Testing for Tensions Between Datasets

David Parkinson
University of Queensland

In collaboration with
Shahab Joudaki (Oxford)

Outline

- Introduction
- Statistical Inference
- Methods
- Linear models
- Example using WL and CMB data
- Conclusions

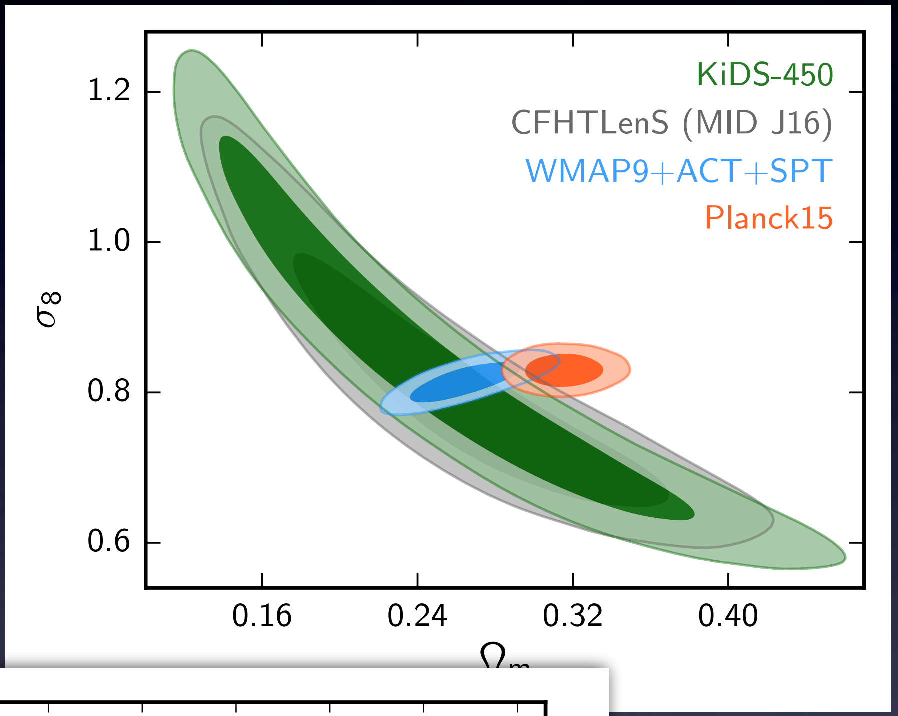
What is Probability?

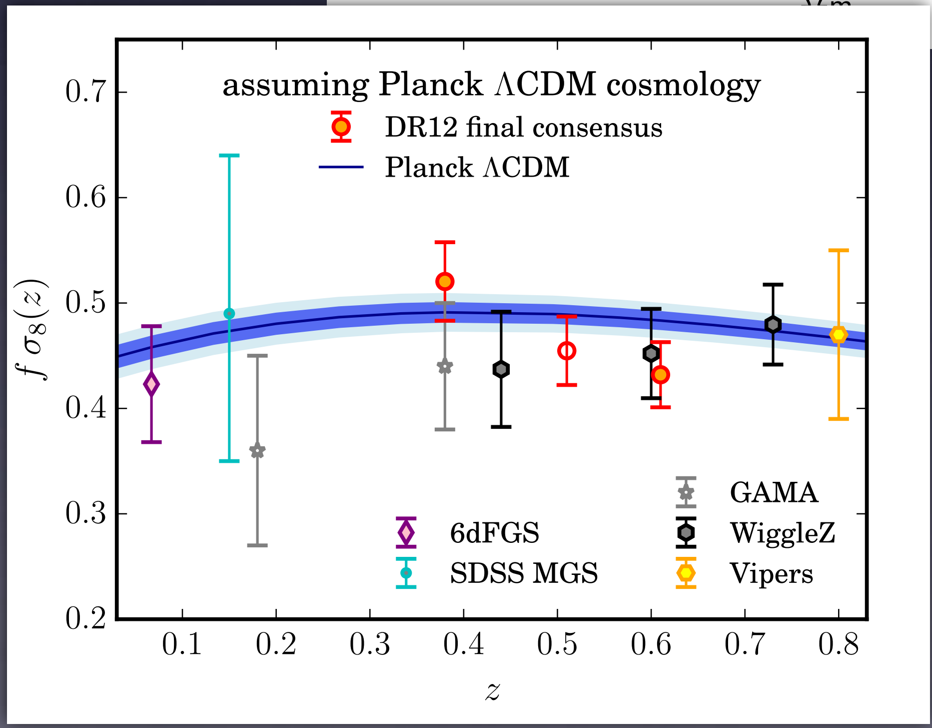
- In 1812 Laplace published *Analytic Theory of Probabilities*
- He suggested the computation of *"the probability of causes and future events, derived from past events"*
- *"Every event being determined by the general laws of the universe, there is only probability relative to us."*
- *"Probability is relative, in part to [our] ignorance, in part to our knowledge."*
- So to Laplace, probability theory is applied to our level of knowledge

Pierre-Simon Laplace

Comparing datasets

- As there is only one Universe (setting aside the Multiverse), we make observations of un-repeatable ‘experiments’
- Therefore we have to proceed by inference
- Furthermore we cannot check or probe for biases by repeating the experiment - we cannot ‘restart the Universe’ (however much we may want to)
- If there is a tension (i.e. if two data sets don’t agree), can’t take the data again. Need to instead make inferences with the data we have

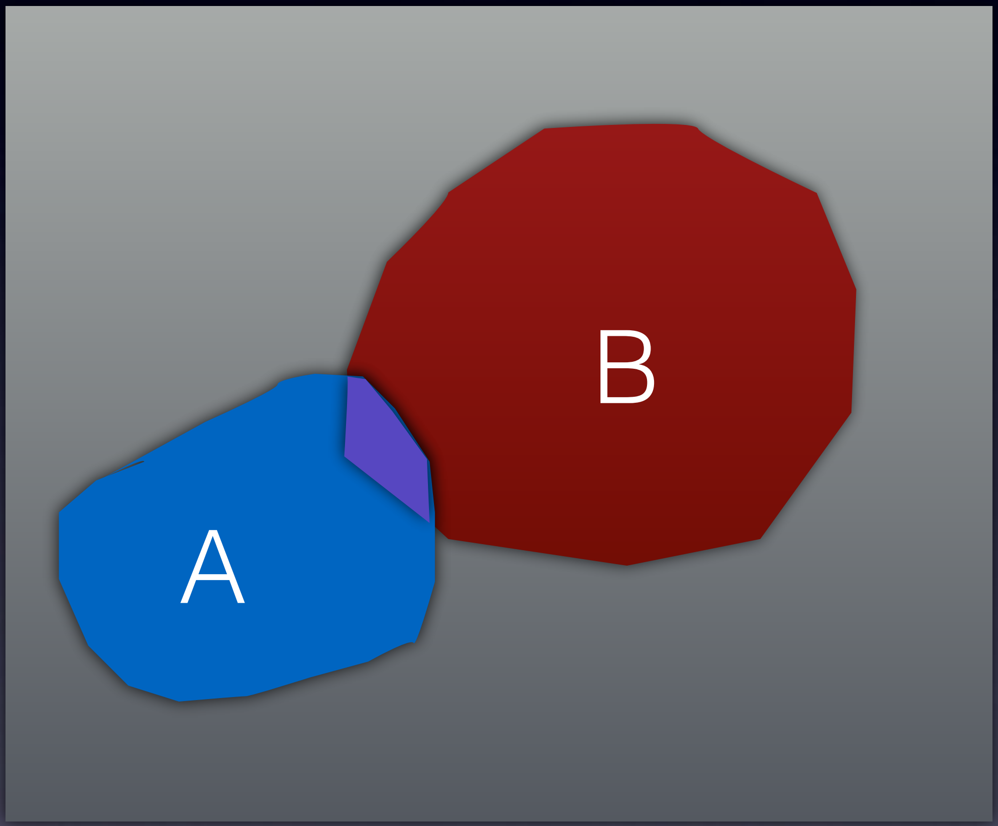




Alam et al 2016

Rules of Probability

- We define Probability to have numerical value
- We define the lower bound, of logical absurdities, to be zero, $P(\varnothing)=0$
- We normalize it so the sum of the probabilities over all options is unity, $\sum P(Ai)\equiv 1$



Sum Rule: $P(A\cup B)=P(A)+P(B)-P(A\cap B)$

Product Rule: $P(A\cap B)=P(A)P(B|A)=P(B)P(A|B)$

Bayes Theorem

- Bayes theorem is easily derived from the product rule

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

- We have some model M, with some unknown parameters θ, and want to test it with some data D

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

posterior: $P(\theta|D,M)$; likelihood: $P(D|\theta,M)$; prior: $P(\theta|M)$; evidence: $P(D|M)$

- Here we apply probability to models and parameters, as well as data

Model Selection

- If we marginalize over the parameter uncertainties, we are left with the marginal likelihood, or evidence

evidence likelihood prior

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

- If we compare the evidences of two different models, we find the Bayes factor

Model posterior evidence Model prior

$$\frac{P(M_1|D)}{P(M_2|D)}=\frac{P(D|M_1)P(M_1)}{P(D|M_2)P(M_2)}$$

- Bayes theorem provides a consistent framework for choosing between different models

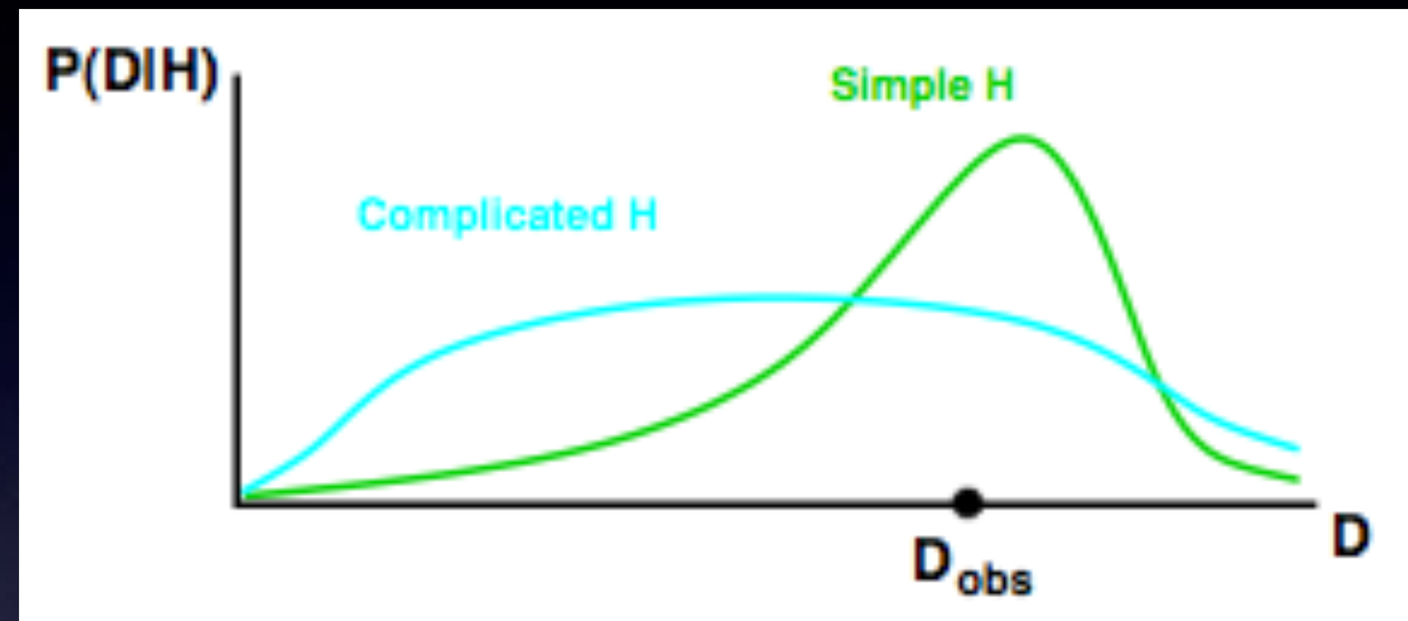
Occam's Razor

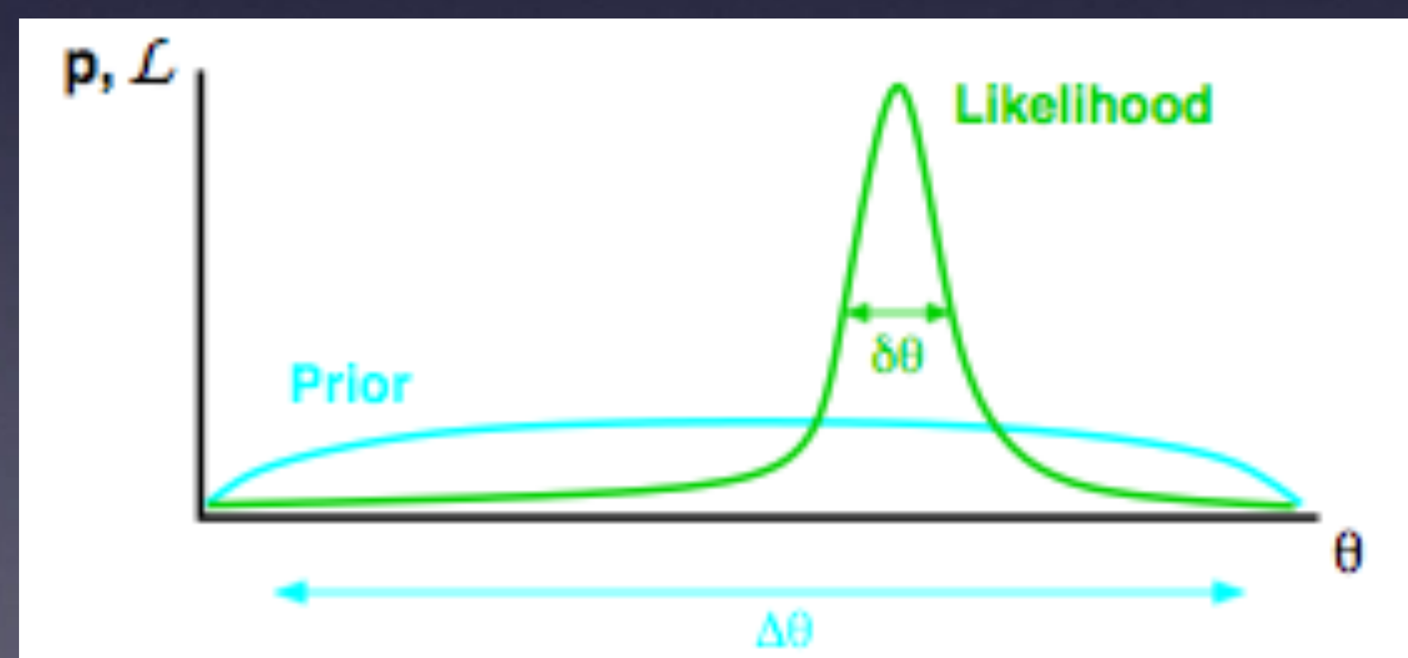
$$E = \int d\theta P(D|\theta, \mathcal{M}) P(\theta|\mathcal{M})$$

$$\approx P(D|\hat{\theta}, \mathcal{M}) \times \frac{\delta\theta}{\Delta\theta}$$

Best fit likelihood

Occam factor





- Occam factor rewards the model with the least amount of wasted parameter space ("most predictive")

Bayesian Model Comparison

- Jeffrey's (1961) scale:

Difference	Jeffrey	Trotta	Odds
$\Delta\ln(E)<1$	No evidence	No	3:1
$1<\Delta\ln(E)<2.5$	substantial	weak	12:1
$2.5<\Delta\ln(E)<5$	strong	moderate	150:1
$\Delta\ln(E)>5$	decisive	strong	>150:

- If model priors are equal, evidence ratio and Bayes factor are the same

Information Criteria

- Instead of using the Evidence (which is difficult to calculate accurately) we can approximate it using an Information Criteria statistic
- Ability to fit the data (chi-squared) penalised by (lack of) predictivity
- Smaller the value of the IC, the better the model
- Bayesian Information Criterion

$$\mathrm{BIC} = \chi^2(\hat{\theta}) + k \ln N$$

  - k is the number of free parameters and N is the number of data points
- Deviance Information Criterion (Spielgelhalter et al. 2002)

$$\mathrm{DIC} = \chi^2(\hat{\theta}) + 2c$$

- Here c is the complexity, which is equal to number of well measured parameters

Complexity

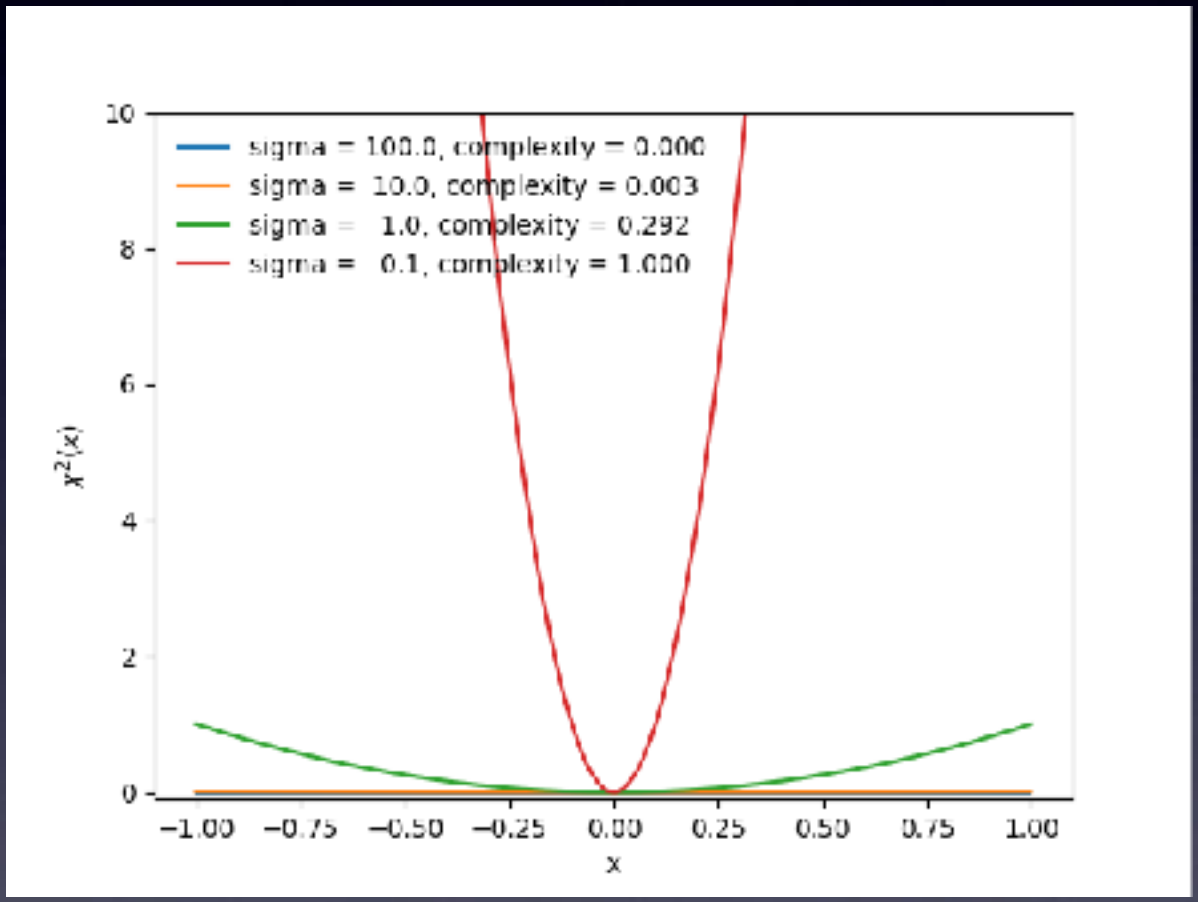
- The DIC penalises models based on the *Bayesian complexity*, the number of well-measured parameters
- This can be computed through the information gain (KL divergence) between the prior and posterior, minus a point estimate

$$\mathcal{C}_b = -2\left(D_{\mathrm{KL}}\left[P(\theta|D,\mathcal{M})P(\theta|\mathcal{M})\right] - \widehat{D_{\mathrm{KL}}}\right)$$

- For the simple gaussian likelihood, this is given by

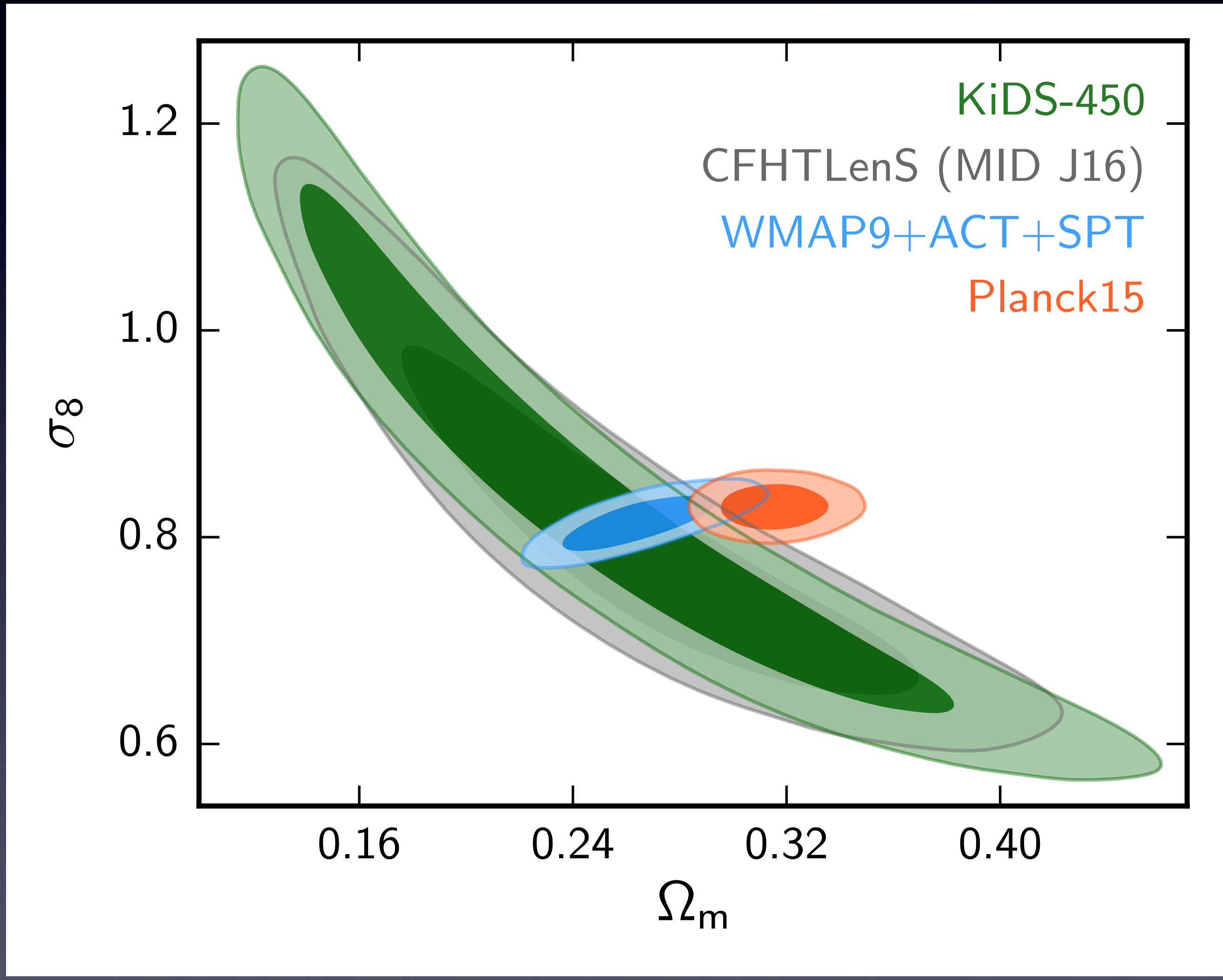
$$\mathcal{C}_b = \overline{\chi^2(\theta)} - \chi^2(\bar{\theta})$$

- Average is over posterior



Tensions

- Tensions occur when two datasets have different preferred values (posterior distributions) for some common parameters
- This can arise due to
  - random chance
  - systematic errors
  - undiscovered physics



Diagnostic statistics

- Need to diagnose not if the model is correct, but if the tension is significant
- Simple test χ^2 per degree of freedom
  - Equivalent to p-value test on data
  - Only a point estimate though
- Raveri (2015): the evidence ratio

$$\mathcal{C}(D_1, D_2, \mathcal{M}) = \frac{P(D_1 \cup D_2 | \mathcal{M})}{P(D_1|\mathcal{M})P(D_2|\mathcal{M})}$$

- Joudaki et al (2016): change in DIC

$$\Delta\mathrm{DIC} = \mathrm{DIC}(D_1 \cup D_2) - \mathrm{DIC}(D_1) - \mathrm{DIC}(D_2)$$

Linear evidence

2 3

$$P(D|\mathcal{M}) = \mathcal{L}_0 \frac{|F|^{-1/2}}{|\Pi|^{-1/2}} \exp\left[-\frac{1}{2}(\theta_L^T L \theta_L + \theta_\pi^T \Pi \theta_\pi - \bar{\theta}^T F \bar{\theta})\right]$$

1

- Evidence in linear case dependent on
  1. likelihood normalisation
  2. Occam factor (compression of prior into posterior)
  3. Displacement between prior and posterior
- In linear case, final Fisher information matrix is sum of prior and likelihood (F=L+Π)
- If prior is wide, Π is small (so displacement minimised), but Occam factor larger

Simple linear model

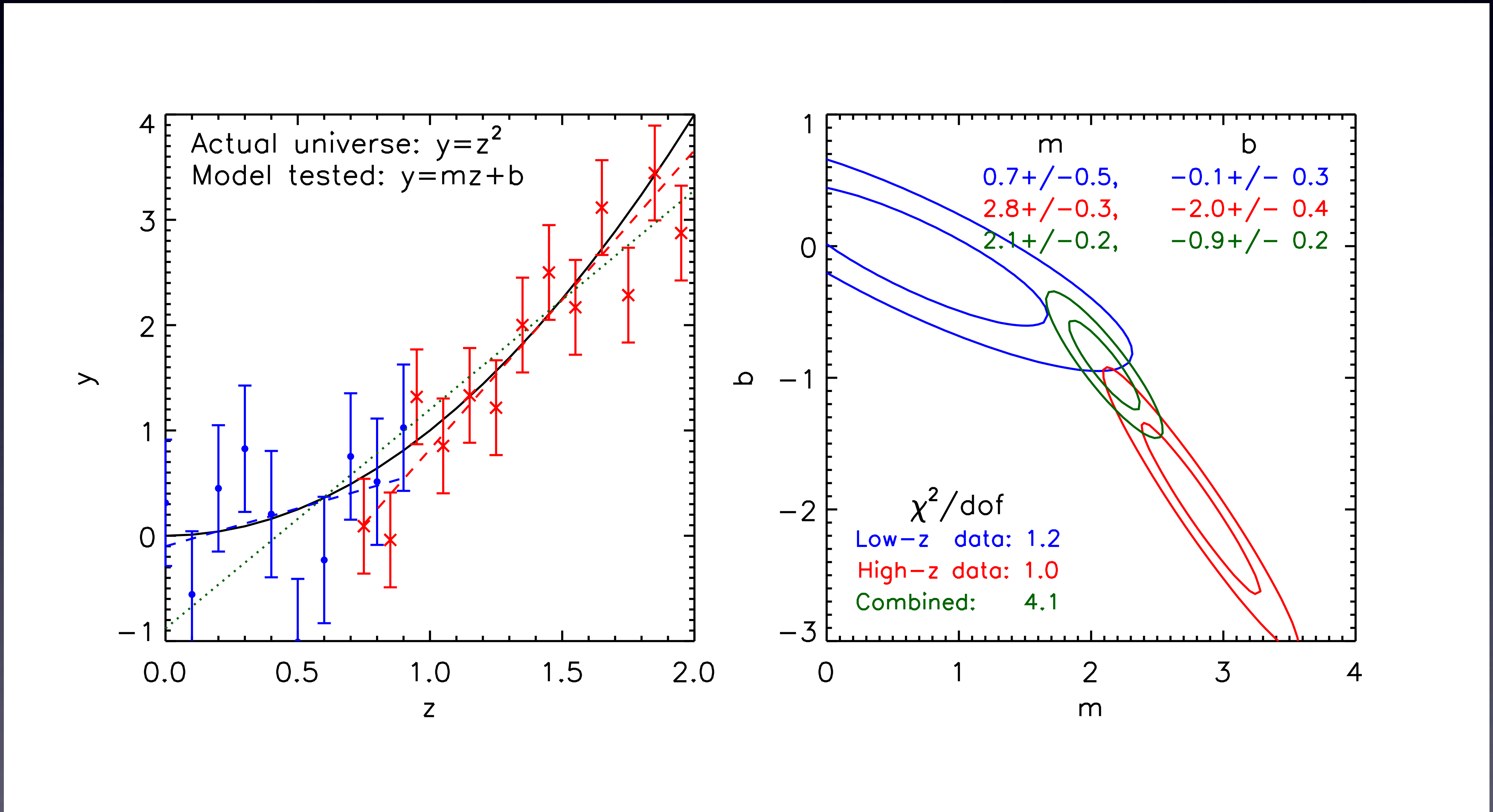


Image credit: Tamara Davis

Diagnostics II: The Surprise

- Seehars et al (2016): the ‘Surprise’ statistic, based on cross entropy of two distributions
- Cross entropy given by KL divergence between original (D_1) and updated dataset (D_2)

$$D_{\mathrm{KL}}\left(P(\theta|D_2)||P(\theta|D_1)\right) = \int P(\theta|D_2)\log\left[\frac{P(\theta|D_2)}{P(\theta|D_1)}\right]$$

- Surprise is difference of observed KL divergence relative to expected
  - where expected assumes consistency

$$S \equiv D_{\mathrm{KL}}\left(P(\theta|D_2)||P(\theta|D_1)\right) - \langle D \rangle$$

- One data set is assumed to be ‘ground-truth’, and information gain is considered in light on updating, or additional

Linear tension

$$\frac{P(D_{1+2}|\mathcal{M})}{P(D_1|\mathcal{M})P(D_2|\mathcal{M})} = \frac{\mathcal{L}_0^{1+2}}{\mathcal{L}_0^1\mathcal{L}_0^2} \times \frac{|F_{1+2}|^{-1/2}}{|F_1|^{-1/2}|F_2|^{-1/2}} \times \text{displacement terms}$$

- Displacement terms equivalent to `Surprise' - relative entropy between two distributions
- Occam factor independent of tensions
- Tensions manifest in first and third terms - best fit likelihood and displacement

Linear DIC

- ΔDIC statistic has two components
  - Difference in mean parameter (best fit) likelihood

$$\Delta\chi^2 = \chi^2_{1+2} - \chi^2_1 - \chi^2_2$$

  - Difference in penalty term (complexity)

$$\Delta\mathcal{C}_b = \mathcal{C}_{b1+2} - \mathcal{C}_{b1} - \mathcal{C}_{b2}$$

- In linear case, final Fisher matrix is the sum of individual matrices, so complexity doesn't change
  - Tension statistic (in linear case) driven entirely by difference in best likelihood

Linear Surprise

- Surprise is difference between information gain (going from data set D_1 to D_2) and expected information gain
- In the linear case, KL divergence can be

$$D_{\rm KL} = -\frac{1}{2}\left[\overline{\chi^2_{1+2}(\theta)} - \overline{\chi^2_1(\theta)}\right]$$

- For the expectation of the information gain, need to average over possible outcomes for the combined data set
  - But in the linear case, this corresponds to the maximum likelihood, where the information gain is evaluated at the posterior maximum
  - $$\langle D \rangle = -\frac{1}{2}\left[\chi^2_{1+2}(\bar\theta) - \chi^2_1(\bar\theta)\right]$$
- This is not the same as the complexity change, even though it looks similar, as the averaging process happens over the final posterior, not individual ones

$$S = D_{\rm KL} - \langle D \rangle = \frac{1}{2}\left[\chi^2_{1+2}(\bar\theta) - \chi^2_1(\bar\theta) - (\overline{\chi^2_{1+2}(\theta)} - \overline{\chi^2_1(\theta)})\right]$$

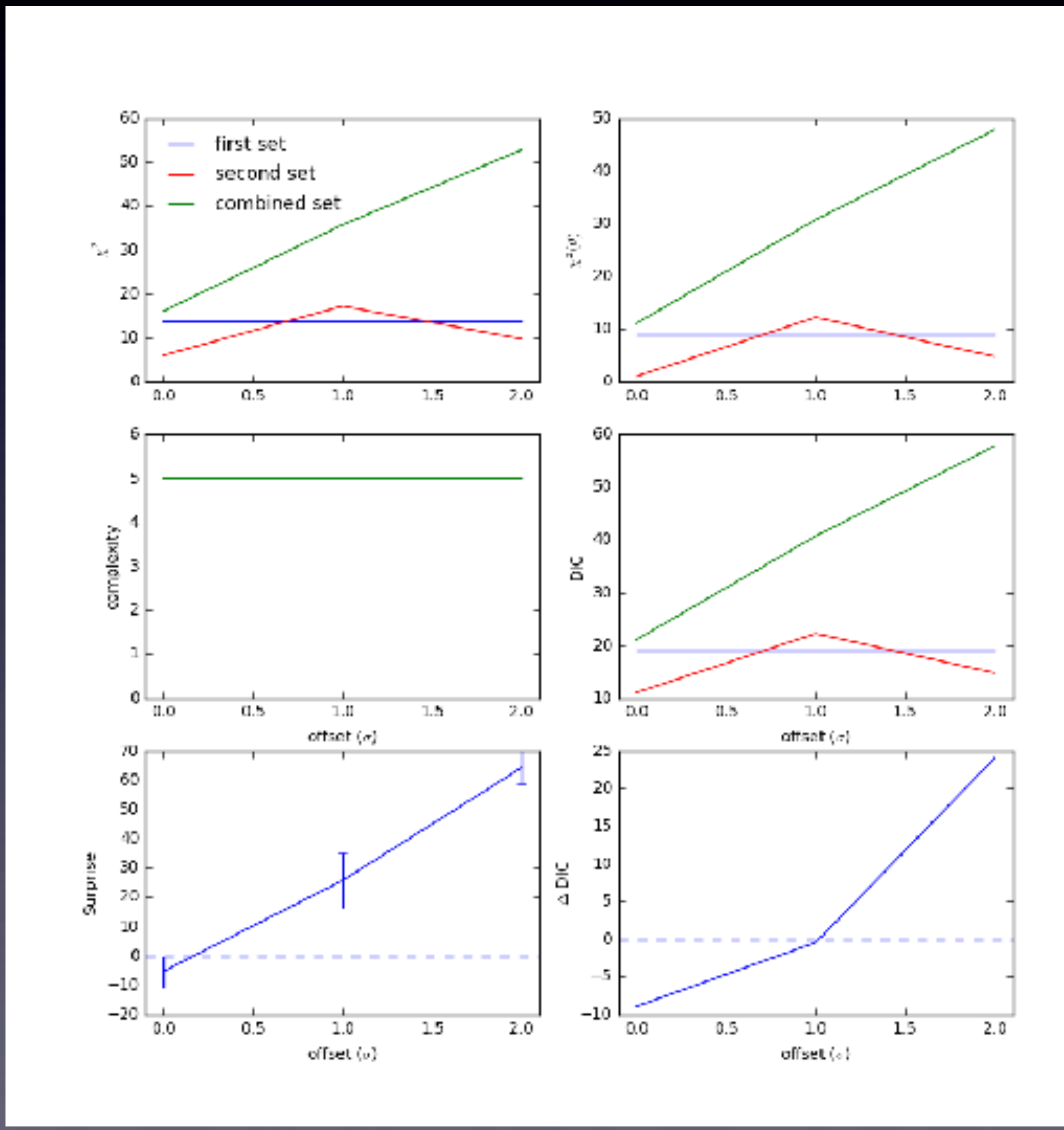
Pros and Cons

Approach	Like ratio	Evidence	DIC	Surprise
Average over parameters	No	Yes	Yes	Yes
From MCMC chain	Yes	No	Yes	Yes
Probabalistic	Yes	Yes	Yes	No
Symmetric	Yes	Yes	Yes	No

DIC

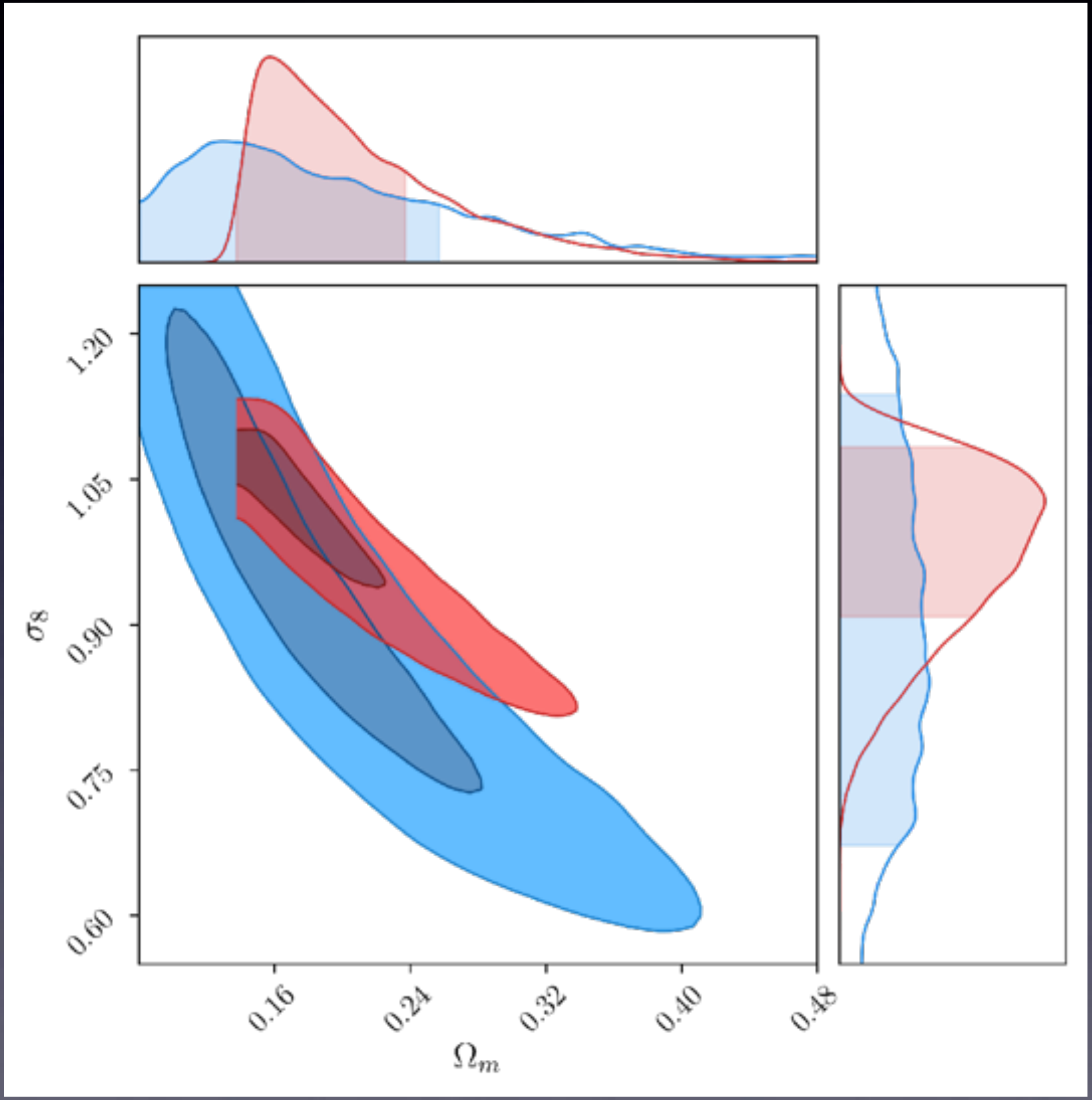
- Simple 5th order polynomial model, with second data set offset from the first
- Complexity of each individual data, and also combined data, is the same
  - Both measure the 5 free parameters well
- DIC only changes due to worsening of χ^2
- The ΔDIC goes from negative (agreement) to positive (tension) as the offset increases
- Odds ratio of agreement

$$\mathcal{I}(D_1, D_2) \equiv \exp\{-\Delta\mathrm{DIC}(D_1, D_2)/2\}$$



KiDS vs Planck

- All tensions considered here are in light of a particular model
- If the model is changed, the tension may be alleviated
- This is not the same as model selection

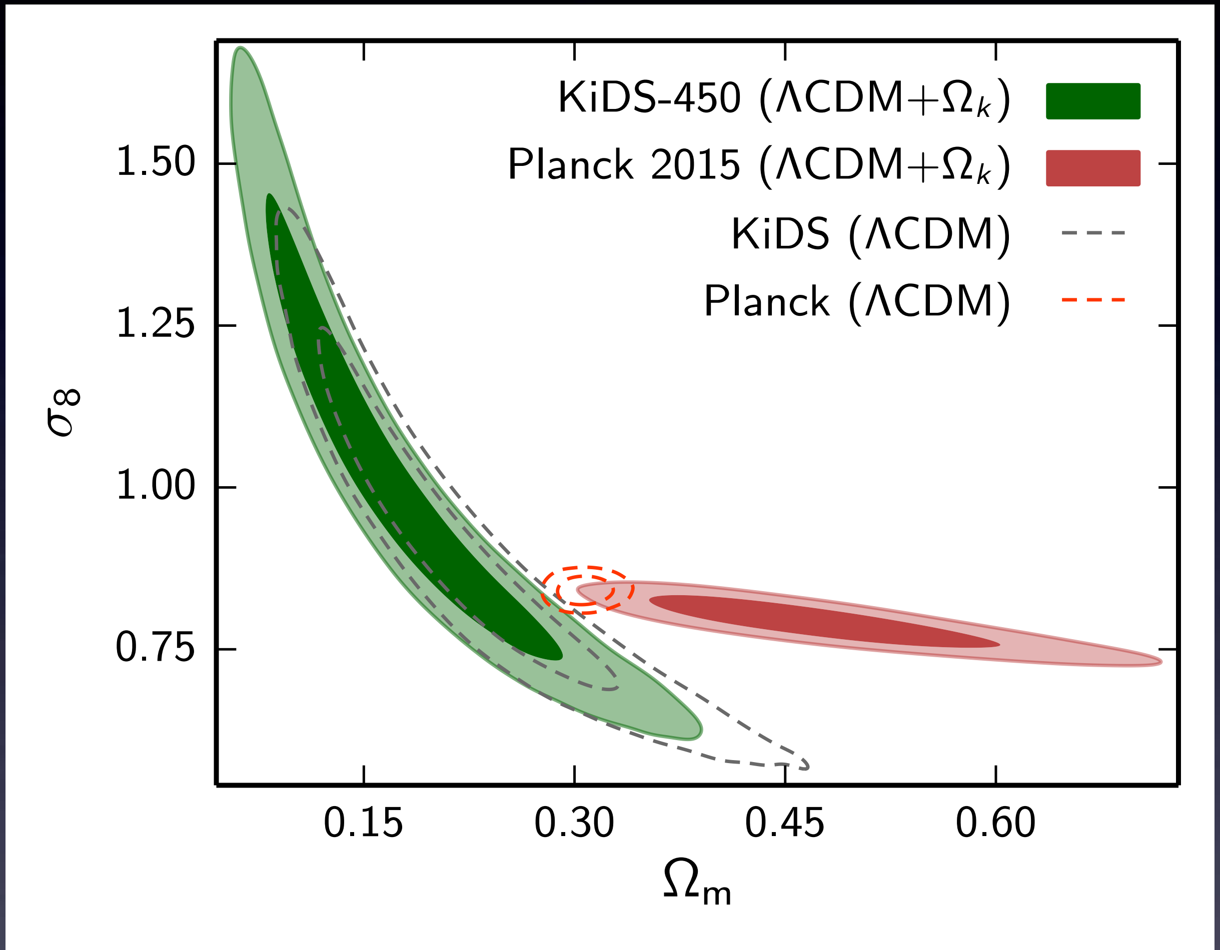


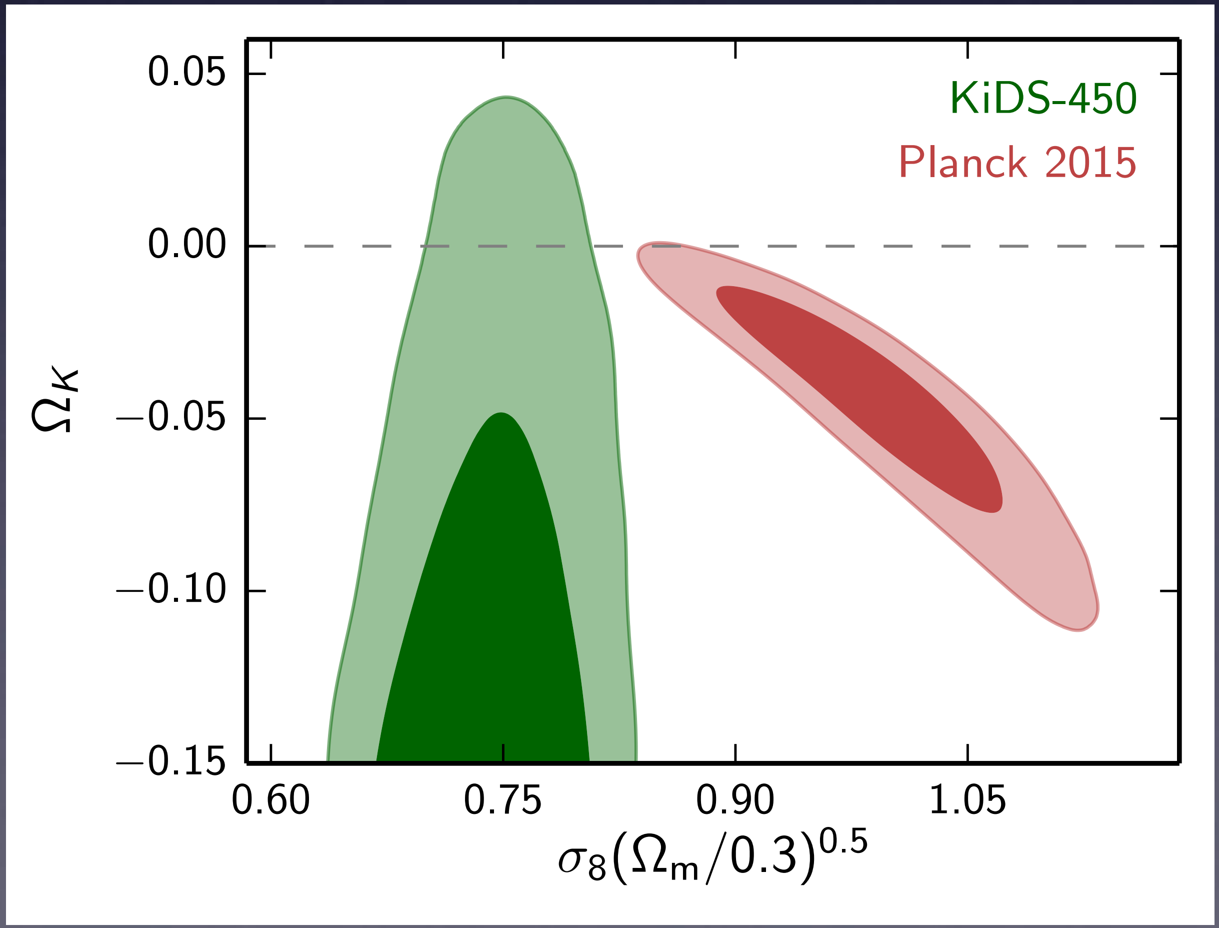
Application to lensing data

- In Joudaki et al (2016) they compared the cosmological constraints from Planck CMB data with KiDS-450 weak lensing data
- Including curvature worsened tension, but allowing for dynamical dark energy improved agreement

Model	$T(S_8)$	ΔDIC	
ΛCDM			
— fiducial systematics	2.1σ	1.26	Small tension
— extended systematics	1.8σ	1.4	Small tension
— large scales	1.9σ	1.24	Small tension
Neutrino mass	2.4σ	0.022	Marginal case
Curvature	3.5σ	3.4	Large tension
Dark Energy (constant w)	0.89σ	-1.98	Agreement
Curvature + dark energy	2.1σ	-1.18	Agreement

Curvature


KiDS-450 (ΛCDM+Ω_k)
Planck 2015 (ΛCDM+Ω_k)
KiDS (ΛCDM)
Planck (ΛCDM)
1.50
1.25
1.00
0.75
σ_8
0.15
0.30
0.45
0.60
Ω_m


KiDS-450
Planck 2015
0.05
0.00
−0.05
−0.10
−0.15
Ω_K
0.60
0.75
0.90
1.05
$\sigma_8(\Omega_m/0.3)^{0.5}$

Summary

- We can estimate the relative probability of tensions between data sets using ratios of model likelihood (evidence)
- The Deviance Information Criteria is a simple method, symmetric to evaluate tensions, being sensitive to likelihood ratio, but calibrated against parameter confidence regions
- Comparing tension between CMB and weak lensing tomography, we find these data sets give better agreement when dynamical dark energy is included in the model