Bayes in the sky:

Statistical challenges in cosmology

Roberto Trotta

University of Oxford, Astrophysics
St Anne's College

Lockyer Fellow of the Royal Astronomical Society

Contemporary Physics, Vol. 00, No. 00, Month–Month 2006, 1–41

(On every Arxiv near you next week)

Bayes in the sky: Bayesian inference and model selection in cosmology

Roberto Trotta*

Oxford University, Astrophysics Department
Denys Wilkinson Building, Keble Rd, Oxford, OX1 3RH, UK

(*March 17, 2008*)

The application of Bayesian methods in cosmology and astrophysics has flourished over the past decade, spurred by data sets of increasing size and complexity. In many respects, Bayesian methods have proven to be vastly superior to more traditional statistical tools, offering the advantage of higher efficiency and of a consistent conceptual basis for dealing with the problem of induction in the presence of uncertainty. This trend is likely to continue in the future, when the way we collect, manipulate and analyse observations and compare them with theoretical models will assume an even more central role in cosmology.

This review is an introduction to Bayesian methods in cosmology and astrophysics and recent results in the field. I first present Bayesian probability theory and its conceptual underpinnings, Bayes' Theorem and the role of priors. I discuss the problem of parameter inference and its general solution, along with numerical techniques such as Monte Carlo Markov Chain methods. I then review the theory and application of Bayesian model comparison, discussing the notions of Bayesian evidence and effective model complexity, and how to compute and interpret those quantities. Recent developments in cosmological parameter extraction and Bayesian cosmological model building are summarized, highlighting the challenges that lie ahead.

Keywords: Bayesian methods; model comparison; cosmology; parameter inference; data analysis; statistical methods.

The cosmological concordance 'model'

		1D 68%	Best fit
	Cosmological parameters		
Baryon density	$\Omega_b h^2 \times 10^2$	$2.23^{+0.08}_{-0.06}$	2.28
Cold dark matter density	$\Omega_c h^2$	0.106 ± 0.004	0.107
Angular size of sound horizon	Θ_*	1.043 ± 0.003	1.042
Optical depth to reionization	τ	$0.084^{+0.014}_{-0.013}$	0.087
Expansion rate	H_0 [Km s^{-1}Mpc^{-1}]	74.3 ± 2.1	73.1
	Power spectra parameters		
Amplitude of fluctuations	$\ln(P_s^0 \times 10^{10})$	$3.11^{+0.06}_{-0.11}$	3.15
Scale dependence of fluctuations	n_0	0.973 ± 0.019	0.961

"I'm a theorist - why should I bother?"

1. Increasingly complex models and data: "chi-square by eye" simply not enough
2. "If it's real, better data will show it": but all the action is in the "discovery zone" around 3-4 sigma significance
3. Don't waste time explaining effects which are not there (e.g., reionization at $z \sim 16$)
4. Plan for the future: which is the best strategy? (survey design & optimization)
5. In some cases, there will be no better data! (cosmic variance)

Bayesian inference chain

1. Select a model
   (parameters and range)

2. Predict observational signature
   (as a function of parameters)

3. Compare with data

   a) derive parameters constraints — PARAMETER INFERENCE

   b) compute relative model probability — MODEL COMPARISON

4. Go back to 1

Challenge #1

Using the right tool for each question

or

How to distinguish between

parameter constraint and model selection tasks

Modelling it all

- *Primordial fluctuations*

A, n_s, dn/dln k, features, ...

10x10 matrix M (isocurvature)

isocurvature tilts, running, ...

Planck scale (B, ω, ϕ, ...)

Inflation (V, V', V'', ...)

Gravity waves (r, n_T, ...)

- *Matter-energy budget*

Ω_κ, Ω_Λ, Ω_{cdm}, Ω_{wdm}, Ω_ν, Ω_b

neutrino sector (N_ν, m_ν, c^2_{vis}, ...)

dark energy sector (w(z), c_s^2, ...)

baryons (Y_p, Ω_b)

dark matter sector (b, m_χ, σ, ...)

strings, monopoles, ...

- *Astrophysics*

Reionization (τ, x_e, history)

Cluster physics

Galaxy formation history

- *Exotica*

Branes, extra dimensions

Alignements, Bianchi VII models

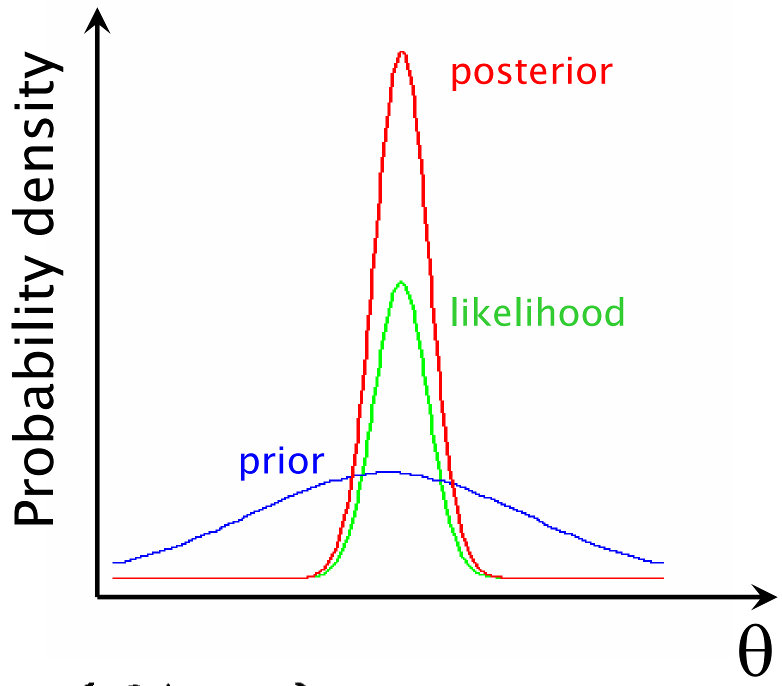
Quintessence, axions, ...

Bayes Theorem

$$\mathcal{P}(A|B) = \frac{\mathcal{P}(B|A)\mathcal{P}(A)}{\mathcal{P}(B)}$$

$A \rightarrow \boldsymbol{\theta}$: parameters

$B \rightarrow \mathrm{d}$: data



$$\mathcal{P}(\boldsymbol{\theta}|\mathbf{d}) = \frac{\mathcal{P}(\mathbf{d}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\mathcal{P}(\mathbf{d})}$$

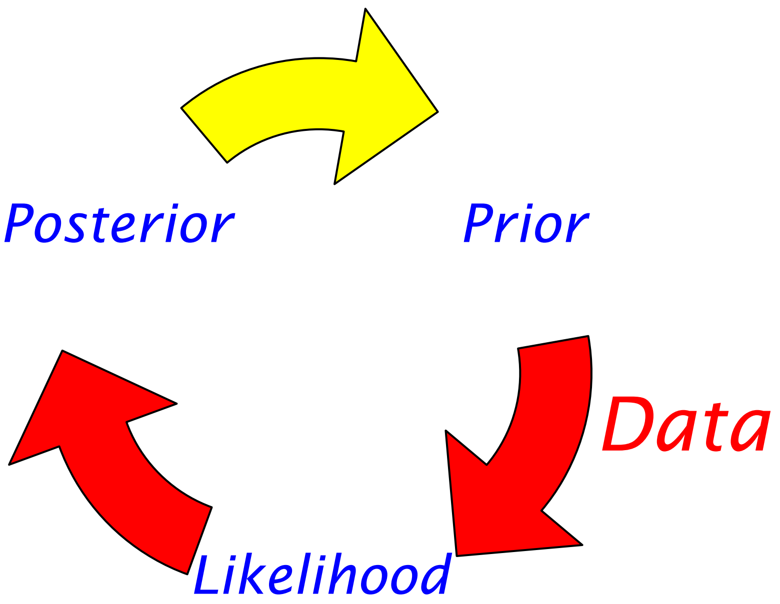
$$\mathcal{P}(\boldsymbol{\theta}|\mathrm{d},\mathcal{M}) = \frac{\mathcal{P}(\mathrm{d}|\boldsymbol{\theta},\mathcal{M})\pi(\boldsymbol{\theta}|\mathcal{M})}{\mathcal{P}(\mathrm{d}|\mathcal{M})}$$

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}}$$

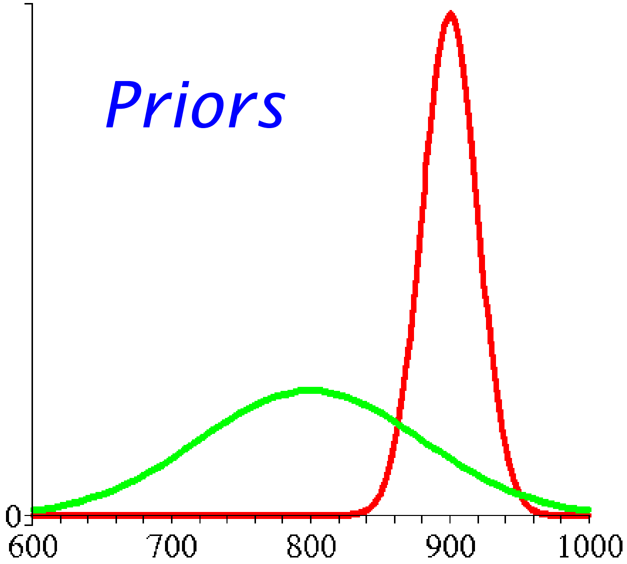
The role of priors

Parameter inference: (relatively) unproblematic

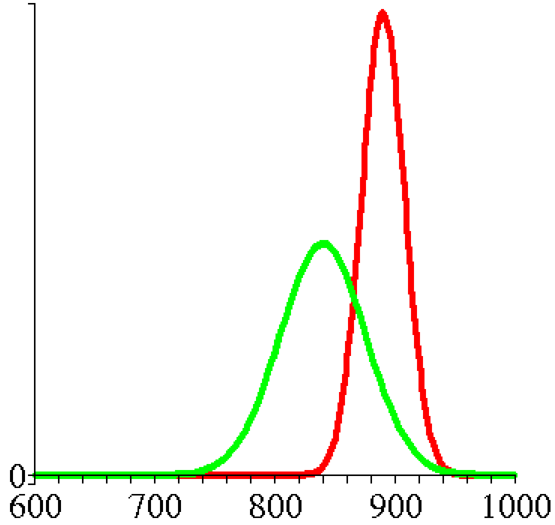
Prior as "state of knowledge"
Different people will have different priors
Updated to posterior through the data & Bayes Theorem
Will eventually go away as data become better



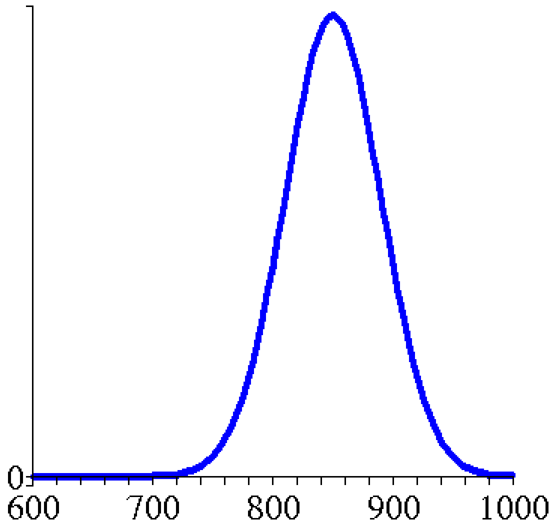
Converging views



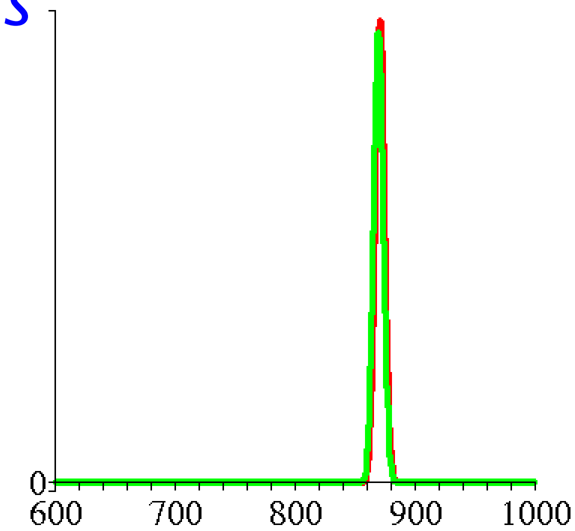
Posterior after 1 datum



Likelihood (1 datum)



Posterior after 100 data points



Bayesian model comparison

Goal: to compare the "performance" of models against the data

$$\mathcal{P}(\boldsymbol{\theta}|\mathrm{d},\mathcal{M}) = \frac{\mathcal{P}(\mathrm{d}|\boldsymbol{\theta},\mathcal{M})\pi(\boldsymbol{\theta}|\mathcal{M})}{\mathcal{P}(\mathrm{d}|\mathcal{M})}$$

The model likelihood ("Bayesian evidence")

$$\mathcal{P}(\mathrm{d}|\mathcal{M}) = \int_{\Omega} \mathcal{P}(\mathrm{d}|\boldsymbol{\theta},\mathcal{M})\pi(\boldsymbol{\theta}|\mathcal{M})\mathrm{d}\boldsymbol{\theta}$$

The posterior probability for the model

$$\mathcal{P}(\mathcal{M}|\mathrm{d}) \propto \mathcal{P}(\mathrm{d}|\mathcal{M})\pi(\mathcal{M})$$

The change in odds is given by the Bayes factor

$$\frac{\mathcal{P}(\mathcal{M}_0|\mathrm{d})}{\mathcal{P}(\mathcal{M}_1|\mathrm{d})} = \boxed{\frac{\mathcal{P}(\mathrm{d}|\mathcal{M}_0)}{\mathcal{P}(\mathrm{d}|\mathcal{M}_1)}} \frac{\pi(\mathcal{M}_0)}{\pi(\mathcal{M}_1)}$$

Jeffreys' scale for the strength of evidence

The Bayes factor

$$B_{01} = \frac{\mathcal{P}(\mathbf{d}|\mathcal{M}_0)}{\mathcal{P}(\mathbf{d}|\mathcal{M}_1)}$$

Interpretation: Jeffreys' scale for the strength of evidence

$\|\ln B_{01}\|$	Odds	Probability (2 models)	#σ	Interpretation
< 1.0	*< 3:1*	*< 0.750*	*1.15*	*not worth the mention*
< 2.5	*< 12:1*	*0.923*	*1.77*	*weak*
< 5.0	*< 150:1*	*0.993*	*2.70*	*moderate*
>5.0	*> 150:1*	*> 0.993*	*> 2.70*	*strong*

Priors and model comparison

Model comparison takes Bayesian inference to the model level - it complements parameter inference

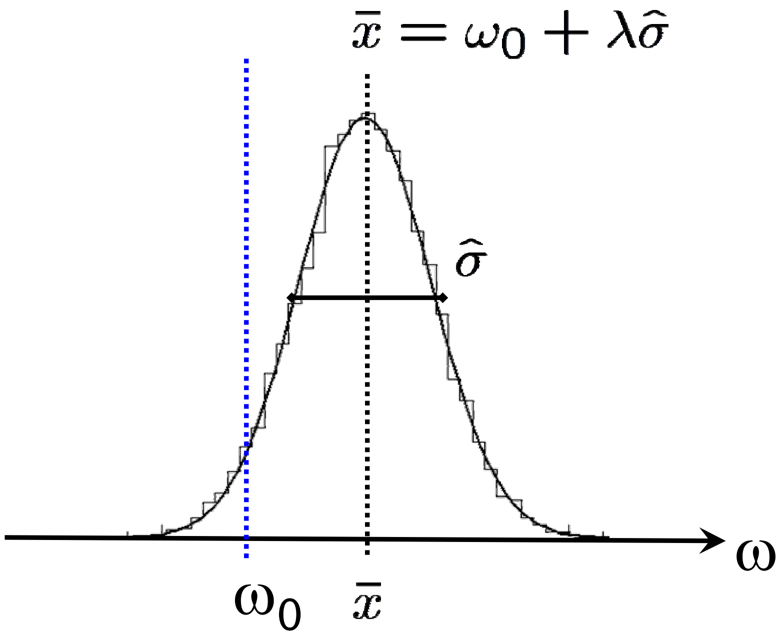
Prior choice is inherent to model specification
Gives available model parameter space
Related to physical insight into the model

A model is a choice of parameters and their ranges

$$\mathcal{M} \equiv \{\theta, \pi(\theta|\mathcal{M})\}$$

An automatic Occam's razor

The Bayes factor balances quality of fit vs extra model complexity. It rewards highly predictive models



Model 0: $\omega = \omega_0$

Model 1: $\omega \neq \omega_0$ with $\pi(\omega)$

For "informative" data

$$\ln B_{01} = I - \frac{\lambda^2}{2}$$

I = ln(prior width / likelihood width) $\geq$ 0

= "wasted" volume of parameter space

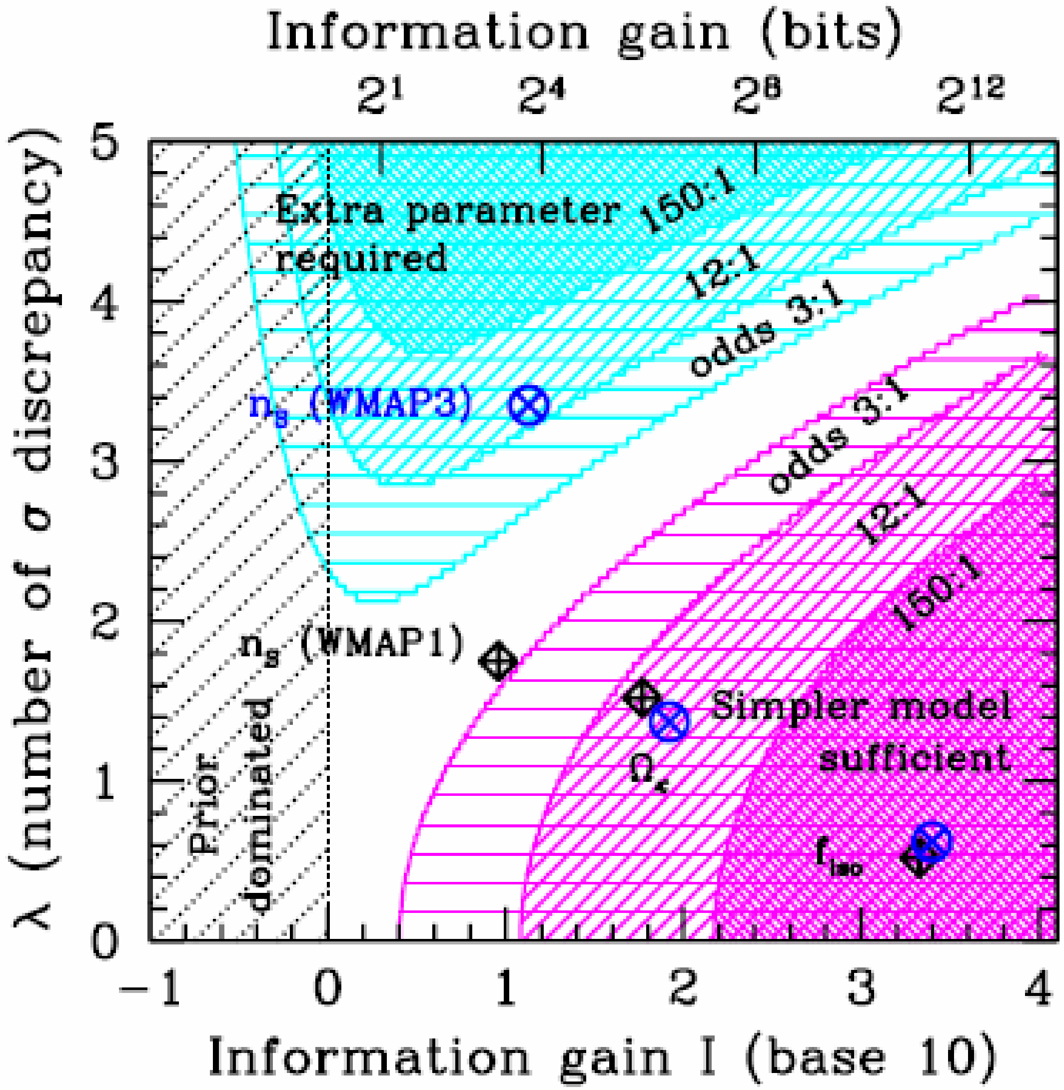
= amount by which our knowledge has increased

Cosmological applications

Spectral index:
$P(n_s=1)$ vs $0.8 < n_s < 1.2$

WMAP1+:
$\ln B_{01} = 0.7$ (2:1)

WMAP3+:
$\ln B_{01} = -2.9$ (1:17)



Trotta 2007, MNRAS astro-ph/0504022
Trotta 2007, MNRAS Lett, astro-ph/0703063

ΛCDM is in the lead

Trotta (2008)

Bayes factor: ln B < 0 favours ΛCDM

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

Challenge #2

What is a "significant" effect?

or

How not to loose your sleep over

2-sigma "detections" (and why)

Frequentist hypothesis testing

Frequentist hypothesis testing (eg: likelihood ratio) is not what you think it is

A 2-sigma result does not wrongly reject the null hypothesis 5% of the time: at least 29% of 2-sigma results are wrong!

Take an equal mixture of H_0, H_1
Simulate data, perform hypothesis testing for H_0
Select results rejecting H_0 at 1-α CL
What fraction of those results did actually come from H_0 ("true nulls", should not have been rejected)?

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

What went wrong?

For details see: Sellke, Bayarri & Berger, *The American Statistician*, 55, 1 (2001)

The fundamental mistake is to confuse

$$\underbrace{\mathcal{P}(\text{data}|\text{hypothesis})}_{\text{p-value}} \neq \underbrace{\mathcal{P}(\text{hypothesis}|\text{data})}$$

p-value
Frequentist hypothesis testing

Requires Bayes theorem!

1) Hypothesis (θ): is a random person female?
2) Gather data: “pregnant = Y/N”
3) ...Don’t get confused!

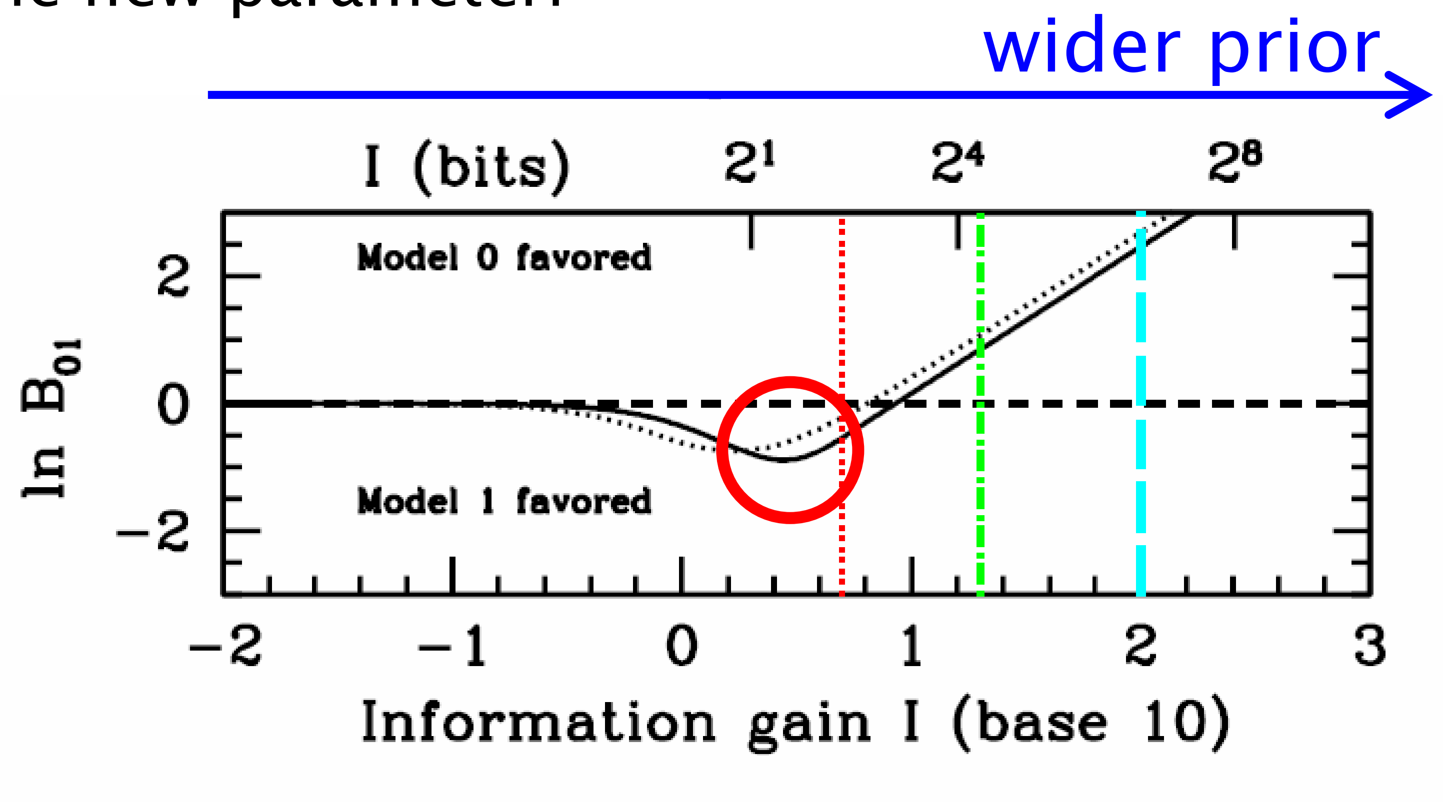
$$\mathcal{P}(\mathbf{d}=\mathbf{Y}|\boldsymbol{\theta}=\mathbf{F})=\mathbf{0.03}$$

$$\mathcal{P}(\boldsymbol{\theta}=F|\mathbf{d}=\mathbf{Y})\gg\mathbf{0.03}$$

Maximising support for the alternative

A Bayesian step is required to obtain the probability of the hypothesis ("model")

When a meaningful prior is not easy to derive, we can still employ an upper bound on the evidence in favour of the new parameter:



Bayesian calibrated p-values

Sellke & Berger (1987), Gordon & Trotta (2007), MNRASLett, arxiv:0706.3014

p-value: $\wp = p(t \geqslant t_{\rm obs}(x)|M_0)$

For a wide class of unimodal, symmetric priors around ω_0 one can prove that, for all priors

$$B \leqslant \bar{B} = \frac{-1}{e\wp \ln \wp}$$

where

$$B = B_{10} = \frac{\mathcal{P}(\mathrm{d}|\mathcal{M}_1)}{\mathcal{P}(\mathrm{d}|\mathcal{M}_0)}$$

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

A conversion table

Gordon & Trotta (2007), MNRASLett

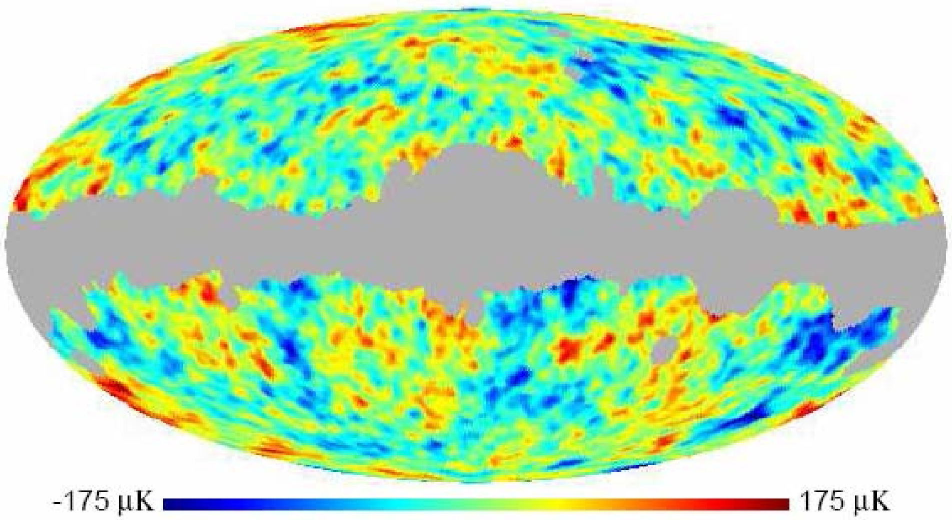
Significance p–value	Bayesian evidence bound $\bar{B}$	$\ln \bar{B}$	sigma	Interpretation (Jeffreys' scale) category
0.05	2.5	0.9	2.0	
0.04	2.9	1.0	2.1	'weak' at best
0.01	8.0	2.1	2.6	
0.006	12	2.5	2.7	'moderate' at best
0.003	21	3.0	3.0	
0.001	53	4.0	3.3	
0.0003	150	5.0	3.6	'strong' at best
6×10^{-7}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

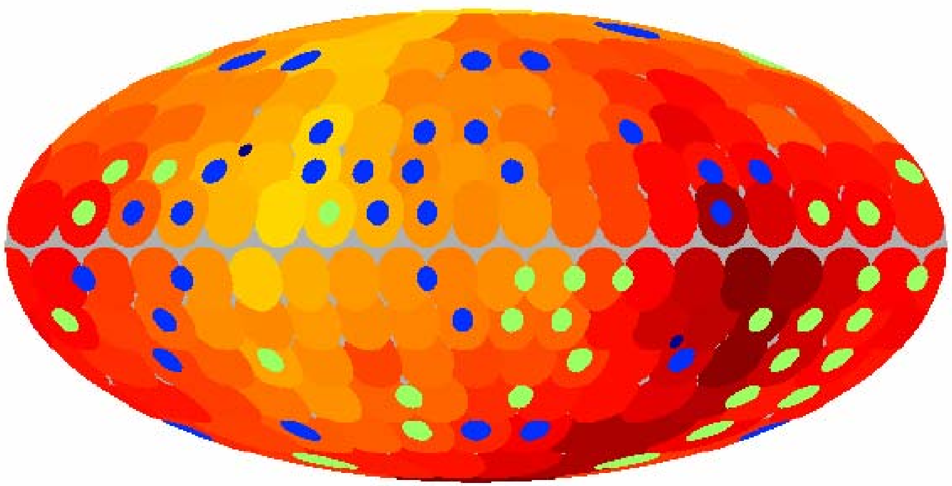
Rule of thumb:

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

Hemispheric asymmetry in the CMB

WMAP1 ILC maps





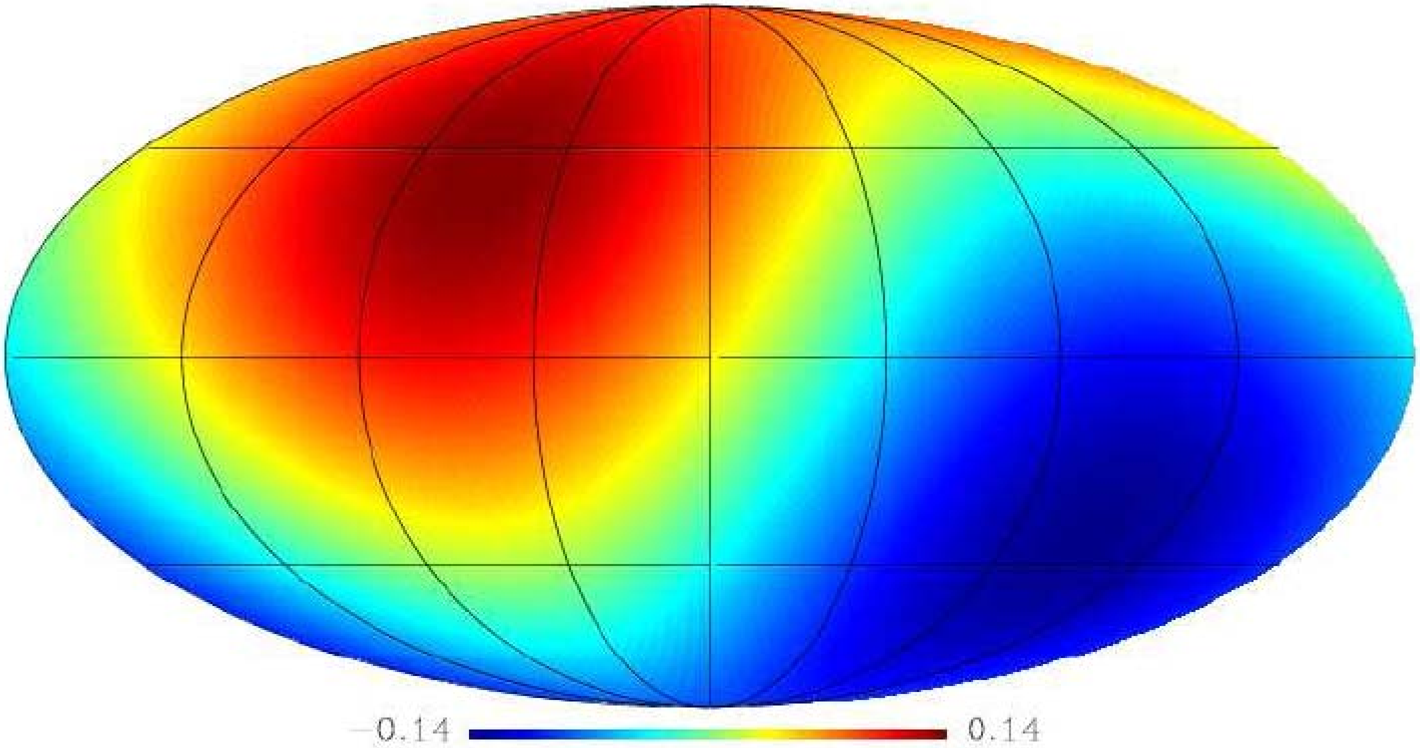
Eriksen et al (2004)

WMAP-3 maps

Gordon (2007)
Gordon & Trotta (2007)

Introduce dipolar modulating function

$$\Delta T(\hat{\mathbf{n}}) = \Delta T_{\rm iso}(\hat{\mathbf{n}})(1 + A\hat{\mathbf{n}} \cdot \hat{\mathbf{d}})$$



$\Delta\chi^2 = 9$ for 3 extra parameters: is this significant?
Bayesian evidence upper bound of 9:1 (weak support)

Challenge #3

Predicting the outcome of future observations

or

How to exploit the known unknowns

Predictive distributions

Goal: Probability distribution for the outcome of a future observation averaging over current parameters and model uncertainty

Multi-model inference: Bayesian generalization of Fisher matrix forecast

o: current experiment
e: future experiment
θ: future max like value

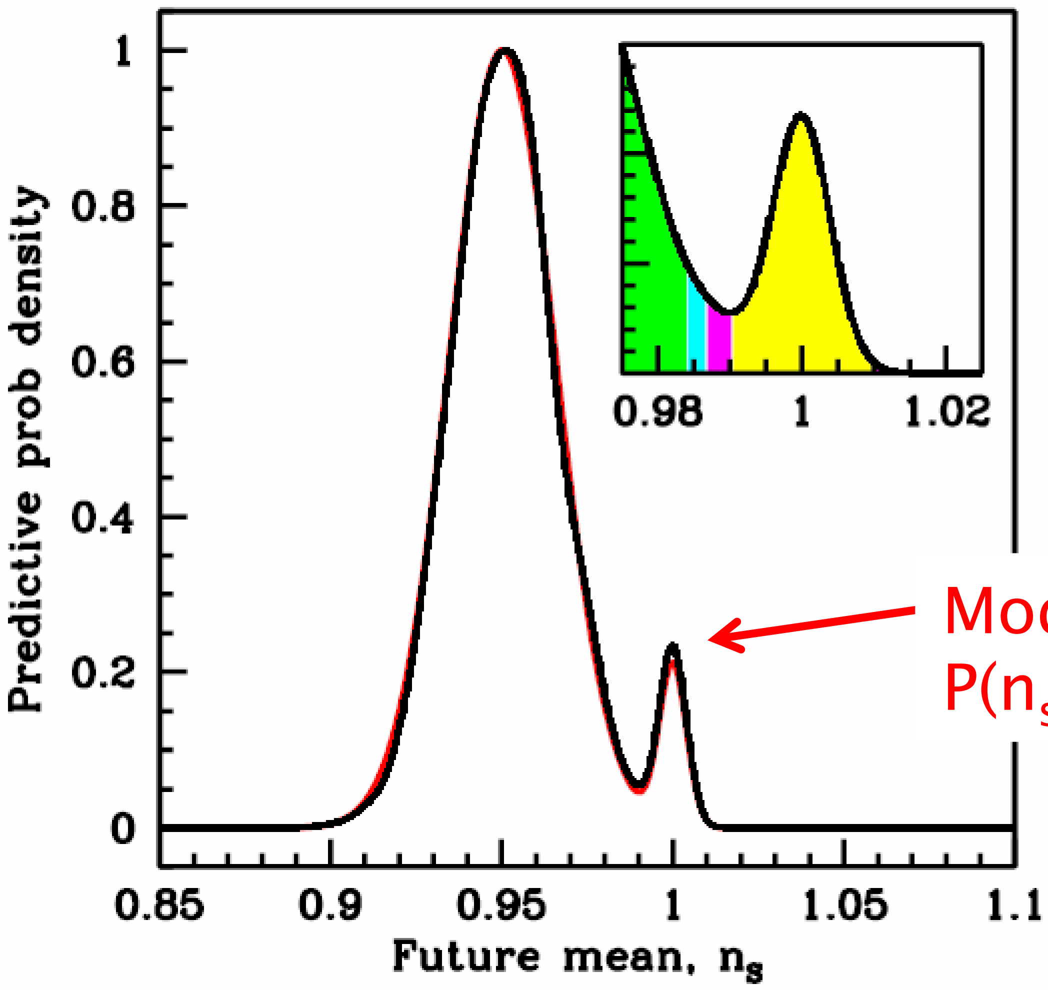
$$P(\theta|o,e) = \sum_i P(\theta|o,e,\mathcal{M}^{(i)})P(\mathcal{M}^{(i)}|o)$$
$$= \sum_i P(\mathcal{M}^{(i)}|o) \int P(\theta|\hat{\theta}^{(i)},e,\mathcal{M}^{(i)})P(\hat{\theta}^{(i)}|o,\mathcal{M}^{(i)})\mathrm{d}\hat{\theta}^{(i)},$$

current model posterior — Fisher matrix for fiducial point — present posterior (weight)

Application: the PPOD technique

Trotta 2007, MNRAS, astro-ph/0703063

*Prediction of the value of n_s from Planck data given current data (WMAP3+others): **Predictive Posterior Odds Distribution***



Planck forecasts

$n_s = 1$ vs $0.8 < n_s < 1.2$

P(lnB < -5) = 0.93

P(-5<lnB<0) = 0.01

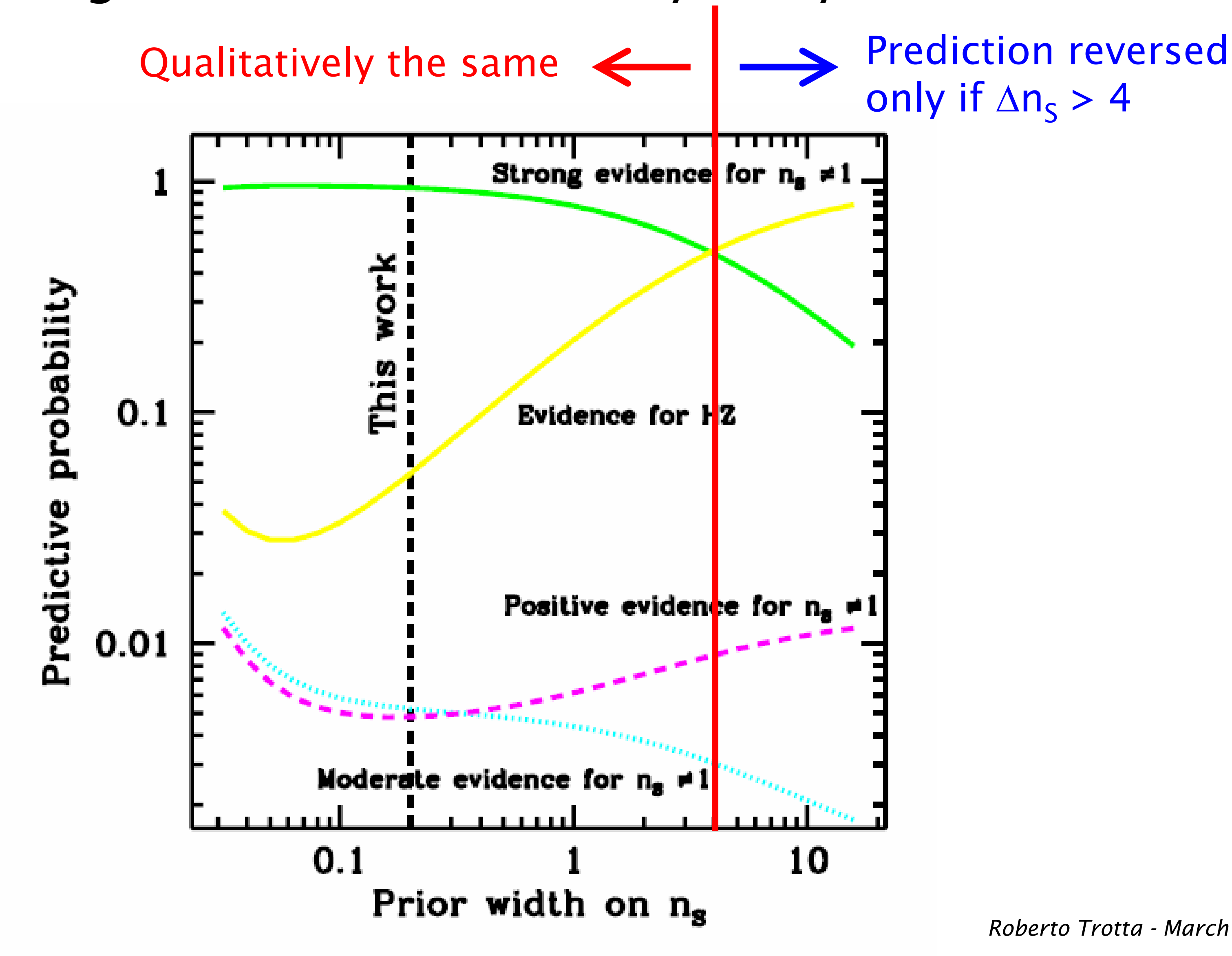
P(lnB > 0) = 0.06

Model uncertainty

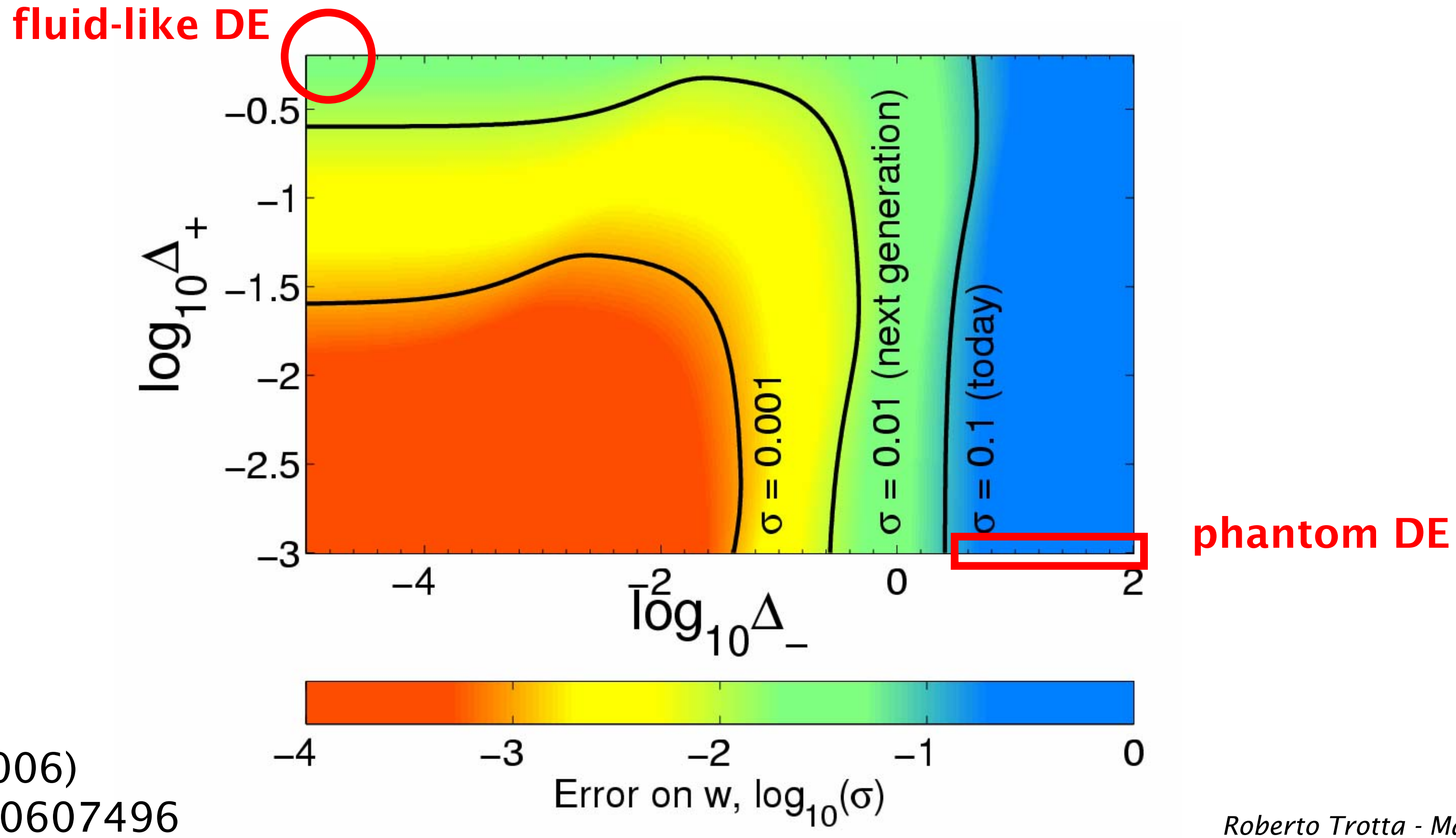
$P(n_s=1|WMAP3+) = 0.05$

Prior dependence

Prior scaling of PPOD derived analytically from SDDR



Model	(Δ_+, Δ_-)	Required σ for evidence level moderate ($\ln B = 3.0$)	strong ($\ln B = 5.0$)
Phantom	$(0, 10)$	0.4	$5 \cdot 10^{-2}$
Fluid–like	$(2/3, 0)$	$3 \cdot 10^{-2}$	$3 \cdot 10^{-3}$
Small departures	$(0.01, 0.01)$	$4 \cdot 10^{-4}$	$5 \cdot 10^{-5}$



Trotta (2006)
astro-ph/0607496

Conclusions - meeting the challenge

Bayesian tools provide a framework for new questions & approaches:

- *Model building: phenomenologically work which is the "best" model. Needs model insight (prior).*
- *Experiment design: what is the best strategy to discriminate among models?*
- *Performance forecast: how well must we do to reach a certain level of evidence?*
- *Science return optimization: use present-day knowledge to optimize future searches*