Edinburgh Research Explorer

# Cosmic Bayes

PAPER IN PHILOSOPHY OF SCIENCE IN PRACTICE

# Cosmic Bayes. Datasets and priors in the hunt for dark energy

Michela Massimi[1] (ORCID)

## Abstract

Bayesian methods are ubiquitous in contemporary observational cosmology. They enter into three main tasks: (I) cross-checking datasets for consistency; (II) fixing constraints on cosmological parameters; and (III) model selection. This article explores some epistemic limits of using Bayesian methods. The first limit concerns *the degree of informativeness* of the Bayesian priors and an ensuing methodological tension between task (I) and task (II). The second limit concerns the choice of *wide flat* priors and related tension between (II) parameter estimation and (III) model selection. The Dark Energy Survey (DES) and its recent Year 1 results illustrate both these limits concerning the use of Bayesianism.

## 1 Introduction. Bayesianism in observational cosmology

Cosmology has witnessed a surge of interest among philosophers keen to explore experimental, statistical, and methodological practices in the current searches for dark matter and dark energy (Smeenk 2013; Anderl 2016, 2018; Beisbart 2009; Ruphy 2016; de Baerdemaeker forthcoming). Bayesianism has become the default approach in observational cosmology (see Marshall et al. 2006; Trotta 2008; Verde 2014) to deliver on three main tasks:

✉  Michela Massimi
    michela.massimi@ed.ac.uk

1    School of Philosophy, Psychology and Language Sciences, The University of Edinburgh, Dugald Stewart Building, 3 Charles Street, Edinburgh EH8 9AD, UK

I.   cross-checking the consistency of independent datasets coming from different cosmological probes (from within the same survey, and/or from different cosmological surveys);
II.  fixing constraints on important cosmological parameters;
III. selecting among different possible cosmological models.

Despite its proven usefulness and ubiquity, some critics have highlighted the epistemic limits of using Bayesian inferences in cosmology (e.g. Benétreau-Dupin 2015). My goal in what follows is to draw attention to two main epistemic limits affecting the use of Bayesianism in delivering tasks (I)–(III) in observational cosmology. These limits become evident when looking at the role that Bayesianism plays in the very high precision measurements currently being carried out in large cosmological surveys. The increasing and preponderant role of Bayesianism in grounding metrological practices in observational cosmology remains an unexplored topic in philosophy of science.

Understanding how Bayesianism enters observational cosmology from the ground up allows us to appreciate both the pervasiveness of Bayesian methods in cosmology and, most importantly, some of its methodological limits too. In the following three Sections, I introduce and explain what I take to be two main epistemic limits. The first concerns what I am going to call *the degree of informativeness* of the Bayesian priors that enter both into the consistency cross-checks between cosmological datasets and into parameter estimation. The choice of more or less informative priors causes a methodological tension between task (I) and task (II). The second and related limit concerns the choice of *wide flat* priors and the ensuing tension between (II) parameter estimation and (III) model selection in cosmology.

In Section 2, I briefly review some Bayesian techniques at play in delivering on tasks (I)–(III). The Dark Energy Survey and its recently released Year 1 results are presented in Section 3 as an illustration of these Bayesian techniques at work. And in Section 4, I return to the aforementioned epistemic limits, substantiate them, and offer a cautionary tale about using Bayesianism to interpret evidence for cosmological models. To be clear, it is not the goal and aim of this paper to contribute to the logic or formal methods of Bayesianism in cosmology (for recent work in the area see e.g. Charnock et al. 2017). My more modest *philosophical* goal is to contribute to the *epistemology* of Bayesian methods in cosmology by assessing the evidential conclusions warranted by their use.

The two main take-home messages of this exercise are the following. First, there is no *absolute* measure of evidence in cosmology, and how much evidence is evidence enough for a given cosmological model is always *relative* to which other model we are considering and how we interpret the Bayes factor for the two models along the Jeffreys scale. Thus, the use of Bayesian methods in cosmology should come with a warning to avoid the so-called "fallacy of acceptance" (to echo Spanos 2013): to accept a cosmological model $M_1$ (because no inconsistent evidence has been found *against* it) should not be conflated with there being substantial/decisive evidence *for* the model $M_1$. Moreover, to accept a cosmological model $M_1$ (because no inconsistent evidence has been found *against* it) does not license the further claim that therefore model $M_1$ is true, or that the parameter estimates made in it are the 'true' one.

A more promising way of looking at the use of Bayesian methods (especially the Bayes factor) in cosmology is as "guiding an evolutionary model-building process" (Kass and Raftery 1995, p. 773) whereby there is a clear

continuity and almost evolution in building and assessing model selection. But to start with, let us consider the nature of data and evidence in contemporary observational cosmology.

## 2 Cosmic Bayes. Datasets consistency, parameter estimation, and model selection

This section offers a brief Bayesian primer about the three aforementioned tasks in observational cosmology, starting with datasets consistency. Datasets in cosmology take very different forms and typically come from a number of diverse cosmological probes within the same cosmological survey. Within the Dark Energy Survey (DES), for example, cosmologists compare for consistency datasets about galaxy clusters[1] with datasets about gravitational lensing.[2] But cosmologists are also interested in comparing and integrating say gravitational lensing datasets from DES with the Baryon Acoustic Oscillation (BAO)[3] datasets from the 6dF Galaxy Survey and the Baryon Oscillation Spectroscopic Survey (BOSS). Or with datasets about cosmic microwave background (CMB)[4] from Planck; and/or with Supernovae Ia datasets[5] from the Joint Lightcurve Analysis (JLA), just to mention a few examples.

Datasets from different cosmological probes are very diverse in nature, and are designed to measure very different features of the universe. Some (e.g. gravitational lensing and galaxy clusters) are designed to measure the 'clumpiness' of matter in the universe (i.e. how matter clumped to form large-scale structure of galaxies and clusters of galaxies over time). Others give a measure of the relative rate of expansion of the universe (using BAO as 'standard rulers' and Supernovae Ia as 'standard candles'). How is it possible to cross-check for consistency datasets of such bewildering variety as supernova explosions, remnants of sound waves in the early universe, and galaxies' shears via lensing? How to extract from this plurality of diverse signals *evidence for* the universe's rate of expansion and growth of structure?

---

[1] Under the action of gravity and what is believed to be dark matter, galaxies form 'clusters' over time, and by observing the distribution of galaxy clusters at different historical epochs after the Big Bang, important information can be gained about the structure formation of the universe over time.

[2] When light from a far-away galaxy passes in the proximity of a high concentration of galaxies, light bends and the shape of the galaxy displays a distinctive distortion ('shear') when observed from a telescope. By measuring the shears of very many galaxies, it is possible to infer how clumpy the universe is at different epochs.

[3] BAO refers to the remnants of original sound waves travelling at almost the speed of light shortly after the Big Bang and before the universe started cooling down and atoms formed. This phenomenon resulted in the formation of what appears in the sky today as an over-dense region of galaxies forming a ring with a radius around a given galaxy. By knowing the radius of the ring (which is a 'standard ruler'), cosmologists can measure the angle subtended from the Earth vantage point and probe the rate of expansion of the universe.

[4] The CMB from Planck (see Ade et al. 2016) shows initial density fluctuations in the hot plasma at the time of last scattering. The over-dense blue regions in these maps indicate the seeds that led to the growth of structure, and the gradual formation of galaxies and rich galaxy clusters over time.

[5] Supernovae Ia are gigantic explosions of stars that have come to the end of their lifetime (called white dwarfs), and whose brightness tends to stay the same, and depends only on their distance from us. Hence, they are routinely used in cosmology as 'standard candles' to measure on the basis of their brightness and redshift, the rate of expansion of the universe.

Within each cosmological survey there are sub-groups whose expertise is entirely dedicated to harvesting data from one single probe (e.g. gravitational lensing) and to run statistical analyses, which then have to be compared and integrated with the measurement outcomes of other sub-groups working on other probes and datasets (e.g. galaxy clusters). Ultimately, the task is to assess the ongoing validity of the standard cosmological model, i.e. Lambda Cold Dark Matter (ΛCDM), which postulates dark matter and dark energy to explain structure formation and the rate of expansion.

If datasets cross-checks were to reveal a discrepancy in some expected values, the consequences would be far-reaching (see Charnock et al. 2017). It could be evidence that there might be something wrong with our currently accepted cosmological model and that the very notion of dark energy (as a non-zero value of the vacuum energy density) would have to be reconsidered. Given that such high stakes in the foundations of cosmology rest on harvesting data and statistically analysing them, it comes as no surprise that recent decades have seen a surge of investments in the establishment of many large cosmological surveys (e.g., DES, Gama, KiDS, DESI, Euclid, just to mention a few of them) whose goal is to measure with increasing accuracy and high precision the value of relevant cosmological parameters and feed them into model selection. And this is where Bayesianism comes in.

Cross-checking for consistency large datasets from different cosmological probes typically requires the use of so-called *Bayesian evidence*. Bayesian evidence assesses how likely it is to observe the datasets $D$ that are actually observed, given a certain model $M_1$ whose constrained cosmological parameters $\theta_i^{M_1}$ all range over certain intervals of possible values:

$$p(D|M_1) = \int p(D|\theta_i^{M_1}, M_1) p(\theta_i^{M_1}|M_1) d^n \theta_i^{M_1} \qquad (1)$$

The *Bayesian evidence* (Equ. 1) takes the form of an (analytically very complex to solve) marginal or integrated likelihood that gives the probability of finding the datasets $D$ by integrating over the parameter space $\theta_i^{M_1}$ of model $M_1$ with $p(\theta_i^{M_1}|M_1)$ being the priors for those parameters. If we want to assess whether two independent datasets $D_1$ and $D_2$ are consistent with one another (conditional on a single underling model, $M_1$ which in this case typically is the ΛCDM model), one possible option is to use what is sometimes called *R statistic* (see Marshall et al. 2006; for its use in DES Y1 see Abbott et al. 2018, and Handley and Lemos 2019 for a discussion). *R statistic* is defined as follows:

$$R = \frac{p(D_1, D_2|M_1)}{p(D_1|M_1)p(D_2|M_1)} \qquad (1*)$$

and it measures the ratio between fitting model $M_1$ to both datasets simultaneously vis-à-vis fitting the model to each dataset individually (with the probabilities $p$ defined as in Equ. 1). It is worth pointing out already at this stage how strongly prior-dependent $R$ is: it depends on the priors of the constrained parameters—i.e. $p(\theta_i^{M_1}|M_1)$— which are shared between the marginal likelihoods as defined in Eq. (1) for each individual dataset (but, of course, $R$ is not dependent on the priors of possible additional unconstrained parameters).

Consider now two models $M_1$ and $M_2$ with slightly different intervals of values for the $n$ constrained parameters $\theta$. To assess how likely a dataset $D$ that is actually observed is given either $M_1$ or $M_2$, cosmologists resort to the ratio of the Bayesian evidences for the two models—this is called the *Bayes factor* and is usually (and confusingly enough) also indicated with '*R*' (but not to be confused with *R statistic* defined by Eq. 1* which assumes one single model $M_1$). The Bayes factor $R$ is given by

$$R = \frac{\int p\big(D|\theta_i^{M_1}, \ M_1\big)p\big(\theta_i^{M_1}|M_1\big)d^n\theta_i^{M_1}}{\int p\big(D|\theta_i^{M_2}, M_2\big)p\big(\theta_i^{M_2}|M_2\big)d^n\theta_i^{M_2}} \tag{2}$$

where again $D$ is a given dataset, $\theta_i^{M_1}$ are $n$ theoretical parameters that are shared between model $M_1$ and $M_2$, $p\big(\theta_i^{M_1}|M_1\big)$ are the prior probabilities of the parameters in model $M_1$ (similarly for $M_2$), and $p\big(D|\theta_i^{M_1}, \ M_1\big)$ is the likelihood (i.e. how likely the dataset $D$ is, given the range of possible values for $\theta_i^{M_1}$; the same applies to $M_2$).

Cosmological models contain parameters $\theta_i^{M_1}$ whose possible values $i$ range over an interval to be determined, hence the marginal likelihoods for the models are obtained by integrating over the parameter space of each model, rather than trying to best-fit models to the data as in frequentist approaches. The advantage of adopting Bayesian rather than frequentist approaches in this context is that the former do not unduly penalise models that—albeit interesting to explore—might nonetheless have not very well constrained theoretical parameters (see Amendola and Tsujikawa 2010, pp. 363–4). Such models would be discarded by frequentist best-fit analyses, which would tend to maximise fit between the model and the available data.

But Bayesianism is ubiquitous and enters also into parameter estimation and model selection. When cosmologists want to fix more rigorous constraints on the main cosmological parameters (assuming, say, only one model $M_1$), they resort to the Bayes theorem. To calculate the posterior probabilities (Eq. 3) for, say, parameter $\theta_i^{M_1}$ (which, let us assume for simplicity, can range over $i = 1, 2$), using the Bayes theorem cosmologists proceed as follows:

$$\frac{p\big(\theta_1^{M_1}|D,M_1\big)}{p\big(\theta_2^{M_1}|D,M_1\big)} = \frac{p\big(D|\theta_1^{M_1},M_1\big)p\big(\theta_1^{M_1}|M_1\big)}{p(D|M_1)} \times \frac{p(D|M_1)}{p\big(D|\theta_2^{M_1},M_1\big)p\big(\theta_2^{M_1}|M_1\big)}$$

$$= \frac{p\big(D|\theta_1^{M_1},M_1\big)p\big(\theta_1^{M_1}|M_1\big)}{p\big(D|\theta_2^{M_1},M_1\big)p\big(\theta_2^{M_1}|M_1\big)} \tag{3}$$

where the Bayesian evidence $p(D|M_1)$ cancels out and the posterior probability is given by the likelihood of the dataset $D$ and the priors for $\theta_1^{M_1}$ (same for $\theta_2^{M_1}$), given model $M_1$. In practice, these Bayesian inferences for cosmological parameters are carried out with numerical simulations in the form of Monte Carlo Markov Chain techniques (see Trotta 2008).

And when it comes to model comparison and model selection, Bayesianism allows for the calculation of the respective posterior probabilities of two rival models $M_1$ and $M_2$ given the same observed dataset $D$ as follows:

$$\frac{p(M_1|D)}{p(M_2|D)} = \frac{p(D|M_1)p(M_1)}{p(D|M_2)p(M_2)} \tag{4}$$

In the Bayes's theorem behind Eq. (4), the denominators $p(D)$ cancel each other out; equal priors are usually assumed for $p(M_1)$ and $p(M_2)$; and the ratio of the likelihoods $\frac{p(D|M_1)}{p(D|M_2)}$ is again given by the all-powerful Bayes factor of Eq. (2).

The Bayes factor $R$ tells us that if $R$ is less than 1, then the evidence *in favour* of $M_1$ over $M_2$ is weak. But if $R$ is more than 1, then there *is* evidence *in favour* of $M_1$ over $M_2$. How much evidence *in favour* of $M_1$ is evidence enough? Some interesting philosophical questions come into play here concerning the use of the Bayes factor in assessing cosmological evidence (see Skilling 2011, p. 33). The Bayes factor in cosmology offers a standard for assessing evidence always *relative to* two rival models, rather than a standard of evidence in *absolute terms.* It measures how likely the evidence for a given model $M_1$ (let us call it the null hypothesis) is vis-à-vis a rival model $M_2$. But it is not enough to establish what $R$ is going to look like. A scale for reading and interpreting such values is also required (and, crucially, the same holds for the $R$ statistic at play in consistency cross-checks for datasets, Eq. 1*). The scale in question is typically the Jeffreys scale (Jeffreys 1939/1961).

In the original Appendix to Jeffreys' textbook, the Jeffreys scale considers a Bayes factor of <1 as not significant; between 1 and 2.5 as moderate evidence; between 2.5 and 5 as strong evidence; and above 5 as decisive evidence for $M_1$ over $M_2$. But the Jeffreys scale can be adjusted and adapted to fit evidential needs in different contexts of inquiry. Cosmologists typically adopt a slightly expanded version of the Jeffreys scale because of the large degree of uncertainty affecting the choice of priors in cosmology (see Liddle et al. 2009, p. 90). Typically in cosmology a Bayes factor above 5 (rather than above 2.5) is regarded as strong (but not decisive) evidence for $M_1$ over $M_2$; and a Bayes factor above 10 is taken as very strong evidence (as we shall see in the following section concerning the DES case study). However, even with a Bayes factor $R > 5$ as "strong evidence", cosmologists warn that the "terminology is purely suggestive and not to be taken literally. We can consider it as a practical bookkeeping device." (Amendola and Tsujikawa 2010, p. 366).

Before worrying about how to read and interpret the values of the Bayes factor along the Jeffreys scale, those values need to be calculated. Calculating the marginal likelihoods for rival models in the Bayes factor (Eq. 2) is a non-trivial matter and typically requires prior distributions for the relevant theoretical parameters— $p\left(\theta_i^{M_1}|M_1\right)$ and $p\left(\theta_i^{M_2}|M_2\right)$. By contrast with subjective Bayesianism, the priors in this context are not cosmologists' subjective degrees of belief. They are typically fixed either on the basis of theoretical considerations or by using existing data coming from previous cosmological surveys. This practice—in and of itself—is of course unproblematic (for it reflects in a way prior knowledge based on available evidence). Yet the particular choice of priors raises interesting philosophical questions. One is *how informative* we want our priors to be.

How much information each prior packs depends both on (i) the nature of the priors and (ii) the source of the priors. On (i), some priors are Gaussian priors with a mean and a variance; others are top-hat flat priors assigning the same probability within an allowed range of values. And re (ii), some priors originate from pre-existing measurements or galaxy catalogues while others are motivated mostly by theoretical considerations. Both *data-dependent* and *theory-dependent* priors enter into datasets cross-checks, parameter estimation, and model selection.

Examples of what I call *theory-dependent* priors are, for example, the priors for the baryon energy density $\Omega_b$ which are reasonably expected to be top-hat flat (i.e. to have equal probability) within the range 0.03–0.05, as we shall see in Section 3. This range of admissible flat priors is justified by theoretical considerations about Big Bang nucleosynthesis, which allow cosmologists to establish what the baryon-to-photon ratio might have been at the time of last scattering after the Big Bang. Similarly, it is reasonable to expect that the matter energy density $\Omega_m$ ranges over an interval of top-hat flat priors between 0.1 and 0.9,[6] given present-day estimates from $\Lambda$CDM. Clearly, whether these priors are exportable to other rival models is precisely one of the problems behind *theory-dependent* priors that are going to affect tasks (I)-(III). Datasets are cross-checked for consistency (via Eqs. 1 and 1*) *granted the assumption* (embedded by those aforementioned priors) that we live indeed in a universe with a geometrically flat metric and a matter density less than 1, which suggests implicitly the existence of both dark energy and dark matter (the latter is assumed to compensate for the discrepancy between the estimated value for the overall matter energy density $\Omega_m$ and the baryon energy density $\Omega_b$).

Other priors, especially those for nuisance parameters (e.g. photo-$z$, shear calibration, among others), are obtained from previous systematic-error analyses from galaxy catalogues in already existing databases.[7] I am going to call them *data-dependent* priors. Choices are made every step of the way about which galaxy catalogue to use as a sample to inform those priors, and which sample might be the most 'representative' for the specific datasets cross-check consistency. Data-dependent and theory-dependent priors encode more or less information for the task at hand either by providing a mean and a variance for the spread of the nuisance parameters (as with Gaussian priors having a broad or narrow peak); or by remaining agnostic about where exactly in a given range of physically allowed values the 'most likely' value of the cosmological parameter might lie (as with flat priors that can have a large or short top-hat width).

But how informative should the priors be for delivering on the relevant taks? We want them to be as informative as possible when it comes to datasets cross-checks (using Bayesian evidence, Equ. 1) for the purpose of eliminating systematic uncertainties and what is called galaxy bias, for example. But we also want them to be less informative when it comes to parameter estimation because the posterior probabilities of these parameters (in Equ. 3) should not be too sensitive to the choice of the priors.

---

[6] $\Omega_m$ indicates the matter energy density of the universe and so on theoretical grounds it can only range between 0 (no matter in the universe) and 1 (everything in the universe is matter).

[7] Priors for nuisance parameters tend to be Gaussian (rather than top-hat flat priors) because the idea is to have more informative priors to better control galaxy bias and systematic uncertainties. By contrast, in parameter estimation, flat priors are privileged over Gaussian ones because they are less informative about where the real value lies and the posterior probability has to be less sensitive to the choice of the priors. Although top-hat flat priors have a centre and width, they assume *equal* probabilities for all the values covered by the top-hat range, whereas Gaussian priors single out a mean where the probability is higher than everywhere else in the range.

And since the priors are the same for datasets consistency cross-checks and parameter estimation (and necessarily so since the universe we are studying is the same and the data and the relevant parameters are the same for the two tasks), there is bound to be a tension about how (more or less) informative the priors are set to be. Statistically, one cannot use different priors for different problems concerning the same data and the same parameters, especially since—as Section 3 explains—the priors at stake here are theory-dependent and data-dependent, but they are not subjective degrees of belief of cosmologist A vs. cosmologists B.

To be more precise, the tension in question is the product of the specific feature of the $R$ statistic used for consistency cross-checks in Eq. (1*), which, as already noted, is strongly prior-dependent. If we go for informative priors to reduce systematic uncertainty, and hence try to reduce the width of the possible range for the constrained parameters priors— i.e. $p\left(\theta_i^{M_1}|M_1\right)$—, the Bayesian evidence (Equ. 1) increases. However, this very same move has the effect of decreasing the value for the $R$ statistic in (1*), which has one Bayesian evidence in the numerator—i.e. $p(D_1, D_2| M_1)$— and two in the denominator, i.e. $p(D_1| M_1)p(D_2| M_1)$. A low $R$ ($<< 1$) along the Jeffreys scale indicates inconsistency among datasets given model $M_1$. Thus, informative, custom-taylored priors are good for the Bayesian evidence but bad for the $R$ statistic used to measure consistency across independent datasets. The narrower the range of the priors, the more precise the Bayesian evidence as to how a given model $M_1$ fits a given dataset, the lower the chances of the dataset being consistent with another independent dataset that might be fitted to the same model when (1*) is adopted for consistency cross-checks. So much worse for informative priors, one might say. Let us stick with uninformative wide-ranging priors instead.

Not so fast. For uninformative wide-ranging flat priors might bump up the $R$ statistic and suggest datasets consistency when in fact there might be none. Second, uninformative wide-ranging flat priors might result in a mostly empty posterior volume in most of the space allowed by the prior's width when it comes to parameter estimation. More in general, the informativeness of priors engenders a *methodological bootstrap* between task (I)—i.e. cross-checking the consistency of diverse datasets via the $R$ statistic (Eq. 1*) where prior distributions of cosmological parameters enter—*and* task (II)—i.e. refining and improving the estimates of these very same cosmological parameters (as per Eq. 3) using the already-cross-checked-datasets, as I am going to illustrate in Section 3 and 4.

A second interesting question concerns *how widely* the 'top-hat' flat priors should range. As the DES case shows in Section 3, and as is further discussed in Section 4, wide flat priors in the Bayes factor (Eq. 2) cause a tension between parameter estimation (II) and model selection (III). The tension arises from the specific choice of equal probability (flat) ranging over a sufficiently 'wide' spectrum of possible values for the dark energy equation of state parameter $w$ (whose maximal posterior probability needs be estimated using the Bayes theorem as per Eq. 3). Wide flat priors do in turn affect model selection because they tend to favour the so-called 'null hypothesis' (namely, the default hypothesis which in the case of cosmology is the standard $\Lambda$CDM model) when it comes to the comparative assessment of evidence between different models (Eq. 4). This phenomenon is known in statistics as Bartlett's paradox (see Section 4.2), and I illustrate it with reference to a salient example coming from the Dark Energy Survey (DES), to which I turn next.

# 3 Some lessons from the Dark Energy Survey year 1 results

The Dark Energy Survey (DES) is one of the largest cosmological surveys mapping the 14-billion-year cosmic expansion of the universe and the rate of growth of large-scale structure. DES is a photometric survey. In what follows I concentrate on the data already publicly available and released in the summer 2017 concerning Y1 results (Abbott et al. 2018).[8]

DES resorts to a total of four different probes. Two probes measure the rate of expansion of the universe at different epochs: Supernovae Ia as standard candles and BAO as standard rulers. The other two probes (weak gravitational lensing and galaxy clusters) measure the rate of growth of large-scale structure; or, if you like, the 'clumpiness' of matter in the universe. By using this four-probe approach DES hopes to find out more about the nature of dark energy at work in these two phenomena.

But DES also integrates datasets coming from different cosmological surveys: BAO from 6dF Galaxy Survey and BOSS; datasets about CMB from Planck; Supernovae Ia datasets from the JLA, just to mention a few examples. Year 1 results do not include all four probes but only a combination of two main probes: namely, galaxy clustering (not to be confused with galaxy clusters – clustering is the distribution of galaxy positions) and weak gravitational lensing.

Galaxies were put to a twofold use to obtain these results. Some were used as 'lens galaxies' for measuring the angular distribution of galaxies. Others were used as 'source galaxies' to estimate the so-called cosmic shear, i.e. how foreground large-scale structure distorts the shape of far-away galaxies when observed through weak lensing. A number of systematic uncertainties enter into these data measurements: for example, possible errors in the photometric redshifts and in shear calibration. In galaxy clustering, systematic uncertainty creeps in the form of what is called 'galaxy bias', namely how galaxy space distribution may or may not fit with the expected matter distribution on theoretical grounds.

Once collected, calibrated and cross-checked (task I), DES Year 1 data are put to a twofold use. The first is to compare the $\Lambda$CDM model with a rival proxy model (task III), called $w$CDM, which shares with $\Lambda$CDM six main theoretical parameters (and treats a seventh shared one $w$—the dark energy equation of state—as a free parameter). The second is to fix more rigorous constraints on the estimates of the seven main theoretical parameters and twenty additional nuisance parameters (task II). In $\Lambda$CDM, the seven main parameters are as follows: the matter energy density ($\Omega_m$); the assumed spatial flatness of the universe with ($\Omega_\Lambda = 1 - \Omega_m$); the baryon energy density ($\Omega_b$); the massive neutrinos' energy density $\Omega_\nu$; the reduced Hubble parameter ($h$) defined as the Hubble constant in units of 100 km s$^{-1}$Mpc$^{-1}$ (i.e. if H$_0$ = 70 km s$^{-1}$Mpc$^{-1}$, the reduced Hubble parameter is $h$ = 0.7); the dark energy equation of state $w$, which is fixed to −1; and the amplitude and the spectral index of the primordial scalar density perturbations, $A_s$ and $n_s$.[9]

$w$CDM is a phenomenological proxy for a variety of physical models that have some dark energy evolution. It treats the equation of state parameter $w$ not as fixed at −1 (as it would be in $\Lambda$CDM), but as a free parameter that can take a range of possible values. In

---

[8] In what follows I build and expand upon Massimi (2020). I am very grateful to Ofer Lahav and DES members for allowing me to participate in the June 2017 DES Collaboration meeting at the University of Chicago, and for helpful comments and discussions from which this research originated.

[9] Others cosmological parameters include: the tensor-to-scalar ratio for primordial perturbations $r$ that is assumed to be zero; and a two-parameter primordial power spectrum of adiabatic and Gaussian fluctuations.

addition to these key theoretical parameters there are, as noted, twenty nuisance parameters, which are common to both ΛCDM and wCDM, and include parameters for lens galaxy bias, photo-z shifts[10] for both lens galaxies and source galaxies, and shear calibration. Table 1 gives the priors for all these cosmological and nuisance parameters. Priors are key in the methodological procedure that follows.

DES clearly made a choice for "flat priors that span the range of values well beyond the uncertainties reported by recent experiments." (Abbott et al. 2018, p. 043526–12). Having a wide flat prior might not be very telling in and of itself, but the methodological principle for DES has been that priors "should not impact our final results, and in particular that the tails of the posterior parameter distributions should not lie close to the edges of the priors" (Abbott et al. 2018, p. 043526–13). The priors for nuisance parameters (e.g. photo-z, shear calibration) are obtained from previous systematic-error analyses from galaxy catalogues in already existing databases.[11] For example, the priors constraints on the lens and source photo-z shifts in Table 1 were obtained from selecting and sampling galaxies from already existing databases (e.g. COSMOS) which were taken as "representative of the DES sample with successful shape measurements based on their color, magnitude, and preseeing size." (Abbott et al. 2018, p. 043526–8). These are examples of what I previously called *data-dependent* priors. Other priors come from data analysis of the Sloan Digital Sky Survey, whose spectroscopic redshift feeds in cross-correlation of DES RedMaGiC software at work for lens photo-z. Choices are made about which galaxy catalogue to use as a sample to inform those priors, and which sample might be the most 'representative' for the specific datasets consistency cross-checks.

With these priors in place, DES fixes new constraints on the main seven parameters in ΛCDM and wCDM (task II). These are calculated as posterior probabilities (via Eq. 3) by using the priors listed in Table 1 and by considering likelihoods for datasets that have been cross-checked for consistency (via Eq. 1*) with a plurality of external datasets (CMB data from Planck; BAO data from 6dF Galaxy Survey; BOSS Data Release; SNe Ia data from the JLA). The refined estimated values for the parameters (with their margins of error) are shown in Table 2.

---

[10] Photo-z are estimates of photometric redshifts that affect both lens galaxies and source galaxies in weak lensing. As such they also affect measurements and calibration of cosmic shears. Such measurements are challenging due to noise and systematic errors (not all galaxy images are high resolution, and there might be small, faint galaxies that are very difficult to measure accurately). One way of estimating shear is via fitting models, where a model with parameters (known as 'shear estimator') is used to calculate the gravitationally distorted shape of the galaxy by fitting the model to the galaxy surface brightness profile. Obviously, if there are a lot of parameters involved in such fitting models, the 'shear estimator' might itself be subject to 'noise bias' and in need of further calibration. But to calibrate 'noise bias' often another image of a galaxy is used that is itself subject to noise bias. In DES, these photo-z estimates (and their priors) are obtained by a METACALIBRATION galaxy catalogue, which measures the shapes of galaxies via a Gaussian fit to the pixel data for all available band exposures and then calculates the possible gravitational shear.

[11] And they are Gaussian priors with a mean and a variance (rather than wide flat priors) because the idea is to have more informative priors for nuisance parameters to better control galaxy bias. By contrast, in parameter estimation, flat priors are privileged because they are less informative about where the real value of each of those parameters lies within the allowed width and the posterior probability has to be less sensitive to the choice of the priors.

As announced in Section 2, theory-dependent priors for cosmological parameters and data-dependent priors for nuisance parameters cause however a tension between the task (I) of cross-checking datasets for consistency and the task (II) of fixing constraints on parameters. Parameter estimation has to be insensitive to the choice of priors, hence wide top-hat flat priors are chosen that span a reasonably large set of possible values. An example is the matter energy density $\Omega_m$ whose priors range over 0.1–0.9 and whose posterior probabilities in $\Lambda$CDM are computed as in Table 2. To measure these posterior probabilities for the parameters in Table 2, cosmologists have to rely on a variety of datasets coming from different probes (DES + Planck, DES + JLA + BAO, etc., as per the second column in Table 2) that have already been cross-checked as consistent within the $\Lambda$CDM-model via the *R statistic* in Eq. (1*). Priors for cosmological parameters enter into the Bayesian evidence (Eq. 1) and hence into *R statistic* in Eq. (1*).

**Table 1** Priors for the seven main cosmological parameters and nuisance parameters. Reprinted Table 1 with permission from T. M. C. Abbott et. al. (Dark Energy Survey Collaboration) "Dark Energy Survey year 1 results: cosmological constraints from galaxy clustering and weak lensing", Physical Review D 98, 043526-7, 2018] Copyright (2018) by the American Physical Society. https://doi.org/10.1103/PhysRevD.98.043526

| Parameter | Prior |
|---|---|
| Cosmology | |
| $\Omega_m$ | Flat (0.1, 0.9) |
| $A_s$ | Flat ($5\times10^{-10}$, $5\times10^{-9}$) |
| $n_s$ | Flat (0.87, 1.07) |
| $\Omega_b$ | Flat (0.03, 0.07) |
| $h$ | Flat (0.55, 0.91) |
| $\Omega_\nu h^2$ | Flat ($5\times10^{-4}$, $10^{-2}$) |
| $w$ | Flat ($-2$, $-0.33$) |
| Lens galaxy bias | |
| $b_i (i = 1, 5)$ | Flat (0.8, 3.0) |
| Intrinsic alignment | |
| $A_{IA}(z) = A_{IA}[(1 + z)/1.62]^{\eta_{IA}}$ | |
| $A_{IA}$ | flat ($-5$, 5) |
| $\eta_{IA}$ | flat ($-5$, 5) |
| Lens photo-$z$ shift (red sequence) | |
| $\Delta z_l^1$ | Gauss (0.008, 0.007) |
| $\Delta z_l^2$ | Gauss ($-0.005$, 0.007) |
| $\Delta z_l^3$ | Gauss (0.006, 0.006) |
| $\Delta z_l^4$ | Gauss (0.000, 0.010) |
| $\Delta z_l^5$ | Gauss (0.000, 0.010) |
| Source photo-$z$ shift | |
| $\Delta z_s^1$ | Gauss ($-0.001$, 0.016) |
| $\Delta z_s^2$ | Gauss ($-0.019$, 0.013) |
| $\Delta z_s^3$ | Gauss ($+0.009$, 0.011) |
| $\Delta z_s^4$ | Gauss ($-0.018$, 0.022) |
| Shear calibration | |
| $m^i_{\mathrm{METACALIBRATION}}\ (i = 1, 4)$ | Gauss (0.012, 0.023) |
| $m^i_{\mathrm{IM3SHAPE}}(i = 1, 4)$ | Gauss (0.0, 0.035) |

Table 2 Constraints and margins of error for the cosmological parameters in $\Lambda$CDM and $w$CDM. Reprinted Table 2 with permission from T.M.C. Abbott et. al. (Dark Energy Survey Collaboration) "Dark Energy Survey year 1 results: cosmological constraints from galaxy clustering and weak lensing", Physical Review D 98, 043526-17, 2018. Copyright 2018) by the American Physical Society. https://doi.org/10.1103/PhysRevD.98.043526

| Model | Data sets | $\Omega_m$ | $S_8$ | $n_s$ | $\Omega_b$ | $h$ | $\sum m_\nu (eV)$ (95% CL) | $w$ |
|---|---|---|---|---|---|---|---|---|
| $\Lambda$CDM | DES Y1 $\xi_\pm(\theta)$ | $0.260^{+0.065}_{-0.037}$ | $0.782^{+0.027}_{-0.027}$ | ... | ... | ... | ... | ... |
| $\Lambda$CDM | DES Y1 $w(\theta)+\gamma_t$ | $0.288^{+0.045}_{-0.026}$ | $0.760^{+0.033}_{-0.030}$ | ... | ... | ... | ... | ... |
| $\Lambda$CDM | DES Y1 $3\times2$ | $0.267^{+0.030}_{-0.017}$ | $0.773^{+0.026}_{-0.020}$ | ... | ... | ... | ... | ... |
| $\Lambda$CDM | Planck (no lensing) | $0.334^{+0.037}_{-0.026}$ | $0.841^{+0.027}_{-0.025}$ | $0.958^{+0.008}_{-0.005}$ | $0.0503^{+0.0046}_{-0.0019}$ | $0.658^{+0.019}_{-0.027}$ | ... | ... |
| $\Lambda$CDM | DES Y1 + Planck (no lensing) | $0.297^{+0.016}_{-0.012}$ | $0.795^{+0.020}_{-0.013}$ | $0.972^{+0.006}_{-0.004}$ | $0.0477^{+0.0016}_{-0.0012}$ | $0.686^{+0.009}_{-0.014}$ | <0.47 | ... |
| $\Lambda$CDM | DES Y1 + JLA + BAO | $0.295^{+0.018}_{-0.014}$ | $0.768^{+0.018}_{-0.023}$ | $1.044^{+0.019}_{-0.087}$ | $0.0516^{+0.0050}_{-0.0080}$ | $0.672^{+0.049}_{-0.034}$ | ... | ... |
| $\Lambda$CDM | Planck + JLA + BAO | $0.306^{+0.007}_{-0.007}$ | $0.815^{+0.015}_{-0.013}$ | $0.969^{+0.004}_{-0.005}$ | $0.0483^{+0.0008}_{-0.0006}$ | $0.678^{+0.007}_{-0.005}$ | <0.22 | ... |
| $\Lambda$CDM | DES Y1 + Plank+JLA + BAO | $0.298^{+0.007}_{-0.007}$ | $0.802^{+0.012}_{-0.012}$ | $0.973^{+0.005}_{-0.004}$ | $0.0479^{+0.0007}_{-0.0008}$ | $0.685^{+0.005}_{-0.007}$ | <0.26 | ... |
| $w$CDM | DES Y1 $\xi_\pm(\theta)$ | $0.274^{+0.073}_{-0.042}$ | $0.777^{+0.036}_{-0.038}$ | ... | ... | ... | ... | $-0.99^{+0.33}_{-0.39}$ |
| $w$CDM | DES Y1 $w(\theta)+\gamma_t$ | $0.310^{+0.049}_{-0.036}$ | $0.785^{+0.040}_{-0.072}$ | ... | ... | ... | ... | $-0.79^{+0.22}_{-0.39}$ |
| $w$CDM | DES Y1 $3\times2$ | $0.284^{+0.033}_{-0.030}$ | $0.782^{+0.036}_{-0.024}$ | ... | ... | ... | ... | $-0.82^{+0.21}_{-0.20}$ |
| $w$CDM | Planck (no lensing) | $0.222^{+0.069}_{-0.024}$ | $0.810^{+0.029}_{-0.036}$ | $0.960^{+0.005}_{-0.007}$ | $0.0334^{+0.0099}_{-0.0032}$ | $0.801^{+0.045}_{-0.097}$ | ... | $-1.47^{+0.31}_{-0.22}$ |
| $w$CDM | DES Y1 + Planck (no lensing) | $0.233^{+0.025}_{-0.033}$ | $0.775^{+0.021}_{-0.021}$ | $0.971^{+0.004}_{-0.006}$ | $0.0355^{+0.0050}_{-0.0039}$ | $0.775^{+0.056}_{-0.040}$ | <0.65 | $-1.35^{+0.16}_{-0.17}$ |
| $w$CDM | Planck + JLA + BAO | $0.303^{+0.010}_{-0.008}$ | $0.816^{+0.014}_{-0.013}$ | $0.968^{+0.004}_{-0.006}$ | $0.0479^{+0.0016}_{-0.0014}$ | $0.679^{+0.013}_{-0.008}$ | <0.27 | $-1.02^{+0.05}_{-0.05}$ |
| $w$CDM | DES Y1 + Plank+JLA + BAO | $0.301^{+0.007}_{-0.010}$ | $0.801^{+0.011}_{-0.012}$ | $0.974^{+0.005}_{-0.005}$ | $0.0483^{+0.0013}_{-0.0016}$ | $0.680^{+0.013}_{-0.008}$ | <0.31 | $-1.00^{+0.05}_{-0.04}$ |
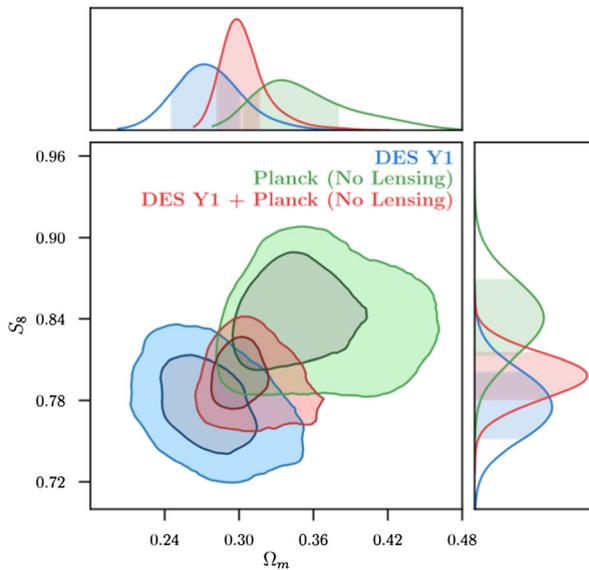
**Fig. 1** Mapping DES Y1 dataset with Planck datases on the $\Omega_m$ —$S_8$ space. Reprinted Fig. 10 with permission from T.M.C. Abbott et. al. (Dark Energy Survey Collaboration) "Dark Energy Survey year 1 results: cosmological constraints from galaxy clustering and weak lensing", Physical Review D 98, 043526-20, 2018. Copyright (2018) by the American Physical Society. https://doi.org/10.1103/PhysRevD.98.043526

Here is an example of dataset comparison from two different probes: CMB from Plank and lensing from DES Y1. The datasets are plotted in a one-dimensional space defined by the matter energy density $\Omega_m$ and another parameter $S_8$ defined as follows

$$S_8 \equiv \sigma_8 \left( \frac{\Omega_m}{0.3} \right)^{0.5} \tag{5}$$

which measures the root mean square amplitude of mass fluctuations, $\sigma_8$, or in other words the present-day clumpiness of the universe. Both these two parameters $\Omega_m$ and $S_8$ can be determined from either Planck CMB data or DES lensing data, so it is possible to see whether the ΛCDM predictions are correct. It is like taking two snapshots of the universe. The CMB dataset gives an image of the universe's growth of structure, when the universe was only 380,000 years old, while DES Y1 dataset gives an image of the universe ten billion years later. Any tension between DES dataset and Planck dataset might imply that the ΛCDM-based predictions on the growth of structure might not be correct (assuming uncertainties and systematics have been correctly estimated). The result of this comparison can be found in Fig. 1 from Abbott et al. 2018.

There is some visible tension between the DES Y1 data and the Planck data, and this is addressed in the following comment that accompanies the figure:

The two-dimensional constraints shown in Fig. 1 [Fig. 10 in original] visually hint at tension […] However, a more quantitative measure of consistency in the full 26-parameter space is the Bayes factor. […] The Bayes factor for combining

DES and Planck (no lensing) in the ΛCDM model is R = 6.6 indicating
"substantial" evidence for consistency on the Jeffreys scale, so any inconsistency
apparent in Fig. 1 [Figure 10 in original] is not statistically significant according
to this metric. (Abbott et al. 2018, 043526–20)

With these caveats, DES concludes that the red contour in the figure captures the "true
parameters", where "it is not unlikely for two independent experiments to return the
blue and green contour regions" (ibid., 043526–20). Similarly:

The DES + BAO + SNe solution shows good consistency in the $\Omega_m$ —$w$— $S_8$
subspace and yields our final constraint on the dark energy equation of state:

$$w = -1^{+0.05}_{-0.04}$$

….The evidence ratio $R_w = 0.1$ for this full combination of data sets, disfavouring
the introduction of $w$ as a free parameter. (Abbott et al. 2018, 043526–23)

In the next Section, I take a look at two epistemic limits arising from the use of
Bayesianism in observational cosmology as a reminder that conclusions about "true
parameters" derived from datasets via Bayesian methods should always be taken with
some caution.

## 4 Two epistemic limits of using Bayesianism in observational cosmology

### 4.1 Methodological bootstrapping and context-dependence of standards of evidence

A distinctive kind of *methodological bootstrap* is at play in delivering on task (I)
datasets consistency cross-checks and task (II) parameter estimation. For the choice of
priors that feed into the Bayesian evidence (Eq. 1), and indirectly in Eq. (1*) to deliver
on task (I) is affected by pre-existing choices with regard to not just controlling
systematic uncertainties in relation to nuisance parameters but also estimating important
cosmological parameters. In other words, datasets cross-checks are the outcome of
specific choices of *theory-dependent* and *data-dependent* priors that convey more or
less background information.

In turn, datasets that have been cross-checked for consistency using these priors in the
Bayesian evidence and $R$ statistic (Eqs. 1 and 1*) feed into the calculation of the maximum
posterior probabilities for these very same cosmological parameters in task (II). Maximum
posterior probabilities are the most likely points in the parameter space within the range of
allowed possible values by the priors, and they are listed for all the main cosmological
parameters (in either ΛCDM or $w$CDM) in Table 2 (the margins of error beside each value
reflect a number of systematic uncertainties and errors affecting the datasets listed on the left
and used to update probabilities). In other words, priors for parameters originally set to assess
how consistent datasets are with respect to a given model $M_1$ are subsequently used to

calculate (via Eq. 3) a new round of estimates for the very same theoretical parameters (and in our example, to conclude for example from the joint DES Y1 and Planck datasets that the value for the parameter $h$ must be around 0.686).

These remarks are neither meant to cast doubts on the validity of the DES's statistical analysis, nor to suggest any vicious circularity. For it is part of the Bayesian framework that maximum conditional probabilities for cosmological parameters are updated as more datasets are cross-checked and found to be consistent given a model $M_1$. The remarks are instead meant to highlight a distinctive epistemological feature concerning Bayesian analysis in observational cosmology: namely, that there is no 'empirically rock-solid' ground in observational cosmology and that model-building and model-confirmation via Bayesian statistics work as "an evolutionary process" (to echo Kass and Raftery 1995, p. 773). I'd like to think of this 'evolutionary process' in analogy with Neurath's boat as a methodological stance in observational cosmology: there are no first foundations, there is no starting from scratch, and building is effectively always a rebuilding (i.e. rebuilding the boat while adrift at sea). Analogously, model-building and model-selection in cosmology is an exercise in rebuilding, refining and improving on existing parameter estimates of current models via new, expanded, more diversified datasets within the constraints of Bayesian methods.

Consider, for example, the priors for the Hubble parameter $h$ in Table 1. They are originally chosen to be flat and to range *uninformatively* between 0.55 and 0.91. This choice of priors width is intentionally uninformative with an eye to avoiding being caught up in the current controversy about the degeneracy of the value for the Hubble constant $H_0$ where different tests have produced slightly diverging measurements. Using SNe Ia calibrated by Cepheids, Riess et al. (2016) measured the value for the Hubble constant at $73.24 \pm 1.74$ km s$^{-1}$ Mpc$^{-1}$. This value is in 3.4σ tension with the latest news from Planck CMB data (see Aghanim et al. 2016; and Bernal et al. 2016 for an excellent discussion). And to complicate matters still further, in July 2019 Wendy Freedman and collaborators have used measurements of luminous red giant stars to give a new value of the Hubble constant at $69.8 \pm 1.9$ km s$^{-1}$ Mpc$^{-1}$, which is roughly half-way between Planck and the H0LiCOW values (Freedman et al. 2019).

Thus, DES Y1's choice to set the theory-dependent priors for $h$ flat between 0.55–0.91 is intended to be as uninformative as possible about where the actual value for $h$ might lie in this allowed spectrum (let us pretend we are under a veil of ignorance). But the fact is that we do *know* from the aforementioned discrepant measurements of the Hubble constant that the reduced Hubble parameter $h$ must be peaked somewhere around 0.7. That means that most of the posterior volume of the DES Y1 $h$ (as compatible with the chosen range 0.55–091 for the priors) is bound to be empty. Uninformative theory-dependent flat priors risk having a mostly empty posterior volume. And the problem with an empty posterior volume is that if we are trying to establish how likely the evidence $D$ is given a model $M_1$ (via Bayesian evidence in Equ. 1), it is desirable to have better constrained parameters $\theta_i^{M_1}$ than loosely constrained ones with a mostly empty posterior volume.

On the other hand, if we try to improve the fit with the model in the Bayesian evidence (Equ. 1) by shortening the range of the prior to custom tailor it to the available *known* measurements, although the parameter estimation will not be affected, dataset comparison will be strongly affected by a narrower range of values. This is an undesired feature of using the $R$ statistic (Equ. 1*) for cross-checking dataset consistency.

Recall that since $R$ (Equ. 1*) depends on the priors of the shared parameters —i.e., $p\left(\theta_i^{M_1}|M_1\right)$ —decreasing the width of the range of the priors (to improve the fit of the model to the Bayesian evidence, Equ. 1) has the side-effect of decreasing $R$ and the associated ability to cross-check for consistency the two datasets. Thus, choosing the right width for the flat priors is paramount. Too large a width for uninformative priors reduces the ability to fit the relevant model to the Bayesian evidence (Eq. 1). Too narrow a width for more informative priors improves the fit of the model to the Bayesian evidence for an individual dataset, at the cost of decreasing the consistency with other independent datasets. Ideal priors must lie somewhere in the Bayesian Goldilock region, metaphorically speaking: their width must be neither too narrow nor too wide, but 'just right'. Indeed, their width must be the narrowest allowed range that does not force $R$ to fall below 1, i.e. that does not skew consistency cross-checks.

Now, one possible strategy to mitigate this prior-dependency in datasets consistency cross-check has recently been proposed by Handley and Lemos (2019). They propose to interpret the $R$ statistic as consisting of two parts: (a) what might be called the *information ratio I* defined by the Kullback-Leibler divergence that gives a logarithmic information ($\log I$) measure of how unlikely it is that the two datasets might match given a certain choice of the priors; and (b) a logarithmic measure of the mismatch between two datasets that Handley and Lemos call *suspiciousness S* (or $\log S$) and it is defined as the difference between the logarithmic version of $R$ (i.e. $\log R$) and the Kullback-Leibler divergence ($\log I$). Suspiciousness $S$ is designed to remove or at least mitigate the dependence on the choice of priors that affect both $\log R$ *and* $\log I$ as illustrated by the following Table 3.

Reinterpreting $R$ along these lines implies rethinking DES Y1 outcomes and especially the tension between the DES Y1 weak galaxy lensing datasets and Planck datasets in Fig. 1. The jury on this specific tension is still very much out at this point

**Table 3** Comparing prior dependency for $\log R$, $\log I$, and $\log S$, with $\widetilde{d}$ as the Bayesian combined model dimensionality indicating the number of shared constrained parameters between datasets, and $p$ is the tension probability. Reprinted (Table 2) with permission from: Will Handley and Pablo Lemos "Quantifying tensions in cosmological parameters: interpreting the DES evidence ratio", *Physical Review D* **100**, 043504-12 (Handley and Lemos 2019). Copyright (2019) by the American Physical Society. https://doi.org/10.1103/PhysRevD.100.043504

| Dataset | Prior | $\log R$ | $\log I$ | $\log S$ | $\widetilde{d}$ | $p(\%)$ |
|---------|-------|----------|----------|----------|-----------------|---------|
| BOSS-*Planck* | default | $6.30 \pm 0.29$ | $6.18 \pm 0.29$ | $0.11 \pm 0.29$ | $2.91 \pm 0.51$ | $42.66 \pm 4.28$ |
| | medium | $4.51 \pm 0.28$ | $4.06 \pm 0.28$ | $0.46 \pm 0.28$ | $3.30 \pm 0.55$ | $55.12 \pm 4.47$ |
| | narrow | $1.30 \pm 0.23$ | $0.69 \pm 0.22$ | $0.61 \pm 0.22$ | $1.67 \pm 0.54$ | $77.12 \pm 14.10$ |
| DES-*Planck* | default | $2.88 \pm 0.35$ | $6.15 \pm 0.34$ | $-3.28 \pm 0.34$ | $3.97 \pm 0.82$ | $3.23 \pm 1.00$ |
| | medium | $0.51 \pm 0.34$ | $4.00 \pm 0.34$ | $-3.49 \pm 0.34$ | $3.31 \pm 0.81$ | $2.04 \pm 0.79$ |
| | narrow | $-1.88 \pm 0.29$ | $0.90 \pm 0.29$ | $-2.78 \pm 0.29$ | $1.15 \pm 0.77$ | $1.44 \pm 0.91$ |
| SH$_0$ES- *Planck* | default | $-2.03 \pm 0.29$ | $1.96 \pm 0.28$ | $-3.99 \pm 0.28$ | $0.78 \pm 0.52$ | $0.25 \pm 0.17$ |
| | medium | $-2.50 \pm 0.28$ | $1.56 \pm 0.28$ | $-4.06 \pm 0.28$ | $1.77 \pm 0.51$ | $0.56 \pm 0.24$ |
| | narrow | $-2.00 \pm 0.23$ | $1.43 \pm 0.23$ | $-3.43 \pm 0.23$ | $1.92 \pm 0.52$ | $1.17 \pm 0.45$ |

in time. But let us be clear about the philosophically interesting point concerning this methodological bootstrap. The choice of the width of the priors (default, medium/ 'Goldilock', or narrow) is going to affect the measure of the dataset consistency as Table 3 clearly highlights. To come back to my main point and sum it up, a peculiar kind of bootstrapping seems to affect the passage from task I to task II. To perform task II, uninformative large flat priors are desirable. But to perform task I, informative narrower flat priors are better as long as they do not skew the consistency cross-checks. *Even* in the best-case scenario of an original choice of physically reasonable 'Goldilock' priors for the relevant parameters of a cosmological model (i.e. the ΛCDM) any joint fit to the model of independent datasets in task I (using Equ. 1*) ends up 'bootstrapping' the original choice of the priors that enter into the next round of parameter estimation (task II). What was an originally educated guess of choosing priors with widths that are neither too large (at the cost of an empty posterior volume) nor too narrow (at the risk of jeopardising cross-checks) ends up sanctioning itself as one moves from task I to task II.

## 4.2 Parameter estimation and model selection: A Bayesian trade-off

A second and different kind of tension arises from the use of Bayesianism in task (II)— i.e. parameter estimation—and task (III)—i.e. model selection—and once again it is caused by the specific choice of priors that enter into both. As we have seen, parameter estimation requires the choice of uninformative / wide top-hat priors to deliver posterior probabilities that are as insensitive as possible to the choice of priors, especially in open-ended and controversial cases (such as the current debate surrounding the measurement of the Hubble constant).

However, the choice of wide flat priors is not just in tension with the Bayesian evidence in task (I), as already explained in Section 4.1. It is also methodologically not innocent when it comes model selection (task III). In particular, a Bayes factor (Equ. 2) that has very wide flat priors for a parameter $\theta_i^{M_2}$   (with $\theta_i^{M_2} \rightarrow \infty$) tends to favour (with $p = 1$) the so-called 'null hypothesis' when it comes to the comparative assessment of evidence in the choice between different models. This phenomenon is known in statistics as Bartlett's paradox (see Raftery 1996 for a discussion). The history of the paradox is slightly complicated as Bartlett (1957) is effectively a commentary on D.V. Lindley (1957), where the so-called Lindley's paradox is presented. The latter concerns a phenomenon originally observed by Jeffreys himself and highlights a conflict between the following two statistical scenarios (frequentist and Bayesian, respectively) concerning testing a hypothesis *H* with some experimental outcome *x*:

(i)   A frequentist significance test for *H* reveals that *x* is significant at 5% level;
(ii)   The Bayesian posterior probability for *H* given *x*, and given a narrow width of prior probabilities for *H*, is as high as 95%.

The original Lindley's paradox is meant to highlight a tension between significance testing and Bayes's theorem when it comes to null hypotheses testing. For example, one can imagine that the hypothesis *H* involves a parameter $\theta_i^H$ which can take *i* possible values and that the null or default hypothesis $H_0$ assumes that the parameter

takes a specific value, e.g. $\theta_i^{H_0} = m$. Suppose we run the relevant experiment several times and collect a random large sample $n$ of experimental outcomes $x = (x_1, x_2, \ldots, x_n)$ with a Gaussian distribution having a mean (call it $m$) and a variance $\sigma^2$. Let the prior probability for the null hypothesis be $p\,(H_0) > 0$ and any alternative scenario from $H_0$ is assigned a top-hat flat prior. In Bayesian terms, the posterior probability $p\,(H_0\,|x)$— namely, the probability that $\theta_i^{H_0} = m$ given experimental outcomes $x$—tends to 1 whenever $n \to \infty$: the null hypothesis tends to be favoured. Lindley's original paradox showed how this Bayesian measure for null hypotheses testing was at odds with the frequentist counterpart, where in an experiment a significance testing at 5% gives in fact very strong reasons to doubt the null hypothesis.

Bartlett's two-page (Bartlett 1957) commentary on Lindley's original paper pointed out a missing extra factor in one of Lindley's formulas concerning the prior distribution over a certain range $I$ for the alternative hypothesis (i.e. $H \neq H_0$). This meant that in situations where one might be tempted to stretch the range $I$ of the uniform prior for the rival hypothesis to infinity, the "silly answer" (Bartlett 1957, 533) follows that the posterior probability for the null hypothesis becomes 1. Thus, effectively, what is known as Bartlett's paradox highlights a specific feature in the choice of the width of the flat priors for the non-null hypothesis that was implicit or better missing in Lindley's original paradox.[12]

Bartlett's paradox becomes particularly pressing in cosmology where the choice of wide flat priors causes a trade-off effectively between parameter estimation (where wide priors are required for the reasons mentioned in Section 4.1) and model selection (where wide priors have the effect of statistically favouring the null hypothesis—in this case the $\Lambda$CDM model—over possible rival ones). And since $w$CDM takes the dark energy equation of state parameter $w$ as free (rather than fixed at $-1$ as in $\Lambda$CDM), the Bayes factor at play in Equ. (4) does not level the playing field in the model selection between $\Lambda$CDM and $w$CDM.

Bartlett's paradox is a reminder of the risk of what is sometimes called "the fallacy of acceptance" (to echo Spanos 2013): it is a fallacy to conflate "accept the null hypothesis" (there is no inconsistent evidence against it) with "there is evidence *for* the null hypothesis". What Bayesian analysis shows is that a plurality of datasets are consistent with $\Lambda$CDM, with $R_w = 0.1$ favouring $\Lambda$CDM over $w$CDM. But this Bayesian way of doing model selection should not of course be read as licensing more general conclusions about which model is 'true' or what the "true parameters" are. In other words, one should avoid reading Premises 1–3

(Premise 1): The probability of finding the dataset $D_1$ (which is actually found) is high, given model $M_1$ [Bayesian evidence, Equ. 1]
(Premise 2): The probability of jointly finding datasets $D_1$ and $D_2$ is high, given model $M_1$ [R statistic, Equ. 1*]

---

[12] The accompanying Editorial Note to Bartlett's paper reads as follows: "The point raised by Prof. Bartlett's second paragraph is related to the difficulty of laying down a uniform prior probability for a parameter of infinite range, a point which in my opinion has not been properly cleared up…The root of this difficulty seems to be that several limiting processes are involved and no clear rules have been laid down as to which, if any, has priority. In any case this point mainly concerns estimation, whereas Mr. Lindley was concerned with testing hypotheses" Bartlett (1957), p. 534.

(Premise 3): The posterior probability of model $M_1$ given joint datasets $D_1$ & $D_2$ is higher than the posterior probability of model $M_2$ given $D_1$ & $D_2$ [using a tweaked version of Equ. 4]

as somehow licensing.

(Conclusion 3): Therefore, there is substantial evidence for $M_1$.

This is just another way of re-stating the more general point that Bayesian methodology gives us only *relative* and not *absolute* measures for model selection and having evidence that increases the posterior probability of a model over a rival one is not one and the same as concluding that therefore $M_1$ is the 'true' model (unless the word 'true' is here used in some very loose and unspecified sense). For there might be other rival models (beyond $w$CDM and the specific issue of the Bartlett's paradox here considered) that have not yet been examined, or whose evidence (for or against) has not yet been evaluated using the Bayes factor. And those rival models remain effectively all live candidates worth exploring and examining in future research.

## 5 Concluding remarks

Bayesianism provides a ubiquitous and very powerful tool to allow comparison among different datasets, and to deliver on parameter estimate and model selection in contemporary observational cosmology. The philosophical goal of this paper was to highlight the power but also the epistemic limits of using Bayesianism in delivering on these different tasks.

## Compliance with ethical standards

**Conflict of interest**   None.

**Ethical approval**   Not applicable.

**Informed consent**    Not applicable.

# References

Abbott, T. M. C., DES Collaboration, et al. (2018). Dark Energy Survey year 1 results: cosmological constraints from galaxy clustering and weak lensing. *Physical Review D, 98*, 1–31. https://doi.org/10.1103/PhysRevD.98.043526

Ade, P. A. R., et al. (2016). Planck 2015 results. XIII. Cosmological parameters. *Astronomy & Astrophysics, 594*(A13), 1–63.

Aghanim, N., Planck Collaboration, et al. (2016). Planck intermediate results. XLVI. Reduction of large scale systematic effects in HFI polarization maps and estimation of the reionization optical depth. *Astronomy and Astrophysics, 596*(A107), 1–52.

Amendola, L. & Tsujikawa, S. (2010). *Dark Energy. Theory and Observations*. Cambridge University Press.

Anderl, S. (2016). Astronomy and astrophysics. In P. Humphreys (Ed.), *The Oxford handbook of philosophy of science* (pp. 652–670). Oxford University Press.

Anderl, S. (2018). Simplicity and simplification in astrophysical modelling. *Philosophy of Science, 85*, 819–831.

Bartlett, M. S. (1957). Comment on 'a statistical paradox' by D. V. Lindley. *Biometrika, 44*, 533–534.

Beisbart, C. (2009). Can we justifiably assume the cosmological principle in order to break model Underdetermination in cosmology? *Journal for General Philosophy of Science, 40*, 175–205.

Benétreau-Dupin, Y. (2015). The Bayesian who knew too much. *Synthese, 192*, 1527–1542.

Bernal, J. L., Verde, L., & Riess, A. G. (2016). The trouble with $H_0$. *Journal of Cosmology and Astroparticle Physics, 10*(019), 1–28.

Charnock, T., Battye, R. A., et al. (2017). Planck data versus large scale structure. Methods to quantify discordance. *Physical Review D, 95*(12), 123535.

De Baerdemaeker, S. (forthcoming). Method-driven experiments and the search for dark matter. *Philosophy of Science*.

Freedman, W., et al. (2019) The Carnegie-Chicago Hubble program. VIII. An independent determination of the Hubble constant based on the Tip of the Red Giant Branch. *The Astrophysical Journal, 882*, 1–29.

Handley, W., & Lemos, P. (2019). Quantifying tensions in cosmological parameters: Interpreting the DES evidence ratio. *Physical Review D, 100*(043504), 1–15.

Jeffreys, H. (1939/1961). *Theory of probability*. 3rd ed. Oxford University Press.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association, 90*(430), 773–795.

Liddle, A., Mukherjiee, P. & Parkinson, D. (2009). Model selection and multi-model inference. In Hobson et al (eds.) *Bayesian Methods in Cosmology.* Cambridge University Press, pp. 79–98.

Lindley, D. V. (1957). A statistical paradox. *Biometrika, 44*, 187–192.

Marshall, P., Rajguru, N., & Slosar, A. (2006). Bayesian evidence as a tool for comparing datasets. *Physical Review D, 73*, 067302.

Massimi, M. (2020). A philosopher's look at DES. Reflections on the use of the Bayes factor in cosmology. In O. Lahav, L. Calder, J. Mayers, & J. Frieman (Eds.), *The dark energy survey. The story of a cosmological experiment* (pp. 357–72). World Scientific.

Raftery, A. E. (1996). Approximate Bayes factors and accounting for model uncertainty in generalised linear models. *Biometrika, 83*(2), 251–266.

Riess, A. G., et al. (2016). A 2.4% determination of the local value of the Hubble constant. *The Astrophysical Journal, 826*(56), 1–31.

Ruphy, S. (2016). *Scientific Pluralism Reconsidered. A new approach to the (dis)unity of science*. Pittsburgh, University of Pittsburgh Press.

Skilling, J. (2011). Foundations and algorithms. In M.P. Hobson (Ed.), *Bayesian Methods in Cosmology,* Cambridge University press, pp. 3–35.

Smeenk, C. (2013). Philosophy of cosmology. In R. Batterman (Ed.), *The Oxford handbook of philosophy of physics*. Oxford University Press.

Spanos, A. (2013). Who should be afraid of the Jeffreys-Lindley paradox? *Philosophy of Science, 80*, 73–93.

Trotta, R. (2008). Bayes in the sky: Bayesian inference and model selection in cosmology. *Contemporary Physics, 49*, 71–104.

Verde, L. (2014). Precision cosmology, accuracy cosmology and statistical cosmology. *Statistical Challenges in 21st century cosmology, Proceedings of the International Astronomical Union, IAU Symposium*, Volume 306, pp. 223–234.