

## Concordance and discordance in cosmology

Marco Raveri and Wayne Hu

*Kavli Institute for Cosmological Physics, Department of Astronomy & Astrophysics,  
Enrico Fermi Institute, The University of Chicago, Chicago, Illinois 60637, USA*



(Received 26 June 2018; published 7 February 2019)

The success of present and future cosmological studies is tied to the ability to detect discrepancies in complex data sets within the framework of a cosmological model. Tensions caused by the presence of unknown systematic effects need to be isolated and corrected to increase the overall accuracy of parameter constraints, while discrepancies due to new physical phenomena need to be promptly identified. We develop a full set of estimators of internal and mutual agreement and disagreement, whose strengths complement each other. These estimators take into account the effect of prior information and compute the statistical significance of both tensions and confirmatory biases. The estimators that we present optimally weight all parameter space directions that are either fully constrained by the data or the prior, allowing for complete and fair degree of freedom counting. We apply them to a wide range of state-of-the-art cosmological probes and show that these estimators can be easily used, regardless of model and data complexity. We derive a series of results that show that discrepancies indeed arise within the standard  $\Lambda$ CDM model. Several of them exceed the probability threshold of 95% and deserve a dedicated effort to understand their origin.

DOI: [10.1103/PhysRevD.99.043506](https://doi.org/10.1103/PhysRevD.99.043506)

### I. INTRODUCTION

Since the discovery of cosmic acceleration [1,2], the description of our Universe based on general relativity with a cosmological constant ( $\Lambda$ ) and cold dark matter (CDM) has provided a successful working model for cosmology. The success of the  $\Lambda$ CDM model relies on its ability to describe a wide array of different cosmological observations ranging from the spectrum of fluctuations in the cosmic microwave background (CMB) to the clustering of galaxies and gravitational lensing observables.

Nevertheless, discrepancies exist between the determination of  $\Lambda$ CDM parameters by different data sets [3–10]. Local measurements of the Hubble constant differ from the value inferred from CMB observations of the Planck satellite [11] by more than  $3.4\sigma$  [12]. Measurements of the galaxy weak lensing correlation function also show disagreement with Planck CMB observations, involving parameters that determine the amplitude of the weak lensing signal, with a statistical significance that ranges from  $1.7\sigma$  to  $2.3\sigma$  for the Dark Energy Survey [13] and the Kilo Degree Survey [14], respectively. Furthermore, the internal consistency of the Planck CMB spectra in both temperature and polarization was analyzed in [11,15,16], revealing some discrepancies between the temperature spectrum and the reconstruction of its lensing signal.

The existence of such discrepancies is in large part due to the advent of precision cosmology and the low statistical errors of large surveys. When facing these and other discrepancies, we have to understand whether they can be attributed to residual systematic effects, an incorrect

modeling of the observables, or new physical phenomena. The next generation of cosmological probes, like Euclid [17], LSST [18], and CMB-S4 [19], are expected to further raise experimental sensitivity. While these may resolve current controversies, their increased modeling complexity will also make it difficult to inspect the data sets or the parameter posteriors to identify future discrepancies. This will make it increasingly difficult to understand whether data sets agree or not, and a failure to do so will compromise their scientific return.

In this paper we discuss several new concordance and discordance estimators (CDEs) that can be used to understand the internal consistency of a data set and its agreement with other cosmological probes. First, inspired by the Bayesian evidence as a measure of goodness of fit (GoF), we introduce a test that exploits the statistics of the likelihood at maximum posterior. Its dependence on the prior distribution allows a proper account of data and prior constrained directions when counting degrees of freedom (d.o.f.), while being significantly easier in practical applications with respect to the evidence. Second, we study the statistics of the evidence ratio test of data set compatibility in order to understand its biases. We show that in practical applications the bias toward agreement of the evidence ratio test is usually as large as its nominal value, making its interpretation on the Jeffreys' scale unreliable in determining agreement or disagreement. Third, we then define an estimator based on the ratio of likelihoods at maximum posterior, which maintains a close relationship with the evidence ratio in limiting cases but allows for an easy

assessment of statistical significance of the reported results. Finally, we consider estimators that quantify shifts in the parameters of two data sets, providing an implementation that works in an arbitrary number of dimensions and priors.

These tools can be straightforwardly applied, regardless of data and model complexity, and are based on a Gaussian linear model for the data likelihood and the posterior distribution that can be easily checked. In addition, they are sensitive to both tensions between data sets and the presence of confirmatory biases.

We illustrate their application on current data sets and analyze known discrepancies between state-of-the-art cosmological probes. More specifically, we investigate the internal consistency of CMB measurements, establishing a set of benchmark results for the next release of the Planck data and showing the following: The cross correlation of the CMB temperature and E-mode polarization is a bad fit to the  $\Lambda$ CDM model due to the likely presence of residual, frequency-dependent systematics or foregrounds; the discrepancy between the CMB spectra and lensing reconstruction is present for both the temperature spectrum and the E-mode polarization spectrum, at about the same statistical significance; the measurement of the large angular scale CMB fluctuations is in tension with the small scale temperature and E-mode spectra with a statistical significance at about the 95% confidence level.

We recover the known tensions between CMB and local measurements of  $H_0$  and weak lensing probes, showing that the latter are slightly larger than those reported in the literature, when considering the Canada-France-Hawaii Telescope Lensing Survey and the Kilo Degree Survey on large linear scales. This tension is also slightly larger than what we estimate by looking at the  $S_8 \equiv \sigma_8 \Omega_m^{0.5}$  parameter since this is not one of the principal components of both parameter covariances while our estimator optimally weights all parameter space directions. We find that the CMB is in tension with probes of the clustering of galaxies, which can be attributed to the SDSS LRG DR4 survey being too good of an internal fit to different values of cosmological parameters.

This paper is organized as follows. In Sec. II we discuss the technical aspects of several CDEs and their application to data. In particular, the first part of the section contains a review of the relevant statistical tools, while the second part contains the discussion of different estimators and contains most of the theoretical results that are new to this paper. In Sec. III we detail the cosmological model and data sets, and we apply the CDEs to them in Sec. IV. We summarize our conclusions in Sec. V. In Appendixes A–G, we derive the statistical properties of the CDEs and give details on their implementation.

## II. CONCORDANCE AND DISCORDANCE ESTIMATORS

In this section we introduce and review the concordance and discordance estimators (CDEs) that we later apply to cosmological data sets. This section is organized as

follows: In Sec. II A we discuss the requirements for and limitations of the probabilistic interpretation of CDEs; in Sec. II B we define the notation of subsequent sections; in Sec. II C we review the Gaussian linear model; in Sec. II D we apply it to quantify internal consistency of data sets; in Secs. II E and II F we use it to discuss pairwise CDEs.

### A. CDE measures

We loosely refer to a CDE as a statement about a data set  $D$  or a collection of data sets  $D = D_1 \cup \dots \cup D_n$ , within a given model  $\mathcal{M}$ , that quantifies agreement or disagreement between the data and the model. In the case of a single data set, these statements should quantify internal consistency (or self-consistency), and in the case of multiple data sets, mutual consistency.

Since we regard data as random, CDEs are random variables as well, distributed over the space of data  $D$ . When defining a CDE,

- (i) we must be able to compute the distribution of the CDE over the space of data realizations  $D$ , where  $D$  can be a single data set or the union of multiple data sets  $D = D_1 \cup \dots \cup D_n$ , depending on the definition of the CDE;
- (ii) we report the probability  $P(\text{CDE} > \text{CDE}_{\text{obs}})$  so that low (high) probabilities identify disagreement (agreement) based on the observed value  $\text{CDE}_{\text{obs}}$ .

The distribution over data space is usually high dimensional and, though it is, in principle, possible to understand it with Monte Carlo techniques, doing so is typically extremely computationally intensive. For this reason we apply, and test the validity of, Gaussian approximations to analytically work out the distribution of these estimators.

Once probabilities over data space are computed, if  $P(\text{CDE} > \text{CDE}_{\text{obs}})$  is too low, then this could point toward the presence of tensions, and if it is too high, it could point to the presence of confirmatory biases. Note that confirmation bias in this sense does not necessarily mean a voluntary human action directed at confirming prior beliefs, but it includes any subtle assumption that can bias results toward accepting a fiducial model. These could include, as an example, overestimating data covariances, assuming a fiducial cosmology in the data reduction (e.g., converting angles and redshifts to distances), calibrating numerical algorithms around a given cosmology, and others. As experimental precision increases, even subtle biases, if not properly counterbalanced, would damage the scientific return of the affected experiment.

Notice that the key point that allows us to define CDEs in a frequentist-like fashion, from a Bayesian perspective, is that the problem of data set compatibility is not a model selection problem. We ask the statistical question of whether two data sets agree or not, within a model, at a fixed model while accounting for the fact that the parameters of the model are unknown.

Many of the commonly employed estimators are presented in the literature without computing their statistics but rather interpreting their observed value as an indication of agreement or disagreement. This does not take into account that CDEs can be biased,  $\langle \text{CDE} \rangle_D \neq 0$ , toward agreement or disagreement. Knowing the distribution over the data space prevents us from being tricked into thinking that there is agreement or disagreement when it is not the case.

We next warn the reader about caveats in interpreting CDEs. CDEs can indicate agreement or disagreement but do not reveal the cause. In particular, in the case of tensions these could result from a problem with the data and unknown systematic effects, a problem with the predictions that stems from an incomplete modeling of the observable, or a more fundamental problem with the model. CDEs do not discriminate one from the other but rather quantify the statistical level of unknowns in the given theoretical and experimental situation.

Another limitation, common to all methods to quantify agreement or disagreement within a model, is that they do not quantify the need for model extensions. It is always possible to relax tension with the addition of extra parameters, which could describe systematic effects or new physical aspects of the model, but doing so carries the danger of overfitting. The methods that we describe in this paper should not be used to justify model extensions directly but rather motivate further studies with the appropriate statistical tools, like Bayesian model selection.

Just as no one CDE gives the probability of the model given the data, not all CDEs result in the same assessment of statistical significance for concordance or discordance. There are multiple ways in which the model can be in tension or agreement with the data. In fact, if the CDE is selected after looking at the data, one can always find some aspect of the data that deviates from the model just by chance fluctuations. It is therefore advantageous to select, before looking at the data, multiple CDEs that correspond to meaningful quantities whose values we would want to be probable given a model.

Finally, when looking at these multiple CDE results, we should not naively combine them into a global probability. To assess this, we would need to know the joint distribution of the multiple tests. For example, the CDEs might be correlated, making multiple concordance or discordance results redundant. Even if the CDEs are uncorrelated, we would expect that out of many tests, one might fail due to chance fluctuations. We instead use the CDEs to flag individual aspects of the data and model for further study and multiple CDEs to assess the robustness of conclusions from any single CDE.

## B. Basic definitions

We now lay out some definitions to clarify the notation of the subsequent sections.

We commonly employ the multivariate Gaussian distribution, over the space of  $\theta$ , which we denote as

$$\mathcal{N}_N(\theta; \bar{\theta}, \mathcal{C}) = (2\pi)^{-N/2} |\mathcal{C}|^{-1/2} e^{-\frac{1}{2}(\theta - \bar{\theta})^T \mathcal{C}^{-1} (\theta - \bar{\theta})}, \quad (1)$$

where  $\det(\cdot) \equiv |\cdot|$ ,  $N$  corresponds to the number of dimensions,  $\bar{\theta}$  is the mean of the distribution, and  $\mathcal{C}$  is the covariance. Generally, throughout this paper we denote parameter covariances as  $\mathcal{C}$  and data covariances as  $\Sigma$ . Given a model  $\mathcal{M}$  and data  $D$ , the probability of the  $N$  model parameters  $\theta$  after the data  $D$  is given by

$$P(\theta|D, \mathcal{M}) = \frac{P(D|\theta, \mathcal{M})P(\theta|\mathcal{M})}{P(D|\mathcal{M})} = \frac{\mathcal{L}(\theta)\Pi(\theta)}{\mathcal{E}}, \quad (2)$$

which we call the parameter's posterior and where  $P(\theta|\mathcal{M}) \equiv \Pi(\theta)$  is the prior probability density function (pdf), normalized to unity over parameter space,  $P(D|\theta, \mathcal{M}) \equiv \mathcal{L}(\theta)$  is the likelihood, and  $P(D|\mathcal{M}) \equiv \mathcal{E}$  is the evidence. Usually, the normalization of the posterior is not computed, and one has to work with the following:

$$\mathcal{P}(\theta) \equiv \mathcal{L}(\theta)\Pi(\theta), \quad (3)$$

which we call the un-normalized posterior. The normalization factor of the un-normalized posterior is the evidence

$$\mathcal{E} \equiv P(D|\mathcal{M}) = \int \mathcal{P}(\theta) d\theta = \int \mathcal{L}(\theta)\Pi(\theta) d\theta. \quad (4)$$

Notice that, within a given model  $\mathcal{M}$ , the evidence defines the prior probability for observing data  $D$ . This is especially relevant in cosmology where we do not have the possibility of having truly different data realizations. Thus, we have to fix the model that then predicts its distribution of data realizations that would be drawn from its evidence. In this sense we can define functions of the data  $D$  and, within a model  $\mathcal{M}$ , we can compute their distributions and, for example, their average over data realizations as

$$\langle f(D) \rangle_D = \int f(D) P(D|\mathcal{M}) dD, \quad (5)$$

where the measure over data space is the evidence of the model. This aspect is key in the definition of CDEs as frequentist-like statements, in a Bayesian context. Since the problem of data set compatibility is posed at a fixed model (with unknown parameters), the evidence gives the probability distribution of the data and allows us to define statistics over data draws and study their distribution. As such, the evidence is always involved in frequentist-like tests.

As for the prior distribution, we use four different functional forms, depending on the application of interest:

- (i) *Flat prior*: Given by a ‘‘top hat’’ function  $\Pi(\theta) = 1/V_\Pi$  when all the  $n$ th parameter components are included between  $\theta_{\max}^{(n)}$  and  $\theta_{\min}^{(n)}$ . The prior volume is  $V_\Pi = \prod_{n=1}^N [\theta_{\max}^{(n)} - \theta_{\min}^{(n)}]$ .

- (ii) *Uninformative flat prior*: A flat prior where the range is chosen so that the prior is uninformative with respect to the data, i.e.,  $(\theta_{\max}^{(n)} - \theta_{\min}^{(n)})^2 \gg C_{\theta^{(n)}\theta^{(n)}}$ .
- (iii) *Gaussian prior*: Given by a multivariate Gaussian  $\Pi(\theta) = \mathcal{N}_N(\theta; \theta_{\Pi}, \mathcal{C}_{\Pi})$  with mean  $\theta_{\Pi}$  and covariance  $\mathcal{C}_{\Pi}$ . These priors are normalized to unity, and their maximum value at  $\theta_{\Pi}$  is  $(2\pi)^{-N/2} |\mathcal{C}_{\Pi}|^{-1/2}$ .
- (iv) *Delta prior*: A Gaussian prior in the limit  $\mathcal{C}_{\Pi} \rightarrow 0$  or  $\Pi(\theta) = \delta(\theta - \theta_{\Pi})$ , a rather stubborn choice which we use for pedagogical purposes.

Uninformative flat prior results can be related to the Gaussian ones by appropriately setting the center parameter as well as in the limit  $\mathcal{C}_{\Pi} \gg \mathcal{C}$ . Moreover, the Gaussian prior volume, which is formally undefined, can be taken as  $V_{\Pi} = (2\pi)^{N/2} |\mathcal{C}_{\Pi}|^{1/2}$  to retain the same normalization as  $\Pi(\theta) = 1/V_{\Pi}$  at the peak.

Flat priors are the ones that are used in most practical applications but, to the best of our knowledge, it is not possible to derive simple analytic results, in general. For this prior choice some directions in parameter space might be constrained by the data. When this is the case, they become uninformative flat priors where analytic results can be derived. When they are much more informative than the data, on the other hand, their effect is closer to the delta prior case. For the intermediate, partially informative case, we approximate flat priors with Gaussian priors, taking into account that, when the prior needs to be directly evaluated, it would give  $\Pi(\theta) = 1/V_{\Pi}$ . This allows us to appreciate the two most important features of flat priors: the shift between the maximum likelihood and the maximum likelihood as constrained by the prior; the information content of the prior, as modeled by the covariance of the bounded flat distribution  $C = (\theta_{\max}^{(n)} - \theta_{\min}^{(n)})^2/12$ . For practical applications we discuss the Gaussian approximation of the Markov chain Monte Carlo (MCMC) sampled posterior in Appendix E.

### C. Gaussian linear model

To understand the statistics of the CDEs discussed in this section, we need to make some simplifying assumptions. We assume that the likelihood of the data is Gaussian distributed in data space, and we expand our model predictions to linear order in their parameter dependence. This results in the Gaussian linear model (GLM), which was discussed in [7,20], and whose treatment we mostly follow. The assumptions of the GLM are somewhat restrictive but find many applications in cosmology. Most of the available data likelihoods are Gaussian distributions in the data, and many probes, notably the CMB, constrain the parameters of the  $\Lambda$ CDM model sufficiently well that the linear approximation is valid. Let us assume that we have  $d$  Gaussian distributed data points  $x$ , with mean  $m$  and covariance  $\Sigma$ . Their likelihood is a Gaussian distribution in data space:

$$\mathcal{L} = \mathcal{N}_d(x; m, \Sigma). \quad (6)$$

Our model  $\mathcal{M}$  would predict  $m$  as a function of  $N$  parameters  $\theta$ . We thus expand in series the prediction around a given parameter value  $\hat{\theta}$ :

$$\begin{aligned} m(\theta) &= m(\hat{\theta}) + \left. \frac{\partial m}{\partial \theta} \right|_{\hat{\theta}} (\theta - \hat{\theta}) + \dots \\ &\equiv \hat{m} + M(\theta - \hat{\theta}) + \dots, \end{aligned} \quad (7)$$

where we defined our central value for the expansion  $\hat{\theta}$ , the corresponding data prediction  $\hat{m} = m(\hat{\theta})$ , and the Jacobian of the transformation between data and parameter space  $M$ . It is worth commenting on the properties of the Jacobian. Since the dimension of the parameter space and data space is usually different, the Jacobian is not square and thus not invertible. We can, however, define

$$\tilde{M} = (M^T \Sigma^{-1} M)^{-1} M^T \Sigma^{-1}, \quad (8)$$

which has the following properties:

- (i)  $\tilde{M}^T = \Sigma^{-1} M (M^T \Sigma^{-1} M)^{-1}$  given that  $M^T \Sigma^{-1} M$  is symmetric;
- (ii)  $\tilde{M} M = M^T \tilde{M}^T = \mathbb{I}_{N \times N}$ .

The two matrices  $M$  and  $\tilde{M}$  can be used to define a projector on the  $m(\hat{\theta})$  tangent space:

$$\mathbb{P} = M \tilde{M}, \quad (9)$$

with the following properties:

- (i)  $\mathbb{P}^2 = \mathbb{P}$ , i.e.,  $\mathbb{P}$  is a projector and its complement is  $\mathbb{I} - \mathbb{P}$ ;
- (ii)  $\mathbb{P} M \theta = M \theta$ , which leaves the tangent space of  $m(\hat{\theta})$  invariant;
- (iii)  $(\mathbb{I} - \mathbb{P}^T) \Sigma^{-1} \mathbb{P} = 0$  so that the complementary projectors are orthogonal in the metric defined by  $\Sigma^{-1}$ .

By decomposing the data residual  $(x - m)$  in a component that is projected along the model,  $\mathbb{P}(x - m)$ , and a component that is orthogonal to the model,  $(\mathbb{I} - \mathbb{P})(x - m)$ , we can now recast Eq. (6) into

$$\mathcal{L} = \mathcal{L}_{\max} \exp \left[ -\frac{1}{2} (\theta - \theta_{\text{ML}})^T \mathcal{C}^{-1} (\theta - \theta_{\text{ML}}) \right], \quad (10)$$

with the maximum likelihood

$$\mathcal{L}_{\max} = \frac{\exp \left[ -\frac{1}{2} (x - \hat{m})^T (\mathbb{I} - \mathbb{P})^T \Sigma^{-1} (\mathbb{I} - \mathbb{P})(x - \hat{m}) \right]}{(2\pi)^{d/2} |\Sigma|^{1/2}}, \quad (11)$$

and maximum likelihood parameters and covariance

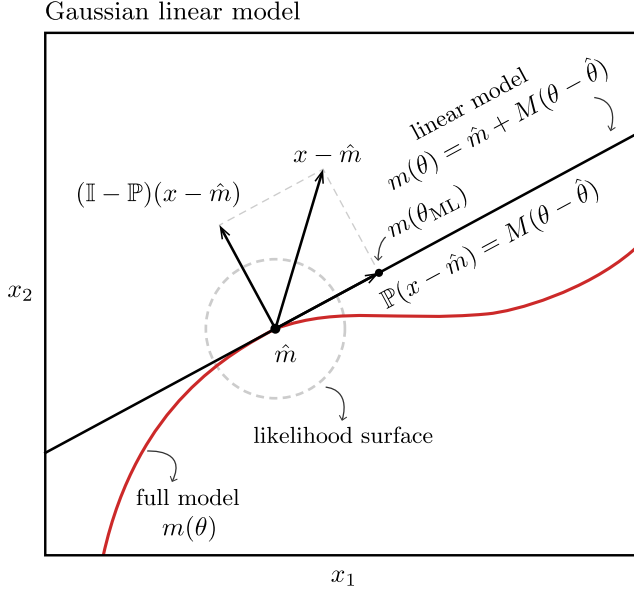


FIG. 1. Geometrical interpretation of the Gaussian linear model. Here,  $(x_1, x_2)$  represents data space and  $m(\theta)$  a one-dimensional model, i.e., a curve in the  $(x_1, x_2)$  space. The figure also shows the linearization of the model and how to decompose differences between a data realization and the model (at fixed parameters) in the direction that is parallel and orthogonal to the model. Note that  $m(\theta_{\text{ML}})$  shows the model corresponding to the best-fit parameter values for the given data realization. The dashed line shows a constant likelihood surface, where we assume, for simplicity, that data covariance is proportional to the identity matrix.

$$\begin{aligned}\theta_{\text{ML}} &= \hat{\theta} + \tilde{M}(x - \hat{m}), \\ \mathcal{C} &= (M^T \Sigma^{-1} M)^{-1}.\end{aligned}\quad (12)$$

Notice that the maximum likelihood parameter value depends on the data realization  $x$ . Figure 1 summarizes the geometrical meaning of the GLM in a two-dimensional data space with a one-parameter model.

Having computed the likelihood, we can get the posterior of the data, for the GLM, with different prior choices. In the case of Gaussian priors the posterior is still Gaussian  $P(\theta|D, \mathcal{M}) = \mathcal{N}_N(\theta; \theta_p, \mathcal{C}_p)$  with

$$\begin{aligned}\mathcal{C}_p &= (\mathcal{C}_\Pi^{-1} + \mathcal{C}^{-1})^{-1} = (\mathcal{C}_\Pi^{-1} + M^T \Sigma^{-1} M)^{-1}, \\ \theta_p &= \mathcal{C}_p [\mathcal{C}_\Pi^{-1} \theta_\Pi + \mathcal{C}^{-1} \theta_{\text{ML}}] \\ &= \mathcal{C}_p [\mathcal{C}_\Pi^{-1} \theta_\Pi + M^T \Sigma^{-1} (x - \hat{m} + M\hat{\theta})].\end{aligned}\quad (13)$$

If we consider an uninformative flat prior, then the posterior is Gaussian  $P(\theta|D, \mathcal{M}) = \mathcal{N}_N(\theta; \theta_{\text{ML}}, \mathcal{C})$ . In the case of a delta prior, instead the posterior is a delta function around the chosen parameter value  $P(\theta|D, \mathcal{M}) = \delta(\theta - \hat{\theta})$ .

The evidence can now be computed in a given model and for a given prior choice. In parameter space and for Gaussian priors, the evidence is given by

$$\begin{aligned}\ln \mathcal{E} &= \ln \mathcal{L}_{\text{max}} + \frac{1}{2} \ln \frac{|\mathcal{C}|}{|\mathcal{C} + \mathcal{C}_\Pi|} \\ &\quad - \frac{1}{2} (\theta_{\text{ML}} - \theta_\Pi)^T (\mathcal{C} + \mathcal{C}_\Pi)^{-1} (\theta_{\text{ML}} - \theta_\Pi),\end{aligned}\quad (14)$$

where the first line contains the familiar Occam's razor term and the second line a penalty for cases where the prior center is not the maximum of the likelihood. We can equivalently express this in terms of the likelihood evaluated at the maximum posterior probability point  $\theta = \theta_p$ ,

$$\ln \mathcal{E} = \ln \mathcal{L}(\theta_p) + \frac{1}{2} \ln |\mathcal{C}_p| + \frac{N}{2} \ln(2\pi) + \ln \Pi(\theta_p). \quad (15)$$

This form also highlights the limit which coincides with the case of uninformative flat priors where  $\theta_p = \theta_{\text{ML}}$ ,  $\mathcal{C}_p = \mathcal{C}$  and  $\Pi(\theta_p) = 1/V_\Pi$ :

$$\ln \mathcal{E} = \ln \mathcal{L}_{\text{max}} + \frac{1}{2} \ln |\mathcal{C}| + \frac{N}{2} \ln(2\pi) - \ln V_\Pi. \quad (16)$$

Likewise, it highlights the delta prior limit,  $\theta_p = \theta_\Pi$ , where  $\ln \mathcal{E} = \ln \mathcal{L}(\theta_\Pi)$ , which is the limiting case of Gaussian priors as the prior covariance goes to zero.

We can now write these results in data space by means of the GLM. Figure 2 shows the graphical interpretation of the GLM evidence, for different prior choices, in our two-dimensional example. In the Gaussian prior case, shown in panel (a) of Fig. 2, the evidence is a Gaussian distribution in data space  $\mathcal{E} = P(D|\mathcal{M}) = \mathcal{N}_d(x; m_\Pi, \Sigma_0)$  with

$$\begin{aligned}m_\Pi &= m(\theta_\Pi), \\ \Sigma_0 &= \Sigma + M \mathcal{C}_\Pi M^T.\end{aligned}\quad (17)$$

In the uninformative flat prior case, the evidence is a Gaussian distribution orthogonal to the projector

$$\mathcal{E} \propto e^{-\frac{1}{2}(x - \hat{m})^T (\mathbb{I} - \mathbb{P})^T \Sigma^{-1} (\mathbb{I} - \mathbb{P})(x - \hat{m})}, \quad (18)$$

with a normalization factor such that the distribution integrates to unity over the data space. Notice that we have not defined this with the corresponding normal distribution since the projection operation is not invertible so that the determinant of  $(\mathbb{I} - \mathbb{P})^T \Sigma^{-1} (\mathbb{I} - \mathbb{P})$  is singular. If we consider delta priors, as in panel (c) of Fig. 2, the evidence is still Gaussian in data space  $\mathcal{E} = \mathcal{N}_d(x; \hat{m}, \Sigma)$ .

When studying the distribution of different quantities over data realizations, the evidence provides the distribution of the data. It is then a noteworthy result that, within a given model  $\mathcal{M}$ , regardless of the parameters, for all the considered prior choices, the data realizations provide

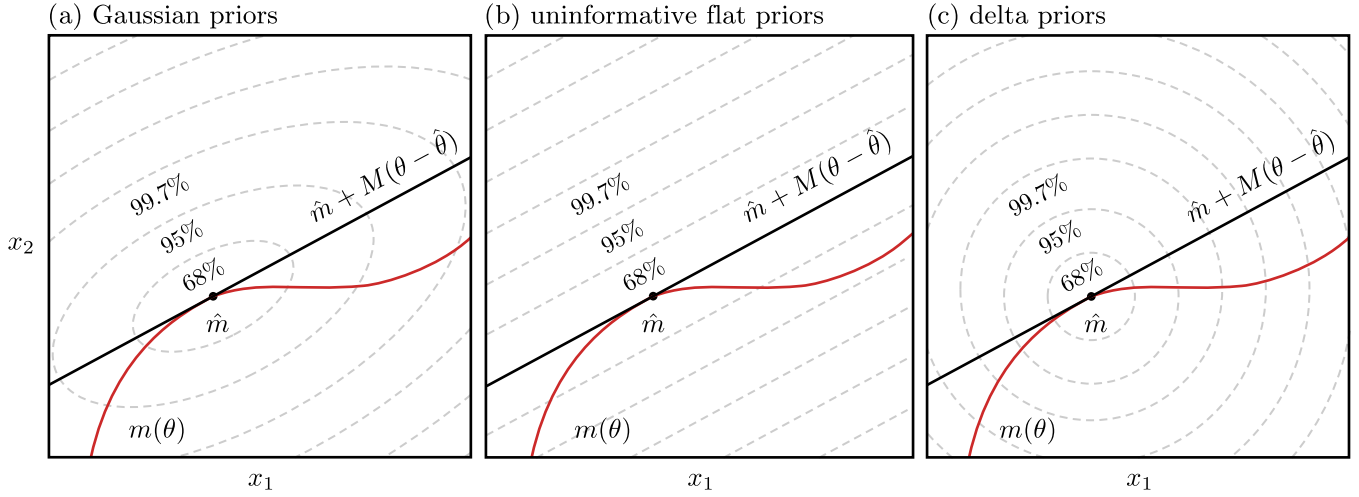


FIG. 2. Geometrical interpretation of the GLM evidence. In all panels  $(x_1, x_2)$  represents data space and  $m(\theta)$  a one-dimensional model, i.e., a curve in the  $(x_1, x_2)$  space. The figure also shows the linearization of the model. The dashed lines correspond to the evidence contours, for different prior choices, and different confidence levels. The contours show that, when drawing data realizations from the evidence, these will be 68% of the time inside the 68% contour, 95% of the time inside the 95% contour, and so on. As in the previous figure we assumed, for simplicity, that data covariance is proportional to the identity matrix. In the Gaussian prior case we also assumed that  $m_{\Pi} = \hat{m}$ .

evidence that is a Gaussian distribution, with different mean and covariance.

The last aspect of the GLM that we discuss is the dependence of the results on the expansion point  $\hat{\theta}$ . When using the GLM to compute the distribution of different estimators, the results do not depend on the arbitrary expansion point unless we purposely make that point special by prior choice.

#### D. Goodness-of-fit-type tests

The first application of the GLM consists in defining GoF type measures. These are the only CDEs that we consider that measure the internal consistency of a single data set, within a model.

We first define the usual maximum likelihood GoF measure as the quadratic form:

$$Q_{\text{ML}} = (x - \hat{m})^T (\mathbb{I} - \mathbb{P})^T \Sigma^{-1} (\mathbb{I} - \mathbb{P})(x - \hat{m}). \quad (19)$$

Note that

$$Q_{\text{ML}} = -2 \ln \mathcal{L}_{\text{max}} + 2 \langle \ln \mathcal{L}_{\text{max}} \rangle_D + \langle Q_{\text{ML}} \rangle_D, \quad (20)$$

where the average is over data realizations; thus, up to these constant offsets,  $Q_{\text{ML}}$  is equivalent to  $-2 \ln \mathcal{L}_{\text{max}}$ , the familiar effective  $\chi^2$  at its minimum. This quadratic form therefore quantifies the distance between the data and the model at its best parameter point. Taken as a CDE, if  $P(Q_{\text{ML}} > Q_{\text{ML}}^{\text{obs}})$  is too low, then the data are a bad fit to the model, and conversely, if it is too high, it is too good a fit to the data, possibly indicating the presence of confirmatory biases.

Equation (19) defines a quadratic form over data space, and its distribution, in general, depends on the evidence, as the probability of data given the model, which in turn depends on the prior. However, in this case the projection  $\mathbb{I} - \mathbb{P}$  in  $Q_{\text{ML}}$  makes its statistical properties independent of the prior and is given by  $Q_{\text{ML}} \sim \chi^2(d - N_{\mathcal{L}})$  (see Appendix B; here and below “ $\sim$ ” denotes “distributed as”). Here  $N_{\mathcal{L}} = \text{rank}[\mathbb{P}]$ , taking into account the fact that the likelihood might not be sensitive to some parameters if  $\partial m(\theta)/\partial \theta = 0$ . If there are no irrelevant parameters,  $N_{\mathcal{L}} = N$ .

Implicit in the use of  $Q_{\text{ML}}$  as a goodness-of-fit statistic is that the likelihood is maximized over all the relevant parameters without reference to or bounds from the prior. However, once the allowed model parameters are constrained by priors, we must adopt a different goodness-of-fit statistic.

The prior distribution usually encodes physical requirements on the model, like  $\Omega_m \geq 0$ , or a vague integration of previous experimental knowledge, like  $20 \leq H_0 [\text{kms}^{-1} \text{Mpc}^{-1}] \leq 100$ . We are not interested in a model that fits the data well while violating physical requirements or accepted previous results. The effect of the prior is to penalize such situations.

To define a GoF measure that takes the effect of the prior into account, we start from the evidence. To see why, we consider a one-parameter ( $\theta$ ) model and a data set that directly measures that parameter. The evidence is then  $\mathcal{E} = \int_{-\infty}^{+\infty} \mathcal{L}(D|\theta) \Pi(\theta) d\theta$ . Under the simplifying assumption that the likelihood depends on the difference between the parameter and the data (which in this example is just the measured value of the parameter), the evidence, as a function

of the data, becomes  $\mathcal{E} = \int_{-\infty}^{+\infty} \mathcal{L}(\theta - D)\Pi(\theta)d\theta$ . This is the convolution integral that gives the probability density of the difference between the prior and the data.

The evidence GoF is then defined by analogy to Eq. (20) as

$$Q_{\mathcal{E}} \equiv -2 \ln \mathcal{E} + 2\langle \ln \mathcal{E} \rangle_D + \langle Q_{\mathcal{E}} \rangle_D. \quad (21)$$

Unlike  $Q_{\text{ML}}$ , the specific quadratic form  $Q_{\mathcal{E}}$  describing the data dependence of the evidence depends on the prior, so we give its explicit form for the various cases below. This statistics quantifies the compatibility between the prior and the likelihood, defining a goodness-of-fit statistics that is effectively conditioned on the prior. We then apply the GLM to Eq. (21) for different prior choices.

If we consider uninformative flat priors, the evidence quadratic form is given by

$$Q_{\mathcal{E}} = (x - \hat{m})^T (\mathbb{I} - \mathbb{P})^T \Sigma^{-1} (\mathbb{I} - \mathbb{P})(x - \hat{m}), \quad (22)$$

just like  $Q_{\text{ML}}$ , and it is chi-squared distributed with  $d - N_{\mathcal{L}}$  d.o.f. This means that the evidence and maximum likelihood GoF statistics are identically distributed in the case of uninformative flat priors as one might expect.

At the other extreme are delta priors. The evidence goodness of fit is determined by

$$Q_{\mathcal{E}} = (x - m_{\Pi})^T \Sigma^{-1} (x - m_{\Pi}), \quad (23)$$

where  $x \sim \mathcal{N}_d(x; m_{\Pi}, \Sigma)$  so that  $Q_{\mathcal{E}} \sim \chi^2(d)$ . Notice that d.o.f. counting is different than in the uninformative flat prior case because the model cannot be optimized over the parameter space.

For Gaussian priors we have that

$$Q_{\mathcal{E}} = (x - m_{\Pi})^T (\Sigma + MC_{\Pi}M^T)^{-1} (x - m_{\Pi}). \quad (24)$$

Since the distribution of data draws is Gaussian,  $x \sim \mathcal{N}_d(x; m_{\Pi}, \Sigma + MC_{\Pi}M^T)$ ,  $Q_{\mathcal{E}} \sim \chi^2(d)$  just like the delta prior case. Although the model can now be optimized over the parameter space,  $Q_{\mathcal{E}}$  pays a compensating penalty from the prior.

These results for the evidence highlight two aspects that are worth commenting on. The first is that the evidence GoF is the optimal estimator to weight differences between the prior and the data. In both the delta and Gaussian prior cases, the difference between the model with priors and the data draws  $x - m_{\Pi}$  is weighted with its inverse covariance. We discuss in Appendix D what makes inverse covariance weighting optimal. In the case of uninformative flat priors, where there is no sense of preferred model parameters, this reduces to the usual maximum likelihood GoF. The second aspect is that there is a direct relationship between the evidence GoF and maximum likelihood based GoF, which is the result of a hidden symmetry. We can always regard priors as external data so that the evidence GoF for

Gaussian priors is the same as the maximum likelihood GoF if we add an additional data point for each Gaussian prior. With Gaussian priors on all  $N$  parameters, the maximum likelihood GoF would be distributed with  $(d + N) - N = d$  d.o.f., as the evidence GoF.

In practical applications we want to define a GoF measure that retains the best properties of both the maximum likelihood GoF and the evidence GoF. We want the former measure to be easy to compute while accounting for limitations that the prior places on optimizing parameters, which the latter measures.

Similar considerations in the literature for assessing Bayesian goodness of fit, for the purpose of model selection, has led to the use of the deviance information criterion (DIC), which measures the improvement of the likelihood, within the region of support of the prior, relative to the number of effective parameters that the data constrain. The DIC is defined as [21–25]

$$\text{DIC} \equiv -2 \ln \mathcal{L}(\theta_p) + 2N_{\text{eff}}, \quad (25)$$

where  $\theta_p$  is an estimate of the true parameters. Here,  $N_{\text{eff}}$  is the Bayesian complexity:

$$N_{\text{eff}} \equiv 2 \ln \mathcal{L}(\theta_p) - 2\langle \ln \mathcal{L} \rangle_{\theta}, \quad (26)$$

where the average is over the posterior. Note that  $\theta_p$  could be fixed to be the parameter mean or the maximum point of the posterior. Note that with the commonly used flat priors, the maximum likelihood point within the prior range is the maximum of the posterior. We therefore take the latter case for generality.

Now we can define, by analogy to Eq. (20), a new GoF statistic

$$Q_{\text{MAP}} = -2 \ln \mathcal{L}(\theta_p) + 2\langle \ln \mathcal{L}(\theta_p) \rangle_D + \langle Q_{\text{MAP}} \rangle_D, \quad (27)$$

for the likelihood at the maximum *a posteriori* (MAP) point. Since the specific quadratic form for  $Q_{\text{MAP}}$  depends on the prior, we now consider each case separately.

In both the delta and uninformative flat prior cases, the likelihood at maximum posterior is distributed as the evidence  $Q_{\text{MAP}} = Q_{\mathcal{E}}$ . In the Gaussian prior case it defines a quadratic form in data space:

$$Q_{\text{MAP}} = (x - m_{\Pi})^T [(\mathbb{I} - \mathbb{P})^T \Sigma^{-1} (\mathbb{I} - \mathbb{P}) + \tilde{M}^T C_{\Pi}^{-1} C_p C_p^{-1} C_{\Pi} \tilde{M}] (x - m_{\Pi}). \quad (28)$$

This case also illuminates the meaning of  $N_{\text{eff}}$ . If some directions in parameter space are not constrained by the data, as it happens in many practical applications, the quadratic form defined by Eq. (28) is lower rank; i.e., the model cannot invest all its nominal parameters in improving the goodness of the fit. Here,  $Q_{\text{MAP}}$  is distributed as a sum of Gamma-distributed variables, and its distribution can be

conservatively approximated by that of a chi-squared distributed variable with  $d - \text{tr}[(C_{\Pi} + C)^{-1}C_{\Pi}]$  d.o.f. The trace term is  $N_{\text{eff}}$  under GLM with Gaussian priors,

$$N_{\text{eff}} = N - \text{tr}[C_{\Pi}^{-1}C_p]. \quad (29)$$

It can be interpreted as the effective parameters that a data set is constraining. To see why, consider the limiting cases. If the prior covariance is much wider than the data covariance, this expression returns the full number of parameters  $N$ , whereas in the opposite limit where all parameters are prior limited, it returns zero. Thus, for any type of prior,  $0 \leq N_{\text{eff}} \leq N$ , making the uninformative flat and delta case bounds on  $N_{\text{eff}}$  and limits of the statistics of  $Q_{\text{MAP}}$ .

For the case of flat priors which may be informative, we can follow a similar procedure of identifying the effective number of parameters using Eq. (29). While this approximation is not exact, it tends to be conservative. Furthermore, being conservative for directions that are weakly constrained by the data mitigates non-Gaussianity in the posterior. Along these directions, it is more likely that the posterior is non-Gaussian with slowly decaying tails.

To summarize, our procedure gives the exact distribution of  $Q_{\text{MAP}}$  for all parameter space directions that are either completely constrained by the prior or the data and, in these limiting cases, reduces to the evidence GoF. Moreover, in the case of completely data constrained parameters, it further reduces to the maximum likelihood GoF measure.

### E. Evidence ratio type tests

We next proceed to the application of the GLM to estimators that aim at quantifying the compatibility of data set couples. One that has been applied in the literature is the evidence ratio estimator of data set compatibility [6,13,26–35].

With the posterior distribution of two different data sets, we want to test whether they can be described with the same set of cosmological parameters. This amounts to comparing the probabilities of two different statements:

- (i)  $\mathcal{I}_0$ : the two data sets are described by the same choice of unknown parameters;
- (ii)  $\mathcal{I}_1$ : the two data sets are described by independent choices of unknown parameters.

Then we compute their probabilities and compare them:

$$C = \frac{P(D_1 \cup D_2 | \mathcal{I}_0, \mathcal{M})}{P(D_1 \cup D_2 | \mathcal{I}_1, \mathcal{M})} = \frac{P(D_1 \cup D_2 | \mathcal{M})}{P(D_1 | \mathcal{M})P(D_2 | \mathcal{M})}, \quad (30)$$

where  $P(D_1 \cup D_2 | \mathcal{M})$  is the joint evidence of the two data sets while  $P(D_1 | \mathcal{M})$  and  $P(D_2 | \mathcal{M})$  are the evidence for the single ones. Since we are working with two data sets,  $D_1$  and  $D_2$ , we use the subscripts 1, 2, and 12 to indicate quantities referring to the first, the second, and the joint data sets, respectively.

Used in the form of Eq. (30), the evidence ratio does not provide an estimate of the statistical significance of the reported results. It is common in the literature to interpret the outcome on a Jeffreys' scale [36,37]:  $\ln C < 0$  indicates tension between the data sets and  $\ln C > 0$  agreement; 3 : 1 odds one way or the other is “substantial,” 10 : 1 is “strong,” 30 : 1 is “very strong,” and 100 : 1 is “decisive.” This has the disadvantage that the Jeffreys' scale is not calibrated on the specific application at hand, and using it might give misleading results [29,38,39].

In the case of uninformative flat priors the Gaussian approximation for the evidence ratio can be immediately read from Eqs. (14) and (16):

$$\begin{aligned} \ln C &= \ln \mathcal{L}_{\text{max}}^{12} - \ln \mathcal{L}_{\text{max}}^1 - \ln \mathcal{L}_{\text{max}}^2 \\ &+ \frac{1}{2} \ln \frac{|C_{12}|}{|C_1||C_2|} + \ln \frac{V(\Pi_1)V(\Pi_2)}{V(\Pi_{12})} \\ &- \frac{N_1 + N_2 - N_{12}}{2} \ln(2\pi), \end{aligned} \quad (31)$$

and this shows that, when averaging this quantity over  $D_1 \cup D_2$  realizations, several terms would not cancel out; i.e., this CDE is biased. In the case of uninformative flat priors, the calculation explicitly gives

$$\begin{aligned} \langle \ln C \rangle_{12} &= -\frac{N_1 + N_2 - N_{12}}{2} [1 + \ln(2\pi)] \\ &+ \frac{1}{2} \ln \frac{|C_{12}|}{|C_1||C_2|} + \ln \frac{V(\Pi_1)V(\Pi_2)}{V(\Pi_{12})}. \end{aligned} \quad (32)$$

Notice that, in practical applications, the Occam's razor factors in the second line of Eq. (32) are much larger than the first line, thus making the evidence ratio biased toward agreement since  $\mathcal{I}_1$  effectively involves two Occam's factors compared with one for  $\mathcal{I}_0$ . Priors are in fact generally chosen to be as uninformative as possible, so the posterior is almost always localized in a small fraction of the prior volume, making the Occam factor due to prior volume very large. This makes the application of the evidence ratio likely to be misleading in practical applications, generally underestimating discrepancies. In the literature a positive evidence ratio,  $\ln C > 0$ , was usually used as a sufficient criterion to claim consistency of two different probes. We stress that one should really expect a very large value of  $\ln C$  if the data are truly consistent and that discrepancies might be hidden by the bias computed in Eq. (32). A smaller value, but still positive, only shows that the data are possibly inconsistent but that the preferred parameter values for the two subsets differ by an amount that is small in comparison with the prior range.

We define the debiased evidence ratio test as

$$\Delta \ln C = -2 \ln C + 2 \langle \ln C \rangle_{12}. \quad (33)$$



If  $\Delta \ln C$  is significantly greater than zero, this indicates tension; if it is smaller than zero, it indicates confirmation bias. The confidence level of the statement can be computed using the GLM. The proofs of the results of this section can be found in Appendix C.

In the case of uninformative flat priors,  $\Delta \ln C$  is, up to an additive constant, chi-squared distributed with  $N_1 + N_2 - N_{12}$  d.o.f., and the observed value can be read from Eqs. (31) and (32). In the case of delta priors the evidence ratio is trivially distributed as  $\Delta \ln C = 0$  for all data draws.

For Gaussian priors, the distribution is more complicated and is, in general, a sum of independent variance-gamma distributed variables (see Appendix C). Notice that in this case, to obtain the distribution of the Gaussian prior evidence ratio from that of the maximum likelihood ratio, treating the prior as additional data, we need to take into account the fact that we add the prior to the analysis of both  $D_1$ ,  $D_2$  and  $D_{12}$ . If we now regard the prior as data, since the prior does not change in the analysis of the different data sets, data draws of  $D_1$  and  $D_2$  would be correlated by the prior. The evidence in the Gaussian prior case is then in correspondence with the maximum likelihood ratio of correlated data sets.

As with GoF in the previous section, we aim at defining a CDE that retains ease of use as the ratio of maximum likelihoods, which does not require heavy use of numerical integration to compute the statistical significance, like the evidence ratio, but at the same time encodes the effect of the prior. This suggests that we again examine the statistics of the various likelihoods at their maximum posterior point.

We therefore consider the difference of log likelihoods at their MAP point

$$Q_{\text{DMAP}} \equiv -2 \ln \mathcal{L}_{12}(\theta_p^{12}) + 2 \ln \mathcal{L}_1(\theta_p^1) + 2 \ln \mathcal{L}_2(\theta_p^2). \quad (34)$$

Note that in this case, the normalization factors in  $\mathcal{L}$ , which provide the offset mean values in Eq. (20), drop out of the difference so long as 1 and 2 are independent data sets. If data are drawn from the evidence with uninformative flat priors and delta priors, the distribution of  $Q_{\text{DMAP}}$  is the same as the distribution of the evidence ratio. In the Gaussian prior case its distribution is conservatively approximated with a chi-squared distribution:

$$Q_{\text{DMAP}} \sim \chi^2(N_{\text{eff}}^1 + N_{\text{eff}}^2 - N_{\text{eff}}^{12}). \quad (35)$$

Its exact distribution in terms of a sum of Gamma distributed variables can be found in Appendix C.

This estimator quantifies the loss in goodness of fit when combining two data sets. When considering single data sets, the model parameters can be separately optimized within the prior; when joining them, there is less freedom in model parameter optimization. The ratio of likelihoods at maximum posterior tells us whether this decrease in

goodness of fit is consistent with expectation from statistical fluctuations or not.

The statistics of  $Q_{\text{DMAP}}$  is the same as the evidence ratio, once Occam's factors are removed, for completely data or prior constrained directions, while it differs over partially constrained directions. Over these directions the statistical significance of agreement or disagreement is underestimated as a mitigation strategy against non-Gaussianities.

This discussion allows us to shed light on the deviance information criterion (DIC) ratio estimator, as introduced in [33] to assess the agreement between CFHTLenS and Planck. Using Eq. (25), we can define the DIC ratio

$$\ln \mathcal{I} = -\frac{1}{2} [\text{DIC}(D_1 \cup D_2) - \text{DIC}(D_1) - \text{DIC}(D_2)]. \quad (36)$$

Similarly to the evidence ratio, Eq. (36) is expected to indicate agreement or disagreement between two posterior distributions if it is found to be negative or positive, respectively. Depending on the evaluation point  $\theta_p$  for DIC, the statistics of the DIC ratio changes accordingly. If  $\ln \mathcal{L}(\theta_p)$  in the DIC statistic is evaluated at maximum posterior, then twice the DIC ratio is distributed as  $Q_{\text{DMAP}}$ , up to a data-independent constant. If the maximum likelihood is taken without regard for the prior, the distribution is chi squared, with  $N_1 + N_2 - N_{12}$  d.o.f., similar to the maximum likelihood ratio and the evidence ratio in the uninformative flat prior case. This clarifies the relationship between the evidence ratio, the DIC ratio, and  $Q_{\text{DMAP}}$ . When the data are either informative or completely uninformative, these three quantities measure the same aspect of agreement or disagreement with different mean values over data space.

## F. Parameter differences

The next application of the GLM is to understand the distribution of quadratic forms in model parameters. These are natural generalizations of the usual rule-of-thumb estimator for tension and contain, as subcases, other estimators that have been proposed in the literature.

If we consider two independent random variables  $\theta_1$  and  $\theta_2$ , the probability density of their difference, in one dimension,  $\Delta\theta \equiv \theta_1 - \theta_2$ , is given by the convolution integral of the two probability densities,  $P_{\theta_1}$  and  $P_{\theta_2}$ , as

$$P(\Delta\theta) = \int_{-\infty}^{+\infty} P_{\theta_1}(\tilde{\theta}) P_{\theta_2}(\tilde{\theta} - \Delta\theta) d\tilde{\theta}. \quad (37)$$

Tension between the measurements would be indicated if  $P(\Delta\theta)$  has most of its support at very negative or positive  $\Delta\theta$ . For the former case, there would be a low probability for the difference to be greater than zero:

$$P(\Delta\theta > 0) = \int_0^{\infty} P(\Delta\theta) d\Delta\theta. \quad (38)$$

To account for the possibility that the observed tension could be in either direction, we take the smaller of  $P(\Delta\theta > 0)$  and  $P(\Delta\theta < 0)$ . The probability of obtaining a 1D parameter shift,  $T_1$ , more extreme than the data, in either direction, is then

$$P(T_1 > T_1^{\text{obs}}) = 2 \min[P(\Delta\theta > 0), P(\Delta\theta < 0)]. \quad (39)$$

We refer to this as the 1D parameter shift tension statistic. This holds for any two independent probability distributions and can be easily evaluated numerically.

If we assume that the two distributions,  $P_{\theta_1}$  and  $P_{\theta_2}$ , are Gaussian, then we can evaluate this probability analytically. Since the convolution of two Gaussians is another Gaussian, with a variance given by the sums of the individual Gaussians, the tension statistic becomes the usual ‘‘rule-of-thumb difference in mean.’’ This consists in comparing the difference in the best-fit values, or means, of one parameter for two different data sets to the quadrature sum of the parameters’ variances:

$$T_1(\theta) \equiv \frac{|\theta(D_1) - \theta(D_2)|}{\sqrt{\sigma_\theta^2(D_1) + \sigma_\theta^2(D_2)}}, \quad (40)$$

where  $\theta(D_i)$  is the parameter best fit (or mean), for a given model and data set  $D_i$ , and  $\sigma_\theta^2(D_i)$  denotes its variance. The statistical significance of the 1D parameter shift then becomes  $P(T_1 > T_1^{\text{obs}}) = \text{erf}(T_1/\sqrt{2})$ , where erf is the error function.

Because of its simplicity, this estimator is an easy and intuitive proxy to understand tensions between data sets and is also accurate if differences in a parameter are manifest at the posterior level. However, there is no guarantee that the overall consistency of two generic data sets is properly signaled: The method needs to pick up the ‘‘right’’ parameter where all tension is expressed; it would not work right away in the multidimensional case as it does not take into account the effect of priors.

When considering more than one dimension, we can turn to the GLM to understand the statistics of tension estimators that, like the rule-of-thumb difference in mean, are defined in parameter space.

We consider differences in the posterior means of two different data sets:

$$\Delta\bar{\theta} \equiv \theta_{p1} - \theta_{p2}, \quad (41)$$

which can be easily computed, from the results of Sec. II C, as

$$\Delta\bar{\theta} = C_{p1}[C_\Pi^{-1}\theta_\Pi + C_1^{-1}\theta_1^{\text{ML}}] - C_{p2}[C_\Pi^{-1}\theta_\Pi + C_2^{-1}\theta_2^{\text{ML}}], \quad (42)$$

for Gaussian priors, and

$$\Delta\bar{\theta} = \theta_1^{\text{ML}} - \theta_2^{\text{ML}}, \quad (43)$$

in the case of uninformative flat priors. Note that under the GLM, the parameter means are the same as the parameters at the maximum posterior point.

Notice that both Eqs. (42) and (43) are defined in terms of the parameters that  $D_1$  and  $D_2$  have in common, so when there are additional parameters describing systematic effects in one data set, the corresponding distributions have to be marginalized over them. When treating Gaussian priors, we assume that the prior center is the same for both data sets and equal to the prior center of their combination, as we assumed in the previous sections.

Since the posterior means depend on the data, we now turn to the computation of the statistics of their differences over the space of joint data draws from  $D_1$  and  $D_2$ .

Since  $\Delta\bar{\theta}$  is a linear combination of correlated Gaussian variables, it is Gaussian distributed. Furthermore, it can be shown that, for both Gaussian and uninformative flat priors,

$$\langle \Delta\bar{\theta} \rangle_{12} = 0. \quad (44)$$

Notice that this holds if the prior center is fixed (to an arbitrary value) for  $D_1$ ,  $D_2$  and  $D_{12}$ . If this is not the case and the prior center is different for the different data sets, the expectation value of the parameter difference is nonzero.

We are then left with computing the covariance of  $\Delta\bar{\theta}$ . In the case of uninformative flat priors, this reads

$$C(\Delta\bar{\theta}) = C_{p1} + C_{p2}, \quad (45)$$

while in the case of Gaussian priors, direct computation from the GLM gives

$$C(\Delta\bar{\theta}) = C_{p1} + C_{p2} - C_{p1}C_\Pi^{-1}C_{p2} - C_{p2}C_\Pi^{-1}C_{p1}. \quad (46)$$

These results can be directly obtained by means of the covariance of the joint data draws reported in Appendix C.

Having computed the distribution of  $\Delta\bar{\theta}$ , we can compute the distribution of a related quantity that carries the same information but has useful properties when applied in practice to data sets with non-Gaussian posteriors. This is the difference between the mean parameters of one data set and the mean parameters of the joint data set.

We refer to this quantity as the update difference in mean since it quantifies the differences in parameters of one data set when updating it with another one. If we assume that the GLM applies to  $D_1$  and  $D_2$ , then it also applies to  $D_{12}$  and we can write the update difference in mean as

$$\Delta\bar{\theta}_U \equiv \theta_{p1} - \theta_{p12} = C_{p1}(C_{p1} + C_2)^{-1}(\theta_{p1} - \theta_2), \quad (47)$$

which still has zero mean and covariance:

$$C(\Delta\bar{\theta}_U) = C_{p1} - C_{p12}, \quad (48)$$

for both uninformative flat priors and Gaussian priors. Since the CDEs discussed in this section are defined in terms of the parameter space posterior only, it is simple to derive all of the above results on parameter differences by considering the two data sets and the prior as independently measuring  $\theta$  directly in parameter space using the projected covariances  $\mathcal{C}$ .

Notice that the previously discussed covariances have to be positive definite or positive semidefinite. While this is true when all distributions are well-defined Gaussians in the application to real data, with covariances from MCMC sampling, it might not be strictly true. We come back to the problem of computing this estimator in Sec. IV D.

We are now in a position to define CDEs based on quadratic forms of parameter differences. Given a positive semidefinite matrix  $A$ , we define two types of quadratic estimators, depending on the vector that we use to define them.

If we consider  $\Delta\bar{\theta}$  we have a difference in mean quadratic CDEs defined as

$$Q_{\text{DM}} = (\Delta\bar{\theta})^T A (\Delta\bar{\theta}), \quad (49)$$

while if we use  $\Delta\bar{\theta}_U$  we have update difference in mean quadratic CDEs defined as

$$Q_{\text{UDM}} = (\Delta\bar{\theta}_U)^T A (\Delta\bar{\theta}_U). \quad (50)$$

All these quadratic forms are central and sometimes degenerate, depending on the rank of  $A$ .

We have two estimators that belong to this family of CDE that have been studied previously. The first one is the difference in mean  $\Delta\bar{\theta}$  with  $A = \mathcal{C}^{-1}(\Delta\bar{\theta})$  (e.g., [40–42]). The second one is the surprise that was introduced in [20] and used in [7,39,43], and corresponds to  $\Delta\bar{\theta}$  with  $A = \mathcal{C}_1^{-1}$ , which is related to the Gaussian approximation of the Kullback-Leibler divergence [44] between different data sets’ posteriors. The consideration of quadratic forms for the update  $\Delta\bar{\theta}_U$  is new to this work as far as we are aware.

For either  $\Delta\bar{\theta}$  or  $\Delta\bar{\theta}_U$ , the optimal choice of  $A$  is the inverse covariance of the parameter difference that is being considered. Other measures, provided that their distribution is properly calculated, can only underestimate rare events by not weighting them properly compared to an optimal measure. We discuss the criterion that makes inverse covariance weighting optimal in Appendix D. This also clarifies why the rule-of-thumb difference in mean in one dimension works so well when all the tension is manifest in one parameter where the choice in weighting of multiple dimensions is absent.

We therefore consider only  $A = \mathcal{C}^{-1}$  in the following. With this choice the quadratic forms of Eqs. (49) and (50) are chi-squared distributed with d.o.f.  $\langle Q_{\text{DM}} \rangle = \text{rank}[\mathcal{C}(\Delta\bar{\theta})]$  and  $\langle Q_{\text{UDM}} \rangle = \text{rank}[\mathcal{C}(\Delta\bar{\theta}_U)]$ , respectively.

### III. MODEL AND DATA SETS

Our baseline model is the six-parameter  $\Lambda$ CDM model defined as follows: cold dark matter density  $\Omega_c h^2$ ; baryon density  $\Omega_b h^2$ ; the angular size of the sound horizon  $\theta_{\text{MC}}$ ; the spectral index of the primordial spectrum of scalar fluctuations  $n_s$  and its amplitude  $\ln(10^{10} A_s)$ ; and the reionization optical depth  $\tau$ . We also include in the model massive neutrinos, fixing the sum of their masses to the minimal value allowed by flavor oscillation measurements  $\sum_\nu m_\nu = 0.06$  eV [45]. We discuss in Appendix F the priors that we use throughout this work.

We analyze the level of agreement of several, publicly available, cosmological data sets within the  $\Lambda$ CDM model. The first data set that we consider consists of the measurements of CMB fluctuations in both temperature (T) and polarization (EB) of the *Planck* satellite [11,46]. We further consider the *Planck* 2015 full-sky lensing potential power spectrum [47] in the multipole range  $40 \leq \ell \leq 400$ . We exclude multipoles above  $\ell = 400$  as CMB lensing, at smaller angular scales, is strongly influenced by the non-linear evolution of dark matter perturbations.

We include the “joint light-curve analysis” (JLA) supernovae sample [48], which combines SNLS, SDSS, and HST supernovae with several low redshift ones. We also use BAO measurements of BOSS in its DR12 data release [49], the SDSS main galaxy sample [50], and the 6dFGS survey [51]. We include the galaxy clustering power spectrum data derived from the SDSS LRG survey DR4 [52] and the WiggleZ Dark Energy Survey galaxy power spectrum as measured from 170,352 blue emission line galaxies over a volume of 1 Gpc<sup>3</sup> [53,54]. For both data sets we exclude all the data points with  $k > 0.08$  h/Mpc.

We consider the measurements of the galaxy weak lensing shear correlation function as provided by the Canada-France-Hawaii Telescope Lensing Survey (CFHTLenS) [55] in their reanalyzed version of [33] and the Kilo Degree Survey (KiDS) [14]. We apply ultraconservative cuts that make both CFHTLenS and KiDS data insensitive to the modeling of nonlinear evolution, and we include uncertainties in the modeling of intrinsic galaxy alignments, as in [14,33]. *A posteriori* we notice that, when considering ultraconservative data cuts, intrinsic alignment parameters are weakly constrained.

We use local measurements of the Hubble constant derived by the Supernovae, H0, for the Equation of State of Dark Energy (SHOES) team [12] with the calibration of [56]. In addition we employ measurements derived from the joint analysis of three multiply-imaged quasar systems with measured gravitational time delays, from the HOLiCOW Collaboration [57].

We combine the previously discussed data sets into families probing similar physical processes: a CMB family composed of CMB temperature, polarization, and CMB lensing reconstruction; a background family joining supernovae and BAO measurements; the combination of SDSS

TABLE I. Summary of data sets and data set combinations used in this work.

Acronym	Data set	Year	Reference
<i>lowl</i>	Planck low- $\ell$ TEB	2015	[46]
<i>CMBTT</i>	Planck high- $\ell$ TT	2015	[46]
<i>CMBEE</i>	Planck high- $\ell$ EE	2015	[46]
<i>CMBTE</i>	Planck high- $\ell$ TE	2015	[46]
<i>CMBL</i>	Planck CMB Lensing	2015	[47]
<i>SN</i>	JLA	2014	[48]
<i>BAO</i>	BOSS DR12 + SDSS MGS + 6dFGS	2011–15	[50,51,60]
<i>LRG</i>	SDSS LRG survey DR4	2006	[52]
<i>WiggleZ</i>	WiggleZ survey	2012	[53,54]
<i>CFHTLenS</i>	CFHTLenS survey	2016	[33]
<i>KiDS</i>	KiDS survey	2016	[14]
<i>H</i>	SH0ES	2016	[12,56]
<i>HSL</i>	H0LiCOW	2016	[57]
<i>CMB</i>	<i>lowl</i> + <i>CMBTTTEEE</i> + <i>CMBL</i>	2015	...
<i>BG</i>	<i>SN</i> + <i>BAO</i>	2011–15	...
<i>GC</i>	<i>LRG</i> + <i>WiggleZ</i>	2006–12	...
<i>WL</i>	<i>CFHTLenS</i> + <i>KiDS</i>	2016	...
<i>H0</i>	<i>H</i> + <i>HSL</i>	2016	...

LRG and WiggleZ measurements probing the clustering of galaxies; CFHTLenS and KiDS joined together in a weak lensing probe; Hubble constant measurements from SH0ES and H0LiCOW.

Notice that the background family does not measure the Hubble constant as SN measurements are analytically marginalized over intrinsic luminosity. The galaxy clustering data set does not measure the present-day amplitude of CDM perturbations  $\sigma_8$  as both power spectrum measurements are separately marginalized over the power spectrum amplitude. The H0LiCOW data set in turn not only measures  $H_0$  but a combination of  $H_0$  and  $\Omega_m$  since we implement the full non-Gaussian likelihood described in [57].

Table I summarizes the data sets, acronyms, and literature references for all the data sets used in this work. We use the CAMB code [58] to compute the predictions for all the cosmological observables described above, and we use MCMC sampling of the posterior of the previously discussed experiments with CosmoMC [59].

Our results rest on two assumptions: linear theory modeling of the observables and the accuracy of the GLM. As a sanity check of the former, we compare the parameter posterior and best-fit prediction of the data, as obtained by neglecting and including nonlinear modeling of the matter distribution, described by Halofit [61], with the updated fitting formulas described in [62,63]. We find that, with the above discussed setup, the parameter posterior and best fits are not noticeably different.

All the techniques considered in this work rely on the applicability of the Gaussian approximation to either the likelihood or the posterior of the considered data set. Most of the considered data sets have Gaussian likelihoods, with the exception of the *HSL* and *lowl* data sets that we exclude

from tests requiring Gaussianity of the data likelihood. We build the Gaussian approximation of the parameter space posterior, and we check whether we can reliably use it, as discussed in Appendix E. We find that the posterior of all combinations of data sets containing the CMB power spectrum can be well approximated by Gaussian distributions in the parameters. Single, weakly constraining data sets, on the other hand, usually result in non-Gaussian parameter posteriors.

## IV. APPLICATION AND RESULTS

In this section we discuss the application of the CDEs in Sec. II to cosmological data. This section is organized as follows: In Sec. IV A we discuss our recommended suite of CDE tests for assessing internal and pairwise data consistency; in Sec. IV B we present the results of internal consistency tests; in Sec. IV C we show the results of the application of compatibility tests for data set couples.

### A. Methodology

To assess the internal consistency of a data set, we consider the likelihood at maximum posterior as a goodness-of-fit measure:

$$\begin{aligned}
 Q_{\text{MAP}} &\equiv -2 \ln \mathcal{L}(\theta_p) - d \ln(2\pi) - \ln(|\Sigma|) \\
 &\sim \chi^2(d - N_{\text{eff}}), \\
 N_{\text{eff}} &\equiv N - \text{tr}(\mathcal{C}_{\Pi}^{-1} \mathcal{C}_p). \tag{51}
 \end{aligned}$$

To test the compatibility of data set couples,  $D_1$  and  $D_2$ , we consider the ratios of likelihoods at their maximum posterior:

$$Q_{\text{DMAP}} \equiv -2 \ln \mathcal{L}_{12}(\theta_p^{12}) + 2 \ln \mathcal{L}_1(\theta_p^1) + 2 \ln \mathcal{L}_2(\theta_p^2) \\ \sim \chi^2(N_{\text{eff}}^1 + N_{\text{eff}}^2 - N_{\text{eff}}^{12}), \quad (52)$$

which measures the decrease in, prior constrained, goodness of fit when combining two data sets.

This is paired with parameter shifts in their update form:

$$Q_{\text{UDM}} \equiv (\theta_p^1 - \theta_p^{12})^T (C_{p1} - C_{p12})^{-1} (\theta_p^1 - \theta_p^{12}) \\ \sim \chi^2(\text{rank}[C_{p1} - C_{p12}]). \quad (53)$$

When possible we apply these CDEs to every data set alone and to sets that define families of physical probes, to test their internal consistency. Then we move to testing the consistency of different families by probing all their possible combinations.

Different tests applied to the same data sets provide complementary information that is helpful in singling out possible problems. Goodness-of-fit-type tests inform us of the internal consistency of the data sets but do not specifically highlight confirmation biases or tensions that look like parameter changes. The ratios of likelihoods at their maximum posterior and parameter shift tests, on the other hand, are designed to isolate problems along parameter modes. In particular, the former estimator is sensitive to shifts in all the parameters that two data sets jointly constrain, while the latter is sensitive to shifts in the constraints that one of the data sets improves over the other.

As an example, the goodness-of-fit test for a data set might fail, indicating a tension. Still, parameter deviations, probed by the other two tests, might not be statistically significant, indicating that possible systematic effects or new physics do not mimic the effect of a change in parameters. In a cosmological context the matter power spectrum could indicate the presence of an additional physical scale resulting in a scale-dependent growth. With sufficient experimental accuracy this will fail goodness-of-fit tests, as growth in the  $\Lambda$ CDM model is scale independent. On the other hand, this may not fail parameter shift tests as none of the nominal  $\Lambda$ CDM parameters can exactly describe this effect. Conversely, a smooth dark energy component will generally result in a scale-independent modification to the growth of structures that might mimic the effect of a change in  $A_s$  or other cosmological parameters. This will not show up at the goodness-of-fit level but might show up at the parameter level when we compare two probes that have different sensitivities to the amplitude of perturbations, for example, measuring it at different redshifts. In this case also, the joint goodness-of-fit test is not guaranteed to fail as it might be dominated by the data set with a larger number of data points.

In addition to these aspects, different tests, when applied in practice, have different responses to the presence of non-Gaussianities in the data and parameter spaces and thus have different failure modes. Testing multiple ones ensures

that these are easily identified. In particular, if the posterior of a given experiment is non-Gaussian because the low probability tails decay slower than a Gaussian distribution, the evidence ratio and parameter shift estimator have different opposite responses. While the first one would overestimate tensions and underestimate confirmation, the second one is built to mitigate this and may underestimate tensions.

In Appendix G we report, in table format, the full results of the application of the CDEs that we discuss to data. In addition, we also report the results that can be obtained with the 1D parameter shifts and rule-of-thumb difference in mean statistics, when evaluated with our data configuration and analysis pipeline, to recover some known results that we use as a benchmark for our estimators.

## B. Goodness-of-fit type tests

In this section we present the application of the goodness-of-fit measures that were discussed in Sec. II D.

In applying these estimators to real data, there are two major challenges. The first one consists in obtaining accurate best-fit estimates. This involves global optimization of the posterior and is complicated by the large number of parameter space dimensions usually involved in cosmological studies. What proves particularly challenging in this respect is the presence of mostly unconstrained parameters that can create multiple local maxima in the posterior. This can be mitigated by having well-converged MCMC parameter chains whose sample best-fit estimate provides a good starting point to eventually find the global minimum with appropriate algorithms.

The second challenging aspect is to correctly estimate the number of parameters that a data set is constraining,  $N_{\text{eff}}$ . Prior distributions are, in practice, often non-Gaussian, for example, when some direct or derived parameter is limited to be in a certain range. Nonetheless, in all cases, we adopt Eq. (29) for its calculation. This is a reasonable approximation in that the comparison of the prior covariance  $C_{\Pi}$  to the posterior covariance  $C_p$  always provides a criterion for when the prior is informative and the parameter cannot be optimized to the data. As a concrete example, consider the top hat prior on a single parameter where  $C_{\Pi} = (\theta_{\text{max}} - \theta_{\text{min}})^2/12$ . Equation (29) tells us that  $N_{\text{eff}} = 1/2$  when the prior variance equals the data variance. For the top hat prior, this occurs when the half-width is  $\sqrt{3}$  times the rms of the data constraint, i.e., between  $1\sigma$  and  $2\sigma$  of a Gaussian data constraint.

Therefore, Eq. (29) suffices for an estimate even for this highly non-Gaussian prior so long as we allow for errors in each partially constrained direction at the level of a few tenths of a parameter. We have verified this error estimate with numerical simulations in one dimension, noticing that the error depends on the value of  $N_{\text{eff}}$ : It is small in the two limits  $N_{\text{eff}} = 0$  and  $N_{\text{eff}} = 1$  where the distribution is exact; it increases as  $N_{\text{eff}}$  decreases from  $N_{\text{eff}} = 0.9$  to

$N_{\text{eff}} = 0.1$  approximately ranging from 0.1 to 0.4; in this same range of  $N_{\text{eff}}$ , the distribution of the  $Q_{\text{MAP}}$  estimator is increasingly conservative.

Evaluating Eq. (29) for  $N_{\text{eff}}$  also requires well-sampled parameter distributions to limit errors in parameter covariance estimates. We thus require the Gelman and Rubin  $R$  test [64,65] to satisfy  $R - 1 < 0.005$  for the worst constrained covariance eigenvalue. We can then check sampling errors on the number of effective parameters as their variance across different MCMC chains, and we find that these are usually of the same order as  $R - 1$ .

In order to have a reliable estimate of  $N_{\text{eff}}$ , we also need a good knowledge of the prior covariance. This is built by joining different blocks. We directly MCMC sample the prior on the base  $\Lambda$ CDM parameters because of priors on derived parameters. Flat priors on nuisance parameters are uncorrelated with priors on the base parameters, and their diagonal entry in the prior covariance is built out of the covariance of the flat distribution. Some nuisance parameters have Gaussian priors that are uncorrelated with other priors. Their covariance entry can be easily set with the variance of the Gaussian prior. Further details about the modeling of the prior distribution can be found in Appendix F.

Once these technical aspects have been properly addressed, we can check the estimate of the number of effective parameters that a data set constrains against physical intuition. We list in Table II the values of  $N_{\text{eff}}$  and the number of nominal parameters for the data sets that we consider.

As we can see, the primary CMB spectra have seven, five, and seven parameters for *CMBTT*, *CMBEE*, and *CMBTE*, respectively, that are not constrained by the data. These are nuisance parameters describing foregrounds and are instead constrained by informative Gaussian priors [46]. CMB lensing has four unconstrained parameters:  $\tau$ ,  $n_s$ ,  $\Omega_b h^2$ , and a calibration parameter. A combination of the other cosmological parameters, mainly  $A_s$  and  $\Omega_c h^2$ , is well constrained by the lensing amplitude, whereas the directions

constraining the shape of the potential are only partially constrained. *SN* constrain three parameters, the total matter density  $\Omega_m$  and two nuisance parameters, the intrinsic supernovae color and stretch. The *BAO* data set constrains three parameters, as it includes redshift space distortion measurements, so that only  $\tau$  and  $n_s$  are unconstrained while  $A_s$  is mostly unconstrained. The *LRG* and *WiggleZ* data sets constrain slightly more than two parameters, which are combinations of  $\Omega_m$ ,  $\Omega_b$ , and  $H_0$ , thanks to the detection of the BAO feature in the matter power spectrum. Both *CFHTLenS* and *KiDS* constrain two parameters, the amplitude of the weak lensing signal and the amplitude of intrinsic alignment. The latter, while not being detected, is slightly constrained over the prior and thus enters in d.o.f. counting.

The number of effective parameters that combinations of these data sets constrain is consistent with what we would expect from these results. Notice that no physical knowledge was input to get the results of Table II, which automatically and accurately recover the physical results to a fraction of a parameter.

We can now turn to the probabilities associated with the values of  $Q_{\text{MAP}}$  in the various cases, as displayed in Fig. 3. In applying these estimators to the data, we cannot use the *lowl* and *HSL* data sets as their likelihood is not Gaussian in the data points. We have to exclude the *H* data set as the full data likelihood is not provided, and we just have the parameter likelihood.

As we can see from both Fig. 3 and Table V, the *CMBTT*, *CMBEE*, *CMBL*, *SN*, *BAO*, and *WiggleZ* data sets are a reasonable fit to the data, showing no tension nor confirmation at high statistical significance. The *CMBL* result showcases the use of maximum posterior as a goodness-of-fit measure. This data set has no irrelevant parameters, and if we were to count all its parameters as being optimized, this would indicate the presence, at a 5% probability to exceed, of tensions. Since the  $\Lambda$ CDM model cannot use all its nominal parameters due to the priors, it is actually still a good fit to the *CMBL* data.

The *CMBTE* data set in turn is not a good fit at high statistical significance. The result is stable against d.o.f. counting since the goodness of fit, in this case, is dominated by the number of data points in the fit. Since, as noted in [46], the coadded frequency spectrum is a good fit, we suspect that this result is dominated by frequency-dependent rather than cosmological effects, e.g., foreground and systematics modeling, especially in the  $100 \text{ GHz} \times 217 \text{ GHz}$  and  $100 \text{ GHz} \times 100 \text{ GHz}$  spectra that have been highlighted in [46], at about the same statistical significance.

The full *CMB* goodness of fit is dominated by the TE results, whose statistical significance gets diluted by the increased number of data points in the joint data set. The results for the *CMBTTTEEE* data set further confirm this, showing that the discrepancy in the fit cannot be attributed to *CMBL* measurements. Moreover, the goodness-of-fit

TABLE II. The number of effective parameters,  $N_{\text{eff}}$ , and the number of nominal parameters,  $N$ , for the different data sets that we consider.

Data set	$N_{\text{eff}}$	$N$
<i>CMBTT</i>	14.3	21
<i>CMBEE</i>	8.1	13
<i>CMBTE</i>	7.9	15
<i>CMBL</i>	2.5	7
<i>SN</i>	3.0	8
<i>BAO</i>	3.1	6
<i>LRG</i>	2.5	6
<i>WiggleZ</i>	1.9	6
<i>CFHTLenS</i>	1.8	7
<i>KiDS</i>	1.8	7

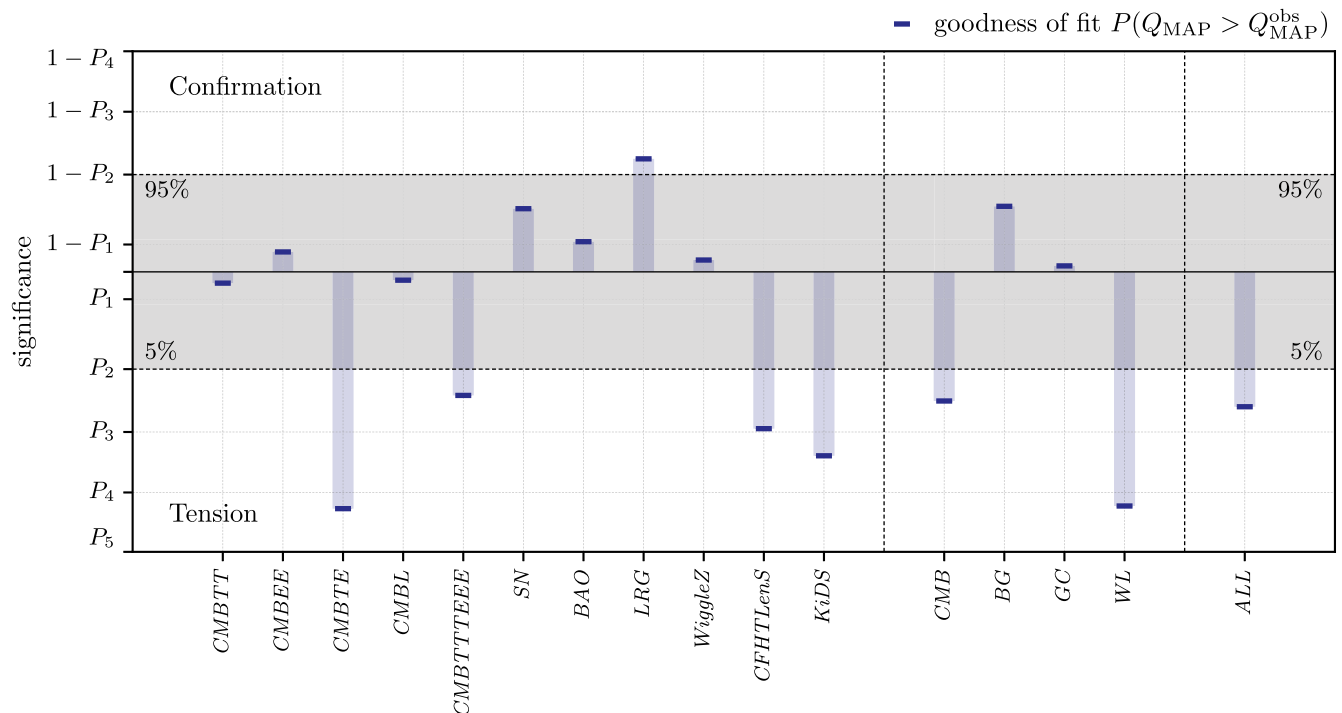


FIG. 3. The statistical significance of the posterior goodness-of-fit estimator,  $Q_{\text{MAP}}$ , from Eq. (51), applied to different data sets and data set combinations. The labels report different levels of statistical significance:  $P_1 \equiv 32\%$ ,  $P_2 \equiv 5\%$ ,  $P_3 \equiv 0.3\%$ ,  $P_4 \equiv 0.007\%$ , and  $P_5 \equiv 0.00006\%$ . The darker shaded area indicates results that are not statistically significant.

results for all data sets joined together (*ALL*) are dominated too by *CMB* results since this is the data set with the largest number of data points.

At slightly lower statistical significance we find that the *CFHTLenS* and *KiDS* data sets are a bad fit, and the goodness of fit of their union further confirms this at high statistical significance. Notice that this result is particularly worrisome since both data sets are cut at linear cosmological scales and thus should not be influenced by the, possibly improper, modeling of nonlinearities. The statistical significance of the goodness of fit to the joint *WL* data set is only slightly lower than the product of the single data sets, showing that the bad fits are almost independent. These results could be, at least in the case of the *KiDS* data set, due to lack of modeling of survey geometry in the covariance, as reported in [66]. The same explanation does not apply to *CFHTLenS* whose covariance was obtained through simulations.

At a statistical significance that is borderline between significant and not significant, we find that the *LRG* data set is confirmation biased. Notice that, in this case, proper d.o.f. counting is crucial to the assessment of such effects. If we were to assume that this data set measures all  $\Lambda$ CDM parameters, this result will not be statistically significant. If we further assume that the two bias parameters that have been marginalized over are also constrained by the data, the statistical significance of confirmation bias would decrease, becoming 96% for  $N_{\text{eff}} = 3.5$  and 93% for  $N_{\text{eff}} = 4.5$ .

Finally, we notice that the *BG* data set is a good fit, while being dominated by the *SN* data set that has more data points with respect to the *BAO* one. The same effect is seen for the *GC* data set where the statistical significance of confirmation in *LRG* measurements is overweighted by the number of data points in the *WiggleZ* data set.

### C. Evidence ratio type tests

In this section we present the application of the ratio of likelihoods at maximum posterior estimator  $Q_{\text{DMAP}}$ , introduced in Sec. II E, and discuss its relationship with the evidence ratio.

The practical challenges in computing the  $Q_{\text{DMAP}}$  estimator are the same as the maximum posterior goodness of fit and are mitigated in the same way that was discussed in the previous section. The only difference is that, in the previous section, errors on  $N_{\text{eff}}$  have a small effect for all data sets that have a large number of data points. In this context, it is crucial to properly identify parameters, as the number of considered data points drops out of d.o.f. counting, as shown in Sec. II E. As we show in Appendix G, see Table VI, the number of effective parameters for the single and joint data sets agrees well with physical intuition, and their difference appropriately reflects the number of parameters that both data sets measure.

Similarly to the previous section, we cannot use the *lowl* and *HSL* data sets as their likelihood is non-Gaussian in the data. In addition, we cannot apply this test to data set

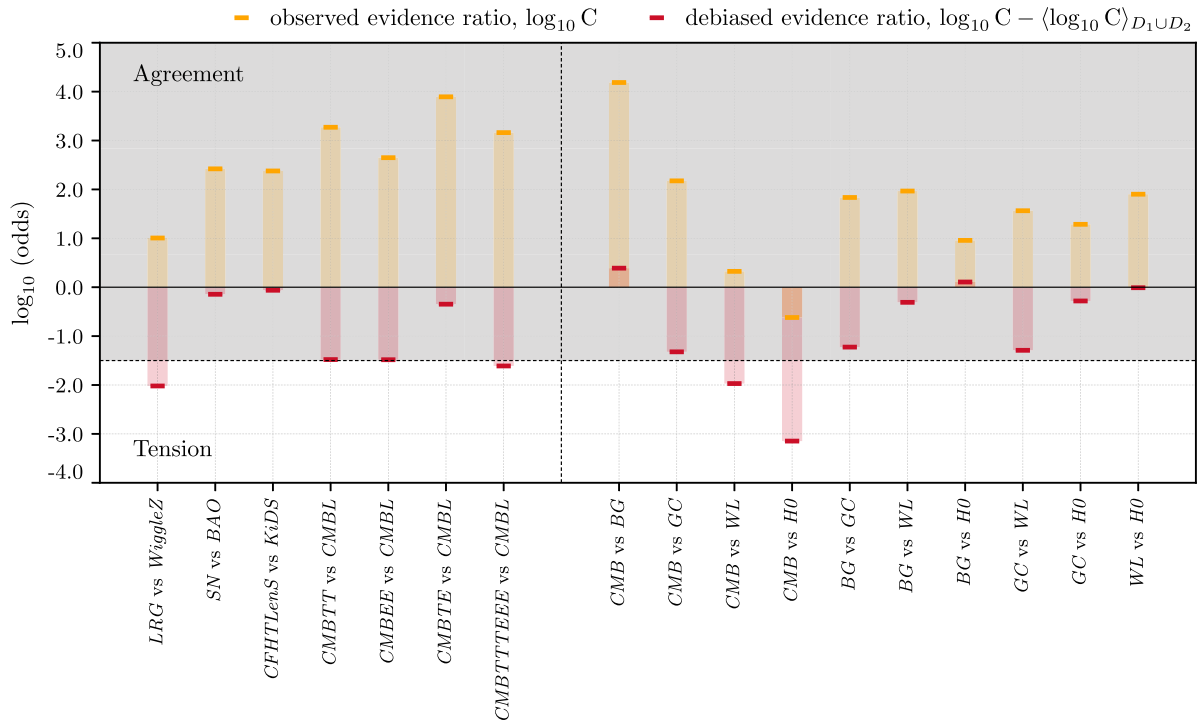


FIG. 4. The evidence ratio estimator applied to different data set couples. We show the nominal observed value of the evidence ratio test and its debiased value. Notice that for most of the data sets, the bias in the evidence ratio estimator is as large as its observed value. The darker shade indicates that results would not be considered statistically significant on the Jeffreys’ scale.

couples that are correlated, and we have to exclude the comparison of the primary CMB spectra.

Before turning to  $Q_{\text{DMAP}}$  we apply the evidence ratio test to several data couples, as shown in Fig. 4, and subtract its bias, as computed within the GLM. The evidence is estimated with the Gaussian approximation to the MCMC posterior, as discussed in Appendix E, and its bias is computed using the statistics of that approximation.

The first noteworthy result that is shown in Fig. 4 is that the observed value of the evidence ratio is usually of the same order of the bias in the evidence ratio. This bias also depends on the data set involved in the comparison and has to be subtracted case by case. This shows the limitations of the evidence ratio test judged on the Jeffreys’ scale. The result is usually so biased that the observed value alone cannot be used to judge agreement or disagreement.

On the other hand, in Fig. 5, we show the statistical significance of the  $Q_{\text{DMAP}}$  estimator. The reported results confirm the picture that comes from the debiased evidence ratio while providing an estimate of statistical significance. The qualitative agreement between the two is due to the fact that, when parameter space directions are either completely constrained by the prior or the data,  $Q_{\text{DMAP}}$  is distributed as the evidence apart for additive factors that do not depend on the data realization and drop out of the statistical significance.

We first consider the internal compatibility of data within the set families. As we can see the *SN* and *BAO* data sets agree

as well as the *CFHTLenS* and *KiDS* data sets, making the *BG* and *WL* families internally consistent. The *LRG* and *WiggleZ* data sets, on the other hand, show a marked indication of disagreement. This is not surprising considering the indication of confirmation bias in the *LRG* data set, and it points toward a significant difference in parameters between the two probes. This difference is not signaled by the rule-of-thumb difference in mean when applied to the  $\Omega_m$  and  $\Omega_b$  parameters, pointing toward a correlated shift in parameters. Notice that, in this case, the bias in the evidence ratio is larger than the observed value. If we were to look at the latter and judge its value of the Jeffreys’ scale, we would draw the wrong conclusion that the two data sets agree.

Other interesting results concern the internal consistency of the *CMB* family. The *CMBTT* and *CMBEE* data sets do not agree with *CMBL* at about a 5% probability to exceed. For both data sets, this is roughly the same statistical significance of the deviation of the amplitude of the lensing parameter,  $A_L$ , from one, as reported in [11,16]. While *CMBTE* and *CMBL* agree, the joint result of *CMBTTTEEE* and *CMBL* is dominated by the tension in the temperature spectrum, consistently with the results in [11,16]. Notice that the evidence ratio result obtained with the Gaussian approximation to *CMBTT* and *CMBL* agrees very well with the result of numerical integration shown in [6].

We next apply the evidence ratio test to understand the compatibility of different families of physical probes.

As we can see in Fig. 5 the *CMB* family agrees well with the *BG* family but disagrees with the other three families of



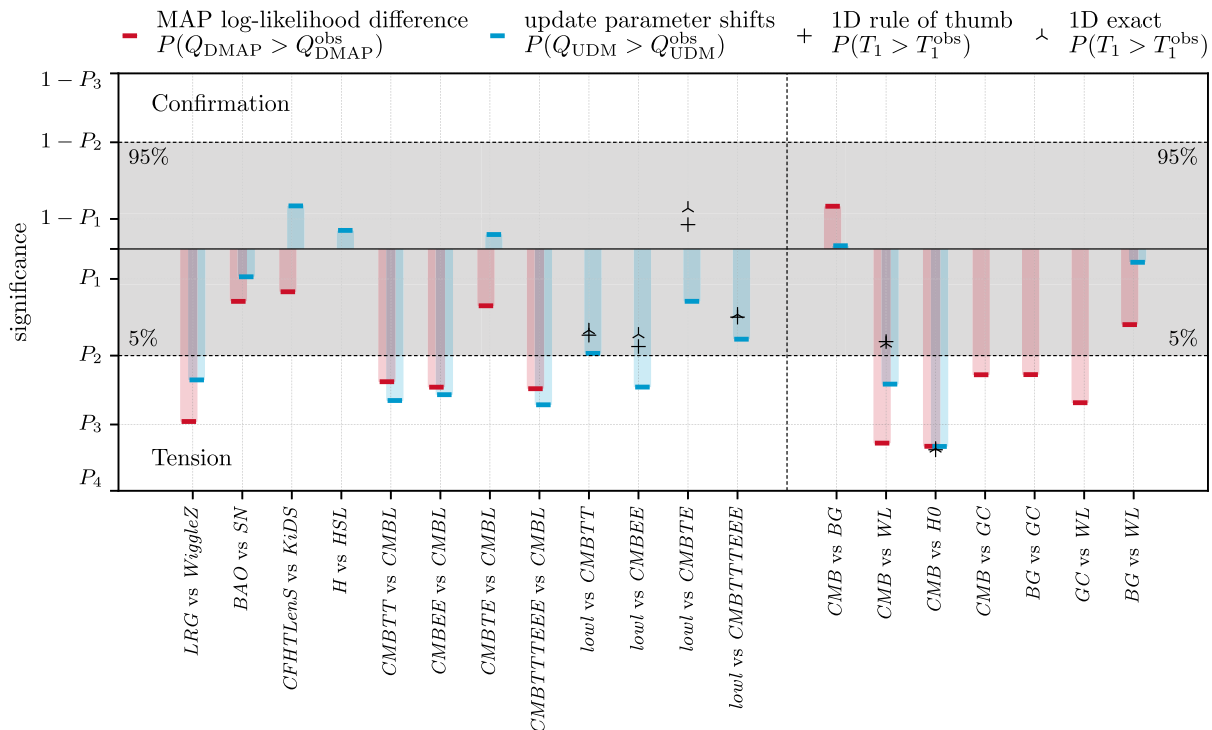


FIG. 5. The statistical significance of different CDEs for various data set couples: the difference in log-likelihood at MAP,  $Q_{\text{DMAP}}$  from Eq. (52); the update parameter shifts test,  $Q_{\text{UDM}}$  from Eq. (53); the exact 1D parameter shifts,  $T_1$  from Eq. (39); and the rule-of-thumb difference in mean, as the Gaussian approximation of  $T_1$  from Eq. (40). Different colors indicate different tests, as shown in the legend. The labels report different levels of statistical significance:  $P_1 \equiv 32\%$ ,  $P_2 \equiv 5\%$ ,  $P_3 \equiv 0.3\%$ ,  $P_4 \equiv 0.007\%$ . Values that are identified as failure modes of one of the estimators are not shown in the figure. The darker shaded area indicates results that are not statistically significant.

data sets that we consider. The disagreement between *CMB* and *GC* families can be understood by considering the indication of a confirmation bias in the *LRG* data set. The statistical significance of the disagreement between these two probes roughly matches the statistical significance of confirmation in the *LRG* data set, pointing toward the hypothesis that the latter data set might be confirmation biased around parameter values that are not the *CMB* ones. The *CMB* data set also shows high statistically significant indications of tensions with the *WL* and *H0* data sets. The tension with Hubble constant measurements is known, and we recover 0.088% probability to exceed compared with the rule-of-thumb difference in mean result applied to  $H_0$  of 0.073% and the exact 1D shift that results in 0.078%. The *WL* result is also known but is usually evaluated using the full scale measurements of weak lensing, including scales that are influenced by the nonlinear evolution of cosmological perturbations. Here we show that this tension persists and remains statistically significant, specifically at 0.1% probability to exceed, when restricting to linear scales. Notice that the evidence ratio between the *CMB* and *H0* data set is the only one that is found to be negative. Still interpreting at face value this ratio on the Jeffreys' scale would lead to the incorrect conclusion that the tension is not significant.

The other data set families considered generally agree. From a physical standpoint we know that they should since they are either measuring different parameters or weakly measuring the same parameters. This aspect is properly recovered, and none of them is found to be in tension or confirmation biased at relevant statistical significance. The only exception is the test applied to the *BG* and *WL* data sets against the *GC* data set. The first is in tension with the latter as a consequence of its agreement with *CMB* at about the same statistical significance. The second one is in tension with the latter due to the fact that both data sets have some problem at the goodness-of-fit level. Their combination, not surprisingly, signals disagreement of some sort.

#### D. Parameter differences

In this section we present the application of the parameter shift CDE discussed in Sec. II F.

The challenges in applying this CDEs to real data are profoundly different than the ones that we discussed in the previous sections. This allows for a larger degree of complementarity between tests and ensures the robustness of conclusions against possible contamination from non-Gaussianities and other estimate problems.

In the following we only use the parameter difference estimator  $Q_{\text{UDM}}$  with Eq. (50), which is defined through the parameter update when combining two data sets. Parameter difference estimators of the form  $Q_{\text{DM}}$  using Eq. (49) have problems that are difficult to overcome in practical applications. In the case of uninformative flat priors, any such test would be ill posed for directions that are unconstrained by one of the data sets. If we consider Gaussian priors, then the  $Q_{\text{DM}}$  itself can be formally defined. However, noise in the determination of the covariances of the two experiments, due to MCMC sampling, makes it difficult to disentangle prior constrained and data constrained directions. In applying it to the data, we find this estimator to be unreliable and numerical noise dominated for a wide variety of algorithms used for the estimate.

Aside from numerical issues, differences in parameter updates also have the clear advantage that corrections due to non-Gaussianities are mitigated if the posterior of the most constraining data set is Gaussian. In our cosmological applications CMB data play this role since parameter posteriors are nearly Gaussian for all  $\Lambda$ CDM parameters. If the second data set has a non-Gaussian posterior, a direct parameter difference would misestimate significance if the mean of the first set is in the tail of the second set. For the parameter update, GLM is effectively applied around the mean of the first set by replacing the non-Gaussian posterior of the second set with a Gaussian approximation locally around that point.

To minimize numerical noise in the  $Q_{\text{UDM}}$  estimates, we use the Karhunen-Loeve (KL) decomposition of the two covariances that are involved. Recall that to compute the observed value of the update parameter shift, we need to evaluate

$$Q_{\text{UDM}} \equiv (\Delta\bar{\theta}_U)^T (\mathcal{C}_{p1} - \mathcal{C}_{p12})^{-1} (\Delta\bar{\theta}_U). \quad (54)$$

The second data set can only add information on top of the first data set, so  $(\mathcal{C}_{p1} - \mathcal{C}_{p12})$  has to be positive definite in the absence of numerical noise. In the presence of numerical noise, it is better to first transform to the KL basis since it is mutually orthogonal in the metrics defined by  $\mathcal{C}_{p1}$  and  $\mathcal{C}_{p12}$ . We solve the generalized eigenvalue problem to find the KL modes,  $\phi^a$ , of the two covariances:

$$\sum_{\nu} \mathcal{C}_{p1}^{\mu\nu} \phi_{\nu}^a = \lambda^a \sum_{\nu} \mathcal{C}_{p12}^{\mu\nu} \phi_{\nu}^a. \quad (55)$$

Here the eigenmodes are defined to be orthonormal in the  $\mathcal{C}_{p12}$  metric

$$\sum_{\mu\nu} \phi_{\mu}^a \mathcal{C}_{p12}^{\mu\nu} \phi_{\nu}^b = \delta^{ab}, \quad (56)$$

and since they are orthogonal in the  $\mathcal{C}_{p1}$  metric, but with variance  $\lambda^a \delta^{ab}$ , the KL basis provides linear combinations

of the parameters that are mutually independent and ordered by the improvement in the variance of 12 over 1. If we now define the linear combination of parameter differences in the KL basis as

$$\Delta p^a = \sum_{\mu} \phi_{\mu}^a \Delta \bar{\theta}_U^{\mu} \quad (57)$$

we obtain

$$Q_{\text{UDM}} \equiv \sum_{a=1}^{N_{\text{KL}}} \frac{(\Delta p^a)^2}{\lambda^a - 1}. \quad (58)$$

While this transformation, when  $N_{\text{KL}}$  is the full set of KL modes, gives exactly the same value as Eq. (54), it also highlights the problem of numerical noise. If 12 does not improve over 1 substantially in a given mode, then  $\lambda^a \approx 1$  and numerical noise in the estimation of covariances creates large errors in  $Q_{\text{UDM}}$ . The KL decomposition allows us to place a well-defined lower cutoff on this improvement in order to remove unwanted numerical noise from the estimator. In practical applications there is a hierarchy of KL modes so that noise and data modes are well separated in the spectrum. We use a simple algorithm to find this separation point and define the optimal cutoff for each data set combination. To minimize numerical noise in the  $Q_{\text{UDM}}$  estimates, we also notice that it is preferable to use the mean of the parameters in the test rather than the best-fit parameters, even though they are the same in the GLM.

We find that for parameter distributions that are well approximated by Gaussian distributions, the cutoff is usually in the range of 5%, while it can be as large as 15% in the case of non-Gaussian posteriors. In all cases we limit the cutoff to be between 2% and 20%, and we cannot extend it to zero or the estimator will be noise dominated. Notice that this prescription also effectively defines

$$\langle Q_{\text{UDM}} \rangle = N_{\text{KL}}, \quad (59)$$

and hence  $Q_{\text{UDM}}$  is chi-squared distributed with  $N_{\text{KL}}$  d.o.f.

With this technique the estimator is stable but is left with one case where the statistic returns a null result. When we combine a data set that is very constraining with a data set that is very weakly constraining, the improvement in the KL modes might be below the threshold that separates data dominated modes and noise dominated modes. In this case the value of  $Q_{\text{UDM}}$  will be zero, and it distributed as a  $\chi^2$  with zero d.o.f., i.e., zero for all data realizations. This simply means that while there may be a true, but tiny, parameter shift, it is too small to measure. In this case, the procedure correctly returns the answer that neither tension nor confirmation bias can be detected.

We start by applying the update difference in mean to assess the consistency of data set families, and we report the results in Fig. 5.

As we can see the disagreement between the *LRG* and *WiggleZ* data sets, at parameter level, is confirmed to be statistically significant, as we found in the previous section. The statistical significance of this result is, however, slightly lower than what is reported by the likelihood at the maximum posterior test. This effect can be attributed to the different sensitivity of the two estimators to the effect of non-Gaussianities in the parameter posteriors.

The parameter update results further confirm the internal consistency of the *BG* and *WL* families, as found in the previous section. On the other hand, we can extend here the study of internal consistency of families of probes to include *H* and *HSL* measurements. While the latter has a likelihood that is non-Gaussian at the data level, at the parameter level it can be well approximated by a Gaussian distribution. As we can see the two data sets agree on the determination of the Hubble constant while not showing indications of tensions or confirmation.

Similarly, here we can extend the study of the *CMB* internal consistency, even though the *lowl* likelihood is non-Gaussian at the data level. The update parameter shift test confirms the tension between *CMBTT*, *CBEE*, and *CMBL* and the agreement between *CMBTE* and *CMBL*, at about the same statistical significance that was found in the previous section.

If we now consider the same set of comparisons, with the addition of the *CMB* large angular scale multipoles, we see that the agreement between the primary *CMB* spectra and *CMBL* improves to the point that it is not statistically significant. This picture is consistent with the results of the update parameter shift test applied between the *lowl* data and the primary *CMB* spectra. As we can see from Fig. 5, all four results are on the tension side and exceed 95% C.L. for the *CBEE* spectrum. These are also in qualitative agreement with the rule of thumb and 1D shift when applied to the  $\tau$  parameter. The tension is reported to be slightly larger because the direction that is selected by the KL decomposition takes into account degeneracies with other cosmological parameters.

The discrepancy between these three probes can be physically understood because at fixed  $A_s e^{-2\tau}$ , lowering  $\tau$  reduces  $A_s$  and hence reduces the gravitational lensing potential and the smoothing of the *CMB* peaks. At high multipoles the *CMB* measurements of Planck have enough precision to be sensitive to gravitational lensing; hence other parameters shift to compensate for the decreased smoothing of the peaks. This is achieved by increasing  $\Omega_m h^2$  and  $A_s e^{-2\tau}$ , while reducing  $n_s$  and  $\Omega_b h^2$ , as discussed in [15]. The best-fit solution to the *lowl* + *CMBTT* has known oscillatory residuals at high multipoles [15] because of the lack of power at large angular scales. Without the *lowl* data set, these oscillatory residuals can be fit by raising  $\tau$ , which is balanced by also raising  $A_s$  and  $\Omega_m h^2$ , which, overall, give a larger *CMB* lensing signal that is in conflict with lensing reconstruction of the *CMBL* data

set. This tension can then be isolated by adding a new parameter that describes the amplitude of the lensing of the *CMB*,  $A_L$ , which allows us to fit the oscillatory residuals in the primary spectra and is found to deviate from unity at about the statistical significance of the tensions that we report here.

We can now proceed to the application of the update parameter shift test to different data set families. These results are largely in agreement with the ones reported in the previous section, with some noticeable differences. As shown in Appendix G these results do not depend strongly on the inclusion of the *lowl* data set that leaves them largely unchanged.

While the tension between *CMB* and *HO* is confirmed and in good agreement with the benchmark results, specifically at 0.087% agreement probability, the tension between the *CMB* and *WL* data sets is markedly lower than the  $Q_{\text{DMAP}}$  result, specifically at 1.6% agreement probability. This is expected since the *WL* data set shows a non-Gaussian posterior with slowly decaying tails. Still, this tension is noticeably higher with respect to the rule-of-thumb estimate applied to the  $S_8$  parameter, which yields 7.1% agreement probability, and the exact 1D shift that takes into account the non-Gaussianity of the posterior and results in 6.7% agreement probability.

In this case the  $Q_{\text{UDM}}$  test indicates, through the number of d.o.f., that this tension is evaluated along one parameter space direction,  $\langle Q_{\text{UDM}} \rangle = 1$ . This direction is built to be the optimal one for both data sets. The  $S_8$  parameter, in turn, does not exactly describe the amplitude of the lensing signal, at the redshifts of the combined *WL* surveys that we are considering, and is not the best constrained parameter. We find that, for the *WL* data set,  $\sigma_8 \Omega_m^{0.7}$  is better constrained, and the rule-of-thumb test signals a tension similar to that of the  $Q_{\text{UDM}}$  estimator.

Finally, we can easily see an example of the null result mode of the estimator by looking at data set combinations involving the *GC* data set. This data set is very weakly constraining, when compared with the *CMB* data set, so its improvement on the parameter constraints cannot be distinguished from numerical noise. Furthermore, the *GC* data set is very weakly constraining along the parameter space directions that are constrained by the *WL* data set, so the result is again dominated by noise and our KL procedure properly identifies this as a null update result.

## V. CONCLUSIONS

We studied statistical estimators of CDEs between cosmological probes and applied them to state-of-the-art cosmological data sets.

We discussed the likelihood at maximum posterior as a measure of the goodness of fit. Unlike the maximum likelihood, this quantity depends on the prior on cosmological parameters and allows us to disentangle parameter space directions that are constrained by the data and by the

prior. This disentanglement provides a fair d.o.f. counting when performing the goodness-of-fit test.

We studied the distribution of the evidence ratio test of data set compatibility over the space of data realizations. This allowed us to uncover the fact that the evidence ratio is usually biased toward agreement and that, in practical applications, this bias is as large as the observed value, making the Jeffreys' scale unreliable as an indicator of agreement or disagreement. We then defined a similar estimator based on the ratio of likelihoods at maximum posterior that allows for an assessment of statistical significance of the reported results. While being equivalent to the evidence ratio in the limiting cases where parameter space directions are completely constrained by either the data or the prior, this estimator is significantly easier to apply.

We investigated the statistics of parameter shifts developing methods that work in an arbitrary number of dimensions. These estimators optimally weight the parameter shifts and mitigate the fact that tensions might not be identifiable at the single parameter level because they are hidden by the process of marginalization over a high dimensional parameter space. We introduce a robust regularization scheme based on the Karhunen-Loeve decomposition which identifies and discounts the small parameter shifts due to sampling noise in MCMC posteriors.

When applying these estimators to cosmological data, we find several noteworthy results. As a benchmark for the estimators we recover the known result regarding tensions between the Planck measurements of the CMB spectra and local measurements of the Hubble constant and the amplitude of the galaxy weak lensing signal. Concerning the latter, we find that, when considering the Canada-France-Hawaii Telescope Lensing Survey and the Kilo Degree Survey on large linear scales, the statistical significance of the disagreement with CMB measurements is between 98.4% and 99.9%. This is somewhat higher than what is estimated by looking at the posterior of the  $S_8 \equiv \sigma_8 \Omega_m^{0.5}$  parameter alone, as we optimally weight all parameter space directions.

We investigated the consistency of CMB measurements of the Planck satellite, establishing a set of results that allow us to prioritize the analysis of the next release of the Planck data. In particular, we find the following: The CMB TE cross correlation is a bad fit and this seems to be related to the presence of residual, frequency-dependent foregrounds; the discrepancy between the CMB TT spectrum and its lensing reconstruction is also present in the E-mode spectrum at about the same statistical significance; the measurements of the large angular scale multipoles,  $\ell < 30$ , are in tension with the small scale temperature and E-mode spectra at about 95% probability.

Moreover, we find CMB results to be in tension with probes of the clustering of galaxies. This disagreement can probably be attributed to the SDSS LRG DR4 survey being slightly confirmation biased toward a different set of cosmological parameters.

We also find that most of the other combinations of data sets are in agreement and can thus be safely combined and, in particular, that there is agreement between SN and BAO, between the two WL surveys that we consider, between strong lensing time delay measurements of the Hubble constant and direct measurements from the distance ladder, and between CMB measurements and SN and BAO.

Overall, we find that the statistical significance of the discrepancies identified in this work is not yet sufficient to firmly establish whether they are due to residual systematic effects or new physical phenomena. In this sense their statistical significance is not yet to the point where we can clearly draw a line between data sets that should not be combined with others because of unaccounted for systematic effects. We highlight that the resolution of these discrepancies, or their unequivocal identification, is likely to come as a result of further improvements of the quality of the data.

The work toward understanding the consistency of present cosmological probes and preparing for the analysis of the next generation of probes is far from complete. Future efforts in these directions include the generalization of the techniques presented in this paper to consider non-Gaussian corrections. Moreover, we need to develop statistical estimators that work on more than two data sets at a time, allowing us to compute the joint distribution of multiple tests. These will allow us to understand the global consistency of the  $\Lambda$ CDM model with a large and diverse set of experimental data. In addition, these would allow us to perform analyses targeted at identifying the most outlying data set, within a larger pool of data sets.

Finally, these tests should be applied as we gather new and more precise cosmological data sets to make sure that inconsistencies due to systematic effects or incomplete modeling of cosmological observables are identified and corrected and that discrepancies due to new physical phenomena are promptly found.

## ACKNOWLEDGMENTS

We thank Saroj Adhikari, Chuck Bennett, Chihway Chang, Tom Crawford, Mike Hobson, Dragan Huterer, Pavel Motloch, Sam Passaglia, Kimmy Wu, and Georgios Zacharegkas for useful comments. M.R. and W.H. are supported by U.S. Department of Energy Contract No. DE-FG02-13ER41958. W.H. was further supported by NASA ATP NNX15AK22G and the Simons Foundation. Computing resources were provided by the University of Chicago Research Computing Center through the Kavli Institute for Cosmological Physics at the University of Chicago.

## APPENDIX A: QUADRATIC FORMS IN GAUSSIAN RANDOM VARIABLES

In this appendix we briefly outline how to practically deal with the statistics of the many quadratic forms that appear in the main text. This material is mostly taken from

[67] and reproduced here to ease the comprehension of the main text.

A quadratic form in the  $p$  dimensional random Gaussian variable  $X$  is defined by

$$Q = X^T A X; \quad X \sim \mathcal{N}_p(x; \mu, \Sigma). \quad (\text{A1})$$

The first two moments of the quadratic form are

$$\begin{aligned} \langle Q \rangle_X &= \text{tr}[A\Sigma] + \mu^T A \mu, \\ \text{Var}(Q) &= 2\text{tr}[(A\Sigma)^2] + 4\mu^T A \Sigma A \mu. \end{aligned} \quad (\text{A2})$$

In the following we only consider the case of central quadratic forms  $\langle X \rangle = \mu = 0$ . We find that all distributions in the main text satisfy this requirement. For the generalization of the following results to the case where  $\mu \neq 0$ , we refer the reader to [67].

Over the subspace where  $\Sigma$  is invertible,  $Q$  admits a decomposition of the form

$$Q = X^T A X = \sum_{j=1}^p \lambda_j U_j^2, \quad (\text{A3})$$

where  $\lambda = \text{eigenval}(A\Sigma)$ ,  $P = \text{eigenvec}(A\Sigma)$  and

$$U = P^T \Sigma^{-1/2} (x - \mu) \sim \mathcal{N}_p(x; 0, \mathbb{I}). \quad (\text{A4})$$

Given that  $U_j$  is a normally distributed variable,  $U_j^2$  is a  $\chi^2(1)$  variable, and so  $Q$  is, in general, distributed as the sum of scaled  $\chi^2(1)$  variables which are themselves known as Gamma distributed variables.

If  $A$  is any projection of  $\Sigma^{-1}$ , i.e.,  $A = \mathbb{P}^T \Sigma^{-1} \mathbb{P}$  where  $\mathbb{P}^2 = \mathbb{P}$ , then all the eigenvalues  $\lambda_j \in 0, 1$  and  $Q$  would be the sum of independent  $\chi^2(1)$  variables,  $Q \sim \chi^2(r)$ , with  $r = \text{rank}(\mathbb{P})$  d.o.f. This includes the trivial case where  $A = \Sigma^{-1}$ .

More generally, if all the eigenvalues  $\lambda_j \geq 0$ , then analytic expressions for the probability density of  $Q$  exist [67], and probabilities can be computed with dedicated algorithms [68] once the eigenvalues of  $A\Sigma$  are obtained.

Alternately, the distribution of  $Q$  can be approximated by that of a chi-squared variable matching some of the moments of  $Q$ . We refer to these as Patnaiks' type approximations [69]. The first approximation matches the mean to a (single) chi-squared distribution:

$$Q = \sum_j \lambda_j X_j^2 \simeq \chi^2(\text{tr}[A\Sigma]), \quad (\text{A5})$$

where  $\simeq$  stands for ‘‘approximately distributed as.’’ The second approximation matches the mean and variance to that of (single) Gamma distribution:

$$Q = \sum_j \lambda_j X_j^2 \simeq c\chi^2(\nu), \quad (\text{A6})$$

where

$$c \equiv \sum_j \lambda_j^2 / \text{tr}[A\Sigma], \quad \nu \equiv (\text{tr}[A\Sigma])^2 / \sum_j \lambda_j^2. \quad (\text{A7})$$

Notice that in both approximations the number of d.o.f. of the (scaled) chi-squared distribution is usually not integer.

We use the first approximation in practice, matching only the mean. When  $0 \leq \lambda \leq 1$ , this approximation is conservative as the second approximation and the full distribution have smaller variance. These types of approximations are usually relevant over parameter space directions that are partially constrained by the prior, where the full posterior of the data set that we consider is usually highly non-Gaussian. Underestimating their contribution to the statistical significance of the reported results is then a mitigation strategy against non-Gaussianities.

## APPENDIX B: PROOFS OF SEC. II D

In this appendix we provide the proofs for the results contained in Sec. II D as a working example of how to use the GLM in practice.

We first consider the maximum likelihood:

$$-2 \ln \mathcal{L}_{\max} = X^T (\mathbb{I} - \mathbb{P})^T \Sigma^{-1} (\mathbb{I} - \mathbb{P}) X + d \ln(2\pi) + \ln(|\Sigma|), \quad (\text{B1})$$

where, here and below,  $X \equiv x - \hat{m}$ , with  $\hat{m} = m_{\Pi}$  for convenience, involves the component of the data vector that is in the complement of the model space  $(\mathbb{I} - \mathbb{P})X$ . For any prior, this data vector is distributed as

$$(\mathbb{I} - \mathbb{P})X \sim \mathcal{N}_r((\mathbb{I} - \mathbb{P})X; 0, (\mathbb{I} - \mathbb{P})\Sigma(\mathbb{I} - \mathbb{P})^T), \quad (\text{B2})$$

where  $r = \text{rank}(\mathbb{I} - \mathbb{P})$ , since the projector nulls the part of the data draw covariance that lives on parameter space. The data dependent piece of the maximum likelihood statistic contains

$$Q_{\text{ML}} = [(\mathbb{I} - \mathbb{P})X]^T \Sigma^{-1} [(\mathbb{I} - \mathbb{P})X], \quad (\text{B3})$$

which is a quadratic form for this data vector. Considering now the results of the previous section, the statistics of  $Q_{\text{ML}}$  is determined by the eigenvalues

$$\lambda = \text{eigenval}[\Sigma^{-1}(\mathbb{I} - \mathbb{P})\Sigma(\mathbb{I} - \mathbb{P})^T] = \text{eigenval}(\mathbb{I} - \mathbb{P}), \quad (\text{B4})$$

which implies  $Q_{\text{ML}} \sim \chi^2(r)$ . If we assume that the model has  $N$  relevant parameters, then  $\mathbb{P}$  projects the data vector onto an  $N$  dimensional subspace and therefore its complement has  $r = d - N$ .

We now turn to the distribution of the evidence. For uninformative flat priors, the evidence quadratic form is identical to the maximum likelihood quadratic form and is therefore also distributed as  $\chi^2(d - N)$ . In the case of delta priors, the evidence quadratic form in data space is

$$Q_{\mathcal{E}} = X^T \Sigma^{-1} X, \quad (\text{B5})$$

where  $X$  is normally distributed with covariance  $\Sigma$ . Thus,  $Q_{\mathcal{E}}$  is chi-squared distributed with full rank  $\chi^2(d)$ .

In the case of Gaussian priors the quadratic form defined by the evidence becomes

$$Q_{\mathcal{E}} = X^T \Sigma_0^{-1} X, \quad (\text{B6})$$

where  $X$  is normally distributed with covariance  $\Sigma_0 = \Sigma + M C_{\Pi} M^T$ . Thus,  $Q_{\mathcal{E}}$  is again distributed as  $\chi^2(d)$ .

We next derive the distribution of the likelihood at maximum posterior for different prior choices. In the case of delta priors, we have

$$\begin{aligned} -2 \ln \mathcal{L}(\theta_p) &= -2 \ln \mathcal{L}(\theta_{\Pi}) \\ &= X^T \Sigma^{-1} X + d \ln(2\pi) + \ln(|\Sigma|), \end{aligned} \quad (\text{B7})$$

which, up to a constant, defines a quadratic form in data space:

$$Q_{\text{MAP}} = X^T \Sigma^{-1} X, \quad (\text{B8})$$

which is distributed as  $\chi^2(d)$ .

In the case of uninformative flat priors the likelihood at maximum posterior is just the maximum likelihood, so  $Q_{\text{MAP}} = Q_{\text{ML}}$ .

Gaussian priors stand in between these cases. The data dependent part,  $Q_{\text{MAP}}$ , of the likelihood at maximum posterior is given by

$$\begin{aligned} Q_{\text{MAP}} &= -2 \ln \mathcal{L}(\theta_p) - d \ln(2\pi) - \ln(|\Sigma|) \\ &= X^T [(\mathbb{I} - \mathbb{P})^T \Sigma^{-1} (\mathbb{I} - \mathbb{P}) \\ &\quad + \tilde{M}^T C_{\Pi}^{-1} C_p C_{\Pi}^{-1} \tilde{M}] X. \end{aligned} \quad (\text{B9})$$

This quadratic form has  $d - N$  eigenvalues with  $\lambda_i = 1$  together with the  $N$  eigenvalues of  $C_{\Pi}^{-1} C_p$  that are bounded as  $0 \leq \lambda_i \leq 1$ . From this set of eigenvalues the exact distribution can be computed or approximated as in Appendix A. If we take the first approximation, Eq. (A5), that matches the mean, then  $Q_{\text{MAP}} \simeq \chi^2(d - N + \text{tr}[C_{\Pi}^{-1} C_p]) = \chi^2(d - N_{\text{eff}})$ . This approximation is exact for all parameter space directions that are data dominated  $C_{\Pi}^{-1} C_p \rightarrow 0$  or completely prior dominated  $C_{\Pi}^{-1} C_p \rightarrow 1$  and approximated for cases in between.

As long as the number of partially constrained directions in parameter space remains small compared with the total number of d.o.f., the approximation works very well. When the number of partially constrained directions is large, Eq. (A5) systematically underestimates the statistical

significance of results. In such cases, however, it is likely that the distributions that are being considered are highly non-Gaussian, so results should be interpreted with caution anyway.

### APPENDIX C: PROOFS OF SEC. II E

In this appendix we report the proofs of the results contained in Sec. II E.

We start by discussing the statistics of the ratio between the maximum likelihoods of two experiments and their joint maximum likelihood. In data space this can be written as

$$-2\Delta \ln \mathcal{L}_{\text{max}} \equiv -2 \ln \mathcal{L}_{\text{max}}^{12} + 2 \ln \mathcal{L}_{\text{max}}^1 + 2 \ln \mathcal{L}_{\text{max}}^2. \quad (\text{C1})$$

Since we assume that the two data sets are uncorrelated, the data independent part cancels so that  $-2\Delta \ln \mathcal{L}_{\text{max}}$  defines a quadratic form in data space:

$$\begin{aligned} Q_{\text{DML}} &= X_{12}^T (\mathbb{I}_{12} - \mathbb{P}_{12})^T \Sigma_{12}^{-1} (\mathbb{I}_{12} - \mathbb{P}_{12}) X_{12} \\ &\quad - X_1^T (\mathbb{I}_1 - \mathbb{P}_1)^T \Sigma_1^{-1} (\mathbb{I}_1 - \mathbb{P}_1) X_1 \\ &\quad - X_2^T (\mathbb{I}_2 - \mathbb{P}_2)^T \Sigma_2^{-1} (\mathbb{I}_2 - \mathbb{P}_2) X_2, \end{aligned} \quad (\text{C2})$$

where  $X_{12} \equiv x_{12} - \hat{m}_{12}$ ,  $X_1 \equiv x_1 - \hat{m}_1$ ,  $X_2 \equiv x_2 - \hat{m}_2$ , and we assume that the two data sets are independent so that  $\Sigma_{12} = \text{diag}(\Sigma_1, \Sigma_2)$ .

The projector  $\mathbb{P}_{12}$  takes data realizations of the joint data set  $(x_1, x_2)$  and projects them on the model tangent space. It can be explicitly written as

$$\begin{aligned} \mathbb{P}_{12} &= M_{12} (M_{12}^T \Sigma_{12}^{-1} M_{12})^{-1} M_{12}^T \Sigma_{12}^{-1} \\ &= \begin{pmatrix} M_1 C_{12} M_1^T \Sigma_1^{-1} & M_1 C_{12} M_2^T \Sigma_2^{-1} \\ M_2 C_{12} M_1^T \Sigma_1^{-1} & M_2 C_{12} M_2^T \Sigma_2^{-1} \end{pmatrix}, \end{aligned} \quad (\text{C3})$$

to verify that it is a projector  $\mathbb{P}_{12}^2 = \mathbb{P}_{12}$  and that it leaves the tangent space of the model invariant  $\mathbb{P}_{12} M_{12} = M_{12}$ . Notice that the projector of the joint data set cannot be written as the direct sum of the two single projectors  $\text{diag}(\mathbb{P}_1, \mathbb{P}_2)$ , but it can be shown by direct calculation that they commute:

$$\mathbb{P}_{12} \text{diag}(\mathbb{P}_1, \mathbb{P}_2) = \text{diag}(\mathbb{P}_1, \mathbb{P}_2) \mathbb{P}_{12} = \mathbb{P}_{12}. \quad (\text{C4})$$

This implies that the subspace that  $\mathbb{P}_{12}$  spans is contained in the subspace that  $\text{diag}(\mathbb{P}_1, \mathbb{P}_2)$  spans,  $\mathbb{P}_{12} \subset \text{diag}(\mathbb{P}_1, \mathbb{P}_2)$ . Conversely,  $\text{diag}(\mathbb{I} - \mathbb{P}_1, \mathbb{I} - \mathbb{P}_2) \subset \mathbb{I} - \mathbb{P}_{12}$ .

Now, by noticing that  $\hat{m}_{12} = (\hat{m}_1, \hat{m}_2)$ , we can write Eq. (C2) as

$$Q_{\text{DML}} = X_{12}^T \left[ (\mathbb{I}_{12} - \mathbb{P}_{12})^T \Sigma_{12}^{-1} (\mathbb{I}_{12} - \mathbb{P}_{12}) - \begin{pmatrix} (\mathbb{I}_1 - \mathbb{P}_1)^T \Sigma_1^{-1} (\mathbb{I}_1 - \mathbb{P}_1) & \mathbb{O} \\ \mathbb{O} & (\mathbb{I}_2 - \mathbb{P}_2)^T \Sigma_2^{-1} (\mathbb{I}_2 - \mathbb{P}_2) \end{pmatrix} \right] X_{12}, \quad (\text{C5})$$

where  $X_{12}$  is distributed according to the evidence of the joint data set.

In the delta prior case the joint evidence is given by  $\mathcal{E}_{12} = \mathcal{N}_{d_1+d_2}(x_{12}; \hat{m}_{12}, \Sigma_{12})$ , while the uninformative flat prior case is  $\mathcal{E}_{12} = \mathcal{N}_{d_1+d_2}(x_{12}; \hat{m}_{12}, [(\mathbb{I} - \mathbb{P}_{12})^T \Sigma_{12}^{-1} (\mathbb{I} - \mathbb{P}_{12})]^{-1})$  and the Gaussian case has  $\mathcal{E}_{12} = \mathcal{N}_{d_{12}}(x_{12}; m_{12\Pi}, \Sigma_{12} + M_{12} \mathcal{C}_\Pi M_{12}^T)$ .

When we compute the eigenvalues of the product of the matrix defining the quadratic form in Eq. (C5) and the covariance of the joint data draws, it is sufficient to notice that in all three cases the projector would null everything along the model joint tangent space and so would  $\text{diag}(\mathbb{I} - \mathbb{P}_1, \mathbb{I} - \mathbb{P}_2)$  since it is contained in  $\mathbb{I} - \mathbb{P}_{12}$ . We can then apply Theorem (5.1.6) in [67] to show that Eq. (C5) is distributed as

$$\begin{aligned} \chi^2(\text{rank}(\mathbb{I} - \mathbb{P}_{12}) - \text{rank}(\mathbb{I} - \mathbb{P}_1) - \text{rank}(\mathbb{I} - \mathbb{P}_2)) \\ = \chi^2(d_{12} - N_{12} - d_1 + N_1 - d_2 + N_2) \\ = \chi^2(N_1 + N_2 - N_{12}), \end{aligned} \quad (\text{C6})$$

where we used the fact that the rank of a block diagonal matrix is the sum of the ranks of the diagonal blocks, having allowed different data sets to have different irrelevant parameters and noticing that  $d_{12} = d_1 + d_2$ .

The same result can be obtained by bringing Eq. (C2) in parameter space to show that

$$Q_{\text{DML}} = (\theta_{\text{ML}}^2 - \theta_{\text{ML}}^1)^T (\mathcal{C}_1 + \mathcal{C}_2)^{-1} (\theta_{\text{ML}}^2 - \theta_{\text{ML}}^1), \quad (\text{C7})$$

and considering  $\theta_{\text{ML}}^1$  and  $\theta_{\text{ML}}^2$  to be drawn independently from a Gaussian distribution with covariance  $\mathcal{C}_1$  and  $\mathcal{C}_2$ , respectively.

We now consider the distribution of the evidence ratio. In the case of delta priors the distribution is trivial:

$$\begin{aligned} -2\Delta \ln \mathcal{E} &\equiv -2 \ln \mathcal{E}_{12} + 2 \ln \mathcal{E}_1 + 2 \ln \mathcal{E}_2 \\ &= -2 \ln \mathcal{L}_{12}(\theta_\Pi) + 2 \ln \mathcal{L}_1(\theta_\Pi) + 2 \ln \mathcal{L}_2(\theta_\Pi) \\ &= 0. \end{aligned} \quad (\text{C8})$$

Notice that we assume that the prior is the same for the analysis of the joint data set and the single data sets. If this is not the case and the prior is changed between the analysis of different data sets, the distributions of this appendix would be more complicated and, in general, noncentral.

In the uninformative flat prior case the distribution of the data dependent part of the evidence ratio follows that of the maximum likelihood  $Q_{\text{DE}} = Q_{\text{DML}}$ .

In the Gaussian case the distribution is more complicated and can be written starting from

$$\begin{aligned} -2\Delta \ln \mathcal{E} &= -2 \ln \mathcal{L}_{12}(\theta_p^{12}) + 2 \ln \mathcal{L}_1(\theta_p^1) + 2 \ln \mathcal{L}_2(\theta_p^2) \\ &\quad - 2 \ln \frac{\Pi_{12}(\theta_p^{12})}{\Pi_{12}^{\max}} + 2 \ln \frac{\Pi_1(\theta_p^1)}{\Pi_1^{\max}} + 2 \ln \frac{\Pi_2(\theta_p^2)}{\Pi_2^{\max}} \\ &\quad + (N_1 + N_2 - N_{12}) \ln(2\pi) \\ &\quad + 2 \ln \frac{V_\Pi^{12}}{V_\Pi^1 V_\Pi^2} + \ln \frac{|\mathcal{C}_{p1}| |\mathcal{C}_{p2}|}{|\mathcal{C}_{p12}|}. \end{aligned} \quad (\text{C9})$$

The data dependent part of Eq. (C9) defines a quadratic form in data space:

$$\begin{aligned} Q_{\text{DE}} &\equiv X_{12}^T A X_{12} \\ &= X_{12}^T \left[ (\Sigma_{12} + M_{12} \mathcal{C}_\Pi M_{12}^T)^{-1} \right. \\ &\quad \left. - \begin{pmatrix} (\Sigma_1 + M_1 \mathcal{C}_\Pi M_1^T)^{-1} & \mathbb{O} \\ \mathbb{O} & (\Sigma_2 + M_2 \mathcal{C}_\Pi M_2^T)^{-1} \end{pmatrix} \right] X_{12}, \end{aligned} \quad (\text{C10})$$

where  $X_{12} \equiv x_{12} - m_{12}(\theta_\Pi)$  and the covariance of the joint data draws is explicitly given by

$$\begin{aligned} \Sigma_{12} + M_{12} \mathcal{C}_\Pi M_{12}^T \\ = \begin{pmatrix} \Sigma_1 & \mathbb{O} \\ \mathbb{O} & \Sigma_2 \end{pmatrix} + \begin{pmatrix} M_1 \mathcal{C}_\Pi M_1^T & M_1 \mathcal{C}_\Pi M_2^T \\ M_2 \mathcal{C}_\Pi M_1^T & M_2 \mathcal{C}_\Pi M_2^T \end{pmatrix}. \end{aligned} \quad (\text{C11})$$

By direct computation of the product between the two matrices,

$$\begin{aligned} \lambda &= \text{eigenval}[A(\Sigma_{12} + M_{12} \mathcal{C}_\Pi M_{12}^T)] \\ &= \pm \sqrt{\text{eigenval}[(\mathbb{I} - \mathcal{C}_\Pi^{-1} \mathcal{C}_{p2})(\mathbb{I} - \mathcal{C}_\Pi^{-1} \mathcal{C}_{p1})]}. \end{aligned} \quad (\text{C12})$$

This means that  $Q_{\text{DE}}$  does not define a positive-definite quadratic form. The expression for the probability density of indefinite quadratic forms can be found in [67]. Here we notice that the decomposition of  $Q_{\text{DE}}$  can be written as

$$Q_{\text{DE}} = \sum_{j=1}^{2N} \lambda_j X_j^2 = \sum_{i=1}^N \lambda_i X_i^2 + \sum_{j=1}^N \lambda_j Y_j^2 = \sum_{i=1}^N |\lambda_i| (X_i^2 - Y_i^2) \quad (\text{C13})$$

where both  $X$  and  $Y$  are independently distributed normal variables with zero mean and unit variance and we exploit the fact that the evidence ratio has two equal eigenvalues of opposite sign. It is now possible to show, by matching the moment-generating function, that the evidence ratio for Gaussian priors is distributed as a sum of independent variance-gamma distributed variables. Summing all the eigenvalues shows that the distribution is zero mean, and in the limit where  $\mathcal{C}_{p1}, \mathcal{C}_{p2} \rightarrow \mathcal{C}_\Pi$ , it recovers delta prior results.

We now turn to the statistics of the ratio of likelihoods at maximum posterior. In both the delta and uninformative flat prior cases, this follows the statistics of the data independent part of the evidence ratio. The Gaussian case instead is given by

$$\begin{aligned} Q_{\text{DMAP}} &\equiv -2\Delta \ln \mathcal{L}_p \\ &= -2 \ln \mathcal{L}_{12}(\theta_{p12}) + 2 \ln \mathcal{L}_1(\theta_{p1}) + 2 \ln \mathcal{L}_2(\theta_{p2}), \end{aligned} \quad (\text{C14})$$

which defines a quadratic form in data space that can be easily written with Eq. (B9). This quadratic form is central and positive definite and, as before, it can be written as the difference of two quadratic forms. By direct calculation it can be shown that its eigenvalues are given by

$$\lambda = \text{eigenval} \begin{pmatrix} A & B \\ C & D \end{pmatrix}, \quad (\text{C15})$$

where

$$\begin{aligned} A &= \mathbb{I} - C_{\Pi}^{-1} C_{p1} - C_{p1}^{-1} C_{p12} + C_{\Pi}^{-1} C_{p12}, \\ B &= C_{p1}^{-1} C_{p12} - C_{\Pi}^{-1} C_{p12} + C_{\Pi}^{-1} C_{p1} - C_{\Pi}^{-1} C_{p1} C_{\Pi}^{-1} C_{p1}, \\ C &= C_{p2}^{-1} C_{p12} - C_{\Pi}^{-1} C_{p12} + C_{\Pi}^{-1} C_{p2} - C_{\Pi}^{-1} C_{p2} C_{\Pi}^{-1} C_{p2}, \\ D &= \mathbb{I} - C_{\Pi}^{-1} C_{p2} - C_{p2}^{-1} C_{p12} + C_{\Pi}^{-1} C_{p12}. \end{aligned} \quad (\text{C16})$$

The quadratic form defining  $Q_{\text{DMAP}}$  is positive definite, so its eigenvalues are all positive and they recover the two limits of uninformative flat priors and delta priors. Equation (C15) can be used if one wants to compute the exact distribution of  $Q_{\text{DMAP}}$ . On the other hand, it is convenient to approximate this distribution by a chi-squared distribution, as discussed in Appendix A, with

$$\sum \lambda = N_{\text{eff}}^1 + N_{\text{eff}}^2 - N_{\text{eff}}^{12} \quad (\text{C17})$$

d.o.f. since this would generally down-weight the contribution of partially constrained parameter space directions.

#### APPENDIX D: OPTIMAL QUADRATIC FORMS

Given that there seems to be no general rule to select the matrix defining the quadratic forms in Eq. (A1), in this appendix we discuss how to choose a quadratic form that is ‘‘optimal’’ in some sense. For this purpose it is worth noticing that a quadratic form defined by Eq. (A1), if rescaled by a positive quantity  $\alpha$ , would give the same statistical significance of results, i.e.,  $P(Q > Q^{\text{obs}}) = P(\alpha Q > \alpha Q^{\text{obs}})$ . This means that, for our purpose, the quadratic forms defined by  $A$  and  $\alpha A$  are equivalent.

As a consequence, all quadratic forms, in one dimension, give the same statistical significance. This explains why the rule-of-thumb difference in mean, discussed in Sec. II F,

when it can be applied and is representative of the full tension, works so well. In one dimension all parameter quadratic forms are equivalent, and the rule of thumb is the one for which we can immediately read the statistical significance.

In multiple dimensions the same does not apply and, apart from a constant rescaling, different choices of the matrix  $A$  would lead to a different statistical significance. We follow [70] in looking for a quadratic form that is optimal according to some criterion. Since the quadratic form defined by Eq. (A1) is central, i.e.,  $\langle X \rangle = 0$ , all the cumulants of the quadratic form pdfs are given by

$$\kappa_m = (m-1)! \text{Tr}[(A\Sigma)^m]. \quad (\text{D1})$$

Starting from this, one can compute all moments.

The mean is given by  $\mu_1 = \kappa_1 = \text{Tr}[(A\Sigma)]$  and the variance by  $\mu_2 = \kappa_2 + \kappa_1^2 = 2\text{Tr}[(A\Sigma)^2]$ . For all other moments we refer the reader to [67].

We define the optimal parameter quadratic form to minimize the variance and all other moments. This can be achieved if the quadratic form minimizes all cumulants. The trivial solution to our optimization problem is  $A = 0$  which is not particularly informative and can be excluded from the solution to our problem. We can look for other solutions by demanding that the quadratic form should not have zero mean. Since, for our purposes, all quadratic forms that are just rescaled by a constant are equivalent, we can assume that they all have the same mean, without loss of generality. Thus we need to minimize

$$f(A) = (m-1)! \text{Tr}[(A\Sigma)^m] + \alpha [\text{Tr}[(A\Sigma)] - \kappa_1], \quad (\text{D2})$$

over all positive matrices  $A$  and for all cumulants greater than one. Notice that we implemented the constraint on the average as a Lagrange multiplier  $\alpha$ . Taking the derivative of  $f$  with respect to the Lagrange multiplier would give us the finite mean constraint. Writing the trace in terms of the  $\lambda_q$  eigenvalues of  $A\Sigma$ , we have

$$f(A) = (m-1)! \sum_{q=1}^N \lambda_q^m + \alpha \left[ \sum_{q=1}^N \lambda_q - \kappa_1 \right], \quad (\text{D3})$$

which has to be minimized over all positive, nonzero  $\lambda_q$ . Setting  $\partial f / \partial \lambda_q = 0$  we can easily find that  $f$  is minimized when all  $\lambda_q$  are the same, so  $A_{\text{opt}} \Sigma = \mathbb{I}$ , which gives

$$A_{\text{opt}} = \Sigma^{-1}. \quad (\text{D4})$$

That is, in multiple dimensions, the quadratic form that minimizes the variance and all moments is the inverse covariance one.

To have an intuition of this result, let us consider a two-dimensional space and two quadratic forms  $Q_{\text{opt}}$  and  $Q_2$ .



The first one is the optimal, inverse covariance weighted, for which  $\kappa_m^{\text{opt}} = 2(m-1)!$ . The second one has a direction rescaled, with respect to the inverse covariance, by a positive constant  $\lambda$  so that all cumulants are given by  $\kappa_m^2 = (\lambda^m + 1)(m-1)!$ . We could now say that we can make the moments of the second form arbitrarily small by properly choosing  $\lambda$ , but this does not take into account invariance under rescaling. We thus rescale the second quadratic form by the ratio of the two averages, in this case  $2/(\lambda+1)$ , so that all cumulants are given by  $\kappa_m^2 = 2^m(\lambda^m + 1)/(\lambda+1)^m(m-1)!$ , and we can see that the second quadratic form has cumulants that are always bigger than the first one.

We can now ask what happens to the statistical significance of the reported results, in our simplified example. Let us suppose that we have two uncorrelated parameters and that the observed difference between them is given by  $\Delta\theta = n(\sigma_1^2, \sigma_2^2)^T$  so that  $Q_{\text{opt}} = 2n^2$  and  $Q_2 = n^2(\lambda+1)$ . The eigenvalues of  $A\Sigma$  in the first case are just (1,1), and  $Q_{\text{opt}}$  is chi-squared distributed with 2 d.o.f. The eigenvalues of  $A\Sigma$  in the second case are given by  $(\lambda, 1)$  so that  $Q_2$  is the sum of a Gamma distributed and a chi-squared distributed variable. Both distributions can be easily integrated numerically to show that statistical significance is the same for  $\lambda \rightarrow 1/\lambda$  and that  $Q_2$ , for all values of positive  $\lambda$ , will underestimate both confirmation biases and tensions. This is why we picked, as a criterion for defining an optimal quadratic form, the minimization of the moments higher than the mean, as this is related to a lower probability of extreme events and would thus make our concordance or discordance estimator more sensitive to the presence of tensions that might be hidden by other estimators.

## APPENDIX E: GAUSSIAN APPROXIMATION OF MCMC POSTERIOR

In this section we describe how we approximate the posterior obtained from MCMC sampling with a multivariate Gaussian. This approximation is useful when computing some of the statistical results of this paper and can be obtained by properly accounting for all the factors that are usually neglected when performing the sampling. While we adopt CosmoMC [59] conventions, similar results would apply for other samplers.

The un-normalized posterior that CosmoMC produces can be approximated by

$$\ln \mathcal{P} = \ln \mathcal{L}(\tilde{\theta}) - \ln V_{\Pi} - \frac{1}{2}(\theta - \tilde{\theta})^T \mathcal{C}_{\tilde{\theta}}^{-1}(\theta - \tilde{\theta}) + \ln \frac{\Pi(\tilde{\theta})}{\Pi_{\text{max}}}, \quad (\text{E1})$$

where  $\tilde{\theta}$  is the parameter around which the expansion is performed,  $\mathcal{C}_{\tilde{\theta}} = \langle (\theta - \tilde{\theta})(\theta - \tilde{\theta}) \rangle_{\theta}$  the covariance of the parameter samples around that point, and  $\mathcal{L}(\tilde{\theta})$  the likelihood at that point. We also included a prior term that takes

into account that some parameters, i.e., some nuisance parameters, might have Gaussian priors. There are mainly three points that we can use to define our Gaussian approximation: the parameters' mean, the maximum posterior parameters, and the parameters from maximum posterior in the MCMC samples.

It is possible to define the best Gaussian approximation by computing the KL divergence [44] between the Gaussian approximation and the full posterior for the three expansion points and select the approximation that has the smallest difference in information content with respect to the full posterior.

Having  $N_s$  samples  $\theta_i$  of the parameter posterior, the KL divergence  $D_{\text{KL}}$  between the (normalized) full posterior  $P_{\text{full}}$  and one of the Gaussian approximations  $P_G$  can be written as

$$D_{\text{KL}}(P_{\text{full}}||P_G) \equiv \int P_{\text{full}}(\theta) \ln \left[ \frac{P_{\text{full}}(\theta)}{P_G(\theta)} \right] d\theta \\ \simeq \frac{1}{N_s} \sum_{i=1}^{N_s} \ln \left[ \frac{\mathcal{P}_{\text{full}}(\theta_i)}{\mathcal{P}_G(\theta_i)} \right] + C \quad (\text{E2})$$

where the samples  $\theta_i$  are drawn from  $P_{\text{full}}$  and  $\mathcal{P}_G$  is easily computed with Eq. (E1). The normalization constant  $C$  is the ratio between the evidence of the full posterior and the evidence of the Gaussian approximation  $C = \ln(\mathcal{E}_G/\mathcal{E}_{\text{full}})$ . Notice that, for the purpose of comparing performances of different Gaussian approximations, there is no need for an accurate estimate of the full posterior evidence. Equation (E2) is trivially generalized to weighted samples.

Given the Gaussian approximation of the MCMC posterior, we can compute the evidence as

$$\ln \mathcal{E} = \ln \mathcal{L}(\tilde{\theta}) - \ln V_{\Pi} + \frac{N}{2} \ln(2\pi) + \frac{1}{2} \ln |\mathcal{C}_{\tilde{\theta}}| + \ln \frac{\Pi(\tilde{\theta})}{\Pi_{\text{max}}}, \quad (\text{E3})$$

which is usually called the Laplace or saddle-point approximation, and we account for Gaussian priors on some parameters.

In general, one cannot test whether a distribution is truly Gaussian, but we can perform several null tests to warn us against non-Gaussianities in parameter space. In particular, we check

- (i) that the marginalized 1D posterior is visually well approximated by the marginalized 1D Gaussian approximation, for all constrained parameters;
- (ii) that the best fit obtained by explicitly minimizing the data residuals, the best fit from MCMC samples, and the mean do not show relevant shifts in units of their covariance, for all the constrained parameters.

Whenever one of the Gaussian approximations fails to comply with these requirements, we flag the results and express caution in interpreting them.

## APPENDIX F: PARAMETER PRIORS

The estimate of most of the results in the main text depends on the prior, especially in quantifying how many directions a data set constrains compared to it. In this appendix we discuss how we approximate the prior distribution.

In many cases these are informative flat priors, and we approximate them with Gaussian priors of the same covariance to compute the statistics discussed in the main text while taking into account that explicit evaluations of the prior would give  $\Pi(\theta) = 1/V_{\Pi}$ . In order to make these approximations, we sample the parameter space for the various flat priors listed in Table III to obtain the covariance and volume. To make our approach more efficient and transparent, we do not sample Gaussian priors but rather account for their variance analytically as described below.

While approximate, this approach works very well in practice. It is faster and computationally less expensive than resampling the parameter posterior and is less noisy with respect to the results obtained by importance sampling the MCMC samples with a Gaussian prior. Its robustness stems from the fact that the most important information that we need to extract from the prior is whether a parameter is constrained or not. Other situations that fall in between are not usually relevant to the end results.

In addition to the parameters in Table III, the likelihood of most experiments will add nuisance parameters describing systematic effects. We include them by analytically augmenting the prior covariance matrix and volume. In the case of flat priors on nuisance parameters, we fill the corresponding entrance in the prior covariance with  $C = (\theta_{\max} - \theta_{\min})^2/12$ , which corresponds to the variance of the flat distribution between  $\theta_{\max}$  and  $\theta_{\min}$ . Some nuisance parameters, noticeably some foreground parameters in CMB observations, have tight uncorrelated Gaussian priors. In this case the corresponding prior covariance entrance can be easily read from the parameter prior variance.

The prior on the base  $\Lambda$ CDM parameters deserves a closer look. We choose a parameter basis that has  $100\theta_{MC}$  instead of the Hubble constant, but we also impose physical constraints on matter density  $\Omega_m$  to be positive definite and

TABLE III. Nominal flat priors on the six cosmological parameters of the  $\Lambda$ CDM model used for all analyses in this work.

Parameter	Prior range
$\Omega_b h^2$	[0.005, 0.1]
$\Omega_c h^2$	[0.001, 0.99]
$100\theta_{MC}$	[0.5, 10]
$\tau$	[0.01, 0.8]
$n_s$	[0.8, 1.2]
$\ln(10^{10} A_s)$	[2, 4]

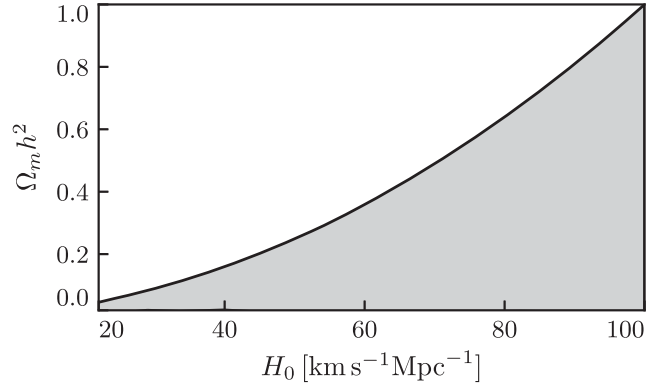


FIG. 6. The two-dimensional marginalized prior distribution of physical matter density  $\Omega_m h^2$  and the Hubble constant  $H_0$ . The shaded area shows parameter choices that satisfy the constraints  $0 \leq \Omega_m \leq 1$  and  $20 \leq H_0 [\text{km s}^{-1} \text{Mpc}^{-1}] \leq 100$ . This projection highlights that priors on derived quantities leave the prior distribution flat but introduce a nontrivial shape to the boundaries of the prior volume.

smaller than unity. Furthermore, we impose a prior cut on the Hubble constant to be between  $20 \text{ km s}^{-1} \text{Mpc}^{-1}$  and  $100 \text{ km s}^{-1} \text{Mpc}^{-1}$ .

These two joint boundary constraints on derived quantities make the prior volume nontrivial in shape and the prior covariance matrix nondiagonal in the base parameters. In Fig. 6 we show the 2D marginalized distribution of the prior to show that the two constraints on derived parameters are still locally flat in the interior, but the shape of the boundary induces a covariance between the parameters.

When we marginalize these flat but shaped priors to obtain the marginal distributions in 1D on the three  $\Lambda$ CDM background parameters, we obtain Fig. 7. As we can see the prior distribution for  $\Omega_b h^2$  and  $\Omega_c h^2$  seems to have curvature on the same scale of the prior range while the prior on  $100\theta_{MC}$  is more constraining. This shape of the 1D prior will not influence the posterior distribution for

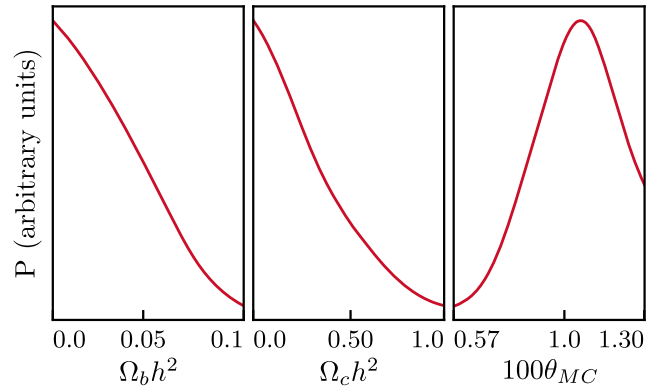


FIG. 7. The one-dimensional marginalized prior distribution on the background  $\Lambda$ CDM parameters. Notice that the actual range of  $100\theta_{MC}$  is much smaller than the nominal one reported in Table III.

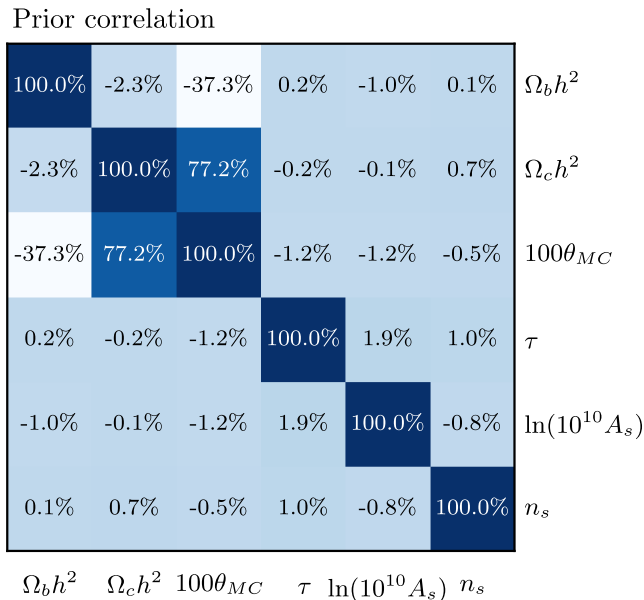


FIG. 8. The correlation between the base  $\Lambda$ CDM parameter priors, as obtained from the prior MCMC samples. Notice the correlation between background parameters induced by the boundaries of the prior volume.

constraining data sets as the prior is locally flat, but it does change the parameter ranges and combinations out to which the prior influences weaker data constraints. In particular, the range of  $100\theta_{MC}$  is modified, with respect to its face value in Table III, and not taking that into account would lead to wrong d.o.f. counting, for data sets that do not constrain it.

Moreover, we show in Fig. 8, the prior correlation between different parameters to highlight that the prior on the background parameters are also correlated because of the nontrivial shape induced by priors on derived quantities. This correlation is also important when judging parameter shifts and counting d.o.f.

The remaining three parameters  $\{\tau, n_s, \ln A_s\}$  have flat distributions and no covariance between themselves or the other parameters. Their covariance is also well approximated by the covariance of the uniform distribution as  $C = (\theta_{\max} - \theta_{\min})^2/12$ . Small correlation values in Fig. 8 are due to the MCMC sampling.

In summary, throughout this work we use a Gaussian approximation to the prior for the six base parameters by using the covariance extracted from the prior MCMC samples.

TABLE IV. The rule-of-thumb difference in mean and 1D exact parameter shift estimators applied to different data sets and data set combinations. The second column indicates the parameter that is being used in the test, while the third and fourth columns report its value and error for the two data sets considered. The last two columns indicate the probability to exceed (P.T.E.) the tests and  $n_s^{\text{eff}}$ , as computed from the results of Sec. II F. All results that are higher than 95% and lower than 5% P.T.E. are highlighted as statistically significant confirmation bias and tension, respectively. This table contains mostly known results that we use as a benchmark for other concordance and discordance estimators.

Data set $D_1$ vs. $D_2$	Parameter	$D_1$ result	$D_2$ result	$P(T_1 > T_1^{\text{obs}})$	
				Rule of thumb	Exact 1D shift
<i>LRG vs WiggleZ</i>	$\Omega_m$	$0.212 \pm 0.043$	$0.37 \pm 0.11$	16.0% (1.4 $\sigma$ )	15.0% (1.4 $\sigma$ )
<i>SN vs BAO</i>	$\Omega_m$	$0.297 \pm 0.034$	$0.358 \pm 0.042$	26.0% (1.1 $\sigma$ )	28.0% (1.1 $\sigma$ )
<i>CFHTLenS vs KiDS</i>	$\sigma_8 \Omega_m^{0.5}$	$0.369 \pm 0.071$	$0.281 \pm 0.087$	43.0% (0.8 $\sigma$ )	43.0% (0.8 $\sigma$ )
<i>H vs HSL</i>	$H_0$	$73.0 \pm 1.7$	$72.3 \pm 2.6$	82.0% (1.3 $\sigma$ )	85.0% (1.4 $\sigma$ )
<i>CMB vs H0</i>	$H_0$	$67.25 \pm 0.73$	$73.0 \pm 1.5$	<b>0.073% (3.4 <math>\sigma</math>)</b>	<b>0.078% (3.4 <math>\sigma</math>)</b>
<i>CMB vs BG</i>	$\Omega_m$	$0.316 \pm 0.01$	$0.32 \pm 0.026$	87.0% (1.5 $\sigma$ )	87.0% (1.5 $\sigma$ )
<i>CMB vs LRG</i>	$\Omega_m$	$0.316 \pm 0.01$	$0.212 \pm 0.043$	<b>2.0% (2.3 <math>\sigma</math>)</b>	<b>4.5% (2.0 <math>\sigma</math>)</b>
<i>CMB vs GC</i>	$\Omega_m$	$0.316 \pm 0.01$	$0.31 \pm 0.075$	94.1% (1.9 $\sigma$ )	77.0% (1.2 $\sigma$ )
<i>CMB vs CFHTLenS</i>	$\sigma_8 \Omega_m^{0.5}$	$0.4595 \pm 0.0071$	$0.369 \pm 0.071$	20.0% (1.3 $\sigma$ )	20.0% (1.3 $\sigma$ )
<i>CMB vs KiDS</i>	$\sigma_8 \Omega_m^{0.5}$	$0.4595 \pm 0.0071$	$0.281 \pm 0.087$	<b>4.0% (2.1 <math>\sigma</math>)</b>	<b>2.1% (2.3 <math>\sigma</math>)</b>
<i>CMB vs WL</i>	$\sigma_8 \Omega_m^{0.5}$	$0.4595 \pm 0.0071$	$0.354 \pm 0.058$	7.1% (1.8 $\sigma$ )	6.7% (1.8 $\sigma$ )
<i>BG vs GC</i>	$\sigma_8 \Omega_m^{0.5}$	$0.448 \pm 0.03$	$0.35 \pm 0.1$	36.0% (0.9 $\sigma$ )	32.0% (1.0 $\sigma$ )
<i>BG vs GC</i>	$\Omega_m$	$0.32 \pm 0.026$	$0.31 \pm 0.075$	90.0% (1.6 $\sigma$ )	76.0% (1.2 $\sigma$ )
<i>BG vs WL</i>	$\sigma_8 \Omega_m^{0.5}$	$0.448 \pm 0.03$	$0.354 \pm 0.058$	15.0% (1.4 $\sigma$ )	16.0% (1.4 $\sigma$ )
<i>BG vs WL</i>	$\Omega_m$	$0.32 \pm 0.026$	$0.3 \pm 0.13$	86.0% (1.5 $\sigma$ )	70.0% (1.0 $\sigma$ )
<i>GC vs WL</i>	$\sigma_8 \Omega_m^{0.5}$	$0.35 \pm 0.1$	$0.354 \pm 0.058$	<b>96.9% (2.2 <math>\sigma</math>)</b>	92.3% (1.8 $\sigma$ )
<i>GC vs WL</i>	$\Omega_m$	$0.31 \pm 0.075$	$0.3 \pm 0.13$	93.1% (1.8 $\sigma$ )	83.0% (1.4 $\sigma$ )
<i>CMBTT vs lowl</i>	$\tau$	$0.137 \pm 0.035$	$0.067 \pm 0.021$	8.6% (1.7 $\sigma$ )	9.9% (1.6 $\sigma$ )
<i>CMBEE vs lowl</i>	$\tau$	$0.191 \pm 0.063$	$0.067 \pm 0.021$	6.1% (1.9 $\sigma$ )	8.8% (1.7 $\sigma$ )
<i>CMBTE vs lowl</i>	$\tau$	$0.094 \pm 0.057$	$0.067 \pm 0.021$	65.0% (0.9 $\sigma$ )	74.0% (1.1 $\sigma$ )
<i>CMBTTEEE vs lowl</i>	$\tau$	$0.115 \pm 0.026$	$0.067 \pm 0.021$	14.0% (1.5 $\sigma$ )	15.0% (1.4 $\sigma$ )

## APPENDIX G: TABLES OF RESULTS

In this appendix we report the full results of the application of the CDEs in Sec. II to cosmological data, in table format. Specifically, we report exact 1D parameter shifts,  $T_1$ , and the rule-of-thumb difference in mean, as the

Gaussian approximation of  $T_1$ , described in Secs. IIF and IV D, in Table IV; the likelihood at maximum posterior goodness of fit  $Q_{\text{MAP}}$ , described in Secs. IID and IV B, in Table V; the difference of log likelihoods at maximum posterior  $Q_{\text{DMAP}}$ , described in Secs. IIE and IV C, in

TABLE V. The likelihood at MAP goodness-of-fit estimator applied to different data sets and combinations. The second column reports the data likelihood at maximum posterior; the third the number of effective parameters  $N_{\text{eff}}$ , as estimated using Eq. (29); the fourth the number of nominal parameters  $N$ ; the fifth the number of data points  $N_{\text{data}} = d$ ; and the seventh the P.T.E. for  $Q_{\text{MAP}}^{\text{obs}}$  assuming our best estimate of the d.o.f.  $N_{\text{data}} - N_{\text{eff}}$ . Values higher than 95% or lower than 5% P.T.E. are highlighted as statistically significant confirmation bias and tension, respectively. The remaining columns list the P.T.E.s assuming the minimal DoF  $N_{\text{data}} - N$  and the maximal DoF  $N_{\text{data}}$ , which place conservative bounds on tension and confirmation, respectively.

Data set $\sigma$	$-2 \ln \mathcal{L}_{\text{MAP}}$	$N_{\text{eff}}$	$N$	$N_{\text{data}}$	$P(Q_{\text{MAP}} > Q_{\text{MAP}}^{\text{obs}})$		
					Min (DoF)	Best (DoF)	Max (DoF)
<i>CMBT</i>	757.6	14.3	21	765	57.0% (0.8 $\sigma$ )	42.0% (0.8 $\sigma$ )	36.0% (0.9 $\sigma$ )
<i>CMBEE</i>	739.8	8.1	13	762	71.0% (1.1 $\sigma$ )	64.0% (0.9 $\sigma$ )	59.0% (0.8 $\sigma$ )
<i>CMBTE</i>	924.6	7.9	15	762	0.0045% (4.1 $\sigma$ )	<b>0.0019% (4.3 <math>\sigma</math>)</b>	0.00089% (4.4 $\sigma$ )
<i>CMBL</i>	5.3	2.5	7	8	73.0% (1.1 $\sigma$ )	44.0% (0.8 $\sigma$ )	2.1% (2.3 $\sigma$ )
<i>CMBTTEEE</i>	2417.1	19.0	33	2289	3.1% (2.2 $\sigma$ )	<b>1.6% (2.4 <math>\sigma</math>)</b>	0.93% (2.6 $\sigma$ )
<i>SN</i>	695.1	3.0	8	740	88.0% (1.6 $\sigma$ )	86.0% (1.5 $\sigma$ )	83.0% (1.4 $\sigma$ )
<i>BAO</i>	5.4	3.1	6	11	90.9% (1.7 $\sigma$ )	70.0% (1.0 $\sigma$ )	37.0% (0.9 $\sigma$ )
<i>LRG</i>	4.1	2.5	6	14	99.49% (2.8 $\sigma$ )	<b>97.5% (2.2 <math>\sigma</math>)</b>	85.0% (1.4 $\sigma$ )
<i>WiggleZ</i>	189.5	1.9	6	196	62.0% (0.9 $\sigma$ )	58.0% (0.8 $\sigma$ )	50.0% (0.7 $\sigma$ )
<i>CFHTLenS</i>	86.8	1.8	7	56	0.52% (2.8 $\sigma$ )	<b>0.32% (2.9 <math>\sigma</math>)</b>	0.07% (3.4 $\sigma$ )
<i>KiDS</i>	58.4	1.8	7	30	0.14% (3.2 $\sigma$ )	<b>0.07% (3.4 <math>\sigma</math>)</b>	0.0064% (4.0 $\sigma$ )
<i>CMB</i>	2432.2	18.9	33	2297	2.5% (2.2 $\sigma$ )	<b>1.2% (2.5 <math>\sigma</math>)</b>	0.71% (2.7 $\sigma$ )
<i>BG</i>	702.2	5.0	8	751	90.0% (1.6 $\sigma$ )	87.0% (1.5 $\sigma$ )	86.0% (1.5 $\sigma$ )
<i>GC</i>	204.7	2.5	6	210	59.0% (0.8 $\sigma$ )	54.0% (0.7 $\sigma$ )	47.0% (0.7 $\sigma$ )
<i>WL</i>	146.5	2.7	8	86	0.0052% (4.0 $\sigma$ )	<b>0.0024% (4.2 <math>\sigma</math>)</b>	0.00044% (4.6 $\sigma$ )
<i>ALL</i>	3516.2	22.8	37	3345	1.9% (2.3 $\sigma$ )	<b>0.96% (2.6 <math>\sigma</math>)</b>	0.59% (2.8 $\sigma$ )

TABLE VI. Evidence ratio type estimators applied to different data set combinations. The first three columns report the number of effective parameters of the first, second, and joint data sets. The fourth column reports the number of effective parameters that both data sets constrain. The fifth column reports the observed value of the evidence ratio and the sixth one its expected value when averaged over data realizations of  $D_1 \cup D_2$ . The last column reports the significance of the observed value of the ratio of likelihoods at maximum posterior (DMAP), as estimated using the results of Sec. IIE. All results that are higher than 95% and lower than 5% P.T.E. are highlighted as statistically significant confirmation bias and tension, respectively.

Data set	$N_1^{\text{eff}}$	$N_2^{\text{eff}}$	$N_{12}^{\text{eff}}$	$\Delta N^{\text{eff}}$	$\log_{10} C$	$\langle \log_{10} C \rangle_{12}$	$P(Q_{\text{DMAP}} > Q_{\text{DMAP}}^{\text{obs}})$
<i>LRG vs WiggleZ</i>	2.5	1.9	2.5	1.8	1.0	3.0	<b>0.31% (3.0 <math>\sigma</math>)</b>
<i>SN vs BAO</i>	3.0	3.1	5.0	1.1	2.4	2.6	20.0% (1.3 $\sigma$ )
<i>CFHTLenS vs KiDS</i>	1.8	1.8	2.7	0.9	2.4	2.4	25.0% (1.2 $\sigma$ )
<i>CMBT vs CMBEE</i>	14.3	8.1	16.9	5.6	10.2	10.6	25.0% (1.2 $\sigma$ )
<i>CMBT vs CMBL</i>	14.3	2.5	14.3	2.5	3.3	4.8	<b>1.8% (2.4 <math>\sigma</math>)</b>
<i>CMBEE vs CMBL</i>	8.1	2.5	8.4	2.3	2.6	4.1	<b>1.4% (2.5 <math>\sigma</math>)</b>
<i>CMBTE vs CMBL</i>	7.9	2.5	8.0	2.4	3.9	4.2	18.0% (1.3 $\sigma$ )
<i>CMBTTEEE vs CMBL</i>	19.0	2.5	18.9	2.6	3.2	4.8	<b>1.3% (2.5 <math>\sigma</math>)</b>
<i>CMB vs BG</i>	18.9	5.0	21.0	3.0	4.2	3.8	75.0% (1.2 $\sigma$ )
<i>CMB vs GC</i>	18.9	2.5	18.9	2.6	2.2	3.5	<b>2.3% (2.3 <math>\sigma</math>)</b>
<i>CMB vs WL</i>	18.9	2.7	20.8	0.8	0.3	2.3	<b>0.1% (3.3 <math>\sigma</math>)</b>
<i>CMB vs H0</i>	18.9	1.4	19.0	1.3	-0.6	2.5	<b>0.088% (3.3 <math>\sigma</math>)</b>
<i>BG vs GC</i>	5.0	2.5	5.5	2.1	1.8	3.1	<b>2.3% (2.3 <math>\sigma</math>)</b>
<i>BG vs WL</i>	5.0	2.7	6.9	0.9	2.0	2.3	12.0% (1.6 $\sigma$ )
<i>GC vs WL</i>	2.5	2.7	4.3	0.9	1.6	2.9	<b>0.74% (2.7 <math>\sigma</math>)</b>
<i>GC vs H0</i>	2.5	1.4	3.2	0.7	1.3	1.6	9.7% (1.7 $\sigma$ )
<i>WL vs H0</i>	2.7	1.4	3.6	0.4	1.9	1.9	22.0% (1.2 $\sigma$ )

TABLE VII. The update difference in mean estimator,  $Q_{\text{UDM}}$ , applied to different data set combinations. The first column reports the observed value, as computed from Eq. (50). The second column is the number of effective KL parameters retained,  $N_{\text{KL}} = \langle Q_{\text{UDM}} \rangle_D$ , for which the second data set significantly improves constraints over the first one. The third column reports the significance of the observed value of the update difference in mean, as estimated using the results of Sec. II F. All results that are higher than 95% and lower than 5% P.T.E. are highlighted as statistically significant confirmation bias and tension, respectively. When  $N_{\text{KL}} = 0$ ,  $Q_{\text{UDM}} = 0$ , and we do not report statistical significance.

Data set	$Q_{\text{UDM}}$	$N_{\text{KL}}$	$P(Q_{\text{UDM}} > Q_{\text{UDM}}^{\text{obs}})$
<i>LRG vs WiggleZ</i>	5.5	1	<b>1.9%</b> ( <b>2.3</b> $\sigma$ )
<i>BAO vs SN</i>	1.0	1	33.0% (1.0 $\sigma$ )
<i>CFHTLenS vs KiDS</i>	0.1	1	75.0% (1.2 $\sigma$ )
<i>H vs HSL</i>	0.3	1	62.0% (0.9 $\sigma$ )
<i>CMBTT vs CMBL</i>	7.0	1	<b>0.82%</b> ( <b>2.6</b> $\sigma$ )
<i>CMBEE vs CMBL</i>	6.6	1	<b>1.0%</b> ( <b>2.6</b> $\sigma$ )
<i>CMBTE vs CMBL</i>	0.3	1	59.0% (0.8 $\sigma$ )
<i>CMBTTTEEE vs CMBL</i>	7.3	1	<b>0.68%</b> ( <b>2.7</b> $\sigma$ )
<i>lowl vs CMBTT</i>	3.9	1	<b>4.9%</b> ( <b>2.0</b> $\sigma$ )
<i>lowl vs CMBEE</i>	8.5	2	<b>1.4%</b> ( <b>2.4</b> $\sigma$ )
<i>lowl vs CMBTE</i>	3.2	2	20.0% (1.3 $\sigma$ )
<i>lowl vs CMBTTTEEE</i>	3.1	1	7.6% (1.8 $\sigma$ )
<i>lowl + CMBTT vs CMBL</i>	1.5	1	22.0% (1.2 $\sigma$ )
<i>lowl + CMBEE vs CMBL</i>	1.1	2	59.0% (0.8 $\sigma$ )
<i>lowl + CMBTE vs CMBL</i>	0.1	1	77.0% (1.2 $\sigma$ )
<i>lowl + CMBTTTEEE vs CMBL</i>	2.0	1	16.0% (1.4 $\sigma$ )
<i>lowl vs CMBTT + CMBL</i>	0.0	1	88.0% (1.6 $\sigma$ )
<i>lowl vs CMBEE + CMBL</i>	2.5	2	29.0% (1.1 $\sigma$ )
<i>lowl vs CMBTE + CMBL</i>	3.1	2	22.0% (1.2 $\sigma$ )
<i>CMB vs BG</i>	0.4	1	52.0% (0.7 $\sigma$ )
<i>CMB vs GC</i>	0.0	0	–
<i>CMB vs WL</i>	5.8	1	<b>1.6%</b> ( <b>2.4</b> $\sigma$ )
<i>CMB vs H0</i>	11.1	1	<b>0.087%</b> ( <b>3.3</b> $\sigma$ )
<i>lowl + CMB vs BG</i>	0.4	1	55.0% (0.8 $\sigma$ )
<i>lowl + CMB vs GC</i>	0.0	0	–
<i>lowl + CMB vs WL</i>	5.9	1	<b>1.5%</b> ( <b>2.4</b> $\sigma$ )
<i>lowl + CMB vs H0</i>	10.7	1	<b>0.11%</b> ( <b>3.3</b> $\sigma$ )
<i>BG vs GC</i>	0.0	0	–
<i>BG vs WL</i>	0.7	1	42.0% (0.8 $\sigma$ )
<i>BG vs H0</i>	0.4	1	55.0% (0.8 $\sigma$ )
<i>GC vs WL</i>	0.0	0	–
<i>GC vs H0</i>	0.6	2	72.0% (1.1 $\sigma$ )
<i>WL vs H0</i>	0.2	2	90.7% (1.7 $\sigma$ )

Table VI; the parameter update  $Q_{\text{UDM}}$ , described in Secs. II F and IV D, in Table VII.

In this appendix we also report probabilities ( $P$ ) in terms of an equivalent number of standard deviations ( $n_\sigma$ ). This should be interpreted as an effective definition corresponding to a Gaussian distribution:

$$n_\sigma^{\text{eff}}(P) \equiv \sqrt{2} \text{erf}^{-1}(1 - \min[P, 1 - P]), \quad (\text{G1})$$

where  $\text{erf}^{-1}$  is the inverse error function. Notice that by defining the correspondence with  $\min[P, 1 - P]$  instead of  $2 \min[P, 1 - P]$  as in Eq. (39), we are equating the tension and confirmation tails of the non-Gaussian CDE distribution separately to the sum of probabilities in the two tails of the Gaussian. As an example, a tension event with probability to exceed of  $P = 4.55\%$  would correspond to a “ $2\sigma$ ” significance. Note that  $n_\sigma^{\text{eff}}$  should not be confused with the number of standard deviations from the mean  $(Q^{\text{obs}} - \langle Q \rangle) / \sqrt{\text{Var}(Q)}$ .

- [1] S. Perlmutter *et al.* (Supernova Cosmology Project), *Astrophys. J.* **517**, 565 (1999).
- [2] A. G. Riess *et al.* (Supernova Search Team), *Astron. J.* **116**, 1009 (1998).
- [3] C. L. Bennett, D. Larson, J. L. Weiland, and G. Hinshaw, *Astrophys. J.* **794**, 135 (2014).
- [4] D. Larson, J. L. Weiland, G. Hinshaw, and C. L. Bennett, *Astrophys. J.* **801**, 9 (2015).
- [5] G. E. Addison, Y. Huang, D. J. Watts, C. L. Bennett, M. Halpern, G. Hinshaw, and J. L. Weiland, *Astrophys. J.* **818**, 132 (2016).
- [6] M. Raveri, *Phys. Rev. D* **93**, 043522 (2016).
- [7] S. Seehars, S. Grandis, A. Amara, and A. Refregier, *Phys. Rev. D* **93**, 103507 (2016).
- [8] G. E. Addison, D. J. Watts, C. L. Bennett, M. Halpern, G. Hinshaw, and J. L. Weiland, *Astrophys. J.* **853**, 119 (2018).
- [9] J. L. Weiland, K. Osumi, G. E. Addison, C. L. Bennett, D. J. Watts, M. Halpern, and G. Hinshaw, *Astrophys. J.* **863**, 161 (2018).
- [10] Y. Huang, G. E. Addison, J. L. Weiland, and C. L. Bennett, *Astrophys. J.* **869**, 38 (2018).
- [11] P. A. R. Ade *et al.* (Planck Collaboration), *Astron. Astrophys.* **594**, A13 (2016).
- [12] A. G. Riess *et al.*, *Astrophys. J.* **826**, 56 (2016).
- [13] T. M. C. Abbott *et al.* (DES Collaboration), *Phys. Rev. D* **98**, 043526 (2018).
- [14] H. Hildebrandt *et al.*, *Mon. Not. R. Astron. Soc.* **465**, 1454 (2017).
- [15] N. Aghanim *et al.* (Planck Collaboration), *Astron. Astrophys.* **607**, A95 (2017).
- [16] P. Motloch and W. Hu, *Phys. Rev. D* **97**, 103536 (2018).
- [17] R. Laureijs *et al.* (EUCLID Collaboration), arXiv:1110.3193.
- [18] P. A. Abell *et al.* (LSST Science and LSST Project), arXiv:0912.0201.
- [19] K. N. Abazajian *et al.* (CMB-S4 Collaboration), arXiv:1610.02743.
- [20] S. Seehars, A. Amara, A. Refregier, A. Paranjape, and J. Akeret, *Phys. Rev. D* **90**, 023533 (2014).
- [21] D. J. Spiegelhalter, N. G. Best, B. P. Carlin, and A. Van Der Linde, *J. R. Stat. Soc. Ser. B* **64**, 583 (2002).
- [22] M. Kunz, R. Trotta, and D. Parkinson, *Phys. Rev. D* **74**, 023503 (2006).
- [23] A. R. Liddle, *Mon. Not. R. Astron. Soc.* **377**, L74 (2007).
- [24] R. Trotta, *Contemp. Phys.* **49**, 71 (2008).
- [25] D. J. Spiegelhalter, N. G. Best, B. P. Carlin, and A. van der Linde, *J. R. Stat. Soc. Ser. B* **76**, 485 (2014).
- [26] P. Marshall, N. Rajguru, and A. Slosar, *Phys. Rev. D* **73**, 067302 (2006).
- [27] F. Feroz, B. C. Allanach, M. Hobson, S. S. AbdusSalam, R. Trotta, and A. M. Weber, *J. High Energy Phys.* **10** (2008) 064.
- [28] M. C. March, R. Trotta, L. Amendola, and D. Huterer, *Mon. Not. R. Astron. Soc.* **415**, 143 (2011).
- [29] L. Amendola, V. Marra, and M. Quartin, *Mon. Not. R. Astron. Soc.* **430**, 1867 (2013).
- [30] L. Verde, P. Protopapas, and R. Jimenez, *Phys. Dark Universe* **2**, 166 (2013).
- [31] J. Martin, C. Ringeval, R. Trotta, and V. Vennin, *Phys. Rev. D* **90**, 063501 (2014).
- [32] N. V. Karpenka, F. Feroz, and M. P. Hobson, *Mon. Not. R. Astron. Soc.* **449**, 2405 (2015).
- [33] S. Joudaki *et al.*, *Mon. Not. R. Astron. Soc.* **465**, 2033 (2017).
- [34] M. P. Hobson, S. L. Bridle, and O. Lahav, *Mon. Not. R. Astron. Soc.* **335**, 377 (2002).
- [35] J. L. Bernal and J. A. Peacock, *J. Cosmol. Astropart. Phys.* **07** (2018) 002.
- [36] H. Jeffreys, *Theory of Probability*, 3rd ed. (Clarendon Press, Oxford, England, 1961).
- [37] R. E. Kass and A. E. Raftery, *J. Am. Stat. Assoc.* **90**, 773 (1995).
- [38] S. Nesseris and J. Garcia-Bellido, *J. Cosmol. Astropart. Phys.* **08** (2013) 036.
- [39] S. Grandis, D. Rapetti, A. Saro, J. J. Mohr, and J. P. Dietrich, *Mon. Not. R. Astron. Soc.* **463**, 1416 (2016).
- [40] R. A. Battye, T. Charnock, and A. Moss, *Phys. Rev. D* **91**, 103508 (2015).
- [41] T. Charnock, R. A. Battye, and A. Moss, *Phys. Rev. D* **95**, 123535 (2017).
- [42] W. Lin and M. Ishak, *Phys. Rev. D* **96**, 023532 (2017).
- [43] S. Grandis, S. Seehars, A. Refregier, A. Amara, and A. Nicola, *J. Cosmol. Astropart. Phys.* **05** (2016) 034.
- [44] S. Kullback and R. A. Leibler, *Ann. Math. Stat.* **22**, 79 (1951).
- [45] A. J. Long, M. Raveri, W. Hu, and S. Dodelson, *Phys. Rev. D* **97**, 043510 (2018).
- [46] N. Aghanim *et al.* (Planck Collaboration), *Astron. Astrophys.* **594**, A11 (2016).
- [47] P. A. R. Ade *et al.* (Planck Collaboration), *Astron. Astrophys.* **594**, A15 (2016).
- [48] M. Betoule *et al.* (SDSS Collaboration), *Astron. Astrophys.* **568**, A22 (2014).
- [49] S. Alam *et al.* (BOSS Collaboration), *Mon. Not. R. Astron. Soc.* **470**, 2617 (2017).
- [50] A. J. Ross, L. Samushia, C. Howlett, W. J. Percival, A. Burden, and M. Manera, *Mon. Not. R. Astron. Soc.* **449**, 835 (2015).
- [51] F. Beutler, C. Blake, M. Colless, D. H. Jones, L. Staveley-Smith, L. Campbell, Q. Parker, W. Saunders, and F. Watson, *Mon. Not. R. Astron. Soc.* **416**, 3017 (2011).
- [52] M. Tegmark *et al.* (SDSS Collaboration), *Phys. Rev. D* **74**, 123507 (2006).
- [53] M. J. Drinkwater *et al.*, *Mon. Not. R. Astron. Soc.* **401**, 1429 (2010).
- [54] D. Parkinson *et al.*, *Phys. Rev. D* **86**, 103518 (2012).
- [55] C. Heymans *et al.*, *Mon. Not. R. Astron. Soc.* **432**, 2433 (2013).
- [56] 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, *Astrophys. J.* **855**, 136 (2018).
- [57] V. Bonvin *et al.*, *Mon. Not. R. Astron. Soc.* **465**, 4914 (2017).
- [58] A. Lewis, A. Challinor, and A. Lasenby, *Astrophys. J.* **538**, 473 (2000).
- [59] A. Lewis and S. Bridle, *Phys. Rev. D* **66**, 103511 (2002).
- [60] H. Gil-Marín *et al.*, *Mon. Not. R. Astron. Soc.* **460**, 4210 (2016).

- [61] R. E. Smith, J. A. Peacock, A. Jenkins, S. D. M. White, C. S. Frenk, F. R. Pearce, P. A. Thomas, G. Efstathiou, and H. M. P. Couchmann (VIRGO Consortium), *Mon. Not. R. Astron. Soc.* **341**, 1311 (2003).
- [62] R. Takahashi, M. Sato, T. Nishimichi, A. Taruya, and M. Oguri, *Astrophys. J.* **761**, 152 (2012).
- [63] A. Mead, J. Peacock, C. Heymans, S. Joudaki, and A. Heavens, *Mon. Not. R. Astron. Soc.* **454**, 1958 (2015).
- [64] A. Gelman and D. B. Rubin, *Stat. Sci.* **7**, 457 (1992).
- [65] L. An, S. Brooks, and A. Gelman, *J. Comput. Graph. Stat.* **7**, 434 (1998).
- [66] M. A. Troxel *et al.* (DES Collaboration), *Mon. Not. R. Astron. Soc.* **479**, 4998 (2018).
- [67] A. Mathai and S. Provost, *Quadratic Forms in Random Variables*, Statistics: A Series of Textbooks and Monographs (Taylor & Francis, London, 1992).
- [68] P. Duchesne and P. L. de Micheaux, *Computational Statistics and Data Analysis* **54**, 858 (2010).
- [69] P. B. Patnaik, *Biometrika* **37**, 78 (1950).
- [70] D. S. Grebenkov, *Phys. Rev. E* **84**, 031124 (2011).