Bayesian Model Comparison

Roberto Trotta - www.robertotrotta.com

ICIC Data Analysis Workshop, Sept 2013

ICIC
Imperial Centre
for Inference & Cosmology

Imperial College
London

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

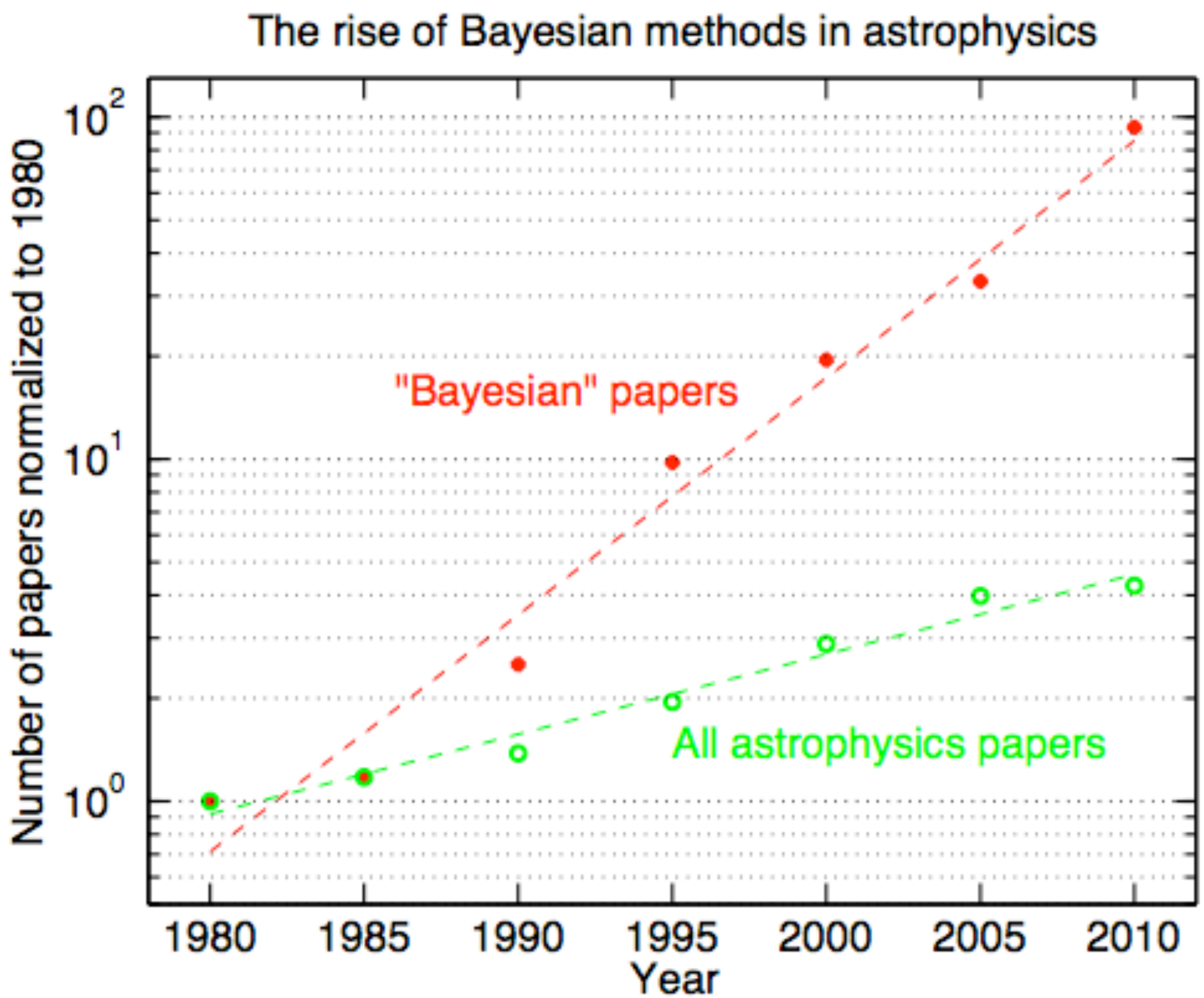
- **Important:** A 2-sigma result does not wrongly reject the null hypothesis 5% of the time: **at least 29% of 2-sigma results are wrong!**
  - Take an equal mixture of H_0, H_1
  - Simulate data, perform hypothesis testing for H_0
  - Select results rejecting H_0 at (or within a small range from) 1-α CL (this is the prescription by Fisher)
  - 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 reading:
Sellke, Bayarri & Berger, *The American Statistician*, 55, 1 (2001)

Bayesian methods on the rise

Imperial College London



ICIC

Bayes' theorem

Imperial College London

posterior — likelihood — prior — evidence

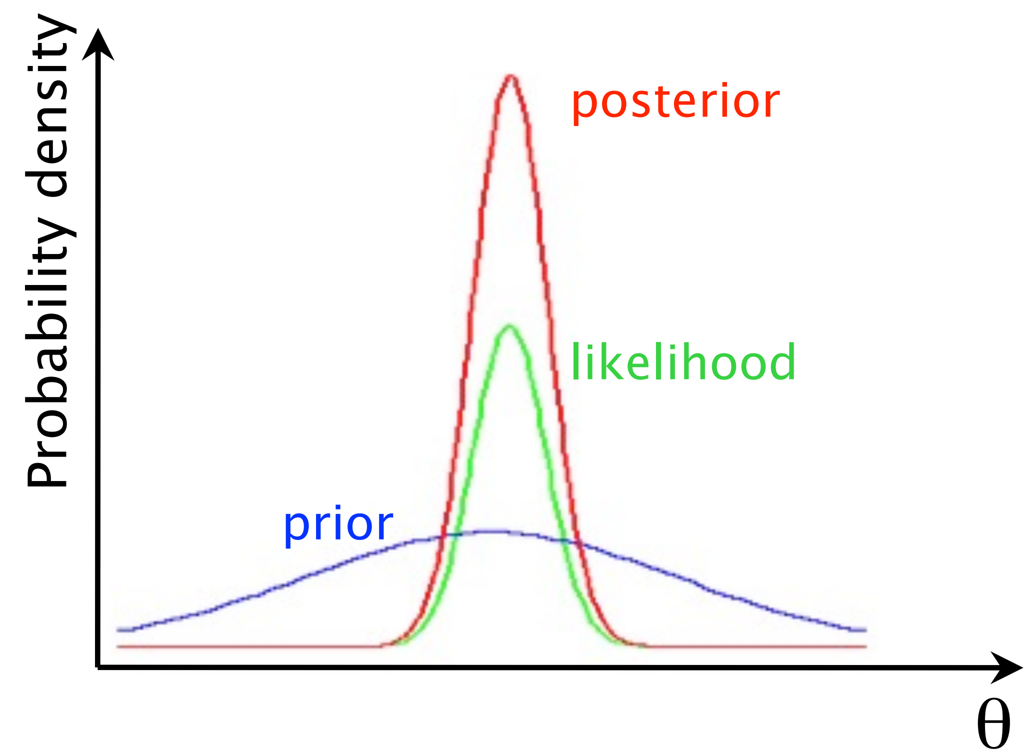
$$P(\theta|d,I) = \frac{P(d|\theta,I)P(\theta|I)}{P(d|I)}$$

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

For parameter inference it is sufficient to consider

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

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



ICIC

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

LEVEL 1
I have selected a model M and prior P(θ|M)

LEVEL 2
Actually, there are several possible models: M_0, M_1,...

LEVEL 3
None of the models is clearly the best

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

Parameter inference
(assumes M is the true model)

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

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

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

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

Level 2: model comparison

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

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

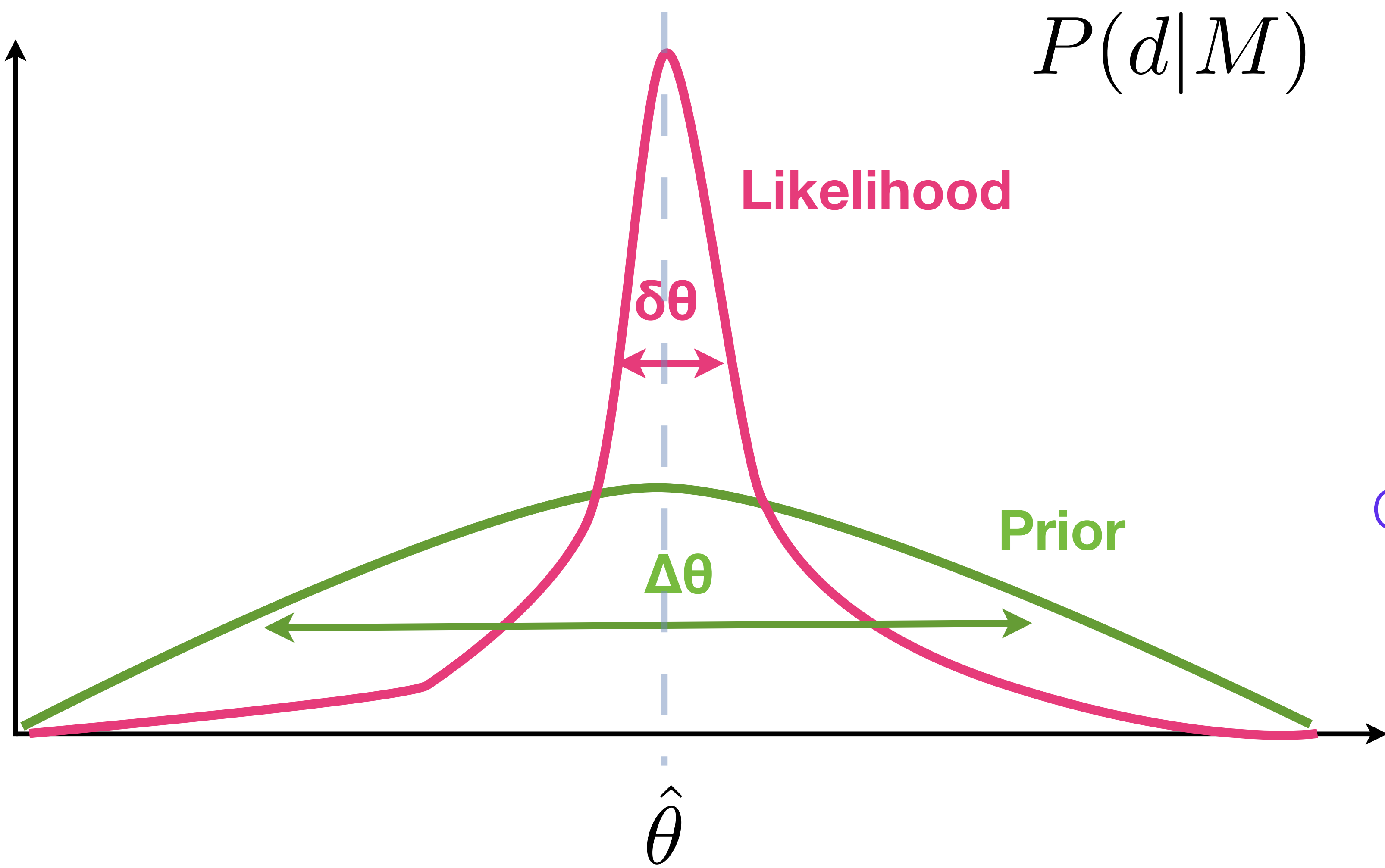
- A (slightly modified) Jeffreys' scale to assess the strength of evidence (**Notice:** this is empirically calibrated!)

$\|\ln B\|$	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



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

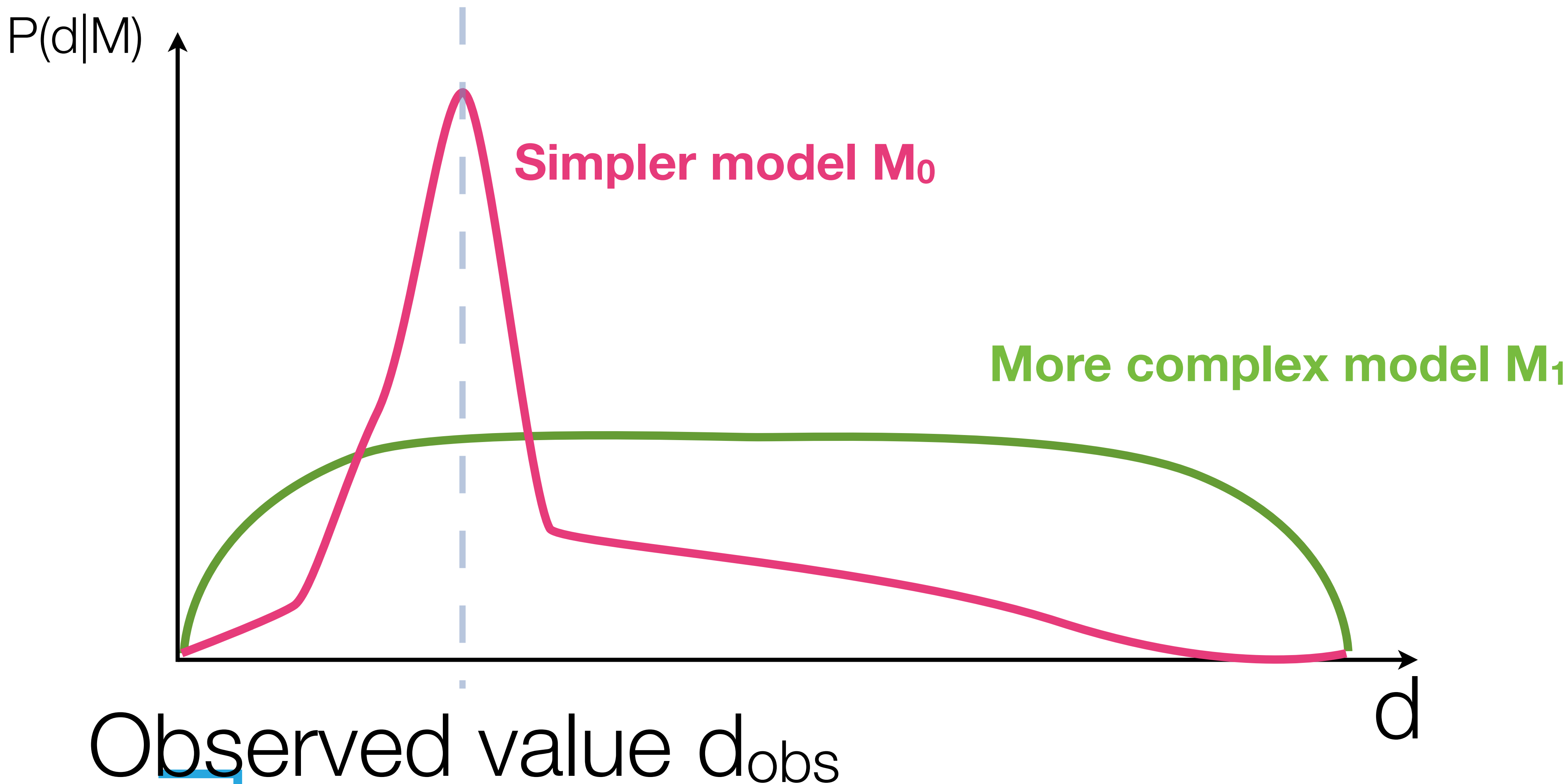
Occam's factor

ICIC

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:

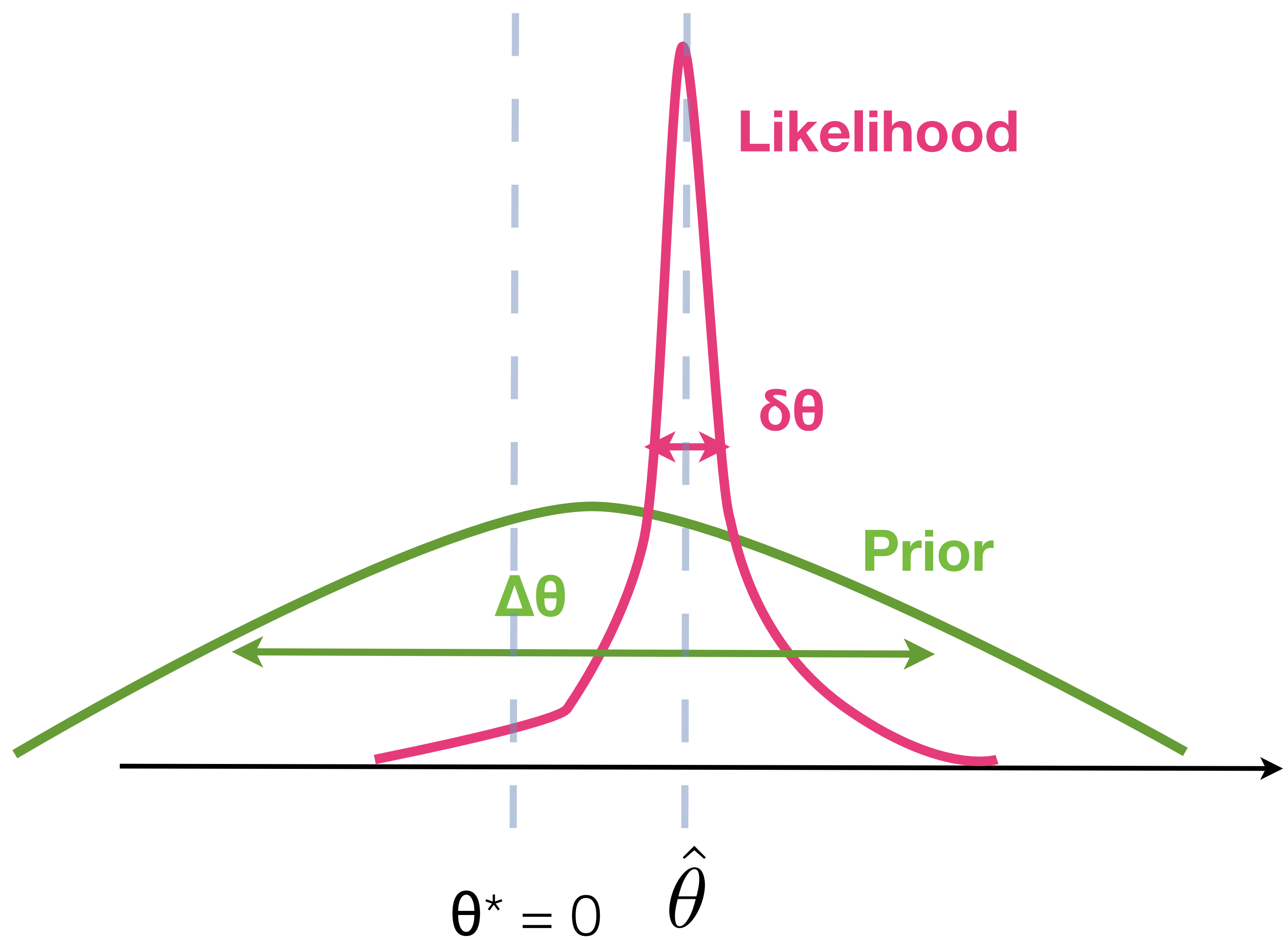


ICIC

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

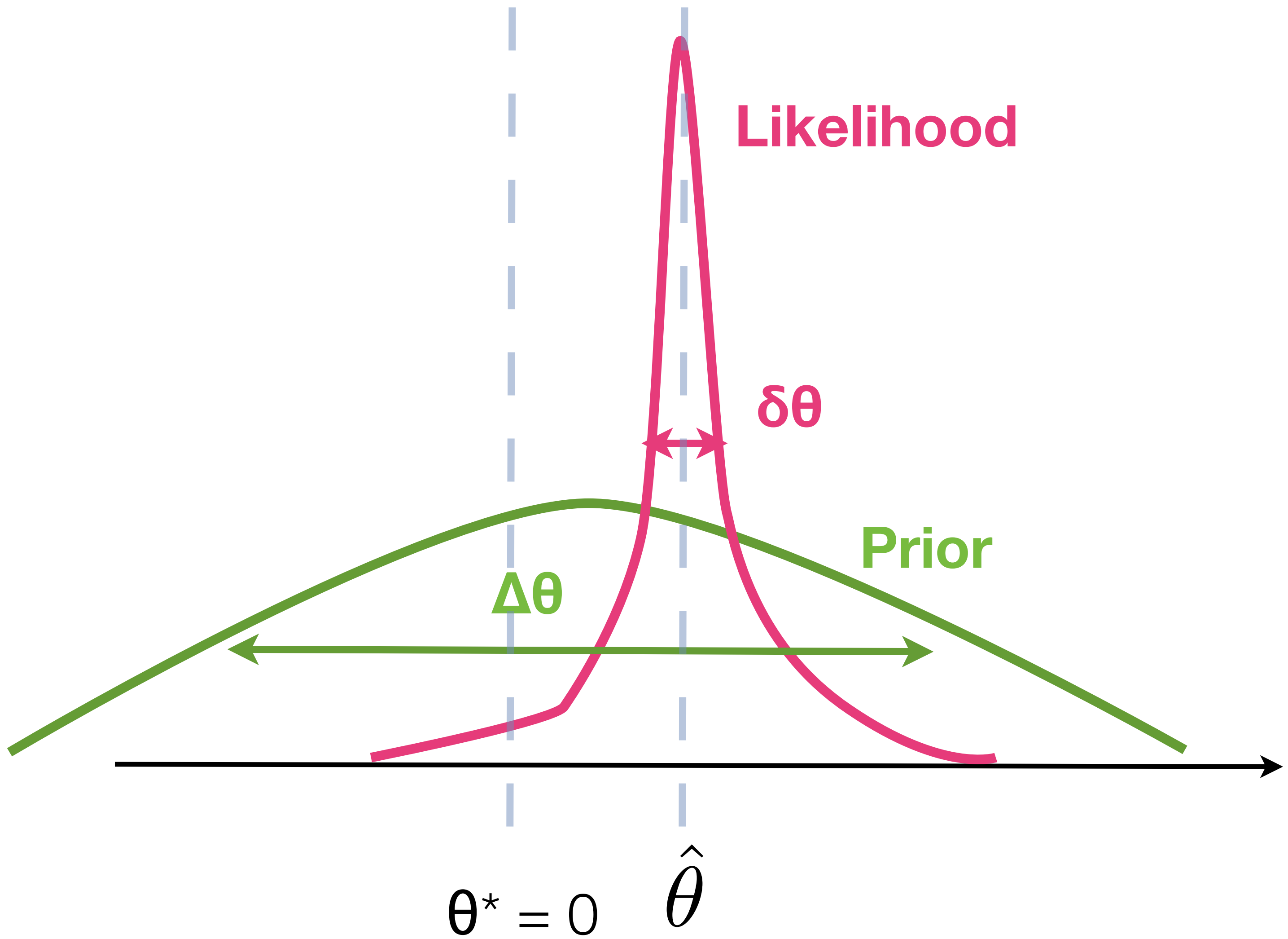
Define: $\lambda \equiv \frac{\hat{\theta} - \theta^\star}{\delta\theta}$

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)



Question 1

- You use Bayesian model comparison to compare:

  Model 0: The coin is fair

  Model 1: The coin is biased, with uniform prior between [0,1] for the probability of heads

- Data: 5 heads out of 10 tosses.

- The Bayes factor will:

  A. Favour model 0

  B. Favour model 1

  C. Favour neither model

Question 2

- You use Bayesian model comparison to compare:
  Model 0: The coin is fair
  Model 1: The coin has two heads
- Data: 9 heads out of 10 tosses.
- The Bayes factor will:

  A. Favour model 0

  B. Favour model 1

  C. Favour neither model

Question 3

- You use Bayesian model comparison to compare:

  Model 0: The coin is fair

  Model 1: The coin is fair, and the height h of the Shard is 326m, with a uniform prior for h in [0,1000] m.

- Data: 5 heads out of 10 tosses.

- The Bayes factor will:

  A. Favour model 0

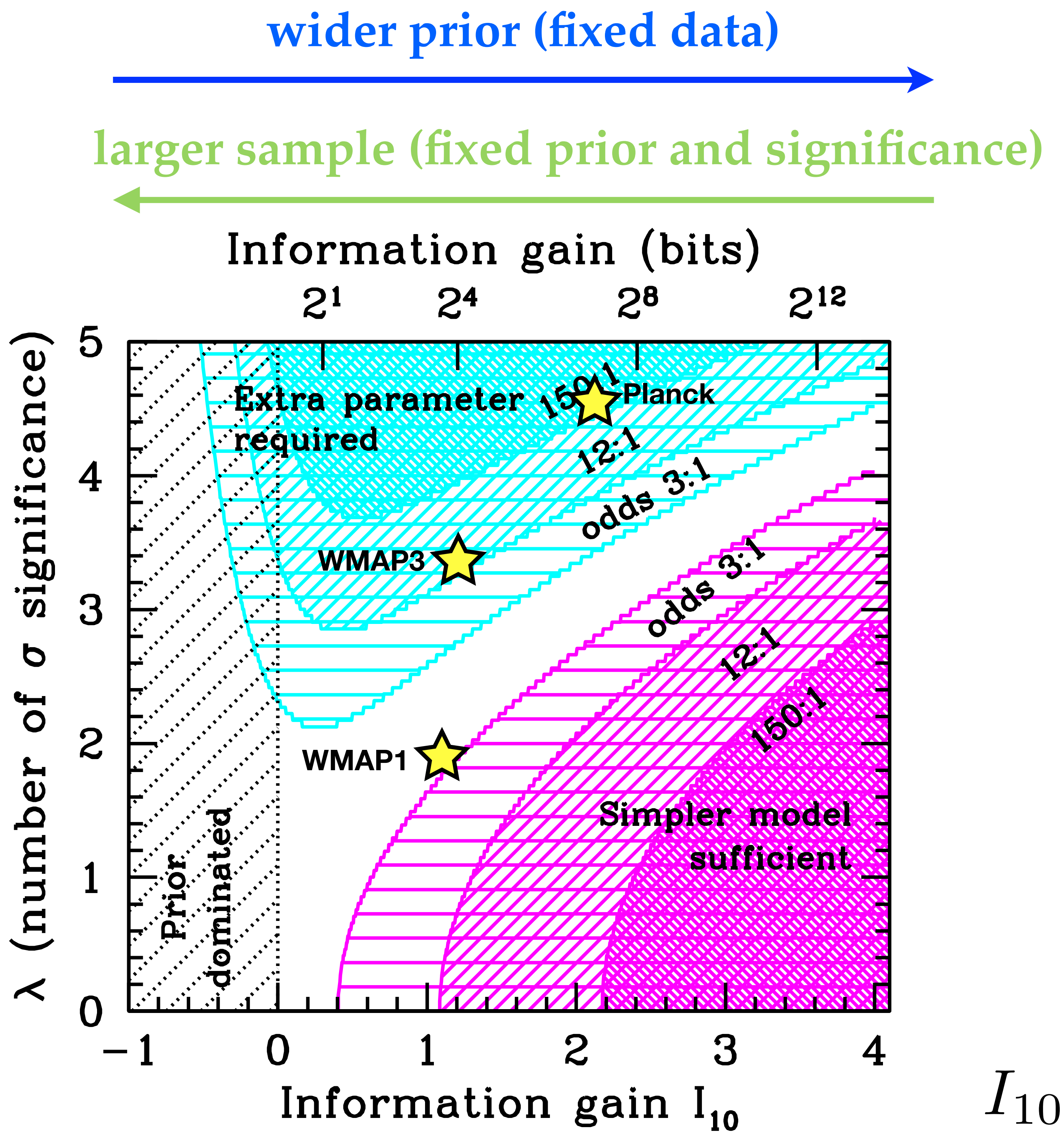
  B. Favour model 1

  C. Favour neither model

The rough guide to model comparison

Imperial College London

Trotta (2008)



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

ICIC

Roberto Trotta

Thursday, 12 September 13

Information criteria

- Several information criteria exist for approximate model comparison
  k = number of fitted parameters
  N = number of data points,
  -2 ln(L_{max}) = best-fit chi-squared
- **Akaike Information Criterium (AIC):** $$\mathrm{AIC} \equiv -2\ln\mathcal{L}_{\mathrm{max}} + 2k$$
- **Bayesian Information Criterium (BIC):** $$\mathrm{BIC} \equiv -2\ln\mathcal{L}_{\mathrm{max}} + k\ln N$$
- **Deviance Information Criterium (DIC):** $$\mathrm{DIC} \equiv -2\widehat{D_{\mathrm{KL}}} + 2\mathcal{C}_b.$$

Notes on information criteria

- The best model is the one which minimizes the AIC/BIC/DIC
- **Warning:** AIC and BIC penalize models differently as a function of the number of data points N.
  For N>7 BIC has a more strong penalty for models with a larger number of free parameters k.
- BIC is an approximation to the full Bayesian evidence with a default Gaussian prior equivalent to 1/N-th of the data in the large N limit.
- DIC takes into account whether parameters are measured or not (via the Bayesian complexity, see later).
- When possible, computation of the Bayesian evidence is preferable (with explicit prior specification).

Computing the evidence

Imperial College London

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

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

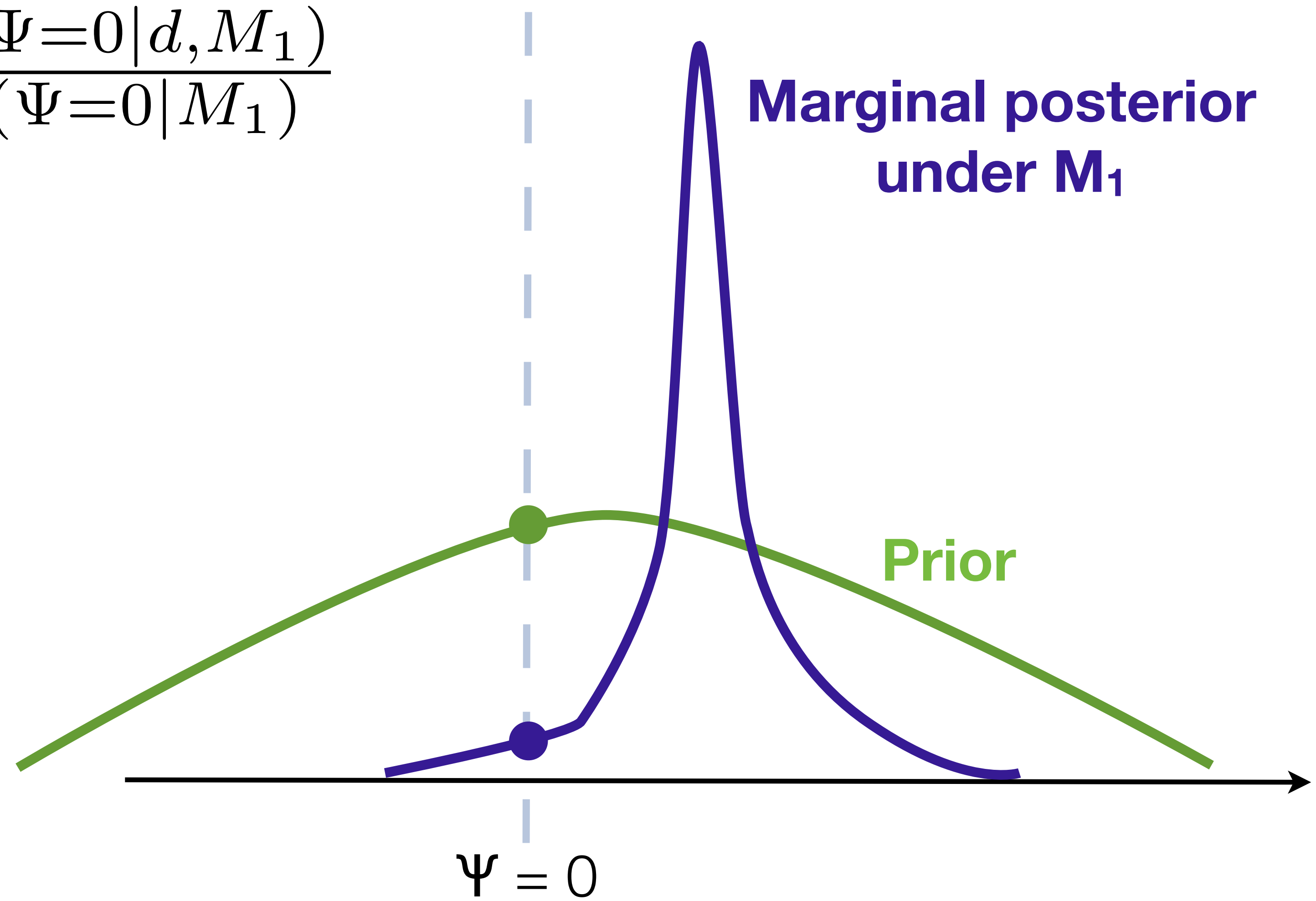
- Usually computational demanding: multi-dimensional integral!
- Several techniques available:
  - Brute force: **thermodynamic integration**
  - **Laplace approximation:** approximate the likelihood to second order around maximum gives Gaussian integrals (for normal prior). Can be inaccurate.
  - **Savage-Dickey 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. Ψ =0) and separable priors (i.e. the prior $P(\theta,\Psi|M_1)$ is uncorrelated with $P(\theta|M_0)$)
- Result:
- **Advantages**:
  - analytical
  - often accurate
  - clarifies the role of prior
  - does not rely on Gaussianity

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

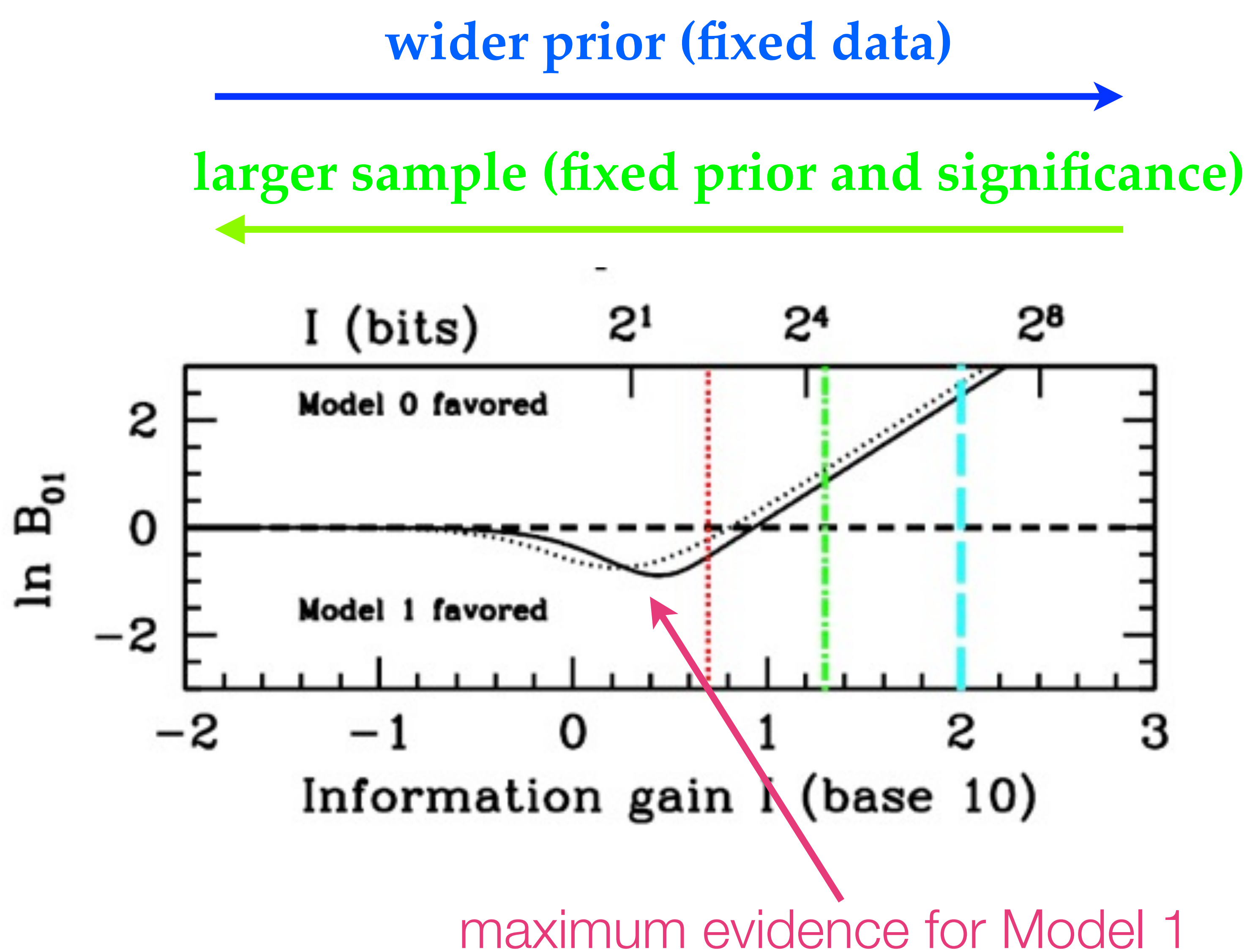


ICIC

“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:



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”

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

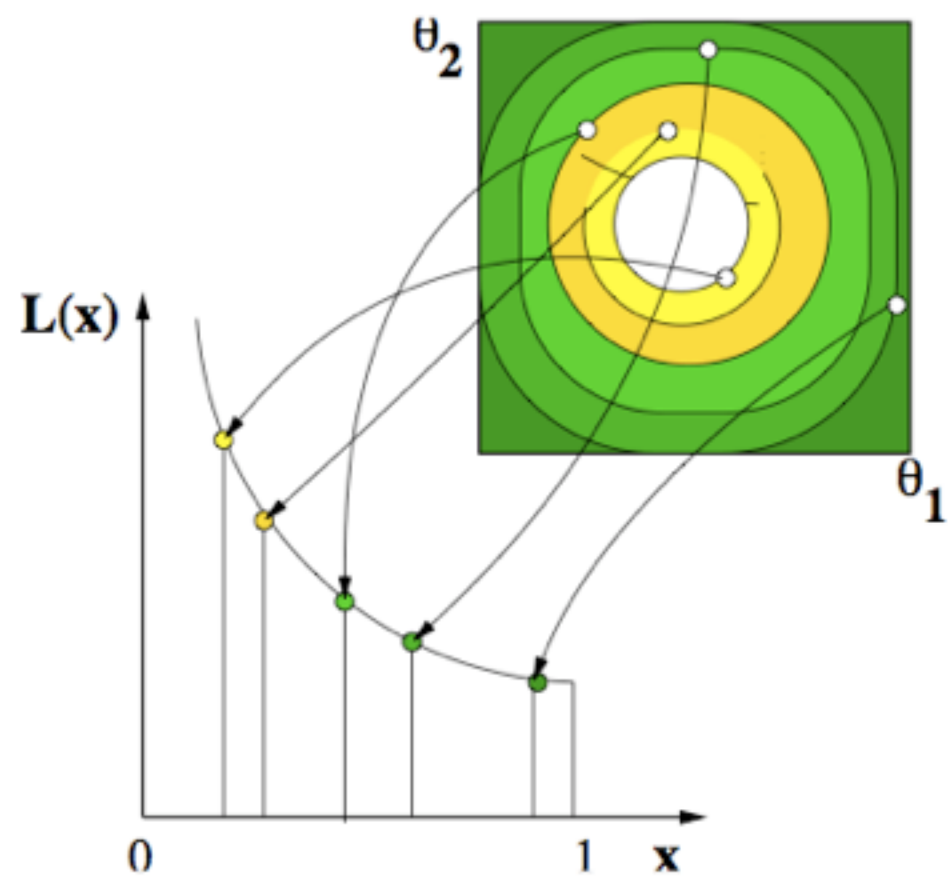
p–value	$\bar{B}$	$\ln \bar{B}$	sigma	category
0.05	2.5	0.9	2.0	
0.04	2.9	1.0	2.1	'weak' at best
0.01	8.0	2.1	2.6	
0.006	12	2.5	2.7	'moderate' at best
0.003	21	3.0	3.0	
0.001	53	4.0	3.3	
0.0003	150	5.0	3.6	'strong' at best
6×10^{-7}	43000	11	5.0	

Rule of thumb:

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

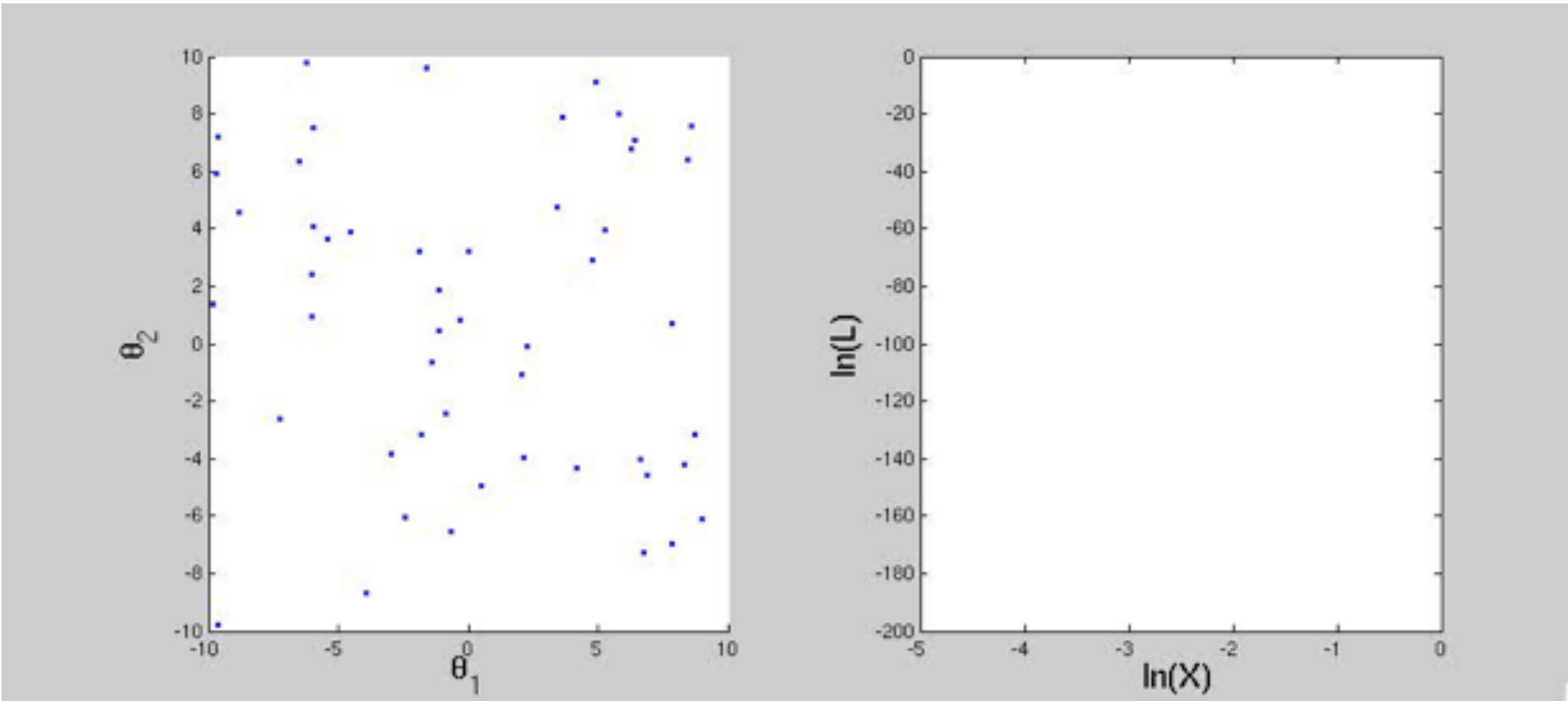
Nested sampling

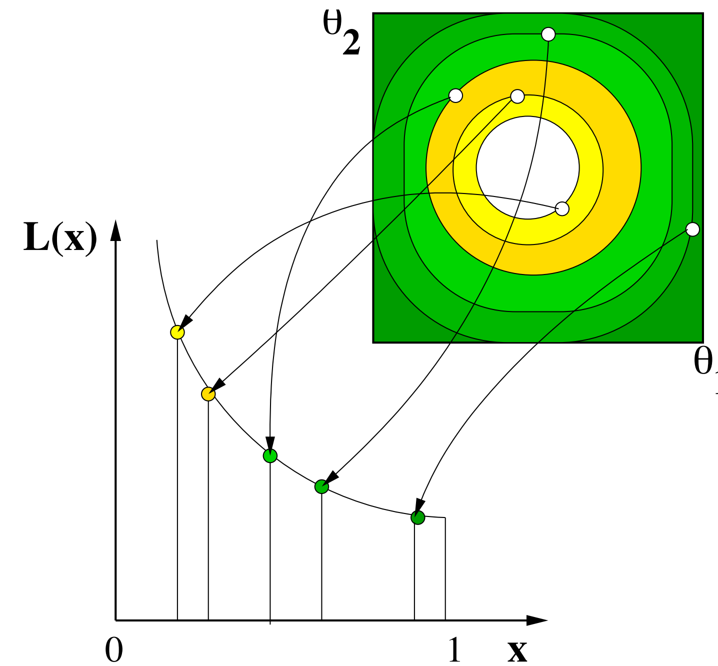
- Perhaps **the** method to compute the evidence
- At the same time, it delivers samples from the posterior: it is a highly efficient sampler! (much better than MCMC in tricky situations)
- Invented by John Skilling in 2005: the gist is to convert a *n*-dimensional integral in a 1D integral that can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



Liddle et al (2006)

Nested sampling





(animation courtesy of David Parkinson)

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

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

$$P(d) = \int d\theta L(\theta) P(\theta) = \int_0^1 L(X) dX$$

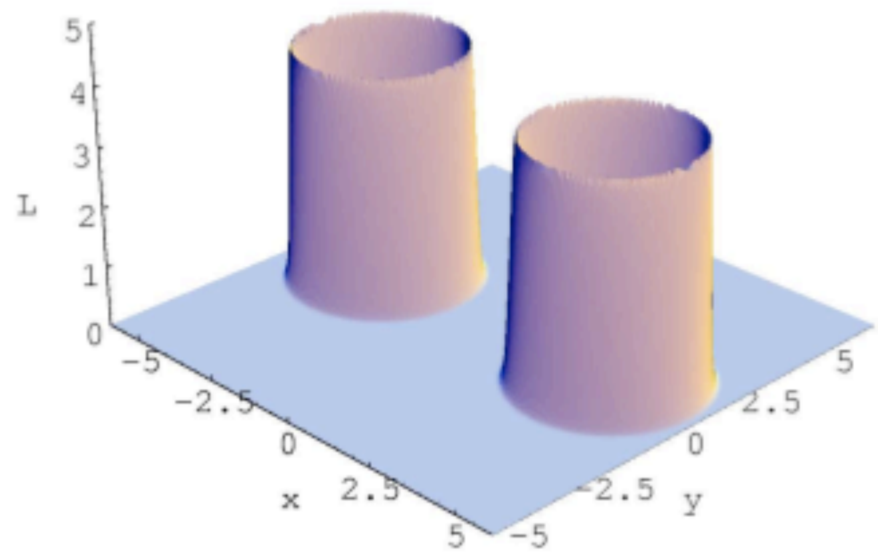
The MultiNest algorithm

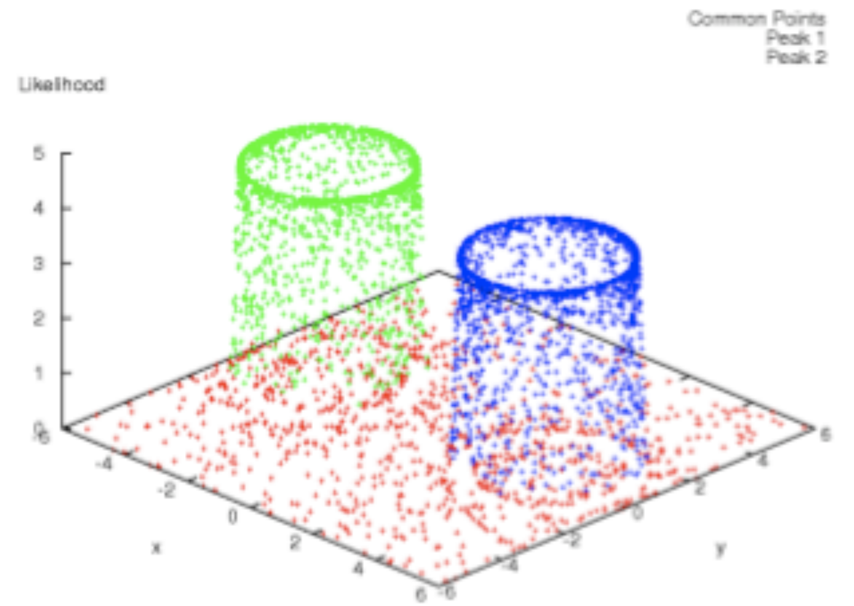
Imperial College London

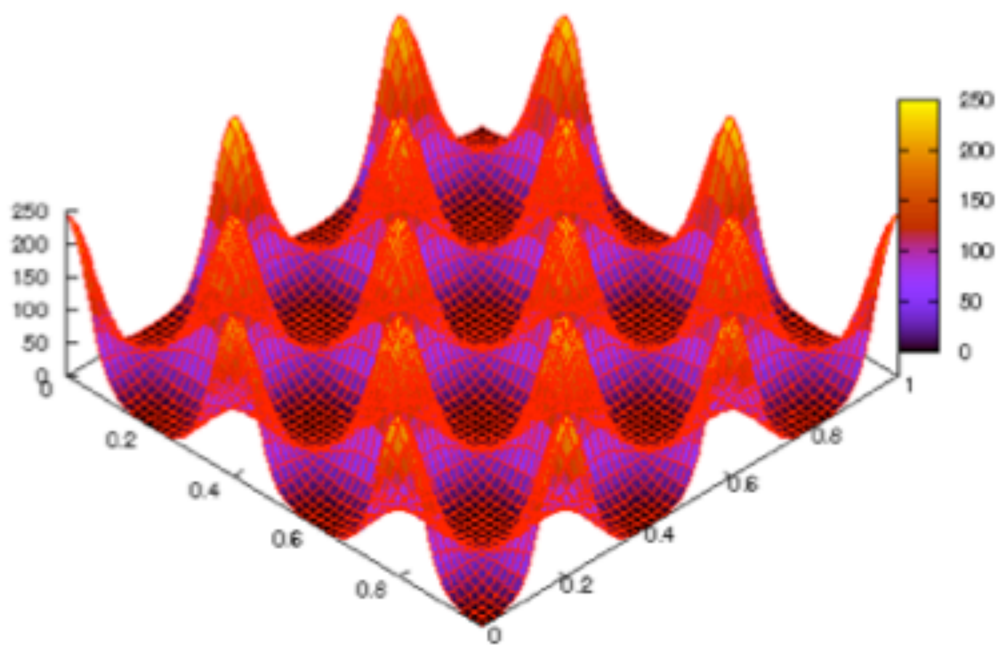
- Feroz & Hobson (2007)

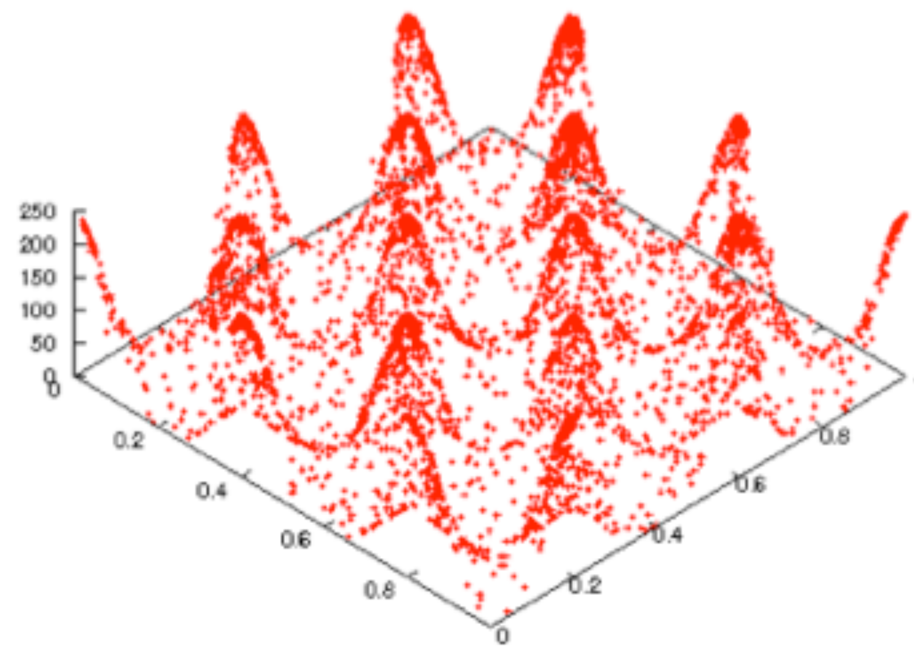
Target

Reconstructed







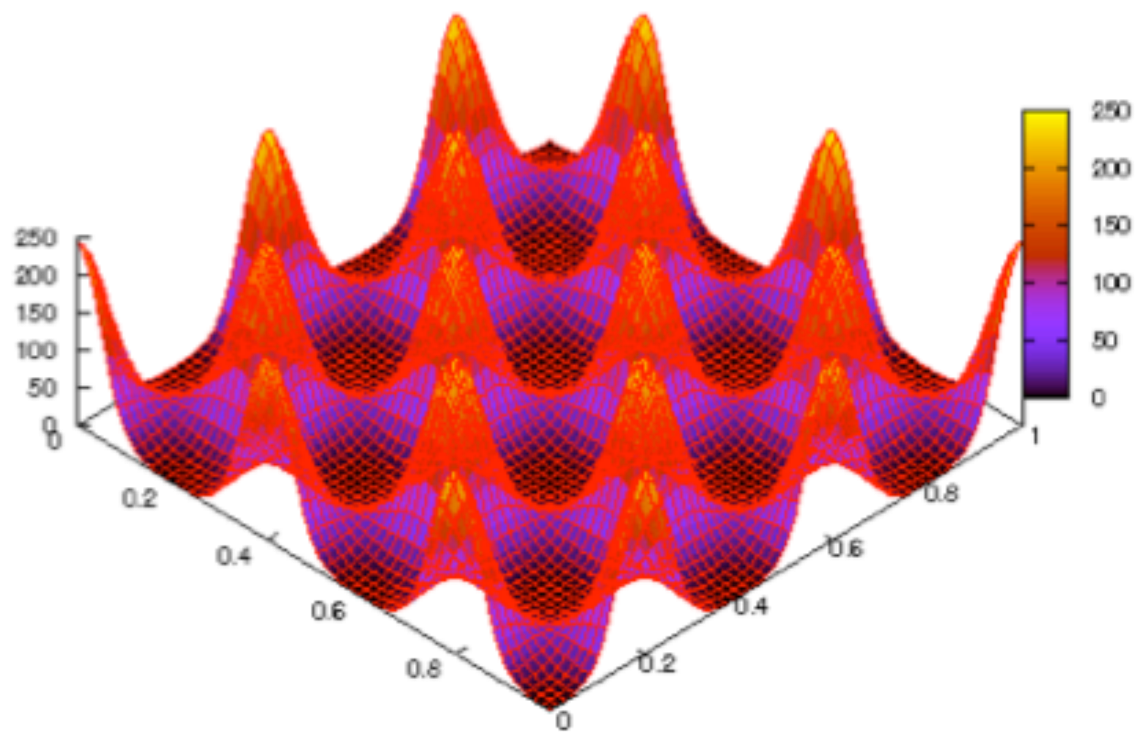


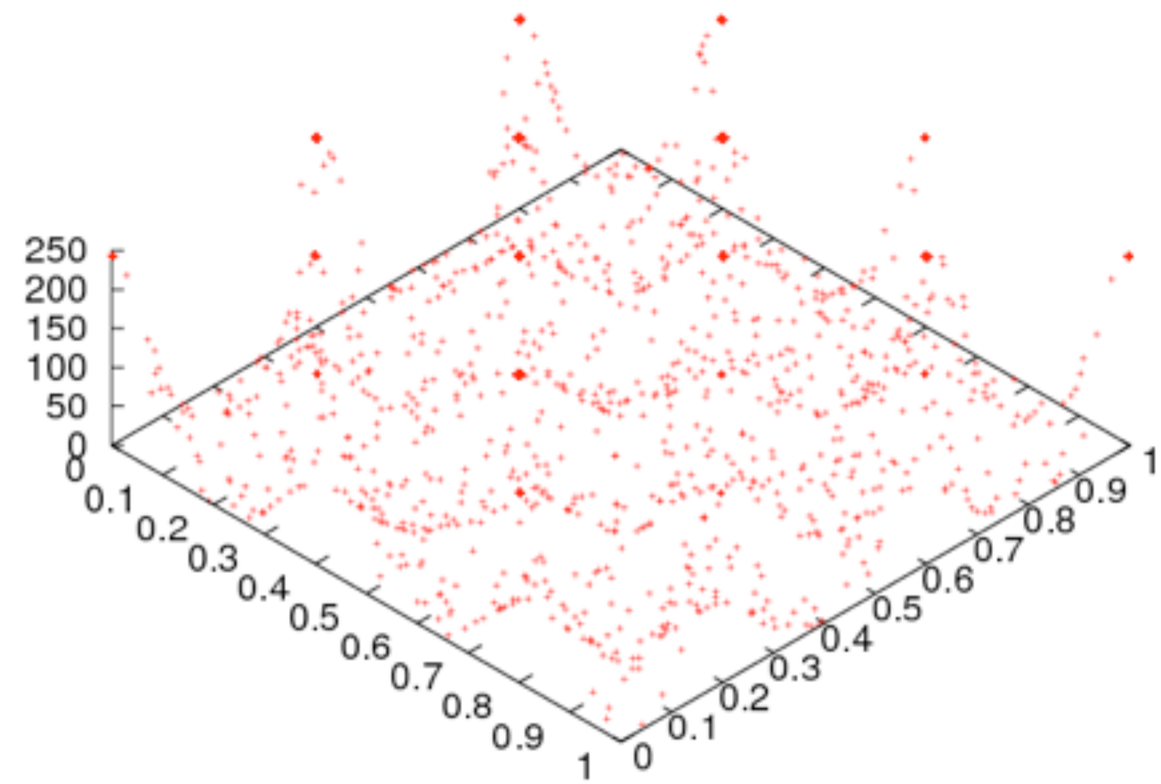
Courtesy Mike Hobson

The egg-box example

Imperial College London

- MultiNest reconstruction of the egg-box posterior:

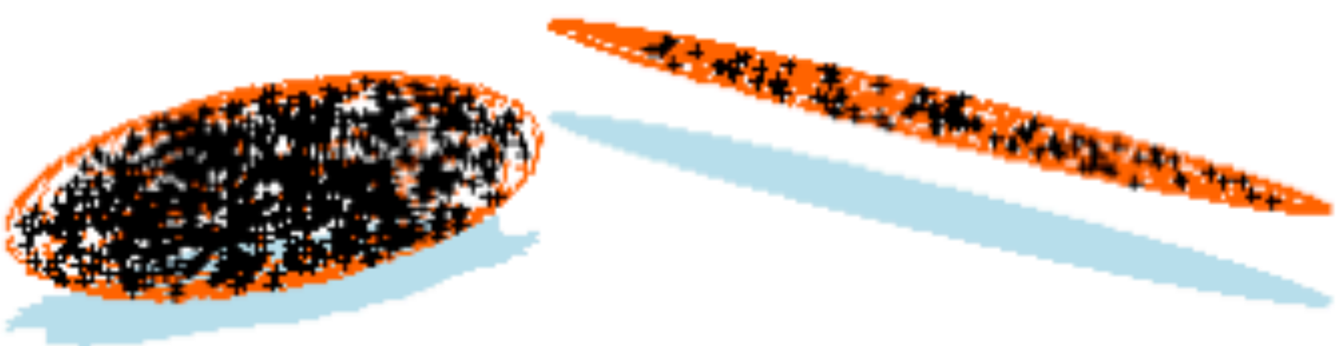


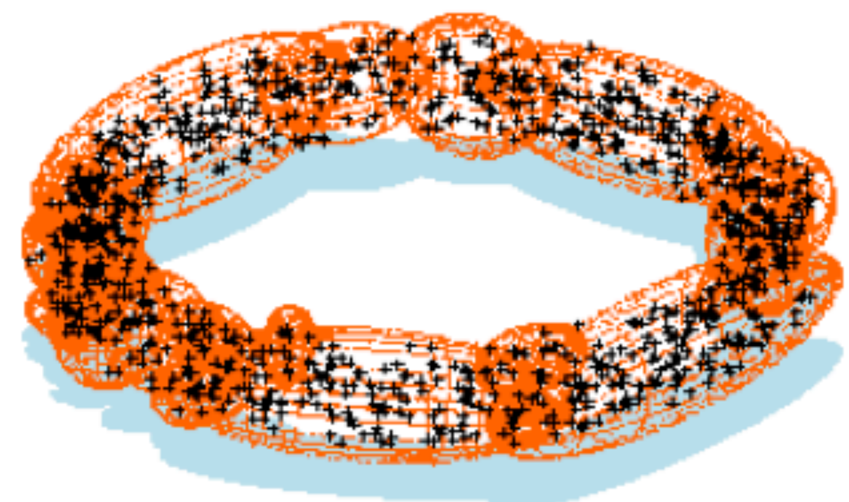


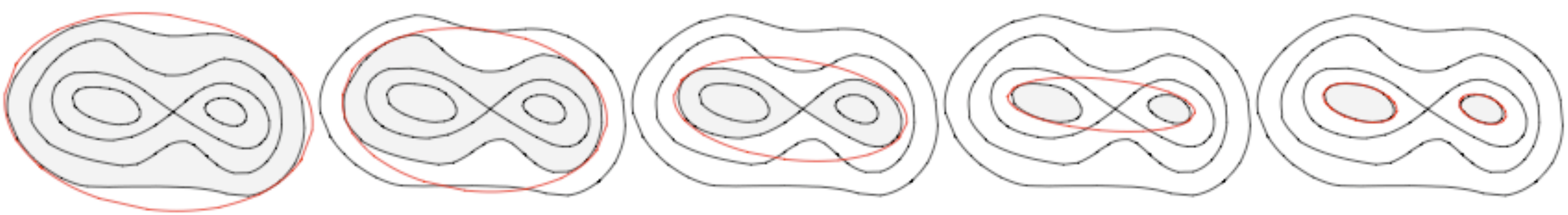
ICIC

Ellipsoidal decomposition

Unimodal distribution | Multimodal distribution



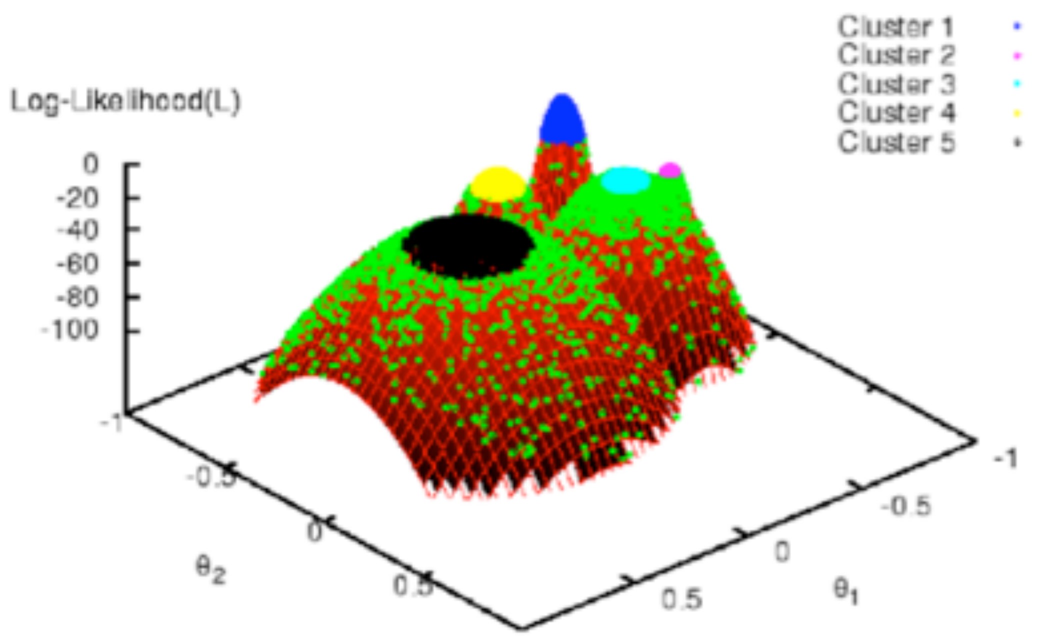




Courtesy Mike Hobson

Multinest: Efficiency

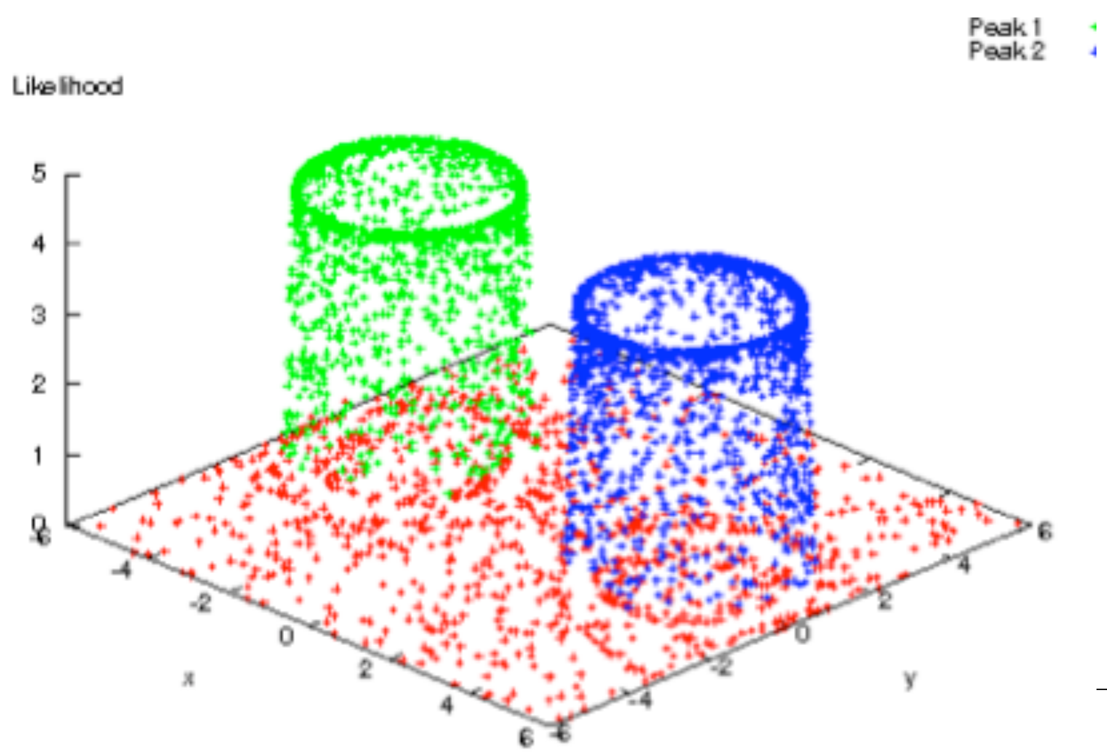
Imperial College London



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

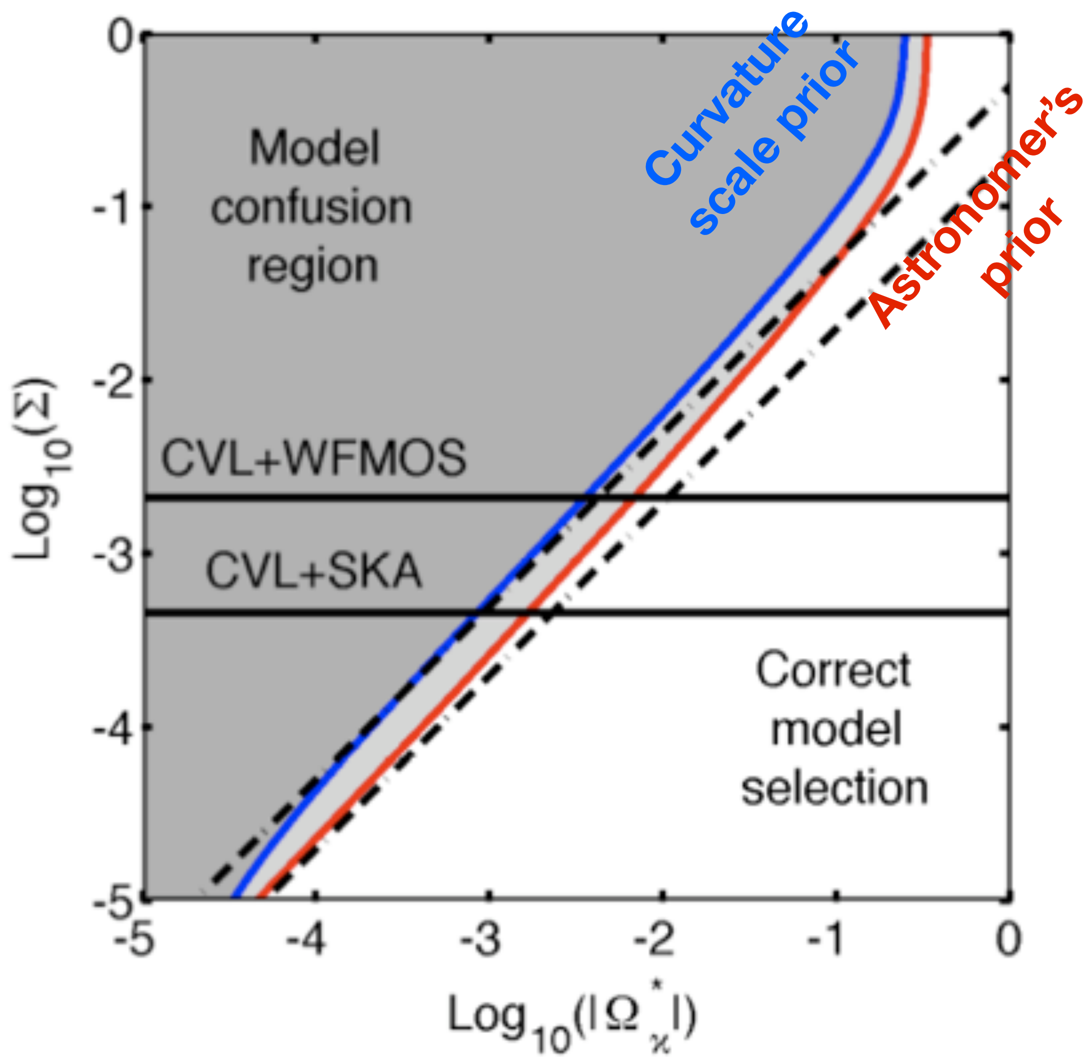


D	N_{like}	efficiency	likes per dimension
2	7000	70%	83
5	18000	51%	7
10	53000	34%	3
20	255000	15%	1.8
30	753000	8%	1.6

Application: the spatial curvature

Imperial College London

- Is the Universe spatially flat? (Vardanyan, Trotta and Silk, 2009)
- A three-way model comparison: **$\Omega_k = 0$ vs $\Omega_k < 0$ vs $\Omega_k > 0$** (with either the Astronomer's prior or Curvature scale prior)
- Result: odds range from moderate evidence (lnB = 4) for flatness to undecided (lnB = 0.4) depending on the choice of prior
- Probability(infinite Universe) = 98% (Astronomer's prior)
  Probability(infinite Universe) = 45% (Curvature scale prior)
- Upper bound: **odds of 49:1 at best for $n \neq 1$** (Gordon and Trotta 2007)



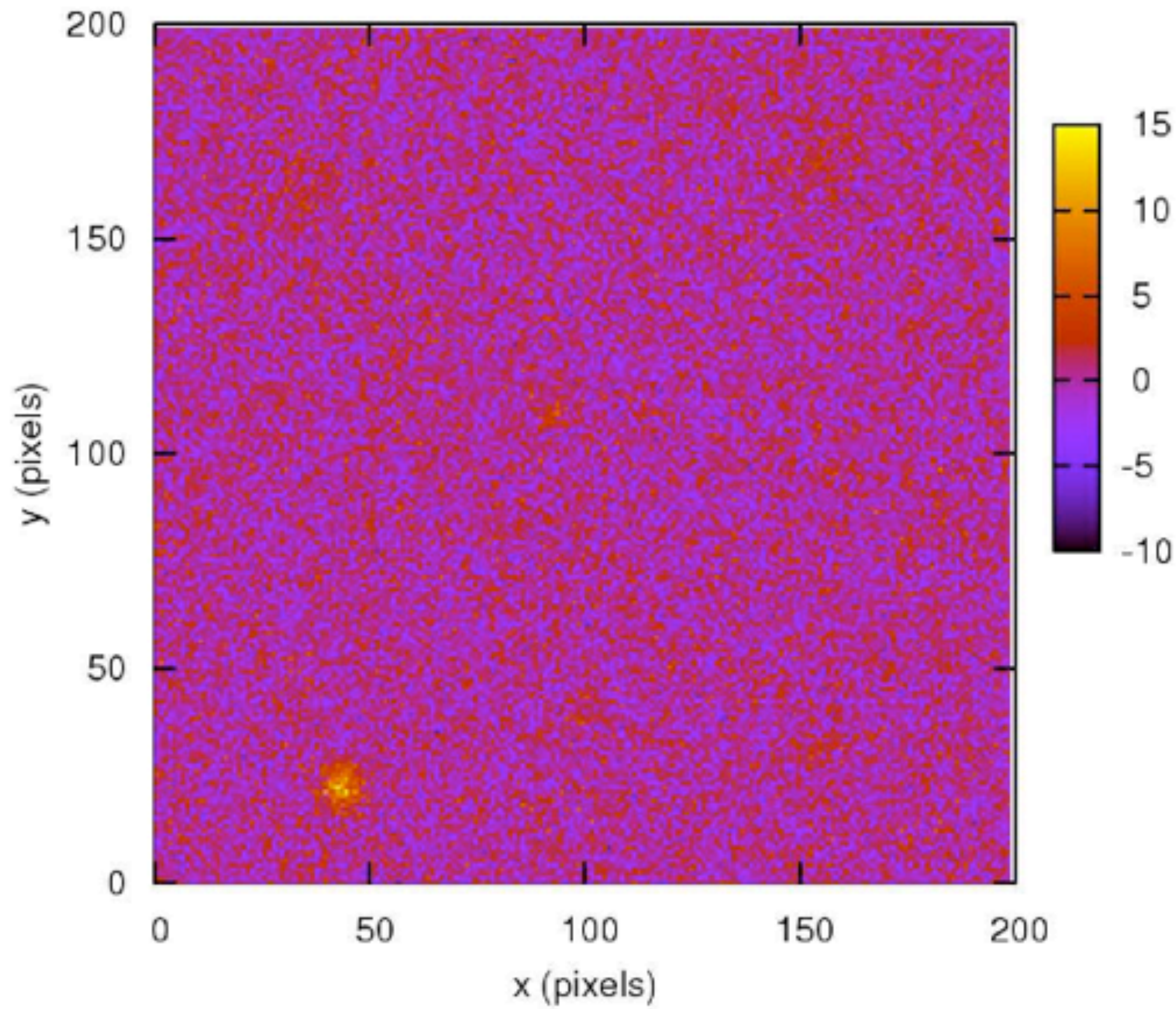
ICIC

A “simple” example: how many sources?

Imperial College London

Feroz and Hobson (2007)

Signal + Noise

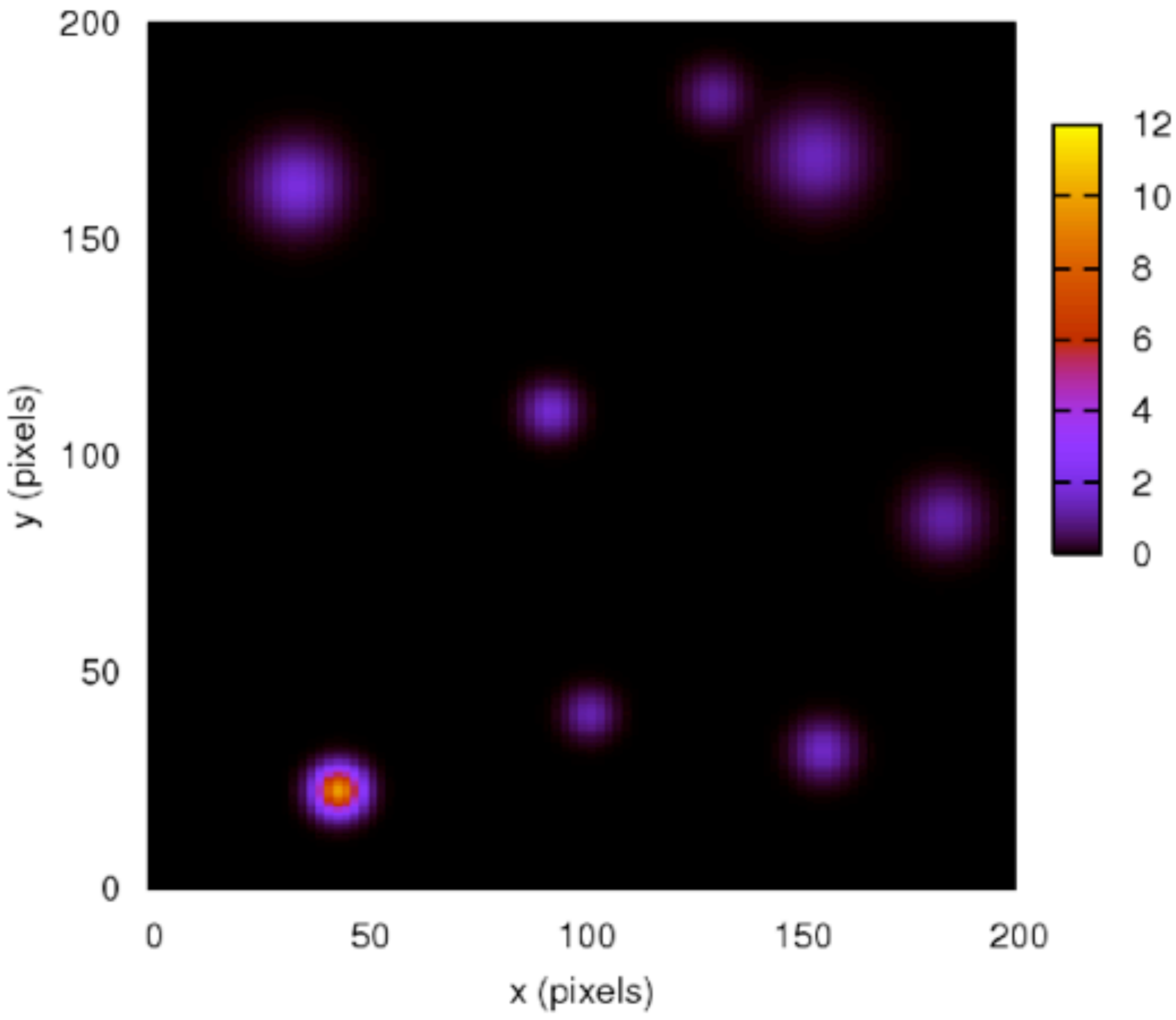


A "simple" example: how many sources?

Imperial College London

Feroz and Hobson (2007)

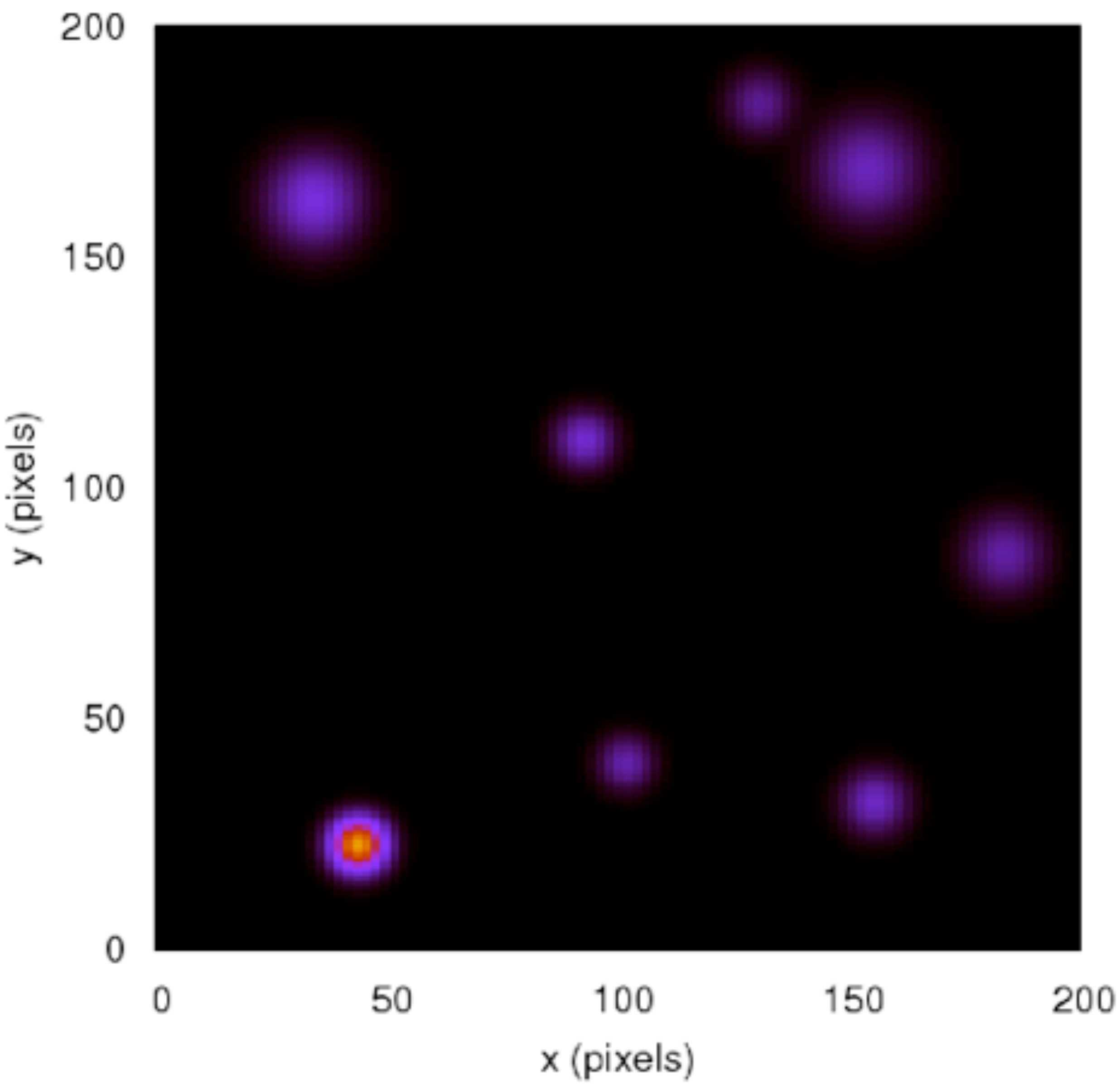
Signal: 8 sources



A “simple” example: how many sources?

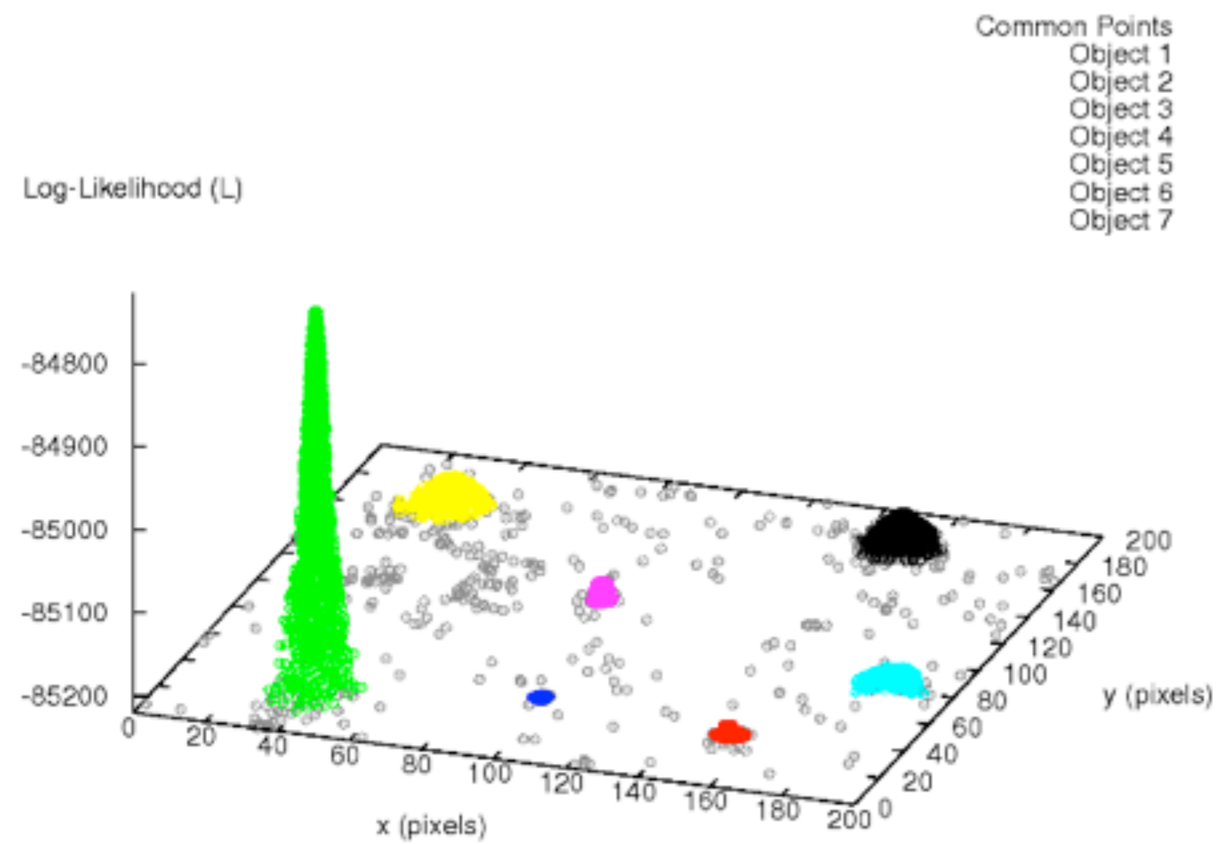
Imperial College London

Feroz and Hobson (2007)



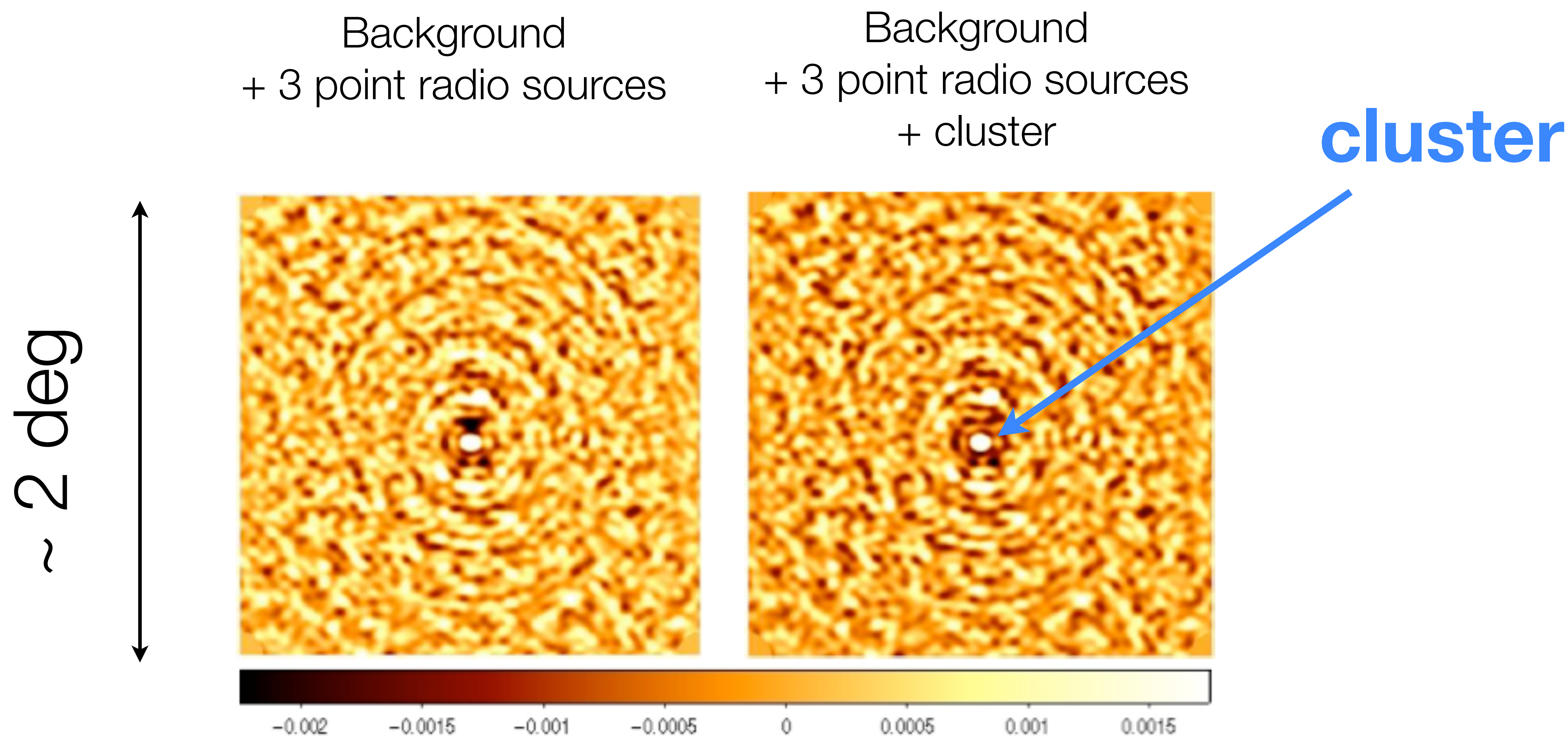
Bayesian reconstruction

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



Roberto Trotta

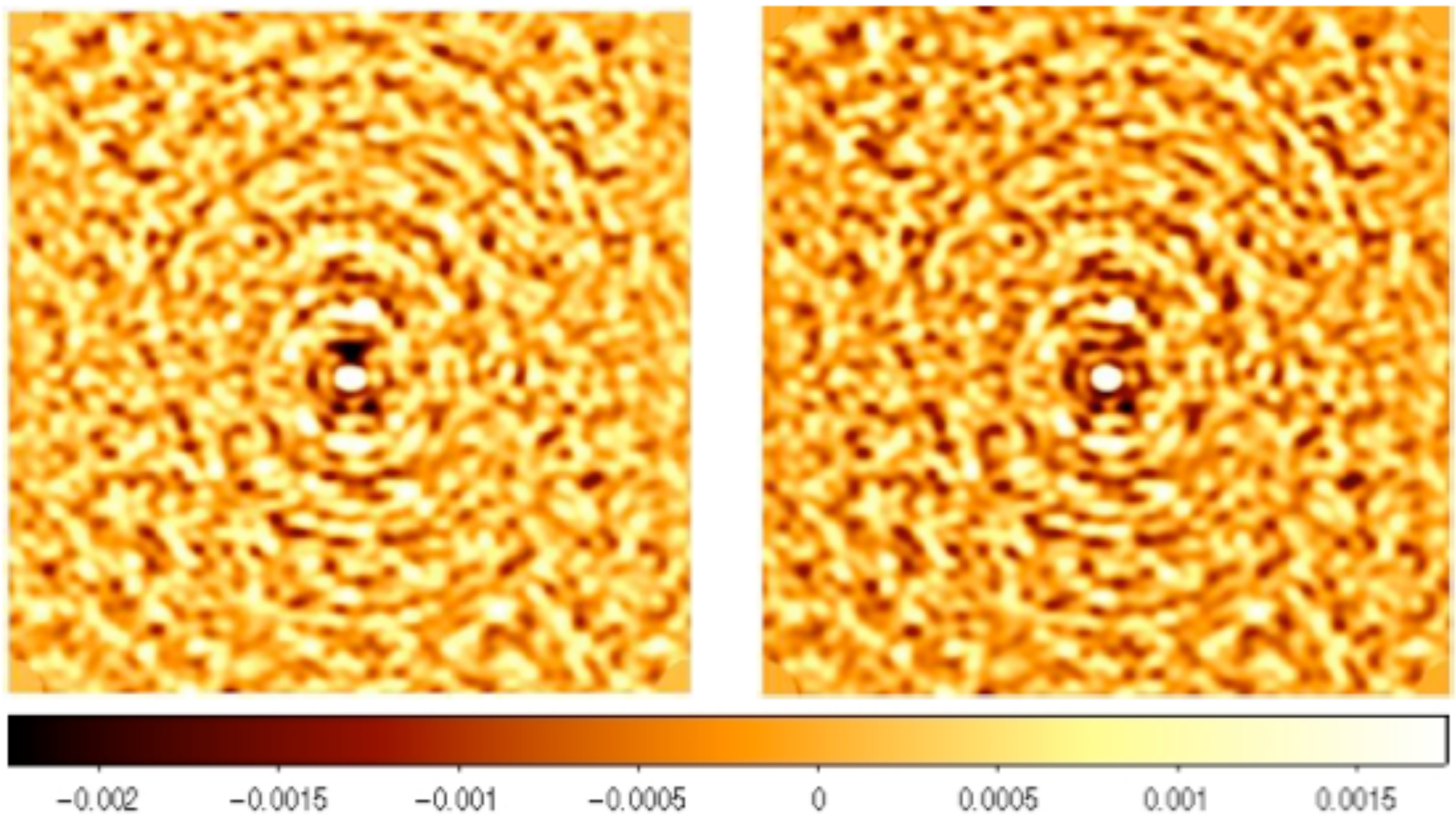
Cluster detection from Sunyaev-Zeldovich effect in cosmic microwave background maps



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)

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

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

The cosmological concordance model

Imperial College London

Competing model	ΔN_{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)

$\ln B < 0$: favours ΛCDM

ICIC

Roberto Trotta

Model complexity

- "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:

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

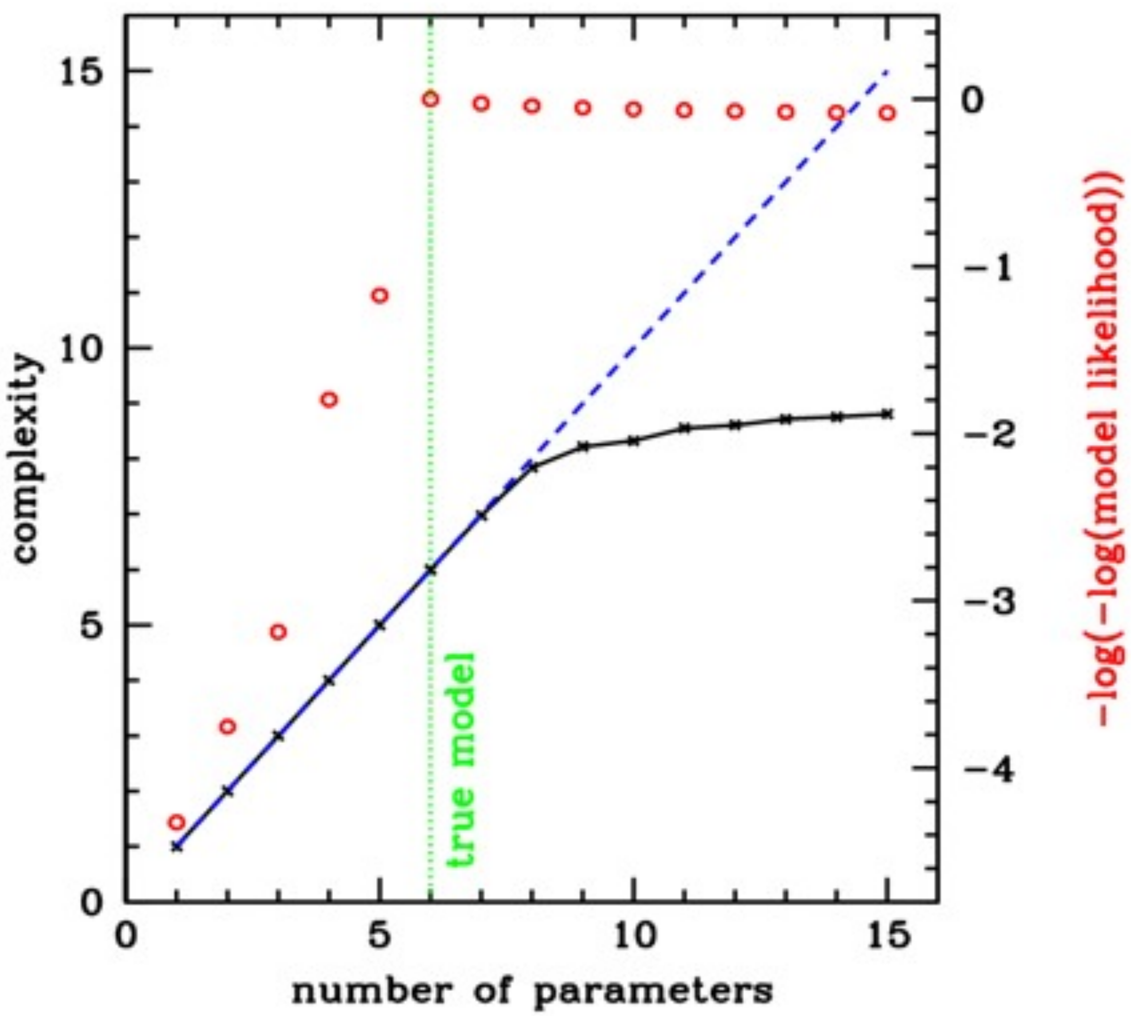
Kunz, RT & Parkinson, astro-ph/0602378, Phys. Rev. D 74, 023503 (2006)
Following Spiegelhalter et al (2002)

Polynomial fitting

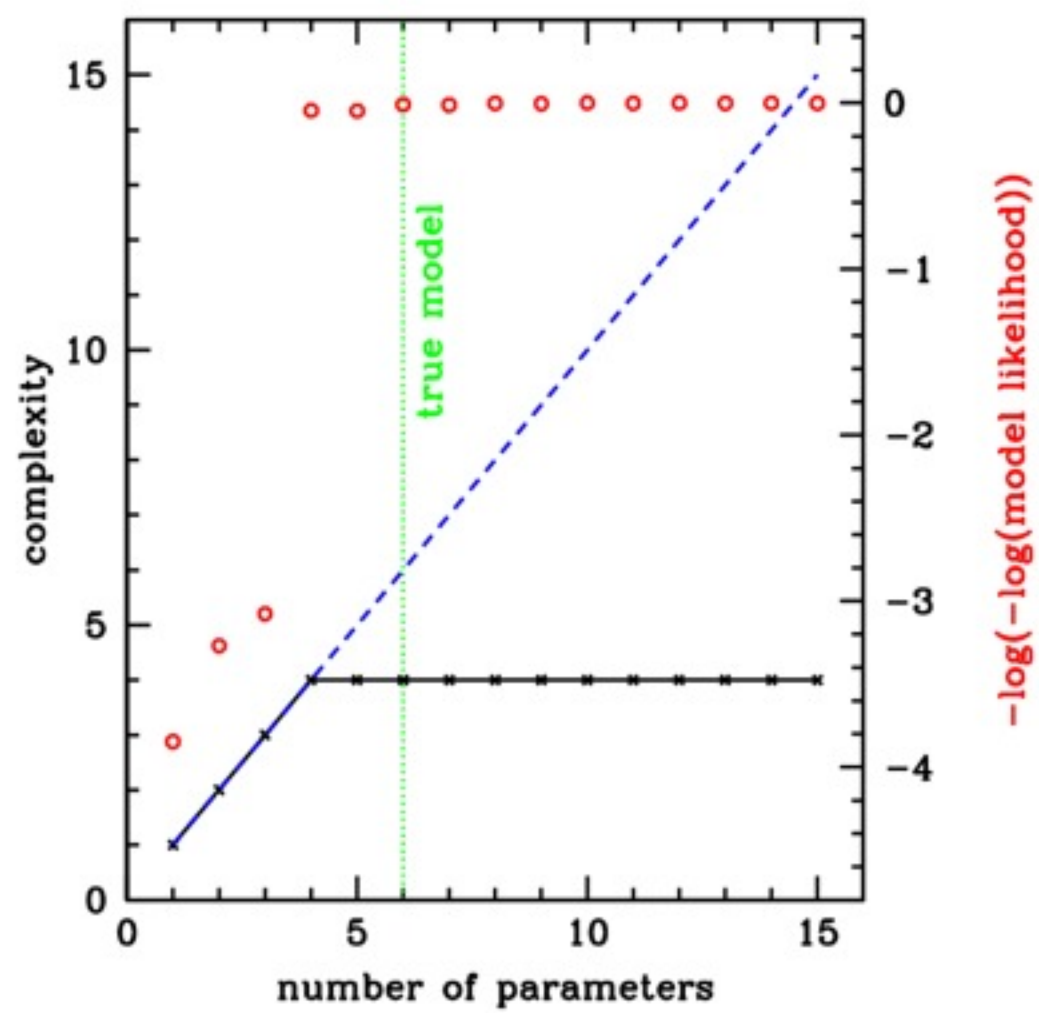
Imperial College London

- Data generated from a model with n = 6:

GOOD DATA
Max supported complexity ~ 9

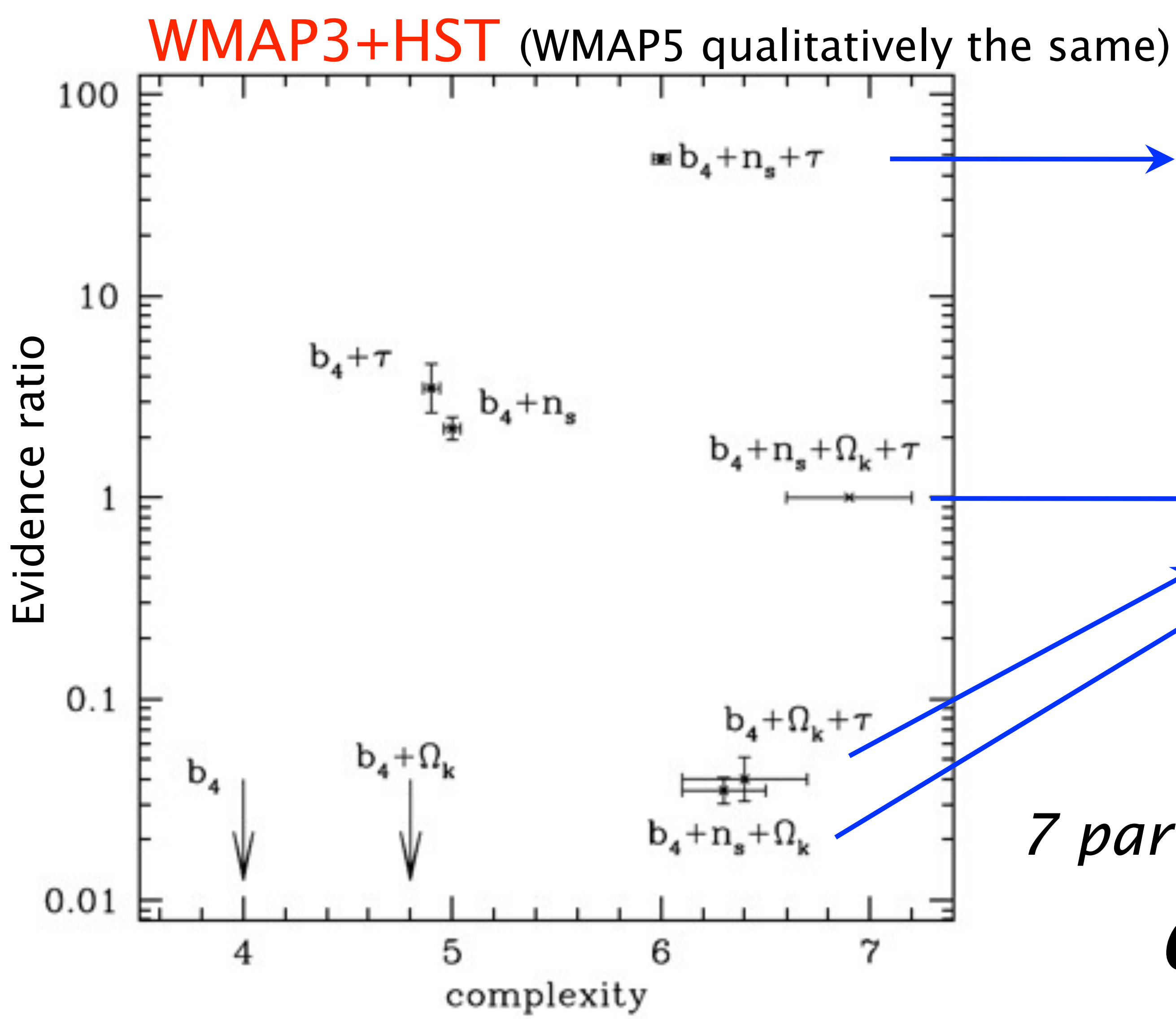


INSUFFICIENT DATA
Max supported complexity ~ 4



How many parameters does the CMB need?

Imperial College London



b_4+n_s+τ measured & favoured

Ω_κ *measured & unnecessary*

7 params measured

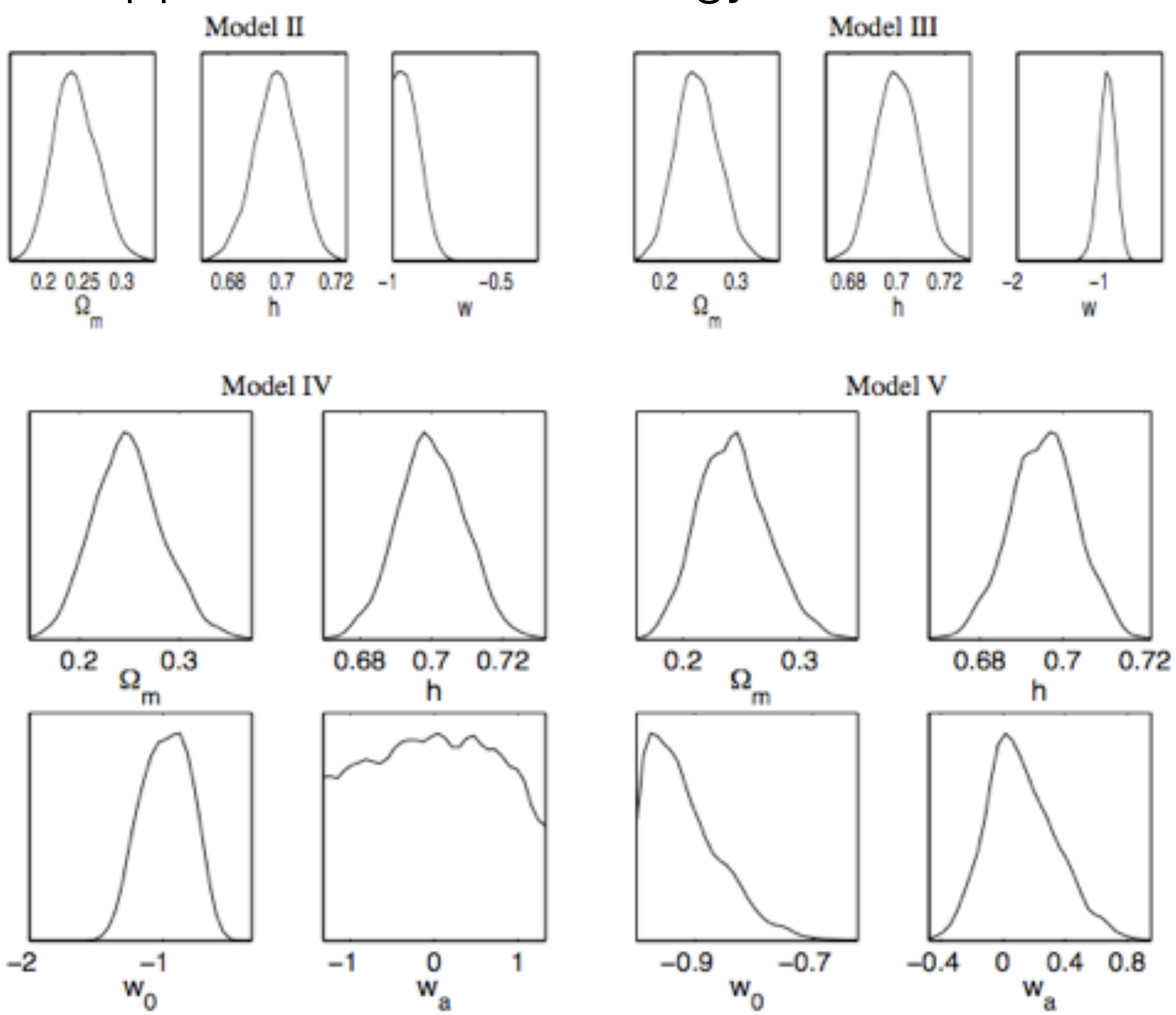
only 6 sufficient

Bayesian Model-averaging

Imperial College London

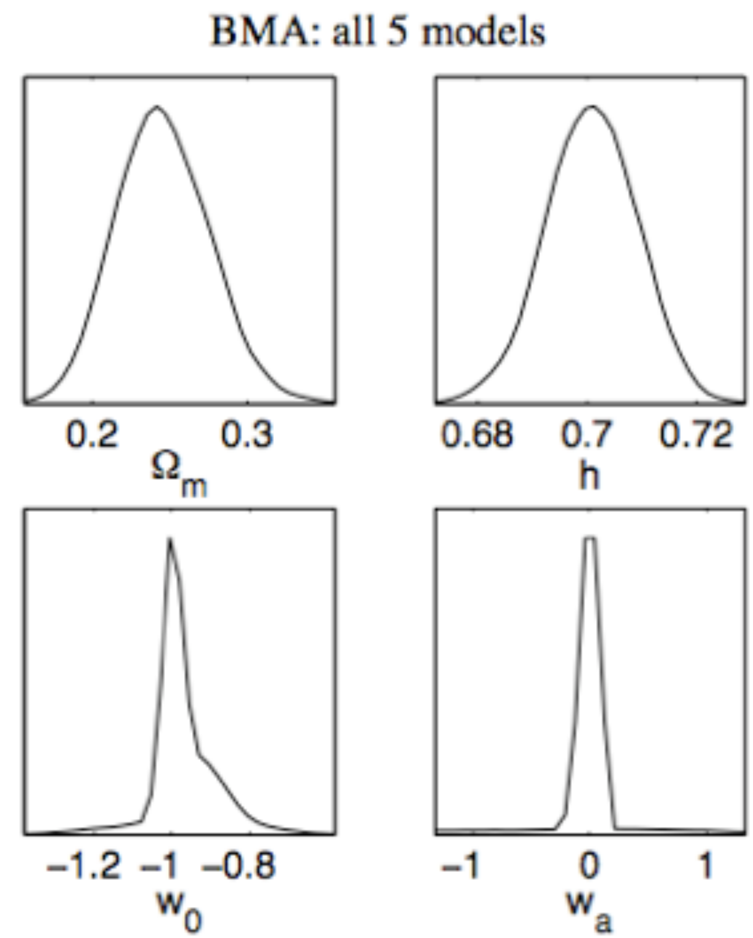
$$P(\boldsymbol{\theta}|d) = \sum_i P(\boldsymbol{\theta}|d,M_i)P(M_i|d)$$

An application to dark energy:



Liddle et al (2007)

Model averaged inferences



ICIC

Roberto Trotta

Key points

- 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.