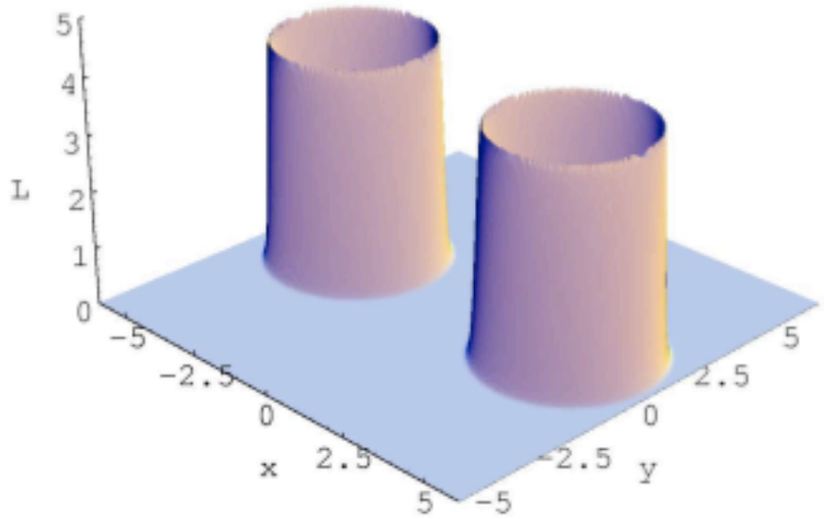
$$P(d) = \int d\theta \mathcal{L}(\theta) P(\theta) = \int_0^1 X(\lambda) d\lambda$$

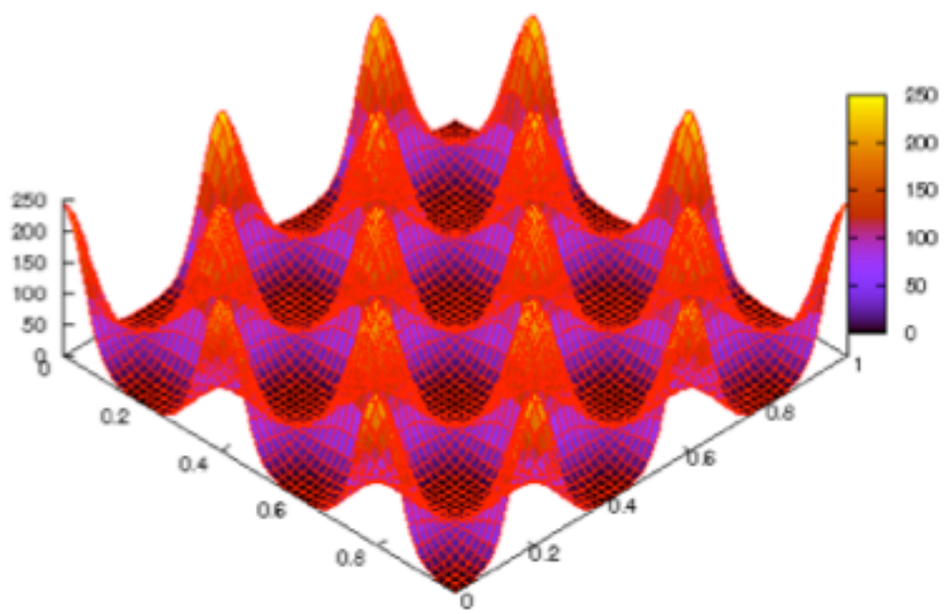
The MultiNest algorithm

Imperial College London

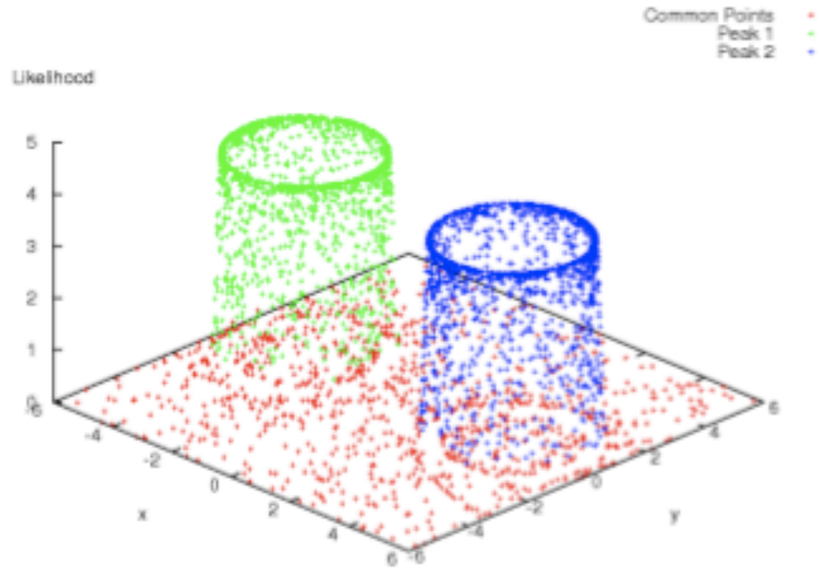
- Feroz & Hobson (2007)

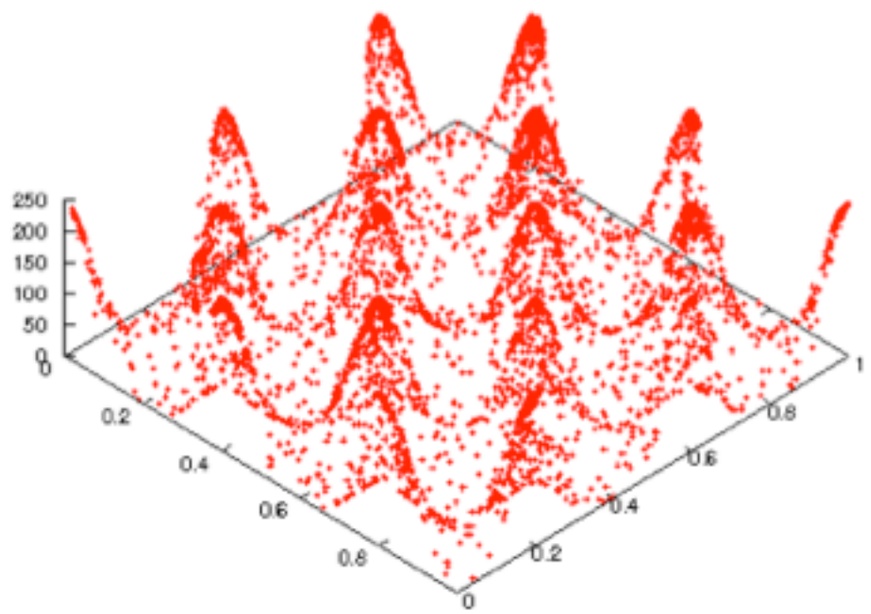
Target





Reconstructed



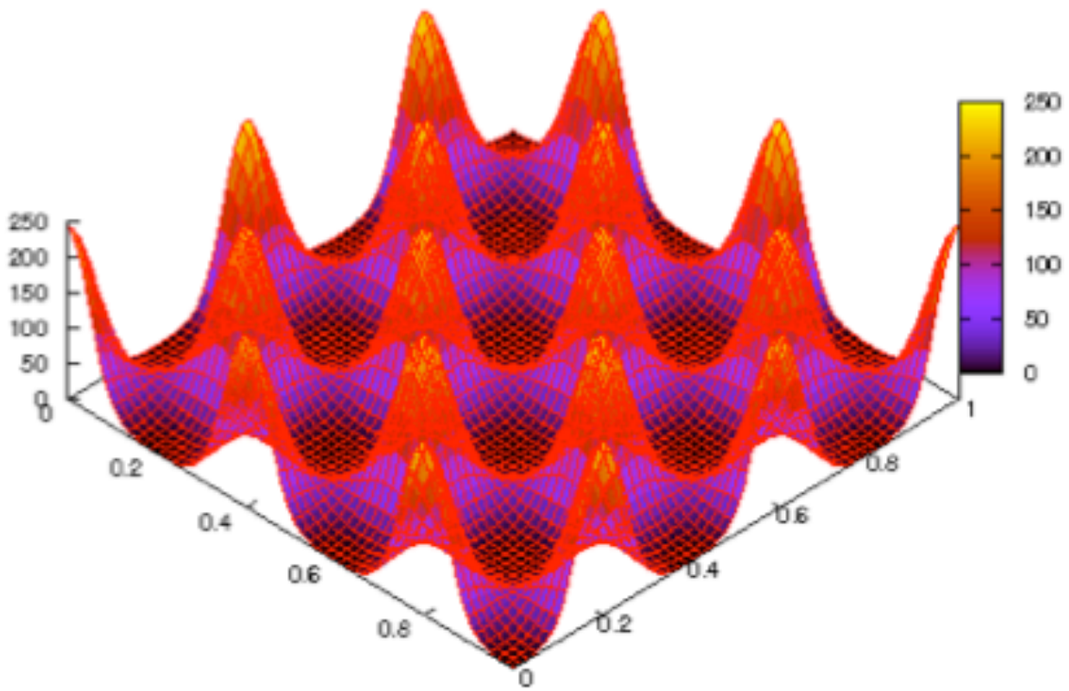


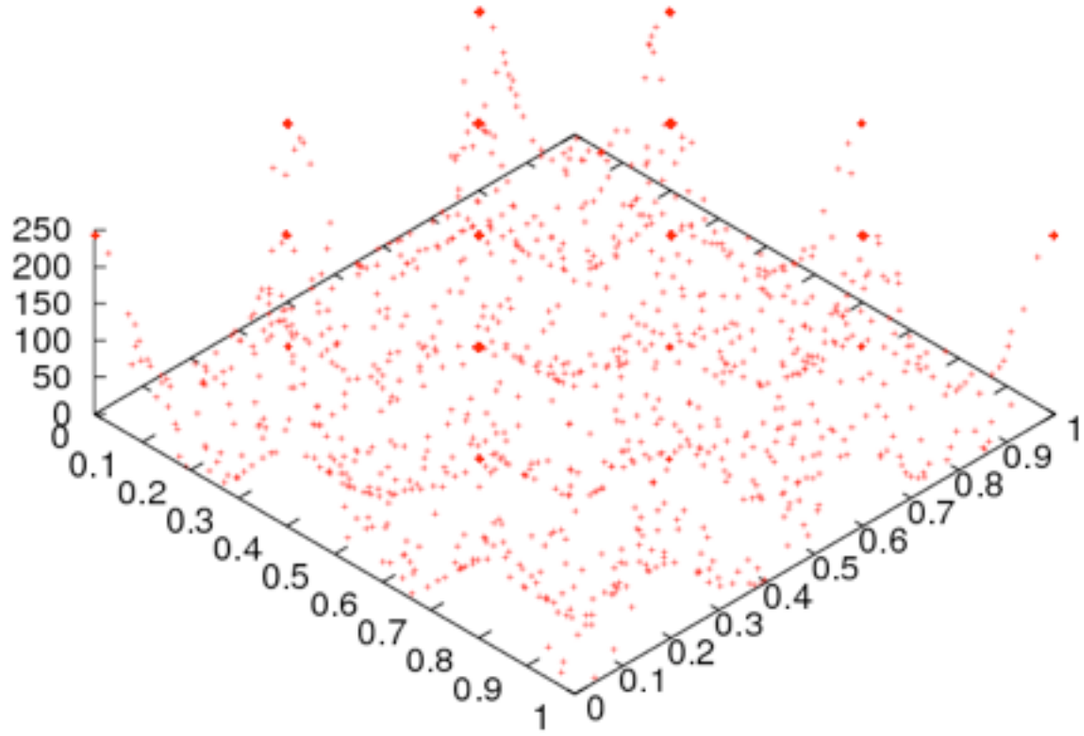
Courtesy Mike Hobson

The egg-box example

Imperial College London

- MultiNest reconstruction of the egg-box likelihood:



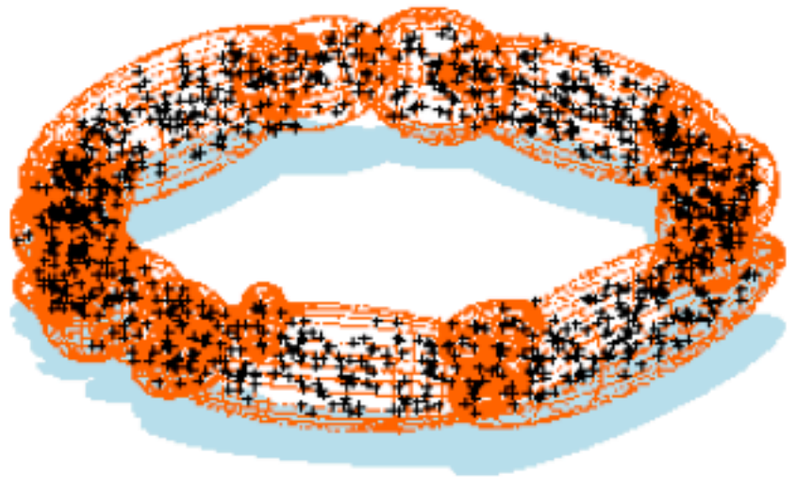


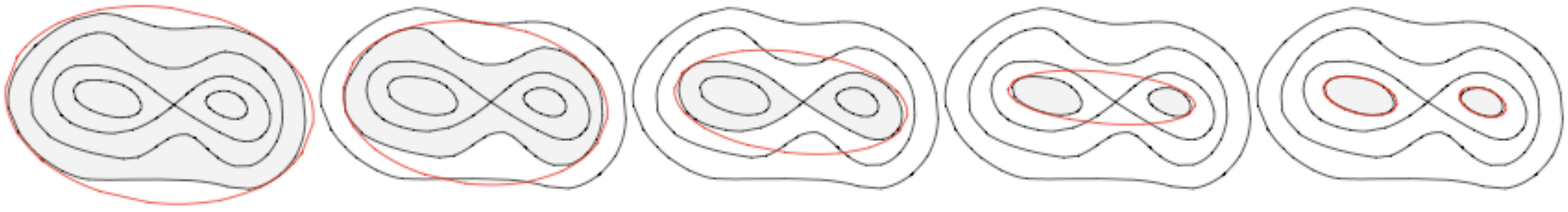
Ellipsoidal decomposition

Imperial College London

Unimodal distribution

Multimodal distribution

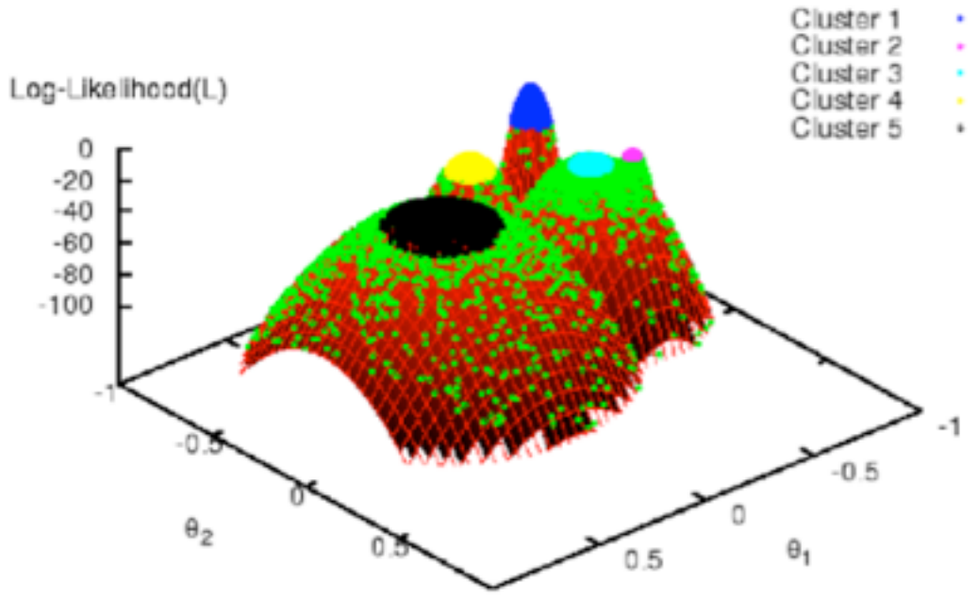




Courtesy Mike Hobson

Multinest: Efficiency

Imperial College London



Gaussian mixture model:

True evidence: log(E) = -5.27

Multinest:

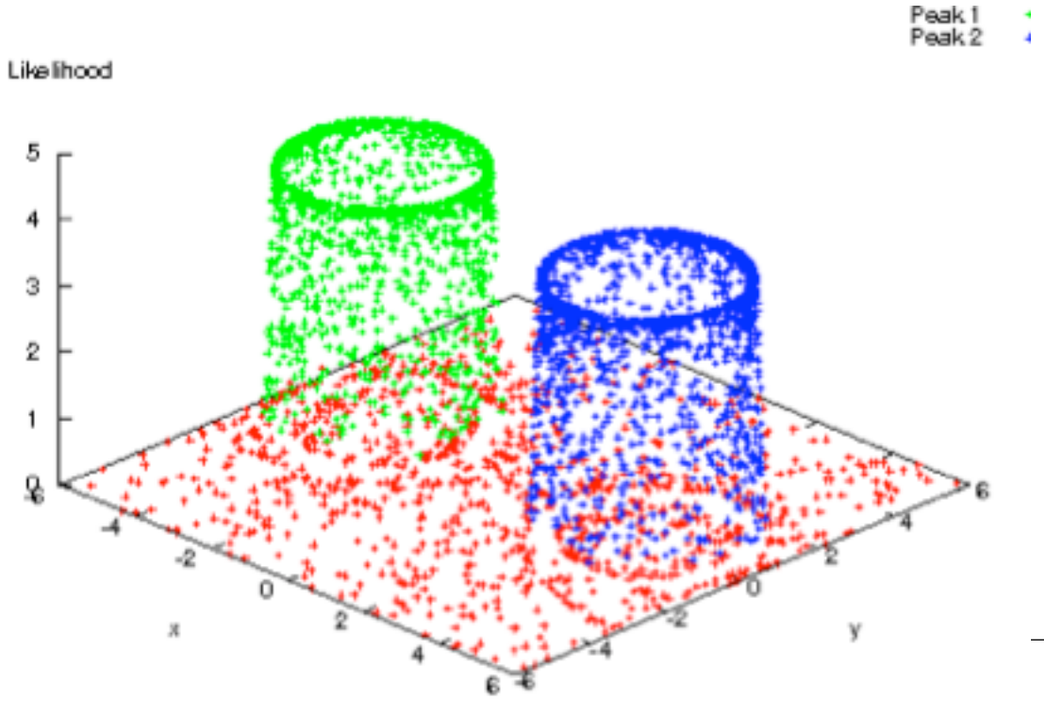
Reconstruction: log(E) = -5.33 ± 0.11

Likelihood evaluations ~ 10^4

Thermodynamic integration:

Reconstruction: log(E) = -5.24 ± 0.12

Likelihood evaluations ~ 10^6



Courtesy Mike Hobson

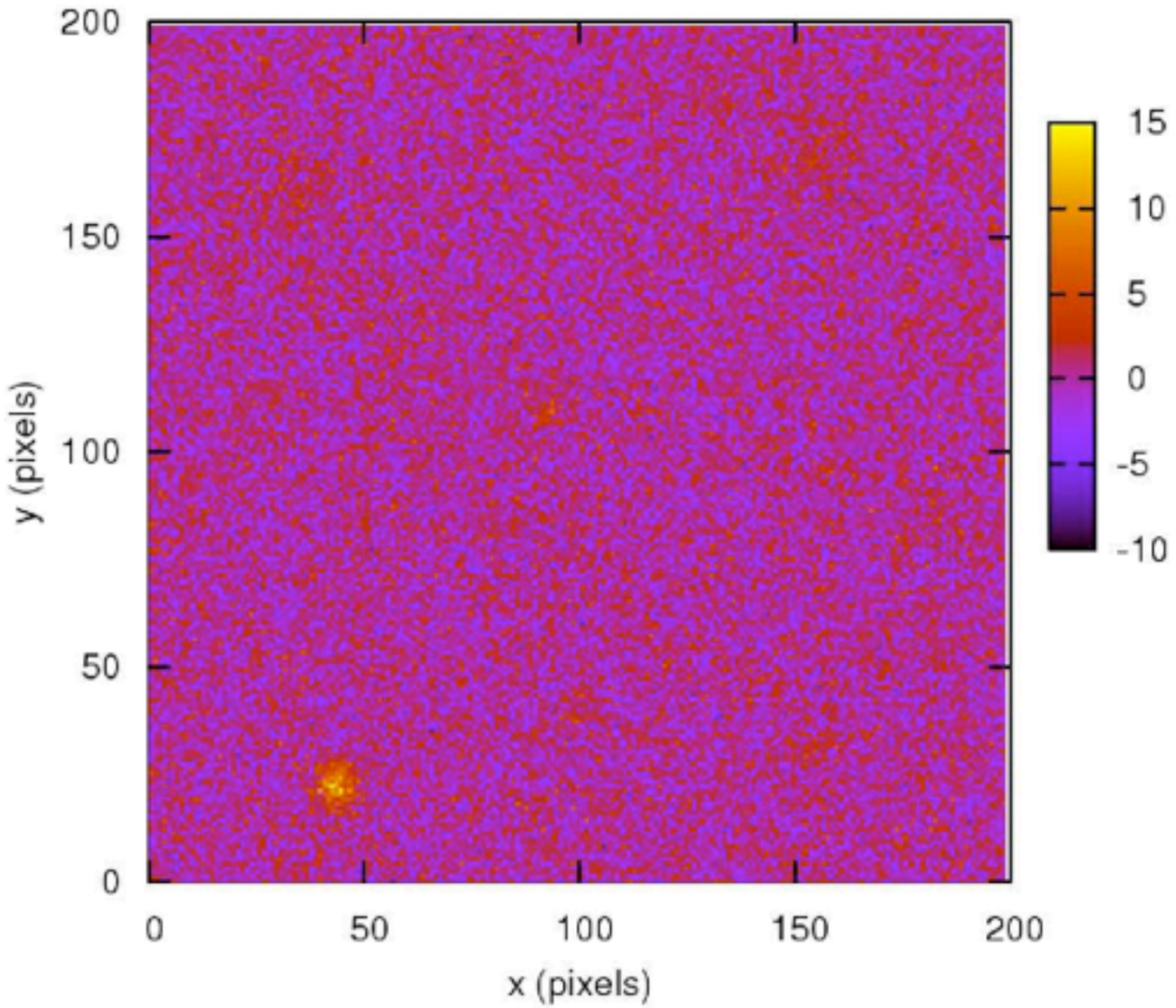
D	N_{like}	efficiency	likes per dimension
2	7000	70%	83
5	18000	51%	7
10	53000	34%	3
20	255000	15%	1.8
30	753000	8%	1.6

A “simple” example: how many sources?

Imperial College London

Feroz and Hobson (2007)

Signal + Noise

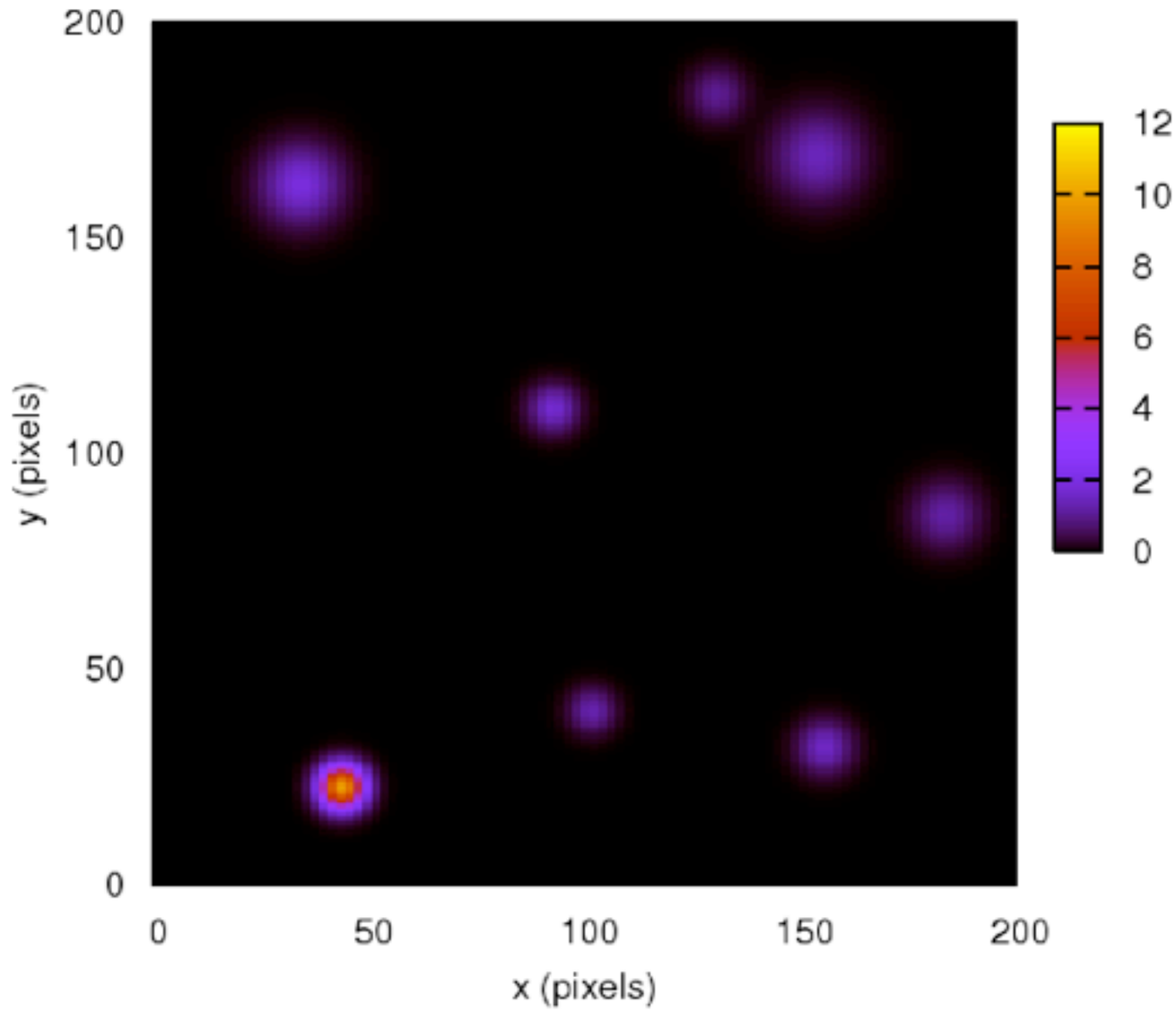


A “simple” example: how many sources?

Imperial College London

Feroz and Hobson (2007)

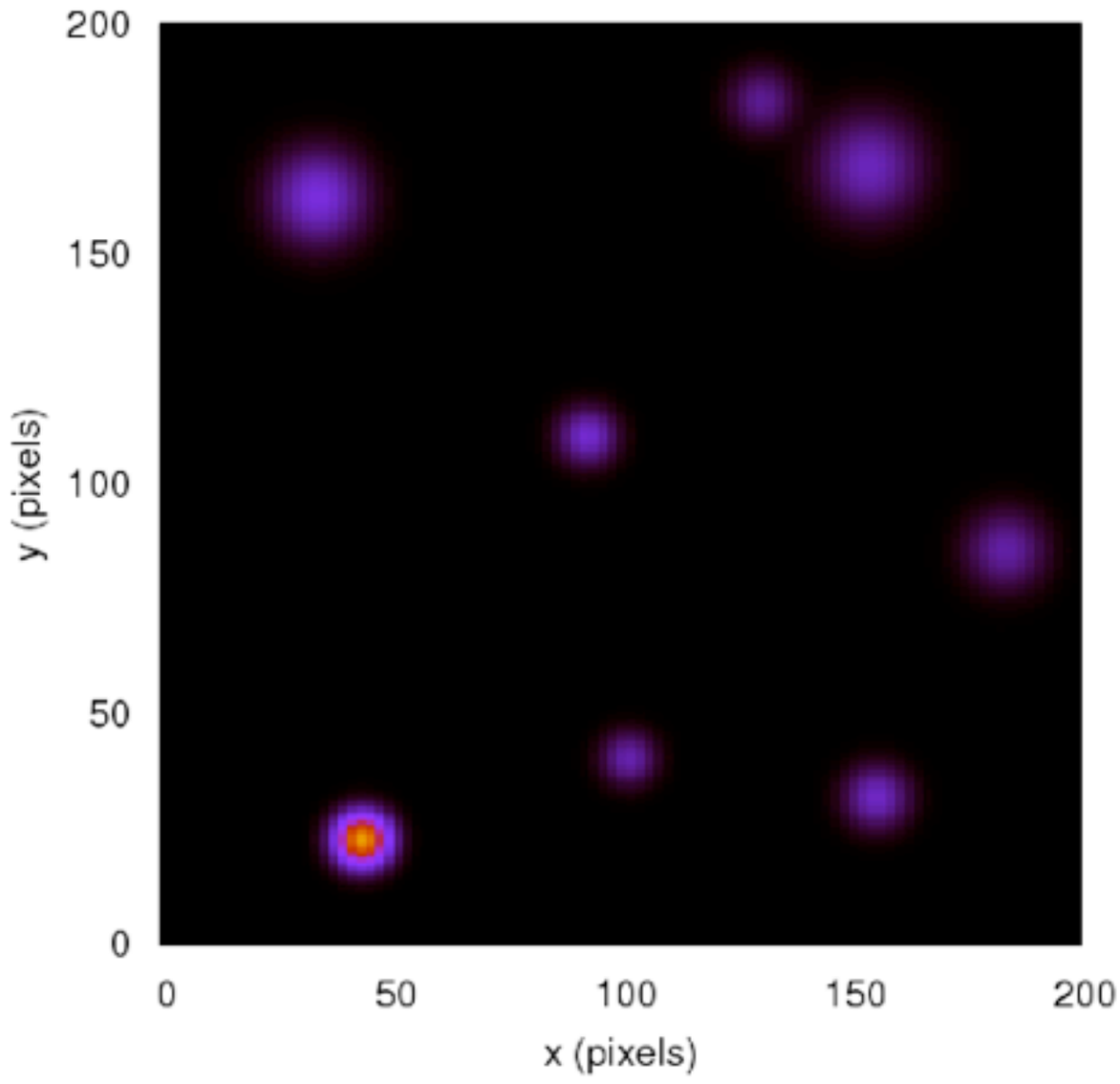
Signal: 8 sources



A “simple” example: how many sources?

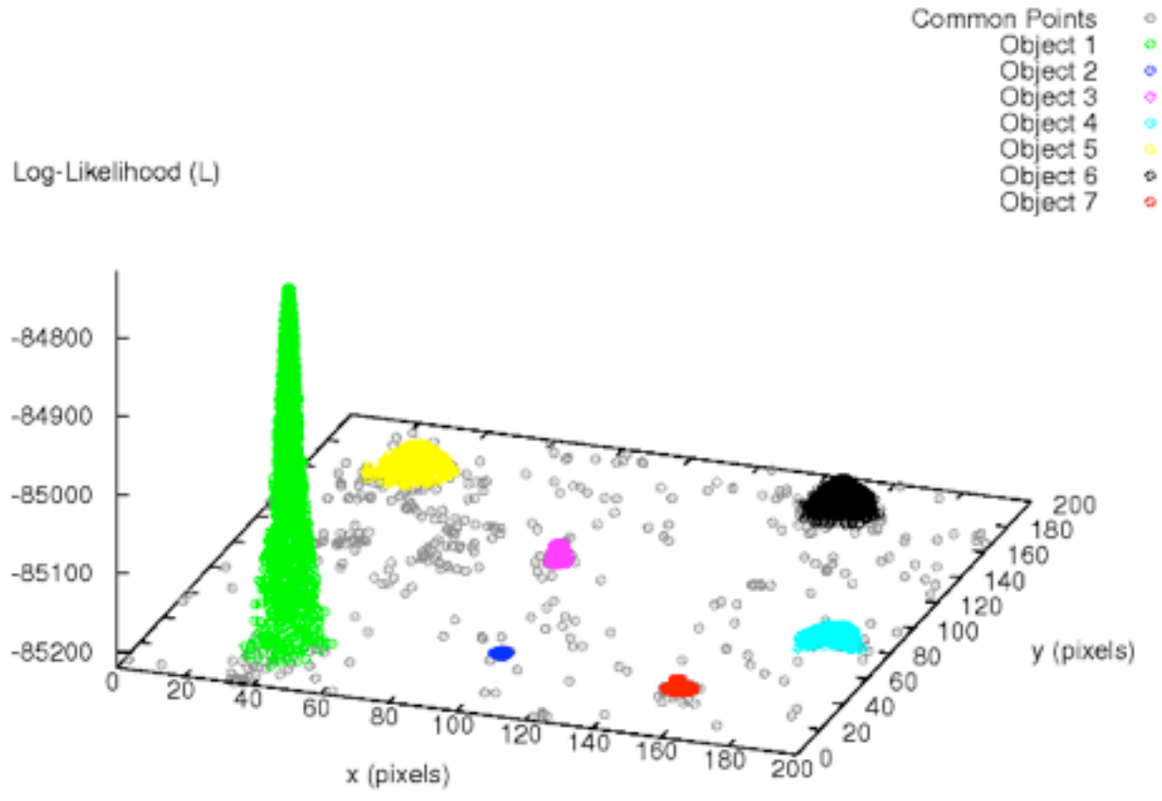
Imperial College London

Feroz and Hobson (2007)

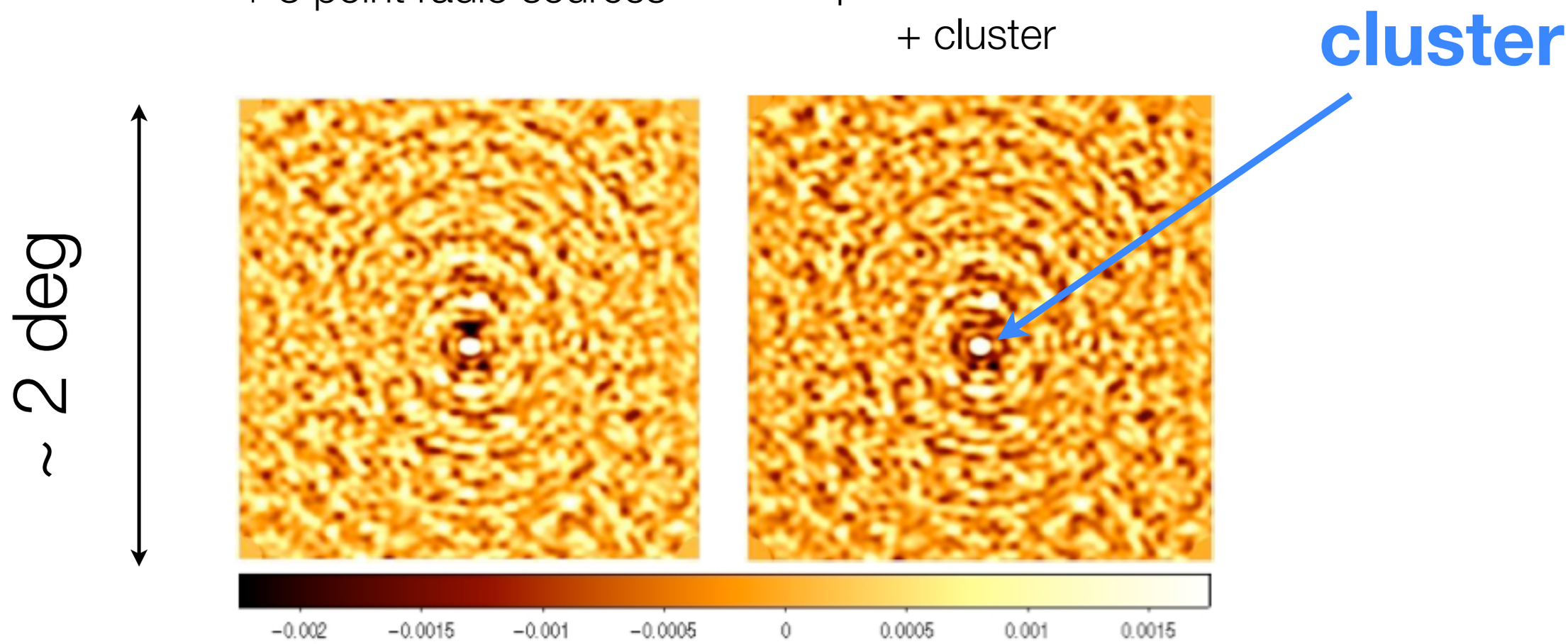


Bayesian reconstruction

7 out of 8 objects correctly identified.
Mistake happens because 2 objects very close.



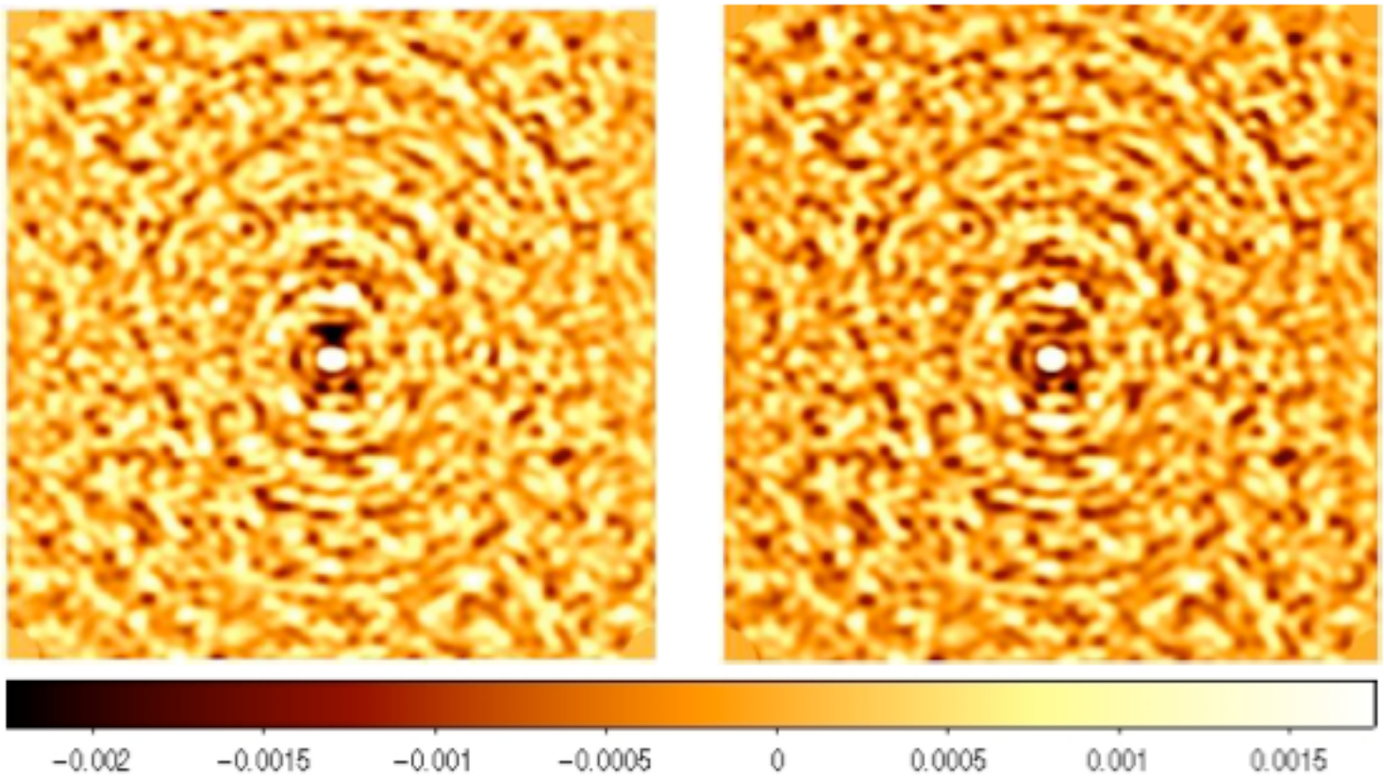
Cluster detection from Sunyaev-Zeldovich effect in cosmic microwave background maps



Feroz et al 2009

Background
+ 3 point radio sources

Background
+ 3 point radio sources
+ cluster



Bayesian model comparison:

R = P(cluster | data)/P(no cluster | data)

$R = 0.35 \pm 0.05$ $R \sim 10^{33}$

Cluster parameters also recovered (position, temperature, profile, etc)

The cosmological concordance model

Imperial College London

Competing model	$\Delta N_{\rm par}$	$\ln B$	Ref	Data	Outcome
Initial conditions					
Isocurvature modes					
CDM isocurvature	+1	−7.6	[58]	WMAP3+, LSS	Strong evidence for adiabaticity
+ arbitrary correlations	+4	−1.0	[46]	WMAP1+, LSS, SN Ia	Undecided
Neutrino entropy	+1	$[-2.5, -6.5]^p$	[60]	WMAP3+, LSS	Moderate to strong evidence for adiabaticity
+ arbitrary correlations	+4	−1.0	[46]	WMAP1+, LSS, SN Ia	Undecided
Neutrino velocity	+1	$[-2.5, -6.5]^p$	[60]	WMAP3+, LSS	Moderate to strong evidence for adiabaticity
+ arbitrary correlations	+4	−1.0	[46]	WMAP1+, LSS, SN Ia	Undecided
Primordial power spectrum					
No tilt ($n_s = 1$)	−1	+0.4	[47]	WMAP1+, LSS	Undecided
		$[-1.1, -0.6]^p$	[51]	WMAP1+, LSS	Undecided
		−0.7	[58]	WMAP1+, LSS	Undecided
		−0.9	[70]	WMAP1+	Undecided
		$[-0.7, -1.7]^{p,d}$	[186]	WMAP3+	$n_s = 1$ weakly disfavoured
		−2.0	[185]	WMAP3+, LSS	$n_s = 1$ weakly disfavoured
		−2.6	[70]	WMAP3+	$n_s = 1$ moderately disfavoured
		−2.9	[58]	WMAP3+, LSS	$n_s = 1$ moderately disfavoured
		$< -3.9^c$	[65]	WMAP3+, LSS	Moderate evidence at best against $n_s \neq 1$
Running	+1	$[-0.6, 1.0]^{p,d}$	[186]	WMAP3+, LSS	No evidence for running
		$< 0.2^c$	[166]	WMAP3+, LSS	Running not required
Running of running	+2	$< 0.4^c$	[166]	WMAP3+, LSS	Not required
Large scales cut–off	+2	$[1.3, 2.2]^{p,d}$	[186]	WMAP3+, LSS	Weak support for a cut–off
Matter–energy content					
Non–flat Universe	+1	−3.8	[70]	WMAP3+, HST	Flat Universe moderately favoured
		−3.4	[58]	WMAP3+, LSS, HST	Flat Universe moderately favoured
Coupled neutrinos	+1	−0.7	[193]	WMAP3+, LSS	No evidence for non–SM neutrinos
Dark energy sector					
$w(z) = w_{\rm eff} \neq -1$	+1	$[-1.3, -2.7]^p$	[187]	SN Ia	Weak to moderate support for Λ
		−3.0	[50]	SN Ia	Moderate support for Λ
		−1.1	[51]	WMAP1+, LSS, SN Ia	Weak support for Λ
		$[-0.2, -1]^p$	[188]	SN Ia, BAO, WMAP3	Undecided
		$[-1.6, -2.3]^d$	[189]	SN Ia, GRB	Weak support for Λ
$w(z) = w_0 + w_1 z$	+2	$[-1.5, -3.4]^p$	[187]	SN Ia	Weak to moderate support for Λ
		−6.0	[50]	SN Ia	Strong support for Λ
		−1.8	[188]	SN Ia, BAO, WMAP3	Weak support for Λ
$w(z) = w_0 + w_a(1-a)$	+2	−1.1	[188]	SN Ia, BAO, WMAP3	Weak support for Λ
		$[-1.2, -2.6]^d$	[189]	SN Ia, GRB	Weak to moderate support for Λ
Reionization history					
No reionization ($\tau = 0$)	−1	−2.6	[70]	WMAP3+, HST	$\tau \neq 0$ moderately favoured
No reionization and no tilt	−2	−10.3	[70]	WMAP3+, HST	Strongly disfavoured

from Trotta (2008)

lnB < 0: favours ΛCDM

Interlude: frequentist hypothesis testing

Imperial College London

- **Warning:** frequentist hypothesis testing (e.g., likelihood ratio test) cannot be interpreted as a statement about the probability of the hypothesis!

- **Example:** to test the null hypothesis H_0: $\theta = 0$, draw *n* normally distributed points (with known variance σ^2). The χ^2 is distributed as a chi-square distribution with *(n-1)* degrees of freedom (dof). Pick a significance level α (or p-value, e.g. $\alpha = 0.05$). If $P(\chi^2 > \chi^2_{obs}) < \alpha$ reject the null hypothesis.

- This is a statement about the likelihood of observing data as extreme or more extreme than have been measured *assuming the null hypothesis is correct*.

- **It is not a statement about the probability of the null hypothesis itself and cannot be interpreted as such! (or you'll make gross mistakes)**

- *The use of p-values implies that a hypothesis that may be true can be rejected because it has not predicted observable results that have not actually occurred.* (Jeffreys, 1961)

The significance of significance

Imperial College London

- **Important:** A 2-sigma result does not wrongly reject the null hypothesis 5% of the time: **at least 29% of 2-sigma results are wrong!**
  - Take an equal mixture of H_0, H_1
  - Simulate data, perform hypothesis testing for H_0
  - Select results rejecting H_0 at 1-α CL
  - What fraction of those results did actually come from H_0 ("true nulls", should not have been rejected)?

p–value	sigma	fraction of true nulls	lower bound
0.05	1.96	0.51	0.29
0.01	2.58	0.20	0.11
0.001	3.29	0.024	0.018

Recommended: Sellke, Bayarri & Berger, *The American Statistician*, 55, 1 (2001)

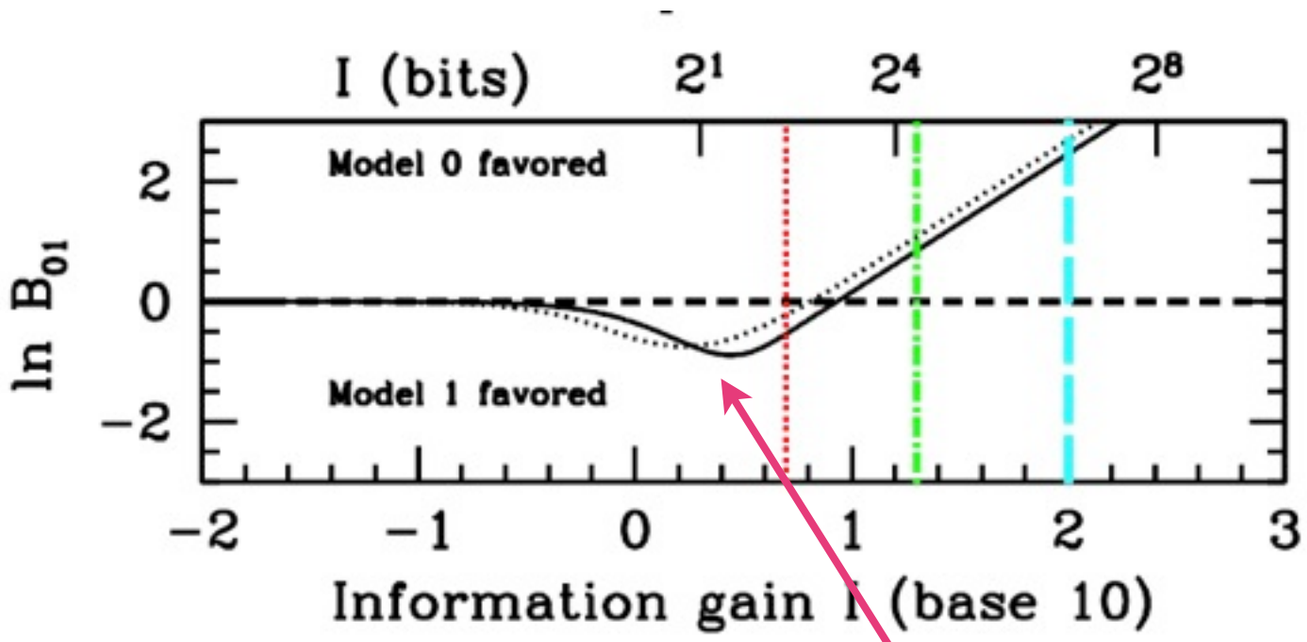
“Prior-free” evidence bounds

Imperial College London

- What if we do not know how to set the prior? For nested models, we can still choose a prior that will maximise the support for the more complex model:

wider prior (fixed data)

larger sample (fixed prior and significance)



maximum evidence for Model 1

Maximum evidence for a detection

Imperial College London

- **The absolute upper bound:** put all prior mass for the alternative onto the observed maximum likelihood value. Then

$$B < \exp(-\chi^2/2)$$

- **More reasonable class of priors:** symmetric and unimodal around Ψ=0, then (α = significance level)

$$B < \frac{-1}{\exp(1)\alpha \ln \alpha}$$

If the upper bound is small, no other choice of prior will make the extra parameter significant.

Sellke, Bayarri & Berger, *The American Statistician*, 55, 1 (2001)

How to interpret the “number of sigma’s”

Imperial College London

α	sigma	Absolute bound on lnB (B)	“Reasonable” bound on lnB (B)
0.05	2.0	2.0 (7:1) weak	0.9 (3:1) undecided
0.003	3.0	4.5 (90:1) moderate	3.0 (21:1) moderate
0.0003	3.6	6.48 (650:1) strong	5.0 (150:1) strong

A conversion table

Imperial College London

p–value	$\bar{B}$	$\ln \bar{B}$	sigma	category
0.05	2.5	0.9	2.0	
0.04	2.9	1.0	2.1	'weak' at best
0.01	8.0	2.1	2.6	
0.006	12	2.5	2.7	'moderate' at best
0.003	21	3.0	3.0	
0.001	53	4.0	3.3	
0.0003	150	5.0	3.6	'strong' at best
6×10^{-7}	43000	11	5.0	

Rule of thumb:

a n-sigma result should be interpreted as a n-1 sigma result

Application: the spectral tilt

Imperial College London

- Is the spectrum of primordial fluctuations scale-invariant?

- Model comparison:
  n = 1 vs n ≠ 1 (with inflation-motivated prior)

- Results:
  n ≠ 1 favoured with odds of 17:1 (Trotta 2007)
  n ≠ 1 favoured with odds of 15:1 (Kunz, Trotta & Parkinson 2007)
  n ≠ 1 favoured with odds of 7:1 (Parkinson 2007 et al 2006)

- Upper bound: **odds of 49:1 at best for n ≠ 1** (Gordon and Trotta 2007)

Application: dipole modulation

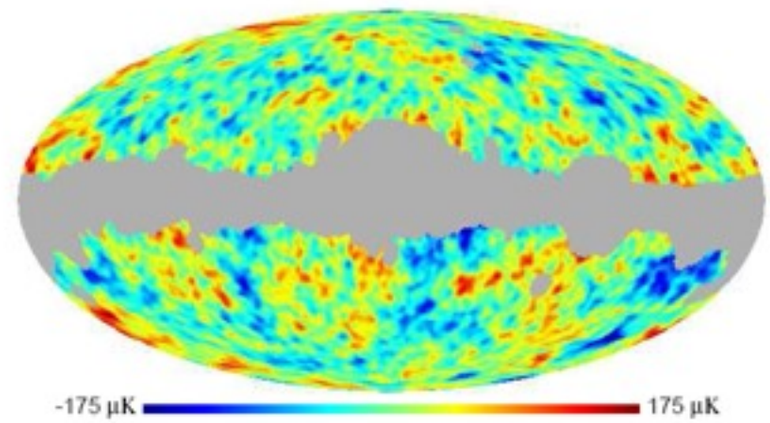
Imperial College London

- Eriksen et al (2004) found hints for a dipolar modulation in WMAP1 ILC map

- Adding a phenomenological dipole pattern improves the chi-square by 9 units (for 3 extra parameters)

- Is this significant evidence?

- Not really: upper bound on B is odds of 9:1
  The absolute upper bound is about the same
  (Gordon and Trotta 2007)



Model complexity

Imperial College London

- "Number of free parameters" is a relative concept. The relevant scale is set by the prior range
- How many parameters can the data support, regardless of whether their detection is significant?
- **Bayesian complexity** or effective number of parameters:

$$\mathcal{C}_b = \overline{\chi^2(\theta)} - \chi^2(\hat{\theta})$$
$$= \sum_i \frac{1}{1 + (\sigma_i/\Sigma_i)^2}$$

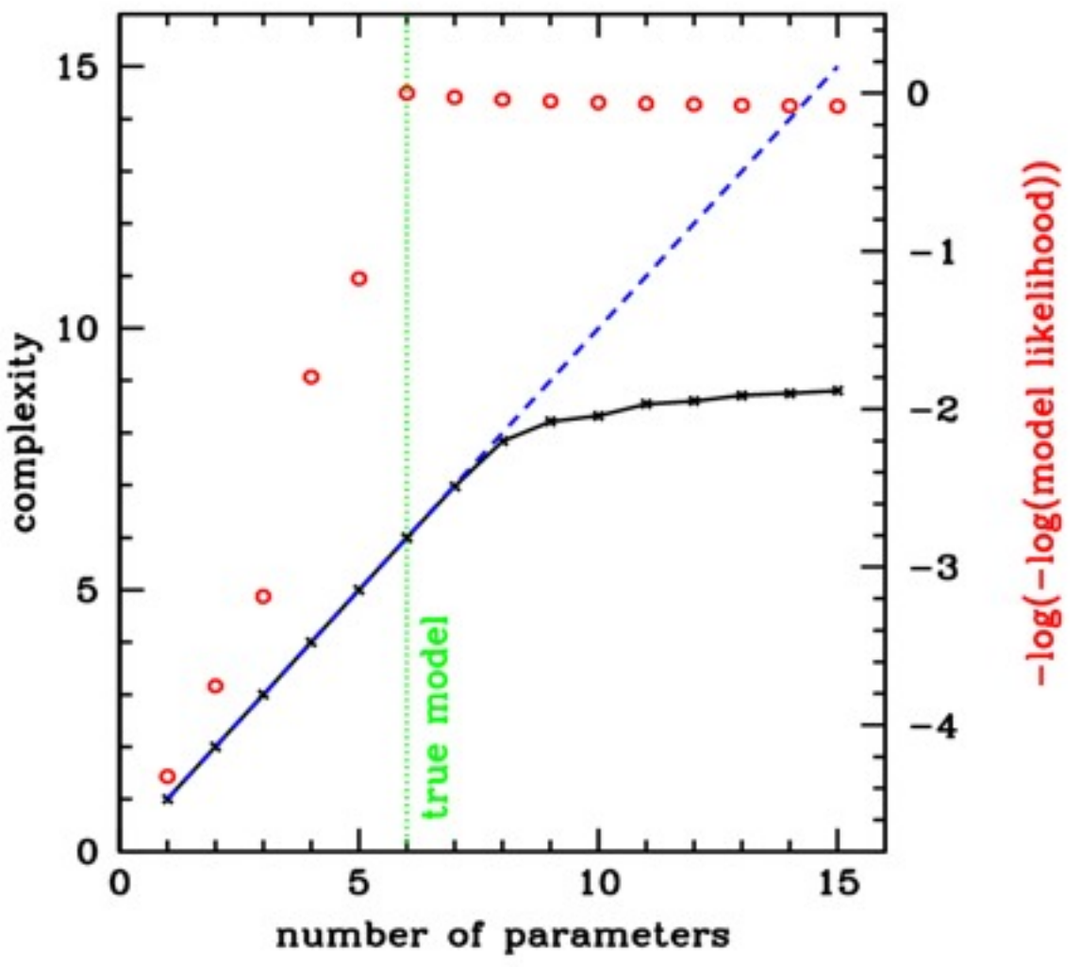
Kunz, RT & Parkinson, astro-ph/0602378, Phys. Rev. D 74, 023503 (2006)
Following Spiegelhalter et al (2002)

Polynomial fitting

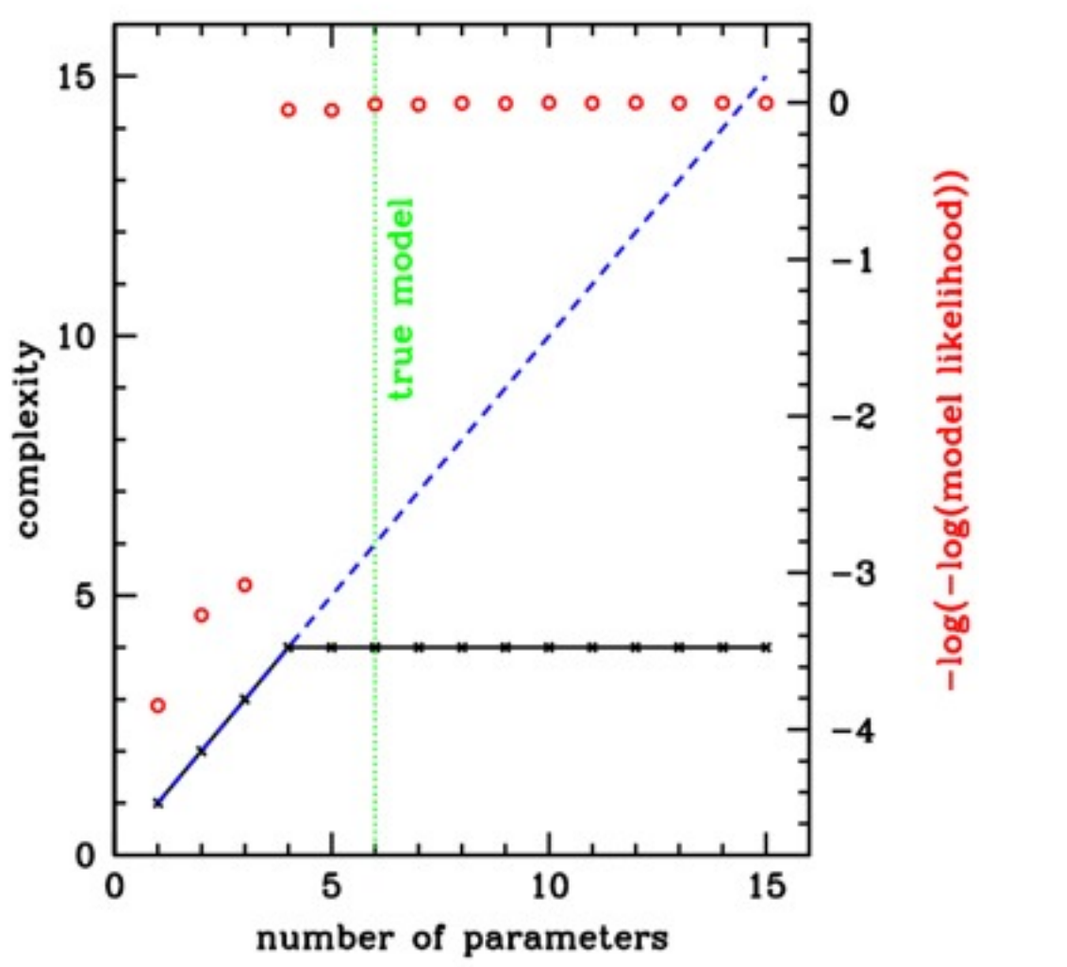
Imperial College London

- Data generated from a model with n = 6:

GOOD DATA
Max supported complexity ¼ 9



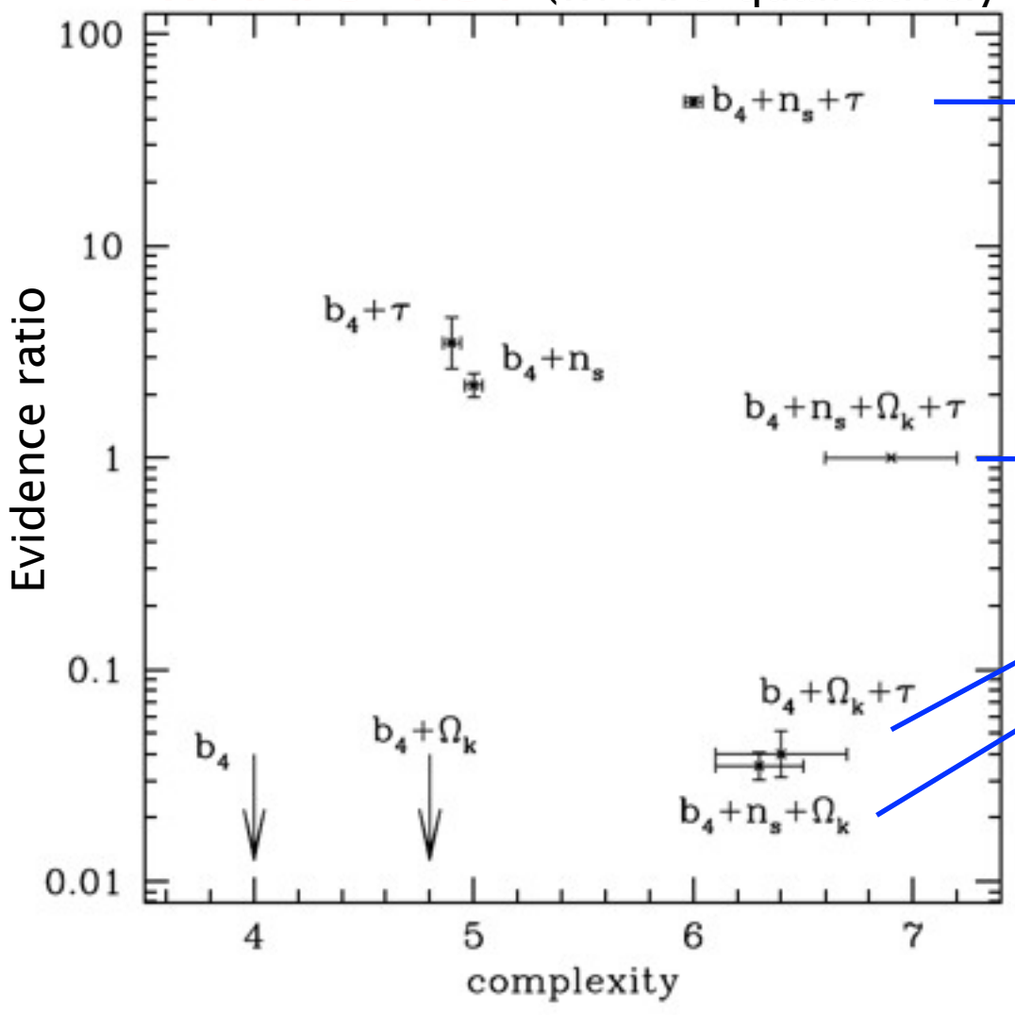
INSUFFICIENT DATA
Max supported complexity ¼ 4



How many parameters does the CMB need?

Imperial College London

WMAP3+HST (WMAP5 qualitatively the same)



$b_4+n_s+\tau$ *measured & favoured*

Ω_κ *measured & unnecessary*

7 params measured

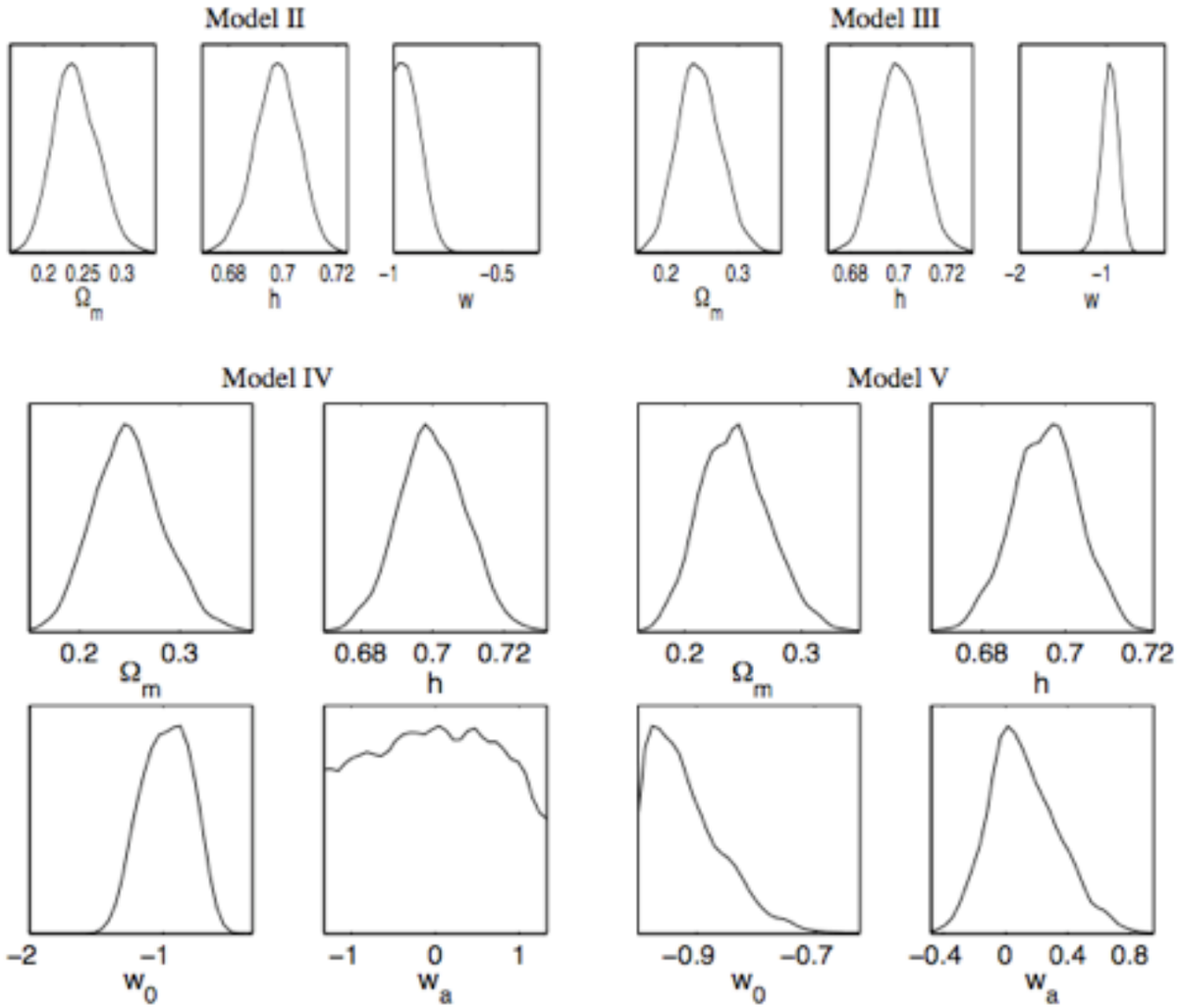
only 6 sufficient

Bayesian Model-averaging

Imperial College London

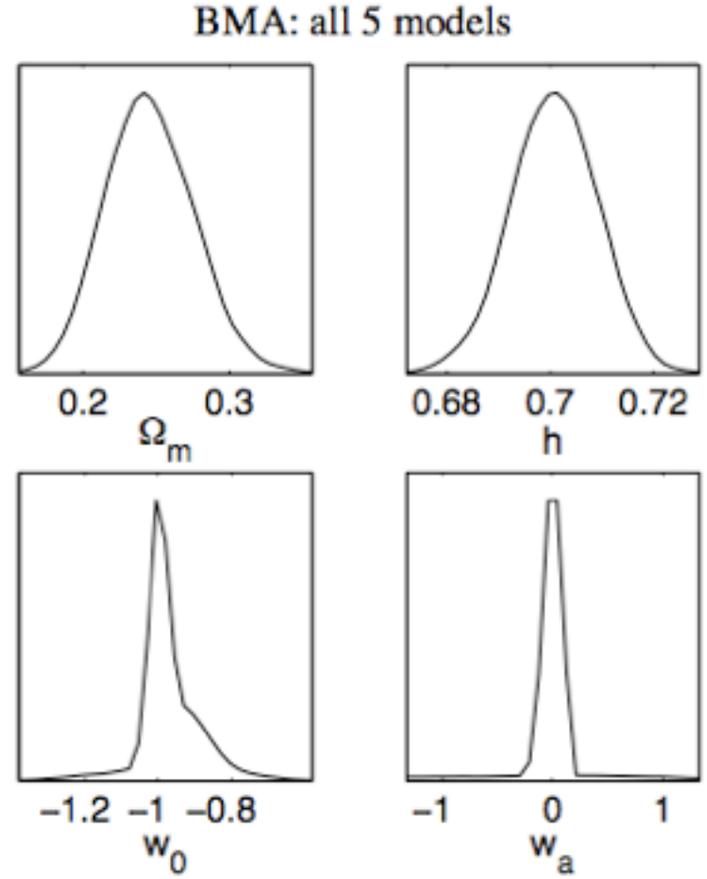
$$P(\theta|d) = \sum_i P(\theta|d,M_i)P(M_i|d)$$

An application to dark energy:



Liddle et al (2007)

Model averaged inferences



Key points

Imperial College London

- Bayesian model comparison extends parameter inference to the space of models
- The Bayesian evidence (model likelihood) represents the change in the degree of belief in the model after we have seen the data
- Models are rewarded for their predictivity (automatic Occam's razor)
- Prior specification is for model comparison a key ingredient of the model building step. If the prior cannot be meaningfully set, then the physics in the model is probably not good enough.
- Bayesian model complexity can help (together with the Bayesian evidence) in assessing model performance.

Prediction and optimization

The Bayesian perspective

Imperial College London

- In the Bayesian framework, we can use present-day knowledge to produce probabilistic forecasts for the outcome of a future measurement

- This is **not** limited to assuming a model/parameter value to be true and to determine future errors

- Many questions of interest today are of model comparison: e.g.

  - is dark energy Lambda or modified gravity?

  - is dark energy evolving with time?

  - Is the Universe flat or not?

  - Is the spectrum of perturbations scale invariant or not?

Predictions for future observations

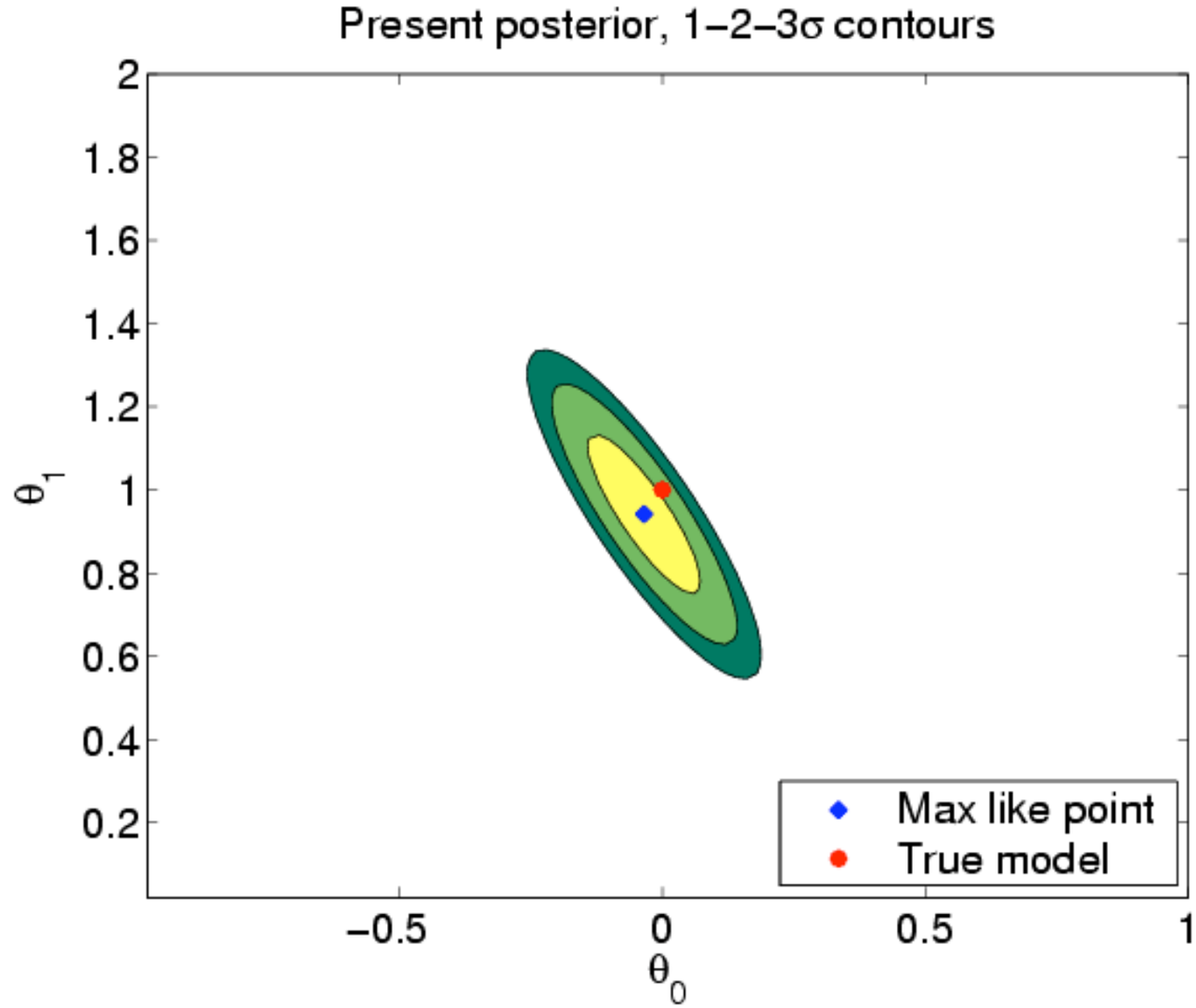
- Toy model: the linear Gaussian model (see Exercices 7-9)

  $y = \theta_0 + x\theta_1$

  $y - Fx = \varepsilon$

- Gaussian noise on ε

- True values: $(\theta_0, \theta_1) = (0,1)$



The predictive distribution

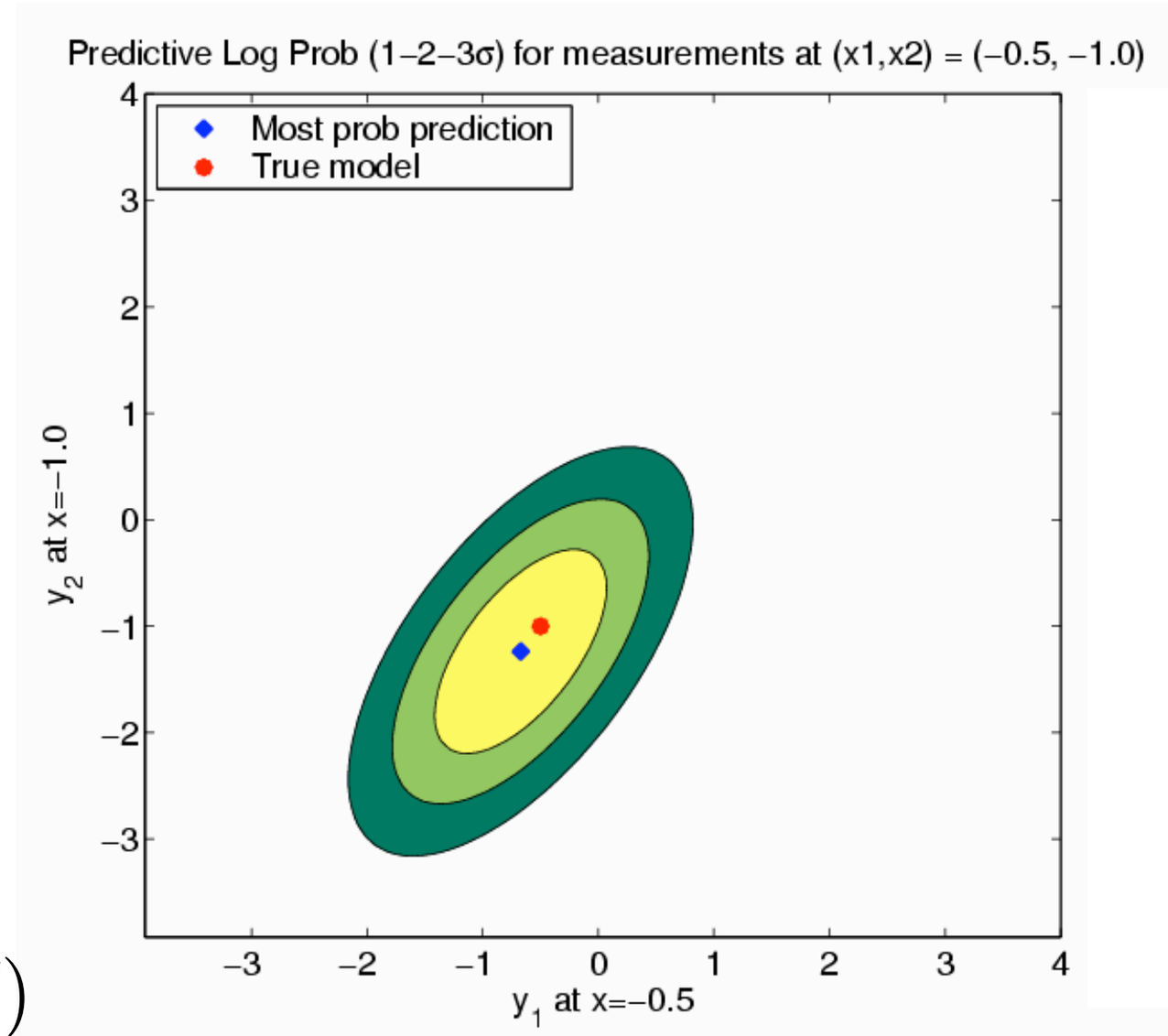
- Use present knowledge (**and uncertainty!)** to predict what a future measurement will find (with corresponding probability)

- True values: $(\theta_0, \theta_1) = (0,1)$

- Present-day data: d

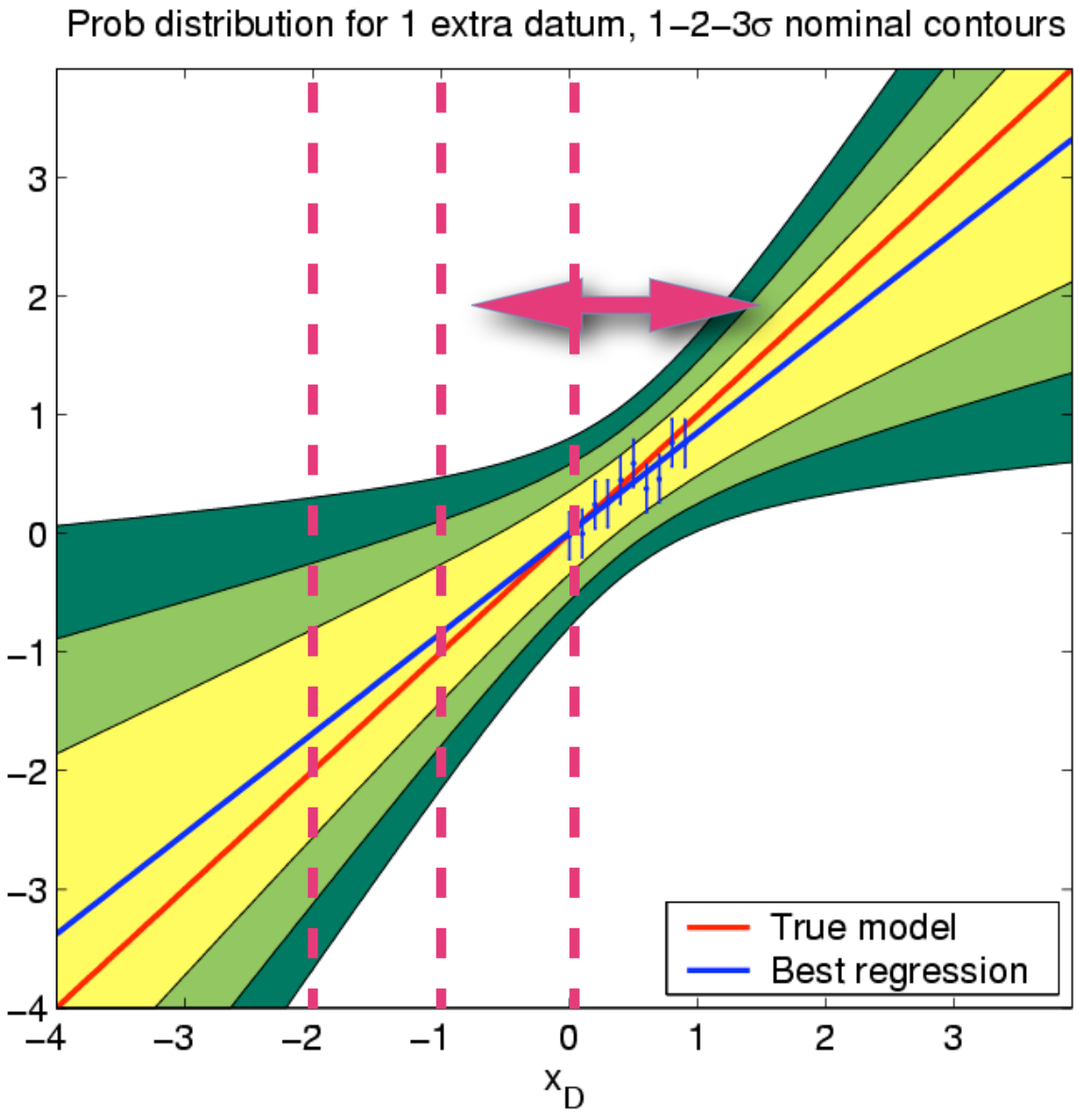
- Future data: D

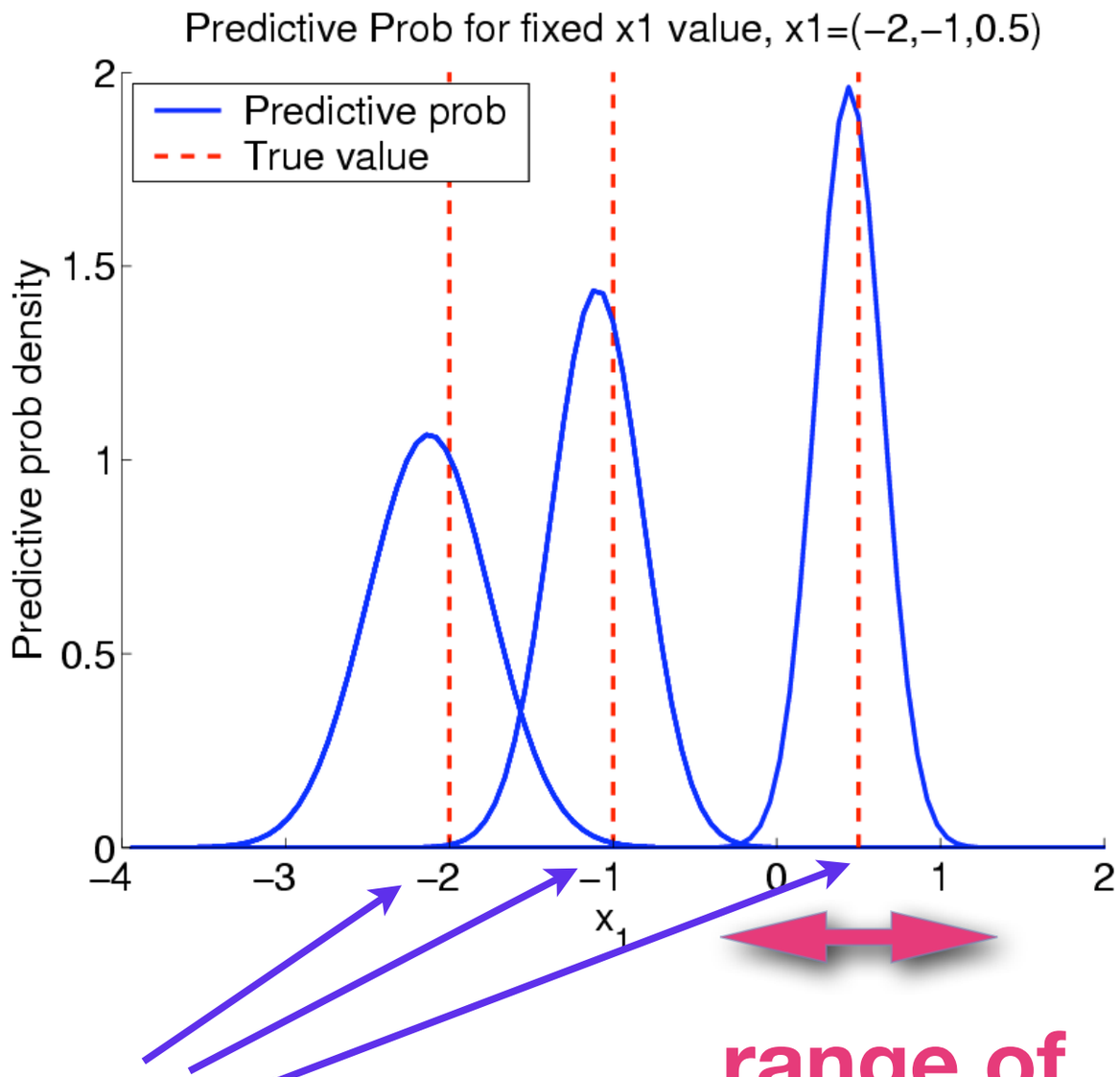
$$P(D|d) = \int d\theta P(D|\theta) P(\theta|d)$$



Predictive probability = future likelihood weighted by present posterior

Predictive distribution





Possible locations of future measurements

range of present-day data

Extending the power of forecasts

Imperial College London

- Thanks to predictive probabilities we can increase the scope and power of forecasts:

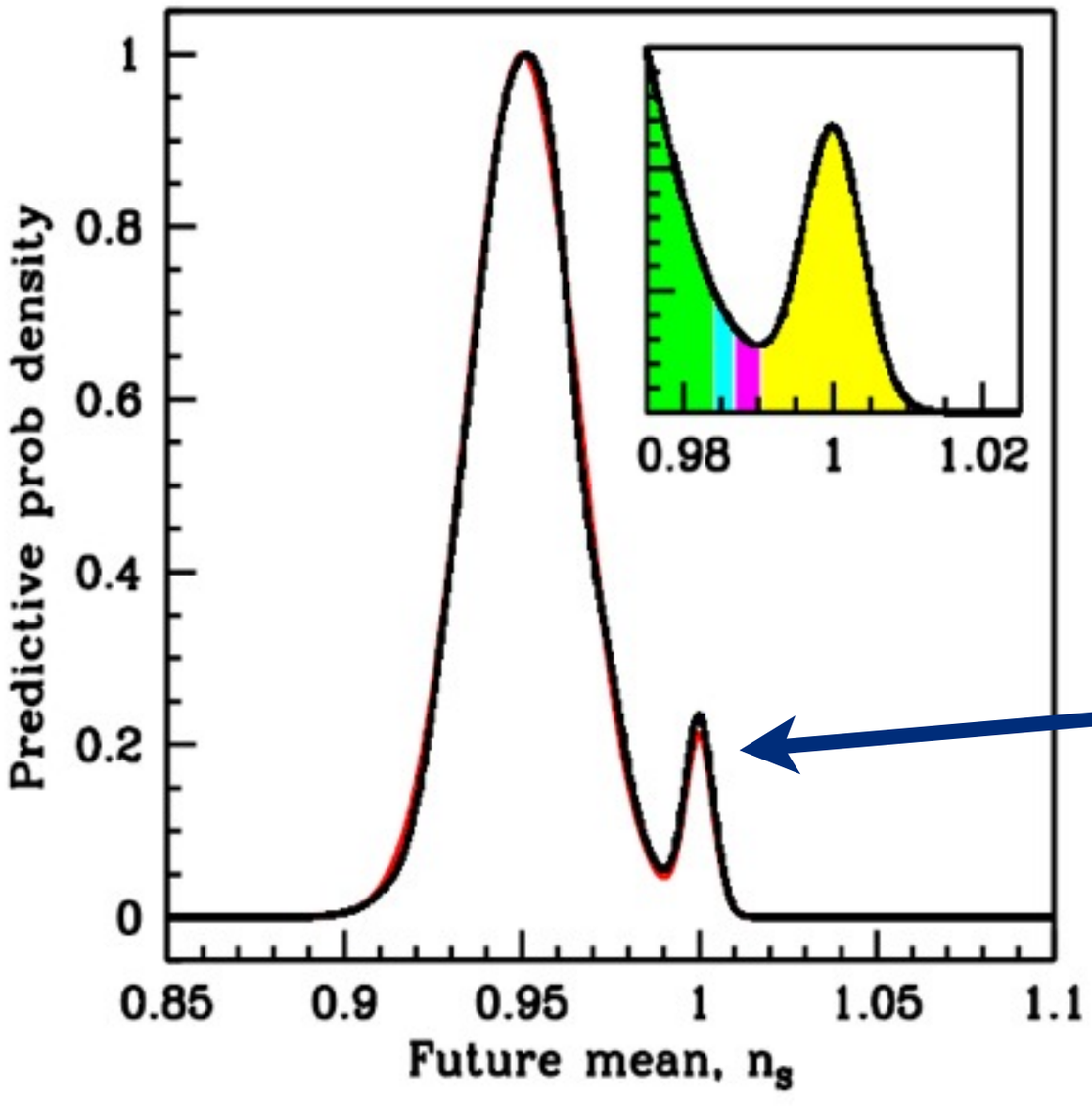
- **Level 0:** assume a model M and a fiducial value for the parameters, θ^*
  produce a forecast for the errors that a future experiment will find *if M and* θ^* *are the correct choices*

- **Level 1:** average over current parameter uncertainty within M

- **Level 2:** average over current model uncertainty: replace M by M_1, M_2,...

Predictive posterior odds distribution

Bayes factor forecast for Planck



n_s=1 vs 0.8<n_s<1.2

P(lnB < -5) = 0.93
P(-5<lnB<0) = 0.01
P(lnB > 0) = 0.06

Model uncertainty
P(n_s=1|WMAP3+) = 0.05

Trotta (2008), Parkinson et al (2006), Pahud et al (2006)

Experiment design

Utility and optimization

Imperial College London

- The optimization problem is fully specified once we define a **utility function U** depending on the outcome *e* of a future observation (e.g., scientific return). We write for the utility U(e, o, θ), where *o* is the current experiment and θ are the true values of the parameters of interest

- We can then evaluate the **expected utility:**

$$\mathcal{E}[U|e,o] = \int d\theta U(\theta, e, o) P(\theta|o)$$

Example: an astronomer measures y = θ x (with Gaussian noise) at a few points 0 < x < 1. She then has a choice between building 2 equally expensive instruments to perform a new measurement:

1. Instrument (e) is as accurate as today's experiments but extends to much larger values of x (to a maximum x_{max})
2. Instrument (a) is much more accurate but it is built in such a way as has to have a "sweet spot" at a certain value of y, call it y*, and much less accurate elsewhere

Which instrument should she go for?

Answer

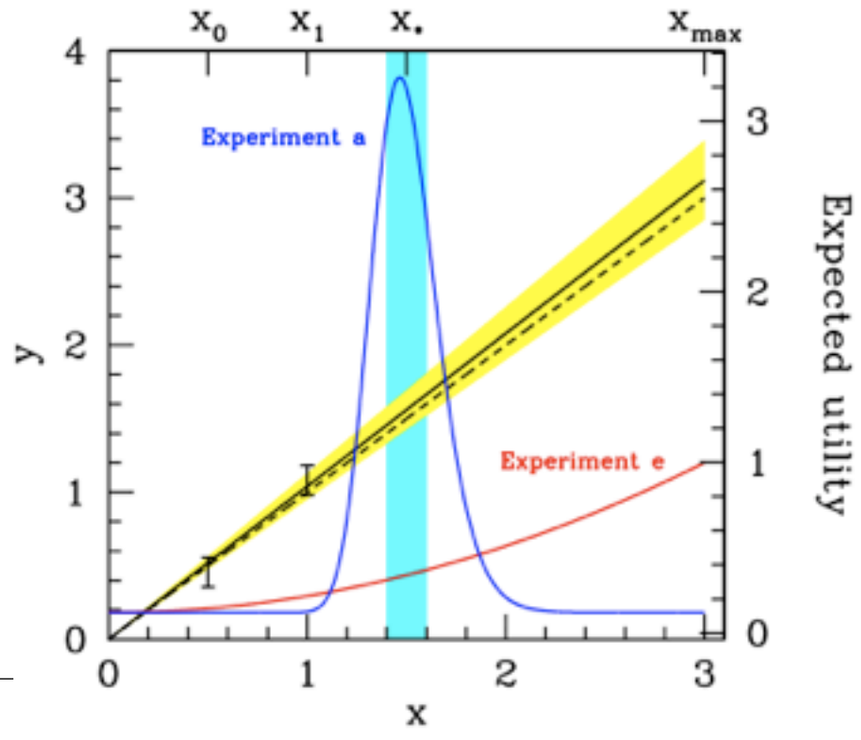
Imperial College London

- The answer depends on how good her current knowledge is - i.e. is the current uncertainty on θ* small enough to allow her to target accurately enough x=x* so that she can get to the "sweet spot" y*= θ*x*?
(try it out for yourself! *Hint: use for the utility the inverse variance of the future posterior on θ and assume for the noise levels of experiment (a) the toy model:*

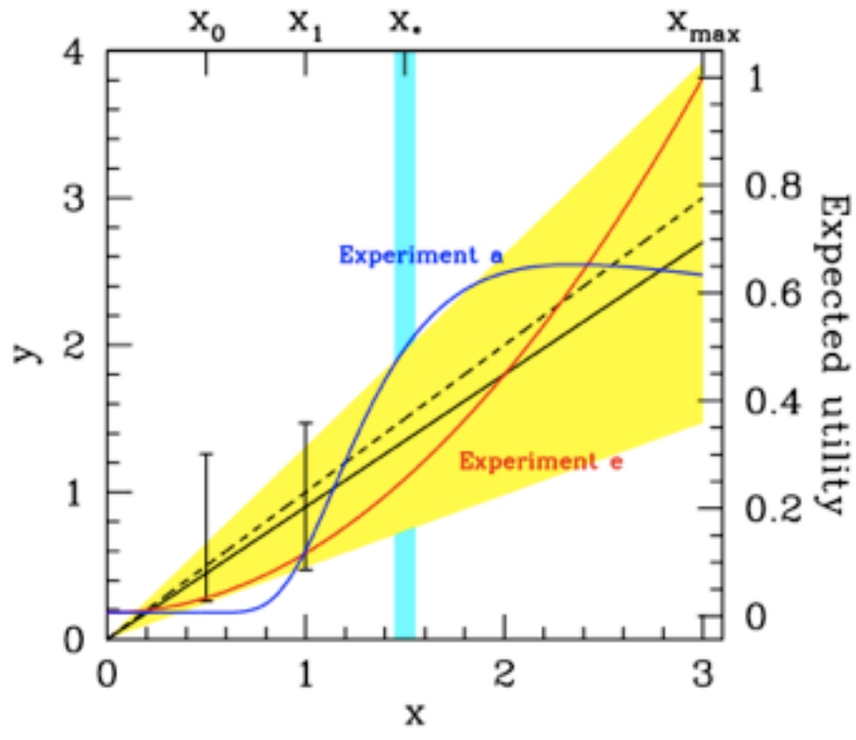
$$\tau_a^2 = \tau_*^2 \exp\left(\frac{(y-y_\star)^2}{2\Delta^2}\right)$$

where y is the location of the sweet spot and Δ is the width of the sweet spot)*

Small uncertainty



Large uncertainty



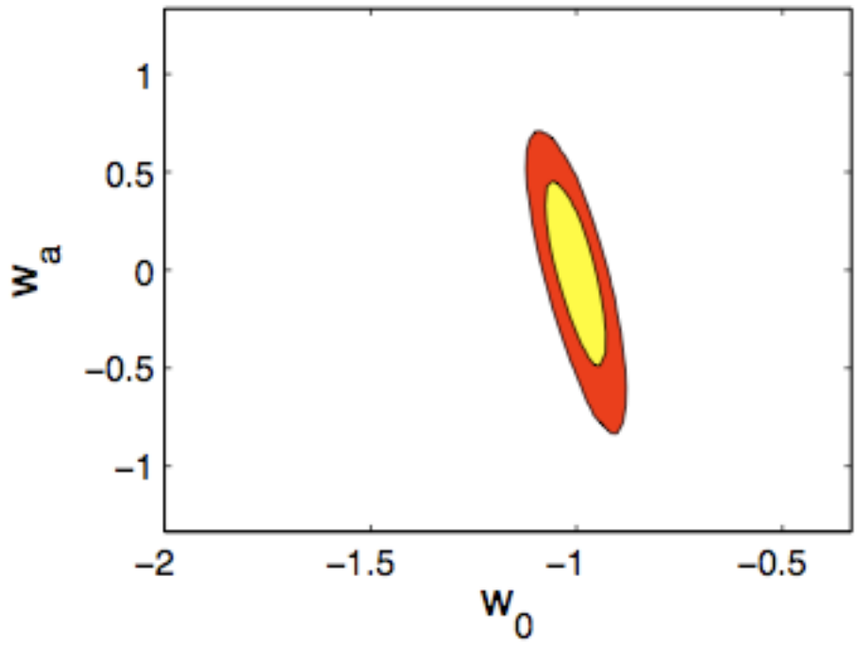
Trotta et al, in Bayesian methods in cosmology (CUP, 2010)

Making predictions: Dark Energy

Imperial College London

A model comparison question: is dark energy Lambda, i.e. $(w_0, w_a) = (-1, 0)$?
How well will the future probe SNAP be able to answer this?

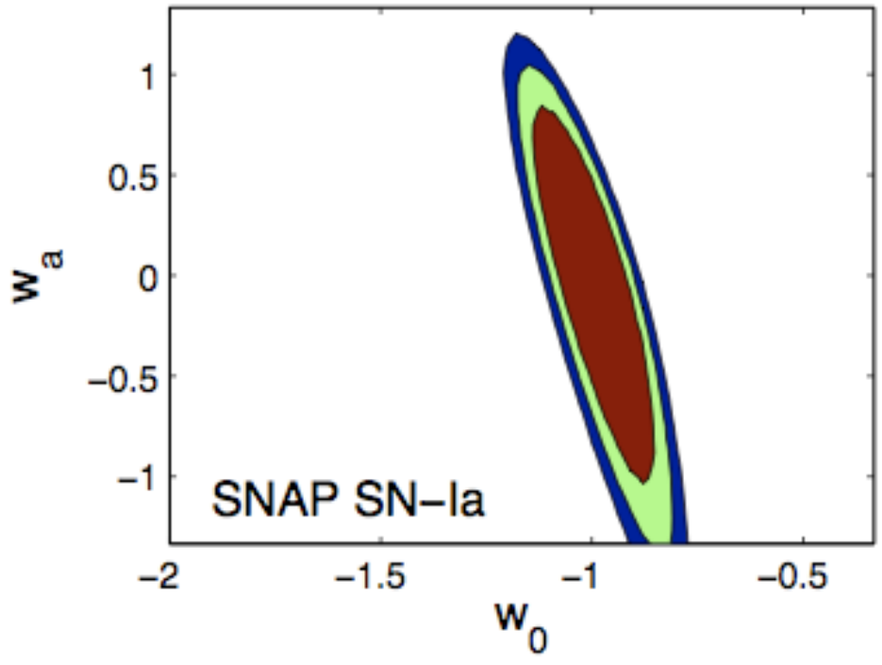
Fisher Matrix



Mukherjee et al (2006)

Simulates from LCDM
Assumes LCDM is true
Ellipse not invariant when changing model assumptios

Bayesian evidence



Simulate from all DE models
Assess “model confusion”
Allows to discriminate against LCDM

Key points

Imperial College London

- Predictive distributions incorporate present uncertainty in forecasts for the future scientific return of an experiment

- Experiment optimization requires the specification of an utility function. The “best” experiment is the one that maximises the expected utility.