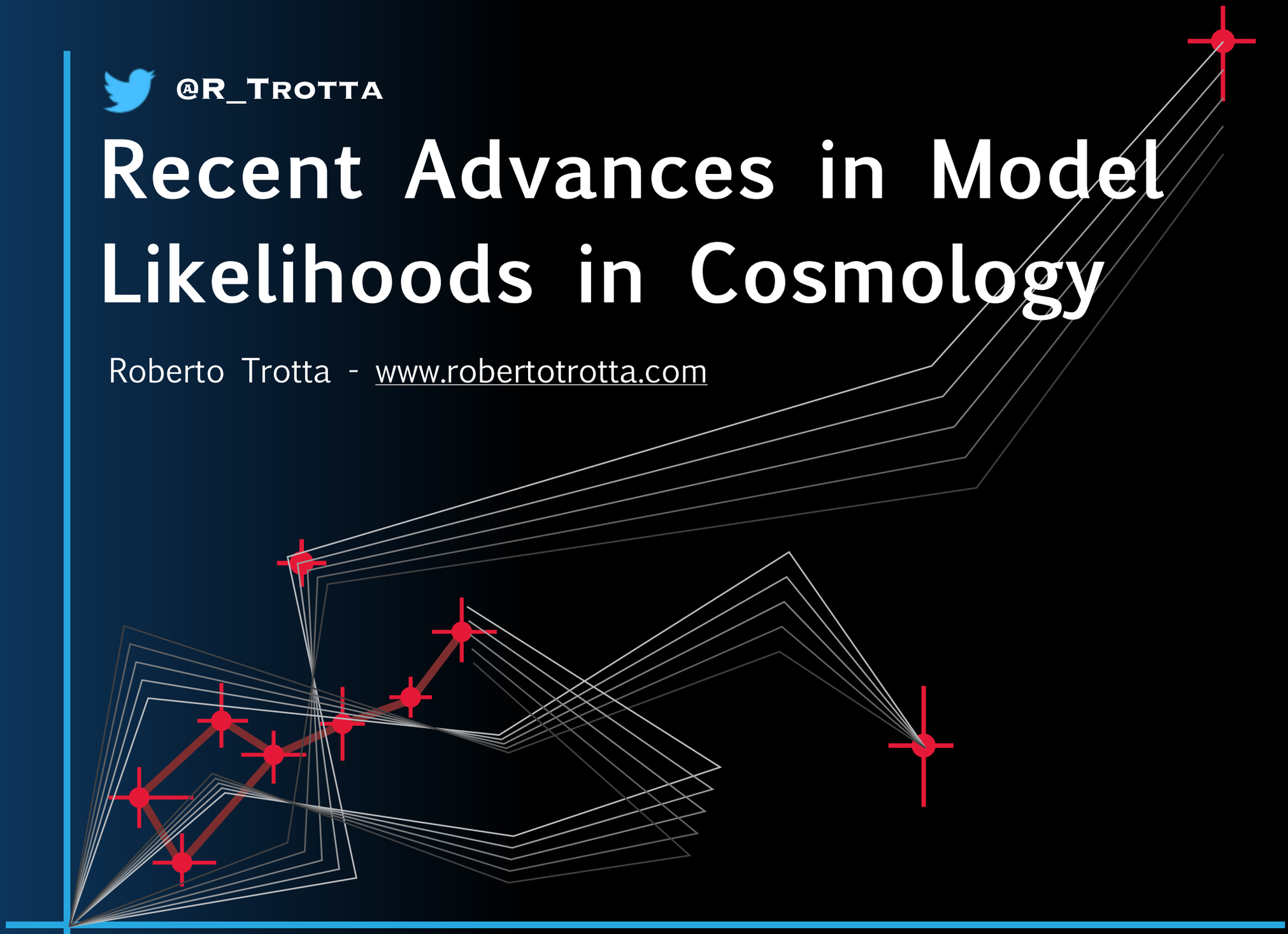
@R_Trotta

Recent Advances in Model Likelihoods in Cosmology

Roberto Trotta - www.robertotrotta.com



ICIC
Imperial Centre
for Inference & Cosmology

CRiSM Estimating Constants Workshop
Warwick, Apr 20th 2016

Imperial College
London

Contents

Imperial College London

The cosmological concordance model

Imperial College London

The ΛCDM cosmological concordance model is built on three pillars:

1. **INFLATION:**
   A burst of exponential expansion in the first ~10^{-32} s after the Big Bang, probably powered by a yet unknown scalar field.
2. **DARK MATTER:**
   The growth of structure in the Universe and the observed gravitational effects require a massive, neutral, non-baryonic yet unknown particle making up ~25% of the energy density.
3. **DARK ENERGY:**
   The accelerated cosmic expansion (together with the flat Universe implied by the Cosmic Microwave Background) requires a smooth yet unknown field with negative equation of state, making up ~70% of the energy density.

The next 5 to 10 years are poised to bring major observational breakthroughs in each of those topics!

Radiation era

Dark matter era

Dark energy era

Big Bang

TODAY

End of the visible cosmos

SN Type Ia

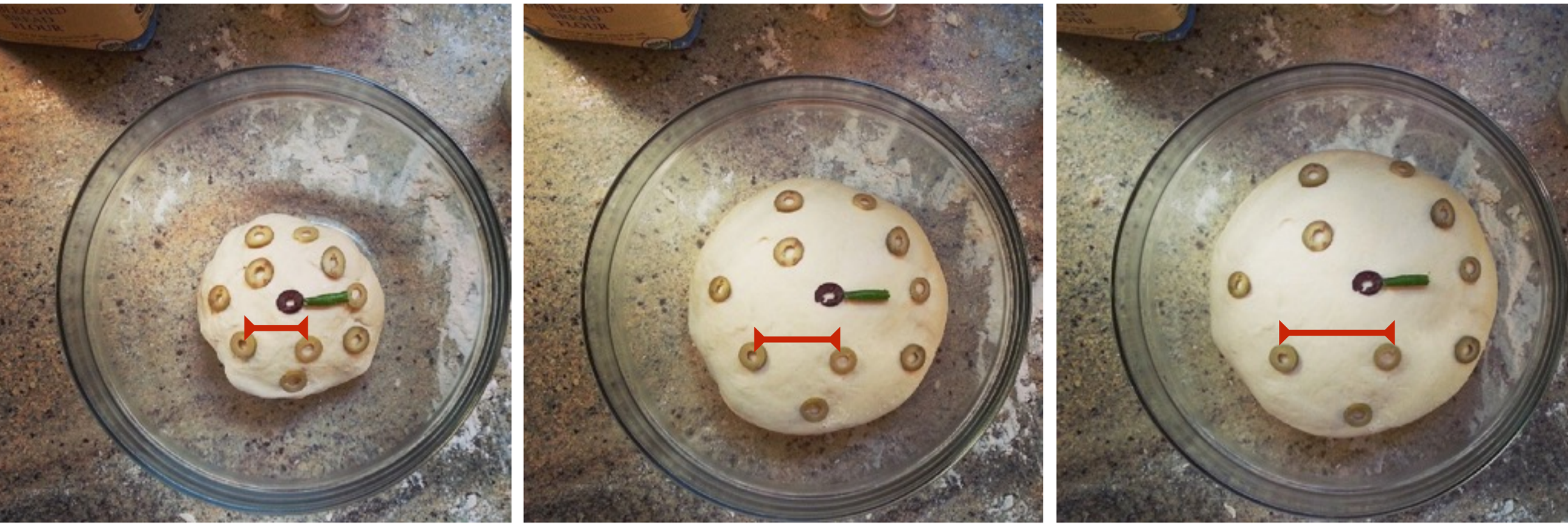
time

NASA/WMAP Science

Cosmography

Imperial College London

The expansion history of the (isotropic, homogeneous) Universe is described by the "scale factor" a(t):



time

Physical separation = a(t) × "coordinate distance"

The cosmological parameters

Imperial College London

The scale factor a(t) is the solution to an ODE containing a number of free parameters: the **"cosmological parameters"**

The cosmological parameters need to be measured observationally. They describe the past history of the Universe and how it will expand in the future.

Dark matter: $\Omega_{\rm m} = 0.315 \pm 0.017$

possibly a new particle beyond the Standard Model, interacting via gravity and weak interaction.

$$\Omega_\Lambda = 0.686 \pm 0.020$$

Dark energy:

a form of vacuum energy with repulsive effect.
Compatible with Einstein's cosmological constant.

Spatial curvature: $\Omega_\kappa = 0.0005 \pm 0.0060$

the Universe is flat, as predicted by the model of inflation.

The 3 levels of inference

Imperial College London

LEVEL 1
I have selected a model M and prior $P(\theta|M)$

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

LEVEL 3
None of the models is clearly the best

Parameter inference
What are the favourie values of the parameters? (assumes M is true)

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

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

$$P(\theta|d,M) = \frac{P(d|\theta,M)P(\theta|M)}{P(d|M)} \qquad \text{odds} = \frac{\mathrm{P(M_0|d)}}{\mathrm{P(M_1|d)}} \qquad P(\theta|d) = \sum_i P(M_i|d)P(\theta|d,M_i)$$

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

ICIC

Scale for the strength of evidence

Imperial College London

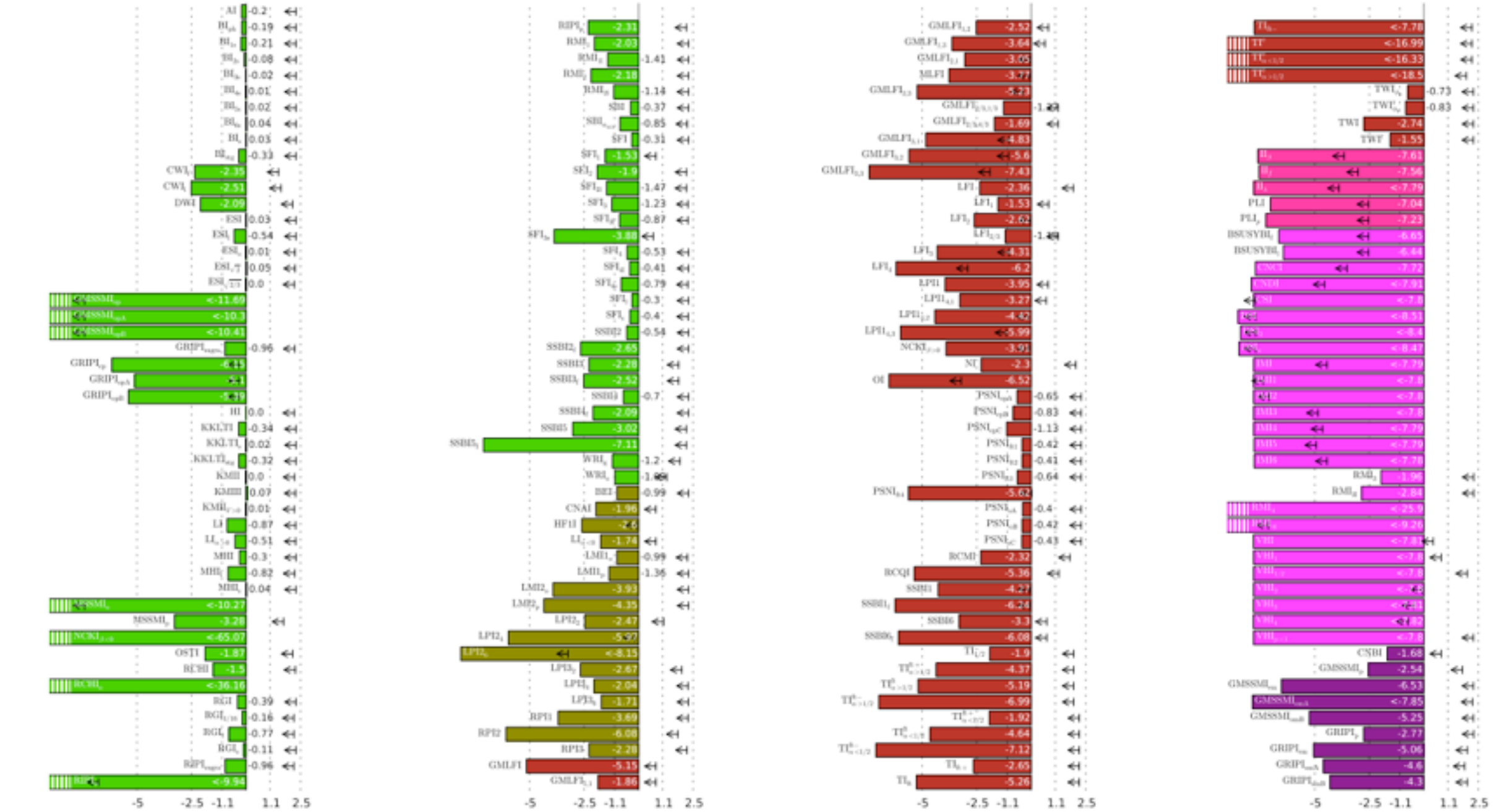
- A (slightly modified) Jeffreys' scale to assess the strength of evidence

|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

Bayesian model comparison of 193 models
Higgs inflation as reference model

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

Martin,RT+14



Schwarz-Terrero-Escalante Classification:

1 | 1-2 | 2 | 2-3 | 3 | 1-2-3

J.Martin, C.Ringeval, R.Trotta, V.Vennin
ASPIC project

Displayed Evidences: 193

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

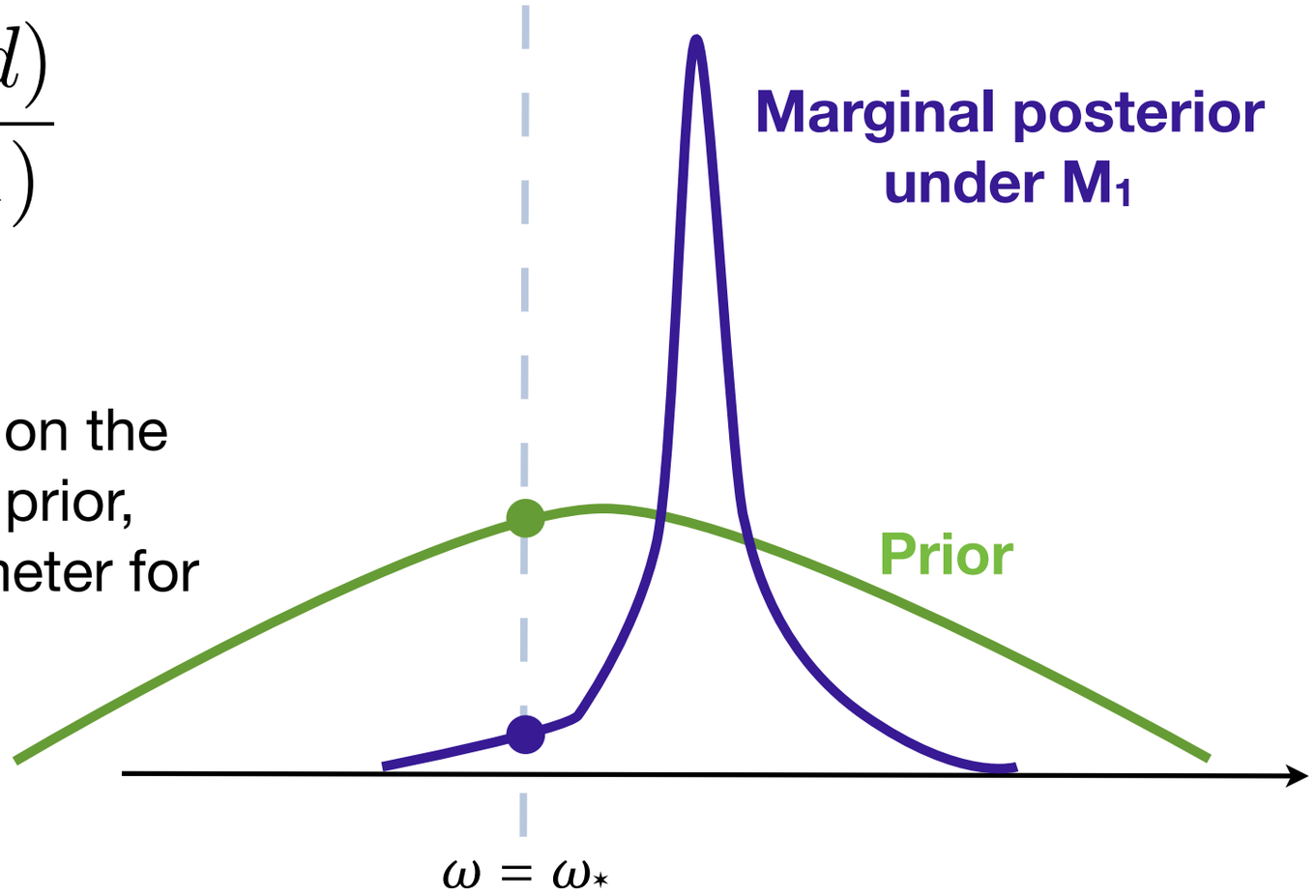
Imperial College London

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.



ICIC

Derivation of the SDDR

Imperial College London

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:

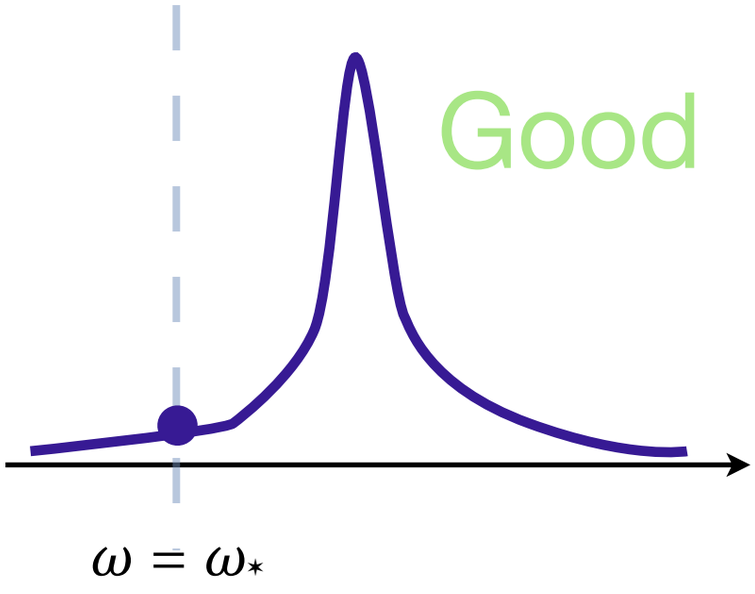
$$\pi_1(\omega, \Psi) = \pi_1(\omega)\pi_0(\Psi)$$

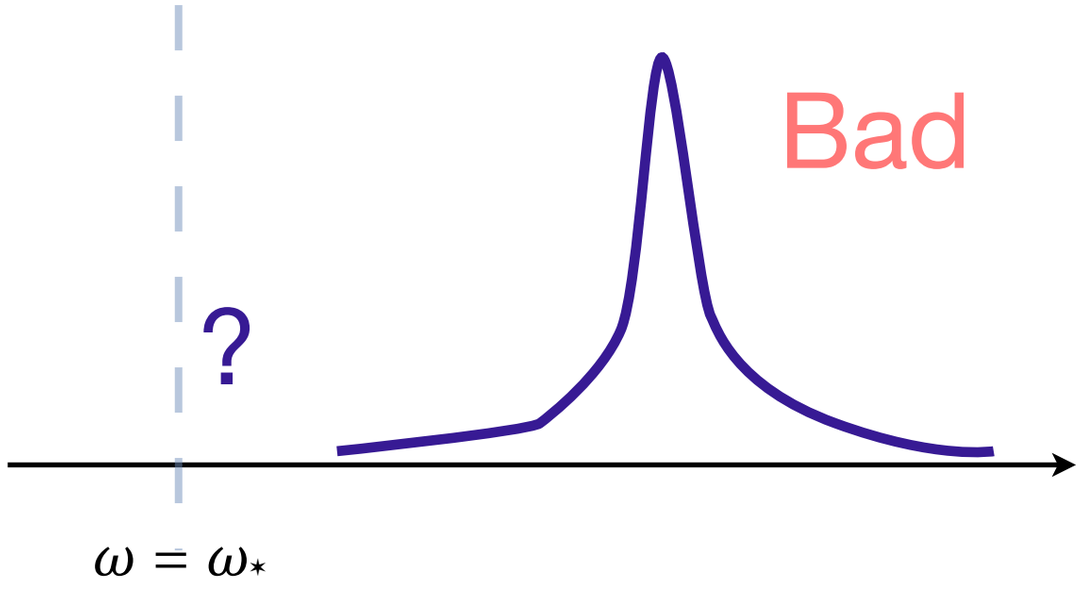
$$B_{01} = \frac{p(\omega_\star|d)}{\pi_1(\omega_\star)} \int d\Psi p(\Psi|\omega_\star, d) = \frac{p(\omega_\star|d)}{\pi_1(\omega_\star)}$$

SDDR: Some comments

Imperial College London

- 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 $\omega=\omega_*$.

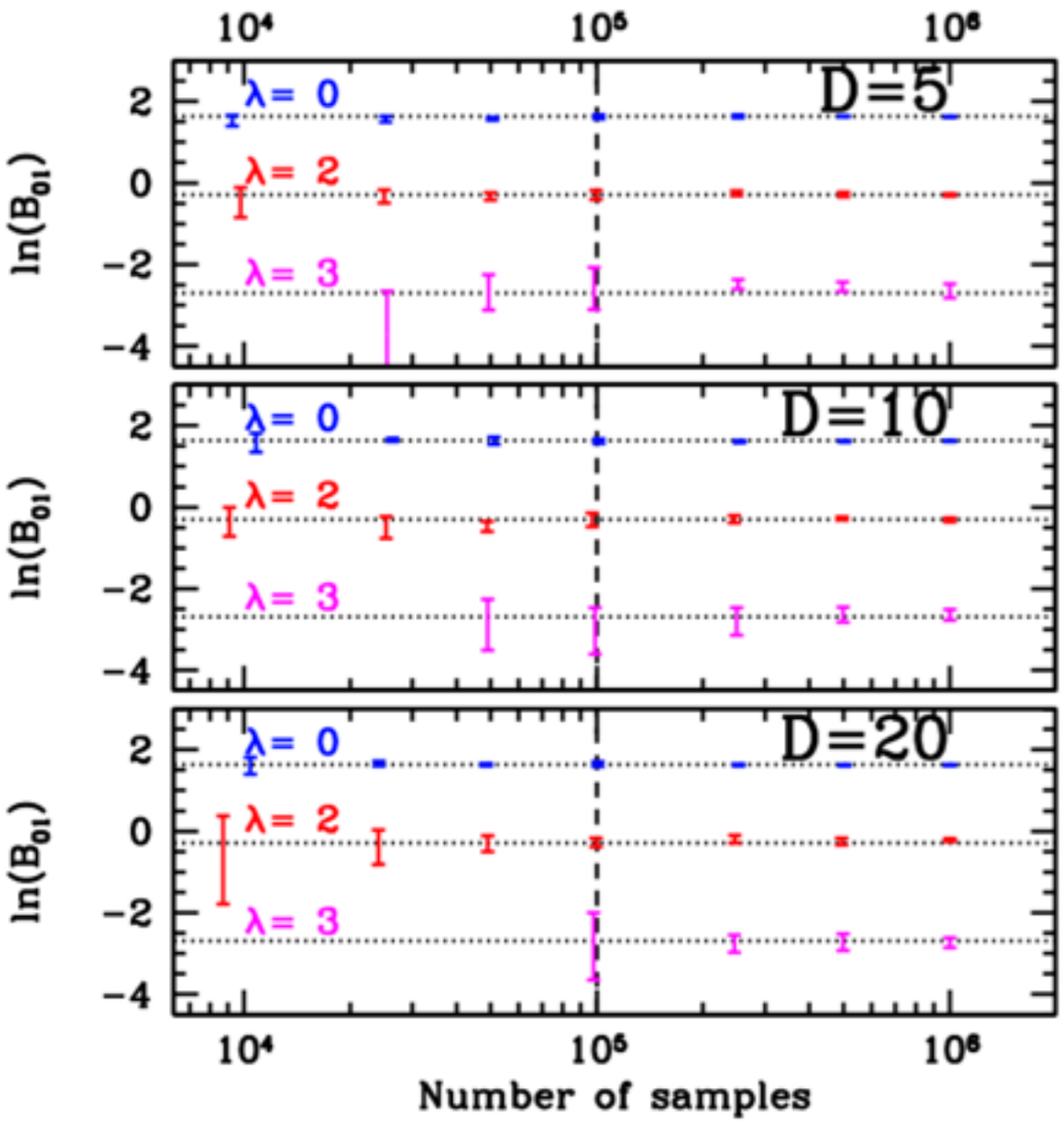


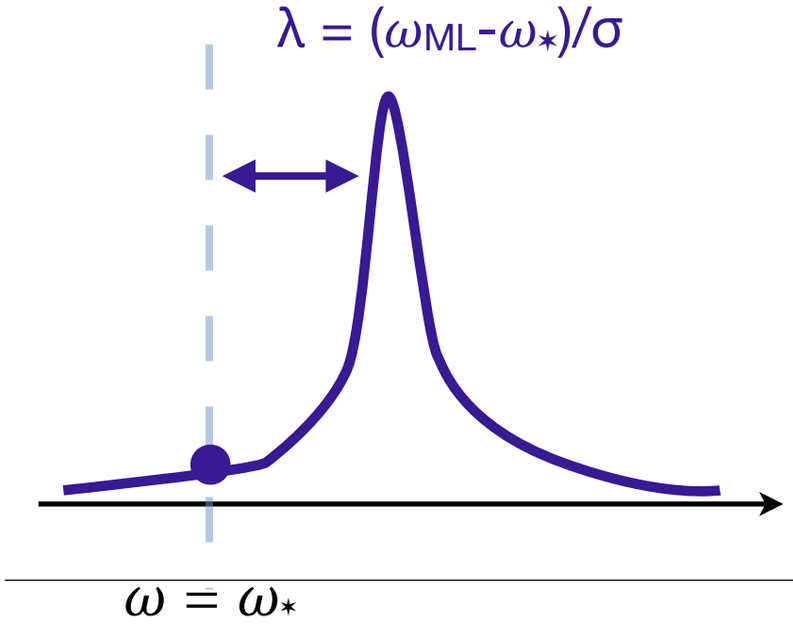


Accuracy tests (Normal case)

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)

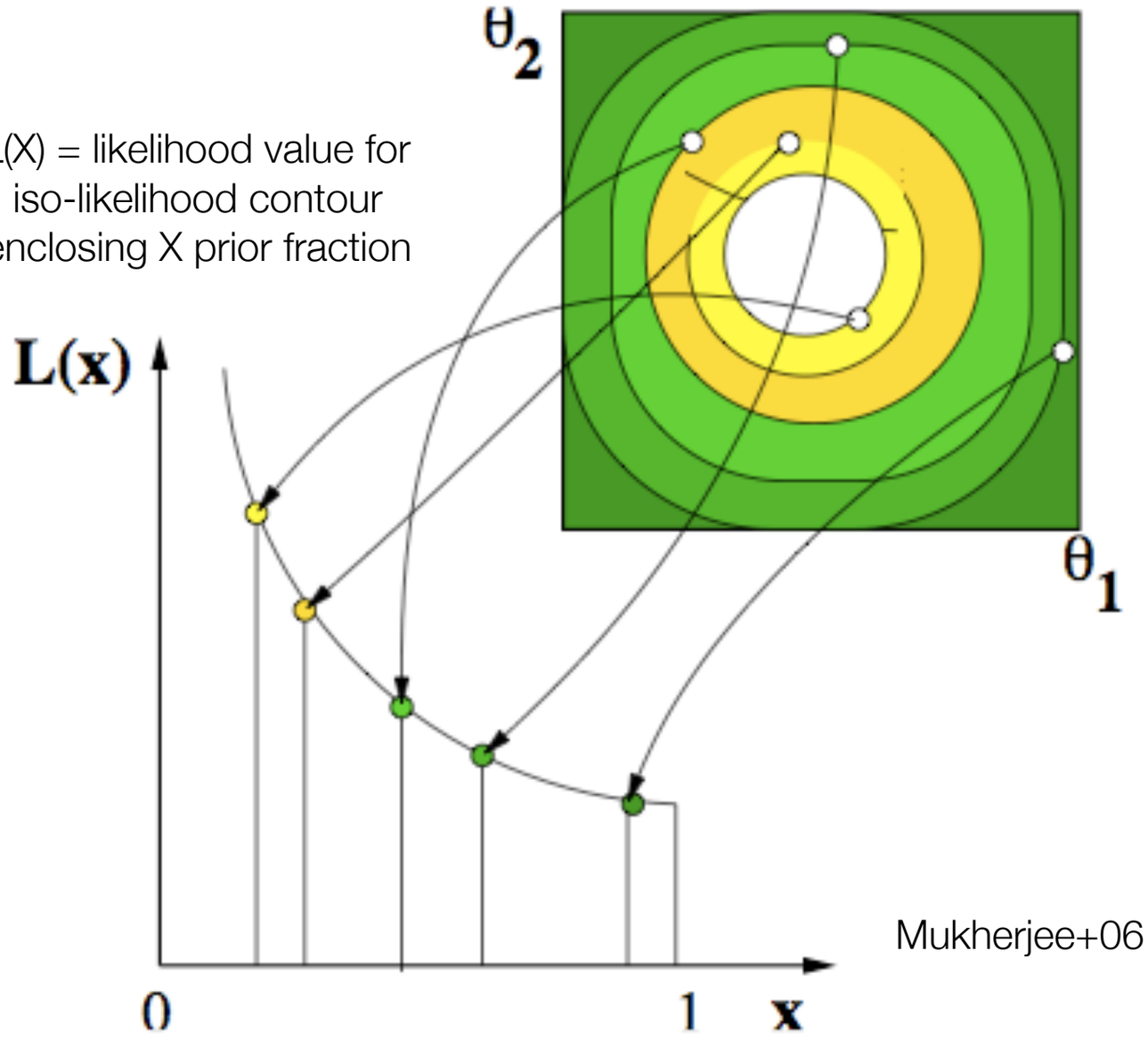
ICIC

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

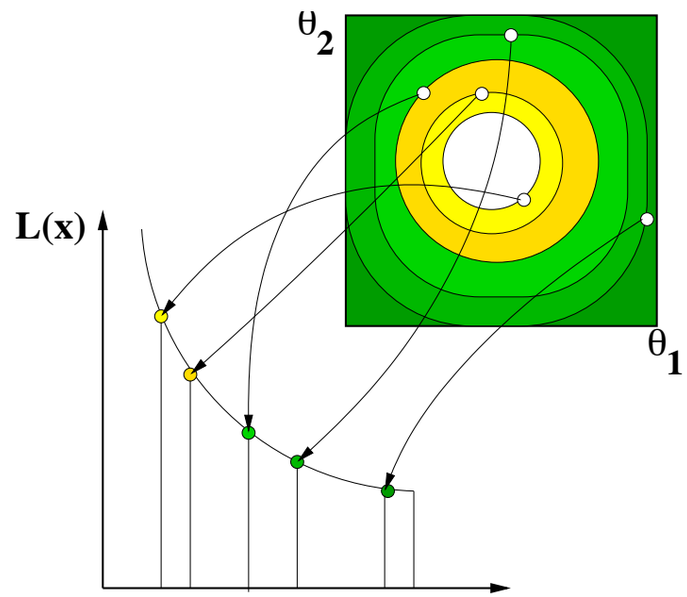
Nested Sampling basics

Imperial College London

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:

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

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

Imperial College London

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:

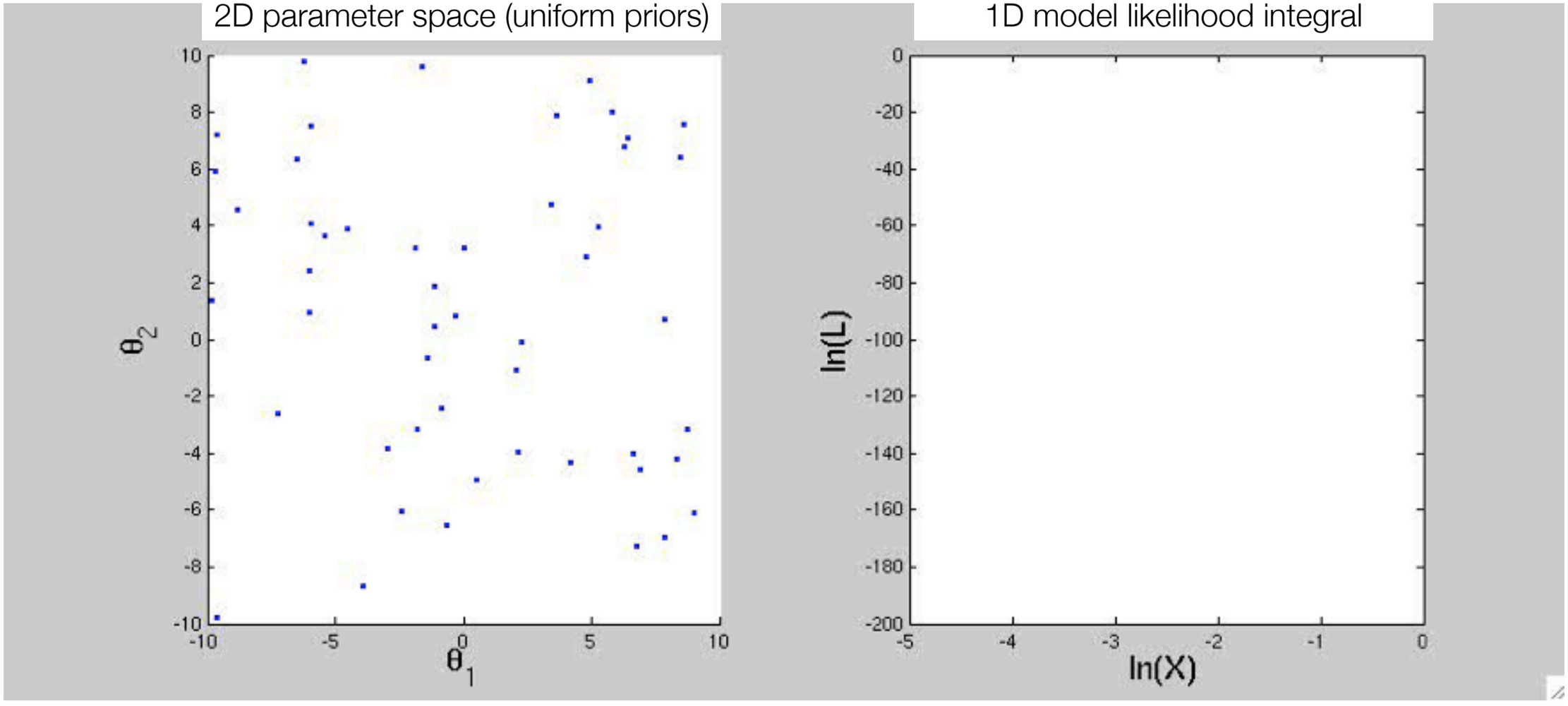
$$w_j = \frac{1}{2}(X_{j-1} - X_{j+1})$$

Nested Sampling in Action

Imperial College London

(animation courtesy of David Parkinson)

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

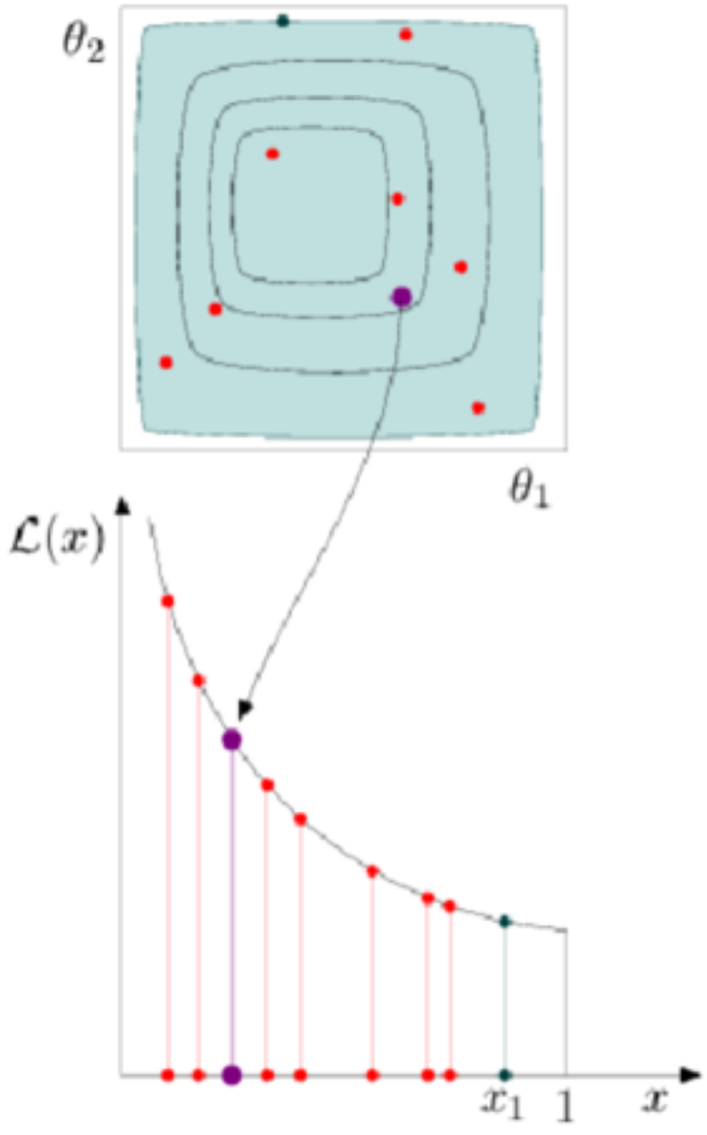


X = Prior fraction

MultiNest sampling approach

Imperial College London

(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 $i \rightarrow i + 1$
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 $E \rightarrow E + L_i w_i$
7. Remove lowest point from active set
8. Replace with new point sampled from $\pi(\boldsymbol{\theta})$ within hard-edged region $L(\boldsymbol{\theta}) > L_i$
9. If $L_{\max} X_i < \alpha E$ (where some tolerance) $\Rightarrow E \rightarrow E + X_i \sum_{j=1}^{N} L(\boldsymbol{\theta}_j)/N$; stop else goto 3

Hard!

Nested Sampling: Sampling Step

Imperial College London

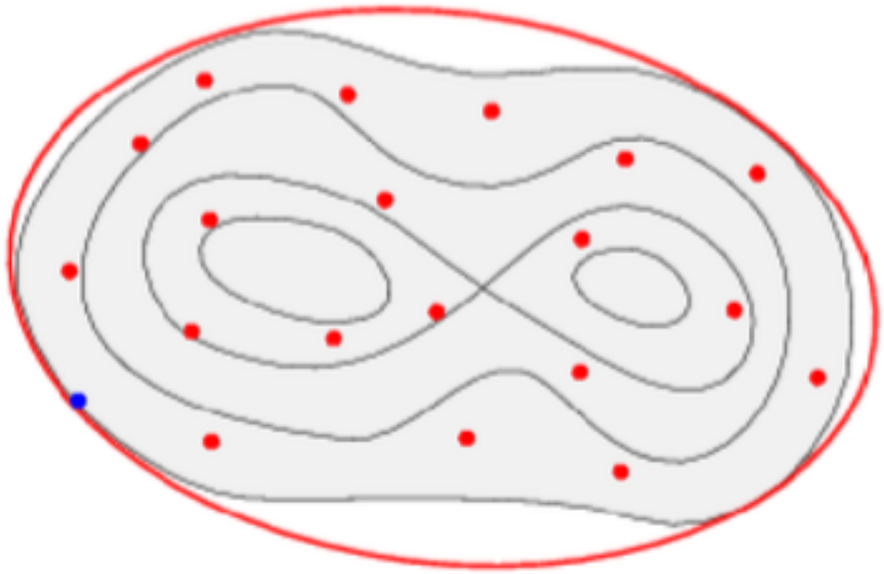
- The hardest part is to sample uniformly from the prior subject to the hard constraint that the likelihood needs to be above a certain level.
- Many specific implementations of this sampling step:
  - Single ellipsoidal sampling (Mukherjee+06)
  - Metropolis nested sampling (Sivia&Skilling06)
  - Clustered and simultaneous ellipsoidal sampling (Shaw+07)
  - Ellipsoidal sampling with k-means (Feroz&Hobson08)
  - Rejection sampling (MultiNest, Feroz&Hobson09)
  - Diffusion nested sampling (Brewer+09)
  - 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)

… there will be others, no doubt.

Sampling Step: Ellipsoid Fit

Imperial College London

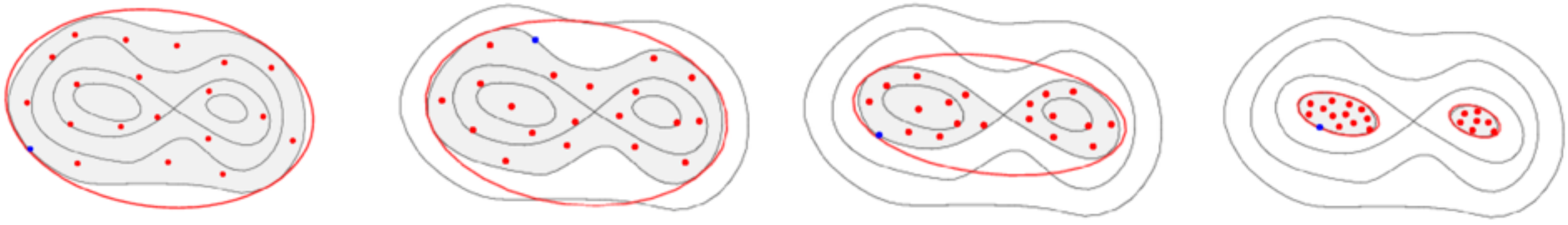
- Simple MCMC (e.g. Metropolis-Hastings) works but can be inefficient
- Mukherjee+06: Take advantage of the existing live points. Fit an ellipsoid to the live point, enlarge it sufficiently (to account for non-ellipsoidal shape), then sample from it using an exact method:



- 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

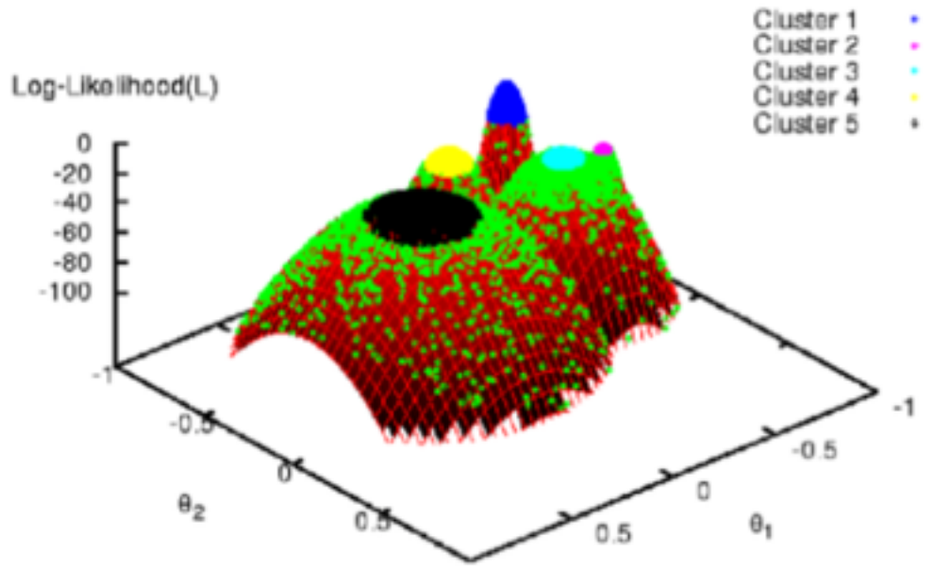


- Feroz&Hobson08; Feroz+08: At each nested sampling iteration
  - Partition active points into clusters
  - Construct ellipsoidal bounds to each cluster
  - Determine ellipsoid overlap
  - Remove point with lowest L_i from active points; increment evidence.
  - Pick ellipsoid randomly and sample new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

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

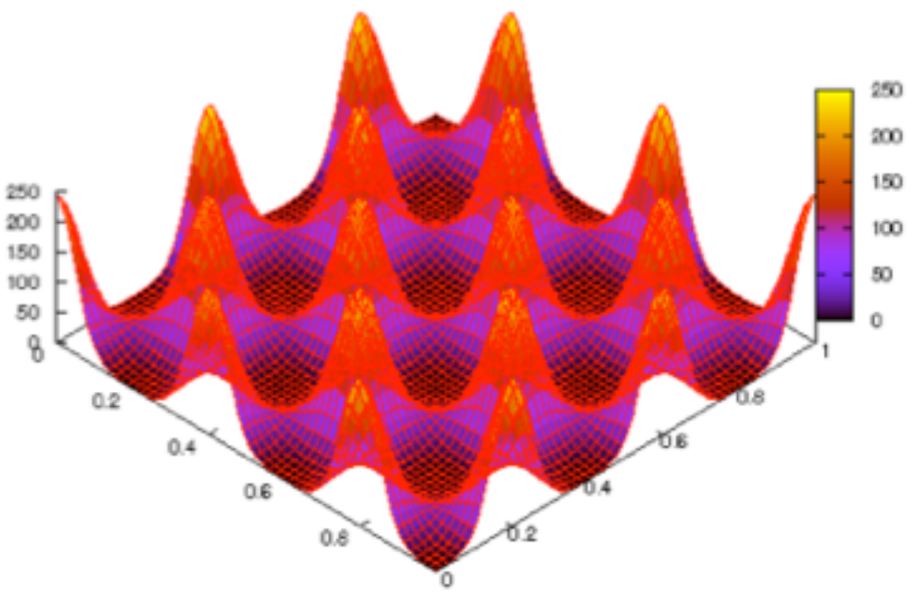
(Animation: Farhan Feroz)

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

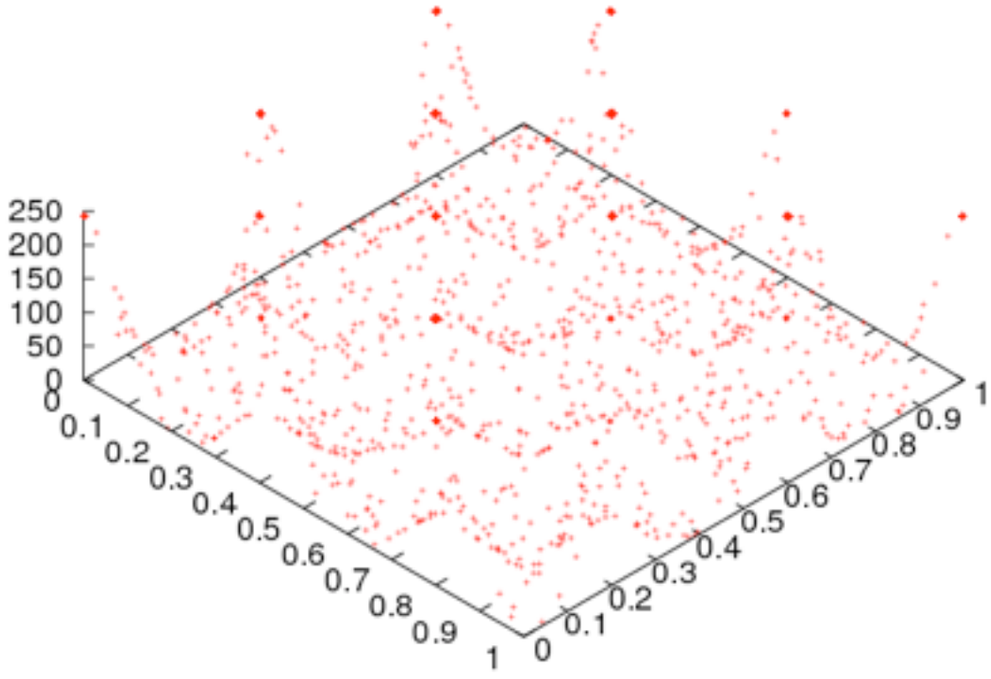
$$\mathcal{L}(\theta_1, \theta_2) = \exp\left(2 + \cos\left(\frac{\theta_1}{2}\right)\cos\left(\frac{\theta_2}{2}\right)\right)^5$$

- 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

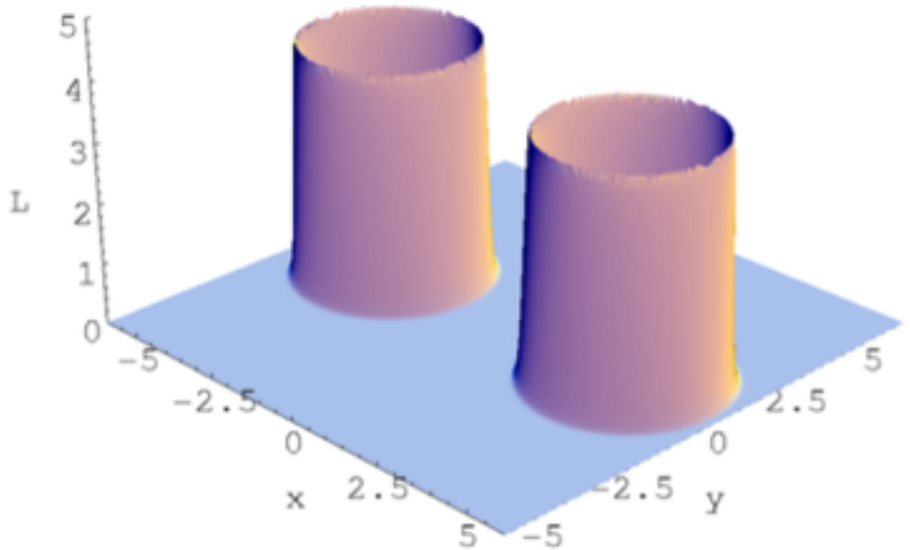


Sampling (30k likelihood evaluations)

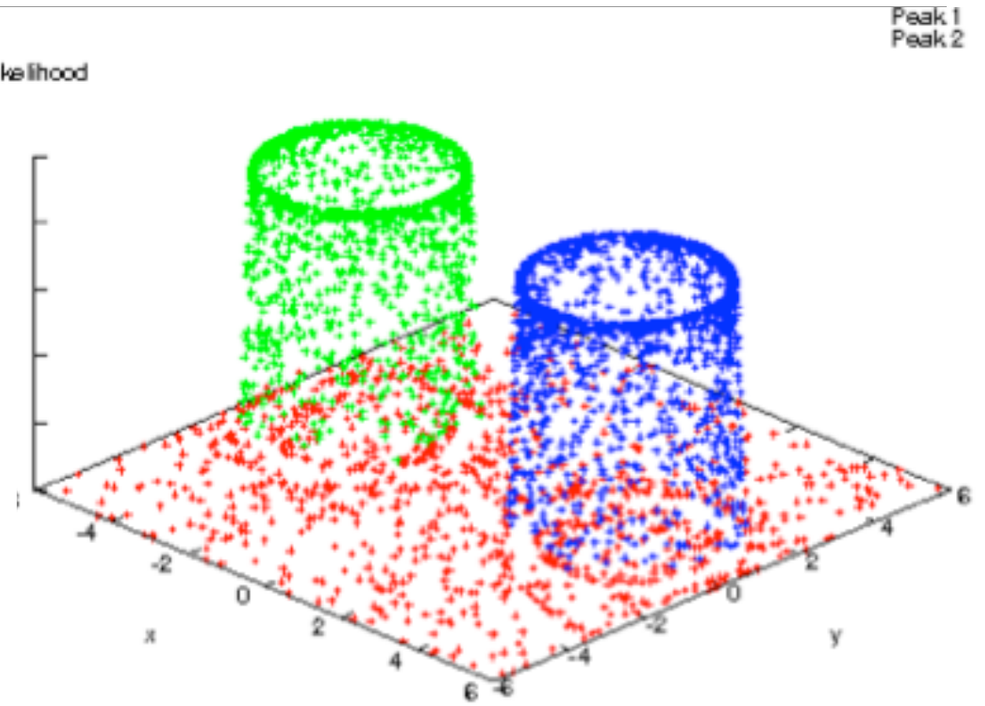
Test: Multiple Gaussian Shells

Imperial College London

Courtesy Mike Hobson



Likelihood



Sampling

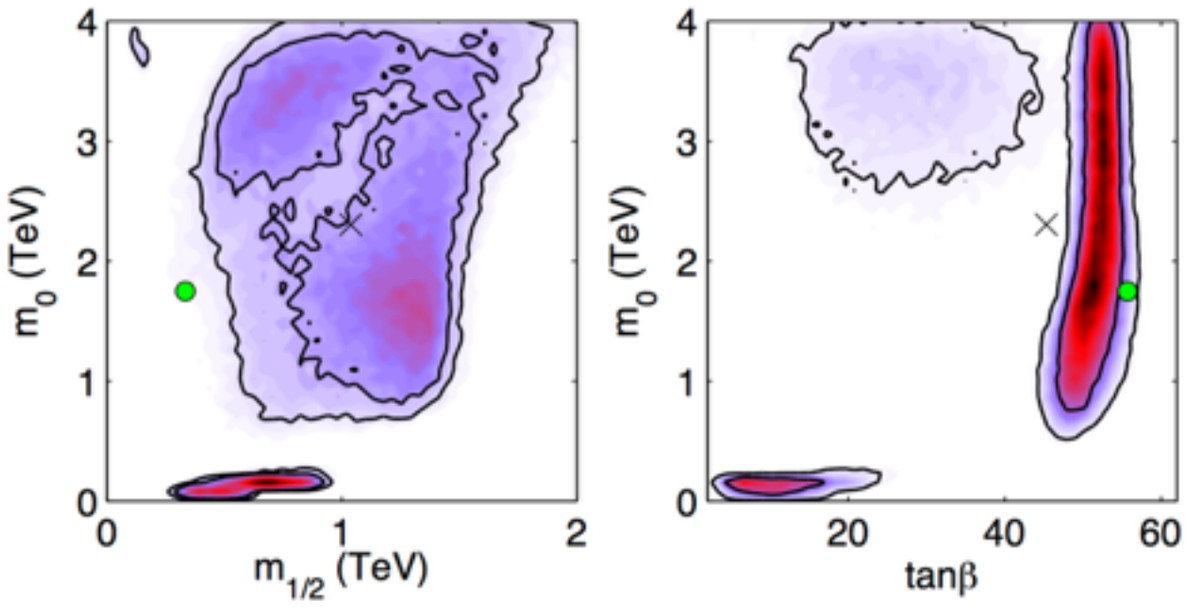
D	N_{like}	Efficiency
2	7000	70%
5	18000	51%
10	53000	34%
20	255000	15%
30	753000	8%

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)

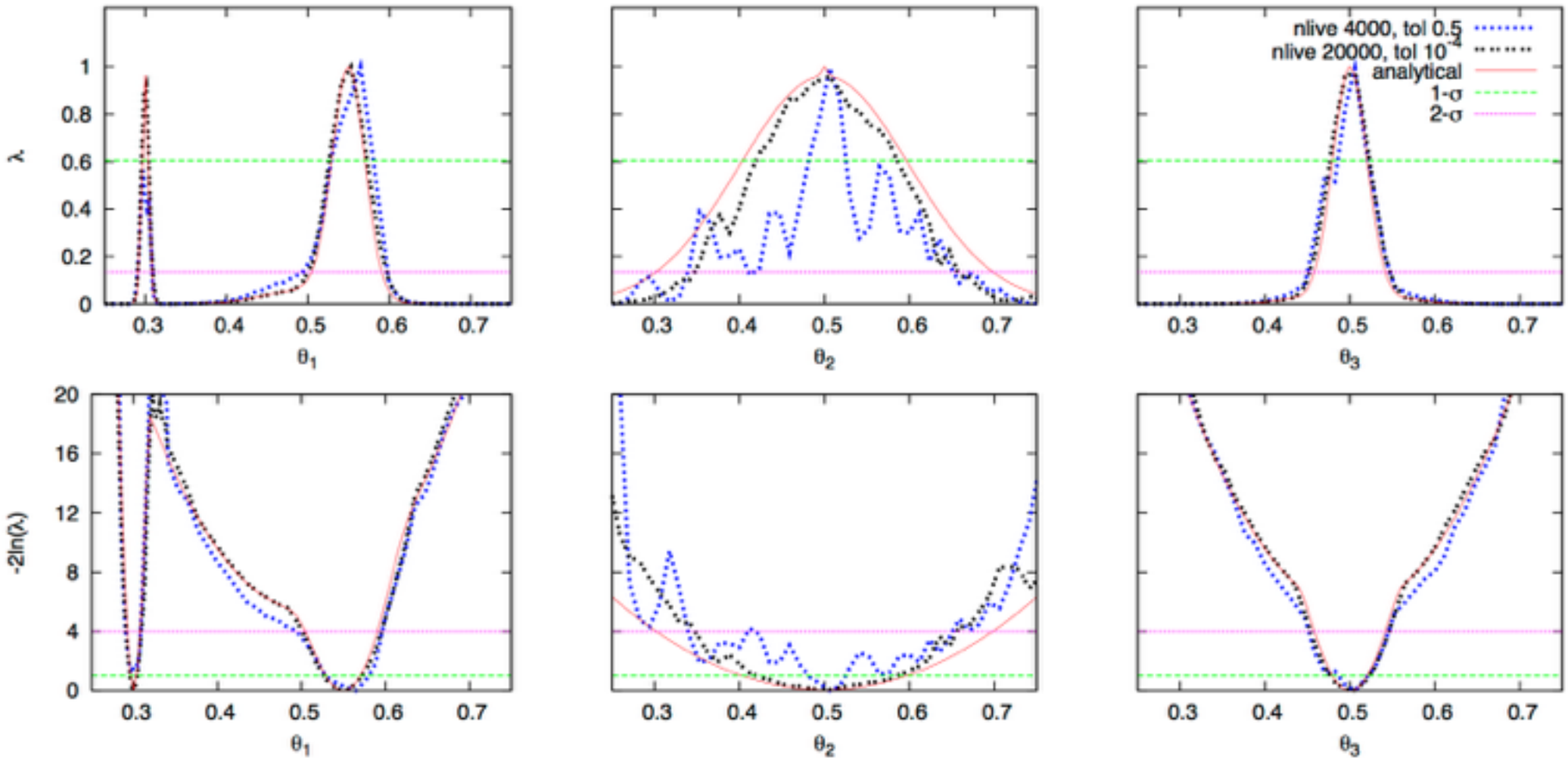
Aside: Profile Likelihood

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 $N_{CPU} >> 1/\text{efficiency}$

  - 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

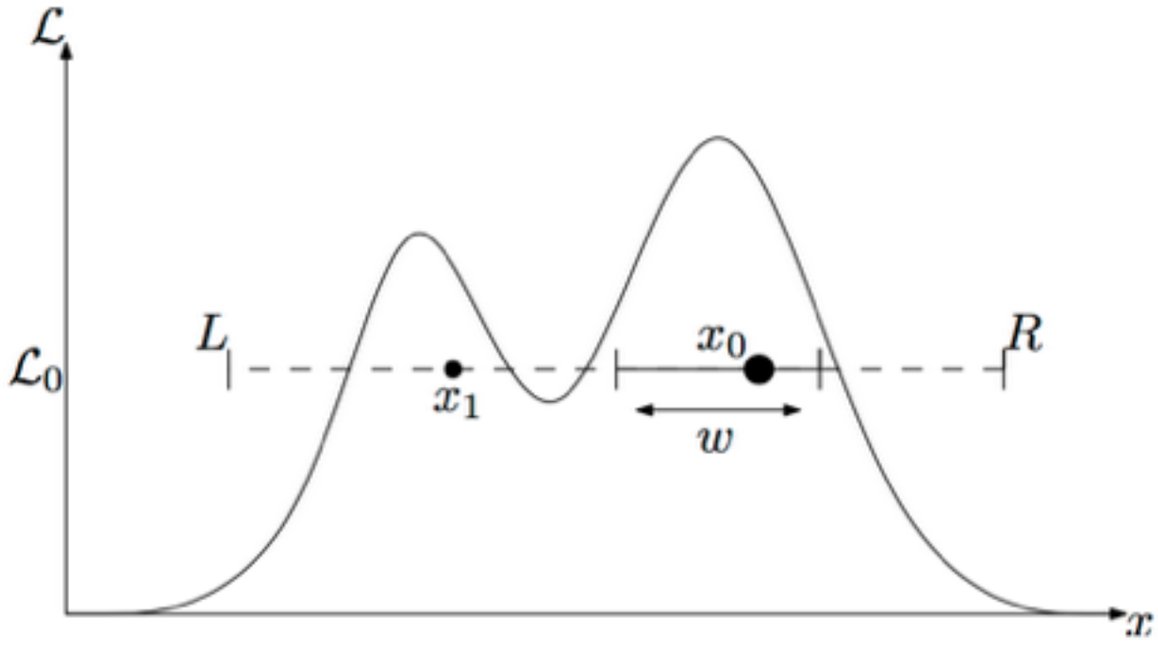
- **Slice Sampling (Neal00) in 1D:**

  - 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

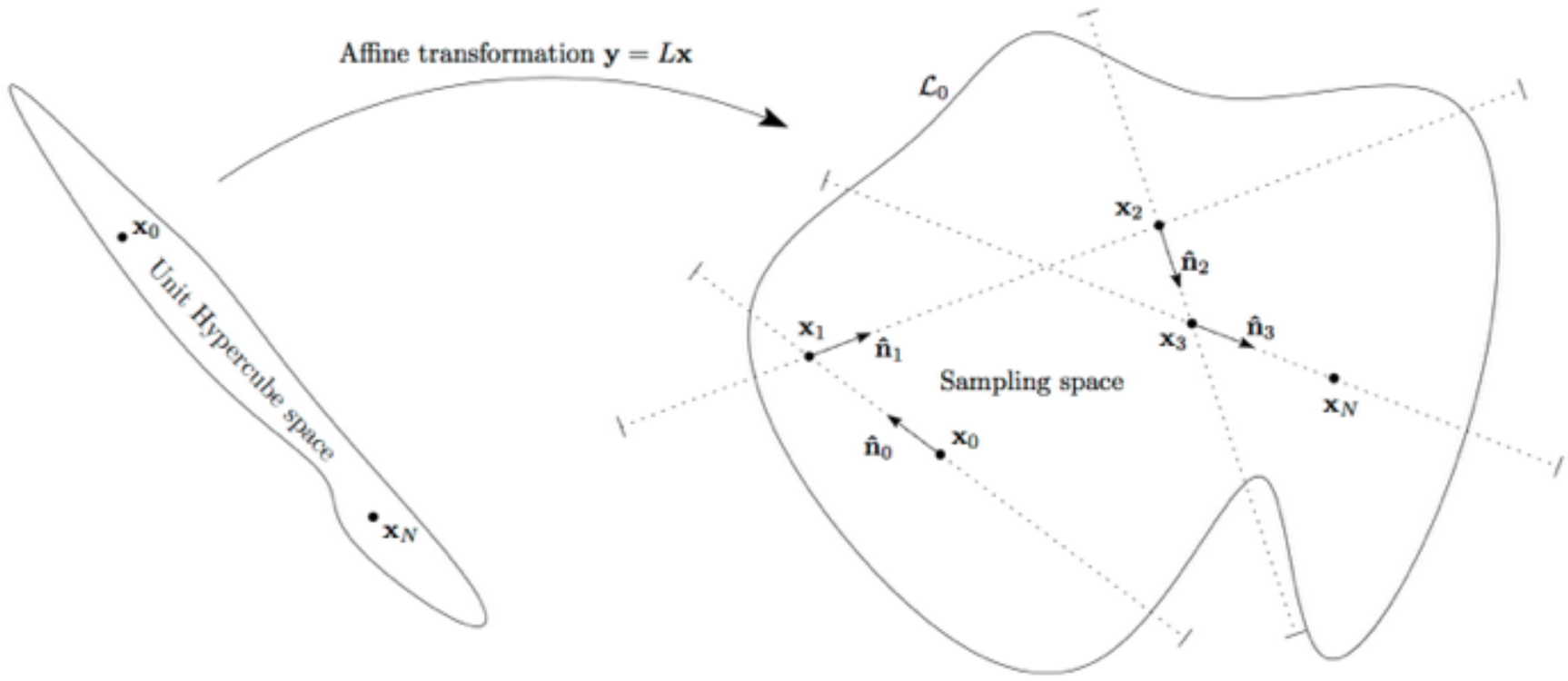
  - If x_1 not in the slice, contract bound down to x_1 and re-sample x_1



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

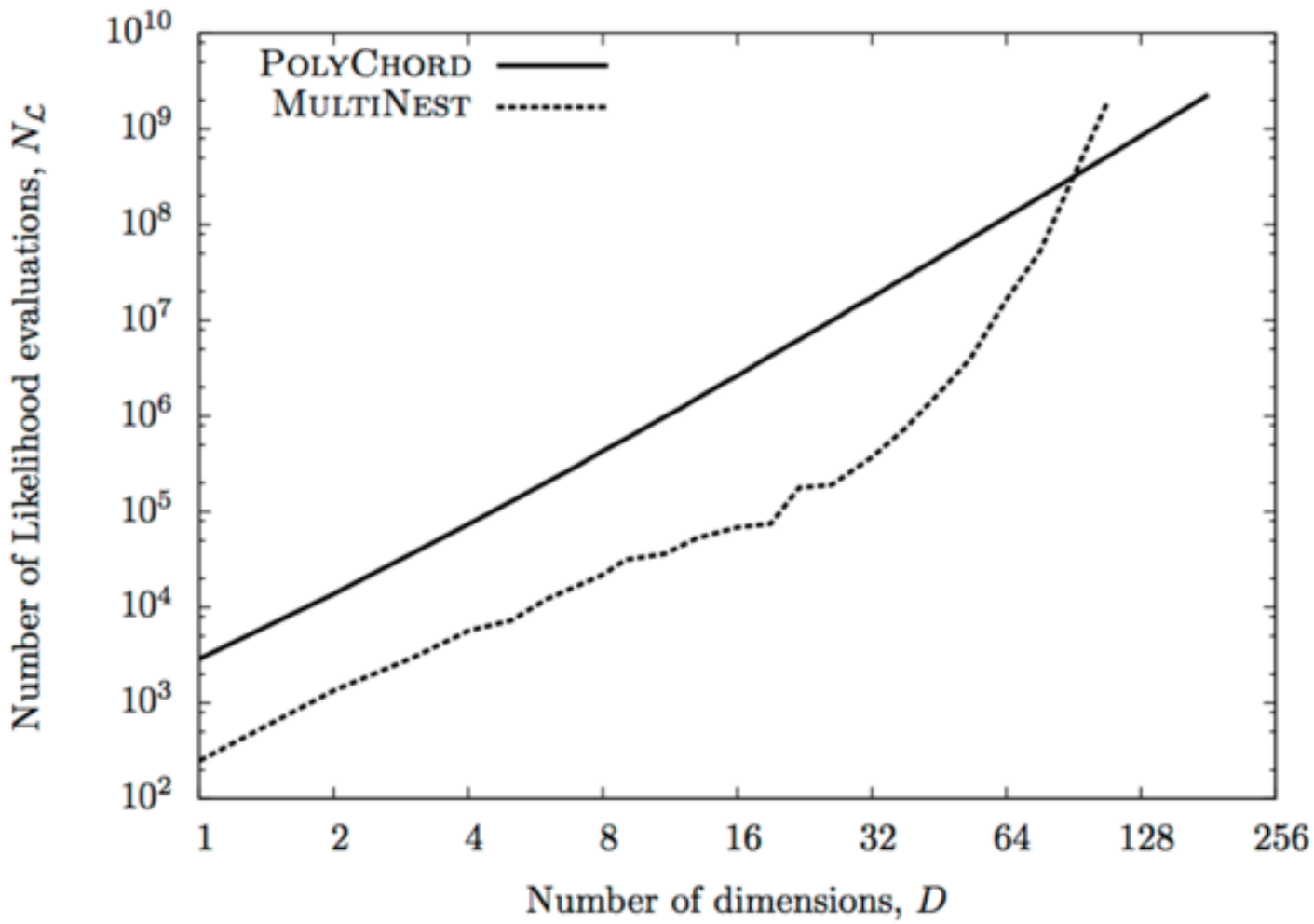


Handley+15

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



ICIC

Summary and Conclusions

Imperial College London

- **Bayesian model comparison** in cosmology requires the evaluation of model likelihoods, often on an industrial scale.
- Many cases of interest involve **nested models:** In this case, the Savage-Dickey Density Ratio offers a computationally inexpensive way of evaluating the Bayes Factor between nested models (with mild assumptions and caveats about sampling accuracy).
- **Nested Sampling** has emerged as a powerful tool for model likelihood computation, giving posterior samples (and accurate profile likelihood estimates) as by-product.
- In the **MultiNest** implementation, nested sampling works well up to ~ 30 dimensions, with O(100) savings in computational time wrt e.g. thermodynamic integration (or standard MCMC for inference).
- Handling larger dimensionality (>> 30) requires better sampling techniques, e.g. **PolyChord** implementing multi-D slice sampling.

Thank you!

www.robertotrotta.com

@R_Trotta

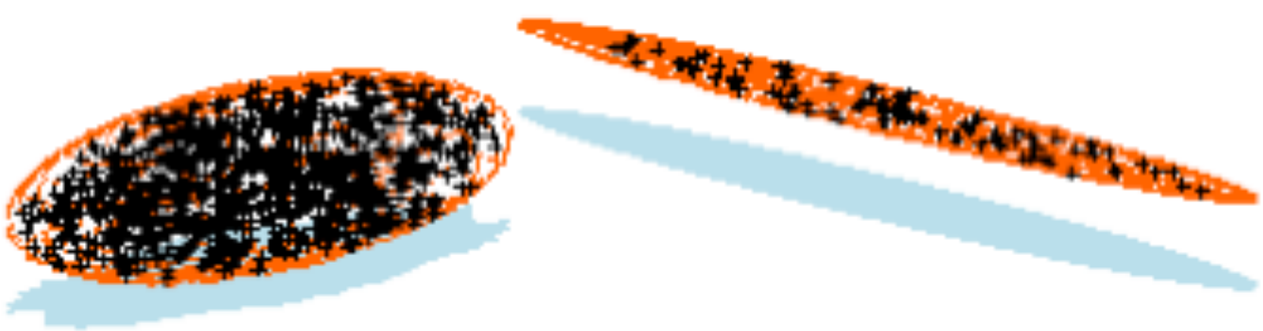
ICIC

astro.ic.ac.uk/icic

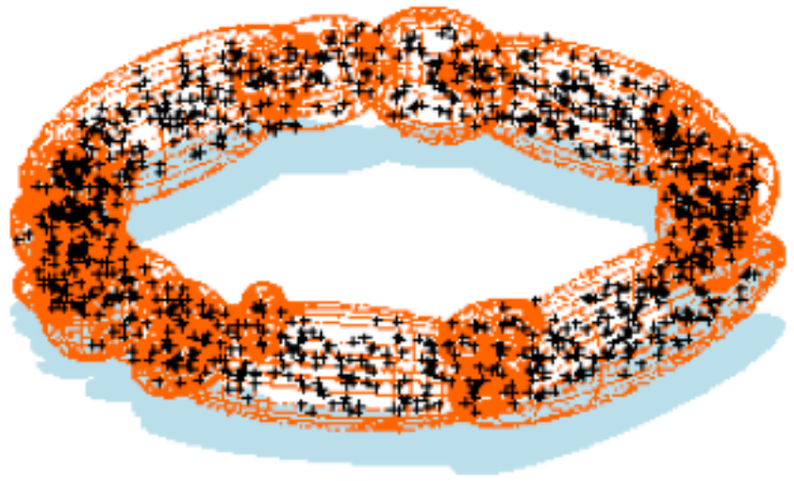
Ellipsoidal decomposition

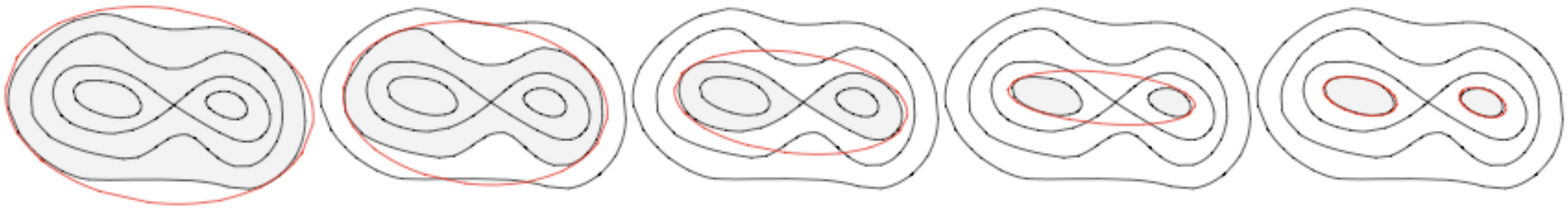
Imperial College London

Unimodal distribution



Multimodal distribution





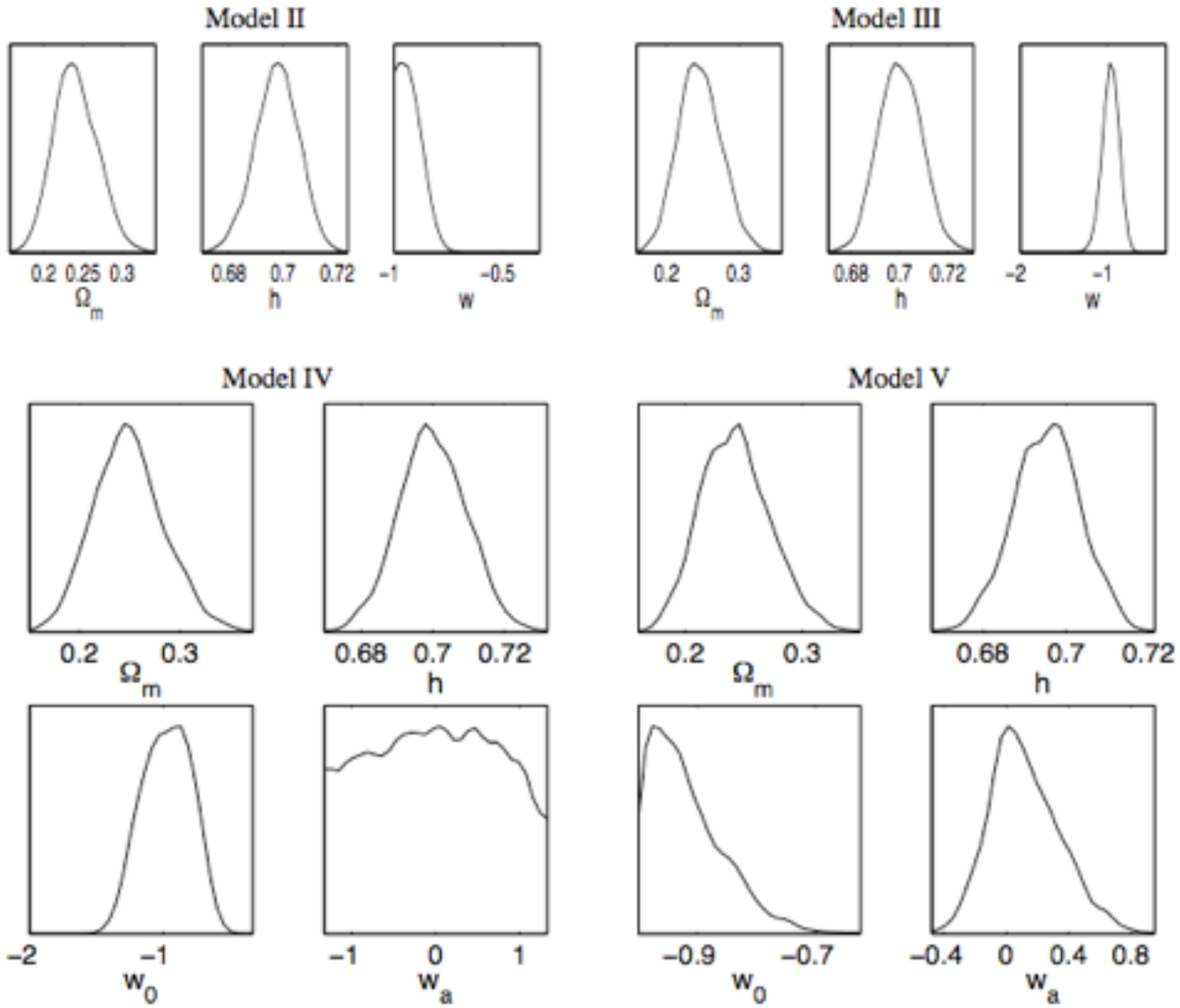
Courtesy Mike Hobson

Bayesian Model-averaging

Imperial College London

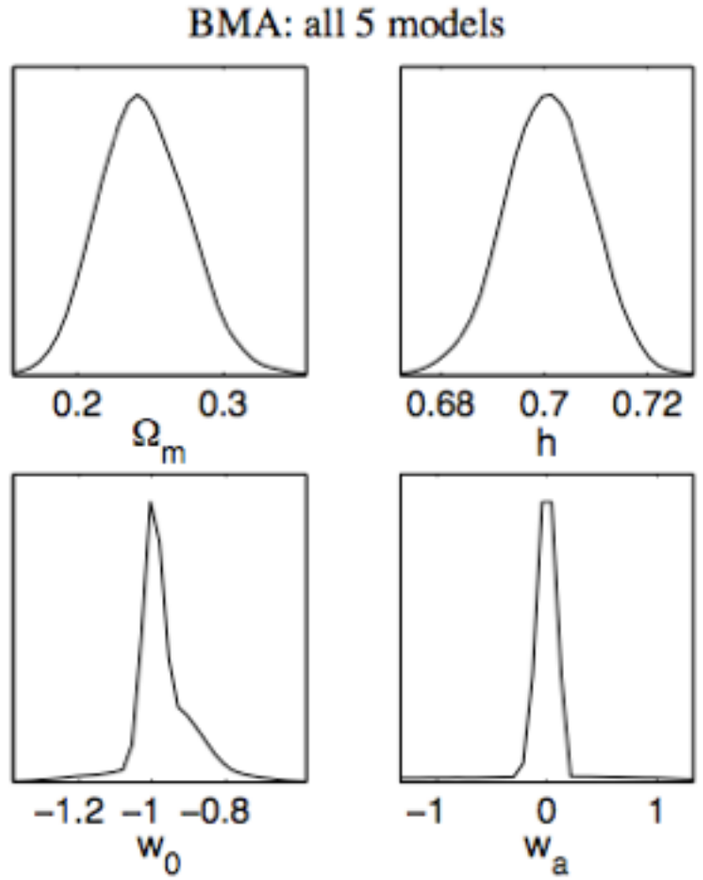
$$P(\theta|d) = \sum_i P(\theta|d,M_i)P(M_i|d)$$

An application to dark energy:



Liddle et al (2007)

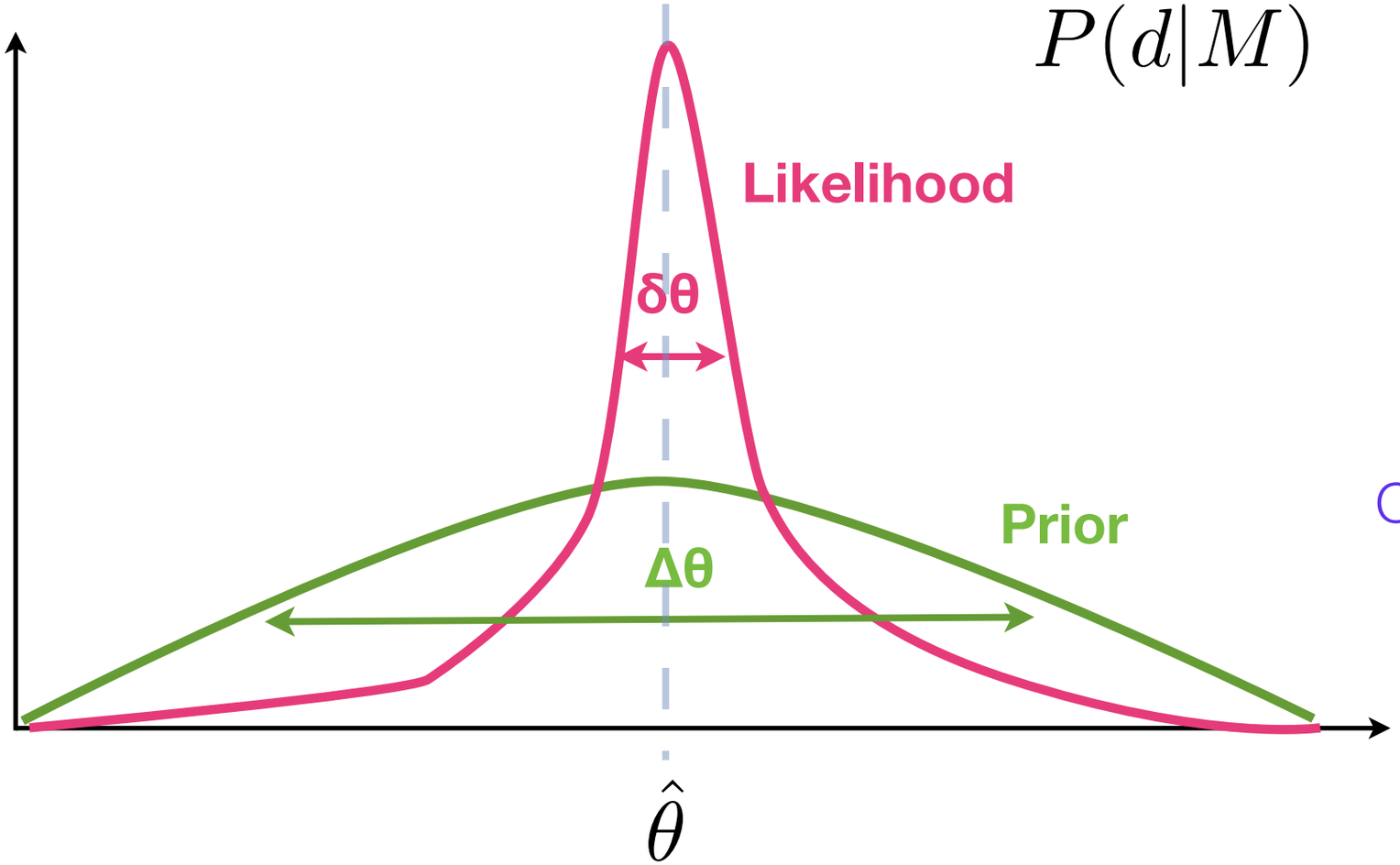
Model averaged inferences



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



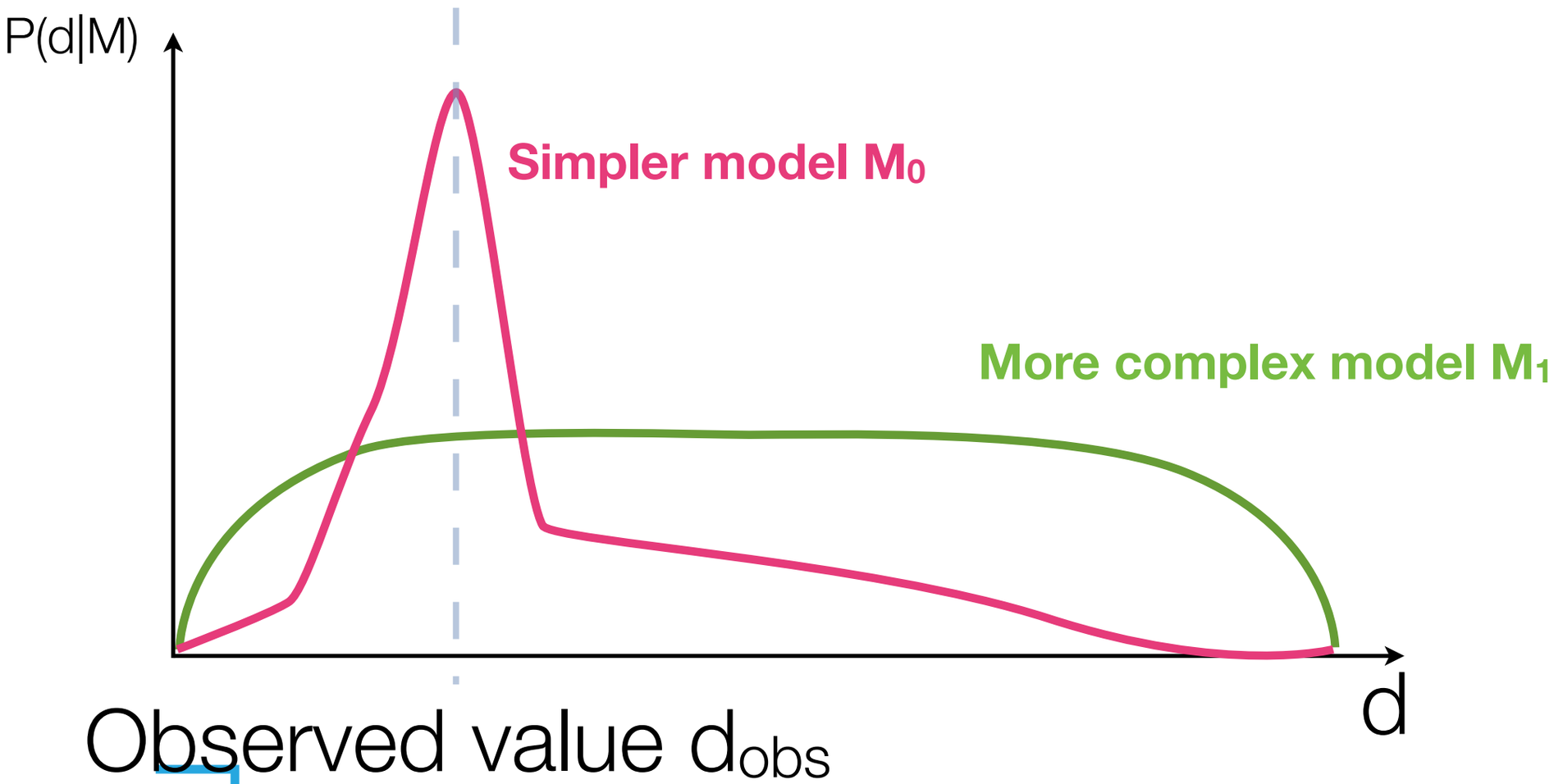
$$P(d|M) = \int d\theta L(\theta) P(\theta|M)$$
$$\approx P(\hat{\theta}) \delta\theta L(\hat{\theta})$$
$$\approx \frac{\delta\theta}{\Delta\theta} L(\hat{\theta})$$

Occam's factor

The evidence as predictive probability

Imperial College London

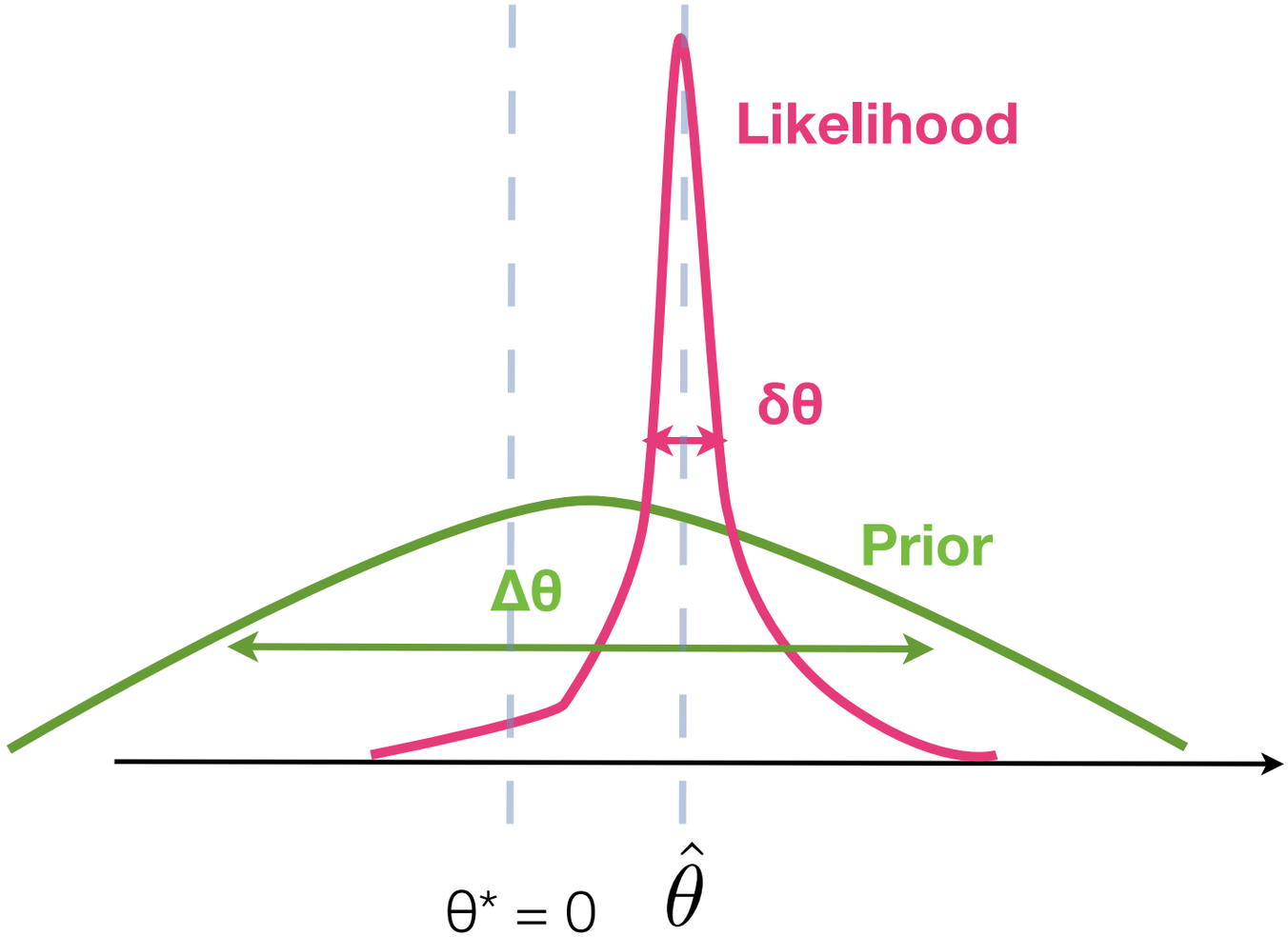
- The evidence can be understood as a function of d to give the predictive probability under the model M:



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

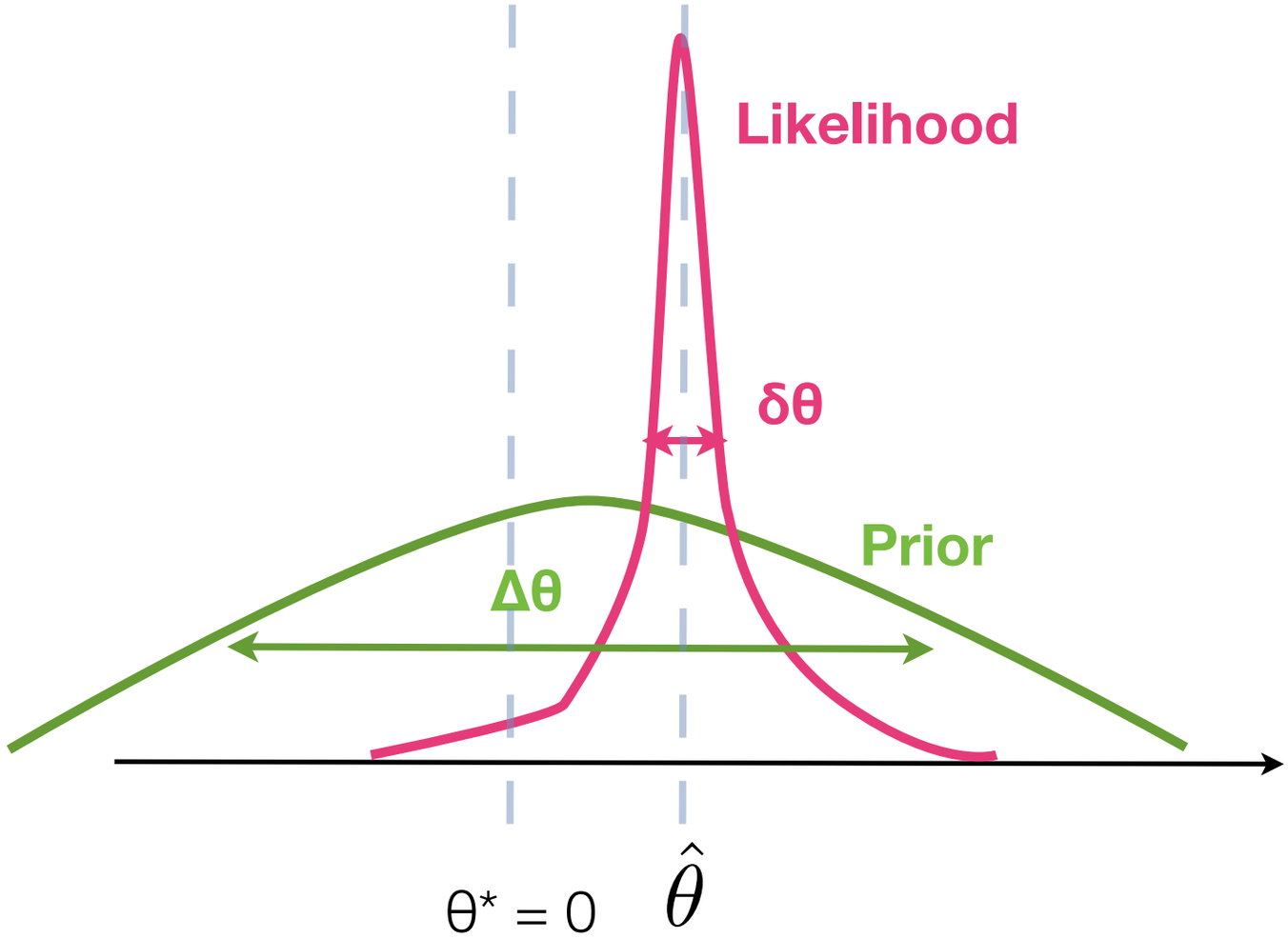
Define: $\lambda \equiv \frac{\hat{\theta} - \theta^\star}{\delta\theta}$

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)



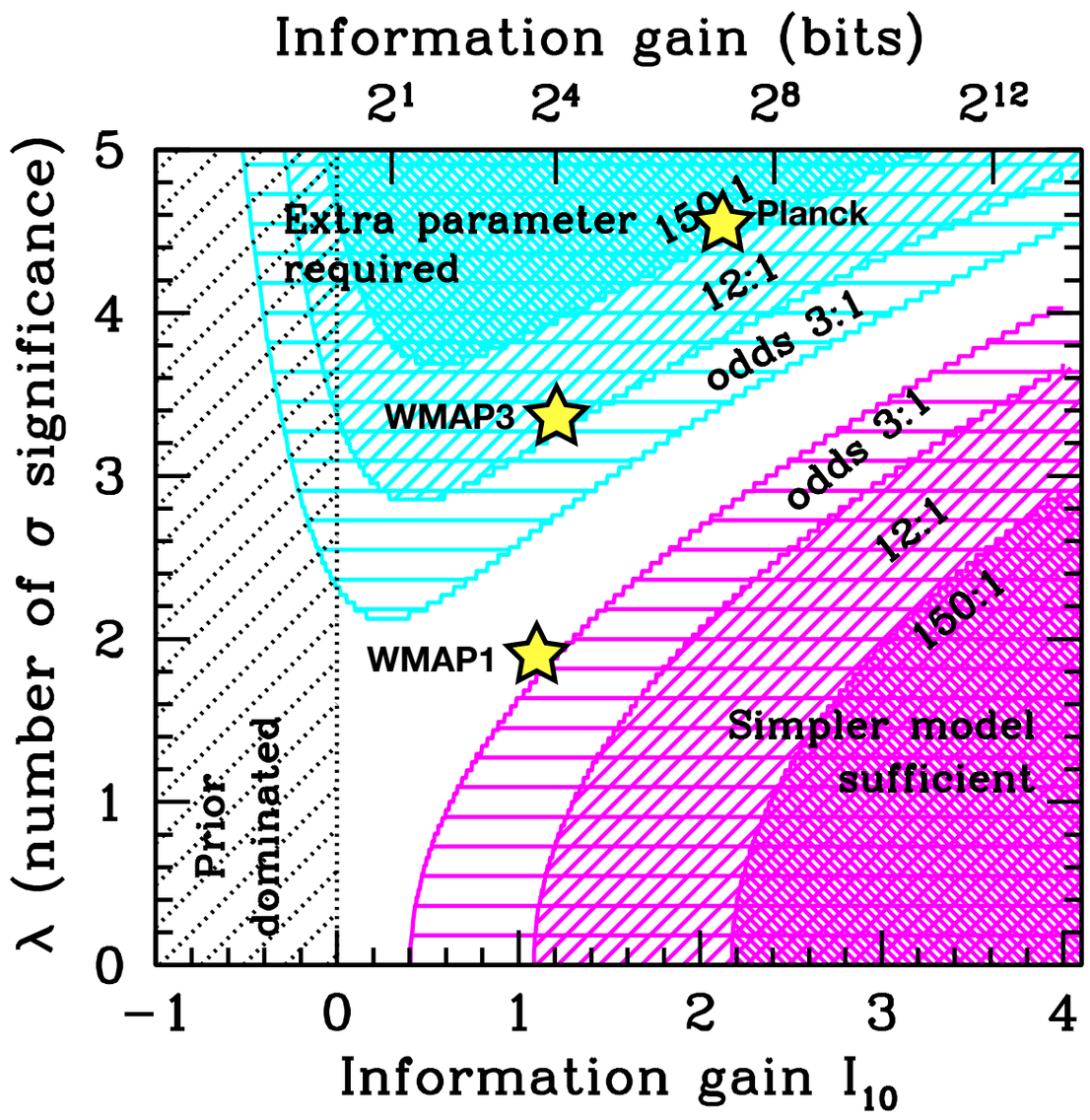
The rough guide to model comparison

Imperial College London

wider prior (fixed data)

larger sample (fixed prior and significance)

Trotta (2008)



Δθ = Prior width
$\delta\theta$ = Likelihood width

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

ICIC