Quantifying dimensionality: Bayesian cosmological model complexities

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

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

(Received 26 April 2019; published 15 July 2019)

We demonstrate a measure for the effective number of parameters constrained by a posterior distribution in the context of cosmology. In the same way that the mean of the Shannon information (i.e., the Kullback-Leibler divergence) provides a measure of the strength of constraint between prior and posterior, we show that the variance of the Shannon information gives a measure of dimensionality of constraint. We examine this quantity in a cosmological context, applying it to likelihoods derived from the cosmic microwave background, large-scale structure and supernovae data. We show that this measure of Bayesian model dimensionality compares favorably both analytically and numerically in a cosmological context with the existing measure of model complexity used in the literature.

DOI: 10.1103/PhysRevD.100.023512

I. INTRODUCTION

With the development of increasingly complex cosmological experiments, there has been a pressing need to understand model complexity in cosmology over the last few decades. The ΛCDM model of cosmology is surprisingly efficient in its parameterization of the background Universe and its fluctuations, needing only six parameters to successfully describe individual observations from all cosmological datasets [1]. However, different observational techniques constrain distinct combinations of these parameters. In addition, the systematic effects that affect various observations introduce a large number of additional nuisance parameters, around 20 in both the analyses of the Dark Energy Survey [2] and *Planck* Collaborations [3].

These nuisance parameters are not always chosen in an optimal way from the point of view of sampling, with known degeneracies between each other and with the cosmological parameters. This complicates quantifying the effective number of parameters constrained by the data. Examples of these parameter degeneracies are the degeneracy between the amplitude of the primordial power spectrum A_s and the optical depth to reionization τ in the combination $A_s e^{-2\tau}$ in temperature anisotropies of the cosmic microwave background (CMB) or the degeneracy between the intrinsic alignment amplitude and the parameter combination $S_8 \equiv \sigma_8 (\Omega_m/0.3)^{0.5}$ in cosmic shear measurements, where Ω_m is the present-day matter density and σ_8 is the present-day linear root-mean-square amplitude of the matter power spectrum [4–6].

Quantifying model complexity is important beyond increasing our understanding of the data. It is necessary to measure the effective number of constrained parameters to quantify tension between datasets. The authors found this in Handley and Lemos [7]. The preprint version of [7] used the Bayesian model complexity (BMC) introduced in Spiegelhalter *et al.* [8], which the authors found unsatisfactory. Motivated by this, in this work we examine an improved *Bayesian model dimensionality* (BMD) to quantify the effective number of dimensions constrained by the data. While the BMD measure has been introduced in the past by numerous authors [9–15], in this work we provide novel interpretations in terms of information theory and compare its performance with the BMC in a modern numerical cosmological context.

In Sec. II we introduce the notation and mathematical formalism and some of the relevant quantities such as the Bayesian evidence, the Shannon information and the Kullback-Leibler divergence. We also discuss some of the problems associated with principle component analyses (PCA), that have been used to quantify model complexity in cosmology in the past.

In Sec. III we discuss dimensionality in a Bayesian framework, describing the Bayesian model complexity of Spiegelhalter *et al.* [8] and introducing the Bayesian model dimensionality. We explain the usage of model dimensionality in the context of some analytical examples. Finally, in

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

Sec. IV, we apply Bayesian model dimensionality to real data, using four different cosmological datasets. We summarize our conclusions in Sec. V.

II. BACKGROUND

In this section we establish notation and introduce the key inference quantities used throughout this paper. For a more detailed account of Bayesian statistics, the reader is recommended the paper by Trotta [16] or the textbooks by MacKay [17] and Sivia and Skilling [18].

A. Bayes theorem

In the context of Bayesian inference, a predictive model $\mathcal{M}$ with free parameters θ can use data D to both provide constraints on the model parameters and infer the relative probability of the model via the Bayes theorem:

$$P(D|\theta) \times P(\theta) = P(\theta|D) \times P(D), \tag{1}$$

$$\mathcal{L} \times \pi = \mathcal{P} \times \mathcal{Z}, \tag{2}$$

which should be read as "likelihood times prior is posterior times evidence." While traditionally Bayes' theorem is rearranged in terms of the posterior $\mathcal{P} = \mathcal{L}\pi/\mathcal{Z}$, Eq. (2) is the form preferred by Skilling [11] and has since been used by other cosmologists [19]. In Skilling's form it emphasizes that the inputs to inference are the model, defined by the likelihood and the prior, while the outputs are the posterior and evidence, used for parameter estimation and model comparison, respectively.

B. Shannon information

The Shannon information [20] is defined as

$$\mathcal{I}(\theta) = \log \frac{\mathcal{P}(\theta)}{\pi(\theta)} \tag{3}$$

and is also known as the information content, self-information or surprisal of θ. The Shannon information represents the amount of information gained in nats (natural bits) about θ when moving from the prior to the posterior.

The Shannon information has the fundamental property that for independent parameters the information is additive:

$$\begin{aligned} \mathcal{P}(\theta_1, \theta_2) &= \mathcal{P}_1(\theta_1)\mathcal{P}_2(\theta_2), \\ \pi(\theta_1, \theta_2) &= \pi_1(\theta_1)\pi_2(\theta_2), \\ \Rightarrow \mathcal{I}(\theta_1, \theta_2) &= \mathcal{I}_1(\theta_1) + \mathcal{I}_2(\theta_2). \end{aligned} \tag{4}$$

Indeed it can be easily shown that the property of additivity defines Eq. (3) up to the base of the logarithm: i.e., if one wishes to define a measure of information provided by a posterior that is additive for independent parameters, then one is forced to use Eq. (3). Additivity is an important concept used throughout this paper, as it forms the underpinning of a measurable quantity. For more detail, see Skilling's chapter in [21].

C. Kullback-Leibler divergence

The Kullback-Leibler divergence [22] is defined as the average Shannon information over the posterior

$$\mathcal{D} = \int \mathcal{P}(\theta) \log \frac{\mathcal{P}(\theta)}{\pi(\theta)} \mathrm{d}\theta = \left\langle \log \frac{\mathcal{P}}{\pi} \right\rangle_{\mathcal{P}} = \langle \mathcal{I} \rangle_{\mathcal{P}} \tag{5}$$

and therefore quantifies in a Bayesian sense how much information is provided by the data D. Since the Shannon information is defined relative to the prior, the Kullback-Leibler divergence naturally has a strong prior dependency [7]. It has been widely utilised in cosmology [12,23–32] for a variety of analyses.

Since the Kullback-Leibler divergence is a linear function of the Shannon information, $\mathcal{D}$ is also measured in nats and is an additive quantity for independent parameters.

Posterior averages such as Eq. (5) in some cases can be numerically computed using samples generated by techniques such as Metropolis-Hastings [33], Gibbs sampling [34] or Hamiltonian Monte Carlo [35]. However, computation of the Kullback-Leibler divergence is numerically more challenging, since it requires knowledge of normalized posterior densities $\mathcal{P}$, or equivalently a computation of the evidence $\mathcal{Z}$, which requires more intensive techniques such as nested sampling [11].

D. Bayesian model complexity

While the Kullback-Leibler divergence provides a well-defined measure of the overall compression from posterior to prior, it marginalizes out any individual parameter information. As such, $\mathcal{D}$ tells us nothing of which parameters are providing us with information or, equally, how many parameters are being constrained by the data.

As a concrete example, consider the two posteriors illustrated in Fig. 1. In this case, both distributions have the same Kullback-Leibler divergence but give very different parameter constraints. For the first distribution, both parameters are well constrained. In the second distribution, the one-dimensional marginal distributions show that the first parameter is slightly constrained, while the second parameter is completely unconstrained and identical to the prior. The full two-dimensional distribution tells a different story, showing that both parameters are heavily correlated and that there is a strong constraint on a specific combination of parameters. In reality this is therefore a one-dimensional constraint that has been garbled across two parameters.

For the two-dimensional case in Fig. 1 we can by eye determine the number of constrained parameters, but in practical cosmological situations this is not possible. The cosmological parameter space of ΛCDM is six- (arguably

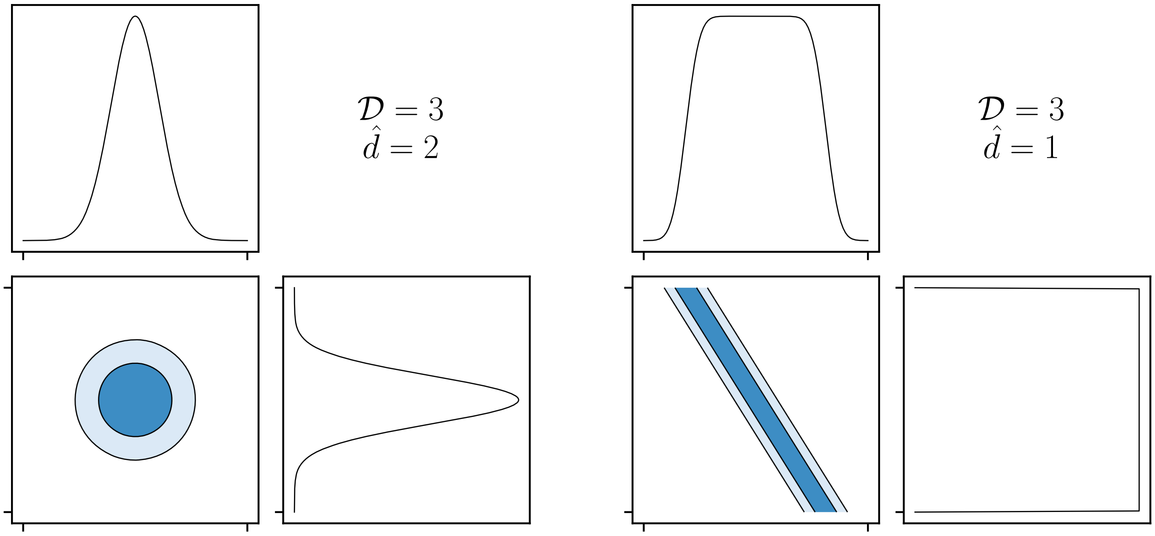

FIG. 1. Distributions with the same Kullback-Leibler divergence but differing dimensionalities. Both the right- and left-hand plots indicate two-dimensional probability distributions. In each plot, the lower left panel is a two-dimensional contour plot indicating the isoprobability contours enclosing 66% and 95% of the probability mass. The upper and lower right panels indicate the one-dimensional marginal probability distributions. There is an implicit uniform prior over the ranges indicated by the axis ticks.

seven-) dimensional [1], and modern likelihoods introduce a host of nuisance parameters to combat the influence of foregrounds and systematics. For example the Planck likelihood [36] is in total 21-dimensional, the DES likelihood [2] is 26-dimensional, and their combination 41-dimensional (Table I). While samples from the posterior distribution represent a near lossless compression of the information present in this distribution, it goes without saying that visualizing a 40-dimensional object is challenging. Triangle and corner plots [37] represent marginalized views of this information and can hide hidden correlations and constraints between three or more parameters. The fear is that one could misdiagnose a dataset that has powerful constraints if Fig. 1 occurred in higher dimensions. It would be helpful if there were a number d similar to the Kullback-Leibler divergence $\mathcal{D}$ which quantifies the effective number of constrained parameters.

TABLE I. Number of parameters sampled over in cosmological likelihoods. d_{Cosmo} is the number of cosmological parameters, d_{Nuis} is the number of nuisance parameters, and $d_{\mathrm{Total}} = d_{\mathrm{Cosmo}} + d_{\mathrm{Nuis}}$ is the total number. Note that we sample over the same six cosmological parameters for all likelihoods, even though we know that some likelihoods cannot constrain certain parameters. For the combinations of two likelihoods, the total number is $d_{\mathrm{Total}}^{A,B} = d_{\mathrm{Cosmo}} + d_{\mathrm{Nuis}}^{A} + d_{\mathrm{Nuis}}^{B}$.

Likelihood	d_{Cosmo}	d_{Nuis}	d_{Total}
SH$_0$ES	6	0	6
BOSS	6	0	6
DES	6	20	26
Planck	6	15	21

To this end, Spiegelhalter *et al.* [8,10] introduced the Bayesian model complexity, defined as

$$\frac{\hat{d}}{2} = \log\frac{P(\hat{\theta})}{\pi(\hat{\theta})} - \left\langle \log\frac{\mathcal{P}}{\pi}\right\rangle_{\mathcal{P}} = \mathcal{I}(\hat{\theta}) - \langle\mathcal{I}\rangle_{\mathcal{P}}. \tag{6}$$

In this case, the model complexity measures the difference between the information at some point $\hat{\theta}$ and the average amount of information. It thus quantifies how much overconstraint there is at $\hat{\theta}$ or, equivalently, the degree of model complexity. This quantity been historically used in several cosmological analyses [7,13,14,38].

There is a degree of arbitrariness in Eq. (6) via the choice of point estimator $\hat{\theta}$. Typical recommended choices include the posterior mean

$$\hat{\theta}_{\mathrm{m}} = \int \theta\mathcal{P}(\theta)\mathrm{d}\theta = \langle\theta\rangle_{\mathcal{P},} \tag{7}$$

the posterior mode

$$\hat{\theta}_{\mathrm{mp}} = \max_{\theta}\mathcal{P}(\theta), \tag{8}$$

or the maximum likelihood point

$$\hat{\theta}_{\mathrm{ml}} = \max_{\theta}\mathcal{L}(\theta) = \max_{\theta}\mathcal{I}(\theta). \tag{9}$$

For the multivariate Gaussian case, $\hat{d}$ coincides with the actual dimensionality d for all three of these estimators.

Unlike the Kullback-Leibler divergence, the BMC is only weakly prior dependent, since the evidence contributions in Eq. (6) cancel:

$$\hat{d} = 2\log\mathcal{L}(\hat{\theta}) - \langle 2\log\mathcal{L}\rangle_{\mathcal{P}}. \tag{10}$$

The model dimensionality thus only changes with prior π if the posterior bulk is significantly altered by changing the prior. For example $\hat{d}$ does not change if one merely expands the widths of a uniform prior that encompasses the posterior (in contrast to the evidence and Kullback-Leibler divergence).

Finally, the model complexity in Eq. (6) has the advantage of an information-theoretic backing and, like the Shannon information and Kullback-Leibler divergence, is additive for independent parameters.

E. The problem with principle component analysis

Intuitively from Fig. 1 one might describe the distribution as having one "component" that is well constrained and another component for which the posterior provides no information.

The approach that is then followed by many researchers is to perform a PCA, which proceeds thus.

(1) Compute the posterior covariance matrix:

$$\Sigma = \langle(\theta-\bar{\theta})(\theta-\bar{\theta})^T\rangle_{\mathcal{P}}, \qquad \bar{\theta} = \langle\theta\rangle_{\mathcal{P}}. \tag{11}$$

(2) Compute the real eigenvalues $\lambda^{(i)}$ and eigenvectors $\Theta^{(i)}$ of Σ, defined via the equation

$$\Sigma\Theta^{(i)} = \lambda^{(i)}\Theta^{(i)}. \tag{12}$$

(3) The eigenvectors with the smallest eigenvalues are the best constrained components, while the eigenvectors with large eigenvalues are poorly constrained.

One could therefore define an alternative to Eq. (6) based on the number of small eigenvalues, although this itself would depend on the eigenvalue cutoff used to define "unconstrained."

Principle component analysis has intuitive appeal due in large part to the weight given to eigenvectors and eigenvalues early in a physicist's undergraduate mathematical education. However, in many contexts that PCA is applied, the procedure is invalid almost to the point of nonsense.

The issue arises from the fact that the PCA procedure is not invariant under linear transformations. Typically the vectors θ have components with differing dimensionalities, in which case (12) is dimensionally invalid.[1] Equivalently, changing the units that the axes are measured in changes both the eigenvalues and eigenvectors.

For example, for CosmoMC the default cosmological parameter vector is

$$\theta_{\mathrm{cosmo}} = (\Omega_c h^2, \Omega_b h^2, 100\theta_{MC}, \tau, \log 10^{10}A_s, n_s), \tag{13}$$

the first and second components have dimensions of $10^{-4}\ \mathrm{km}^2\,\mathrm{s}^{-2}\,\mathrm{Mpc}^{-2}$, and the third is measured in units of 10^{-2} rad, while the final three are dimensionless. If one were to choose a different unit or scale for any one of these (somewhat arbitrary) dimensionalities, the eigenvalues and eigenvectors would change. To be clear, if all parameters are measured in the same units (as is the case for a traditional normal mode analysis), then PCA is a valid procedure.

Given these observations, the real question is not "is PCA the best procedure?", but in fact "why does PCA usually work at all?" The answer to this question, and an information-theoretically valid PCA, will be developed in an upcoming paper.

There are two ways in which one could adjust the naive PCA procedure to be dimensionally valid. The first is simply to normalize all inputs by the prior, say by computing the prior covariance matrix:

$$\Sigma_0 = \langle(\theta-\bar{\theta})(\theta-\bar{\theta})^T\rangle_{\pi}, \qquad \bar{\theta} = \langle\theta\rangle_{\pi}, \tag{14}$$

and then performing posterior PCA in a space normalized in some sense by this prior.

The second dimensionally valid approach would be to apply the PCA procedure to $\log\theta$. There is an implicit scale that one has to divide each component by in order to apply a logarithm, but this choice only alters the transformation by an additive constant, which PCA is in fact insensitive to. This amounts to finding components that are multiplicative combinations of parameters. A good example of such a combination is $\Omega_b h^2$, or $S_8 = \sigma_8\sqrt{\Omega_m/0.3}$, indicating that physicists are used to thinking in these terms.

F. The anatomy of a Gaussian

As a concrete example of all of the above ideas, we will consider them in the context of a d-dimensional multivariate Gaussian. Consider a posterior $\mathcal{P}$, with parameter covariance matrix Σ and mean μ, arising from a uniform prior π with volume V which fully encompasses the posterior. It is easy to show that the Kullback-Leibler divergence for such a distribution is

$$\mathcal{D} = \log\frac{V}{\sqrt{|2\pi e\Sigma|}}. \tag{15}$$

Each isoposterior ellipsoidal contour $\mathcal{P}(\theta) = \mathcal{P}$ defines a Shannon information $\mathcal{I} = \log\mathcal{P}/\pi$. The posterior distribution $\mathcal{P}(\theta)$ induces an offset, rescaled, χ^2_d distribution on the Shannon information:

[1] Those that believe it is should try to answer the question: What is the dimensionality of each eigenvalue $\lambda^{(i)}$?

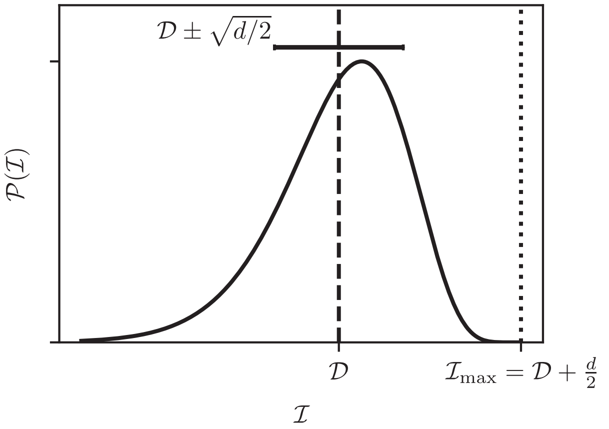

FIG. 2. The typical set of a d-dimensional Gaussian distribution can be visualized by plotting the posterior probability distribution of the Shannon information $\mathcal{I}$. The posterior has mean $\mathcal{D}$ and variance $\frac{d}{2}$. The posterior maximum occurs at $\mathcal{I} = \mathcal{D} + 1$, and the domain is $(-\infty, \mathcal{I}_{\mathrm{max}}]$. The above plot is shown for $d = 16$ in analogy with the *Planck* likelihood from Fig. 8 and Table III.

$$\mathcal{P}(\mathcal{I}) = \frac{1}{\Gamma(d/2)} e^{\mathcal{I}-\mathcal{I}_{\mathrm{max}}} (\mathcal{I}_{\mathrm{max}} - \mathcal{I})^{(d/2)-1}, \tag{16}$$

$$\mathcal{I}_{\mathrm{max}} = \log \frac{V}{\sqrt{|2\pi\Sigma|}} = \mathcal{D} + \frac{d}{2}, \tag{17}$$

$$\mathcal{I} \in (-\infty, \mathcal{I}_{\mathrm{max}}], \qquad \mathcal{I} \approx \mathcal{D} \pm \sqrt{d/2}, \tag{18}$$

which may be seen graphically in Fig. 2.[2] This distribution has mean $\mathcal{D}$ by the definition of the Kullback-Leibler divergence and standard deviation $\sqrt{d/2}$. The region for which the distribution $\mathcal{P}(\mathcal{I})$ is significantly nonzero defines the typical set of the posterior, indicating the Shannon information of points that would be typically drawn from the distribution $\mathcal{P}$. For this Gaussian case, the maximum posterior $\hat{\theta}_{\mathrm{mp}}$, likelihood $\hat{\theta}_{\mathrm{ml}}$ and mean $\hat{\theta}_{\mathrm{m}}$ parameter points coincide and have Shannon information $I_{\mathrm{max}} = \mathcal{D} + \frac{d}{2}$.

III. BAYESIAN MODEL DIMENSIONALITY

A. The problem with Bayesian model complexity $\hat{d}$

While the BMC is widely used in the statistical literature and recovers the correct answer in the case that the posterior distribution is Gaussian, there are three key problems that should be noted.

First, it is clear that the arbitrariness regarding the choice of estimator is far from ideal, and as we shall show in Sec. IV differing choices yield distinct and contradictory answers. A proper information theoretic quantity should be unambiguous.

Second, and most importantly in our view, estimators are not typical posterior points. In general, point estimators such as the maximum likelihood, posterior mode or mean have little statistical meaning in a Bayesian sense, since they occupy a region of vanishing posterior mass. This can be seen in Fig. 2, which shows that while an estimator may represent a point of high information, it lies in a zero posterior mass region—if $d > 2$, one can see from Eq. (16) that $\mathcal{P}(\mathcal{I}_{\mathrm{max}}) = 0$. A physical example familiar to undergraduate quantum physicists is that of the probability distribution of an electron in a 1s orbital: The most likely location to find an electron is the origin, while the radial distribution function shows that the most likely region to find an electron is at the Bohr radius a_0.

A practical consequence of these observations is that if you choose the highest likelihood point from a Markov chain Monte Carlo (MCMC) chain, it will lie at a likelihood some way below the true maximum, and in general one should not expect points in the MCMC chain to lie close to the mean, mode or maximum likelihood point in likelihood space. In general, to compute these point estimators an additional calculation must be performed such as an explicit posterior and likelihood maximization routine or a mean and likelihood computation.

Third, most estimators are parameterization dependent. Namely, if one were to transform the variables and distribution to a different coordinate system via

$$\theta \to \tilde{\theta} = f(\theta), \tag{19}$$

$$\mathcal{P}(\theta) \to \tilde{\mathcal{P}}(\tilde{\theta}) = \mathcal{P}(f^{-1}(\tilde{\theta}))|\partial\theta/\partial\tilde{\theta}|, \tag{20}$$

$$\pi(\theta) \to \tilde{\pi}(\tilde{\theta}) = \pi(f^{-1}(\tilde{\theta}))|\partial\theta/\partial\tilde{\theta}|, \tag{21}$$

then neither the posterior mean from Eq. (7) nor the posterior mode from Eq. (8) transform under Eq. (19) if the transformation f is nonlinear (i.e., the Jacobian $|\partial\theta/\partial\tilde{\theta}|$ depends on $\tilde{\theta}$). It should be noted that this parameterization variance is not quite as bad as it is for the PCA case, which is dependent on even linear transformations of the parameter vector. The maximum likelihood point from Eq. (9) does correctly transform, since the Jacobian terms in Eqs. (20) and (21) cancel in the Shannon information. Parameterization dependency is a highly undesirable ambiguity, particularly in the context of cosmology where in general the preferred choice of parameterization varies between likelihoods and sampling codes [39–41].

Finally, specifically to the mean estimator, for some cosmological likelihoods there may be no guarantee that the mean even lies in the posterior mass, for example in the $\sigma_8 - \Omega_m$ banana distribution visualized by kilo-degree survey [42]. In cosmology, we do not necessarily have the luxury of Gaussianity or convexity.

[2] Note that in the manipulation for Eq. (17) we have used the fact that $\log\sqrt{|2\pi e\Sigma|} = \log\sqrt{e^d|2\pi\Sigma|} = \frac{d}{2} + \log\sqrt{|2\pi\Sigma|}$.

B. The Bayesian model dimensionality $\tilde{d}$

Considering Fig. 2, the fundamental concept to draw is that the BMC leverages the fact that the difference between the Shannon information $\mathcal{I}$ at the posterior peak and the mean of the posterior bulk is $d/2$ for the Gaussian case.

However, there is a second way of bringing the dimensionality out of Fig. 2 via the variance of the posterior bulk. With this in mind, we define the *Bayesian model dimensionality* as

$$\frac{\tilde{d}}{2} = \int \mathcal{P}(\theta)\left(\log\frac{\mathcal{P}(\theta)}{\pi(\theta)} - \mathcal{D}\right)^2 \mathrm{d}\theta \tag{22}$$

$$= \langle \mathcal{I}^2\rangle_{\mathcal{P}} - \langle \mathcal{I}\rangle_{\mathcal{P}}^2 \tag{23}$$

or equivalently as

$$\tilde{d}/2 = \langle (\log\mathcal{L})^2\rangle_{\mathcal{P}} - \langle \log\mathcal{L}\rangle_{\mathcal{P}}^2. \tag{24}$$

We note that this form for quantifying model dimensionality is discussed in passing by Gelman *et al.* [9] (p. 173) and Spiegelhalter *et al.* [10], who conclude that $\tilde{d}$ is less numerically stable than $\hat{d}$. As we shall discuss in Sec. IV we find that when applied to cosmological likelihoods the opposite is in fact true. This measure of model dimensionality is also discussed briefly in the landmark nested sampling paper by Skilling [11], by Raveri and Hu [13], in a cosmological context in terms of χ^2 in Kunz, Trotta, and Parkinson [14] and Liddle [15], and was used as part of the *Planck* analysis [3].

The definition of $\tilde{d}$ shares all of the desiderata that $\hat{d}$ provides; namely both $\tilde{d}$ and $\hat{d}$ are weakly prior dependent, additive for independent parameters and recover the correct answer in the Gaussian case. We believe that there are several attractive theoretical characteristics of $\tilde{d}$ that we view as advantages over $\hat{d}$.

First, $\tilde{d}$ relies only on points drawn from the typical set, which is highly attractive from a Bayesian and information theoretic point of view and more consistent when used alongside a traditional MCMC analysis of cosmological posteriors.

Second, there is a satisfying progression in the fact that while the mean of the Shannon information $\mathcal{D}$ gives one an overall constraint, the next order statistic (the variance) yields a measure of the dimensionality of the constraint.

Finally, in eschewing estimators this measure is completely unambiguous, as it removes all arbitrariness associated with both estimator and underlying parameterization choice.

It should be noted that the computation of $\mathcal{D}$ requires nested sampling to provide an estimate of $\log\mathcal{Z}$. The dimensionality $\tilde{d}$ on the other hand can be computed from a more traditional MCMC chain via Eq. (24).

C. Thermodynamic interpretation

There is a second motivation for the BMD arising from a thermodynamic viewpoint.[3] The thermodynamic generalisation of Bayes theorem is

$$\mathcal{L}^\beta(\theta)\times\pi(\theta) = \mathcal{P}_\beta(\theta)\times\mathcal{Z}(\beta), \tag{25}$$

$$\mathcal{Z}(\beta) = \int \mathcal{L}^\beta(\theta)\pi(\theta)\mathrm{d}\theta, \tag{26}$$

where on the left-hand side of Eq. (25), the inverse temperature $\beta = \frac{1}{T}$ raises the likelihood $\mathcal{L}$ to the power of β and on the right-hand side the posterior has a nontrivial dependency on temperature, denoted by a subscript β. When the evidence in Eqs. (25) and (26) is a function of β it is usually called the partition function.

The link to thermodynamics comes by considering θ to be a continuous index i over microstates, the negative log-likelihood to be the energy E of a microstate, and the prior to be the degeneracy of microstates g:

$$i\leftrightarrow\theta,\qquad E_i\leftrightarrow -\log\mathcal{L}(\theta),\qquad g_i\leftrightarrow\pi(\theta),$$
$$g_i e^{-\beta E_i}\leftrightarrow\mathcal{L}(\theta)^\beta\pi(\theta). \tag{27}$$

It should be noted that one of the principal advantages of nested sampling is that (other than the stopping criterion) it is blind to β, and therefore samples at all temperatures simultaneously. Nested samplers are best described as partition function calculators rather than as posterior samplers. This will be explored in further detail in an upcoming paper [43].

In its thermodynamic form, the evidence becomes a generating function [44]:

$$\frac{\mathrm{d}^2}{\mathrm{d}\beta^2}\log\mathcal{Z}(\beta) = \frac{\mathrm{d}}{\mathrm{d}\beta}\langle\log\mathcal{L}\rangle_{\mathcal{P}_\beta} = \frac{\tilde{d}}{2}, \tag{28}$$

and we may identify the BMD as being related to the rate of change of average log-likelihood (energy) with inverse temperature. The BMD is therefore proportional to the Bayesian analogue of a heat capacity, $C = \frac{\mathrm{d}}{\mathrm{d}T}\langle E\rangle = \beta^2\frac{\mathrm{d}}{\mathrm{d}\beta}\langle -E\rangle$, and, like all heat capacities, is proportional to system size or, equivalently, to the number of active degrees of freedom (i.e., dimensions).

D. Analytical examples

We apply the BMD from Eq. (23) and the BMC from Eq. (6) to six additional univariate analytical examples: top hat, triangular, cosine, logistic, Laplace and Cauchy. The analytical forms for the probability distribution, Kullback-Leibler divergence, BMD and BMC are listed in Table II

[3] Historically, it was this viewpoint that drew our attention to BMD.

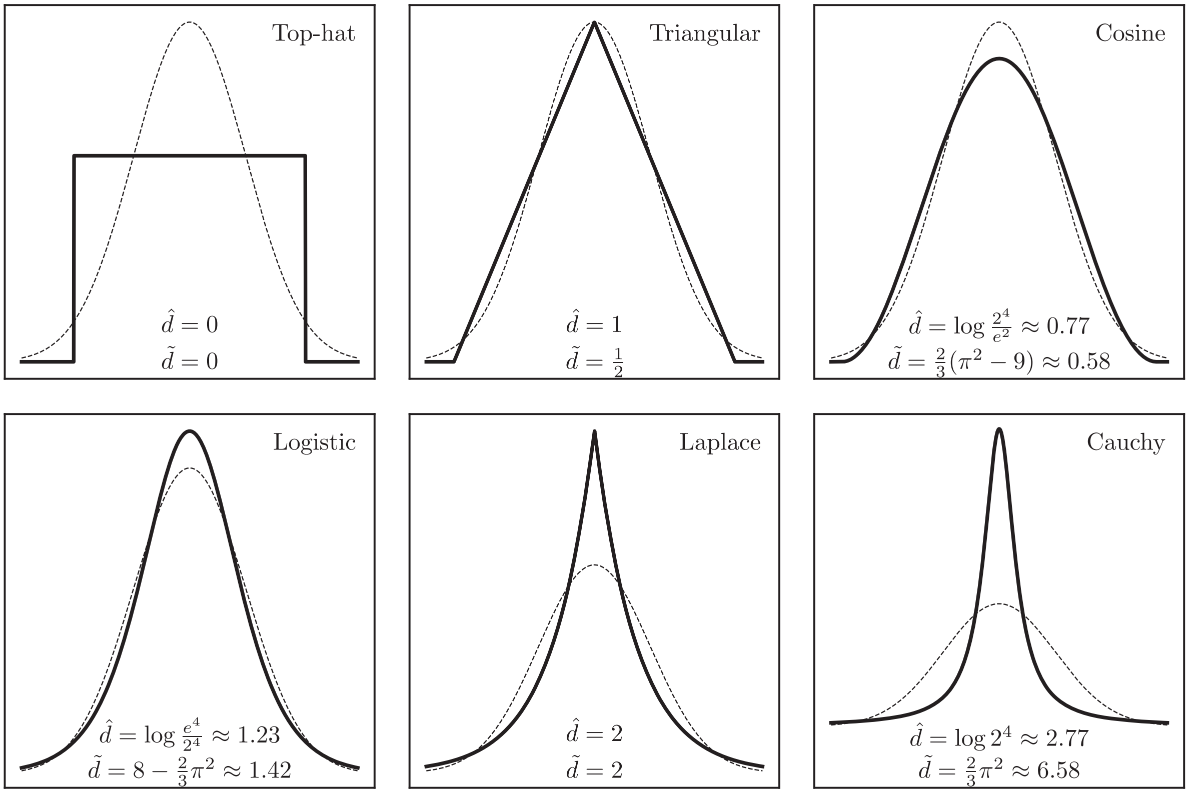

FIG. 3. Bayesian dimensionality for the common one-dimensional distributions in Table II. Widths are normalized so that the distributions all have the same Kullback-Leibler divergence $\mathcal{D}$. The dashed curve in all plots is a Gaussian distribution.

and plotted in Fig. 3. In all cases, we assume a uniform prior of volume V which fully encompasses the posterior.

We find that while the Gaussian distribution gives $\tilde{d} = 1$, distributions that are shorter and fatter give $\tilde{d} < 1$, while distributions that are narrower and taller give $\tilde{d} > 1$. Both measures of $\tilde{d}$ and $\hat{d}$ are in broad agreement. The top-hat (dimensionality 0) and Cauchy distributions (dimensionality ≫ 1) represent pathological cases at either end of the

TABLE II. Dimensionalities for one-dimensional analytic distributions. The first column indicates the unnormalized probability density $\mathcal{P}^*(x)$. An arbitrary width σ can be added by remapping $\mathcal{P}^*(x) \to \frac{1}{\sigma}\mathcal{P}^*(x/\sigma)$. The second column indicates the unnormalized Kullback-Leibler divergence $\mathcal{D}^* = \mathcal{D} - \log V/\sigma$ where the implicit prior is taken to encompass the posterior mass with width $V \gg \sigma$. The final two columns show the BMDs and BMCs, respectively, which are independent of both V and σ. As expected, the Gaussian has dimensionality $\tilde{d} = \hat{d} = 1$, and shorter and fatter distributions have lower dimensionalities, while narrower and taller dimensionalities have dimensions greater than one. This effect can be seen graphically in Fig. 3.

	$\mathcal{P}^*$	$\exp(\mathcal{D}^*)$	$\tilde{d}$	$\hat{d}$
Gaussian	$e^{-x^2/2}$	$1/\sqrt{2\pi e}$	1	1
Top hat	$x \in [-1, 1]$	1	0	0
Triangle	$1 - \lvert x\rvert$	$1/\sqrt{e}$	1/2	1
Cosine	$\cos^2 x$	$e/2\pi$	$\frac{2(\pi^2-9)}{3} \approx 0.58$	$\log \frac{2^4}{e^2} \approx 0.77$
Logistic	$\frac{e^{-x}}{(1+e^{-x})^2}$	$1/e^2$	$\frac{24-2\pi^2}{3} \approx 1.42$	$\log \frac{e^4}{2^4} \approx 1.33$
Laplace	$e^{-\lvert x\rvert}$	$1/2e$	2	2
Cauchy	$(1 + x^2)^{-1}$	$1/4\pi$	$\frac{2\pi^2}{3} \approx 6.58$	$\log 2^4 \approx 2.77$

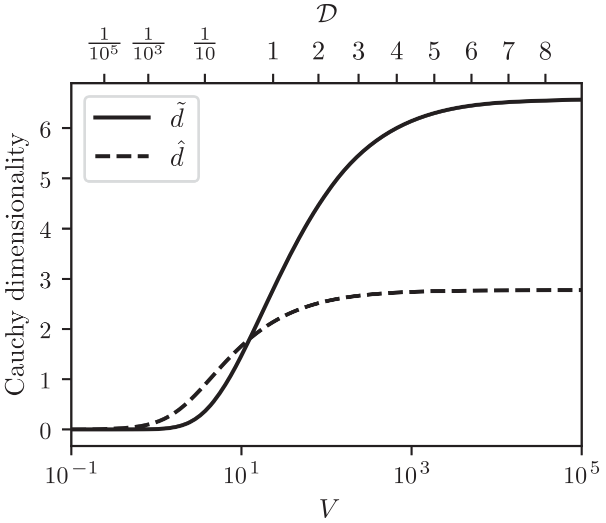

FIG. 4. Dependency of dimensionality and Kullback-Leibler divergence on prior volume for a Cauchy distribution $\mathcal{P}(x) \propto (1 + x^2)^{-1}$. While the BMD and BMC are pathologically large (≫1) if the full domain of the Cauchy distribution is included, truncating the range to a lower prior volume $x \in [-V/2, V/2]$ reduces the dimensionality to more sensible values.

scale, while the remainder all give dimensionalities of order 1. In general, $\hat{d}$ is closer to unity than $\tilde{d}$, on account of the "numerical stability" quoted by Gelman *et al.* [9]. However, accurate computation of $\hat{d}$ is predicated on an exact computation of the maximum, which (as shown in Sec. IV) becomes increasingly unstable in higher dimensions and in cosmological applications.

It should also be noted that while the Cauchy distribution gives a very high dimensionality when integrated over its full infinite domain, if the domain is restricted by the prior, then the dimensionality drops to more sensible values (Fig. 4).

E. Applications

BMDs can be used in a variety of statistical analyses. In this subsection we review a few of the possibilities.

1. The number of constrained cosmological parameters

As detailed in Table I, cosmological likelihoods typically introduce a large number of nuisance parameters in addition to cosmological ones, and they typically constrain a nontrivial combination of parameters. If one has datasets A and B, one can compute the individual model dimensionalities $\tilde{d}_A$ and $\tilde{d}_B$, as well as the model dimensionality of using both datasets together $\tilde{d}_{AB}$. Computing the crossover of these dimensionalities for any choice of d, $\tilde{d}$ or $\hat{d}$:

$$\tilde{d}_{A\cap B} = \tilde{d}_A + \tilde{d}_B - \tilde{d}_{AB} \tag{29}$$

will give the (effective) number of constrained cosmological parameters shared between the datasets, since any parameters constrained by just one of the datasets subtract out of the above expression. This quantity forms a cornerstone of part of the tension analysis in Handley and Lemos [7], and cosmological examples can be seen in the lower section of Table III.

2. Penalizing the number of model parameters

Bayesian evidences are traditionally used in model comparison via the Bayes theorem for models:

$$P(\mathcal{M}_i) = \frac{P(D|\mathcal{M}_i)P(\mathcal{M}_i)}{\sum_j P(D|\mathcal{M}_i)P(\mathcal{M}_j)} = \frac{\mathcal{Z}_i\Pi_i}{\sum_j \mathcal{Z}_j\Pi_j}, \tag{30}$$

where $\Pi_i = P(\mathcal{M}_i)$ are the model priors, which are typically taken to be uniform. Often the data may not be discriminative enough to pick an unambiguously best model via the model posteriors. The correct Bayesian approach in this case is to perform model marginalization over any future predictions [45]. However, in other works [46,47] the Kullback-Leibler divergence has been used to split this degeneracy. The strong prior dependency of the KL divergence can make this a somewhat unfair choice for splitting this degeneracy, and users may find that the model dimensionality is a fairer choice.

One implementation of this approach would be to apply a *post hoc* model prior of

$$\Pi_i(\lambda) = \lambda e^{-\lambda\tilde{d}_i} \tag{31}$$

using for example $\lambda = 1$. This amounts to a logarithmic Bayes factor between models of

$$\log \mathcal{B}^i_j = (\log \mathcal{Z}_i - \lambda\tilde{d}_i) - (\log \mathcal{Z}_j - \lambda\tilde{d}_j). \tag{32}$$

This approach is not strictly Bayesian, since $\tilde{d}_i$ is computed from the data and $\Pi_i(\lambda)$ is therefore not a true prior. However readers familiar with the concepts of sparse reconstructions [48] will recognize the parallels between sparsity and this approach, as one is effectively imposing a penalty factor that promotes models that use as few parameters as necessary to constrain the data.

3. Information criteria

While the authors' preferred method of model comparison is via the Bayesian evidence, other criteria have been used in the context of cosmology [15,49]: The Akaike information criterion (AIC) [50] and Bayesian information criterion (BIC) [51] are defined, respectively, via

$$\mathrm{AIC} = -2\log\mathcal{L}_{\max} + 2k, \tag{33}$$

$$\mathrm{BIC} = -2\log\mathcal{L}_{\max} + k\ln N, \tag{34}$$

where k is the number of parameters in the model and N is the number of data points used in the fit. These criteria could be modified in a Bayesian sense by replacing k with the BMD $\tilde{d}$. A similar modification has been discussed in the context of the deviance information criterion [14,15].

IV. NUMERICAL EXAMPLES

A. Cosmological likelihoods

We test our method on real data by quantifying the effective number of constrained parameters in four publicly available cosmological datasets, assuming a six-parameter ΛCDM cosmological model. We use the following six sampling parameters to describe this model: the density of baryonic matter $\Omega_b h^2$, the density of cold dark matter $\Omega_c h^2$, θ_{MC} an approximation of the ratio of the sound horizon to the angular diameter distance at recombination, the optical depth to reionization τ and the amplitude and tilt of the primordial power spectrum A_s and n_s, respectively. This is the default parameterisation for CosmoMC [40] and is chosen to maximize the efficiency of Metropolis-Hastings sampling codes for CMB data. The possible effects of this parameterization choice in non-CMB constraints will be explored in future work.

We use four key datasets in our analysis. First, we use measurements of temperature and polarization anisotropies

TABLE III. Bayesian model dimensionalities for cosmological datasets. The first column indicates the Kullback-Leibler divergence $\mathcal{D}$ from Eq. (5), and the second column shows the Bayesian model dimensionality $\tilde{d}$ from Eq. (23). The remaining three columns show the Bayesian model complexity $\hat{d}$ from Eq. (6) with the estimator chosen as the posterior mean, posterior mode and maximum likelihood point, respectively. The final three rows show the intersection statistics, computed using the equivalents of Eq. (29).

Dataset	$\mathcal{D}$	$\tilde{d}$	$\hat{d}_{\mathrm{m}}$	$\hat{d}_{\mathrm{mp}}$	$\hat{d}_{\mathrm{ml}}$	d
$\mathrm{SH_0ES}$	2.52 ± 0.03	0.93 ± 0.03	-40.12 ± 0.02	0.96 ± 0.02	0.96 ± 0.02	6
BOSS	5.06 ± 0.05	2.95 ± 0.07	-9.55 ± 0.05	2.93 ± 0.05	2.93 ± 0.05	6
DES	22.82 ± 0.15	14.03 ± 0.30	10.79 ± 0.14	14.45 ± 0.14	17.85 ± 0.14	26
Planck	44.48 ± 0.20	15.84 ± 0.38	14.91 ± 0.16	15.68 ± 0.16	18.91 ± 0.16	21
$\mathrm{SH_0ES}$ + *Planck*	45.02 ± 0.20	15.97 ± 0.36	14.64 ± 0.15	15.39 ± 0.15	18.40 ± 0.15	21
BOSS + *Planck*	43.36 ± 0.20	15.89 ± 0.38	15.11 ± 0.17	15.57 ± 0.17	18.89 ± 0.17	21
DES + *Planck*	61.13 ± 0.25	25.88 ± 0.62	20.79 ± 0.25	23.54 ± 0.25	29.30 ± 0.25	41
$\mathrm{SH_0ES} \cap$ *Planck*	1.99 ± 0.29	0.80 ± 0.52	-39.84 ± 0.23	1.25 ± 0.23	1.48 ± 0.23	6
BOSS $\cap$ *Planck*	6.18 ± 0.30	2.91 ± 0.54	-9.75 ± 0.23	3.04 ± 0.23	2.96 ± 0.23	6
DES $\cap$ *Planck*	6.17 ± 0.36	3.98 ± 0.77	4.91 ± 0.32	6.59 ± 0.32	7.46 ± 0.32	6

in the CMB measured by *Planck* in the form of the publicly available *Planck* 2015 data [36,52]. Second, we use local cosmic distance ladder measurements of the expansion rate, using type Ia SNe calibrated by variable Cepheid stars and implemented as a Gaussian likelihood with the mean and standard deviation given by the latest results obtained by the $\mathrm{SH_0ES}$[4] Collaboration [53]. Third, we use the Dark Energy Survey (DES) year 1 combined analysis of cosmic shear, galaxy clustering and galaxy-galaxy lensing (a combination commonly referred to as "3 × 2") [2]. Finally, we use baryon acoustic oscillation measurements from the Baryon Oscillation Spectroscopic Survey (BOSS) [54] DR12 [55–57]. The number of parameters that we sample over for each likelihood is described in Table I.

B. Nested sampling

To compute the log-evidence $\log \mathcal{Z}$, Kullback-Leibler divergence $\mathcal{D}$ and Bayesian model dimensionality $\tilde{d}$, we use the outputs of a nested sampling run produced by CosmoChord [40,58–61] and the equations

$$\mathcal{Z} \approx \sum_{i=1}^{N} \mathcal{L}_i \times \frac{1}{2}(X_{i-1} - X_{i+1}),$$

$$\mathcal{D} \approx \sum_{i=1}^{N} \frac{\mathcal{L}_i}{\mathcal{Z}} \log \frac{\mathcal{L}_i}{\mathcal{Z}} \times \frac{1}{2}(X_{i-1} - X_{i+1}),$$

$$\frac{\tilde{d}}{2} \approx \sum_{i=1}^{N} \left(\frac{\mathcal{L}_i}{\mathcal{Z}} \log \frac{\mathcal{L}_i}{\mathcal{Z}} - \mathcal{D} \right)^2 \times \frac{1}{2}(X_{i-1} - X_{i+1}),$$

$$X_i = t_i X_{i-1}, \qquad X_0 = 1, \qquad X_{N+1} = 0,$$

$$P(t_i) = n_i t_i^{n_i - 1} \; [0 < t_i < 1], \tag{35}$$

where $\mathcal{L}_i$ are the N likelihood contours of the discarded points, X_i are the prior volumes, and n_i are the number of live points t_i are real random variables. We compute 1000 batches of the samples $\{t_i : i = 1 \ldots N\}$. Code for performing the above calculation is provided by the Python package anesthetic [62]. For our final runs, we used the CosmoChord settings $n_{\mathrm{live}} = 1000$, $n_{\mathrm{prior}} = 10\,000$, with all other settings left at their defaults for CosmoChord version 1.15. For more detail, see Skilling [11] or Handley and Lemos [7].

In order to compute the maximum likelihood and posterior points, we found that the most reliable procedure was to use a Nelder-Mead simplex method [63] with the initial simplex defined by the highest likelihood live points found before termination.

C. Results

Our main results are detailed in Table III, where we report the Bayesian model dimensionality $\tilde{d}$ obtained from Eq. (23), compared with the values obtained for the Bayesian model complexity using Eq. (6) using three different estimators from Eqs. (7)–(9): the posterior mean, posterior mode and maximum likelihood. We use the four individual datasets described in Sec. IV A, as well as in combination with *Planck*. We also report the shared dimensionalities from Eq. (29) using *Planck* as the common baseline in the bottom three rows of the table.

The BMDs produce sensible values in all cases. It should be noted that in general the BMDs are lower than the number of dimensions that are sampled over (Table III): $\mathrm{SH_0ES}$ constrains only one parameter (H_0), BOSS constrains three ($\Omega_b h^2$, $\Omega_c h^2$ and a degenerate $H_0 - A_s$ constraint), and DES and *Planck* constrain only some combinations of cosmological and nuisance parameters as shown by Figs. 5, 6 and 7.

The shared dimensionalities also match cosmological intuition. For example, $\tilde{d}_{\mathrm{DES} \cap Planck}$ shows that DES only constrains four cosmological parameters, as it provides no constraint on τ and only constrains a combination of n_s and

[4] Supernovae and H_0 for the equation of state.

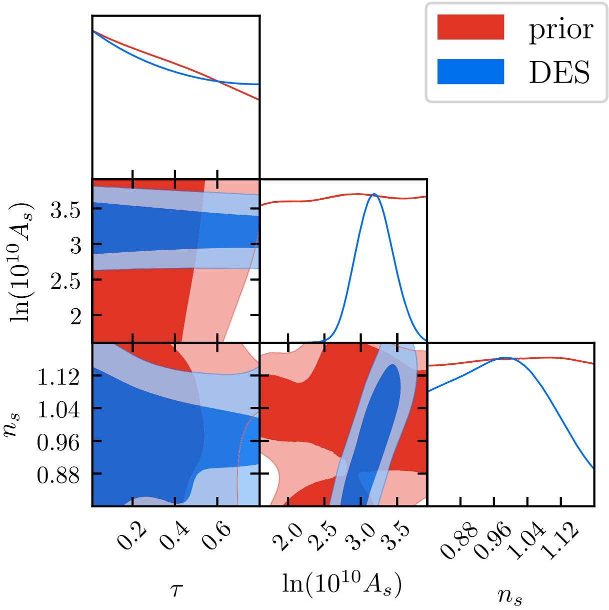

FIG. 5. Cosmological parameters unconstrained by DES. While DES provides constraints on four of the cosmological parameters, it tells us nothing of τ and little of a correlated combination of $\ln 10^{10} A_s$ and n_s. This figure should be compared with Fig. 1.

$\log 10^{10} A_s$. This is shown graphically in Fig. 5, which should be compared with Fig. 1.

All error bars on the dimensionalities arise from nested sampling's imperfect knowledge of prior volumes used to compute the posterior weights. It is likely that the error could be lowered by using a more traditional MCMC run [39–41], although care must be taken with the MCMC error estimation since marginalizing over the likelihood is numerically more unstable than that of traditional expectation values. The process of computing the Bayesian model dimensionalities and their errors is visualized in Fig. 8, which should be compared with Fig. 2.

The results for Bayesian model complexities on the other hand are nowhere near as satisfactory. The arbitrariness in the choice of estimator can be seen clearly, and in general we find $\hat{d}_{\mathrm{m}} < \hat{d}_{\mathrm{mp}} < \hat{d}_{\mathrm{ml}}$. This variation in the choice of estimator is demonstrated graphically in Fig. 9. The two maximization estimators come out a little high, with the most extreme example being that the maximum likelihood estimator claims that there are 7.5 shared dimensions between DES and *Planck*, which is concerning given that there are only six parameters that are shared between them. The fact that the maximum likelihood estimator consistently produces dimensionalities that are too large is unfortunate, given that it is the best motivated of all three estimators.

The mean estimator on the other hand is generally a little lower than expected and produces nonsensical results for SH$_0$ES and BOSS alone, where as mentioned in Sec. III A the parameterization variance of the estimator makes the mean extremely unreliable. In the case of SH$_0$ES, we are sampling over six cosmological parameters but only constrain H_0, which is only one combination of those six. As a consequence, the value of H_0 derived from the means of the mostly unconstrained cosmological parameters is completely prior dominated.

The fact that Fig. 9 shows that the estimators are most consistent for *Planck* data is also very telling. This is not caused by any properties of the *Planck* data; instead it is a consequence of the parameterization choice: All of these posteriors are obtained using the CosmoMC parameterization, which is chosen to be optimal for CMB analyses.

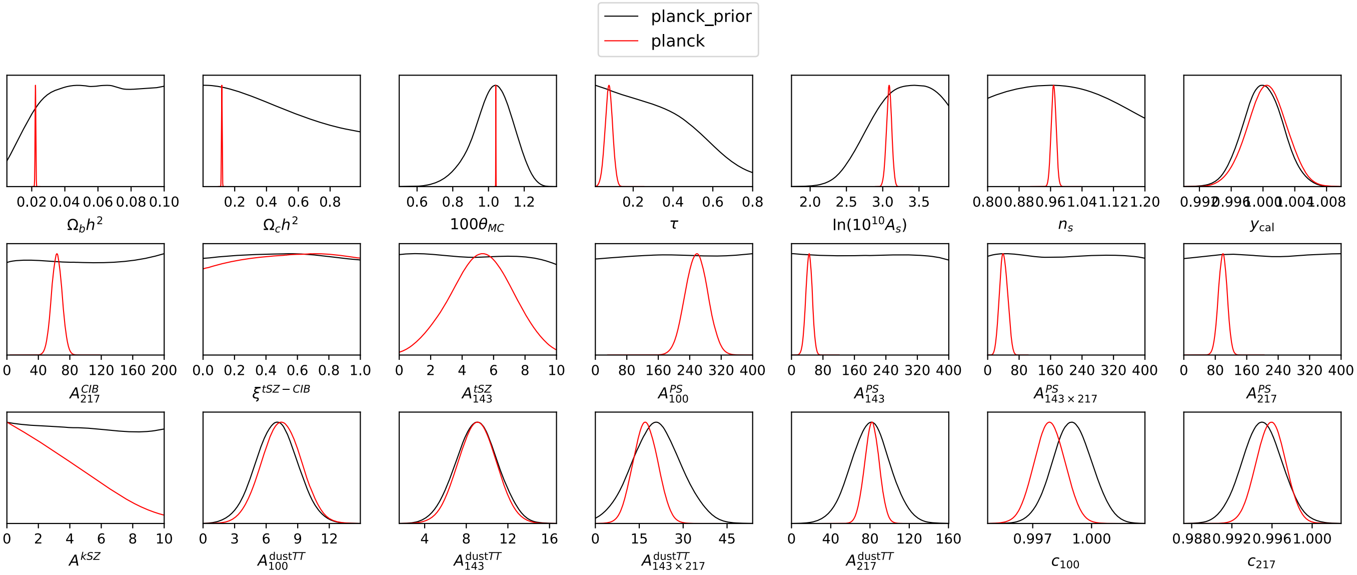

FIG. 6. One-dimensional marginalized default prior (black) and *Planck* posterior (red). The Bayesian model dimensionality of $\tilde{d}_{Planck} \approx 16$ is reflected by the fact that only a subset of the nuisance parameters are constrained by the data.

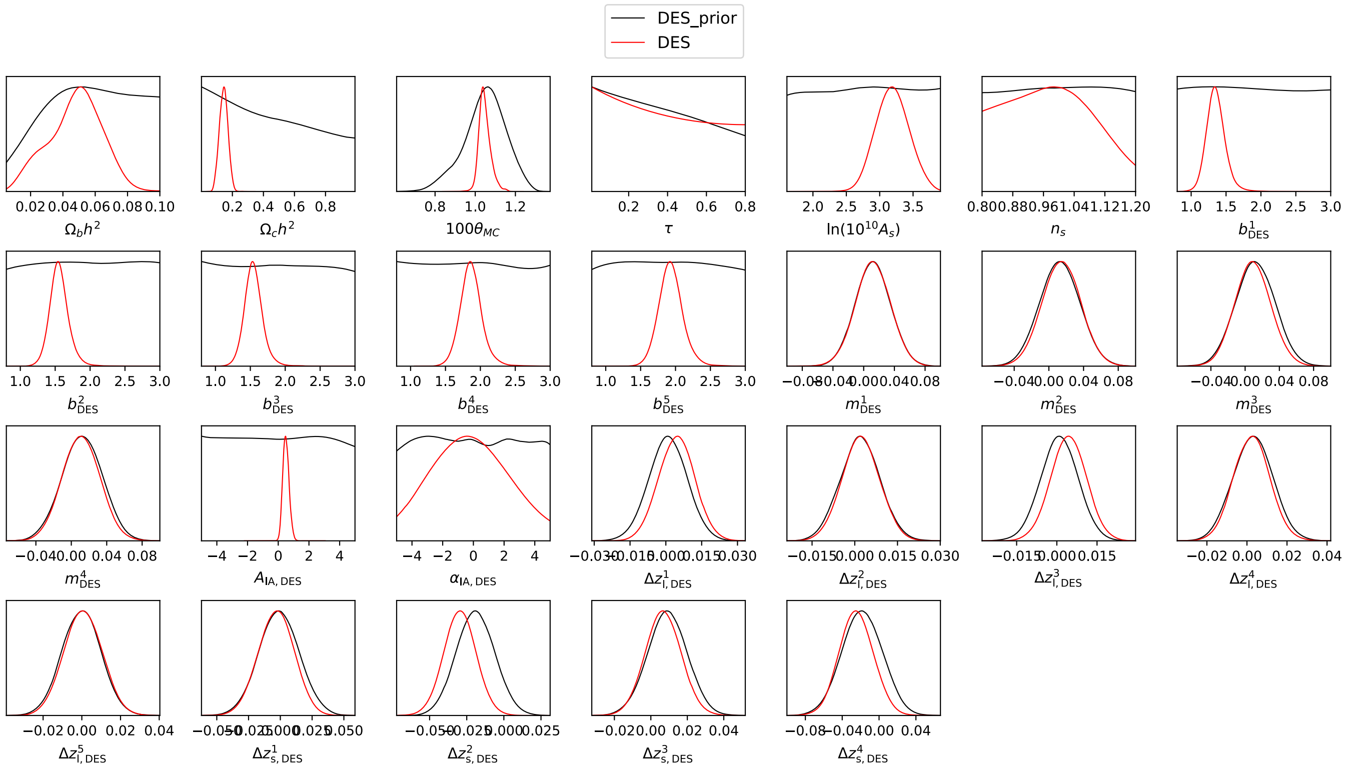

FIG. 7. One-dimensional marginalized default prior (black) and DES Y1 posterior (red). The Bayesian model dimensionality of $\tilde{d}_{\mathrm{DES}} \approx 14$ is reflected by the fact that only a combination of the cosmological parameters and a subset of the nuisance parameters are constrained by the data.

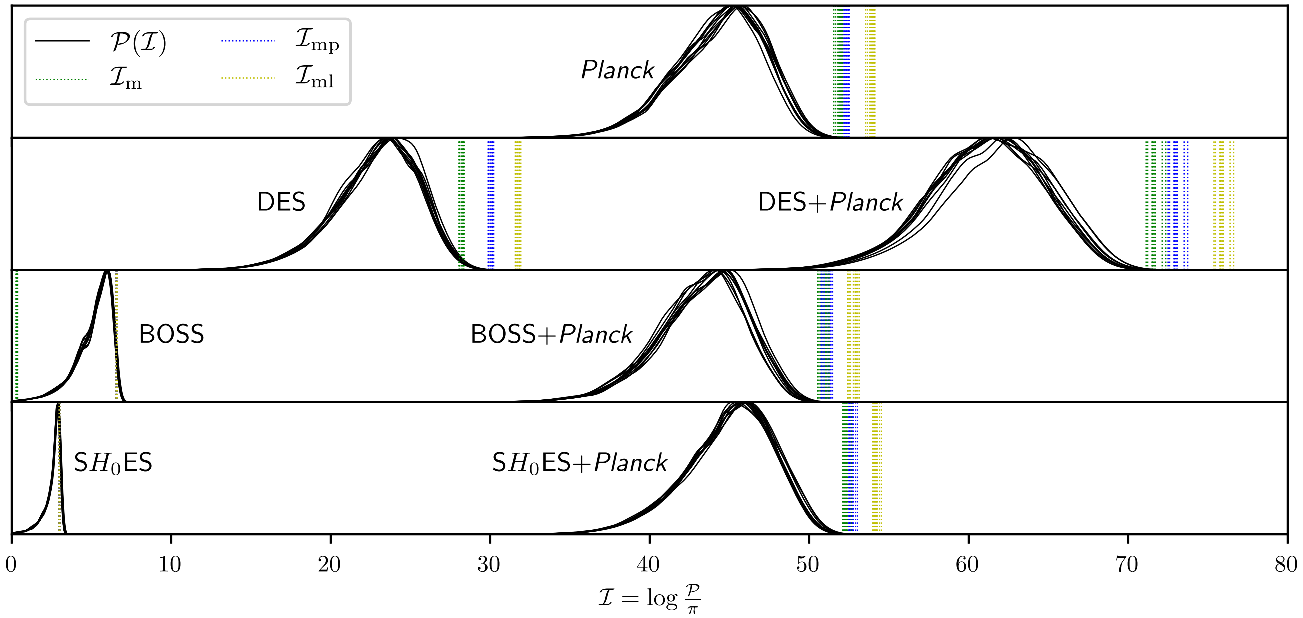

FIG. 8. Shannon information for the numerical examples considered in this paper. These plots are laid out in the same manner as Fig. 2, with the mean of each distribution representing the Kullback-Leibler divergence and the variance the Bayesian model dimensionality. The main difference between these plots and Fig. 2 is that the posterior mean $\mathcal{I}_{\mathrm{m}}$, mode $\mathcal{I}_{\mathrm{mp}}$ and maximum likelihood $\mathcal{I}_{\mathrm{ml}}$ points no longer coincide on account of the nonuniform priors and nontrivial parameterization involved in cosmological modeling. The multiple curves for $\mathcal{P}(\mathcal{I})$ represent independent samples from the distribution of nested sampling prior volumes used to compute the Shannon information, and the spread in these curves accounts for the errors in estimating the quantities detailed in Table III.

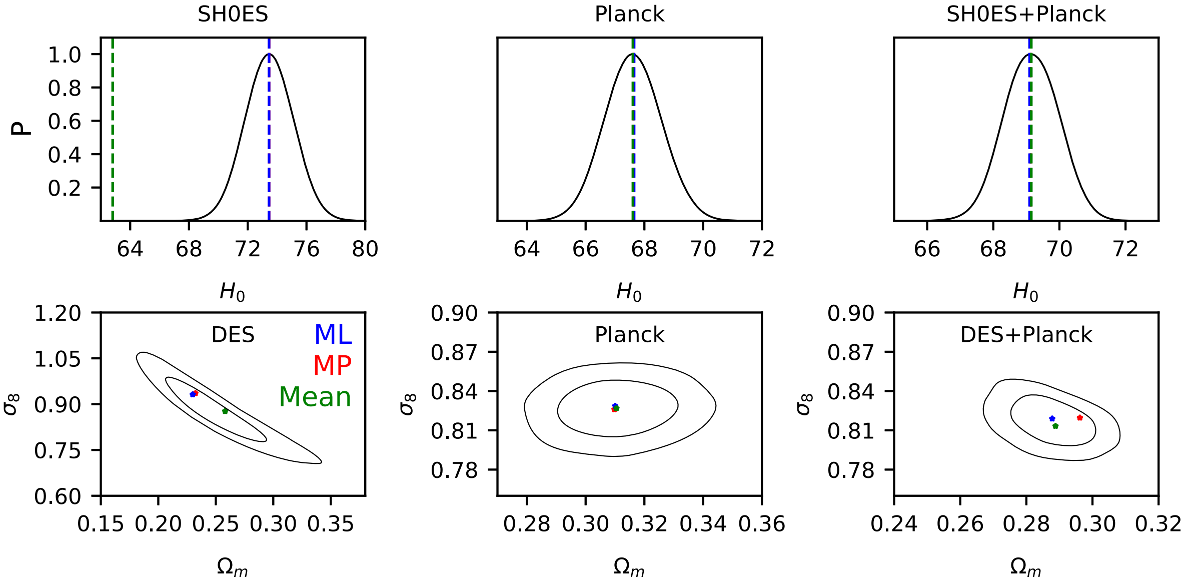

FIG. 9. Marginalized posterior likelihoods (in black), maximum likelihood points (ML, in blue), maximum posterior points (MP, in red) and means (in green), for some of the numerical examples used in this paper. The top plots detail the one-dimensional marginalized posterior on the Hubble parameter, while the lower plots show the two-dimensional marginalized posterior on σ_8 and Ω_m. The top left shows the SH$_0$ES likelihood, the top center *Planck*, and the top right the combination of both. The bottom left shows the DES posterior, the bottom center *Planck*, and the bottom right their combination.

The parameters that other surveys like DES and SH$_0$ES constrain are obtained as derived parameters, which changes both the mean and the maximum posterior.

V. CONCLUSION

In this paper we interpret the variance in the Shannon information as a measure of Bayesian model dimensionality and present it as an alternative to Bayesian model complexity currently used in the literature. We compared these two measures of dimensionality theoretically and in the context of cosmological parameter estimation and found that the Bayesian model dimensionality proves more accurate in reproducing results consistent with intuition.

While the Bayesian model dimensionality has been acknowledged in the literature in different forms, it has yet not been widely used in cosmology. Given the ease with which the Bayesian model dimensionality can be computed from MCMC chains, we hope that this work persuades cosmologists to use this crucial statistic as a part of their analyses. For those using nested sampling, we hope that in the future the reporting of the triple of evidence, Kullback-Leibler divergence and Bayesian model dimensionality $(\mathcal{Z}, \mathcal{D}, \tilde{d})$ becomes a scientific standard.

ACKNOWLEDGMENTS

This work was performed using resources provided by the Cambridge Service for Data Driven Discovery (CSD3) operated by the University of Cambridge Research Computing Service, provided by Dell EMC and Intel using Tier-2 funding from the Engineering and Physical Sciences Research Council (capital Grant No. EP/P020259/1), and Distributed Research using Advanced Computing (DiRAC) funding from the Science and Technology Facilities Council. We thank Dan Mortlock for the discussion of model dimensionalities via email correspondence following the release of Handley and Lemos [7] which directly led to the key idea behind this paper. W. H. thanks Rafi Blumenfeld for lunchtime conversations on alternative PCA approaches. W. H. thanks Gonville & Caius College for their continuing support via a research fellowship. P. L. thanks Science and Technology Facilities Council and UCL for their support via a Science and Technology Facilities Council Consolidated Grant.

[1] Douglas Scott, The Standard Model of cosmology: A skeptic's guide, arXiv:1804.01318.

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

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

[4] M. A. Troxel and M. Ishak, The intrinsic alignment of galaxies and its impact on weak gravitational lensing in an era of precision cosmology, Phys. Rep. **558**, 1 (2015).

[5] B. Joachimi, M. Cacciato, T. D. Kitching, A. Leonard, R. Mandelbaum, B. Malte Schäfer, C. Sifón, H. Hoekstra, A. Kiessling, D. Kirk, and A. Rassat, Galaxy alignments: An overview, Space Sci. Rev. **193**, 1 (2015).

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

[7] W. Handley and P. Lemos, Quantifying tension: Interpreting the DES evidence ratio, arXiv:1902.04029.

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

[9] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin, *Bayesian Data Analysis*, 2nd ed. (Chapman and Hall/CRC, London, 2004).

[10] D. J. Spiegelhalter, N. G. Best, B. P. Carlin, and A. van der Linde, The deviance information criterion: 12 years on, J. R. Stat. Soc. Ser. B **76**, 485 (2014).

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

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

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

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

[15] A. R. Liddle, Information criteria for astrophysical model selection, Mon. Not. R. Astron. Soc. **377**, L74 (2007).

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

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

[18] D. S. Sivia and J. Skilling, *Data Analysis: A Bayesian Tutorial* (Oxford Science Publications, Oxford University Press, New York, 2006).

[19] S. Hannestad and T. Tram, Optimal prior for Bayesian inference in a constrained parameter space, arXiv:1710.08899.

[20] C. E. Shannon and W. Weaver, *The Mathematical Theory of Communication* (University of Illinois Press, 1949).

[21] *Bayesian Methods in Cosmology* (Cambridge University Press, Cambridge, England, 2009).

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

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

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

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

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

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

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

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

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

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

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

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

[34] S. Geman and D. Geman, Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images, IEEE Trans. Pattern Anal. Mach. Intell. **PAMI-6**, 721 (1984).

[35] M. Betancourt, A conceptual introduction to Hamiltonian Monte Carlo, arXiv:1701.02434.

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

[37] D. Foreman-Mackey, corner.py: Scatterplot matrices in python, J. Open Source Software **1**, 24 (2016).

[38] M. Bridges, F. Feroz, M. P. Hobson, and A. N. Lasenby, Bayesian optimal reconstruction of the primordial power spectrum, Mon. Not. R. Astron. Soc. **400**, 1075 (2009).

[39] J. Zuntz, M. Paterno, E. Jennings, D. Rudd, A. Manzotti, S. Dodelson, S. Bridle, S. Sehrish, and J. Kowalkowski, CosmoSIS: Modular cosmological parameter estimation, Astron. Comput. **12**, 45 (2015).

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

[41] B. Audren, J. Lesgourgues, K. Benabed, and S. Prunet, Conservative constraints on early cosmology: An illustration of the Monte Python cosmological parameter inference code, J. Cosmol. Astropart. Phys. 02 (2013) 001.

[42] E. van Uitert *et al.*, KiDS + GAMA: Cosmology constraints from a joint analysis of cosmic shear, galaxy-galaxy lensing, and angular clustering, Mon. Not. R. Astron. Soc. **476**, 4662 (2018).

[43] W. Handley *et al.*, AEONS: Approximate End of Nested Sampling (to be published).

[44] H. S. Wilf, in *Generatingfunctionology*, edited by A. K. Peters (A K Peters, MA, 2006).

[45] S. Gariazzo and O. Mena, Cosmology-marginalized approaches in Bayesian model comparison: The neutrino mass as a case study, Phys. Rev. D **99**, 021301 (2019).

[46] J. Martin, C. Ringeval, and V. Vennin, Information gain on reheating: The one bit milestone, Phys. Rev. D **93**, 103532 (2016).

[47] CORE Collaboration, Exploring cosmic origins with CORE: Inflation, J. Cosmol. Astropart. Phys. 04 (2018) 016.

[48] E. Higson, W. Handley, M. Hobson, and A. Lasenby, Bayesian sparse reconstruction: A brute-force approach to astronomical imaging and machine learning, Mon. Not. R. Astron. Soc. **483**, 4828 (2019).

[49] A. R. Liddle, How many cosmological parameters?, Mon. Not. R. Astron. Soc. **351**, L49 (2004).

[50] H. Akaike, A new look at the statistical model identification, IEEE Trans. Autom. Control **19**, 716 (1974).

[51] G. Schwarz, Estimating the dimension of a model, Ann. Stat. **6**, 461 (1978).

[52] http://www.cosmos.esa.int/web/planck/pla.

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

[54] http://www.sdss3.org/science/BOSS_publications.php.

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

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

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

[58] W. J. Handley, CosmoChord 1.15, https://doi.org/10.5281/zenodo.2552056 (2019).

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

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

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

[62] W. Handley, Anesthetic: Nested sampling visualisation, J. Open Source Software **4**, 01414 (2019).

[63] J. A. Nelder and R. Mead, A simplex method for function minimization, Computer Journal (UK) **7**, 308 (1965).