Quantifying tensions in cosmological parameters: Interpreting the DES evidence ratio

Will Handley[1,2,3,*] and Pablo Lemos[4,†]

[1]*Astrophysics Group, Cavendish Laboratory, J.J. Thomson Avenue,*
Cambridge CB3 0HE, United Kingdom
[2]*Kavli Institute for Cosmology, Madingley Road, Cambridge CB3 0HA, United Kingdom*
[3]*Gonville & Caius College, Trinity Street, Cambridge CB2 1TA, United Kingdom*
[4]*Department of Physics and Astronomy, University College London, Gower Street,*
London WC1E 6BT, United Kingdom

(Received 12 February 2019; published 1 August 2019; corrected 27 March 2020)

We provide a new interpretation for the Bayes factor combination used in the Dark Energy Survey (DES) first year analysis to quantify the tension between the DES and *Planck* datasets. The ratio quantifies a Bayesian confidence in our ability to combine the datasets. This interpretation is prior dependent, with wider prior widths boosting the confidence. We therefore propose that if there are any reasonable priors which reduce the confidence to below unity, then we cannot assert that the datasets are compatible. Computing the evidence ratios for the DES first year analysis and *Planck*, given that narrower priors drop the confidence to below unity, we conclude that DES and *Planck* are, in a Bayesian sense, incompatible under ΛCDM. Additionally we compute ratios which confirm the consensus that measurements of the acoustic scale by the Baryon Oscillation Spectroscopic Survey (BOSS) are compatible with *Planck*, while direct measurements of the acceleration rate of the Universe by the Supernovae and H_0 for the Equation of State of Dark Energy Collaboration (SH_0ES) are not. We propose a modification to the Bayes ratio which removes the prior dependency using Kullback-Leibler divergences, and using this statistical test we find *Planck* in strong tension with SH_0ES, in moderate tension with DES, and in no tension with BOSS. We propose this statistic as the optimal way to compare datasets, ahead of the next DES data releases, as well as future surveys. Finally, as an element of these calculations, we introduce in a cosmological setting the Bayesian model dimensionality, which is a parametrization-independent measure of the number of parameters that a given dataset constrains.

DOI: 10.1103/PhysRevD.100.043504

I. INTRODUCTION

The analysis of the first year of data from the Dark Energy Survey [1] (henceforth DES Y1) has generated considerable discussion. DES Y1 analyzed data from cosmic shear, galaxy clustering, and galaxy-galaxy lensing (an analysis they refer to as “3x2” since it combines three two-point functions). This data combination is particularly suited to constraining the present-day matter density Ω_m and the parameter σ_8, defined as the present-day linear theory root-mean-square amplitude of the power spectrum of matter fluctuations, averaged in spheres of radius $8\ h^{-1}\,\mathrm{Mpc}$, where h is the Hubble constant in units of $100\ \mathrm{km\,s^{-1}\,Mpc^{-1}}$. Before the publication of DES Y1, this parameter combination measured by weak lensing had already generated controversy, with claims of tensions with respect to the cosmic microwave background (CMB) values measured by *Planck* [2] by both the CFHTLenS and Kilo Degree Survey (KiDS) collaborations [3–5]. While this discrepancy has led to claims of new physics [6], it has also highlighted unknown problems in weak lensing analyses that have reduced these tensions to below significant levels [7–9].

DES Y1 obtained results that appear to be in mild tension with *Planck* (see Fig. 10 of DES Y1), but are reported to be perfectly consistent according to the evidence ratio statistic[1] R used in their analysis to quantify the degree of discordance between 3×2 and CMB data. While this R statistic was proposed some time ago [10], and supported since then by many cosmologists [11–15], it is particularly relevant to consider its precise interpretation in light of present and future tensions arising with increasingly powerful datasets providing ever more

[*] wh260@mrao.cam.ac.uk
[†] pablo.lemos.18@ucl.ac.uk

[1] Here R refers to the Bayes factor combination used in DES Y1 to compare different datasets, not to the Bayes ratio used to compare models.

precise parameter constraints. Other measures of tension between datasets have been proposed in the past [16–26]. A summary of a lot of these methods can be found in [27].

In this paper we argue that R is an appropriate measure of tension, quantifying the Bayesian degree of confidence in the ability to combine the data. However, R has some subtle prior-dependent properties, which has led to its misuse in previous works. We explain these properties and provide Bayesian methods to correctly calibrate the scale on which it sits. We also propose an alternative statistic that preserves the desired properties of R to compare different datasets, including its Bayesian nature, but does not suffer from undesired prior dependences.

The tension between weak galaxy lensing and *Planck* is not the only existing tension in cosmology. Measurements of the expansion rate of the Universe parametrized by the Hubble constant H_0 using type Ia supernovae calibrated by the period-luminosity relation of Cepheids and local distance anchors by the SH_0ES Collaboration [28,29] are in tension with the *Planck* value inferred from the CMB using a ΛCDM cosmology [2]. We use this case as an example of clear tension between experiments. Conversely, the measurements of the baryon acoustic oscillation (BAO) scale and redshift-space distortions (RSD) by BOSS [30] produce values of the parameters Ω_m and σ_8 that are in good agreement with *Planck*. We use this case as an example of no tension between experiments.

The paper is structured as follows: In Sec. II we briefly review the key Bayesian theory and establish notation. In Sec. III we define the logarithmic Bayes and information ratios $\log R$ and $\log I$ and present our new Bayesian interpretation of $\log R$. In Sec. IV we examine analytic examples to aid intuition on the properties of the Bayes and information ratios. In Sec. V we apply our techniques to cosmological datasets, with our key results reported in Table II. We conclude in Sec. VI.

II. BACKGROUND

In general we use the following notation for the quantities in Bayes's theorem:

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)} \Leftrightarrow \mathcal{P}_D(\theta) = \frac{\mathcal{L}_D(\theta)\pi(\theta)}{\mathcal{Z}_D},$$

namely, the posterior $\mathcal{P}$, likelihood $\mathcal{L}$, prior π, and evidence $\mathcal{Z}$. We will retain dataset dependence as a subscript and in general will suppress explicit dependency on θ except where its presence increases clarity. Furthermore there is a suppressed explicit model dependence, which is taken to be ΛCDM for our cosmological examples.

A. Bayesian evidence

Throughout this paper the Bayesian evidence $\mathcal{Z}$, defined as

$$\mathcal{Z}_D = \int \mathcal{L}_D \pi \mathrm{d}\theta, \tag{1}$$

will play a key role. Also known as the marginal likelihood [11], the evidence is a key element of model comparison, and may be computed analytically in some rare cases, but is usually evaluated using a Laplace approximation [31], Savage-Dickey ratio [32], or better still with numerical evidence calculators such as `MCEvidence` [33,34] or nested sampling [35–40].

B. Kullback-Leibler divergence

The Kullback-Leibler divergence [41] is defined as

$$\mathcal{D}_D = \int \mathcal{P}_D(\theta) \log \frac{\mathcal{P}_D(\theta)}{\pi(\theta)} d\theta = \left\langle \log \frac{\mathcal{P}_D}{\pi} \right\rangle_{\mathcal{P}_D}, \tag{2}$$

which quantifies the information gain/compression between prior and posterior and has been used by numerous authors [18,19,42–50]. The angular brackets $\langle f \rangle_p$ in the rightmost expression of Eq. (2) denote the average of f over the distribution p.

C. Bayesian model dimensionality

We define the Bayesian model dimensionality [51] as

$$\frac{\tilde{d}_D}{2} = \left\langle \left(\log \frac{\mathcal{P}_D}{\pi} \right)^2 \right\rangle_{\mathcal{P}_D} - \left\langle \log \frac{\mathcal{P}_D}{\pi} \right\rangle^2_{\mathcal{P}_D}. \tag{3}$$

The quantity $\log[\mathcal{P}_D(\theta)/\pi(\theta)]$ is the Shannon information [52] provided by the posterior relative to the prior at parameter θ, measured in natural bits (nats). As can be seen from Eq. (2), the Kullback-Leibler divergence is the average amount of information provided by the posterior, while Eq. (3) shows that the Bayesian model dimensionality is proportional to the variance of the information provided by the posterior.

It should be noted that an earlier preprint of this paper used an alternative definition of the dimensionality by Spiegelhalter [53], which has several unattractive theoretical qualities when applied to significantly non-Gaussian cases. The fundamental qualitative conclusions remain unchanged from the initial version of this paper, and the newer definition of model dimensionality is examined in greater detail in [51].

D. Combining likelihoods

Independent datasets A and B are combined at the likelihood level via $\mathcal{L}_{AB} = \mathcal{L}_A \mathcal{L}_B$ so that

$$\mathcal{P}_A = \frac{\mathcal{L}_A \pi}{\mathcal{Z}_A}, \qquad \mathcal{P}_B = \frac{\mathcal{L}_B \pi}{\mathcal{Z}_B}, \qquad \mathcal{P}_{AB} = \frac{\mathcal{L}_A \mathcal{L}_B \pi}{\mathcal{Z}_{AB}}. \tag{4}$$

$$\mathcal{Z}_A = \int \mathcal{L}_A \pi \mathrm{d}\theta, \qquad \mathcal{Z}_B = \int \mathcal{L}_B \pi \,\mathrm{d}\theta,$$
$$\mathcal{Z}_{AB} = \int \mathcal{L}_A \mathcal{L}_B \pi \mathrm{d}\theta. \tag{5}$$

In general, new datasets may introduce additional parameters, either because more cosmological parameters are constrained or because additional nuisance parameters associated with foregrounds or instrumentation are required to perform inference. In general θ will be taken to be the span of the entire parameter space of interest.

An important point, often misunderstood by professional practitioners, is that the introduction of unconstrained parameters should not impact proper inference. It is oft quoted that Bayes factors (or equivalently evidences) penalize additional parameters, but in fact Bayes factors only penalize constrained parameters. For example, if one were to perform a model comparison between the six-parameter ΛCDM model and an extension to the model which factored in the age of the cosmologist doing the calculation, then both models would have the same evidence value, since a cosmologist's age is (almost) completely unconstrained by cosmological likelihoods. This is not a bug, but a desirable feature of Bayes factors in their use in consistent inference. The proper Bayesian way to deal with this apparent problem is to exclude such trite models at the model prior level.

III. THE *R* STATISTIC

A. Definition and prior dependence

Given two datasets A and B, the R statistic is defined via the equivalent expressions:

$$R = \frac{\mathcal{Z}_{AB}}{\mathcal{Z}_A \mathcal{Z}_B} = \frac{P(A,B)}{P(A)P(B)} = \frac{P(A|B)}{P(A)} = \frac{P(B|A)}{P(B)}, \tag{6}$$

with all probabilities implicitly conditional on an underlying model (e.g., ΛCDM). A value of $R \gg 1$ is interpreted as both datasets being consistent, while $R \ll 1$ means the datasets are inconsistent. Note that while we assume that the datasets A and B are independent, this does not imply that $R = 1$. Specifically, dataset independence means that likelihoods $\mathcal{L}_D(\theta) = P(D|\theta)$, which are probabilities conditioned on θ, combine by multiplication, but evidences $\mathcal{Z}_D = P(D)$, which are likelihoods marginalized over the prior $\pi(\theta) = P(\theta)$, do not.

In the DES Y1 analysis, R is used to quantify tension, with the Jeffreys scale used as the arbiter for whether models are consistent or not. The interpretation on a Jeffreys scale is somewhat unjustified, as the DES papers do not explain which probability ratio they are placing on the scale.

A second, arguably larger concern is that while R satisfies many of the desiderata that one would hope for from such a quantity (dimensional consistency, symmetry, parametrization invariance, use of Bayesian quantities), it is strongly prior dependent. We can render this dependency explicit by combining Eqs. (4) to (6) yield

$$R = \int \frac{\mathcal{P}_A \mathcal{P}_B}{\pi} \mathrm{d}\theta = \left\langle \frac{\mathcal{P}_B}{\pi} \right\rangle_{\mathcal{P}_A} = \left\langle \frac{\mathcal{P}_A}{\pi} \right\rangle_{\mathcal{P}_B}. \tag{7}$$

Thus, R can be thought of as the posterior average of the ratio of the other posterior to the shared prior. More specifically, R depends on the priors set on constrained parameters shared between likelihoods, but not on the prior on additional nuisance or unconstrained parameters.

It should be noted that this variation is in opposition to the usual evidence prior dependency. Namely, reducing the widths of the prior in general increases evidence. The same reduction of prior widths however will *reduce* the ratio R and increase tension. This is easily understood, since in the R ratio there are two evidences on the denominator with only one in the numerator. In a Bayesian sense this is an attractive balance—you can only evidence-hack at the expense of tension.

It is important to note that the prior dependence of R can only hide existent discordance; i.e., R can indicate that two datasets are in agreement, even when they are not. However, if R indicates that two datasets are discordant, this should be taken seriously, since the prior volume effect only increases the value of R.

B. Bayesian interpretation of *R*

An interpretation that is often posited is that R represents a ratio of probabilities that the shared model parameters come from different universes in comparison with the probability that they come from the same universe. Given that evidences are traditionally used in the context of model comparison, this seems a natural interpretation. However, in order to convert evidences to model probabilities, one requires model priors and for probabilities to be conditioned on the same dataset, which in this case is not true. Raw evidences are probabilities of *data*, not of models.

A correct interpretation can be found by examining the two rightmost expressions in Eq. (6). These expressions show that R represents the relative confidence that we have in dataset A in light of knowing dataset B, compared to the confidence in A alone (and vice versa). If $R > 1$, then B has strengthened our confidence in A by a factor R. If $R \ll 1$, then as Bayesians we should be concerned that there is either a problem with the underlying model or a problem with either or both of the datasets, and therefore avoid combining the two.

Given this interpretation, it is important to understand the prior dependency of R, namely, that decreasing the prior widths on shared parameters reduces our confidence in the ability to combine datasets.

If a Bayesian specifies extremely wide and uniform priors, they are saying that they *a priori* believe the parameter constraints derived from a dataset D could reside anywhere within that region. It is therefore reassuring when two independent datasets result in constraints that are close. We should be proportionally more reassured if our initial prior is wider, as it is proportionally less likely *a priori* that the constraints would give such good agreement.

Some practitioners might consider this prior dependency pathological, rather than the correct behavior of such a probability. In our experience, the primary difference between full Bayesians and other statisticians is that a Bayesian considers this kind of prior-dependent behavior of the analysis a feature rather than a bug.

Given this prior dependency and its sensible interpretation, the approach we advocate is as follows:

Proposition 1. If there are *any* physically reasonable priors which render R significantly less than 1, then as Bayesians we should consider these datasets in tension.

Given that narrowing the priors decreases the value of R, the physically reasonable priors that render the lowest possible value of R are the narrowest priors that do not significantly alter the shape of the posteriors. While such an extreme strategy would provide a definitive lower bound on R, many Bayesians would disagree with such a procedure, as it uses a prior that depends on the posterior. In reality, the most pragmatic approach is to choose reasonable initial priors and then to examine the sensitivity of the conclusions to reasonable alterations to them.

C. Information and suspiciousness

The logarithmic version of Eq. (6) for the Bayes ratio in between two datasets A and B is defined as

$$\log R = \log \mathcal{Z}_{AB} - \log \mathcal{Z}_A - \log \mathcal{Z}_B. \qquad (8)$$

As discussed in the previous section, the Bayesian confidence R has two primary contributions, one from the unlikeliness of two datasets ever matching (proportional to prior) and another in their mismatch. We may quantify the first of these via the information ratio I defined using Kullback-Leibler divergences as

$$\log I = \mathcal{D}_A + \mathcal{D}_B - \mathcal{D}_{AB}. \qquad (9)$$

The remaining part of the Bayesian confidence quantifies the mismatch, which we term the suspiciousness S:

$$\log S = \log R - \log I. \qquad (10)$$

Suspiciousness is unaffected by changing the prior widths as long as this change does not significantly alter the posterior, since the information ratio I and Bayes ratio R transform similarly under prior volume alterations.

It is important to recognize that while $\log S$ is indeed prior independent, in constructing this quantity we have lost the probabilistic interpretation found in $\log R$. More care must be taken to calibrate the scale on which $\log S$ sits, which will be considered at the end of the next section.

IV. ANALYTICAL EXAMPLES

In all of the below, for a graphical understanding, one may substitute $A \leftrightarrow$ *Planck*, $B \leftrightarrow$ SH$_0$ES, DES, or BOSS and consult Figs. 1–3, respectively.

For simplicity, we consider A and B to have the same parameters θ, although the case is easily extended to the case where the likelihoods only share some parameters, in which case our results depend only on those parameters that are shared between likelihoods.

A. Top-hat example

As a simple choice, we consider a top-hat posterior over a multidimensional region $\mathcal{R}_X$, enclosing a volume V_X:

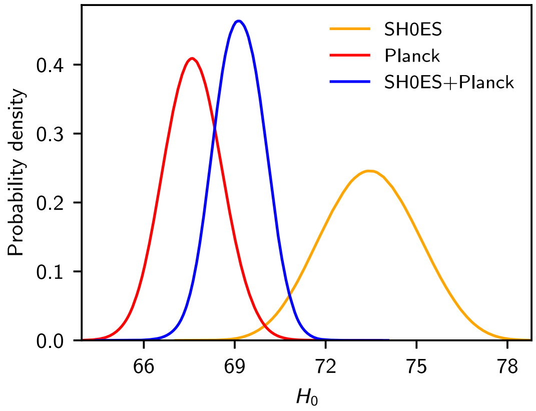


FIG. 1. Tension between the SH$_0$ES and *Planck* datasets as exhibited by examining the posterior parameter constraints on the Hubble constant.

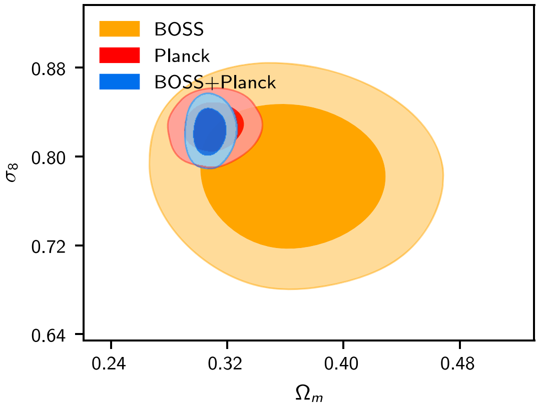


FIG. 2. No tension between BOSS and *Planck* datasets as exhibited by examining the joint posterior parameter constraints on the matter fraction and σ_8.

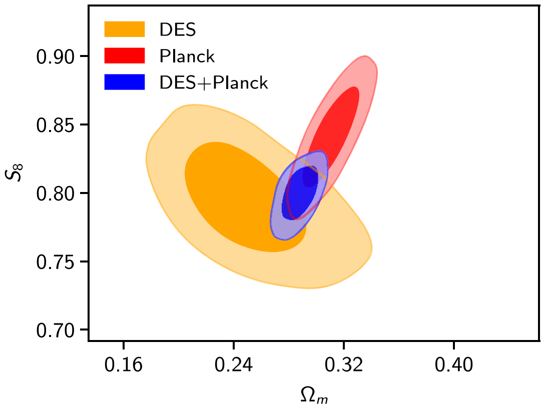


FIG. 3. Possible tension between DES and *Planck* datasets as exhibited by examining the joint posterior parameter constraints on the matter fraction and the parameter combination $S_8 = \sigma_8(\Omega_m/0.3)^{0.5}$.

$$\mathcal{P}_X(\theta) = \begin{cases} V_X^{-1} & : \theta \in \mathcal{R}_X \\ 0 & : \text{otherwise} \end{cases}, \qquad V_X = \int_{\theta\in\mathcal{R}_X} \mathrm{d}\theta. \tag{11}$$

If we have a top-hat prior with volume V_π enclosing two top-hat posteriors P_A and P_B, along with their combined posterior P_{AB}, then

$$\log R = \log I = \log \frac{V_{AB}V_\pi}{V_A V_B}. \tag{12}$$

We can see the explicit prior dependency of R with the presence of the V_π term. Furthermore, we see that R and I are equal in the top-hat posterior, so that the entire contribution to R is in information, and none in suspicion:

$$S = \begin{cases} 1 & : \mathcal{R}_A \cap \mathcal{R}_B \neq \varnothing \\ 0 & : \text{otherwise}. \end{cases} \tag{13}$$

Thus for the uniform case there is no suspiciousness, provided that the posteriors have any overlap region and are thus plausibly consistent.

B. Gaussian example

We now consider a less trivial multivariate Gaussian example [54,55]. A d-dimensional Gaussian likelihood with peak $\mathcal{L}_{\max}$, center μ, and parameter covariance Σ, along with a top-hat enclosing prior over volume V_π, has likelihood, posterior, evidence, and Kullback-Leibler divergence given by the following:

$$\log\mathcal{L}(\theta) = \log\mathcal{L}^{\max} - \frac{1}{2}(\theta-\mu)\Sigma^{-1}(\theta-\mu), \tag{14}$$

$$\log\mathcal{P}(\theta) = -\frac{1}{2}\log|2\pi\Sigma| - \frac{1}{2}(\theta-\mu)\Sigma^{-1}(\theta-\mu), \tag{15}$$

$$\log\mathcal{Z} = \log\mathcal{L}^{\max} + \frac{1}{2}\log|2\pi\Sigma| - \log V_\pi, \tag{16}$$

$$\mathcal{D} = \log V_\pi - \frac{1}{2}(d + \log|2\pi\Sigma|). \tag{17}$$

Note that in the above we have removed explicit dimensionality dependency from the normalization of a Gaussian by exploiting the matrix determinant property $|2\pi\Sigma| = (2\pi)^d|\Sigma|$.

Two likelihoods A and B combine using the relations

$$\log\mathcal{L}_{AB}^{\max} = -\frac{1}{2}(\mu_A-\mu_B)(\Sigma_A+\Sigma_B)^{-1}(\mu_A-\mu_B) + \log\mathcal{L}_A^{\max} + \log\mathcal{L}_B^{\max}, \tag{18}$$

$$\Sigma_{AB}^{-1} = \Sigma_A^{-1} + \Sigma_B^{-1}, \tag{19}$$

$$\mu_{AB} = \Sigma_{AB}[\Sigma_A^{-1}\mu_A + \Sigma_B^{-1}\mu_B]. \tag{20}$$

It should also be noted that

$$(\Sigma_A+\Sigma_B)^{-1} = \Sigma_A^{-1}\Sigma_{AB}\Sigma_B^{-1} = \Sigma_B^{-1}\Sigma_{AB}\Sigma_A^{-1}. \tag{21}$$

We therefore find

$$\log R = -\frac{1}{2}(\mu_A-\mu_B)(\Sigma_A+\Sigma_B)^{-1}(\mu_A-\mu_B) - \frac{1}{2}\log|2\pi(\Sigma_A+\Sigma_B)| + \log V_\pi, \tag{22}$$

and

$$\log I = -\frac{d}{2} - \frac{1}{2}\log|2\pi(\Sigma_A+\Sigma_B)| + \log V_\pi. \tag{23}$$

We thus find that the information content can be used to remove all of the residual prior dependence from $\log R$, giving a suspiciousness

$$\log S = \frac{d}{2} - \frac{1}{2}(\mu_A-\mu_B)(\Sigma_A+\Sigma_B)^{-1}(\mu_A-\mu_B). \tag{24}$$

The numerical value of the suspiciousness is determined by the means and covariances of the posterior distributions A and B. Under a Bayesian interpretation of the posterior, if the “true” value of the measured parameter is θ_0, then both means are drawn from a normal distribution centered on this value with covariance equal to their posterior covariance $\mu_A \sim \mathcal{N}(\theta_0, \Sigma_A)$, $\mu_B \sim \mathcal{N}(\theta_0, \Sigma_B)$, and their difference is drawn from a distribution centered on zero with covariance equal to the sum of the underlying covariances $\mu_A - \mu_B \sim \mathcal{N}(0, \Sigma_A + \Sigma_B)$. One can see that $d - 2\log S$ has a χ_d^2 distribution and that $\log S$ is typically $0 \pm \sqrt{d/2}$. An overly negative value of $\log S$ indicates discordance, and an overly positive value suspicious concordance. More quantitatively, one can use the inverse cumulative χ_d^2 distribution to turn $\log S$ into the tension probability of two datasets being this discordant by chance:

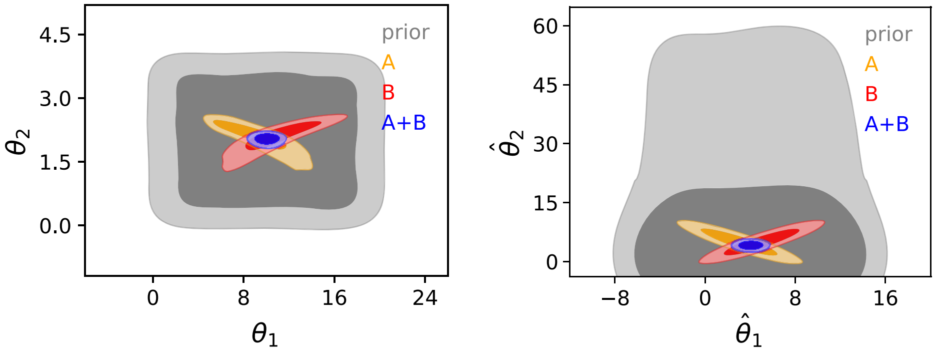


FIG. 4. Many non-Gaussian posteriors (left) may be "Gaussianized" (right) by using Box-Cox transformations.

$$p = \int_{d-2\log S}^{\infty} \chi_d^2(x)\mathrm{d}x = \int_{d-2\log S}^{\infty} \frac{x^{d/2-1}e^{-x/2}}{2^{d/2}\Gamma(d/2)}\mathrm{d}x. \qquad (25)$$

While this procedure is only exact for the Gaussian case, a reasonable proposition for general posteriors would be to compute $\log S$ numerically and then determine tension via a χ^2-like test, in analogy with the Gaussian case:

Proposition 2. If $p \lesssim 0.05$, where p is the tension probability computed from Eq. (25); $\log S$ is computed using numerical evidences and Kullback-Leibler divergences; and $d = \tilde{d}_A + \tilde{d}_B - \tilde{d}_{AB}$ is the Bayesian model dimensionality of the shared constrained parameters computed using Eq. (3), then the datasets should be considered in moderate tension. If $p \lesssim 0.003$, they should be considered in strong tension.[2]

For the case when the posteriors are exactly (or extremely close to) Gaussian, the tension probability p may be interpreted as a probability that one would observe such a discrepancy by chance alone. In the non-Gaussian case, p is only a rough calibration so only extremely small values of p should be regarded with suspicion. The suspiciousness S can be used to determine discordance if $S \ll -\sqrt{\tilde{d}/2}$, and the tension probability p provides a mechanism for putting a number on the concept of $\ll$ in this case. The R statistic, however, is always interpretable as a Bayesian confidence in our ability to combine the data, irrespective of Gaussianity.

It should be noted that many posteriors may be Gaussianized using techniques like Box-Cox transformations [56]. These transformations are nonlinear mappings that can transform complex posteriors into approximately Gaussian ones by changing the parametrization, and they have already been used in the context of cosmology [57,58]. It can be easily proven that these transformations preserve the value of the suspiciousness, although care must be taken to also transform the underlying common prior distribution appropriately (Fig. 4), and that the prior is not significantly distorted by the Box-Cox transformation in the region of the posterior bulk.

Our two propositions for tension quantification are in fact related: one can think of $\log I$ as being the volume of the narrowest prior that does not significantly impinge upon the posterior bulk, and Proposition 2 is one method for quantifying the qualitative statement "any reasonable prior" in Proposition 1. Finally, the interpretation of the Bayesian model dimensionality $\tilde{d}_D$ as the effective number of parameters is made clear in the Gaussian case, since $\tilde{d}_D = d$.

V. NUMERICAL EXAMPLES

We now apply our techniques to the cosmological dataset pairings of CMB data with baryon acoustic oscillations plus redshift-space distortions (BAO + RSD), galaxy clustering and weak lensing (3 × 2), and supernovae (SNe), respectively. This necessitates the numerical computation of evidences and Kullback-Leibler divergences via nested sampling. We find that BAO + RSD observations are fully consistent with CMB, 3 × 2 is in moderate tension, and SNe are in strong tension. Our results are summarized in Table II.

A. Nested sampling computation

To compute the log-evidence $\log \mathcal{Z}$ and the Kullback-Leibler divergence $\mathcal{D}$, we use the outputs of a nested sampling run produced by `CosmoChord` [59], a modified version of `CosmoMC` [60] using `PolyChord` [37,38] as a nested sampler. For a reliable computation of evidences and Kullback-Leibler divergences, we found it essential to use `PolyChord` rather than `MultiNest` [36], due to the high dimensionality of the space of cosmological and nuisance parameters.[3] Furthermore,

[2] The values $p = 0.05$ and 0.003 correspond to 2- and 3-σ Gaussian standard deviations.

[3] A little-known test of the reliability of the evidence estimates reported by `MultiNest` is to check whether two estimates of the evidence (the traditional and importance nested sampling estimation) agree to within the larger error bar. If they do not, then this indicates that the ellipsoidal approximation for generating new live points via rejection sampling is no longer valid. This may be fixed by decreasing the value of the efficiency parameter, with a consequent increase in run time.

PolyChord is able to dramatically speed up nested sampling in the context of cosmology by utilizing the fast-slow hierarchy between nuisance and cosmological parameters [61]. As a historical note, PolyChord was invented as an alternative to MultiNest in the context of the *Planck* Collaboration [62,63] to resolve precisely the issues described above.

The log-evidences and KL divergences are computed using the likelihood contours $\mathcal{L}_i$ of the discarded points from the trapezoidal rule,

$$
\begin{aligned}
&\mathcal{Z} \approx \sum_{i=1}^{N} \mathcal{L}_i \times \frac{1}{2}(X_{i-1} - X_{i+1}),\\
&\mathcal{D} \approx \sum_{i=1}^{N} \frac{\mathcal{L}_i}{\mathcal{Z}} \log \frac{\mathcal{L}_i}{\mathcal{Z}} \times \frac{1}{2}(X_{i-1} - X_{i+1}),\\
&\frac{\tilde{d}}{2} \approx \sum_{i=1}^{N} \frac{\mathcal{L}_i}{\mathcal{Z}} \left(\log \frac{\mathcal{L}_i}{\mathcal{Z}} - \mathcal{D}\right)^2 \times \frac{1}{2}(X_{i-1} - X_{i+1}),\\
&X_i = t_i X_{i-1}, \qquad X_0 = 1, \qquad X_{N+1} = 0,
\end{aligned} \tag{26}
$$

where X_i are the prior volumes of the N likelihood contours and the t_i are real random variables with probability distribution function

$$
P(t_i) = n_i t_i^{n_i - 1} [0 < t_i < 1]. \tag{27}
$$

Here n_i are the (usually constant) number of active live points enclosed by each likelihood contour $\mathcal{L}_i$. To account for the entire correlation between the random variables $\mathcal{D}$ and $\log \mathcal{Z}$, we simulate a set of weights $\{t_i\}$ using Eq. (27) and compute $\mathcal{Z}$, $\mathcal{D}$, and $\tilde{d}$ from Eq. (26) using the same weights. This process is repeated 1000 times to build up a set of samples from the $P(\mathcal{Z}, \mathcal{D}, \tilde{d})$ distribution. Examples of such distributions can be seen graphically in Fig. 5. The log-sum-exp trick must be carefully utilized to avoid overflow errors throughout these computations. For more detail, consult Skilling's original nested sampling paper [35]. Code to compute these quantities is now publicly available as part of the anesthetic pip-installable PYTHON package [64].

For our final runs, we used the CosmoChord settings $n_{\mathrm{live}} = 1000$, $n_{\mathrm{prior}} = 10000$, with all other settings left at their defaults for version 1.15. It is worth remarking that run-time is linear in the number of live points, and that PolyChord (in contrast to MultiNest) can function with extremely low numbers of live points. For low-resolution testing purposes, n_{live} can be set as low as 10, which proves invaluable in the initial exploratory stages of a project when publication-quality runs are not essential.

B. Cosmological likelihoods

For CMB observations we use the publicly available *Planck* 2015 TT + lowl + lowTEB likelihoods[4] [67]. For BAO + RSD observations we use the 6DF + MGS BOSS DR12 final consensus data [30,68,69]. For 3 × 2 data, we use the 1 year final DES dataset [1]. Finally, for SNe data we use a Gaussian likelihood on the Hubble parameter with mean and width indicated by the latest SH_0ES constraints [29].

We follow the notation and parametrization detailed in the respective likelihood papers, and we direct readers to those for further information on the meaning and notation of parameters.

C. Priors

To demonstrate the prior dependencies of $\log R$ and $\log S$, we choose three priors. The first is the default prior provided by CosmoMC. Note that this prior is not a trivial top-hat box prior, since CosmoMC places a model-dependent prior on the parameter space by eliminating regions that are unphysical. This nontrivial shape is shown in Fig. 6. We compare the default with two alternative prior choices; a "narrow" box centered on the posterior mean of *Planck*, with widths extending to 5σ of the *Planck* posterior, and a "medium" box designed to encompass the DES posterior while being a little narrower than the default. The narrow prior is arguably rather tight, but is chosen as the other extreme end of prior choice from the default prior to emphasize the prior dependency of the R statistic. It is worth noting that there is nothing particularly special about the choice of prior provided by the CosmoMC default, which could easily be narrowed or widened without a great deal of consensus objection.

D. Posteriors

The posterior on the Hubble parameter for SH_0ES and *Planck* produced by PolyChord is shown in Fig. 1. By eye it is clear from the individual posteriors that the inferences on the value of H_0 are incompatible and that the combined posterior cannot be trusted.

For BOSS and *Planck*, we show the marginalized posterior on the two parameters σ_8 and Ω_m in Fig. 2. Here there is significant overlap between the two-dimensional marginalized posteriors, and the combined posterior is valid. Note that they do not lie precisely on top on each other,

[4] At the time of writing this article, the *Planck* 2018 likelihoods [2] were not publicly available. The main difference between the *Planck* 2015 and 2018 parameters values is the constraints in the optical depth to reionization τ, that change from $\tau = 0.078 \pm 0.019$ [65] to $\tau = 0.055 \pm 0.009$ [66]. Because this paper is focused on the tension reported in [1], which uses the *Planck* 2015 likelihood, including their value of τ, we do not impose any priors on this parameter and simply use the *Planck* 2015 likelihood.

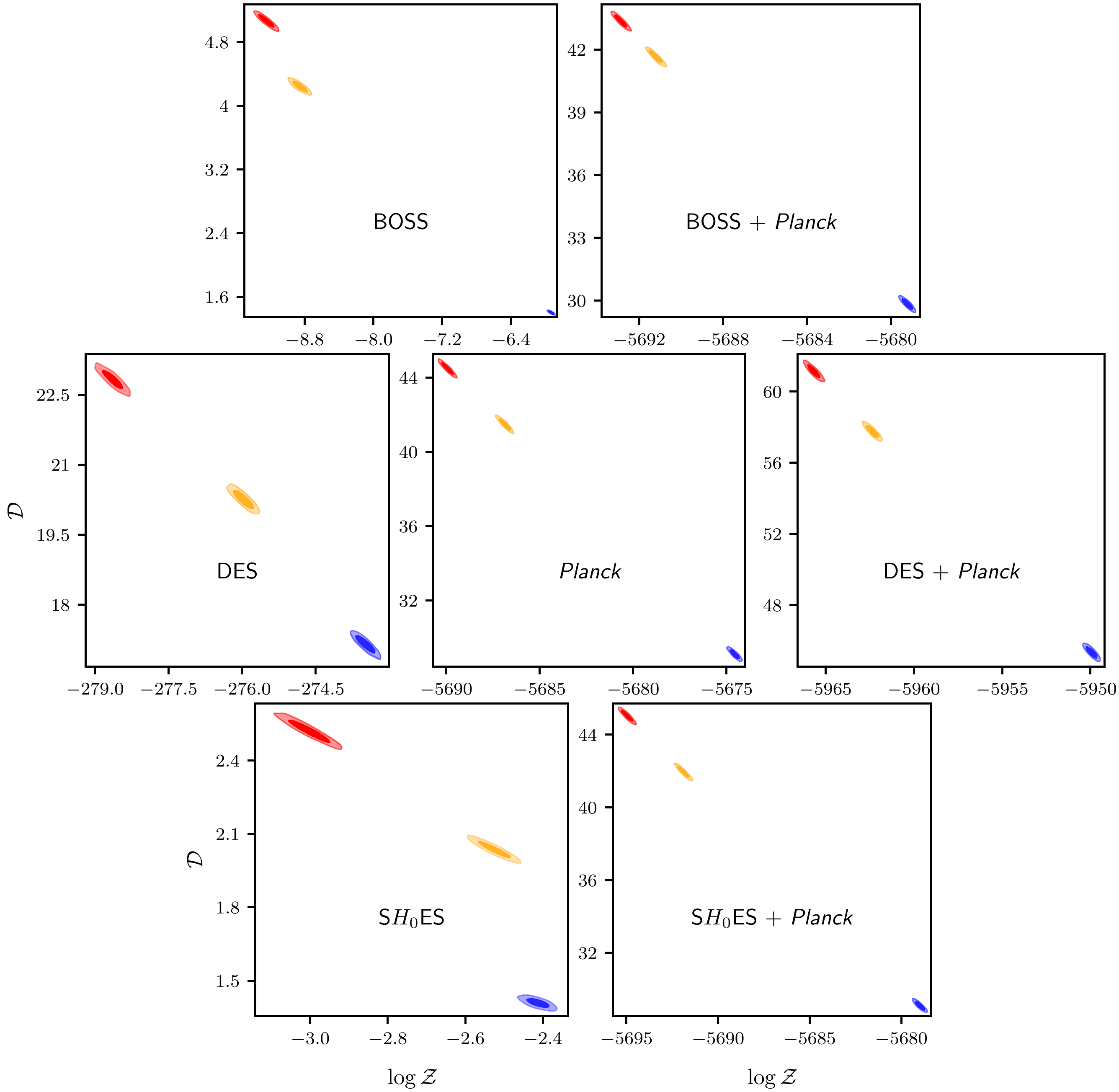
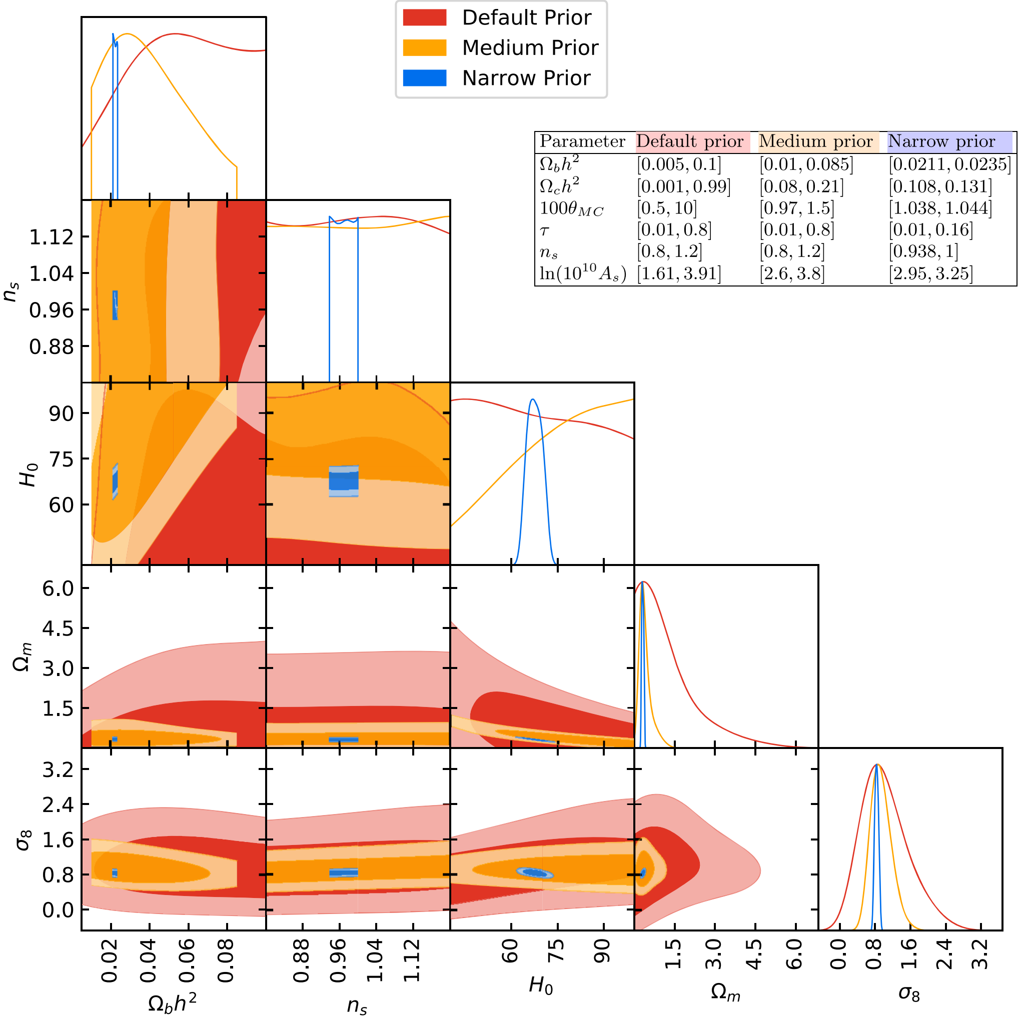


FIG. 5. Log-evidence $\log \mathcal{Z}$ and Kullback-Leibler divergence $\mathcal{D}$ calculations for all datasets and priors considered in this paper. The figures show the numerical values for the log-evidence and Kullback-Leibler divergence for the likelihoods described in Sec. V B under the default and narrow priors summarized in Fig. 6, with red representing results for the default priors, orange medium priors, and blue narrow priors. One can see that narrowing the prior increases the log-evidence and reduces the Kullback-Leibler divergence, but that $\log \mathcal{Z} + \mathcal{D}$ remains constant to within error. It should also be noted that the errors in estimating $\log \mathcal{Z}$ and $\mathcal{D}$ are strongly correlated. These errors arise from the uncertainty inherent in nested sampling's estimate of the volume compression of each likelihood contour and influence both quantities in the same manner. It should be noted that the parameter combination that we are most interested in estimating ($\log \mathcal{Z} + \mathcal{D}$) has the lowest error in its estimation.

which is in itself reassuring as otherwise the datasets would be suspiciously in agreement (and would usually indicate an overestimate of the errors or biases in the analysis).

For DES and *Planck*, we show the marginalized posterior for two parameters similar to those used in the BOSS case. In this case the situation is less clear, with a large proportion of the marginalized posterior bulk in disagreement, but with a small degree of overlap. If one looks at other parameter combinations, the tension becomes better or worse, and indeed it is possible to consider situations where there appears to be excellent overlap in every pair of parameters. However, it should be noted that since tension



Parameter	Default prior	Medium prior	Narrow prior
$\Omega_b h^2$	$[0.005, 0.1]$	$[0.01, 0.085]$	$[0.0211, 0.0235]$
$\Omega_c h^2$	$[0.001, 0.99]$	$[0.08, 0.21]$	$[0.108, 0.131]$
$100\theta_{MC}$	$[0.5, 10]$	$[0.97, 1.5]$	$[1.038, 1.044]$
τ	$[0.01, 0.8]$	$[0.01, 0.8]$	$[0.01, 0.16]$
n_s	$[0.8, 1.2]$	$[0.8, 1.2]$	$[0.938, 1]$
$\ln(10^{10} A_s)$	$[1.61, 3.91]$	$[2.6, 3.8]$	$[2.95, 3.25]$

FIG. 6. The three priors used throughout Sec. V. The priors provided to CosmoMC are shown in the upper right table. These construct an approximate box-prior on the six cosmological parameters. Two of the cosmological parameters are shown in the triangle plot, and indeed a box prior can be seen on the n_s parameter. The $\Omega_b h^2$ prior is not a simple top-hat prior on account of the fact that CosmoMC discards unphysical parameter combinations at the prior level. The remaining parameters are "derived parameters" and in general will not have box-priors, as can be seen in the $(H_0, \Omega_b h^2)$ plot.

is a parameter invariant notion, if one can resolve a significant tension in *any* parameter combination, then this indicates significant discordance that cannot be removed. A toy example of such a posterior is shown in Fig. 7. The advantage of building a general dimensional parametrization-independent prescription to quantify tension is that one can detect discrepancies even if none of the traditional parameters shows obvious tension in its marginalized plot.

E. Evidences and Kullback-Leibler divergences

The numerical evidences and Kullback-Leibler divergences computed from runs produced by PolyChord using the technique described in Sec. VA are reported in Fig. 5.

The first thing to note is that nested sampling does not produce an exact value for the evidence and KL divergence, but instead produces a correlated probability distribution. The correlation is negative, since the dominant error in the evidence estimate is associated with the cumulative Poisson noise in estimating the prior volume contraction at each iteration, and this error contributes equally to both the evidence and KL estimates. Note however that this is advantageous when we wish to compute the $\log S$ ratio, since the error is minimal for the parameters' contribution $\log \mathcal{Z} + \mathcal{D}$, as these prior volume errors cancel out to a large extent.

The second observation that should be made is that as we adjust the priors, the log-evidences increase as the normalization of the prior changes, the Kullback-Leibler divergences decrease since there is less compression between prior and

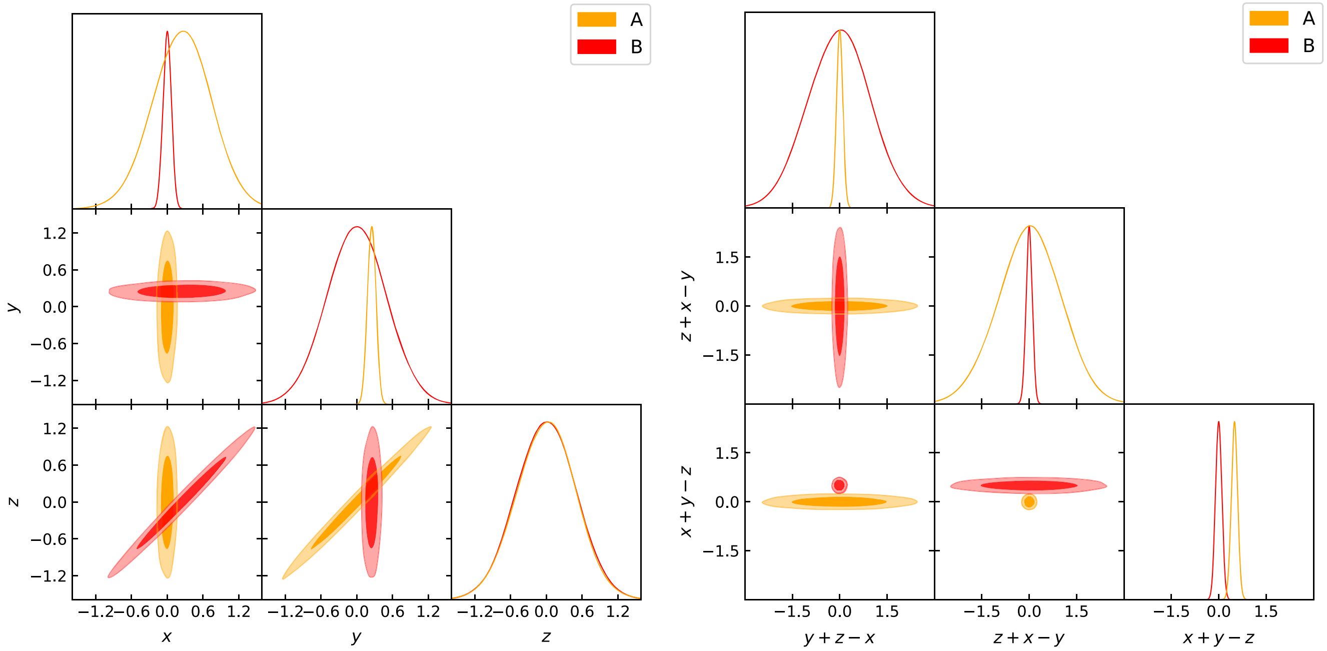


FIG. 7. Hidden tension in a multivariate Gaussian. By eye, the three-dimensional posterior on the left seems to be in reasonable agreement. The one- and two-dimensional marginalized posteriors are clearly consistent. However, upon making the linear transformation indicated on the right, it is clear that the posteriors are in fact disjoint. Such issues become harder in higher-dimensional posteriors and demonstrate the importance of parametrization-independent measures of tension such as those that we demonstrate here.

posterior, but the combination $\log\mathcal{Z}+\mathcal{D}$ remains approximately constant.

F. Bayesian model dimensionalities

The Bayesian model dimensionality for ΛCDM is detailed for each dataset and prior in Table I. As this is the first time such quantities have been utilized in a cosmological setting, they are worthy of some discussion.

First, the model dimensionality of the *Planck* dataset remains stable at $\tilde{d}_{Planck}\approx 15$ for all priors. While the *Planck* 2015 temperature likelihoods nominally have 21 parameters (6 cosmological and 15 nuisance), only a subset of the nuisance parameters are constrained by the data, as can be seen in Fig. 8. The fact that this dimensionality remains constant for all prior choices is due to the fact that the priors enclose the *Planck* posterior bulk in all three cases.

Second, in analogy with *Planck*, the DES Y1 data have a dimensionality of $\tilde{d}_{\mathrm{DES}}\approx 11$. As can be seen in Fig. 9, most of the 20 nuisance parameters and some of the 6 cosmological parameters are unconstrained. Quantifying the dimensionality in this case is made yet harder by the fact that unlike *Planck*, the DES Y1 survey best constrains a nontrivial combination of the sampled parameters (e.g., σ_8). It is for this reason that it is essential to have a parametrization-independent measure of the dimensionality of the constrained parameter space, such as that provided by the Bayesian model dimensionality. Additionally, unlike *Planck*, for DES there is a slight prior dependence of the dimensionality for the narrow priors. This can be understood by the fact that the narrow priors cut a little into the DES posterior, effectively rendering some parameters less constrained relative to the wider prior.

This prior dependency is also mirrored in the SH_0ES and BOSS datasets, although less trivially. For default and medium priors, the dimensionality $\tilde{d}_{SH_0\mathrm{ES}}=1$ reproduces the correct dimensionality given that the likelihood is only a Gaussian on the Hubble parameter. The fact that this rises to $\tilde{d}_{SH_0\mathrm{ES}}=2$ for the narrow prior is as a result of a nontrivial degeneracy that emerges for narrow priors in the combination of $(H_0,\Omega_c h^2)$, meaning that the tension constraint of SH_0ES generates an artificial constraint on $\Omega_c h^2$. The dimensionality of BOSS is yet more complicated,

TABLE I. Bayesian model dimensionality of ΛCDM for all datasets and priors considered in this paper, calculated using Eq. (3).

Prior	SH_0ES	BOSS	DES	*Planck*	SH_0ES + *Planck*	BOSS + *Planck*	DES + *Planck*
Default	0.93 ± 0.03	2.95 ± 0.07	14.01 ± 0.32	15.84 ± 0.38	15.98 ± 0.37	15.89 ± 0.36	25.89 ± 0.63
Medium	0.98 ± 0.03	3.79 ± 0.09	13.35 ± 0.31	15.89 ± 0.38	15.09 ± 0.35	16.38 ± 0.37	26.10 ± 0.65
Narrow	1.68 ± 0.03	1.40 ± 0.02	10.89 ± 0.24	15.96 ± 0.37	15.72 ± 0.37	15.69 ± 0.37	25.69 ± 0.62

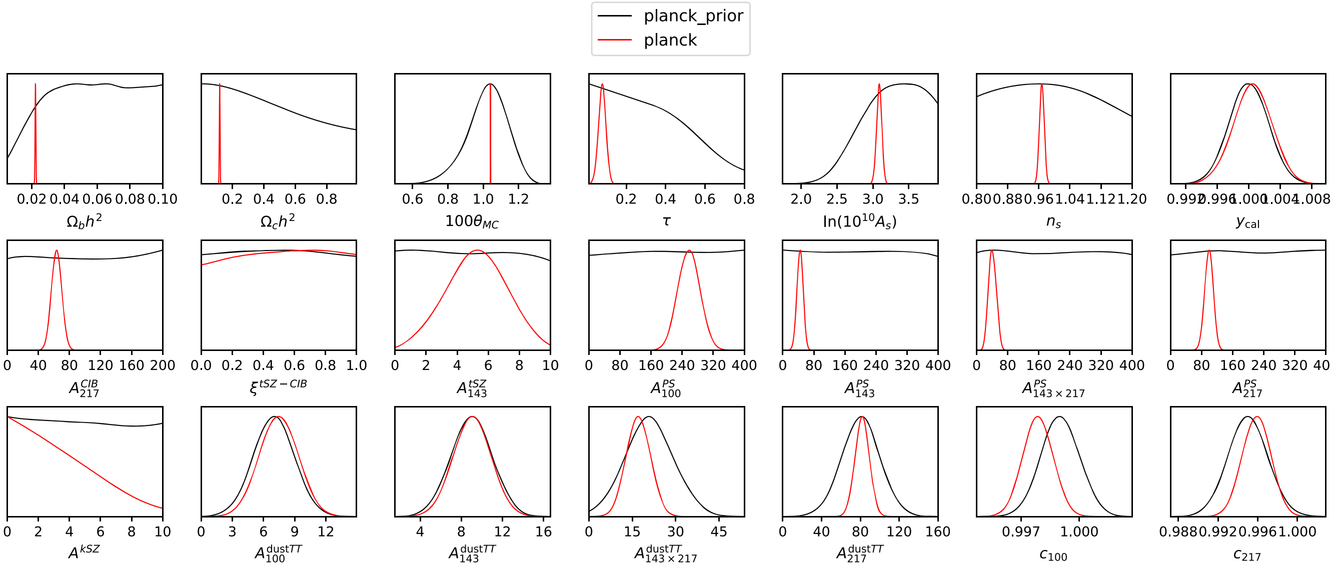


FIG. 8. One-dimensional marginalized default prior (black) and *Planck* posterior (red). The Bayesian model dimensionality of $\tilde{d}_{Planck} \approx 16$ is reflected by the fact that only a subset of the nuisance parameters is constrained by the data.

but consistent with the degeneracies between its likelihood and our prior choice.

Finally, the combined dimensionalities $\tilde{d} = \tilde{d}_A + \tilde{d}_B - \tilde{d}_{AB}$ are detailed in the penultimate column of Table II. These show the number of constrained parameters that the datasets have in common, and we can see that DES and *Planck* share between 1 and 2.5 constrained parameters depending on the prior chosen.

In conclusion, there is a rich structure in Bayesian model dimensionalities, and it is our hope that Bayesian model

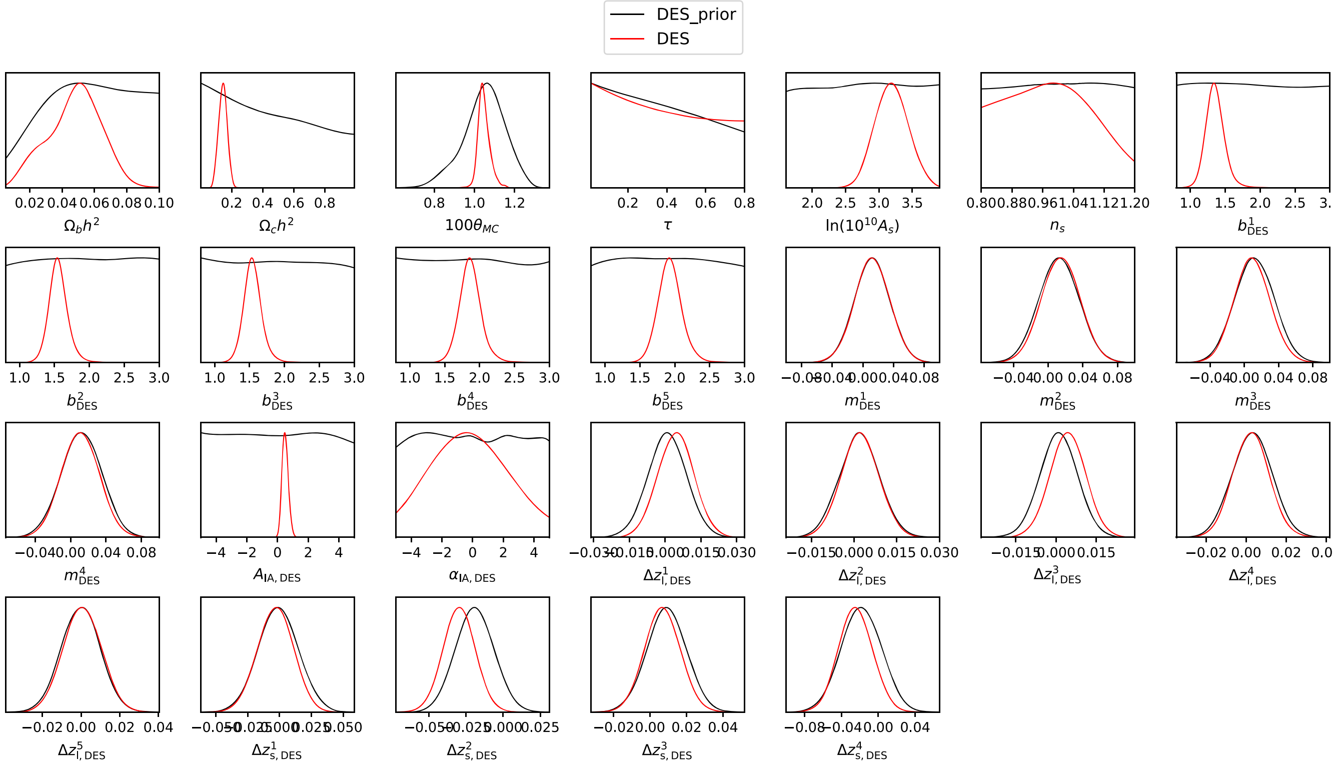


FIG. 9. One-dimensional marginalized default prior (black) and DES Y1 posterior (red). The Bayesian model dimensionality of $\tilde{d}_{\rm DES} \approx 14$ is reflected by the fact that only a combination of the cosmological parameters and a subset of the nuisance parameters are constrained by the data.

TABLE II. Comparison statistics. The values of $\log R$ and $\log I$ are computed via Eqs. (8) and (9), using the evidences and Kullback-Leibler divergences reported in Fig. 5. The suspiciousness statistic is simply $\log S = \log R - \log I$. $\tilde{d}$ is the Bayesian combined model dimensionality from Eq. (3), detailing the number of shared constrained parameters between the datasets, and p is the tension probability computed from Eq. (25). One can see explicitly the prior dependency of $\log R$ and $\log I$, and how this is removed/reduced in $\log S$ and p. In both the Bayes ratio $\log R$ via Proposition 1 and the tension probability p via Proposition 2, we find that the data show no tension between BOSS-*Planck*, moderate discordance between DES-*Planck*, and strong discordance between SH_0ES-*Planck*.

Dataset	Prior	$\log R$	$\log I$	$\log S$	$\tilde{d}$	$p(\%)$
BOSS-*Planck*	Default	6.30 ± 0.29	6.18 ± 0.29	0.11 ± 0.11	2.91 ± 0.51	42.66 ± 4.28
	Medium	4.51 ± 0.28	4.06 ± 0.28	0.46 ± 0.12	3.30 ± 0.55	55.12 ± 4.47
	Narrow	1.30 ± 0.23	0.69 ± 0.22	0.61 ± 0.12	1.67 ± 0.54	77.12 ± 14.10
DES-*Planck*	Default	2.88 ± 0.35	6.15 ± 0.34	-3.28 ± 0.16	3.97 ± 0.82	3.23 ± 1.00
	Medium	0.51 ± 0.34	4.00 ± 0.34	-3.49 ± 0.16	3.13 ± 0.81	2.04 ± 0.79
	Narrow	-1.88 ± 0.29	0.90 ± 0.29	-2.78 ± 0.16	1.15 ± 0.77	1.44 ± 0.91
SH_0ES-*Planck*	Default	-2.03 ± 0.29	1.96 ± 0.28	-3.99 ± 0.12	0.78 ± 0.52	0.25 ± 0.17
	Medium	-2.50 ± 0.28	1.56 ± 0.28	-4.06 ± 0.11	1.77 ± 0.51	0.56 ± 0.24
	Narrow	-2.00 ± 0.23	1.43 ± 0.23	-3.43 ± 0.12	1.92 ± 0.52	1.17 ± 0.45

dimensionality becomes more widely used in cosmological inference.

G. Ratios

We present our key numerical results for the Bayes ratio R and tension probabilities p in Table II.

First, we find that $\log R > 0$ for all priors considered for the BOSS + *Planck* combination, indicating that BAO + RSD datasets are consistent with CMB. More precisely, knowledge of the BOSS dataset boosts our probabilistic confidence in the CMB data by a factor ~500 for the default priors, or ~16 for the narrow priors. We find that $\log S$ is positive and around zero, with a corresponding tension probability $p \gg 5\%$. One should note that $\log S$ and p are not quite prior independent since the narrowed priors impinge somewhat on the posterior bulk of the BOSS dataset.

Second, for SH_0ES + *Planck*, we find that $\log R < 0$ for all priors, with our confidence in CMB data dropping in light of knowing the SNe data for all choices of prior, indicating inconsistency. This is also reflected in the tension probabilities, which indicate $p \sim 0.3\%$ probability of getting such inconsistency by chance.

Finally, for DES data, the default priors show $R \sim 20$, while the narrow priors give $R \sim 0.1$. Under Proposition 1, given that there are some priors which indicate a reduction in confidence in CMB data in light of 3×2 data, we should therefore not regard the datasets as being consistent. Considering the tension statistic, there is a roughly 2% probability of getting such an inconsistency by chance alone. We would therefore consider DES data to be in moderate tension with *Planck*.

H. Comparison with the DES analysis

It should be noted that our conclusion of moderate tension between DES and *Planck* is in contradiction to that presented in DES Y1. In DES Y1, they compute $R = 2.8$ and therefore conclude that there is no tension with CMB data. Hence the datasets are safe to use in conjunction with one another. Aside from a consideration of the precise meaning of R, which is the focus of the first three sections of this paper, there are several issues with their analysis. First, they do not report the errors arising from computing this quantity via nested sampling. Given that they in general use similar settings to ours, it is conceivable that their value of 2.8 is close to being consistent with $R = 1$. Second, they use `MultiNest` to compute this statistic, which renders the value of R that they compute unreliable. Third, they give no consideration to the prior dependency of the R statistic or to the fact that a small adjustment to their priors would have generated $R < 1$. While this dependency is undesirable for some analysts, it should be noted that consistent datasets (e.g., BOSS and *Planck*) in general should have $R \gg 1$, independent of prior choice.

VI. CONCLUSION

In this paper, we examined the Bayes ratio statistic used by DES to quantify the tension between potentially discordant datasets. We provided a novel interpretation of this statistic as a Bayesian quantification of our confidence in our ability to combine the datasets. It represents the factor by which our degree of belief in a dataset is strengthened in light of having incorporated the information provided by another dataset. We explain why this number is prior dependent, and under Proposition 1 say that if there is *any* reasonable prior choice which brings the factor to less than unity, then the datasets should be considered discordant.

For those who mislike the prior dependency of the Bayes ratio, we provide a method of calibrating the statistic using Kullback-Leibler divergences. Inspired by the Gaussian case, Proposition 2 provides a Bayesian tension probability, akin to the frequentist p-value statistic. As discussed in the Introduction, there are several alternative methods for quantifying tensions in the literature, but we claim that this is the only method that preserves all the desiderata of the Bayes ratio, while remaining insensitive to prior volume effects.

We applied these new techniques and interpretations to CMB data from *Planck* combined with the 3×2 data from DES, the BAO + RSD data from BOSS, or the SNe data from SH_0ES. Our technique confirms the consensus view that in comparison with the CMB, there is strong tension with SNe, moderate tension with 3×2, and no tension with BAO + RSD.

We believe that the R statistic is a valuable one for the community to use to compute tension between datasets, but that care must be taken with its interpretation. We hope that these considerations will be taken into account in future DES releases.

ACKNOWLEDGMENTS

This work was performed using resources provided by the Cambridge Service for Data Driven Discovery (CSD3) operated by the University of Cambridge Research Computing Service, provided by Dell EMC and Intel, using Tier-2 funding from the Engineering and Physical Sciences Research Council (Capital Grant No. EP/P020259/1) and Distributed research using advanced computing funding from the Science and Technology Facilities Council. This interpretation arose as a result of the Kavli institute for cosmology conference "Consistency of Cosmological Datasets: Evidence for New Physics?,"[5] from which we are indebted to many participants for extremely fruitful conversations. In particular, we are thankful for discussions with George Efstathiou, Steve Gratton, Ofer Lahav, Mike Hobson, and Anthony Lasenby. We would also like to thank Robert Schuhmann and Benjamin Joachimi for help with the Box-Cox transformations and members of the DES Collaboration for useful discussion, especially Scott Dodelson, Marco Raveri, Andresa Campos, and Vivian Miranda. We would also like to thank Daniel Mortlock, for comments on the first version of this paper which led to the use of a superior measure of model dimensionality, and to Andreu Font-Ribera for comments on the final version of the paper. W. J. H. thanks Gonville & Caius College for their continuing support via a Research Fellowship. P. L. thanks Science and technology facilities council and University college london for their support via a STFC Consolidated Grant.

[5] https://www.ast.cam.ac.uk/meetings/2018/consistency.cosmological.datasets.evidence.new.physics

[1] T. M. C. Abbott *et al.* (Dark Energy Survey Collaboration), Dark energy survey year 1 results: Cosmological constraints from galaxy clustering and weak lensing., Phys. Rev. D **98**, 043526 (2018).

[2] Planck Collaboration, Planck 2018 results. VI. Cosmological parameters, arXiv:1807.06209.

[3] S. Joudaki, C. Blake, C. Heymans, A. Choi, J. Harnois-Deraps, H. Hildebrandt, B. Joachimi, A. Johnson, A. Mead, D. Parkinson, M. Viola, and L. van Waerbeke, CFHTLenS revisited: Assessing concordance with Planck including astrophysical systematics, Mon. Not. R. Astron. Soc. **465**, 2033 (2017).

[4] F. Köhlinger, M. Viola, B. Joachimi, H. Hoekstra, E. van Uitert, H. Hildebrandt, A. Choi, T. Erben, C. Heymans, S. Joudaki, D. Klaes, K. Kuijken, J. Merten, L. Miller, P. Schneider, and E. A. Valentijn, KiDS-450: The tomographic weak lensing power spectrum and constraints on cosmological parameters, Mon. Not. R. Astron. Soc. **471**, 4412 (2017).

[5] H. Hildebrandt *et al.*, KiDS-450: Cosmological parameter constraints from tomographic weak gravitational lensing, Mon. Not. R. Astron. Soc. **465**, 1454 (2017).

[6] S. Joudaki, A. Mead, C. Blake, A. Choi, J. de Jong, T. Erben, I. F. Conti, R. Herbonnet, C. Heymans, H. Hildebrandt, H. Hoekstra, B. Joachimi, D. Klaes, F. Köhlinger, K. Kuijken, J. McFarland, L. Miller, P. Schneider, and M. Viola, KiDS-450: Testing extensions to the standard cosmological model, Mon. Not. R. Astron. Soc. **471**, 1259 (2017).

[7] G. Efstathiou and P. Lemos, Statistical inconsistencies in the KiDS-450 data set, Mon. Not. R. Astron. Soc. **476**, 151 (2018).

[8] M. A. Troxel *et al.*, Survey geometry and the internal consistency of recent cosmic shear measurements, Mon. Not. R. Astron. Soc. **479**, 4998 (2018).

[9] F. Köhlinger, B. Joachimi, M. Asgari, M. Viola, S. Joudaki, and T. Tröster, A Bayesian quantification of consistency in correlated datasets, Mon. Not. R. Astron. Soc. **484**, 3126 (2019).

[10] P. Marshall, N. Rajguru, and A. Slosar, Bayesian evidence as a tool for comparing datasets, Phys. Rev. D **73**, 067302 (2006).

[11] R. Trotta, Bayes in the sky: Bayesian inference and model selection in cosmology, Contemp. Phys. **49**, 71 (2008).

[12] L. Verde, P. Protopapas, and R. Jimenez, Planck and the local Universe: Quantifying the tension, Phys. Dark Universe **2**, 166 (2013).

[13] L. Verde, Precision cosmology, Accuracy cosmology and Statistical cosmology, in *Statistical Challenges in 21st Century Cosmology*, edited by A. Heavens, J.-L. Starck, and A. Krone-Martins, IAU Symposium Volume 306 (Cambridge University Press, England, 2014), pp. 223–234.

[14] M. Raveri, Are cosmological data sets consistent with each other within the Λ cold dark matter model, Phys. Rev. D **93**, 043522 (2016).

[15] S. Seehars, S. Grandis, A. Amara, and A. Refregier, Quantifying concordance in cosmology, Phys. Rev. D **93**, 103507 (2016).

[16] H. F. Inman and E. L. Bradley, Jr., The overlapping coefficient as a measure of agreement between probability distributions and point estimation of the overlap of two normal densities, Commun. Stat., Theory Methods **18**, 3851 (1989).

[17] R. A. Battye, T. Charnock, and A. Moss, Tension between the power spectrum of density perturbations measured on large and small scales, Phys. Rev. D **91**, 103508 (2015).

[18] S. Seehars, A. Amara, A. Refregier, A. Paranjape, and J. Akeret, Information gains from cosmic microwave background experiments, Phys. Rev. D **90**, 023533 (2014).

[19] A. Nicola, A. Amara, and A. Refregier, Consistency tests in cosmology using relative entropy, J. Cosmol. Astropart. Phys. 01 (2019) 011.

[20] M. Kunz, R. Trotta, and D. R. Parkinson, Measuring the effective complexity of cosmological models, Phys. Rev. D **74**, 023503 (2006).

[21] N. V. Karpenka, F. Feroz, and M. P. Hobson, Testing the mutual consistency of different supernovae surveys, Mon. Not. R. Astron. Soc. **449**, 2405 (2015).

[22] N. MacCrann, J. Zuntz, S. Bridle, B. Jain, and M. R. Becker, Cosmic discordance: Are Planck CMB and CFHTLenS weak lensing measurements out of tune, Mon. Not. R. Astron. Soc. **451**, 2877 (2015).

[23] S. Adhikari and D. Huterer, A new measure of tension between experiments, J. Cosmol. Astropart. Phys. 01 (2019) 036.

[24] S. M. Feeney, H. V. Peiris, A. R. Williamson, S. M. Nissanke, D. J. Mortlock, J. Alsing, and D. Scolnic, Prospects for Resolving the Hubble Constant Tension with Standard Sirens, Phys. Rev. Lett. **122**, 061105 (2019).

[25] M. Douspis, L. Salvati, and N. Aghanim, On the tension between large scale structures and cosmic microwave background, Proc. Sci. EDSU2018 (**2018**) 037.

[26] M. Raveri and W. Hu, Concordance and discordance in cosmology, Phys. Rev. D **99**, 043506 (2019).

[27] T. Charnock, R. A. Battye, and A. Moss, Planck data versus large scale structure: Methods to quantify discordance, Phys. Rev. D **95**, 123535 (2017).

[28] A. G. Riess, L. M. Macri, S. L. Hoffmann, D. Scolnic, S. Casertano, A. V. Filippenko, B. E. Tucker, M. J. Reid, D. O. Jones, J. M. Silverman, R. Chornock, P. Challis, W. Yuan, P. J. Brown, and R. J. Foley, A 2.4% determination of the local value of the Hubble constant, Astrophys. J. **826**, 56 (2016).

[29] A. G. Riess, S. Casertano, W. Yuan, L. Macri, J. Anderson, J. W. MacKenty, J. B. Bowers, K. I. Clubb, A. V. Filippenko, D. O. Jones, and B. E. Tucker, New parallaxes of galactic cepheids from spatially scanning the Hubble space telescope: Implications for the Hubble constant, Astrophys. J. **855**, 136 (2018).

[30] S. Alam *et al.*, The clustering of galaxies in the completed SDSS-III Baryon Oscillation Spectroscopic Survey: Cosmological analysis of the DR12 galaxy sample, Mon. Not. R. Astron. Soc. **470**, 2617 (2017).

[31] D. J. C. MacKay, *Information Theory, Inference & Learning Algorithms* (Cambridge University Press, New York, 2002).

[32] R. Trotta, Applications of Bayesian model selection to cosmological parameters, Mon. Not. R. Astron. Soc. **378**, 72 (2007).

[33] A. Heavens, Y. Fantaye, A. Mootoovaloo, H. Eggers, Z. Hosenie, S. Kroon, and E. Sellentin, Marginal likelihoods from Monte Carlo Markov chains, arXiv:1704.03472.

[34] A. Heavens, Y. Fantaye, E. Sellentin, H. Eggers, Z. Hosenie, S. Kroon, and A. Mootoovaloo, No Evidence for Extensions to the Standard Cosmological Model, Phys. Rev. Lett. **119**, 101301 (2017).

[35] J. Skilling, Nested sampling for general Bayesian computation, Bayesian Anal. **1**, 833 (2006).

[36] F. Feroz, M. P. Hobson, and M. Bridges, MULTINEST: An efficient and robust Bayesian inference tool for cosmology and particle physics, Mon. Not. R. Astron. Soc. **398**, 1601 (2009).

[37] W. J. Handley, M. P. Hobson, and A. N. Lasenby, POLYCHORD: Nested sampling for cosmology, Mon. Not. R. Astron. Soc. **450**, L61 (2015).

[38] W. J. Handley, M. P. Hobson, and A. N. Lasenby, POLYCHORD: Next-generation nested sampling. Mon. Not. R. Astron. Soc. **453**, 4384 (2015).

[39] B. J. Brewer, L. B. Pártay, and G. Csányi, Diffusive nested sampling, Stat. Comput. **21**, 649 (2011).

[40] B. J. Brewer and D. Foreman-Mackey, DNest4: Diffusive nested sampling in C++ and Python, arXiv:1606.03757.

[41] S. Kullback and R. A. Leibler, On information and sufficiency, Ann. Math. Stat. **22**, 79 (1951).

[42] A. Hosoya, T. Buchert, and M. Morita, Information Entropy in Cosmology, Phys. Rev. Lett. **92**, 141302 (2004).

[43] L. Verde, P. Protopapas, and R. Jimenez, Planck and the local Universe: Quantifying the tension, Phys. Dark Universe **2**, 166 (2013).

[44] S. Seehars, S. Grandis, A. Amara, and A. Refregier, Quantifying concordance in cosmology, Phys. Rev. D **93**, 103507 (2016).

[45] S. Grandis, S. Seehars, A. Refregier, A. Amara, and A. Nicola, Information gains from cosmological probes, J. Cosmol. Astropart. Phys. 05 (2016) 034.

[46] S. Hee, W. J. Handley, M. P. Hobson, and A. N. Lasenby, Bayesian model selection without evidences: Application to the dark energy equation-of-state, Mon. Not. R. Astron. Soc. **455**, 2461 (2016).

[47] S. Grandis, D. Rapetti, A. Saro, J. J. Mohr, and J. P. Dietrich, Quantifying tensions between CMB and distance data sets in models with free curvature or lensing amplitude, Mon. Not. R. Astron. Soc. **463**, 1416 (2016).

[48] G.-B. Zhao *et al.*, Dynamical dark energy in light of the latest observations, Nat. Astron. **1**, 627 (2017).

[49] A. Nicola, A. Amara, and A. Refregier, Integrated cosmological probes: Concordance quantified, J. Cosmol. Astropart. Phys. 10 (2017) 045.

[50] M. Raveri, M. Martinelli, G. Zhao, and Y. Wang, Information gain in cosmology: From the discovery of expansion to future surveys, arXiv:1606.06273.

[51] W. Handley and P. Lemos, Quantifying dimensionality: Bayesian cosmological model complexities, arXiv:1903.06682.

[52] C. E. Shannon and W. Weaver, The Mathematical Theory of Communication, Bell Syst. Tech. J. **27**, 379 (1948).

[53] D. J. Spiegelhalter, N. G. Best, B. P. Carlin, and A. Van Der Linde, Bayesian measures of model complexity and fit, J. R. Stat. Soc. Ser. B **64**, 583 (2002).

[54] S. M. Feeney, D. J. Mortlock, and N. Dalmasso, Clarifying the Hubble constant tension with a Bayesian hierarchical model of the local distance ladder, Mon. Not. R. Astron. Soc. **476**, 3861 (2018).

[55] S. M. Feeney, H. V. Peiris, A. R. Williamson, S. M. Nissanke, D. J. Mortlock, J. Alsing, and D. Scolnic, Prospects for Resolving the Hubble Constant Tension with Standard Sirens, Phys. Rev. Lett. **122**, 061105 (2019).

[56] G. E. P. Box and D. R. Cox, An analysis of transformations, J. R. Stat. Soc. Ser. B **26**, 211 (1964).

[57] B. Joachimi and A. N. Taylor, Forecasts of non-Gaussian parameter spaces using Box-Cox transformations, Mon. Not. R. Astron. Soc. **416**, 1010 (2011).

[58] R. L. Schuhmann, B. Joachimi, and H. V. Peiris, Gaussianization for fast and accurate inference from cosmological data, Mon. Not. R. Astron. Soc. **459**, 1916 (2016).

[59] W. J. Handley, Cosmochord 1.15, 2019, https://doi.org/10.5281/zenodo.255 2056.

[60] A. Lewis and S. Bridle, Cosmological parameters from CMB and other data: A Monte Carlo approach, Phys. Rev. D **66**, 103511 (2002).

[61] A. Lewis, Efficient sampling of fast and slow cosmological parameters, Phys. Rev. D **87**, 103529 (2013).

[62] Planck Collaboration, Planck 2015 results. XX. Constraints on inflation, Astron. Astrophys. **594**, A20 (2016).

[63] Planck Collaboration, Planck 2018 results. X. Constraints on inflation.

[64] W. Handley, Anesthetic: Nested sampling visualisation, J. Open Source Software **4**, 1414 (2019).

[65] P. A. R. Ade *et al.* (Planck Collaboration), Planck 2015 results. XIII. Cosmological parameters, Astron. Astrophys. **594**, A13 (2016).

[66] N. Aghanim *et al.* (Planck Collaboration), Planck intermediate results. XLVI. Reduction of large-scale systematic effects in HFI polarization maps and estimation of the reionization optical depth, Astron. Astrophys. **596**, A107 (2016).

[67] Planck Collaboration, Planck 2015 results. XI. CMB power spectra, likelihoods, and robustness of parameters, Astron. Astrophys. **594**, A11 (2016).

[68] F. Beutler, C. Blake, M. Colless, D. H. Jones, L. Staveley-Smith, L. Campbell, Q. Parker, W. Saunders, and F. Watson, The 6dF Galaxy Survey: Baryon acoustic oscillations and the local Hubble constant, Mon. Not. R. Astron. Soc. **416**, 3017 (2011).

[69] A. J. Ross, L. Samushia, C. Howlett, W. J. Percival, A. Burden, and M. Manera, The clustering of the SDSS DR7 main Galaxy sample—I. A 4 per cent distance measure at $z = 0.15$, Mon. Not. R. Astron. Soc. **449**, 835 (2015).

Correction: The error figures in the fifth column of Table II and the third line of Eq. (26) were presented inaccurately and have been fixed.