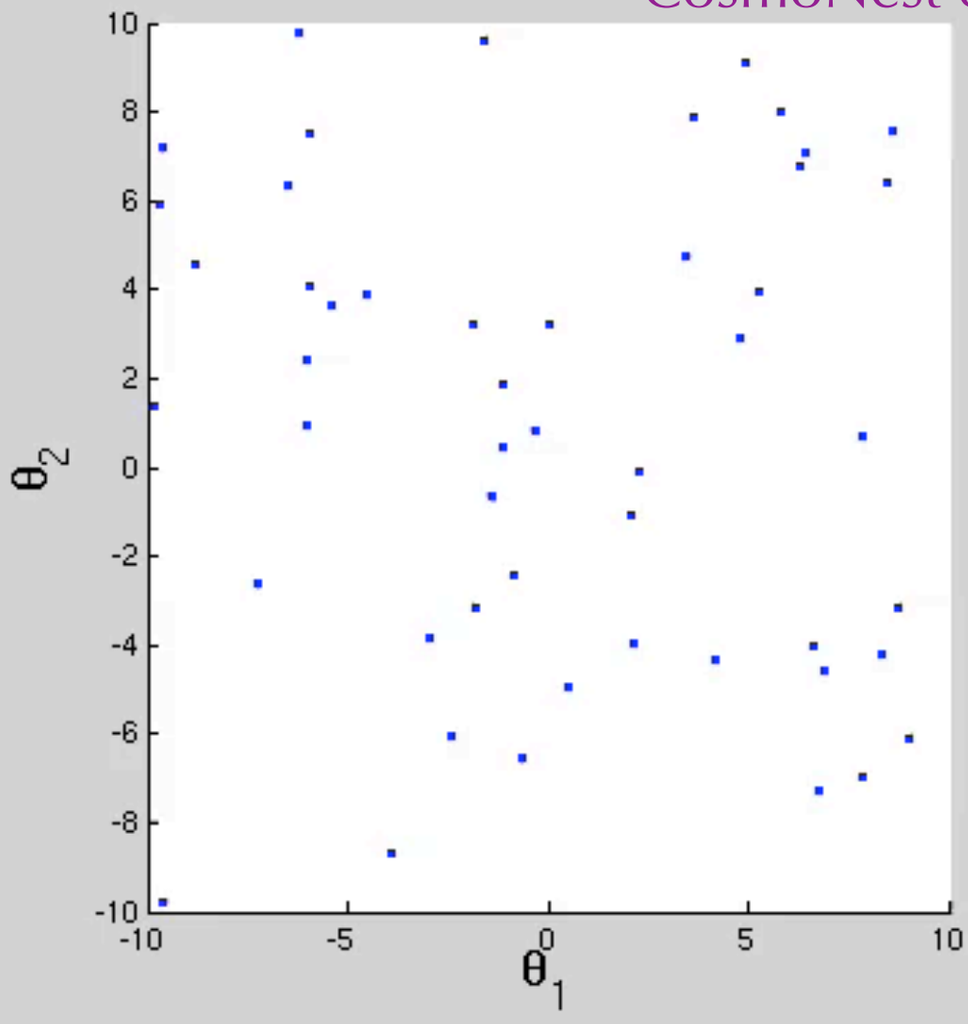
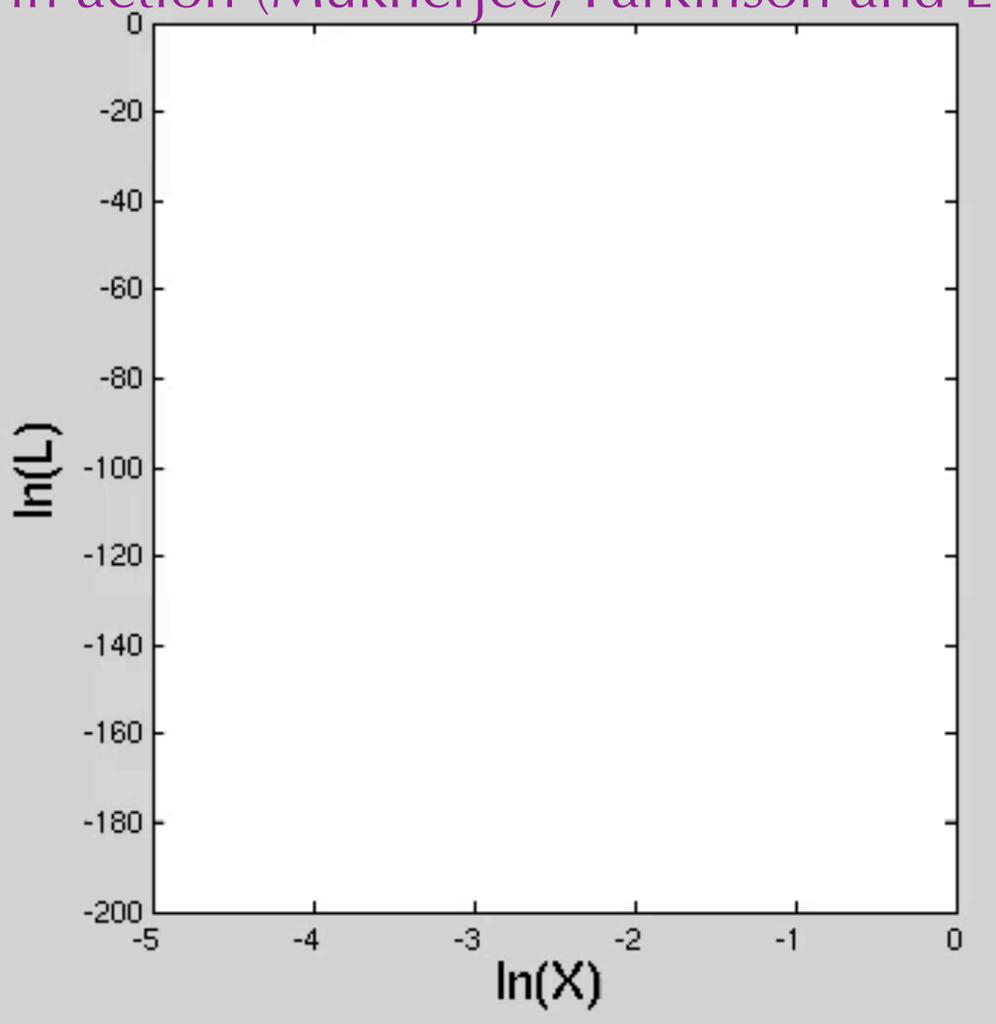
University of Sussex

Model Selection and multi-model inference

CosmoNest code in action (Mukherjee, Parkinson and Liddle)





Levels of Bayesian inference

Parameter Estimation

I've decided what the correct model is.

Now I want to know what values of the parameters are consistent with the data.

I can do this using e.g. Markov Chain Monte Carlo.

Model Selection

Now I think about it, I don't actually know what the correct model is. It could be one of several.

Now I want to know what the best model is.

I can do this by computing the Bayesian Evidence. I can then do parameter estimation using the best model.

Multi-model Inference

Mmm, I did the model selection thing, but there wasn't a single best model.

But I still want to know how probable the parameter values are.

I can do this by combining the parameter likelihoods using Bayesian Model Averaging, adding them together weighted by the model probabilities.

The Bayesian evidence

Bayes theorem again, but conditioned on a model.

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)} \quad \Rightarrow \quad P(\theta|D,M) = \frac{P(D|\theta,M)P(\theta|M)}{P(D|M)}$$

Posterior model probability!

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

Bayesian evidence

How do we calculate it?

$$P(D|M) = \int P(D|\theta,M)P(\theta|M)\,d\theta$$
$$E(M) = \int L(\theta)\,Pr(\theta)\,d\theta$$

This can be evaluated in a number of ways: we use a Monte Carlo integration method called nested sampling.

What does it reward? **Model predictiveness**

Bayesian model selection

Choose dataset

Choose model: Set of parameters to be varied
Prior ranges for those parameters
Compute likelihood function
Obtain posterior parameter distribution

Bayesian model selection

Choose dataset

Choose model M_1:
Set of parameters to be varied
Prior ranges for those parameters
Compute likelihood function
Obtain posterior parameter distribution

Assign model probability $P(M_1)$

Choose model M_2:
Set of parameters to be varied
Prior ranges for those parameters
Compute likelihood function
Obtain posterior parameter distribution

Assign model probability $P(M_2)$

.....

Compute model likelihoods, known as the Bayesian evidence
Update prior model probabilities to posterior ones
[option: multi-model inference by Bayesian model averaging]
Interpret

Model priors

Bayesian inference requires that the prior probabilities be specified, giving the state of knowledge before the data was acquired to test the hypothesis.

We now have to choose model priors too. A common choice is equal prior model probabilities, but this is not obligatory. The important thing is to specify them, so that someone else can see the consequence of their different opinion.

A significant concern is that our set of models may not be complete. There may be better models we haven't thought of yet. Hence we only get an upper limit on the posterior model probability.

In Bayesian model comparison, there is no such thing as a null hypothesis. There are only alternate hypotheses, treated on an equal footing.

Consequently, the concept of Type I (false positive) and Type II (false negative) errors doesn't exist either; as there is no asymmetric hypothesis there is no distinction other than correct and incorrect inference.*

*Nevertheless, one may wish to impose asymmetric criteria, being more willing to accept a wrong answer of one kind than of another.

Ultra-Bayesian view of Higgs searches

The only models which exist that have proven a satisfactory explanation of particle physics data are those that contain a Higgs particle.

Accordingly, the Higgs particle has, in a Bayesian sense, already been detected.

Its mass is rather uncertain, $m_H \simeq (150 \pm 20)$ GeV, but future data are expected to narrow the range.

However at no point in the process envisaged can one say that the Higgs has `been discovered'.

A Bayesian model comparison can lend support to a simpler (nested) model, by disfavouring a more complex alternative which is penalized for lack of predictiveness.

By contrast, frequentist methods can only impose upper limits in such cases, as the more complex model is always able to fit the data at least as well as the simpler one.

Terminology: a nested model is one whose parameters are a subset of the parameters of a more general model.

Interpretational scale

Computing the evidence is often challenging, but feasible due to recent algorithm developments. For guidance in interpretting the evidence, people usually appeal to the Jeffreys' scale.

Jeffreys' Scale:		
	$\Delta \ln E < 1$	Not worth more than a bare mention
	$1 < \Delta \ln E < 2.5$	Substantial evidence
	$2.5 < \Delta \ln E < 5$	Strong to very strong evidence
	$5 < \Delta \ln E$	Decisive evidence

The most useful divisions are 2.5 (odds ratio of 12:1) and 5 (odds ratio of 150:1).

The evidence ratio between two models is called the Bayes Factor, $B_{01} = E_0/E_1$.

A simple example: spatial curvature

WMAP1 said $\Omega_{\rm tot} = 1.02 \pm 0.02$

This was widely interpretted as supporting the idea of a flat Universe, but actually favouring a slightly closed Universe.

Assuming that the density is the only parameter, with a uniform prior from 0.1 to 2, and likelihood

$$\mathcal{L} = \mathcal{L}_0 \exp\left(-\frac{(\Omega - 1.02)^2}{2 \times 0.02^2}\right)$$

- **Flat:** $\text{Evidence} = \mathcal{L}(\Omega = 1) = 0.6\mathcal{L}_0$
- **Curved:** $\text{Evidence} = \frac{1}{1.9}\int \mathcal{L}(\Omega) d\Omega \simeq 0.03\mathcal{L}_0$

According to the evidence, the flat model is a better description of the data, with odds of about 20:1 **against** the curved model. Note that this assumes flat and curved were thought equally likely before the data came along.

A simple example: spatial curvature

WMAP1 said $\Omega_{\rm tot} = 1.02 \pm 0.02$

This was widely interpretted as supporting the idea of a flat Universe, but actually favouring a slightly closed Universe.

Assuming that the density is the only parameter, with a uniform prior from 0.1 to 2, and likelihood

$$\mathcal{L} = \mathcal{L}_0 \exp\left(-\frac{(\Omega - 1.02)^2}{2 \times 0.02^2}\right)$$

- **Flat:** $\text{Evidence} = \mathcal{L}(\Omega = 1) = 0.6\mathcal{L}_0$
- **Curved:** $\text{Evidence} = \frac{1}{1.9}\int \mathcal{L}(\Omega) d\Omega \simeq 0.03\mathcal{L}_0$

Notes:
1) Even if parameter estimation had given $\Omega_{\rm tot} = 1.05 \pm 0.02$ the flat case would still have been preferred.
2) Someone adamantly insisting before WMAP that the total density was 1.02, to the exclusion of all other values, could claim WMAP supported them better than flat.

Calculating the evidence

The evidence is a multi-dimensional integral, a standard numerical problem. However ...

- The parameter space may have a high dimensionality (cosmological examples typically have 6 to 10 parameters simultaneously varying).
- Individual evaluations of the likelihood function may be computationally time-consuming (a few seconds each in typical cosmology examples, ie one CPU-month per million calculations).
- The likelihood function may be sharply peaked, at an unknown location.

Finding the maximum of a function < **Mapping a function in the vicinity of its maximum** < **Integrating a function over its entire domain**

Evidence calculations: `Exact' methods

These are numerical methods which would become exact in the limit of infinite computer time, and may be accurate enough for practical amounts of computer time.

- Thermodynamic integration

  A variant on Metropolis-Hastings, where the chain is given an effective temperature via the substitution $\mathcal{L} \mapsto \mathcal{L}^{1/T}$. As T goes to infinity, the posterior tends to the prior, and so the whole prior space is explored.

- Nested sampling

  Introduced by Skilling in 2004, this is the technique my group has been using - see next slides.

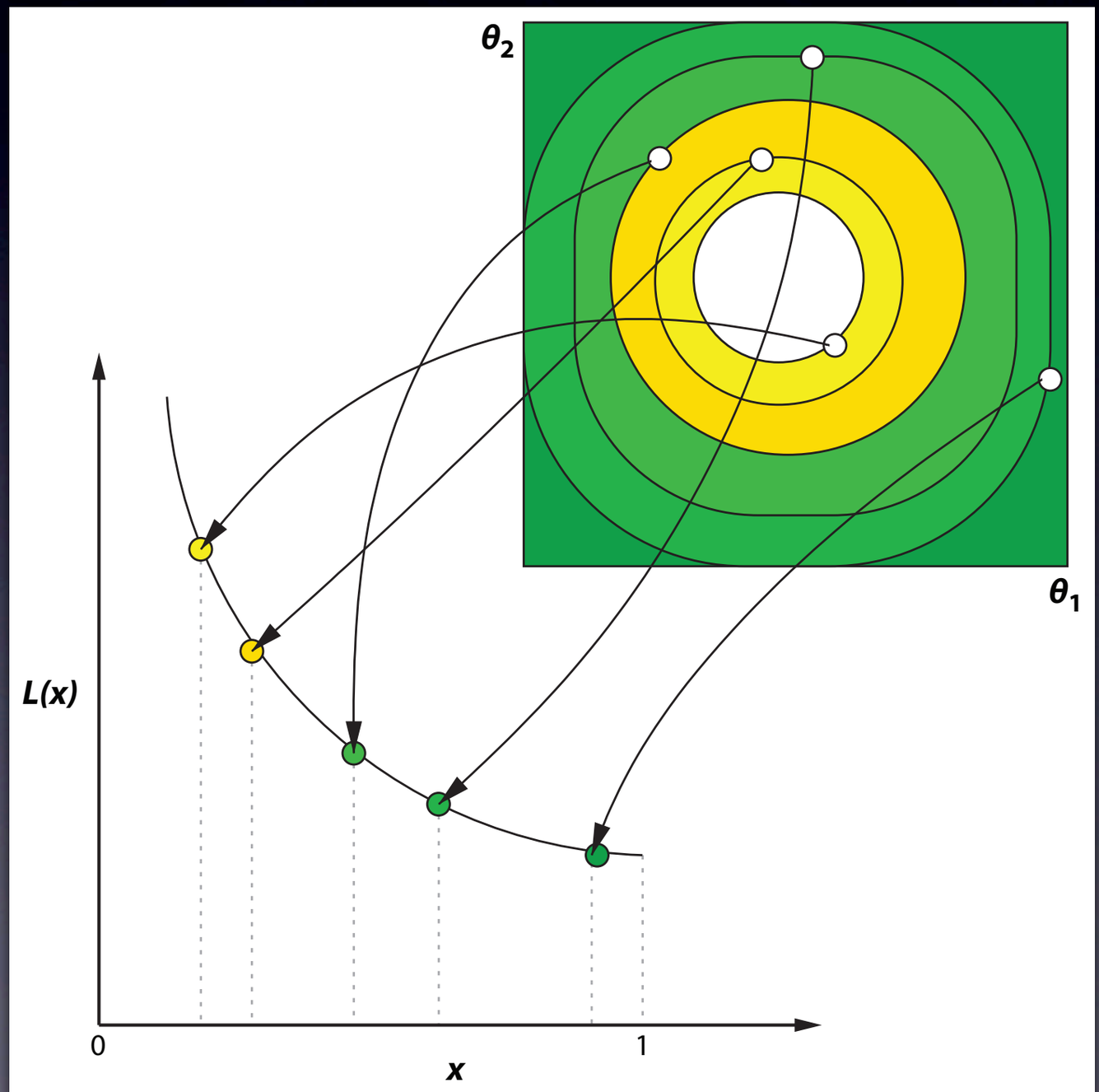
- VEGAS

  A multi-dimensional integrator popular with particle physicists, which shows promise but has been used only once for model selection so far.

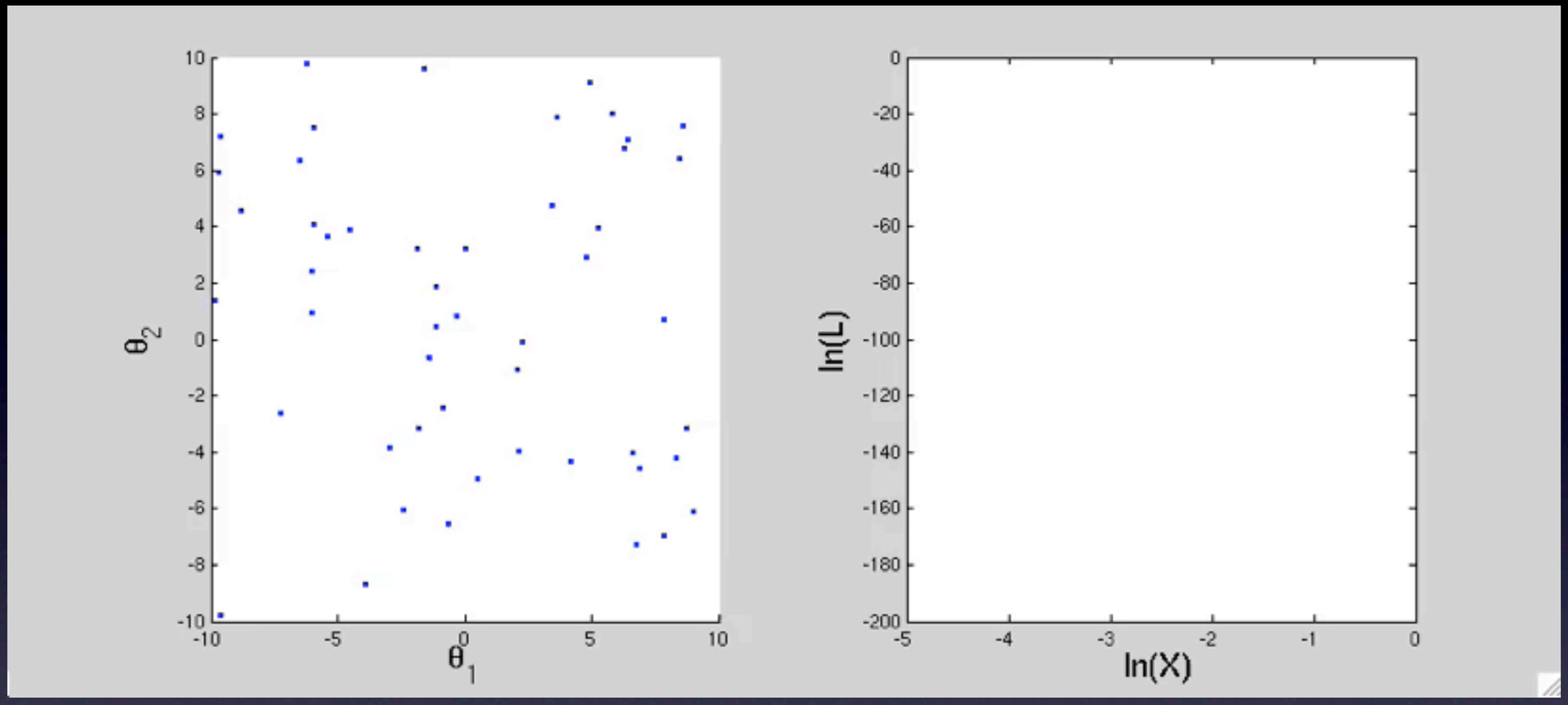
Nested sampling

Nested sampling is a Monte Carlo method, but not a Markov chain one. It computes the evidence by `walking' a set of points towards the maximum.

1. Distribute a set of points randomly within the prior, evaluating the likelihood at each.
2. Discard the lowest likelihood point.
3. Replace with a new point of higher likelihood drawn uniformly from the prior.
4. Accumulate the evidence as a 1-D integral over `prior mass'.
5. Once the remaining points are close enough to the maximum, sum over them.



Nested sampling



A toy calculation (a practical one involves several hundred points).

A substantial benefit of nested sampling is that the points it generates can be processed into a set of posterior samples suitable for parameter estimation, so it carries out both inference tasks simultaneously. These points also sample the posterior more widely than does Metropolis-Hasting.

Evidence calculations: Approximate methods

These are methods which may be faster to calculate, but which are either approximate or hold only in restricted circumstances.

- Laplace approximation

  Expand the likelihood as a multi-variate gaussian about its maximum. Unfortunately the uncertainty is not under control.

- Savage-Dickey density ratio

  For nested models, an exact relation gives the evidence in terms of the marginalized posterior of the more complex model, evaluated at the fixed parameter value(s) of the embedded model. It can be estimated from Markov chains, but sampling error can be a problem.

- Bayesian Information Criterion (BIC)

  $\mathrm{BIC} = -2 \ln \mathcal{L}_{\max} + k \ln N$ (k = number of parameters, N = number of data points).

  The BIC difference approximates the Bayes factor under certain assumptions.

Alternatives to Bayesian methods

There are some model comparison alternatives to Bayesian methods, principally based on information theory. As with the evidence, a number is calculated for each model and used to rank them.

- Akaike Information Criterion (AIC)

  $\mathrm{AIC} = -2 \ln \mathcal{L}_{\max} + 2k$ (k = number of parameters)

- Deviance Information Criterion (DIC)

  $\mathrm{DIC} = -2 \ln \mathcal{L}(\theta_{\mathrm{mean}}) + 2k_{\mathrm{eff}}$ (k_{eff} = parameters constrained by the data).

- Minimum message length/minimum description length

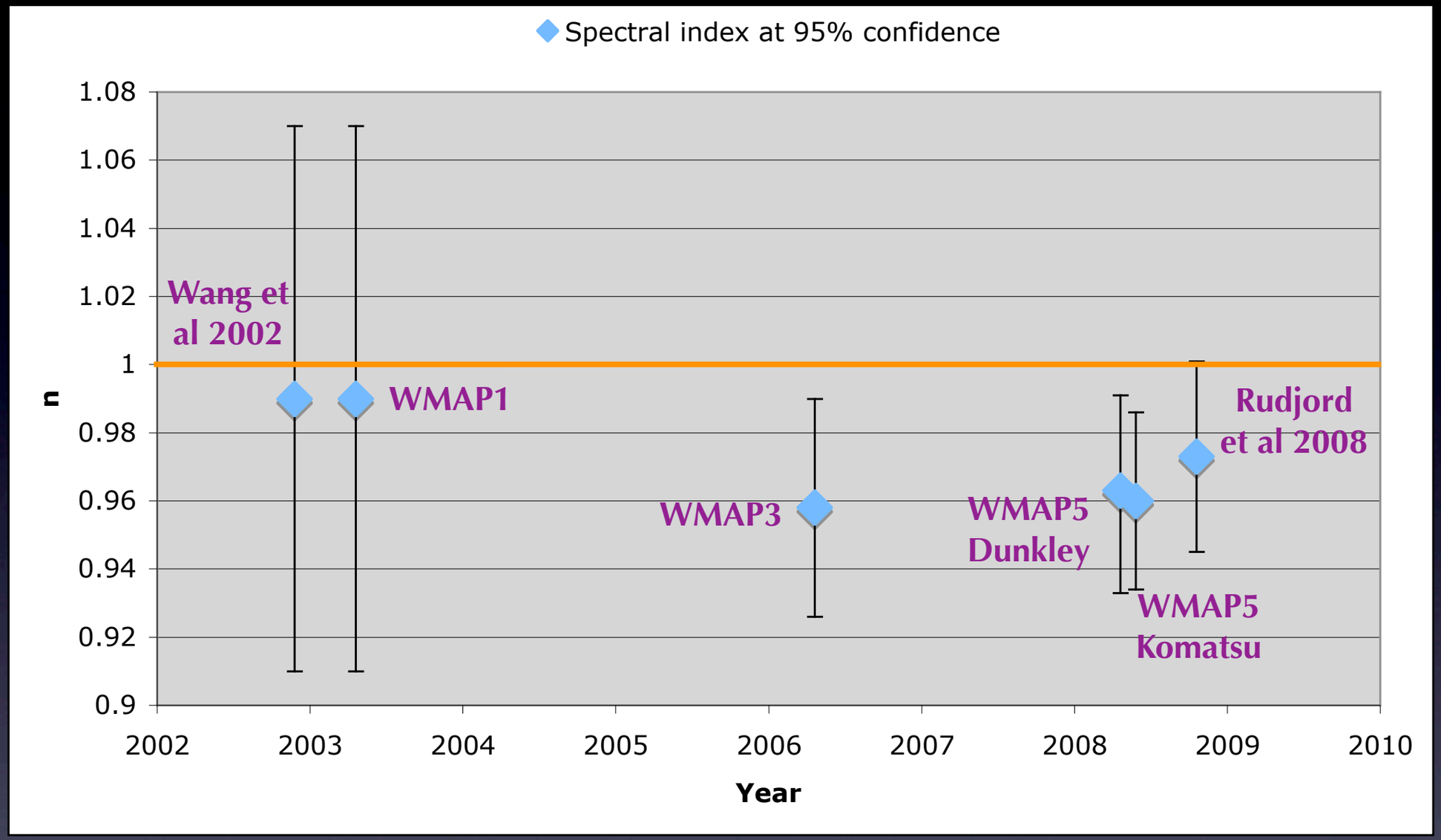
  These state that the best model offers maximum compression of the data, ie model + data residuals can be described in the smallest number of bits.

Some applications

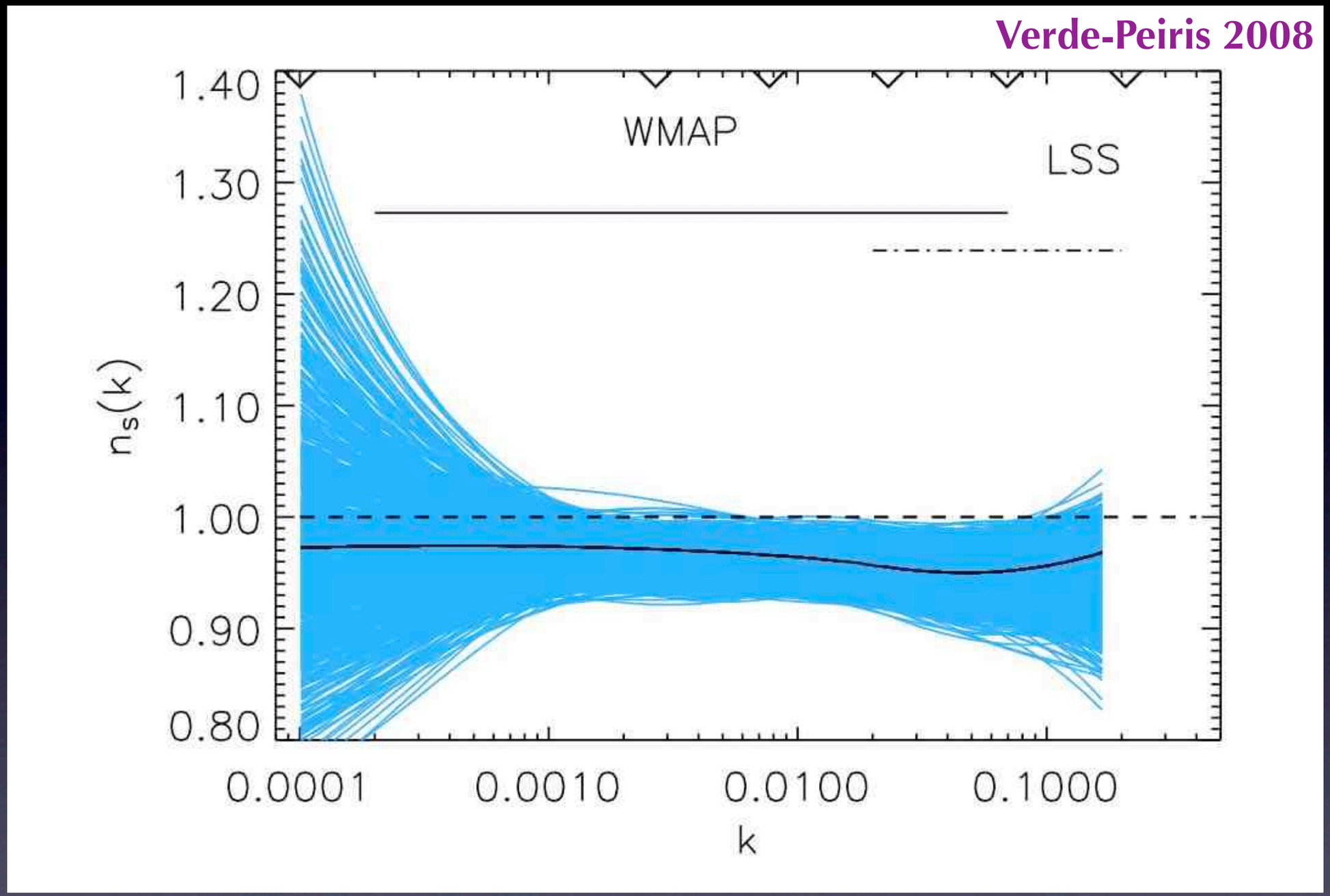
The main current controversy concerning the standard cosmological model is whether the spectral index n is a parameter, or whether instead the Harrison-Zel′dovich choice $n = 1$ is sufficient.

The argument centres around possible data analysis systematics — point source subtraction, WMAP beam profile, SZ effect marginalization — and also around methods for robust definition of the standard cosmological model.

Is *n* different from 1?



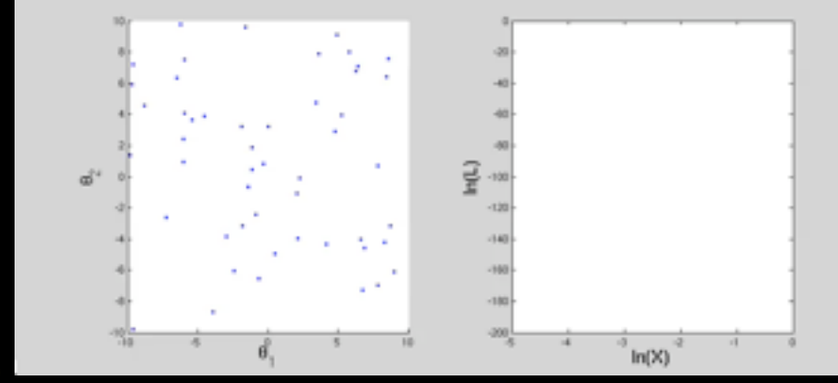
Is *n* different from 1?



Observational analyses put $n = 1$ at around the 2 to 3 sigma position.

But is that the right question to ask?

A model selection example: spectral index from WMAP3



Parkinson, Mukherjee and Liddle, PRD, astro-ph/0605003

WMAP3 has been interpretted as ruling out the Harrison-Zel'dovich $n_S = 1$ spectrum and hence favouring inflation, e.g. $n = 0.958 \pm 0.016$. But this ignores model dimensionality. Using our code CosmoNest we find

Datasets	Model	ln *E*
WMAP only	HZ varying n	0.0 0.34 ± 0.26
WMAP+all	HZ varying *n* *n* and *r* (uniform on *r*) *n* and *r* (log on *r*)	0.0 1.99 ± 0.26 -1.45 ± 0.45 1.90 ± 0.24

Datasets	Model	ln *E*
WMAP only	HZ varying n	0.0 0.34 ± 0.26
WMAP+all	HZ varying *n* *n* and *r* (uniform on *r*) *n* and *r* (log on *r*)	0.0 1.99 ± 0.26 -1.45 ± 0.45 1.90 ± 0.24

1. WMAP alone cannot distinguish between HZ and a varying spectral index.

2. Adding other datasets starts to prefer varying *n*, but only at odds of about 8:1.

3. However inflation predicts we should include both *n* and *r*, which is actually disfavoured as compared to HZ...

4. ... unless you use a logarithmic prior for *r*, which puts you back close to the *r*=0 case.

(Almost) current dark energy data

Liddle, Mukherjee, Parkinson, and Wang, PRD, astro-ph/0610126

Dark energy is the most enigmatic part of the current cosmological model, its fundamental properties being unknown.

Phenomenologically, it can be described by the equation of state w, that relates its pressure to its energy density.

Three models:

- Cosmological constant: $w = -1$.
- Constant w, to be found from fitting to data.
- Evolving w, $w = w_0 + (1-a)w_a$, where a is the scale factor, and the parameters w_0 and w_a are to be fit from data.

In the latter two cases, there are different possible choices of prior depending whether or not one wishes to allow $w < -1$.

(Almost) current dark energy data

Liddle, Mukherjee, Parkinson, and Wang, PRD, astro-ph/0610126

CMB shift+BAO(SDSS)+SN

LambdaCDM

Constant W

W0-Wa

data used				Model
WMAP+SDSS+	$\Delta \ln E$	H	$\chi^2_{\rm min}$	parameter constraints
	Model I: Λ			
Riess04	0.0	5.7	30.5	$\Omega_{\rm m} = 0.26 \pm 0.03$, $H_0 = 65.5 \pm 1.0$
Astier05	0.0	6.5	94.5	$\Omega_{\rm m} = 0.25 \pm 0.03$, $H_0 = 70.3 \pm 1.0$
	Model II: constant w, flat prior $-1 \leq w \leq -0.33$			
Riess04	-0.1 ± 0.1	6.4	28.6	$\Omega_{\rm m} = 0.27 \pm 0.04$, $H_0 = 64.0 \pm 1.4$, $w < -0.81, -0.70^a$
Astier05	-1.3 ± 0.1	8.0	93.3	$\Omega_{\rm m} = 0.24 \pm 0.03$, $H_0 = 69.8 \pm 1.0$, $w < -0.90, -0.83^a$
	Model III: constant w, flat prior $-2 \leq w \leq -0.33$			
Riess04	-1.0 ± 0.1	7.3	28.6	$\Omega_{\rm m} = 0.27 \pm 0.04$, $H_0 = 64.0 \pm 1.5$, $w = -0.87 \pm 0.1$
Astier05	-1.8 ± 0.1	8.2	93.3	$\Omega_{\rm m} = 0.25 \pm 0.03$, $H_0 = 70.0 \pm 1.0$, $w = -0.96 \pm 0.08$
	Model IV: w_0–w_a, flat prior $-2 \leq w_0 \leq -0.33$, $-1.33 \leq w_a \leq 1.33$			
Riess04	-1.1 ± 0.1	7.2	28.5	$\Omega_{\rm m} = 0.27 \pm 0.04$, $H_0 = 64.1 \pm 1.5$, $w_0 = -0.83 \pm 0.20$, $w_a = --^b$
Astier05	-2.0 ± 0.1	8.2	93.3	$\Omega_{\rm m} = 0.25 \pm 0.03$, $H_0 = 70.0 \pm 1.0$, $w_0 = -0.97 \pm 0.18$, $w_a = --^b$
	Model V: w_0–w_a, $-1 \leq w(a) \leq 1$ for $0 \leq z \leq 2$			
Riess04	-2.4 ± 0.1	9.1	28.5	$\Omega_{\rm m} = 0.28 \pm 0.04$, $H_0 = 63.6 \pm 1.3$, $w_0 < -0.78, -0.60^a$, $w_a = -0.07 \pm 0.34$
Astier05	-4.1 ± 0.1	11.1	93.3	$\Omega_{\rm m} = 0.24 \pm 0.03$, $H_0 = 69.5 \pm 1.0$, $w_0 < -0.90, -0.80^a$, $w_a = 0.12 \pm 0.22$

(Almost) current dark energy data

Liddle, Mukherjee, Parkinson, and Wang, PRD, astro-ph/0610126

CMB shift+BAO(SDSS)+SN

data used		$\Delta \ln E$
WMAP+SDSS+		
LambdaCDM	Astier05	0.0
Constant W	Astier05	-1.3 ± 0.1
	Astier05	-1.8 ± 0.1
W_0-W_a	Astier05	-2.0 ± 0.1
	Astier05	-4.1 ± 0.1

Conclusion: LambdaCDM currently favoured but all models still alive

Conclusions

- Bayesian model selection provides a rigorous approach to the comparison of competing models.
- Such techniques can positively support simpler models, and set more stringent conditions for inclusion of new parameters.
- A variety of techniques exist for calculating the Bayesian evidence. The most general can be computationally demanding.
- Alternatives to the Bayesian methodology do exist, relying on ideas from information theory/signal processing.