Testing for Tensions Between Datasets

David Parkinson
University of Queensland

In collaboration with
Shahab Joudaki (Swinburne)

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

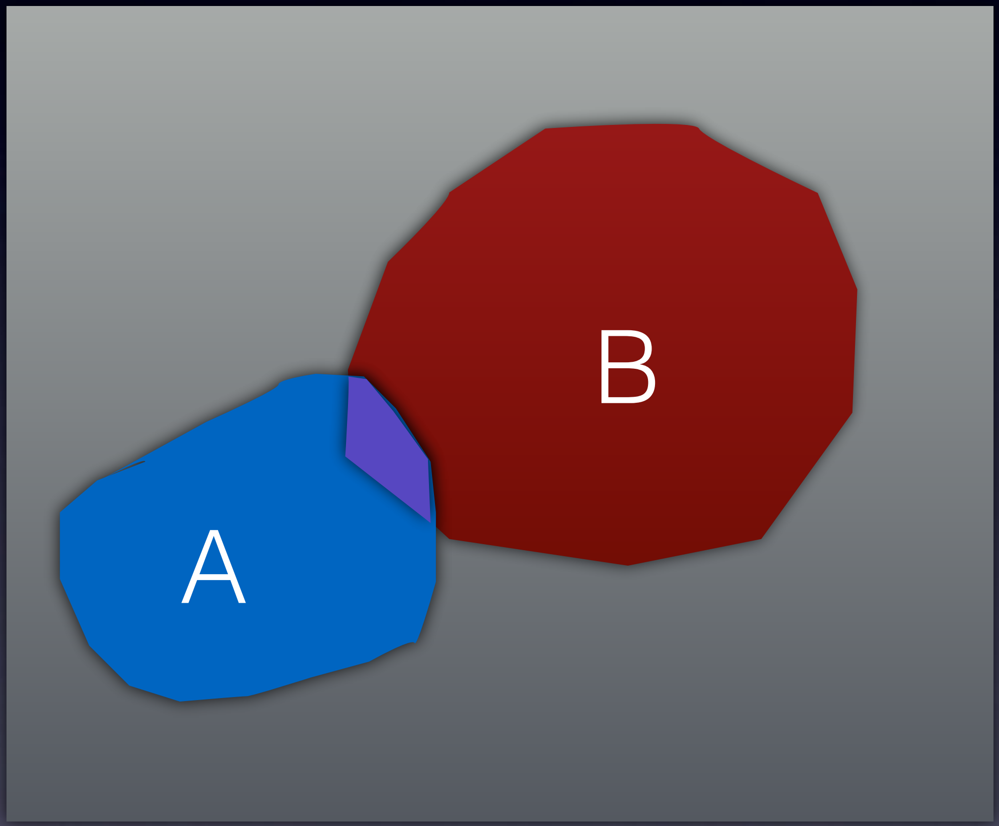
Pierre-Simon Laplace

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

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

likelihood

posterior

prior

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

evidence

- 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

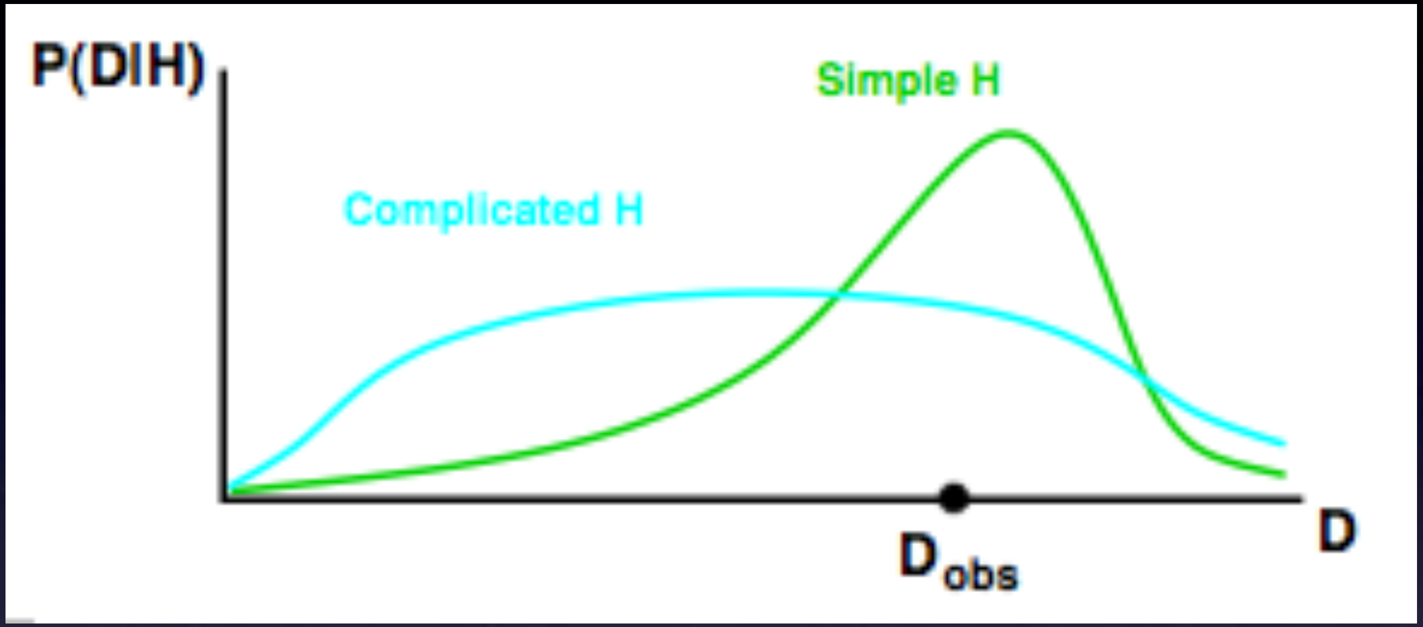
Occam's Razor

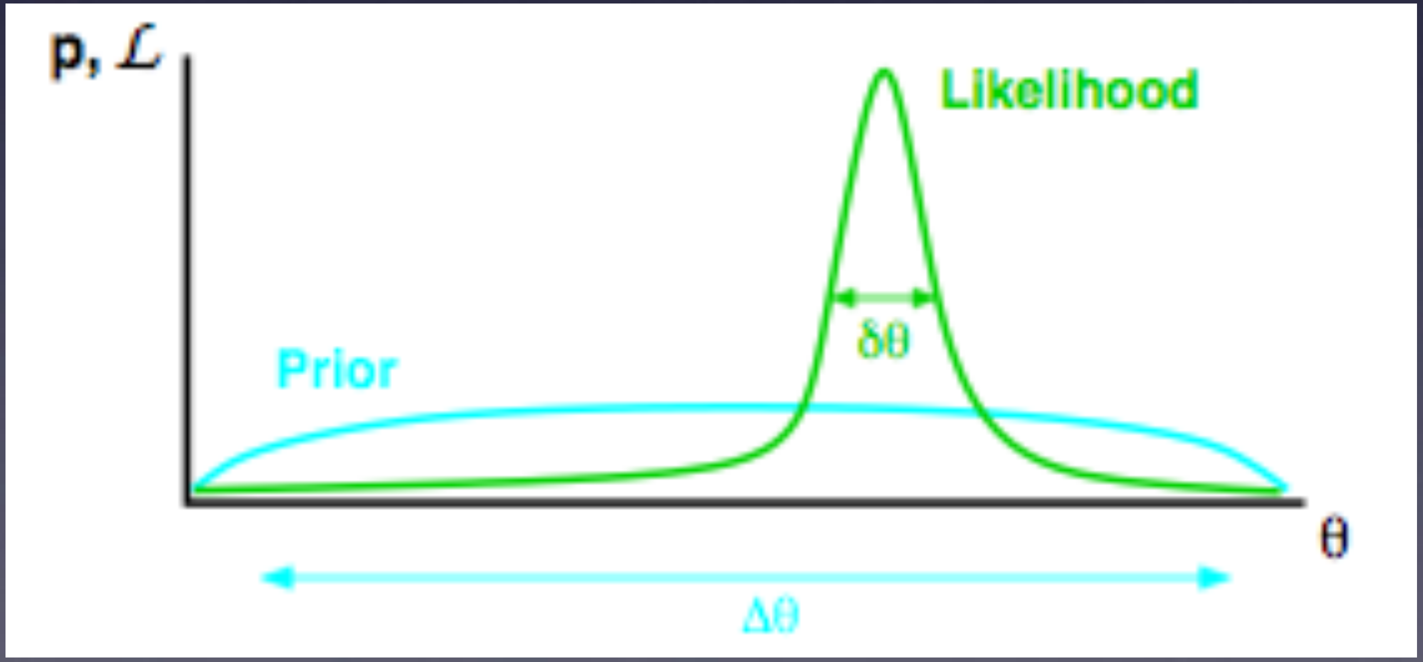
$$E = \int d\theta P(D|\theta, \mathcal{M}) P(\theta|\mathcal{M})$$

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

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 (Schwarz, 1978) - point estimate approximation to the evidence

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

  - k is the number of free parameters and N is the number of data points

Deviance

- Deviance Information Criterion (Spielgelhalter et al. 2002) comes from cross-entropy between prior and posterior

$$D_{\rm KL}\left(P(\theta|D,\mathcal{M})||P(\theta|\mathcal{M})\right) = \int P(\theta|D,\mathcal{M}) \log\left[\frac{P(\theta|D,\mathcal{M})}{P(\theta|\mathcal{M})}\right]$$

- If cross-entropy is small, distance minimised (i.e. model is predictive)
- DIC can be evaluated from MCMC chain

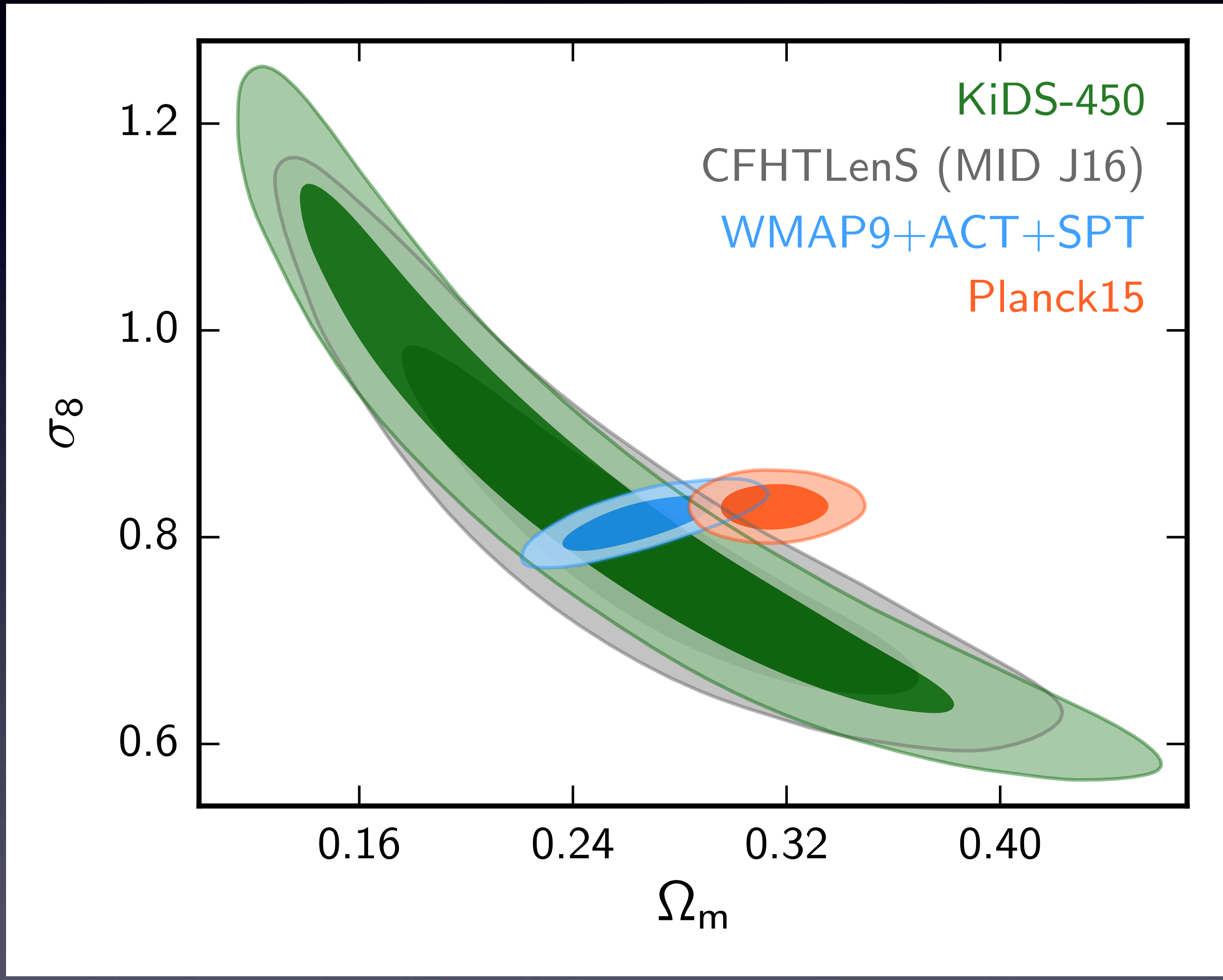
$$\mathrm{DIC} = \chi^2(\hat{\theta}) + 2c$$

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

$$c = -2\left(D_{\rm KL}(P(\theta|D,\mathcal{M})P(\theta|\mathcal{M})) - \widehat{D_{\rm KL}}\right) = \overline{\chi^2(\theta)} - \chi^2(\bar{\theta})$$

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
- 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

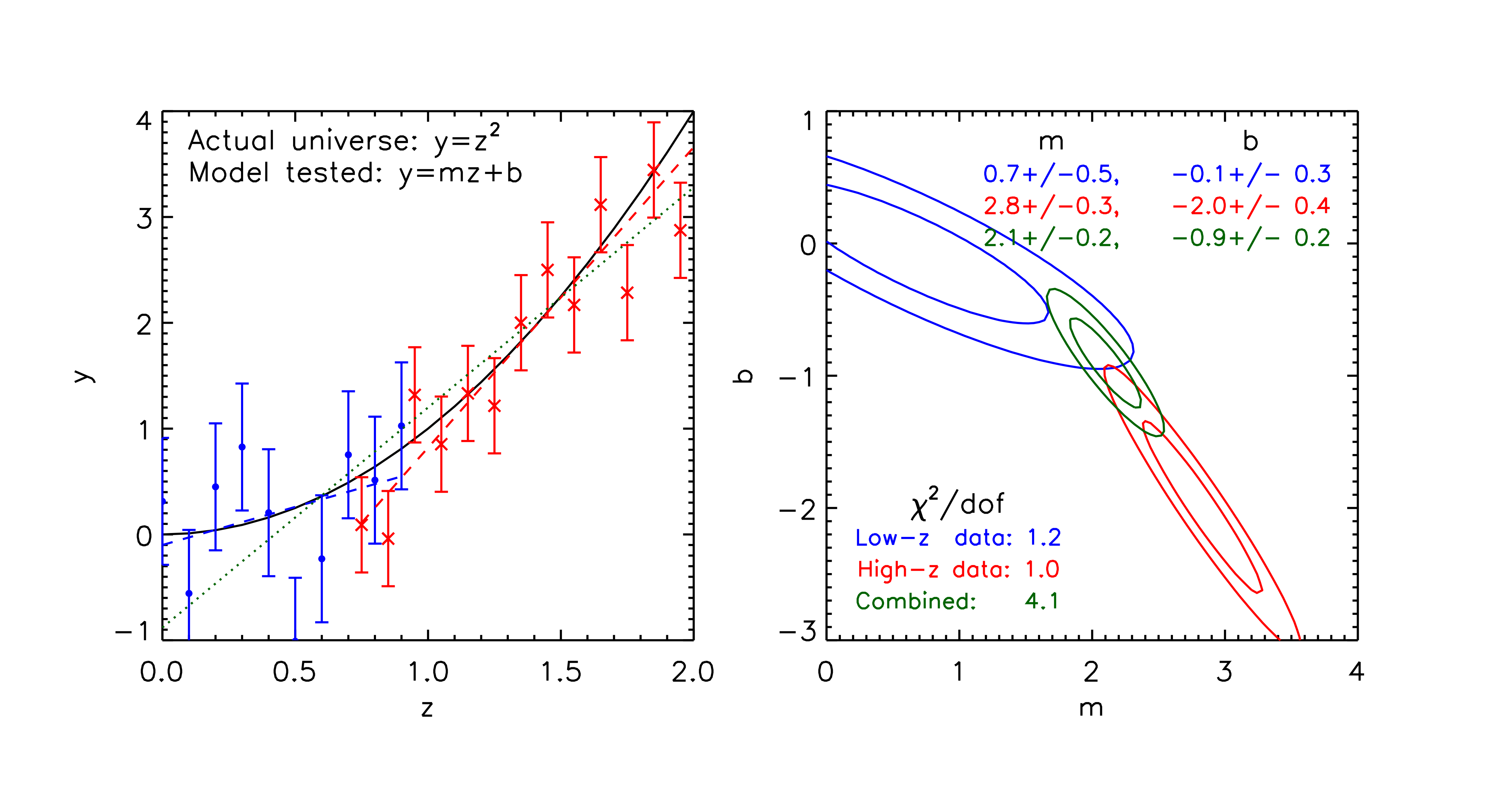
$$\Delta\mathrm{DIC} = \mathrm{DIC}(D_1 \cup D_2) - \mathrm{DIC}(D_1) - \mathrm{DIC}(D_2)$$

Linear evidence

$$P(D|\mathcal{M}) = \underset{1}{\mathcal{L}_0} \overset{2}{\frac{|F|^{-1/2}}{|\Pi|^{-1/2}}} \exp\left[-\frac{1}{2}(\theta_L^T L \theta_L + \overset{3}{\theta_\pi^T \Pi \theta_\pi} - \bar{\theta}^T F \bar{\theta})\right]$$

- 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



Diagnostics II: The Surprise

- Seehars et al (2016): the 'Surprise' statistic, based on cross entropy of two distributions
- Cross entropy given by KL divergence

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

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

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

Linear tension

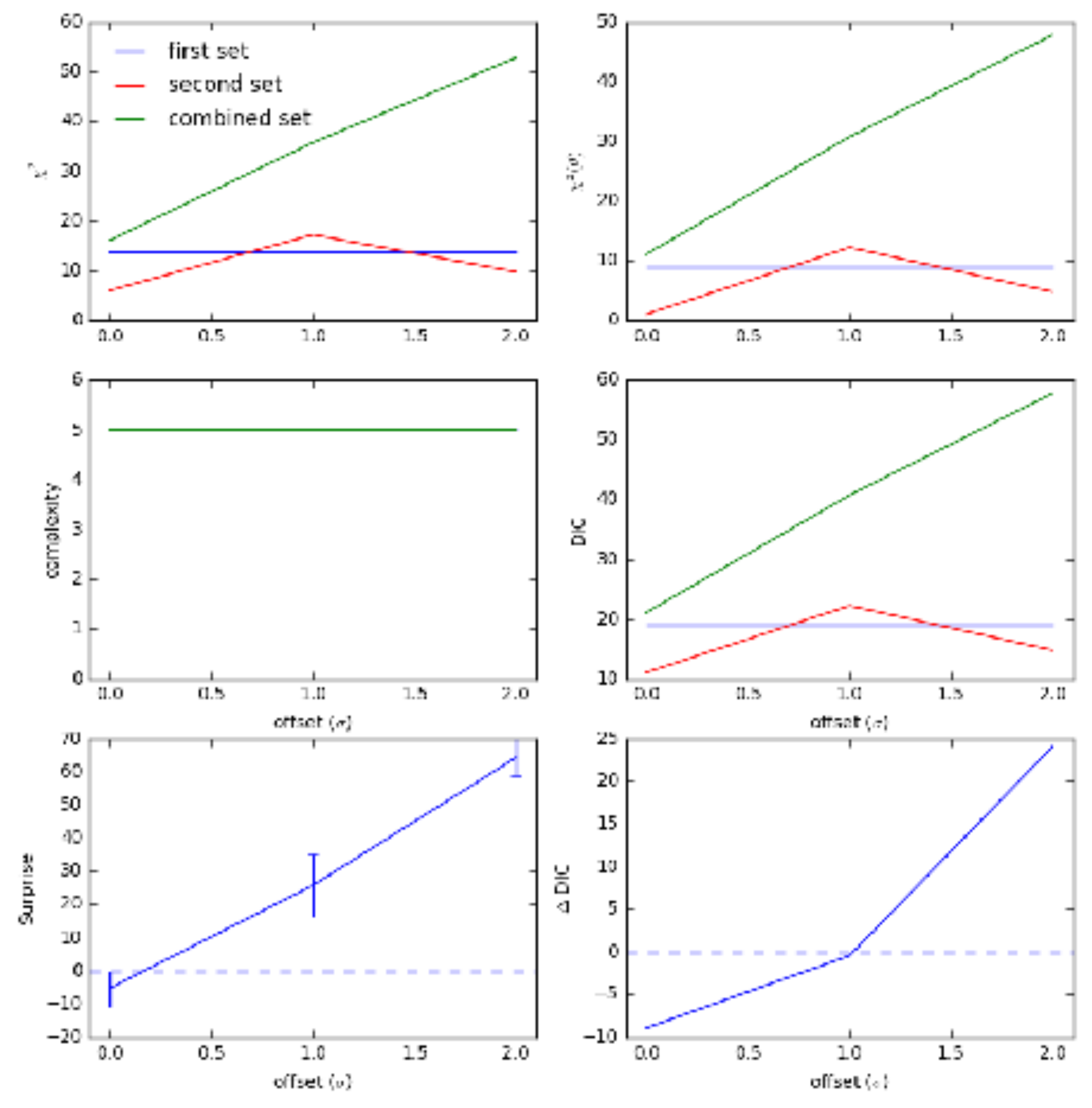
$$\frac{P(D_{1+2}|\mathcal{M})}{P(D_1|\mathcal{M})P(D_2|\mathcal{M})} = \frac{\mathcal{L}_0^{1+2}}{\mathcal{L}_0^1\mathcal{L}_0^2} \times \frac{|F_{1+2}|^{-1/2}}{|F_1|^{-1/2}|F_2|^{-1/2}} \times \text{displacement terms}$$

- Displacement terms equivalent to `Surprise' - relative entropy between two distributions
- Occam factor independent of tensions
- Tensions most manifest in first term - likelihood ratio

DIC vs Surprise

- 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

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



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

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 to evaluate tensions, being sensitive to likelihood ratio, but calibrated against parameter confidence regions
- Surprise is alternative approach to evaluating tensions, also using cross-entropy, though much more sensitive (perhaps overly)
- 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