Shaving Theories with Occam's Razor: Bayesian Model Comparison

@R_Trotta

Roberto Trotta
Astrophysics & Data Science Institute
Imperial College London

ICIC Imperial Centre for Inference & Cosmology

www.robertotrotta.com

Imperial College London

Frequentist hypothesis testing

- **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)

Exercise on hypothesis testing: **Is the coin fair?**

Two experiments are performed:

1. in the Blue Experiment, the coin is flipped N times, recording r heads.

2. in the Red Experiment, the coin is flipped until r heads are recorded.

Both experiments report the same data:

T T H T H T T T T T T H

Blue Team: N=12 is fixed, r the random variable

Red Team: r=3 is fixed, N the random variable

Question: What is the p-value for the null hypothesis?

Solution: Blue Experiment

- N here is fixed, r is the random variable

- The TS is the number of H recorded. Given that r=3 (i.e., smaller than you would expect under the null), a small TS indicates that the data are improbable under the null hypothesis that theta=1/2.

$$P(TS \leq TS_{\rm obs}) = P(r \leq r_{\rm obs} | N, \theta = \frac{1}{2})$$

- Using N = 12, r_{obs} = 3, the p-value is:

$$P(TS \leq TS_{\rm obs}) = \sum_{r=0}^{r_{\rm obs}} P(r|N, \theta = \frac{1}{2}) = \sum_{r=0}^{r_{\rm obs}} \binom{N}{r} \frac{1}{2^N} = 0.073$$

- This result is **not** significant at the 5% level (p-value = 0.05)

Solution: Red Experiment

Imperial College London

- r here is fixed, N is the random variable

- The TS is the number of flips required until we get r=3 heads. In this case, a large value of the TS (i.e., having to wait for a long number of flips) indicates that the data are improbable under the null hypothesis that theta=1/2.

$$P(TS \geq TS_{\rm obs}) = P(N \geq N_{\rm obs}|r, \theta = \frac{1}{2}) = 1 - P(N < N_{\rm obs}|r, \theta = \frac{1}{2})$$

- Using r = 3, N_{obs} = 12, the p-value is:

$$P(N < N_{\rm obs}|r, \theta = \frac{1}{2}) = \sum_{N=r}^{N_{\rm obs}-1} \binom{N-1}{r-1} \frac{1}{2^N} = 0.967$$

$$P(TS \leq TS_{\rm obs}) = 0.033$$

- This result is significant at the 5% level (p-value = 0.05)

The Bayesian Calculation

Imperial College London

- We compare M0 with theta=1/2 to M1 where theta is a free parameter.
- We choose a uniform prior [0,1] for theta under M1 (other choices are possible).
- Compute the Bayesian evidence under M1:

$$P(d|M_1) = \int d\theta \mathcal{L}(\theta) P(\theta|M_1) = \int_0^1 d\theta \binom{N}{r} \theta^r (1-\theta)^{N-r} = \binom{N}{r} \frac{r!(N-r)!}{(N+1)!}$$

- Compute the Bayesian evidence under M0 (notice M0 has no free parameters):

$$P(d|M_0) = \binom{N}{r} \frac{1}{2^N}$$

- The Bayes factor (using N=12, r=3) gives almost no evidence in favour of M1!

$$B_{10} = \frac{P(d|M_1)}{P(d|M_0)} = \frac{r!(N-r)!}{(N+1)!} 2^N = 1.43$$

The significance of significance

- **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 model comparison

The 3 levels of inference

Imperial College London

LEVEL 1

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

Parameter inference

What are the favourite values of the parameters? (assumes M is true)

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

LEVEL 2

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

Model comparison

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

$$\text{odds} = \frac{\mathrm{P}(\mathrm{M}_0|\mathrm{d})}{\mathrm{P}(\mathrm{M}_1|\mathrm{d})}$$

LEVEL 3

None of the models is clearly the best

Model averaging

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

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

Examples of model comparison questions

Imperial College London

ASTROPARTICLE

Gravitational waves detection

Do cosmic rays correlate with AGNs?

Which SUSY model is ‘best’?

Is there evidence for DM modulation?

Is there a DM signal in gamma ray/ neutrino data?

COSMOLOGY

Is the Universe flat?

Does dark energy evolve?

Are there anomalies in the CMB?

Which inflationary model is ‘best’?

Is there evidence for modified gravity?

Are the initial conditions adiabatic?

Many scientific questions are of the model comparison type

ASTROPHYSICS

Exoplanets detection

Is there a line in this spectrum?

Is there a source in this image?

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

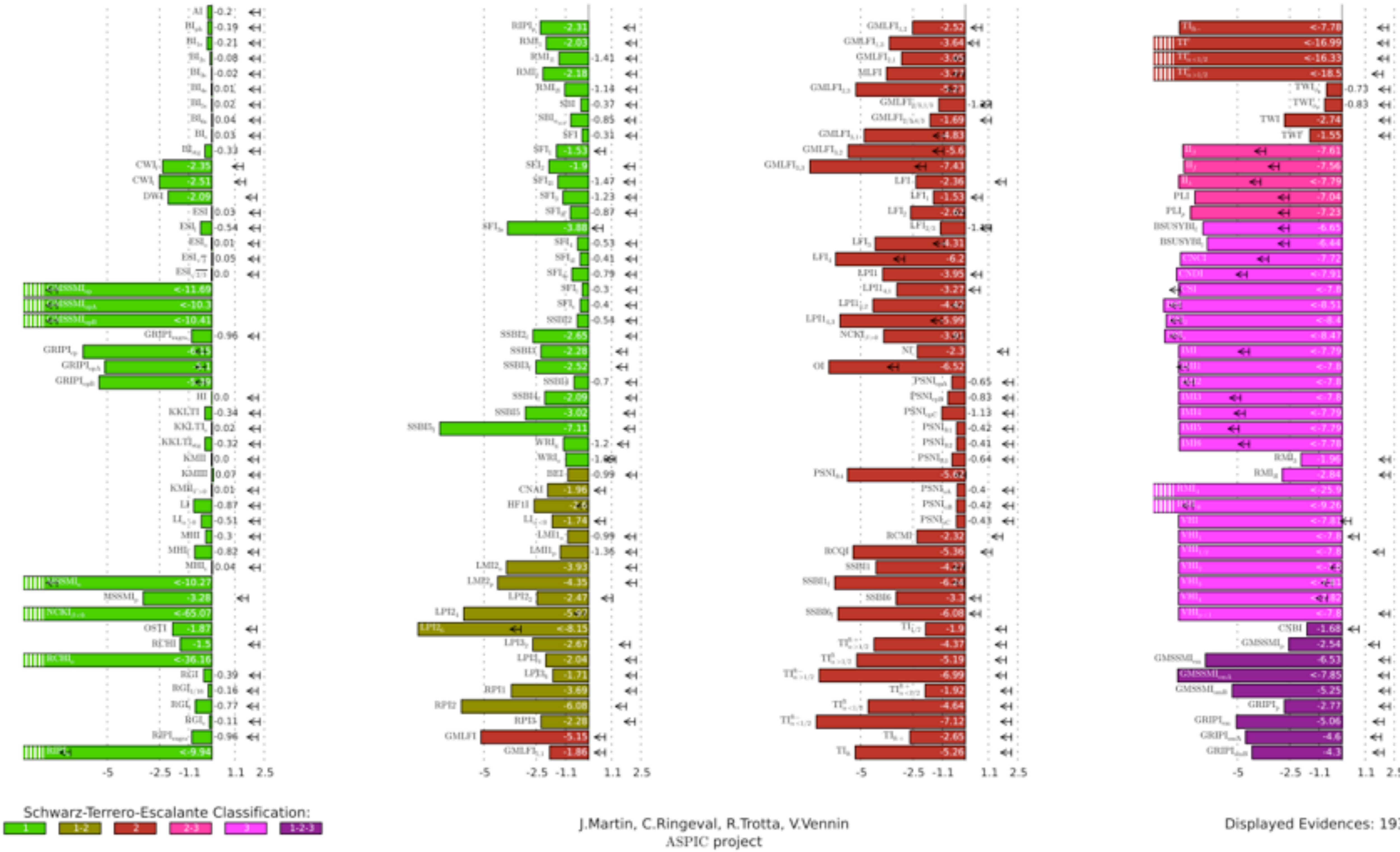
- A (slightly modified) Jeffreys' scale to assess the strength of evidence

$\lvert \ln B \rvert$	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

Bayesian model comparison of 193 models
Higgs inflation as reference model

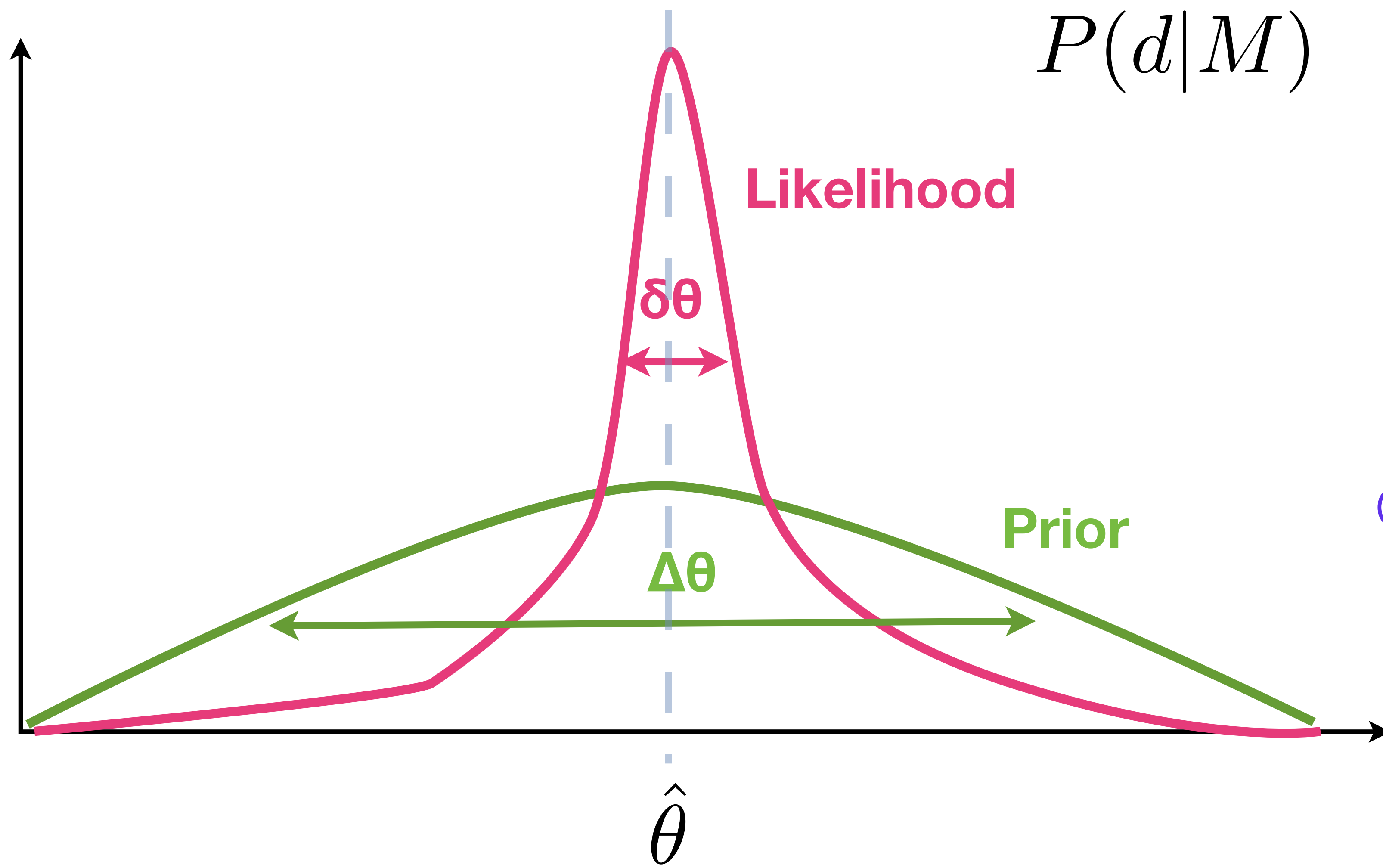
$$\ln(\mathcal{E}/\mathcal{E}_{\mathrm{HI}})$$

Martin,RT+14



An automatic Occam's razor

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

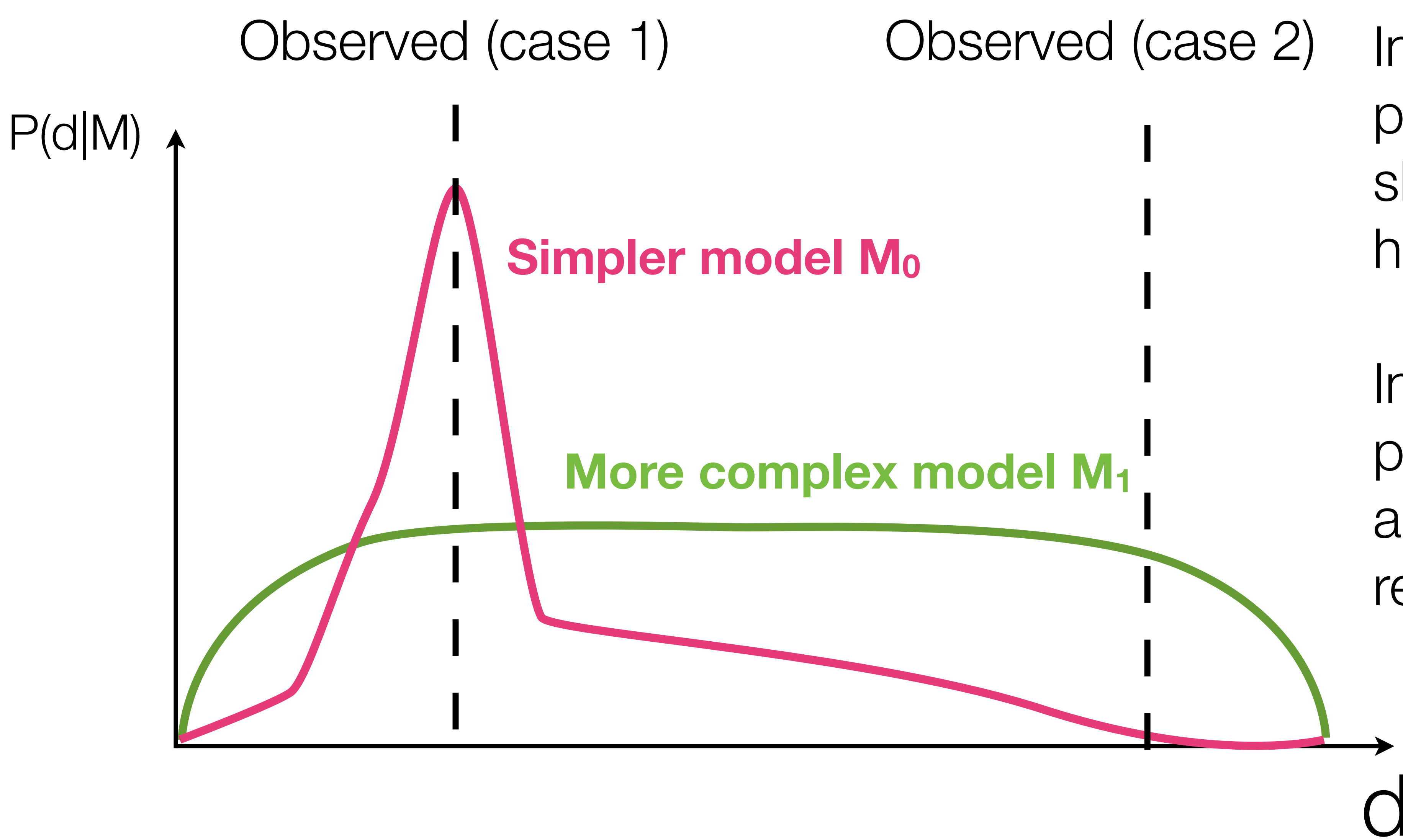


$$
\begin{aligned}
P(d|M) &= \int d\theta L(\theta) P(\theta|M) \\
&\approx P(\hat{\theta}) \delta\theta L(\hat{\theta}) \\
&\approx \frac{\delta\theta}{\Delta\theta} L(\hat{\theta})
\end{aligned}
$$

Occam's factor

The evidence as predictive probability

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



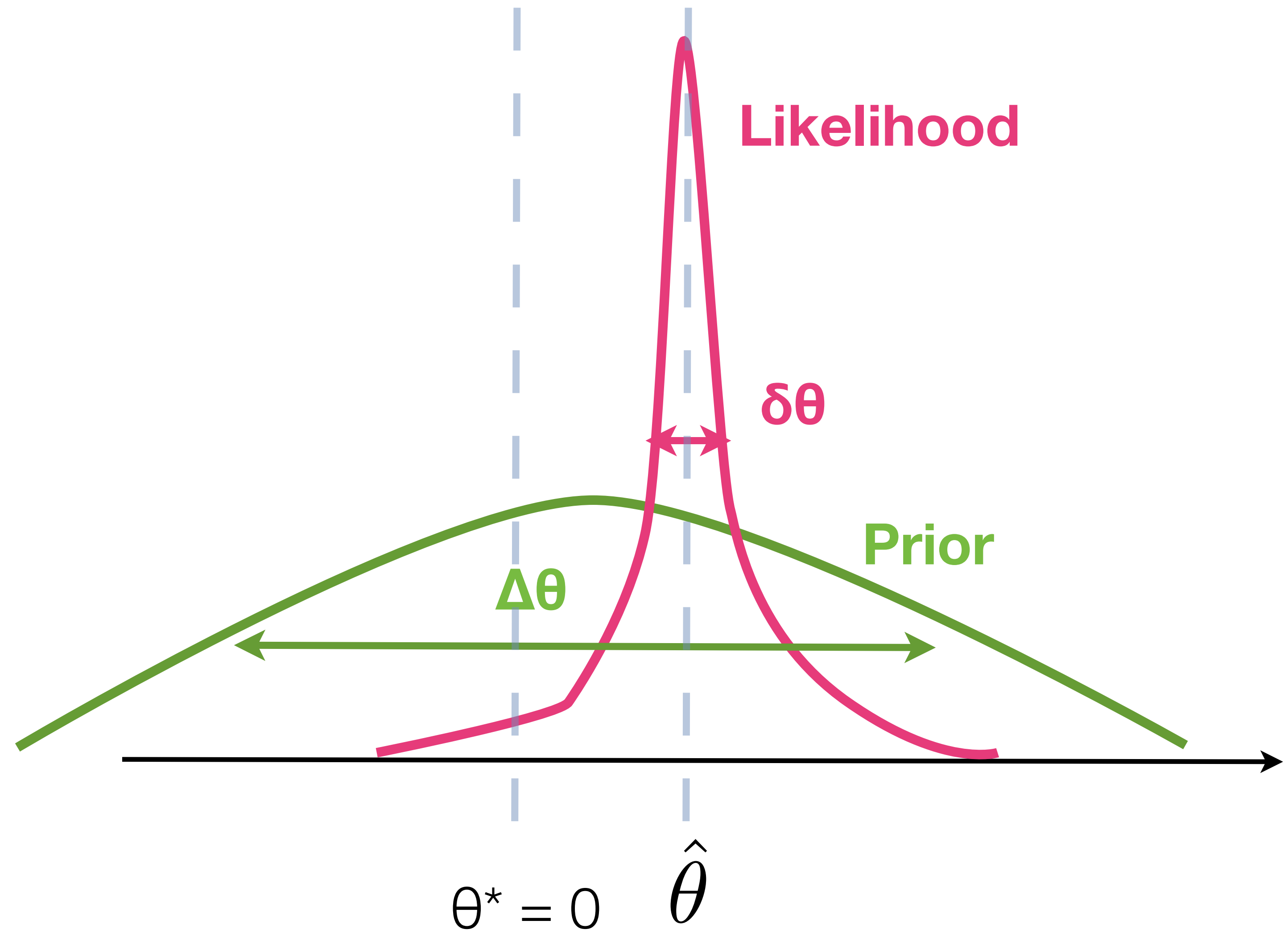
In case 1, M0 is preferred (as it made a sharp prediction that has been verified)

In case 2, M1 is preferred as its additional complexity is required by the data

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

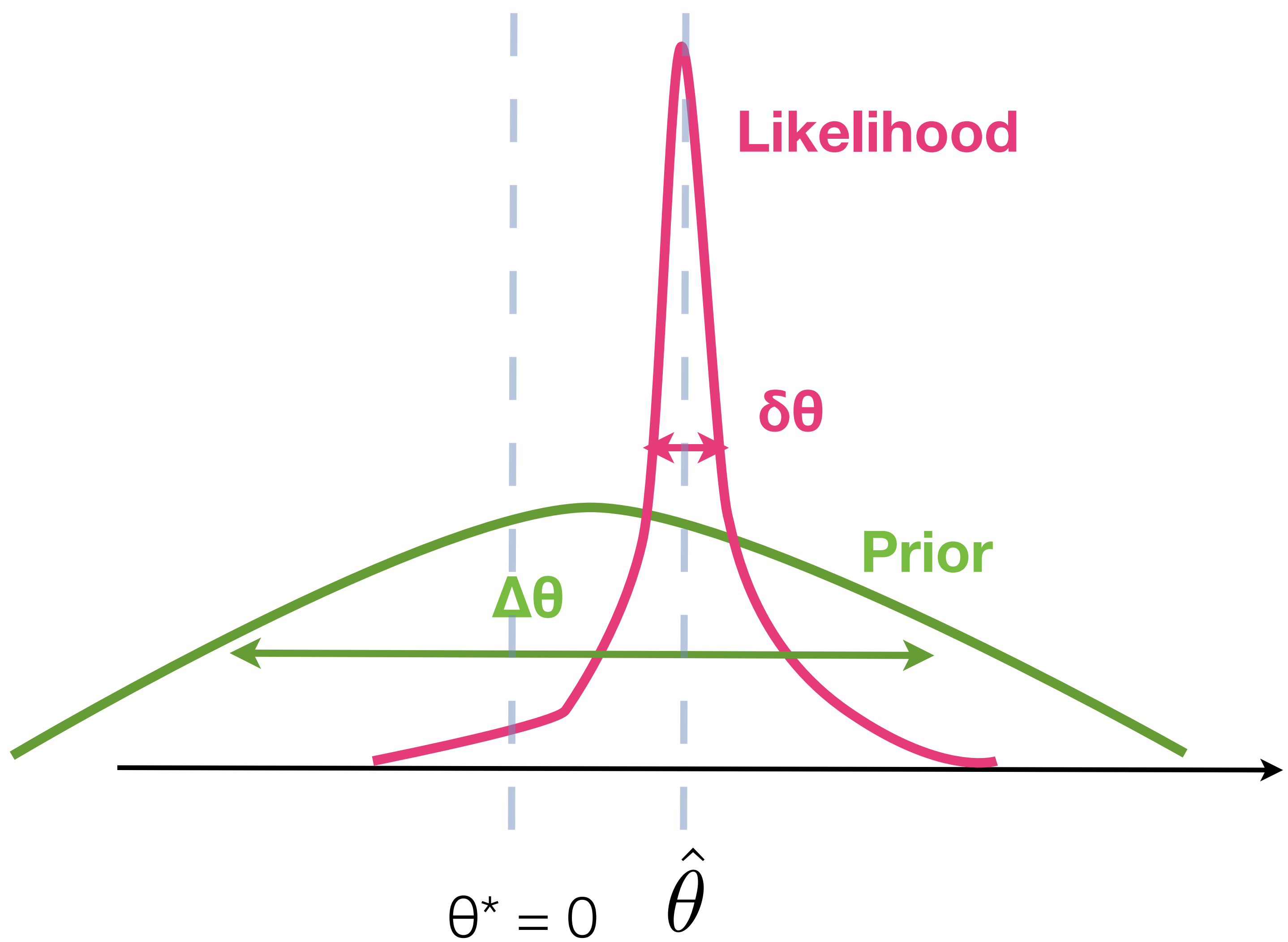
Define: $\lambda \equiv \frac{\hat{\theta}-\theta^\star}{\delta\theta}$

For “informative” data:

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

wasted parameter space (favours simpler model)

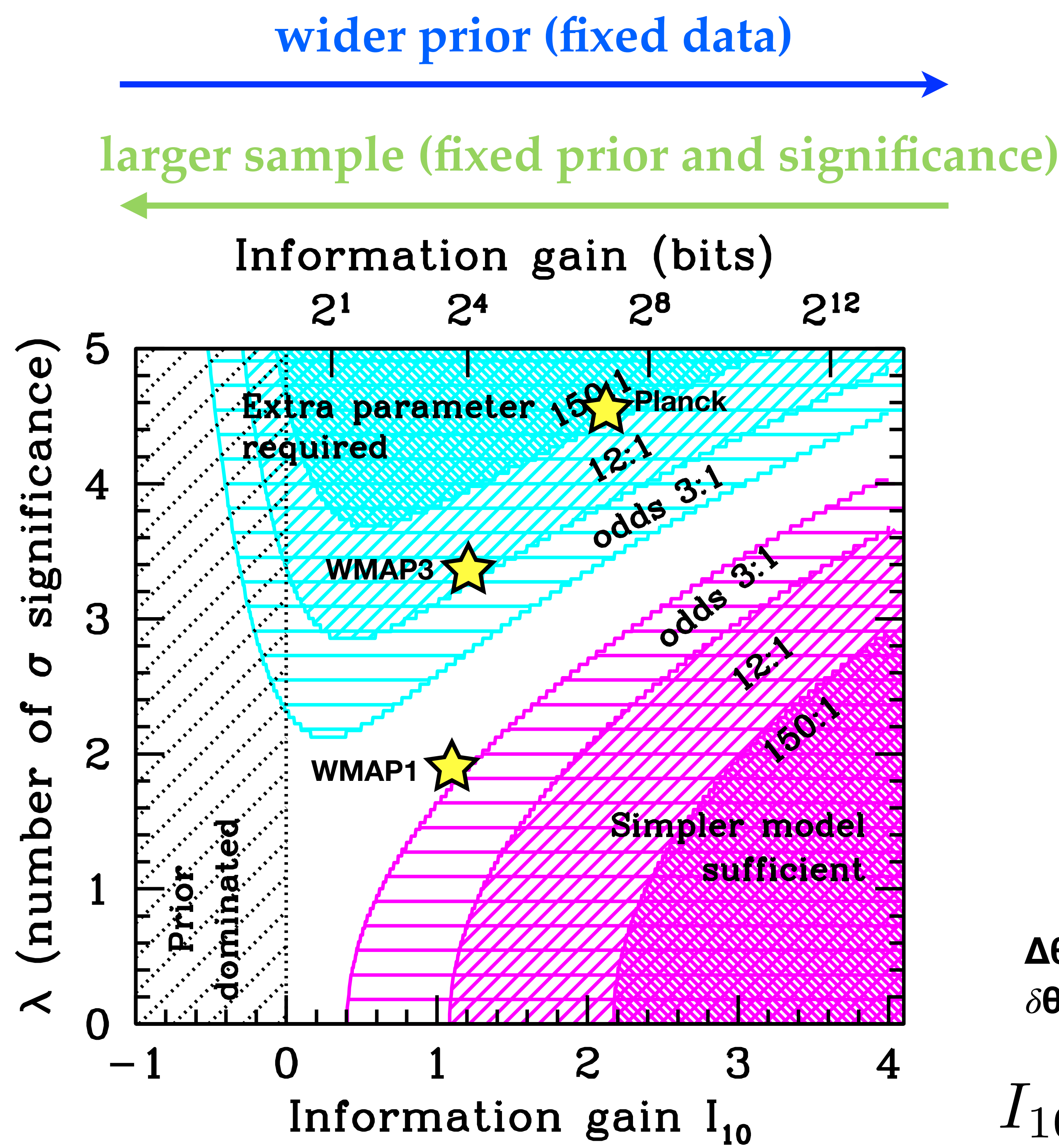
mismatch of prediction with observed data (favours more complex model)



The rough guide to model comparison

Imperial College London

Trotta (2008)



Δθ = Prior width
δθ = Likelihood width

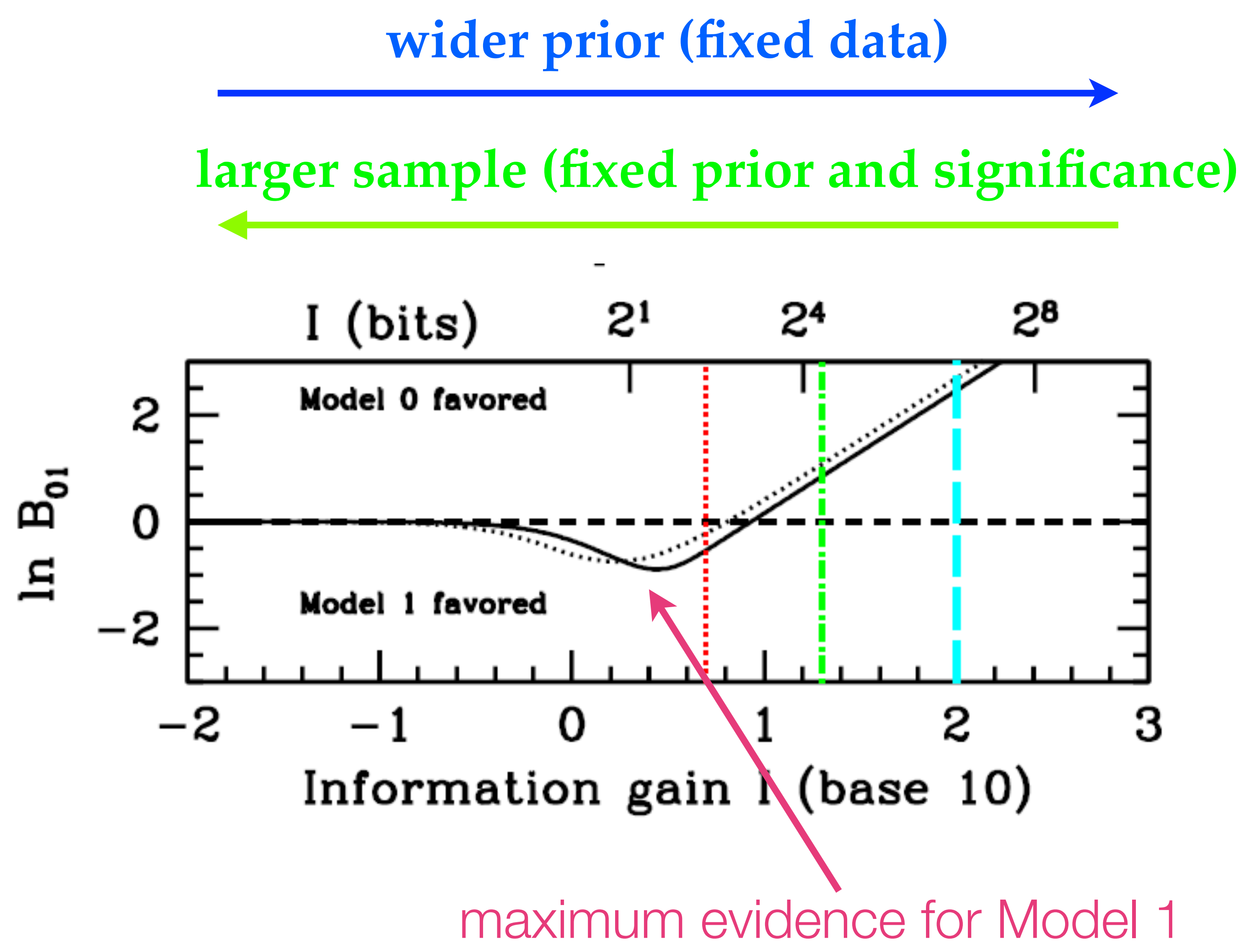
$$I_{10} \equiv \log_{10} \frac{\Delta\theta}{\delta\theta}$$

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	2.0 (7:1) weak	0.9 (3:1) undecided
0.003	3	4.5 (90:1) moderate	3.0 (21:1) moderate
0.0003	3.6	6.48 (650:1) strong	5.0 (150:1) strong

How to assess p-values

Imperial College London

Rule of thumb:
interpret a n-sigma result as a (n-1)-sigma result

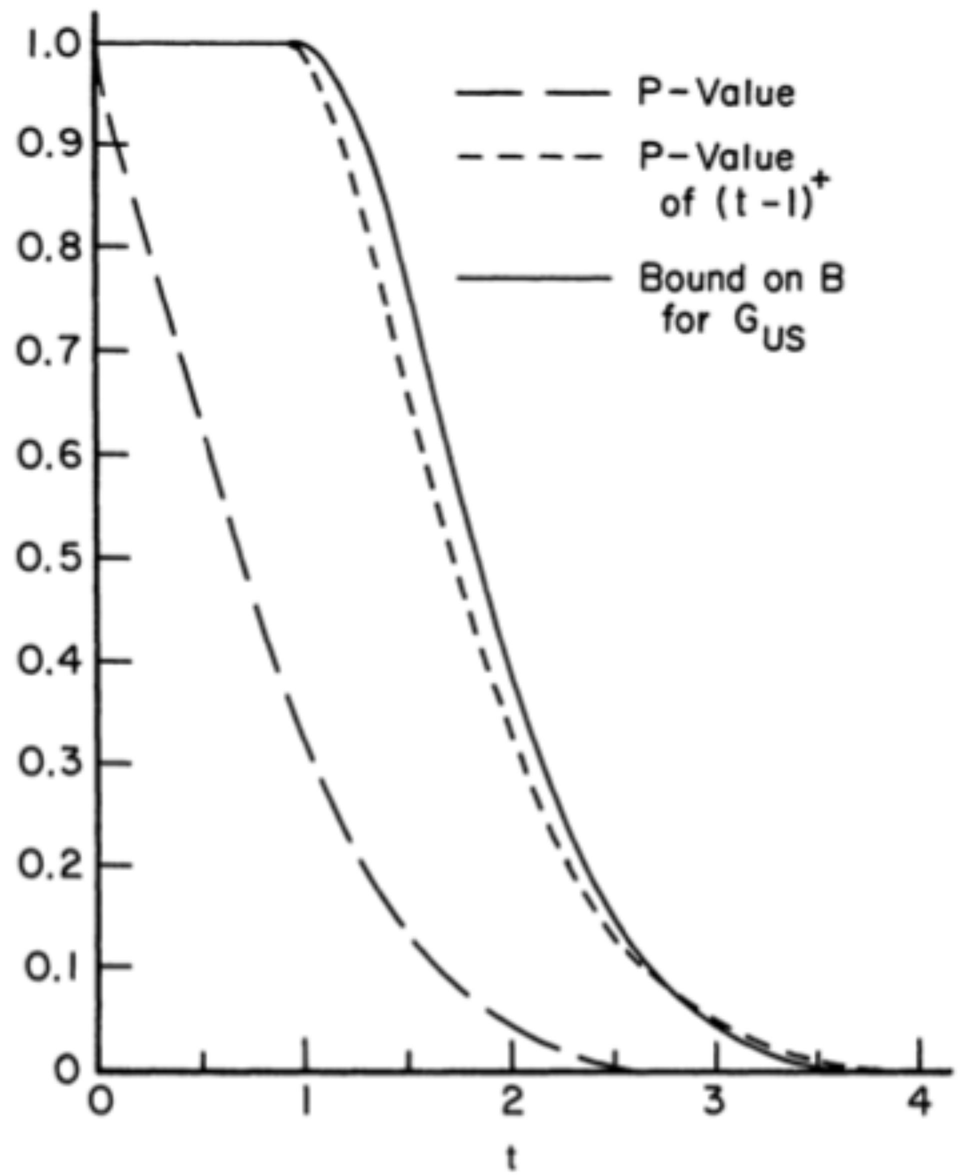


Figure 4. Comparison of $\underline{B}(x, G_{US})$ and P Values.

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

Computing the model likelihood

Imperial College London

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

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

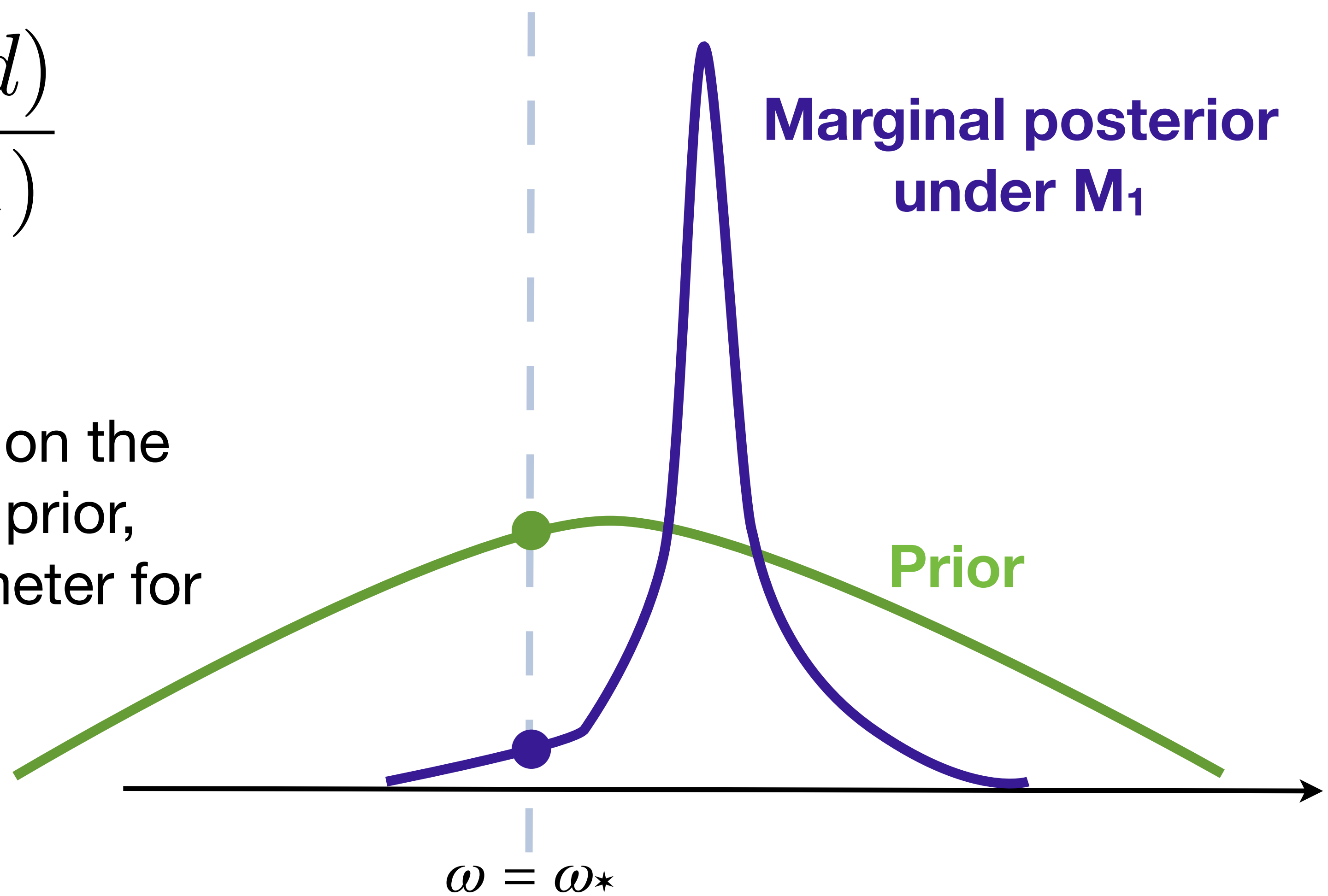
- Usually computational demanding: it's a multi-dimensional integral, averaging the likelihood over the (possibly much wider) prior
- I'll present two methods used by cosmologists:
  - **Savage-Dickey density ratio (Dickey 1971):** Gives the Bayes factor between *nested* models (under mild conditions). Can be usually derived from posterior samples of the larger (higher D) model.
  - **Nested sampling (Skilling 2004):** Transforms the D-dim integral in 1D integration. Can be used generally (within limitations of the efficiency of the sampling method adopted).

The Savage-Dickey density ratio

Imperial College London

Dickey J. M., 1971, Ann. Math. Stat., 42, 204

- This method works for *nested models* and gives the Bayes factor analytically.
- **Assumptions:**
  - Nested models: M_1 with parameters (Ψ,ω) reduces to M_0 for e.g. $\omega=\omega_\star$
  - Separable priors: the prior $\pi_1(\Psi,\omega|M_1)$ is uncorrelated with $\pi_0(\Psi|M_0)$
- **Result:** $$B_{01} = \frac{p(\omega_\star|d)}{\pi_1(\omega_\star)}$$



- The Bayes factor is the ratio of the normalised (1D) marginal posterior on the additional parameter in M_1 over its prior, evaluated at the value of the parameter for which M_1 reduces to M_0.

Derivation of the SDDR

RT, Mon.Not.Roy.Astron.Soc. 378 (2007) 72-82

$$P(d|M_0) = \int d\Psi \pi_0(\Psi) p(d|\Psi, \omega_\star) \quad P(d|M_1) = \int d\Psi d\omega \pi_1(\Psi, \omega) p(d|\Psi, \omega)$$

Divide and multiply B_{01} by:

$$p(\omega_\star|d) = \frac{p(\omega_\star, \Psi|d)}{p(\Psi|\omega_\star, d)}$$

$$B_{01} = p(\omega_\star|d) \int d\Psi \frac{\pi_0(\Psi) p(d|\Psi, \omega_\star)}{P(M_1|d)} \frac{p(\Psi|\omega_\star, d)}{p(\omega_\star, \Psi|d)}$$

Since:

$$p(\omega_\star, \Psi|d) = \frac{p(d|\omega_\star, \Psi)\pi_1(\omega_\star, \Psi)}{P(M_1|d)}$$

$$B_{01} = p(\omega_\star|d) \int d\Psi \frac{\pi_0(\Psi) p(\Psi|\omega_\star, d)}{\pi_1(\omega_\star, \Psi)}$$

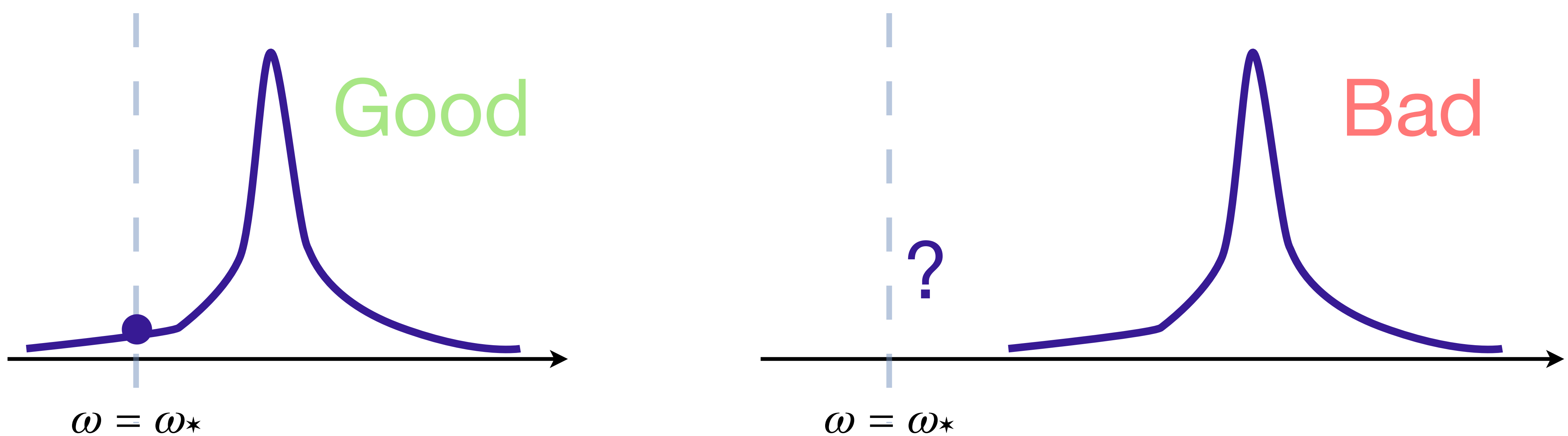
Assuming separable priors:

$$\pi_1(\omega, \Psi) = \pi_1(\omega)\pi_0(\Psi)$$

$$B_{01} = \frac{p(\omega_\star|d)}{\pi_1(\omega_\star)} \int d\Psi p(\Psi|\omega_\star, d) = \frac{p(\omega_\star|d)}{\pi_1(\omega_\star)}$$

SDDR: Some comments

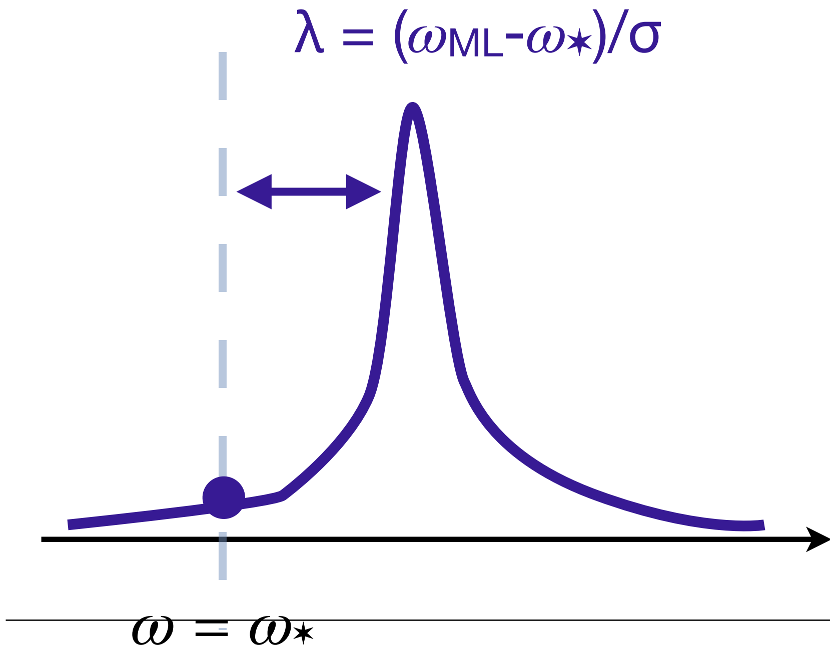
- For separable priors (and nested models), the common parameters do not matter for the value of the Bayes factor
- No need to spend time/resources to average the likelihoods over the common parameters
- Role of the prior on the additional parameter is clarified: the wider, the stronger the Occam's razor effect (due to dilution of the predictive power of model 1)
- Sensitivity analysis simplified: only the prior/scale on the additional parameter between the models needs to be considered.
- Notice: SDDR does not assume Gaussianity, but it does require sufficiently detailed sampling of the posterior to evaluate reliably its value at $\omega=\omega_{\star}$.

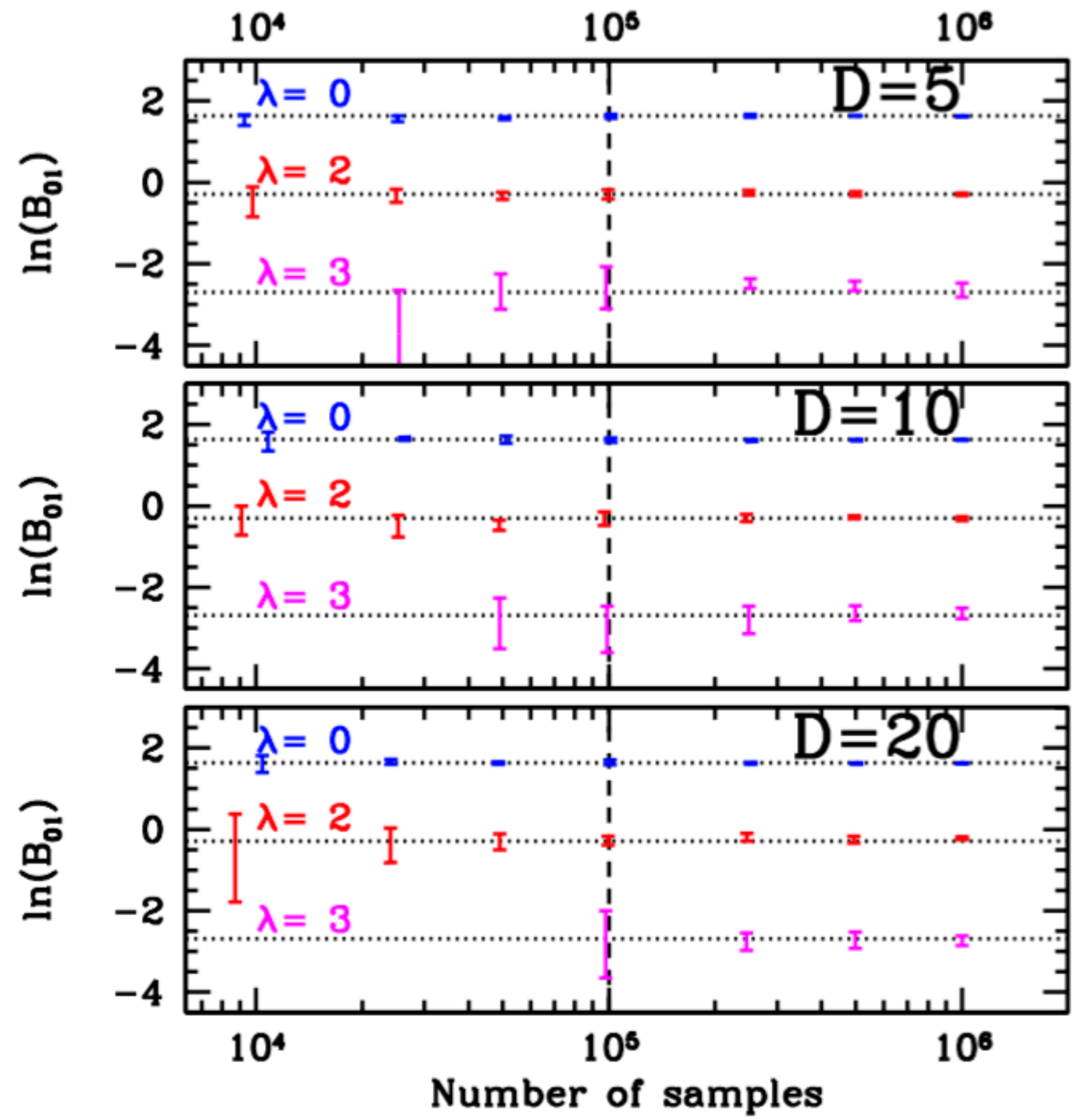


Accuracy tests (Normal case)

Imperial College London

- Tests with variable dimensionality (D) and number of MCMC samples
- λ is the distance of peak posterior from $\omega_\star$ in units of posterior std dev
- SDDR accurate with standard MCMC sampling up to 20-D and λ=3
- Accurate estimates further in the tails might required dedicated sampling schemes





RT, MNRAS, 378, 72-82 (2007)

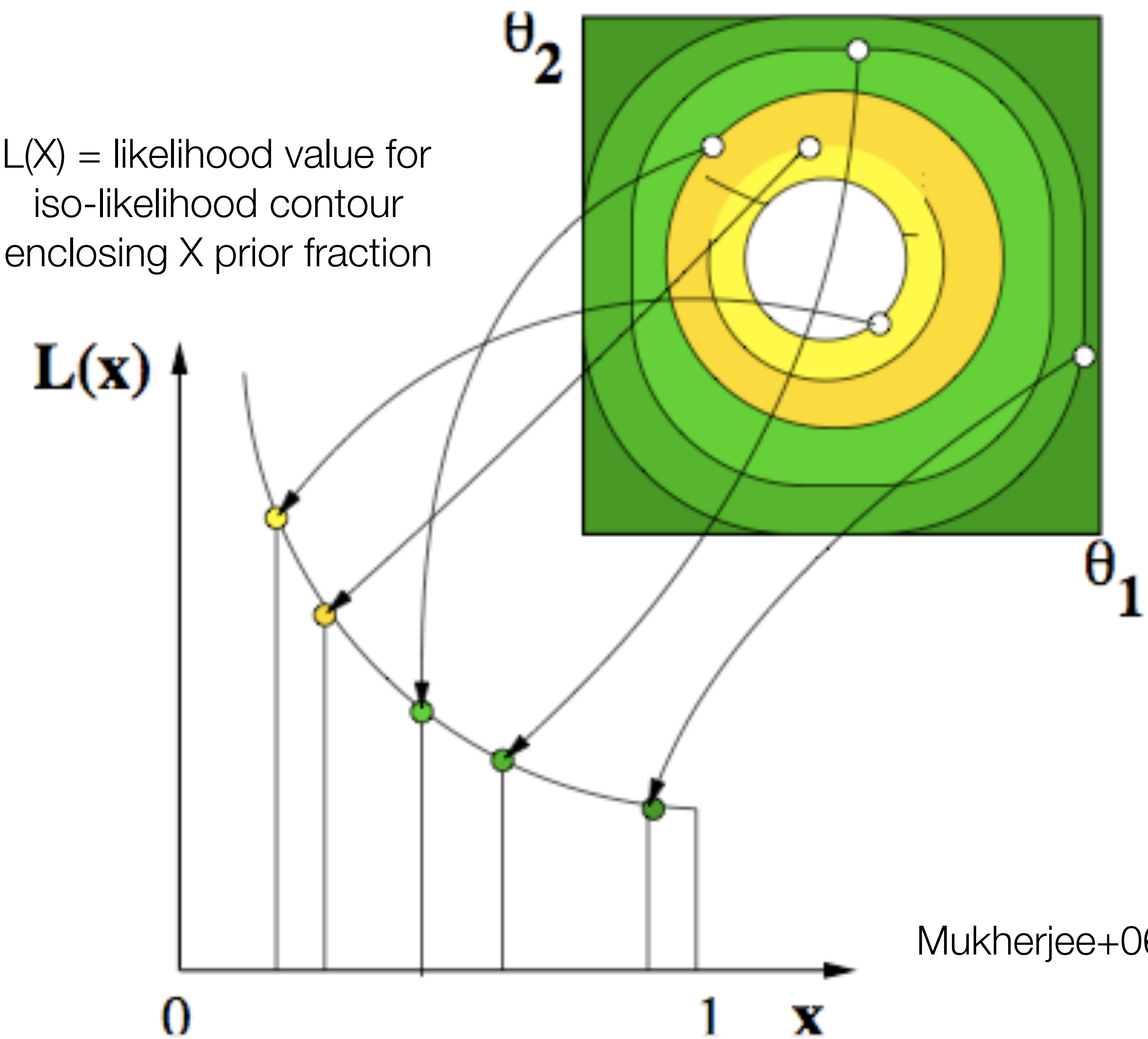
ICIC

Nested Sampling

Imperial College London

- Proposed by John Skilling in 2004: the idea is to convert a D-dimensional integral in a 1D integral that can be done easily.
- As a by-product, it also produces posterior samples: model likelihood and parameter inference obtained simultaneously

L(X) = likelihood value for iso-likelihood contour enclosing X prior fraction

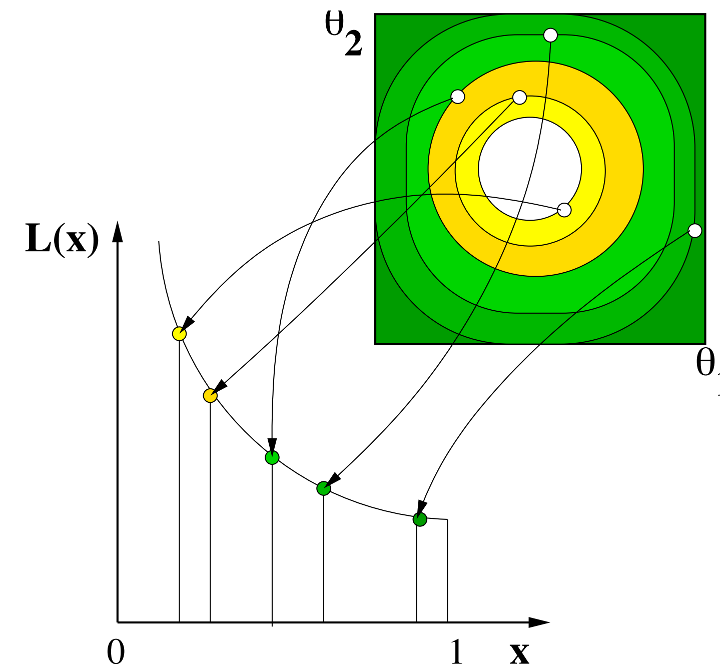


Mukherjee+06

X = Prior fraction

Nested Sampling basics

Skilling, AIP Conf.Proc. 735, 395 (2004); doi: 10.1063/1.1835238



Define X(λ) as the prior mass associated with likelihood values above λ

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

This is a decreasing function of λ:

$$X(0) = 1 \qquad X(\mathcal{L}_{\rm max}) = 0$$

dX is the prior mass associated with likelihoods [λ, λ+dλ]
An infinitesimal interval dX contributes λdx to the evidence, so that:

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

where L(X) is the inverse of X(λ).

Nested Sampling basic

Suppose that we can evaluate $L_j = L(X_j)$, for a sequence:

$$0 < X_m < \cdots < X_2 < X_1 < 1$$

Then the model likelihood P(d) can be estimated numerically as:

$$P(d) = \sum_{j=1}^{m} w_j L_j$$

with a suitable set of weights, e.g. for the trapezium rule:

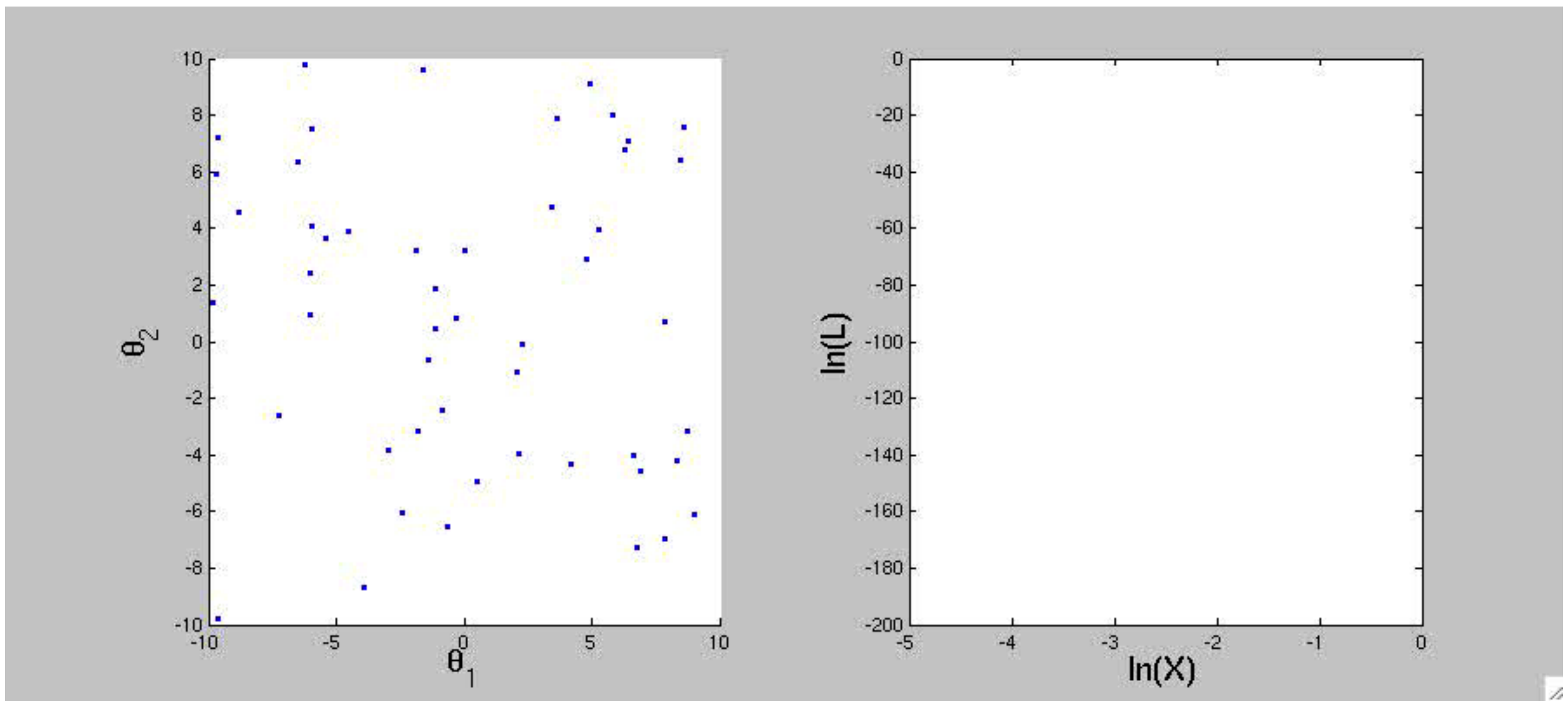
$$w_j = \frac{1}{2}(X_{j-1} - X_{j+1})$$

Nested Sampling in Action

Imperial College London

(animation courtesy of David Parkinson)

2D parameter space (uniform priors) 1D model likelihood integral



Nested sampling pseudo-code

Initialization

- Draw N “live points” from the prior (typically, N ~ 2000)
- Compute the likelihood for each live point

[Loop beings]

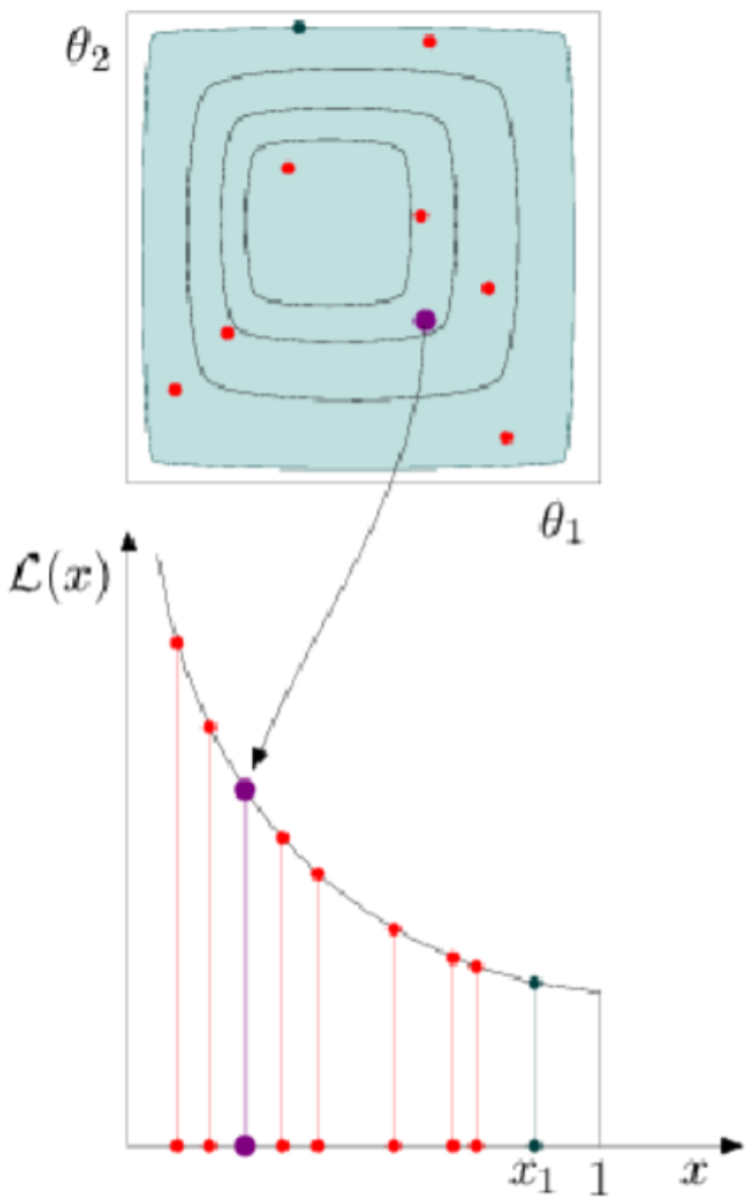
- Select the live point with the lowest likelihood value, μ
- Replace it with a new live point θ drawn from the prior with the constraint L(θ) > μ
- Save the previous live point, together with μ and the prior volume fraction X(μ)
- If L_{max} X < tolerance, exit

[Loop ends]

ICIC

MultiNest sampling approach

(Slide courtesy of Mike Hobson)



Nested sampling approach to summation:

1. Set $i = 0$; initially $X_0 = 1$, $E = 0$
2. Sample N points $\{\boldsymbol{\theta}_j\}$ randomly from $\pi(\boldsymbol{\theta})$ and calculate their likelihoods
3. Set $i \to i + 1$
4. Find point with lowest likelihood value (L_i)
5. Remaining prior volume $X_i = t_i X_{i-1}$ where $\Pr(t_i|N) = N t_i^{N-1}$; or just use $\langle t_i \rangle = N/(N+1)$
6. Increment evidence $E \to E + L_i w_i$
7. Remove lowest point from active set
8. Replace with new point sampled from $\pi(\boldsymbol{\theta})$ within hard-edged region $L(\boldsymbol{\theta}) > L_i$

Hard!

9. If $L_{\max} X_i < \alpha E$ (where some tolerance) $\Rightarrow E \to E + X_i \sum_{j=1}^{N} L(\boldsymbol{\theta}_j)/N$; stop else goto 3

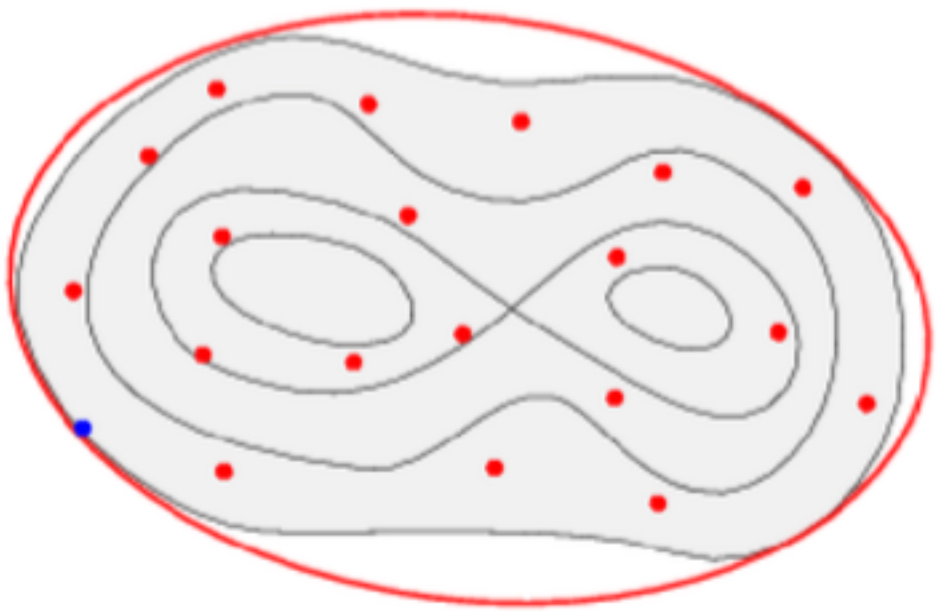
Nested Sampling: Sampling Step

Imperial College London

- The hardest part is to sample uniformly from the prior subject to the hard constraint that the likelihood needs to be above a certain level.
- Many specific implementations of this sampling step:
  - Single ellipsoidal sampling (Mukherjee+06)
  - Metropolis nested sampling (Sivia&Skilling06)
  - Clustered and simultaneous ellipsoidal sampling (Shaw+07)
  - Ellipsoidal sampling with k-means (Feroz&Hobson08)
  - Rejection sampling (MultiNest, Feroz&Hobson09)
  - Diffusive nested sampling (Brewer+11)
  - Artificial neural networks (Graff+12)
  - Galilean Sampling (Betancourt11; Feroz&Skilling13)
  - Simultaneous ellipsoidal sampling with X-means (DIAMONDS, Corsaro&deRidder14)
  - Slice Sampling nested sampling (PolyChord, Handley+15)
  - Dynamic nested sampling (Higson+18)
  - … there will be others, no doubt.

Sampling Step: Ellipsoid Fit

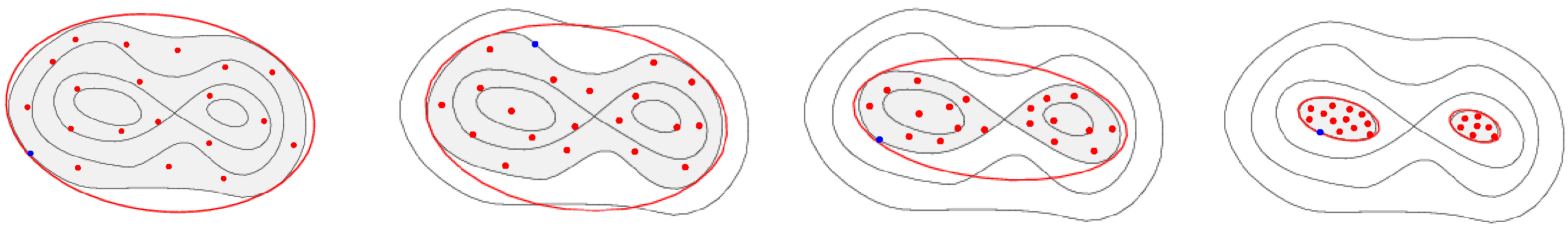
- Simple MCMC (e.g. Metropolis-Hastings) works but can be inefficient
- Mukherjee+06: Take advantage of the existing live points. Fit an ellipsoid to the live point, enlarge it sufficiently (to account for non-ellipsoidal shape), then sample from it using an exact method:



- This works, but is problematic/inefficient for multi-modal likelihoods and/or strong, non-linear degeneracies between parameters.

Sampling Step: Multimodal Sampling

Imperial College London

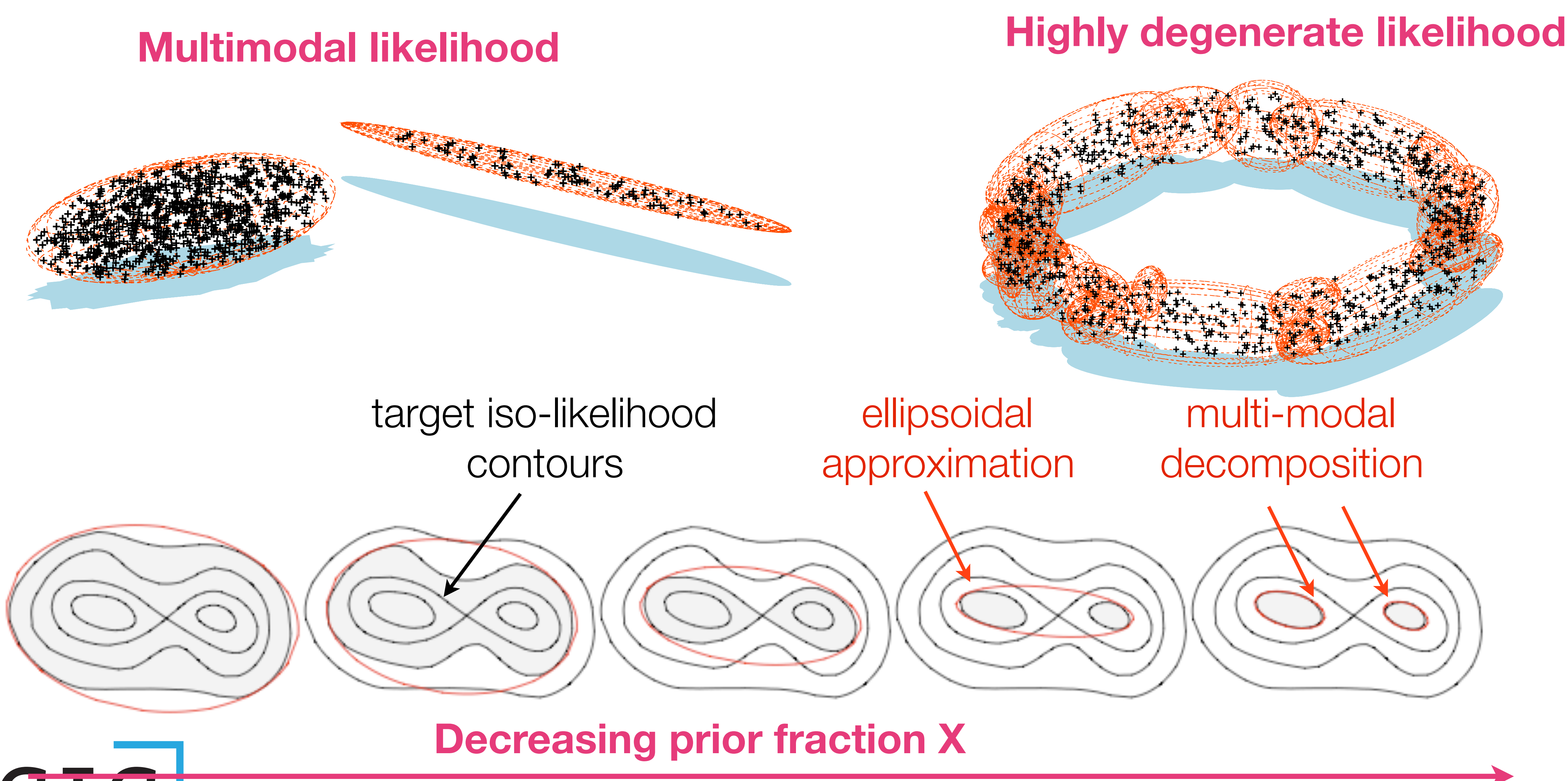


- Feroz&Hobson08; Feroz+08: At each nested sampling iteration
  - Partition active points into clusters
  - Construct ellipsoidal bounds to each cluster
  - Determine ellipsoid overlap
  - Remove point with lowest L_i from active points; increment evidence.
  - Pick ellipsoid randomly and sample new point with $L > L_i$ accounting for overlaps
- Each isolated cluster gives local evidence
- Global evidence is the sum of the local evidences

The MultiNest ellipsoidal sampling

Imperial College London

- The MultiNest algorithm (Feroz & Hobson, 2007, 2008) uses a multi-dimensional ellipsoidal decomposition of the remaining set of “live points” to approximate the prior volume above the target iso-likelihood contour.

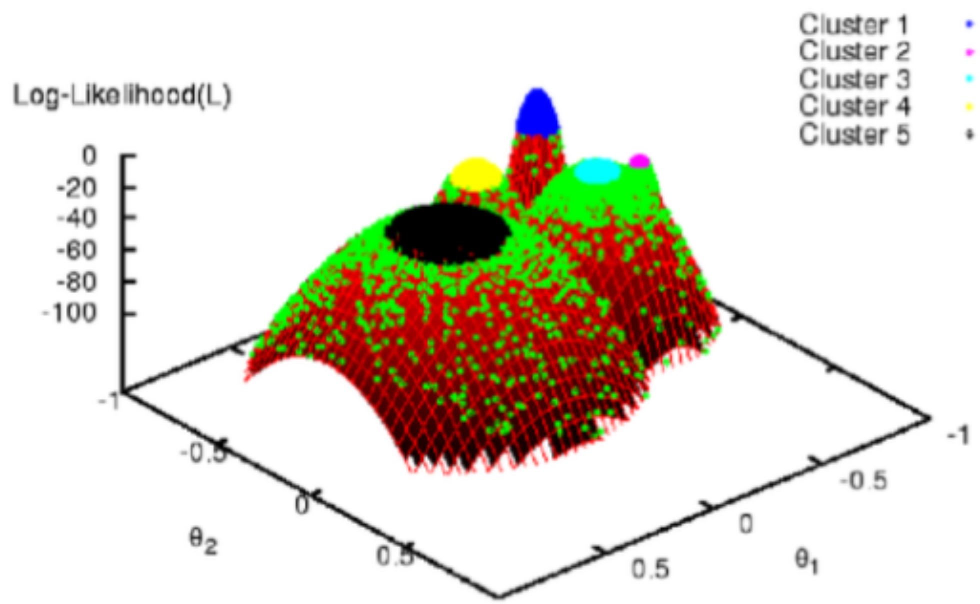


ICIC

Test: Gaussian Mixture Model

Imperial College London

(Slide courtesy of Mike Hobson)



- Likelihood = five 2-D Gaussians of varying widths and amplitudes; prior = uniform
- Analytic evidence integral $\log E = -5.27$
- Multimodal ellipsoidal nested sampling: $\log E = -5.33 \pm 0.11$, $N_{\text{like}} \approx 10^4$
- Metropolis nested sampling: $\log E = -5.22 \pm 0.11$, $N_{\text{like}} \approx 10^5$
- Thermodynamic integration (+ error): $\log E = -5.24 \pm 0.12$, $N_{\text{like}} \approx 4 \times 10^6$

Test: Egg-Box Likelihood

Imperial College London

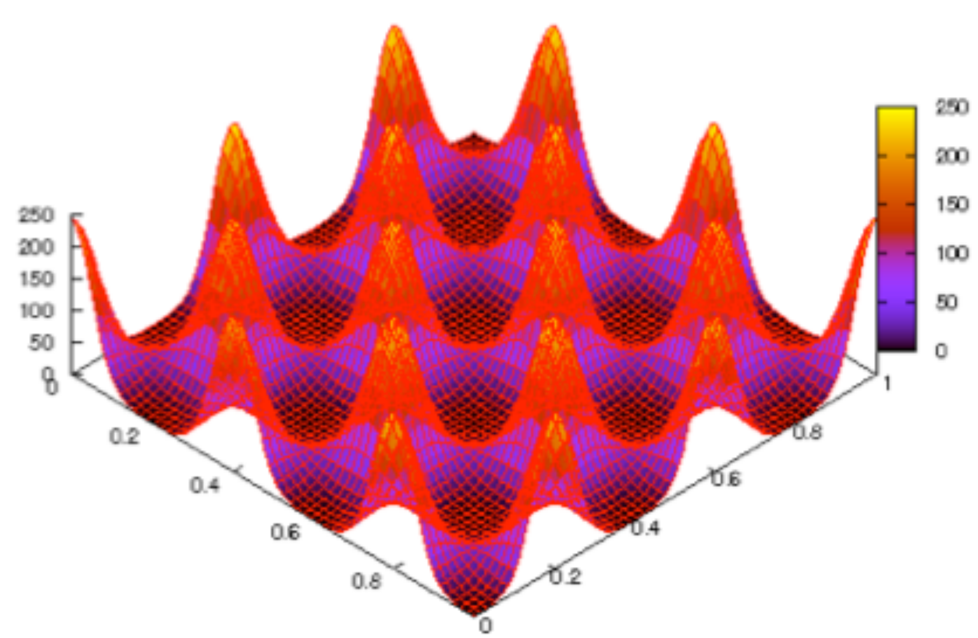
(Animation: Farhan Feroz)

- A more challenging example is the egg-box likelihood:

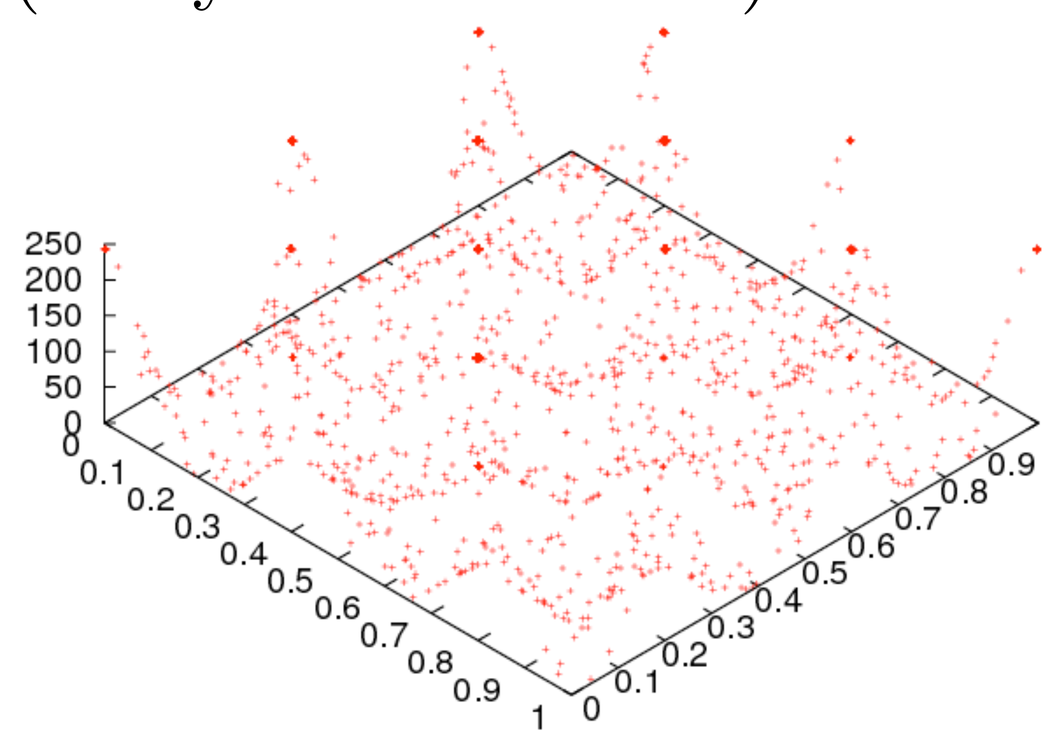
$$\mathcal{L}(\theta_1, \theta_2) = \exp\left(2 + \cos\left(\frac{\theta_1}{2}\right)\cos\left(\frac{\theta_2}{2}\right)\right)^5$$

- Prior: $\theta_i \sim U(0, 10\pi) \quad (i = 1, 2)$

$$\log P(d) = 235.86 \pm 0.06 \quad (\text{analytical} = 235.88)$$



Likelihood

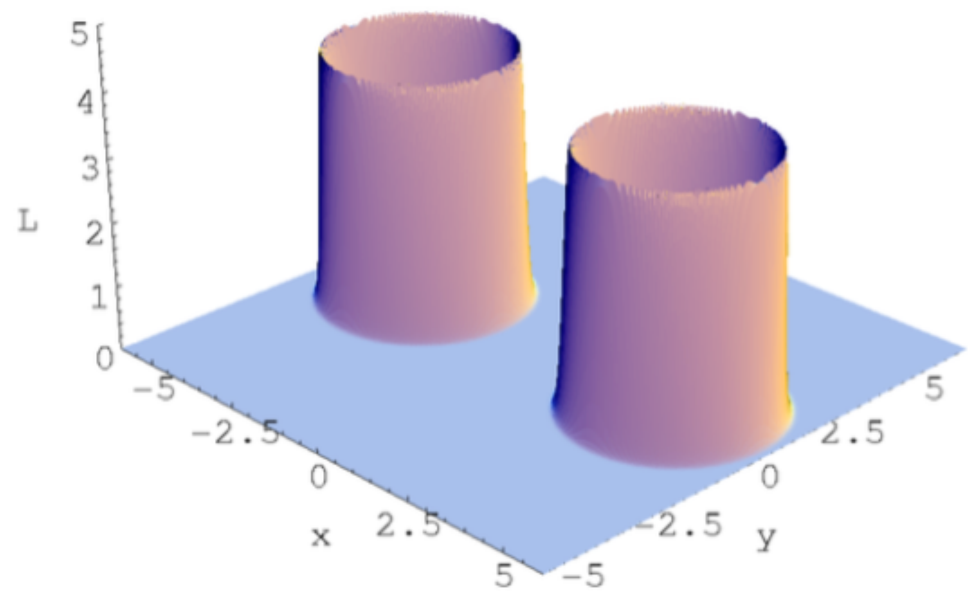


Sampling (30k likelihood evaluations)

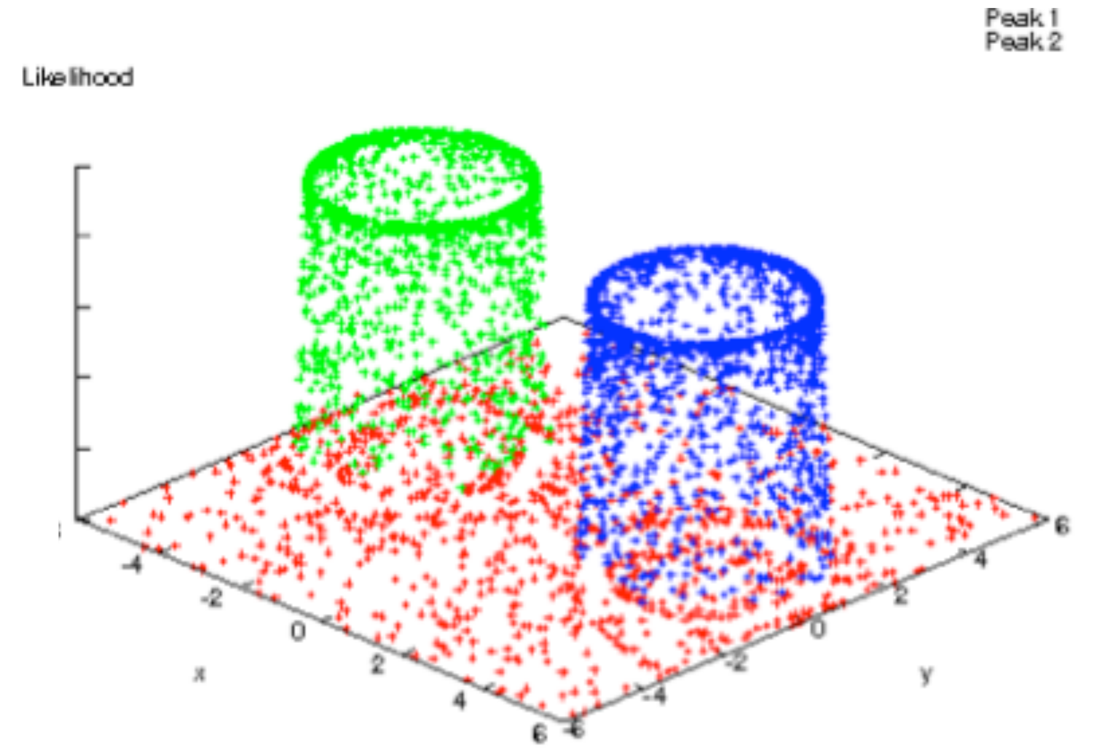
ICIC

Test: Multiple Gaussian Shells

Courtesy Mike Hobson



Likelihood



Sampling

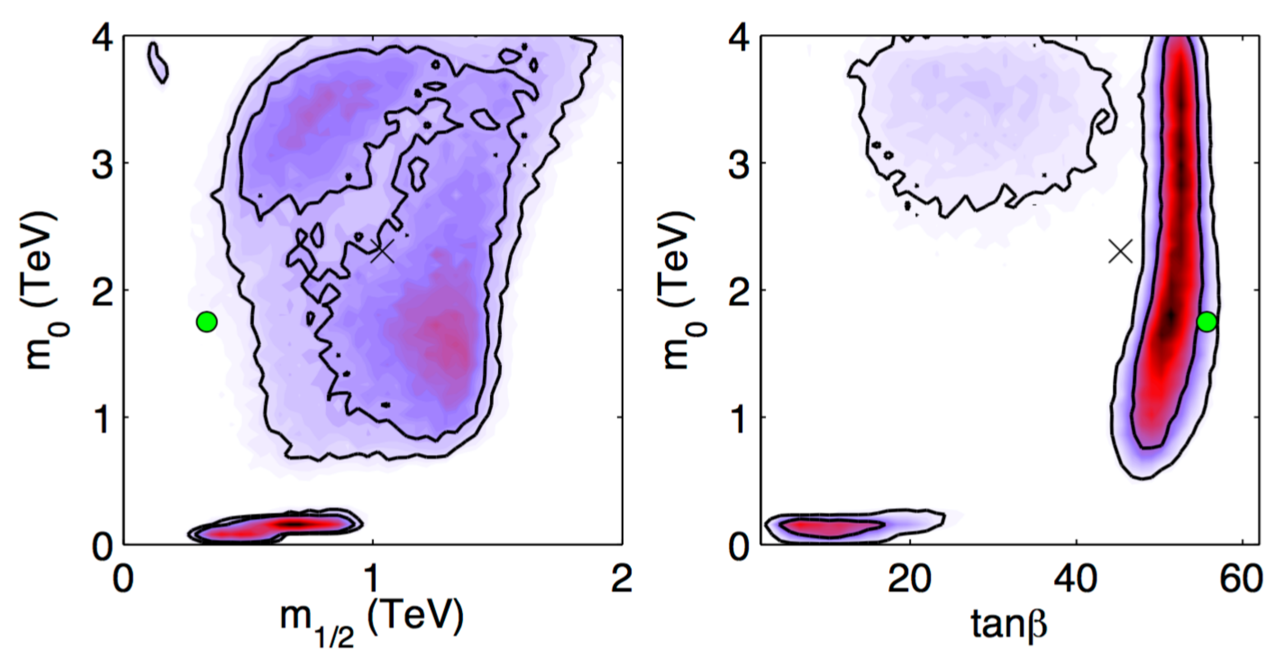
D	N_{like}	Efficiency
2	7000	70%
5	18000	51%
10	53000	34%
20	255000	15%
30	753000	8%

Aside: Posterior Samples

Imperial College London

- Samples from the posterior can be extracted as (free) by-product: take the sequence of sampled points θ_j and weight sample j by $p_j = L_j \omega_j/P(d)$
- MultiNest has only 2 tuning parameters: the number of live points and the tolerance for the stopping criterium (stop if $L_{max} X_i < tol\ P(d)$, where *tol* is the tolerance)
- It can be used (and routinely is used) as fool-proof inference black-box: no need to tune e.g. proposal distribution as in conventional MCMC.

Multi-Modal marginal posterior distributions in an 8D supersymmetric model, sampled with MultiNest (Feroz,RT+11)



(a)

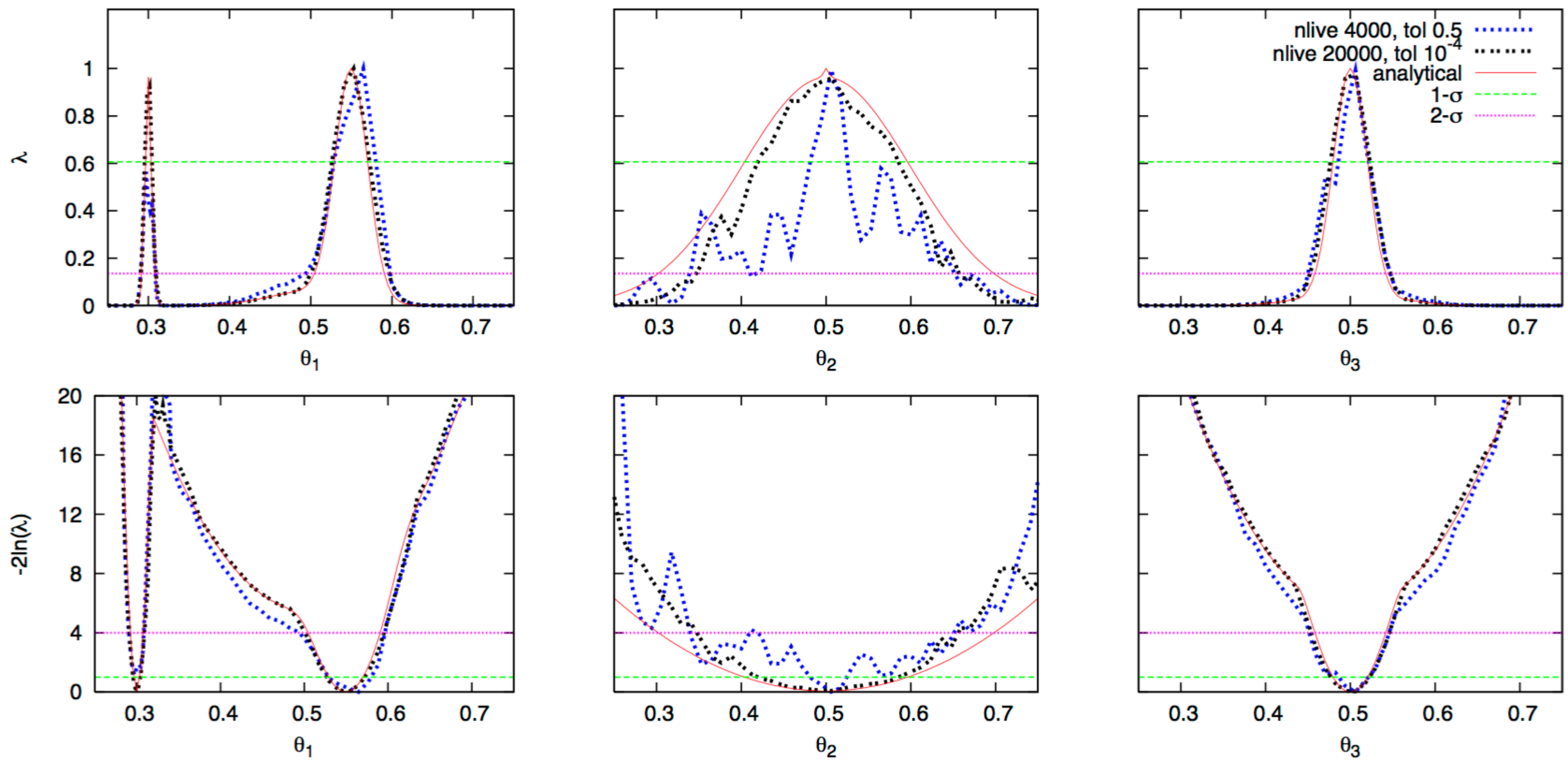
Aside: Profile Likelihood

Imperial College London

- With higher number of live points and smaller tolerance (plus keeping all discarded samples) MultiNest also delivers good profile likelihood estimates (Feroz,RT+11):

8D Gaussian Mixture Model - Profile Likelihood

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



Parallelisation and Efficiency

Imperial College London

- Sampling efficiency is less than unity since ellipsoidal approximation to the iso-likelihood contour is imperfect and ellipsoids may overlap

- **Parallel solution:**

  - At each attempt to draw a replacement point, drawn N_{CPU} candidates, with optimal number of CPUs given by $1/N_{CPU}$ = efficiency

- **Limitations:**

  - Performance improvement plateaus for $N_{CPU} >> 1$/efficiency

  - For D>>30, small error in the ellipsoidal decomposition entails large drop in efficiency as most of the volume is near the surface

  - MultiNest thus (fundamentally) limited to D <= 30 dimensions

Neural Network Acceleration

Imperial College London

Graff+12 (BAMBI) and Graff+14 (SkyNet); Johannesson,RT+16

- A relatively straightforward idea: Use MultiNest discarded samples to train on-line a multi-layer Neural Network (NN) to learn the likelihood function.
- Periodically test the accuracy of predictions: when the NN is ready, replace (possibly expensive) likelihood calls with (fast) NN prediction.
- **SkyNet:** a feed-forward NN with N hidden layers, each with M_n nodes.
- **BAMBI** (Blind Accelerated Multimodal Bayesian Inference): SkyNet integration with MultiNest
- In cosmological applications, BAMBI typically accelerates the model likelihood computation by ~30% — useful, but not a game-changer.
- Further usage of the resulted trained network (e.g. with different priors) delivers speed increases of a factor 4 to 50 (limited by error prediction calculation time).

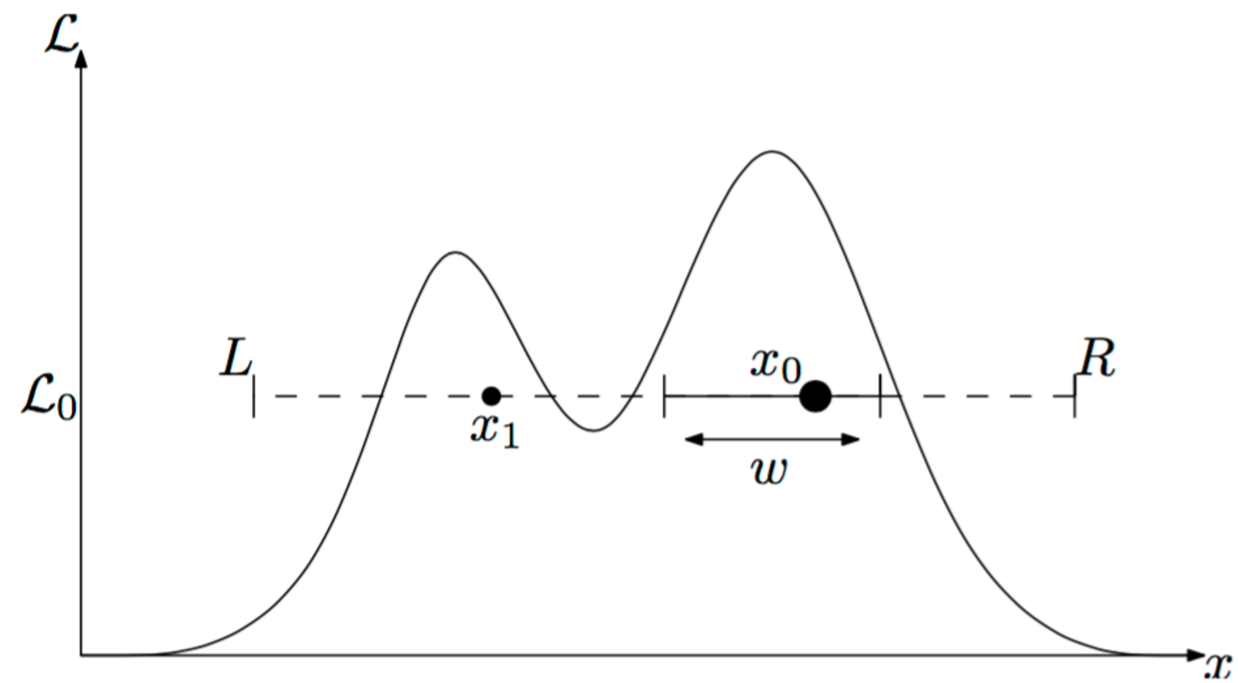
PolyChord: Nested Sampling in high-D

Imperial College London

Handley et al, Mon.Not.Roy.Astron.Soc. 450 (2015)1, L61-L65

- A new sampling step scheme is required to beat the limitations of the ellipsoidal decomposition at the heart of MultiNest

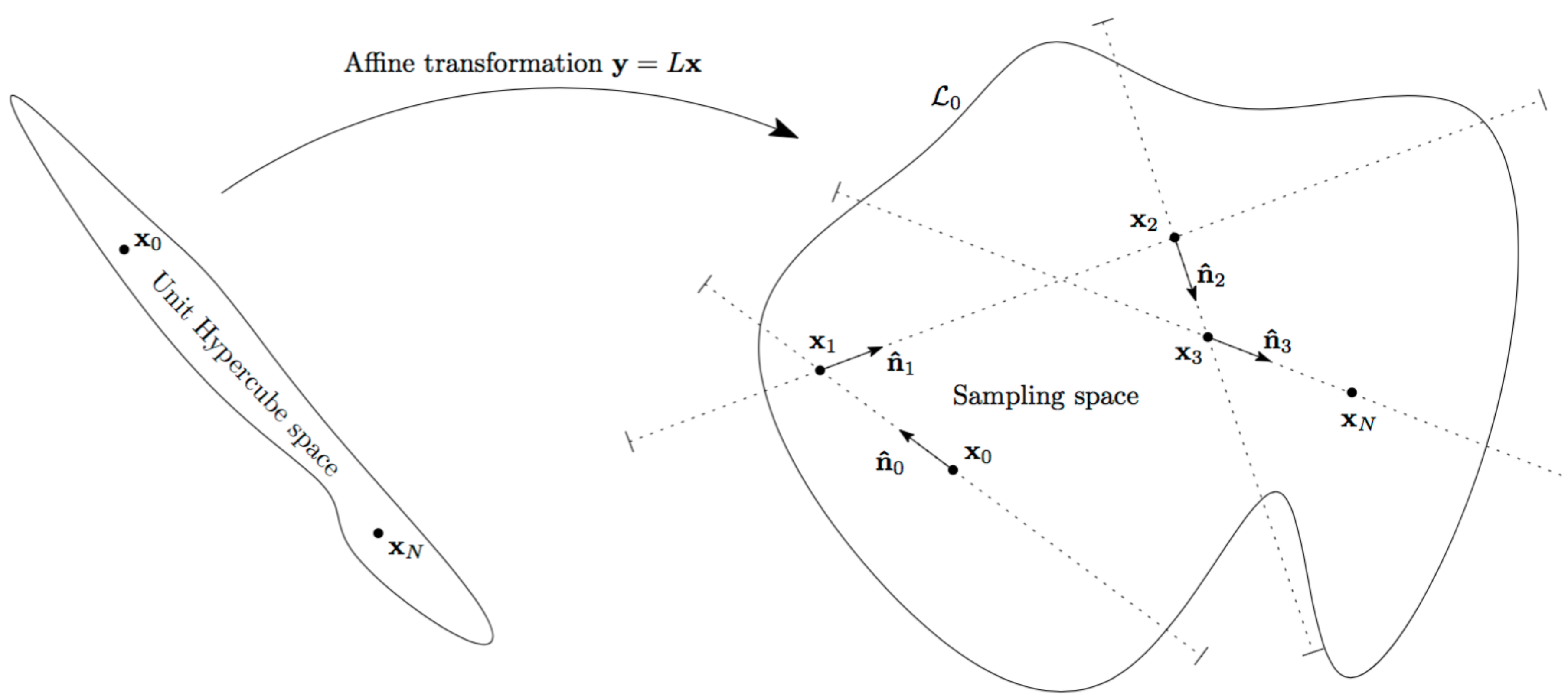
- **Slice Sampling (Neal00) in 1D:**
  - Slice: All points with $L(x)>L_0$
  - From starting point x_0, set initial bounds L/R by expanding from a parameter w
  - Draw x_1 randomly from within L/R
  - If x_1 not in the slice, contract bound down to x_1 and re-sample x_1



High-D Slice Sampling

Imperial College London

- A degenerate contour is transformed into a contour with dimensions of order O(1) in all directions ("whitening")
- Linear skew transform defined by the inverse of the Cholesky decomposition of the live points' covariance matrix
- Direction selected at random, then slice sampling in 1D performed (w=1)
- Repeat N times, with N of order O(D), generating a new point x_N decorrelated from x_0

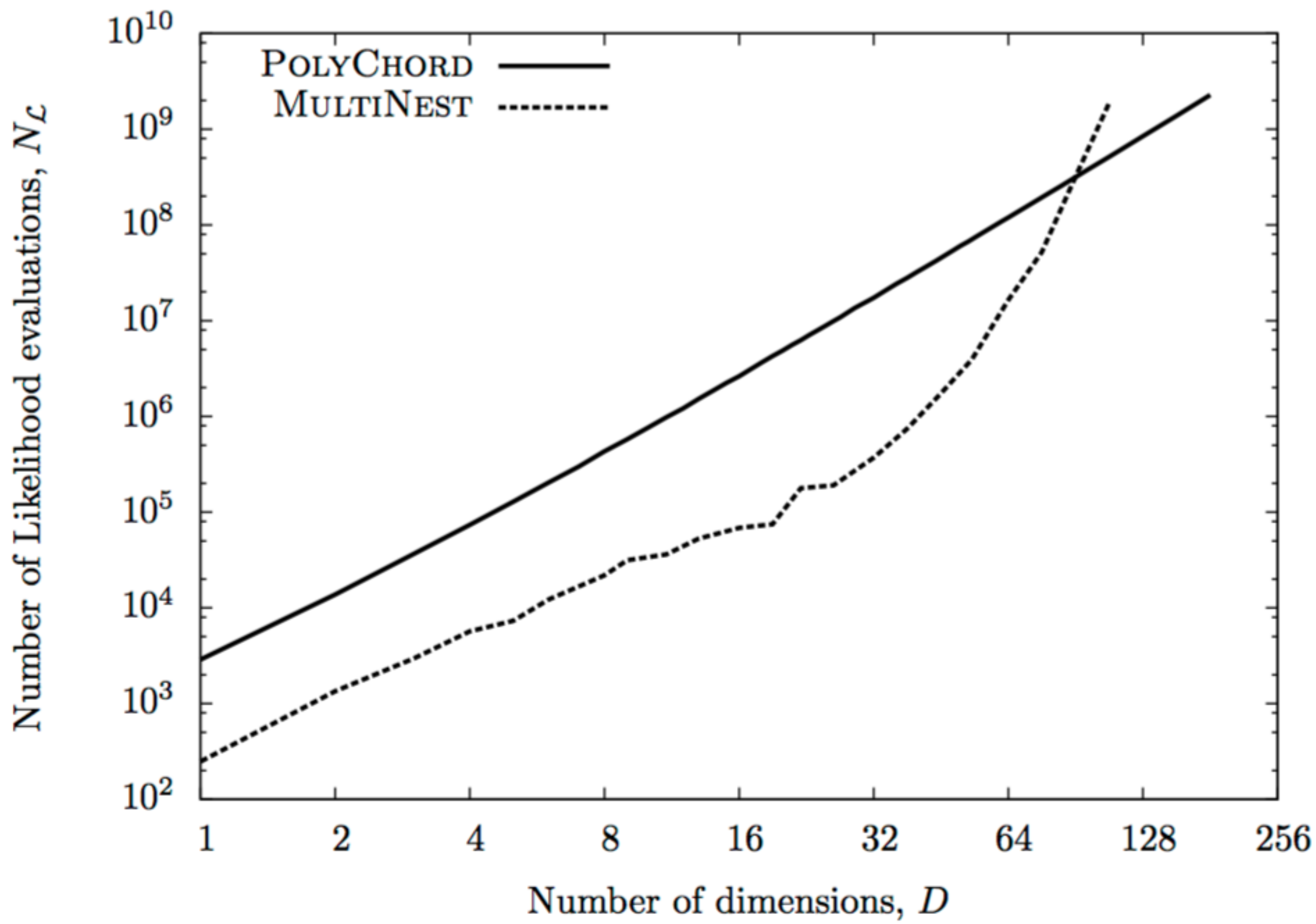


Handley+15

PolyChord: Performance

Imperial College London

- PolyChord number of likelihood evaluations scales at worst as $O(D^3)$ as opposed to exponential for MultiNest in high-D



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

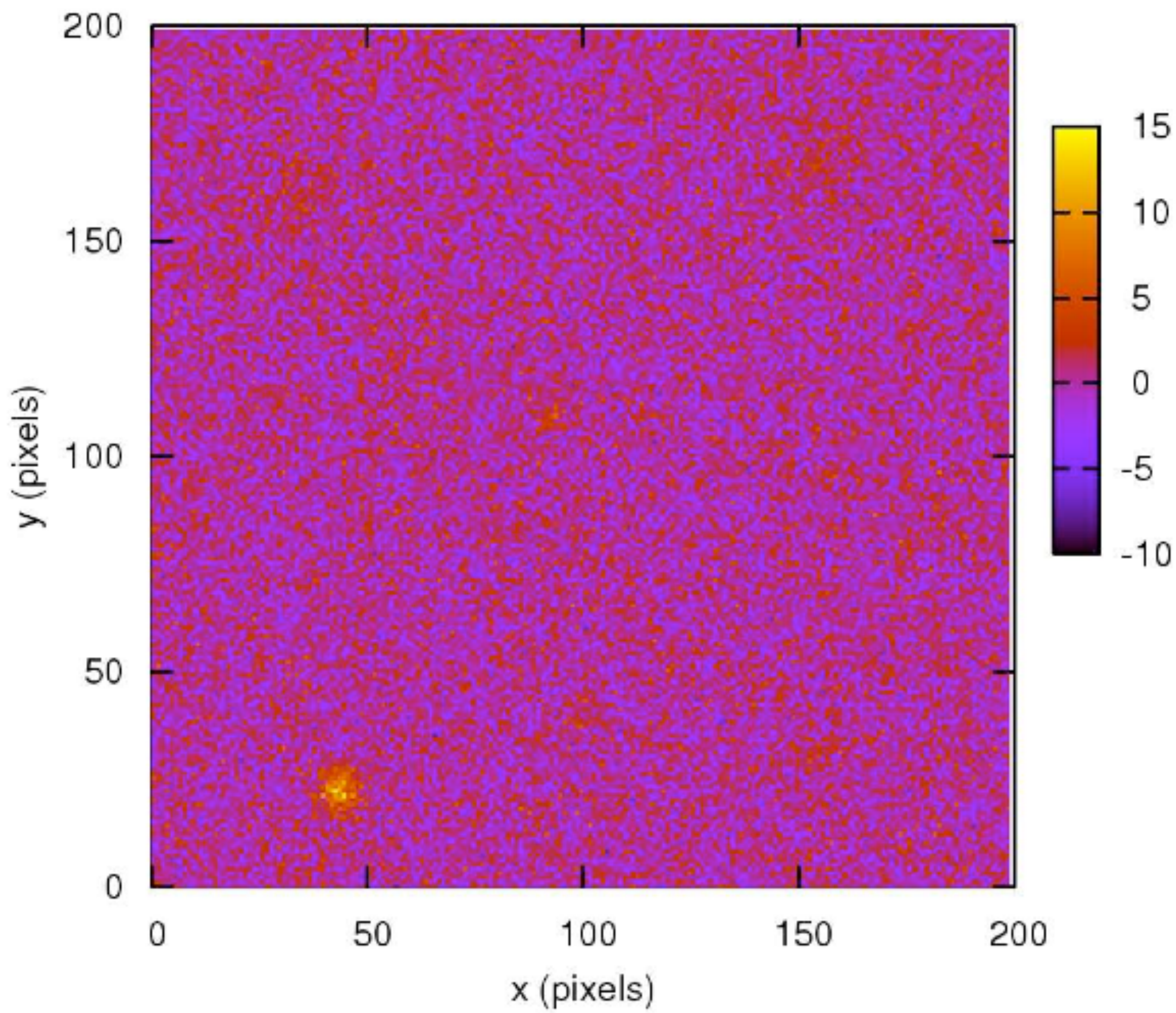
- The best model is the one which minimizes the AIC/BIC/DIC
- **Warning:** AIC and BIC penalize models differently as a function of the number of data points N.
  For N>7 BIC has a more strong penalty for models with a larger number of free parameters k.
- BIC is an approximation to the full Bayesian 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).

A “simple” example: how many sources?

Imperial College London

Feroz and Hobson (2007)

Signal + Noise

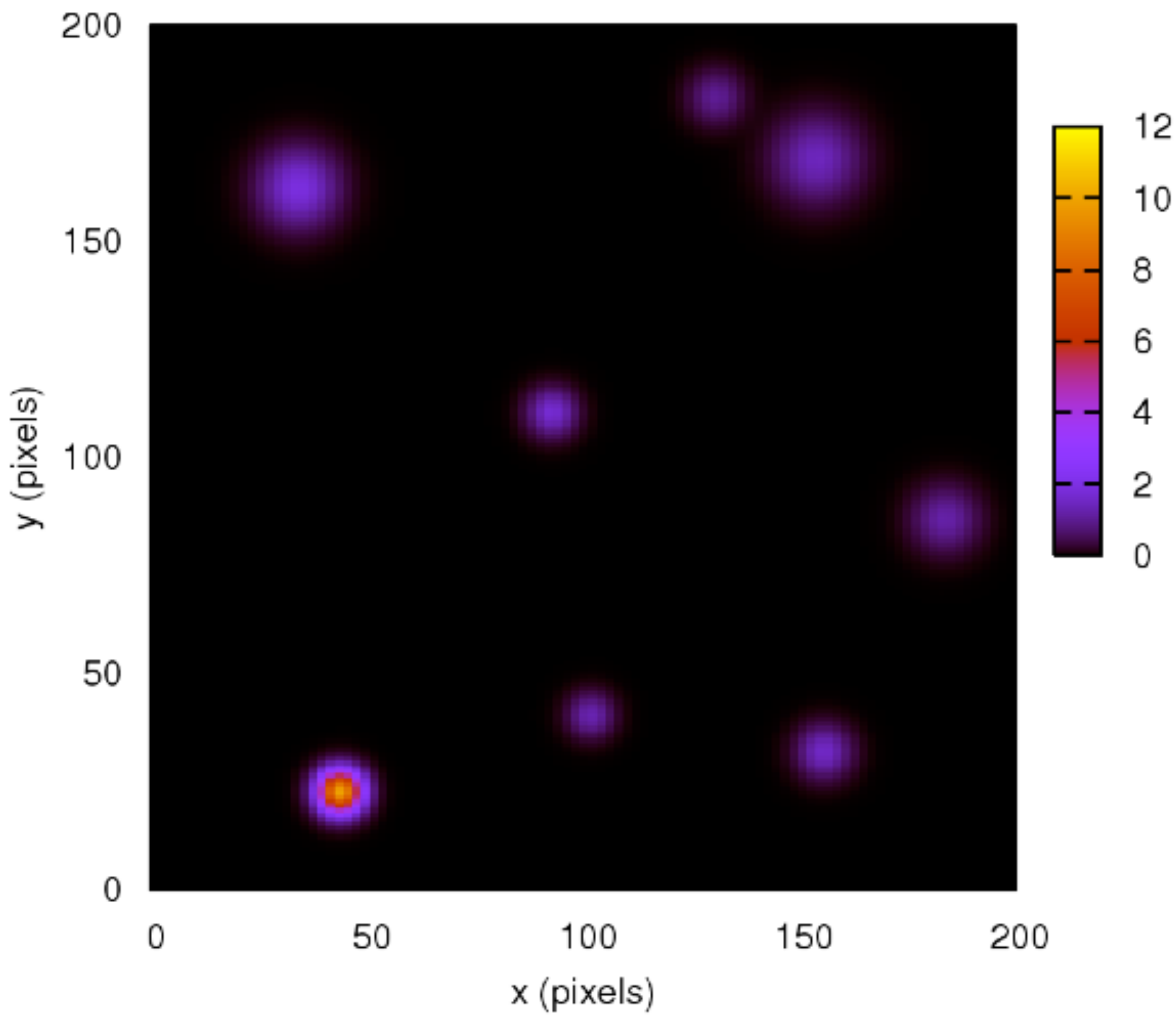


A “simple” example: how many sources?

Imperial College London

Feroz and Hobson (2007)

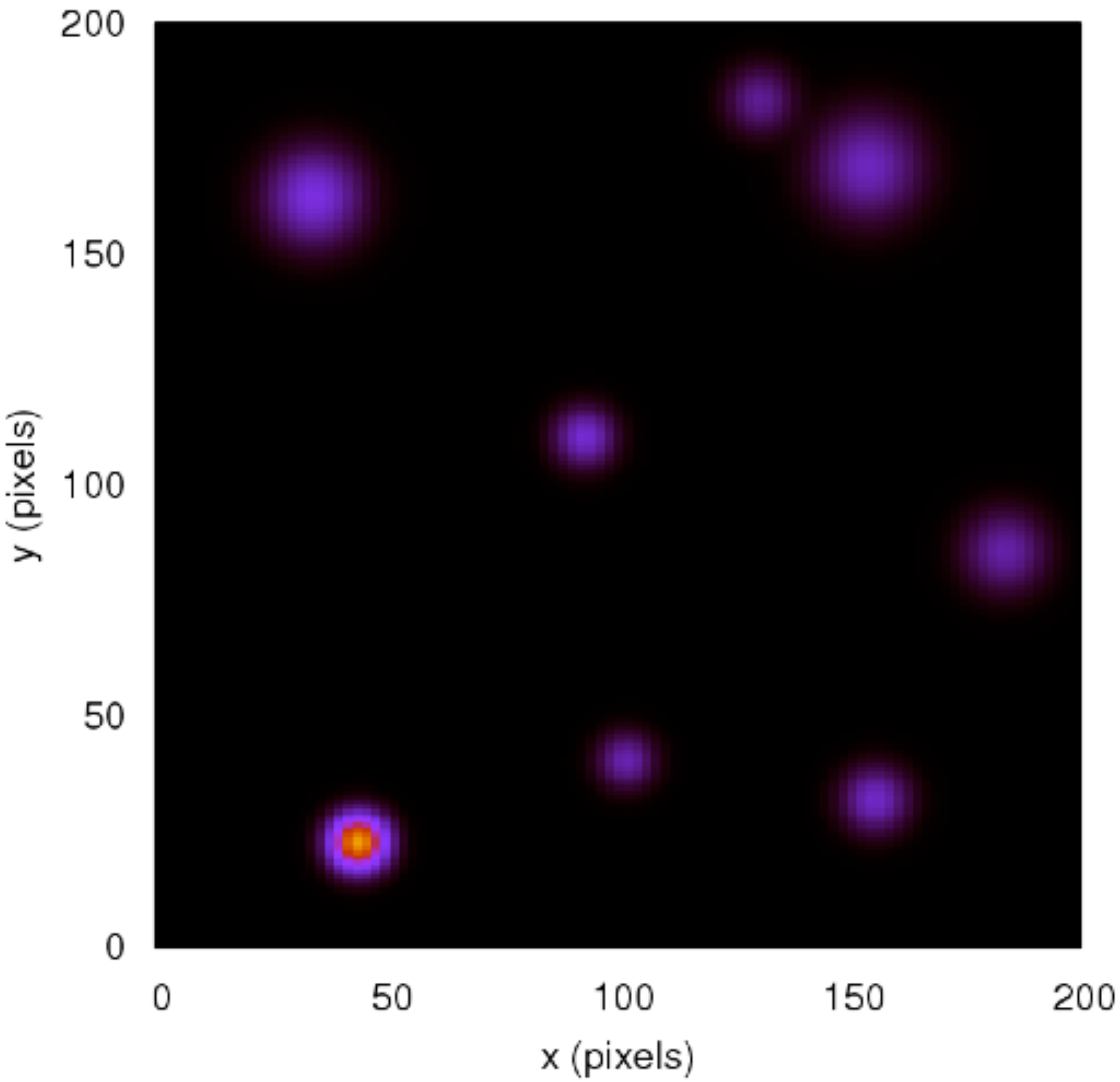
Signal: 8 sources



A “simple” example: how many sources?

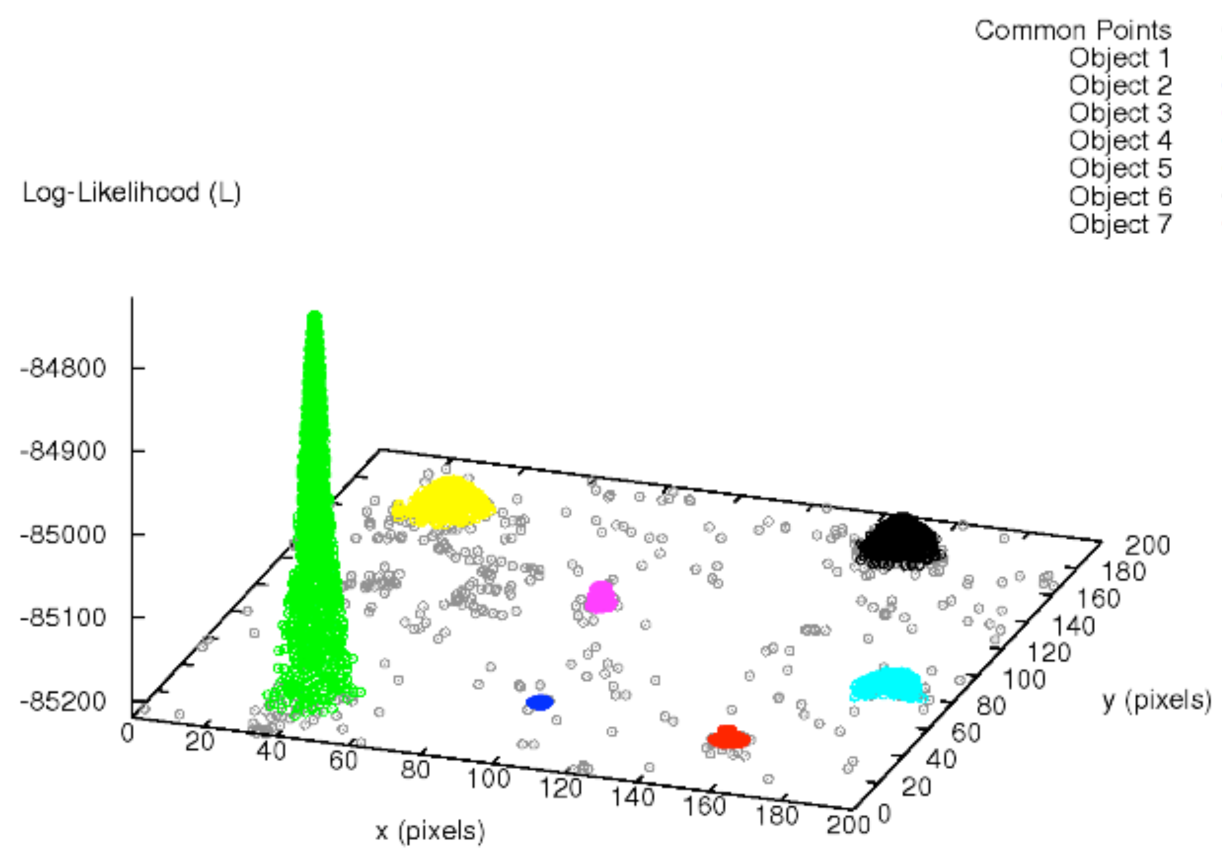
Imperial College London

Feroz and Hobson (2007)

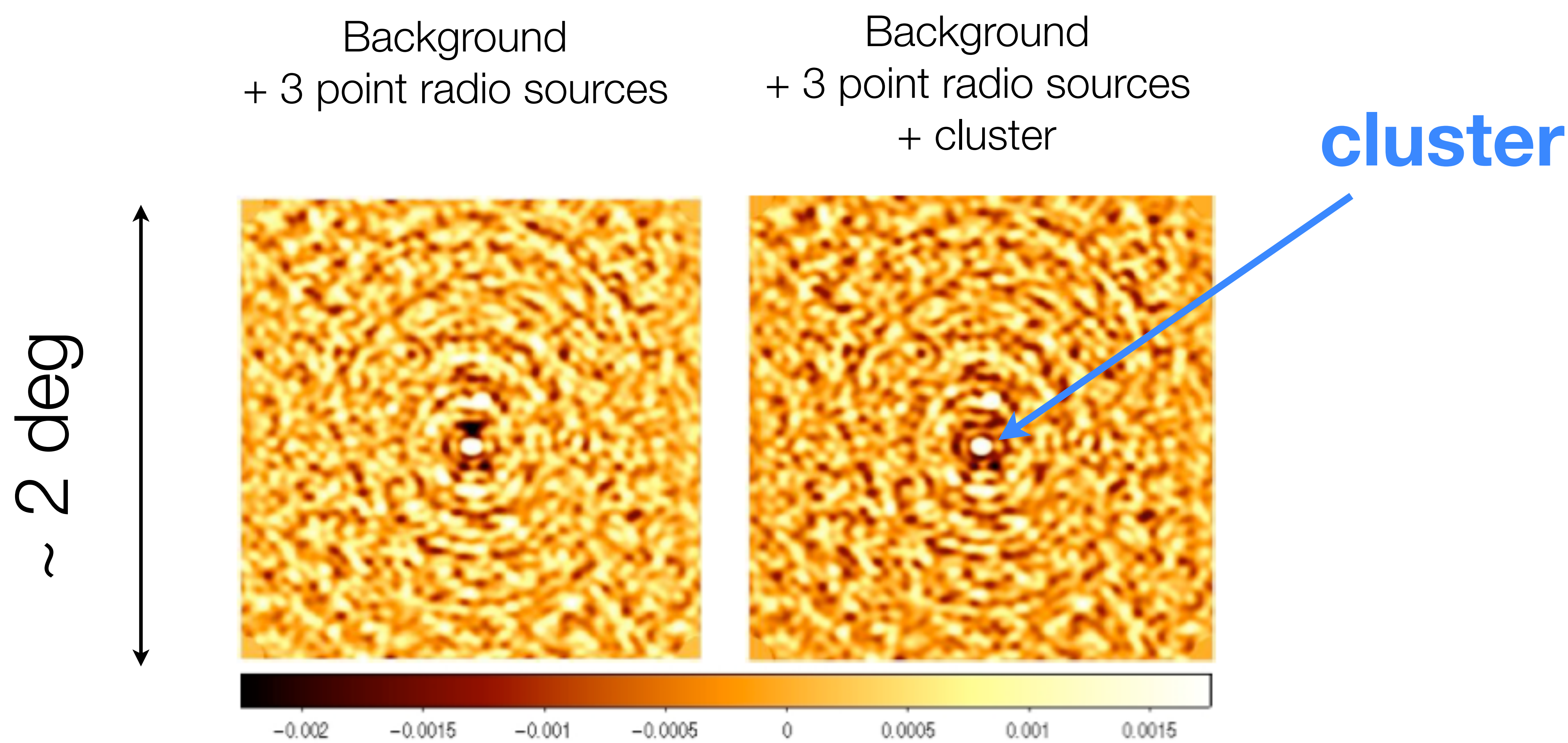


Bayesian reconstruction

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



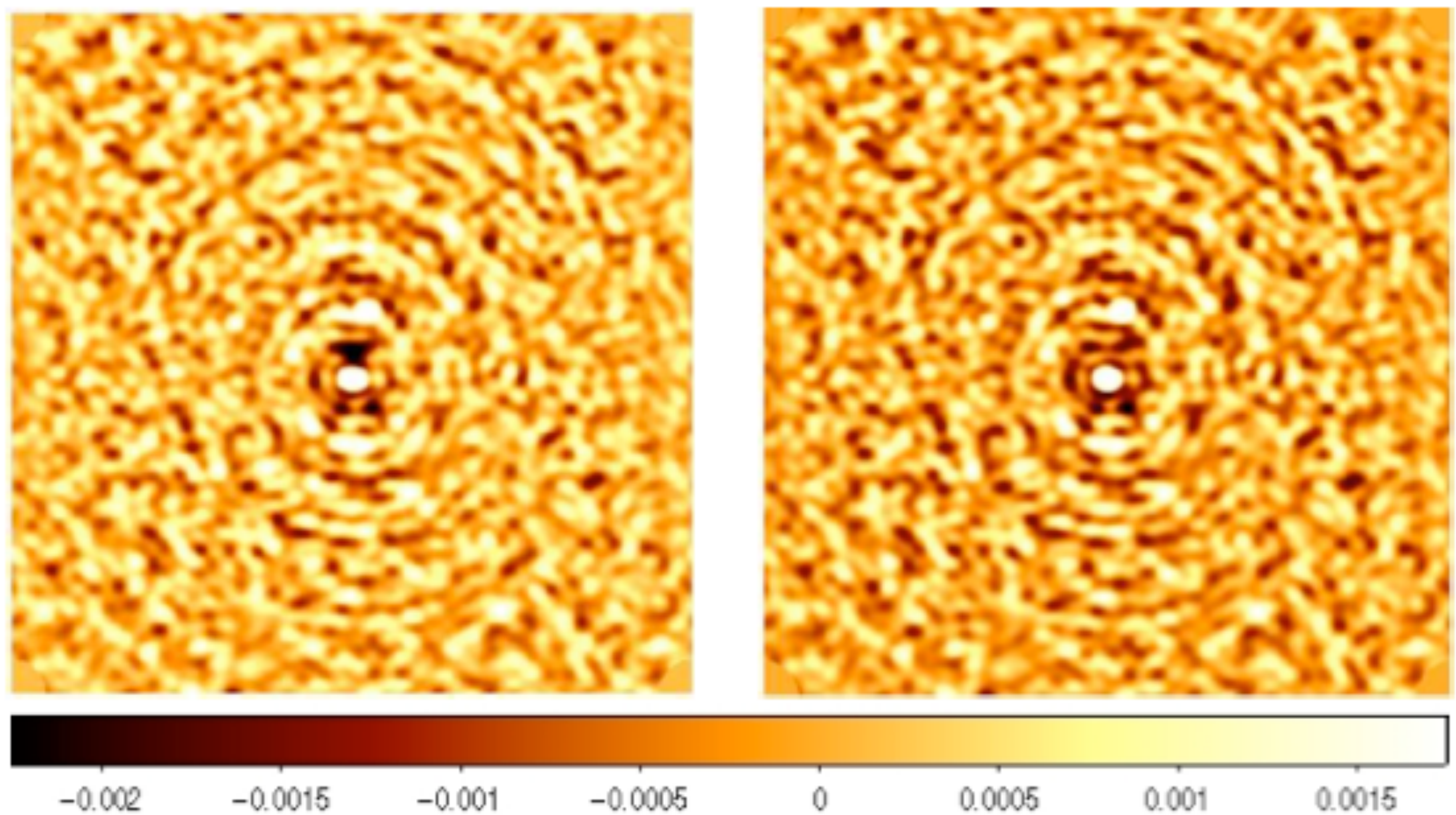
Cluster detection from Sunyaev-Zeldovich effect in cosmic microwave background maps



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)

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

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

The cosmological concordance model

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)

lnB < 0: favours ΛCDM

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:

$$\begin{aligned}\mathcal{C}_b &= \overline{\chi^2(\theta)} - \chi^2(\widehat{\theta}) \\ &= \sum_i \frac{1}{1+(\sigma_i/\Sigma_i)^2}\end{aligned}$$

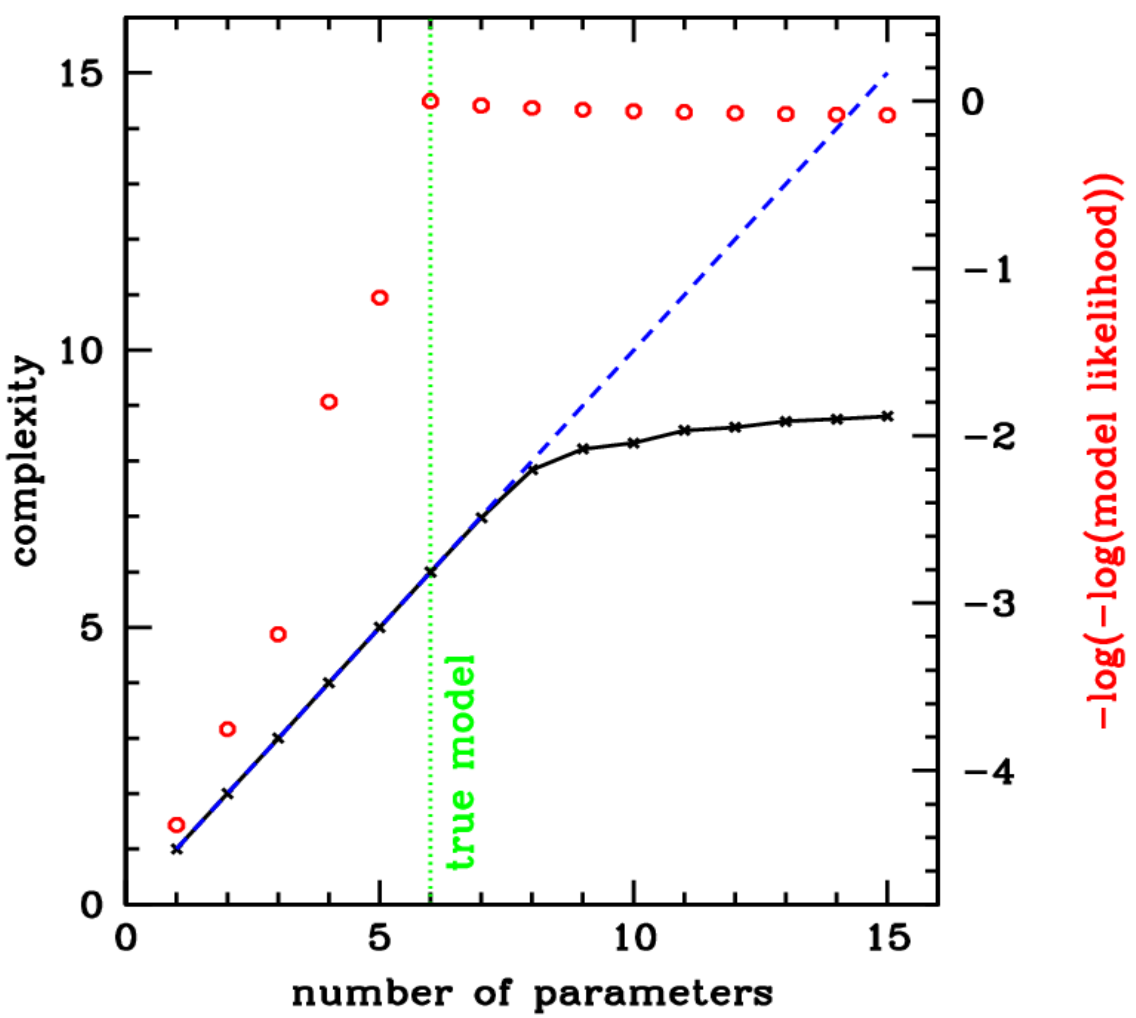
Kunz, RT & Parkinson, astro-ph/0602378, Phys. Rev. D 74, 023503 (2006)
Following Spiegelhalter et al (2002)

Polynomial fitting

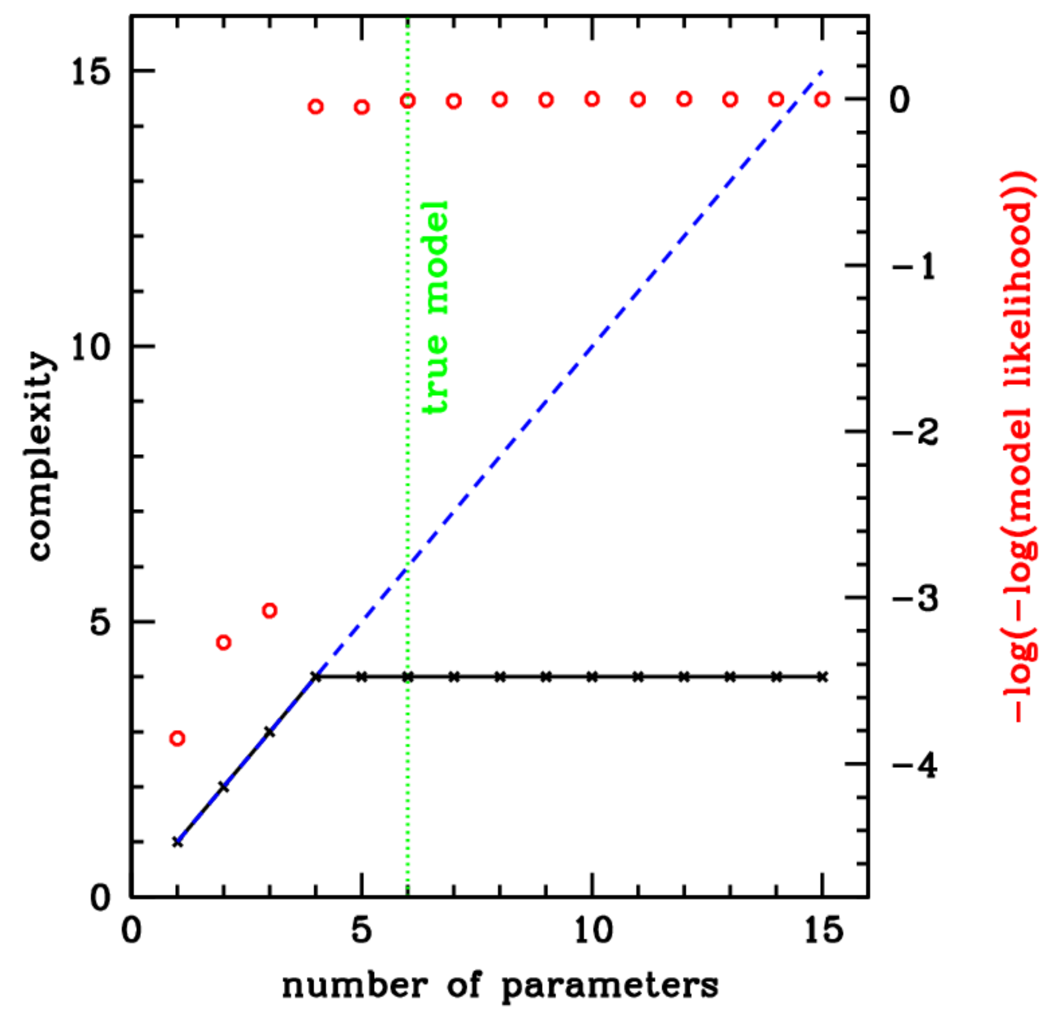
Imperial College London

- Data generated from a model with n = 6:

GOOD DATA
Max supported complexity ~ 9



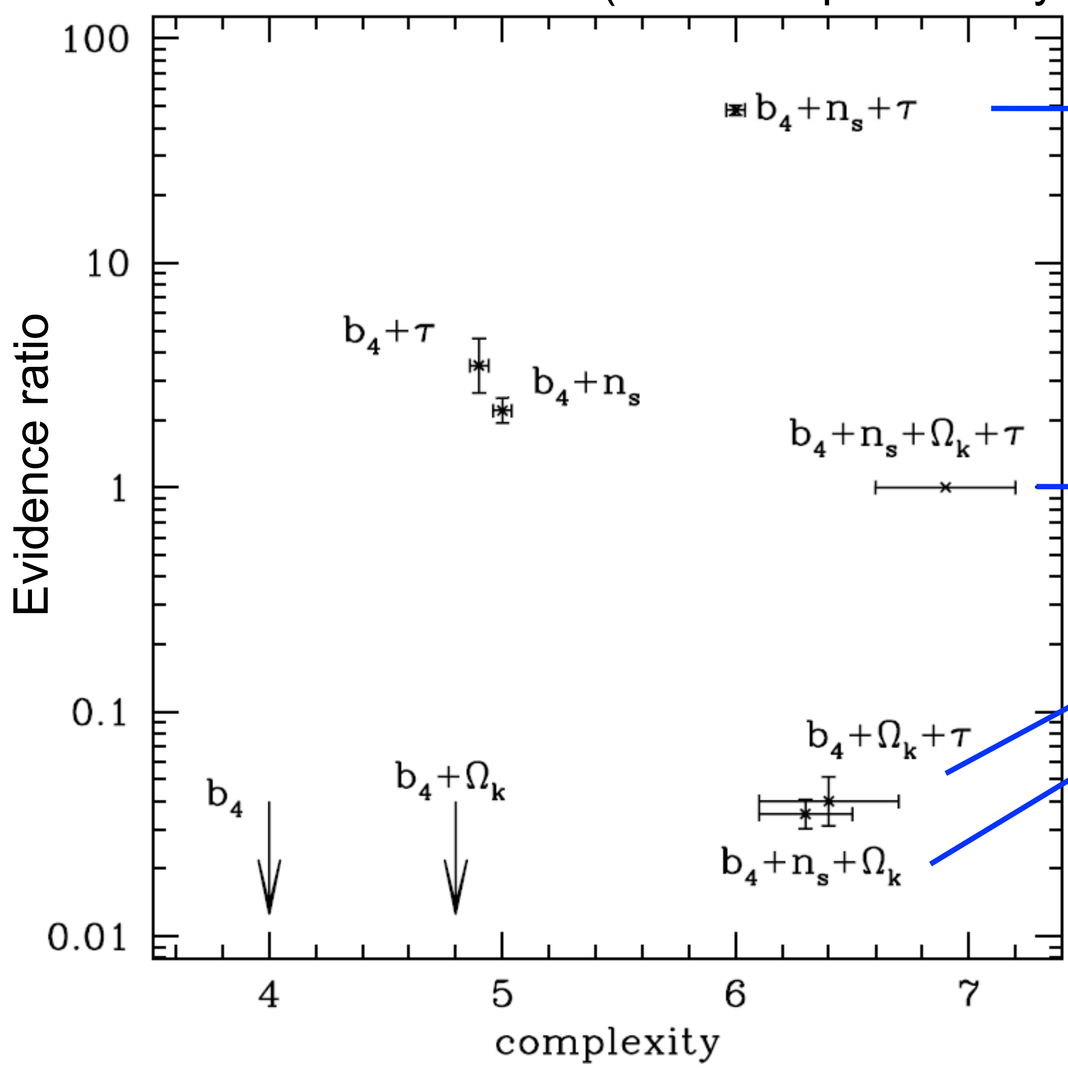
INSUFFICIENT DATA
Max supported complexity ~ 4



How many parameters does the CMB need?

Imperial College London

WMAP3+HST (WMAP5 qualitatively the same)



$b_4+n_s+\tau$ measured & favoured

Ω_κ measured & unnecessary

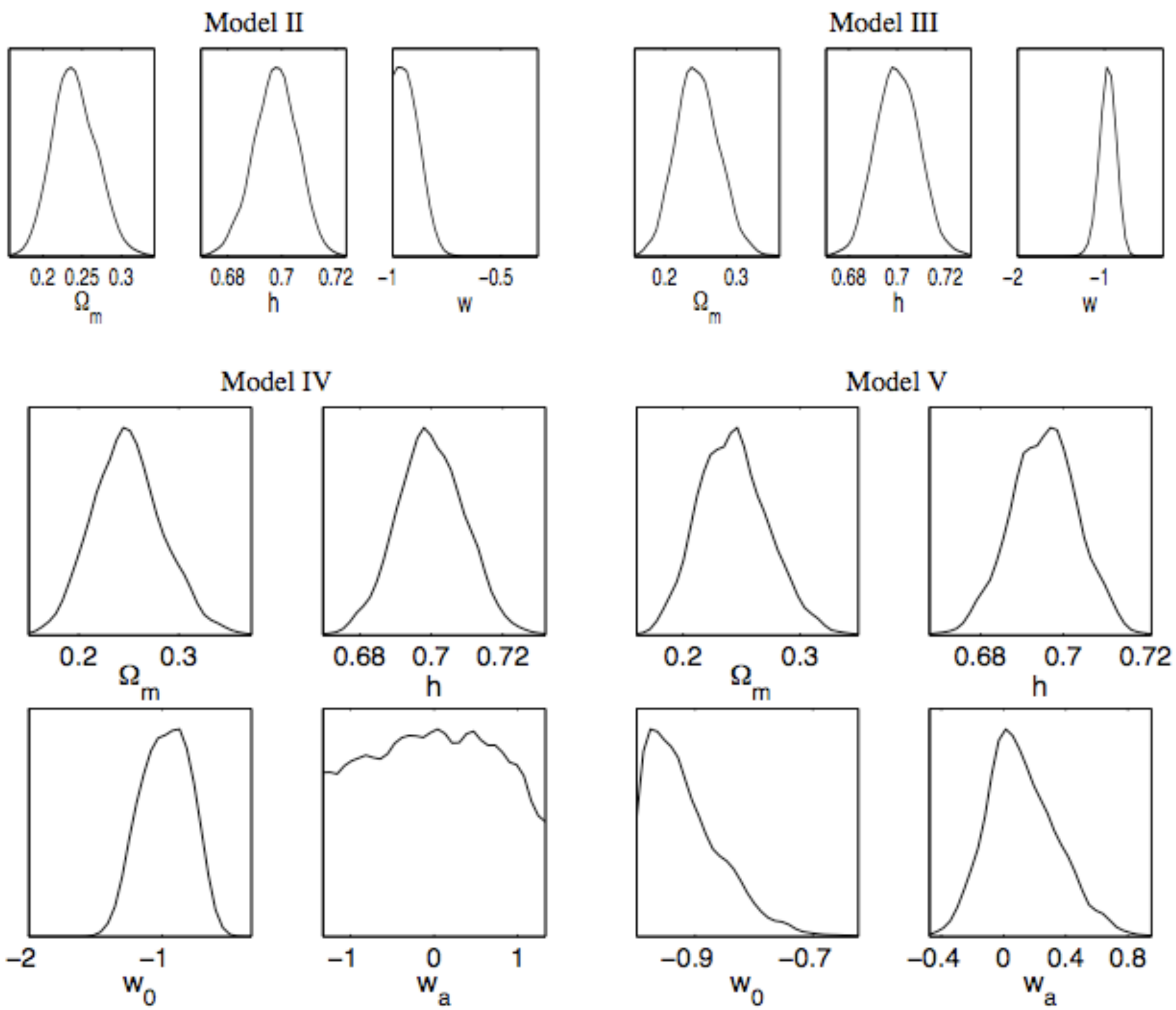
7 params measured

only 6 sufficient

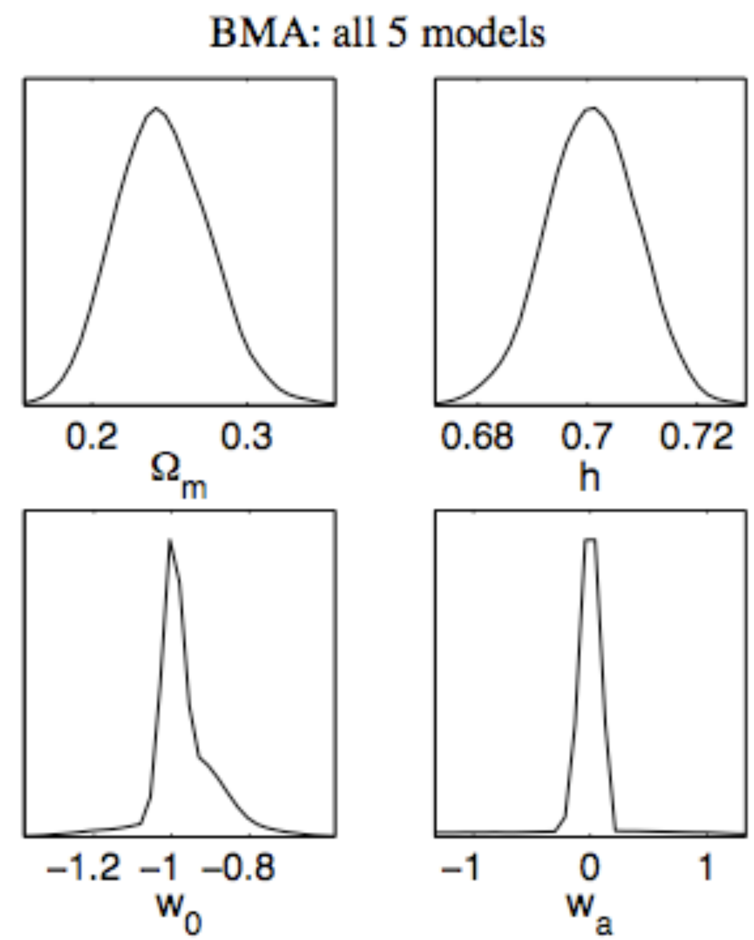
Bayesian Model-averaging

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

An application to dark energy:



Model averaged inferences



Liddle et al (2007)

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.