PHYSICAL REVIEW D **73,** 067302 (2006)

# Bayesian evidence as a tool for comparing datasets

Phil Marshall
*Kavli Institute for Particle Astrophysics and Cosmology, P.O. Box 20450, MS29, Stanford, California 94309, USA*

Nutan Rajguru
*Astrophysics Group, Cavendish Laboratory, Madingley Road, Cambridge, United Kingdom*

Anže Slosar
*Faculty of Mathematics and Physics, University of Ljubljana, Ljubljana, Slovenia*
(Received 9 November 2005; published 17 March 2006)

We introduce a new conservative test for quantifying the consistency of two or more datasets. The test is based on the Bayesian answer to the question, ''How much more probable is it that all my data were generated from the same model system than if each dataset were generated from an independent set of model parameters?'' We make explicit the connection between evidence ratios and the differences in peak chi-squared values, the latter of which are more widely used and more cheaply calculated. Calculating evidence ratios for three cosmological datasets [recent cosmic microwave background data (WMAP, ACBAR, CBI, VSA), SDSS galaxy redshift survey, and the most recent SNe type 1A data] we find that concordance is favored and the tightening of constraints on cosmological parameters is indeed justified.

DOI: 10.1103/PhysRevD.73.067302 PACS numbers: 98.80.Cq

## I. INTRODUCTION

The apparent mutual agreement of a wide range of cosmological observations has led to the current climate of ''concordance'' in cosmology [1–8]. The practice of combining independent datasets, by the multiplication of their associated likelihood functions, in order to increase the precision of the parameters of the world model is now standard, but quantitative consistency checking is emphasized to a much lesser degree. As all physicists will agree, accurate cosmology is preferable to precision cosmology, and it is this that motivates this short communication.

The purpose of this work is to demonstrate one application of Bayesian model selection, that of checking that the far simpler model of a universal set of parameters for modeling all datasets is justified by the data themselves: in doing so we make the connection between the Bayesian formulation of the problem and the pragmatic approach taken at much lower computational cost by the experimental community. In this work we show that, as is so often the case, the standard approach is justified on the grounds of common sense, and demonstrate the reduction of this common sense to calculation via probability theory.

As usual, the route to model selection is via the Bayesian evidence. The evidence for a model $\mathsf{H}$ from data $\boldsymbol{d}$ is just the probability $\Pr(\boldsymbol{d} \mid \mathsf{H})$, and can be calculated in principle by marginalizing the unnormalized posterior probability distribution function (pdf) over all $M$ parameters $\boldsymbol{\theta}$ in the model:

$$\Pr(\boldsymbol{d} \mid \mathsf{H}) = \int \Pr(\boldsymbol{d} \mid \boldsymbol{\theta}, \mathsf{H}) \Pr(\boldsymbol{\theta} \mid \mathsf{H}) d^M \boldsymbol{\theta}. \quad (1)$$

In practice, calculating this integral is rarely feasible, but other techniques exist to provide estimates of the evidence (see, for example, [9]). More detailed introductions to the evidence and its central role in the problem of model selection are available elsewhere [10,11]—here we make the general remarks that the evidence increases sharply with increasing goodness of fit, and decreases with increasing model complexity (quantifying the principle of Occam's razor). We show later explicitly how these two aspects come to the surface and, for the specific case of Gaussian measurement errors, result in model selection proceeding by the comparison of differences in the ubiquitous chi-squared statistic with an ''Occam'' factor which takes the form of an effective number of parameters. The more general approach advocated here is applicable to any likelihood functions [$\Pr(\boldsymbol{d} \mid \boldsymbol{\theta}, \mathsf{H})$], not just those having Gaussian form, and takes into account the full extent of the pdf's involved. It is of course also sensitive to the parameters' prior pdf [$\Pr(\boldsymbol{\theta} \mid \mathsf{H})$]: broader priors represent more complex models and so naturally give lower evidence values. Evidence is the natural tool for comparing datasets in this way: it enables us to quantify such questions as ''Is the mismatch between two experiments large enough to warrant investigation into possible sources of systematics or new physics?''

The simplest model for all the cosmological data in hand is that they provide information on the same set of cosmological parameters: this is the standard assumption made in all the joint analyses to date. Let $\mathsf{H}_0$ represent the hypothesis that ''there is one set of parameters that describes our cosmological model.'' In other words, we believe that we understand both cosmology and our experiments to the extent that there should be no further freedom beyond the parameters specified. However, if we are interested in accuracy as well as precision then we should take care to allow for systematic differences between datasets: the most

 

extreme case would be the one where the observations were in such strong disagreement that they appeared to give conflicting measurements of all the model parameters. In this case one could consider the hypothesis $\mathsf{H}_1$ that ''there is a different set of parameters for modeling each dataset.'' The conservatism of such a model comparison exercise is readily apparent: the large increase in model complexity incurred when moving from $\mathsf{H}_0$ to $\mathsf{H}_1$ means that the joint analysis is intrinsically more favorable. This means that any result in favor of $\mathsf{H}_1$ may be taken as a clear indication of discord between the two experiments. Note also that this test is easily done given that the evidence values will have been calculated for alternative purposes, such as comparing two physical models in the light of each dataset alone.

For checking dataset consistency then the quantity we should calculate is the ratio of probabilities that each model is correct, given the data:

$$\frac{\Pr(\mathsf{H}_0 \mid \boldsymbol{d})}{\Pr(\mathsf{H}_1 \mid \boldsymbol{d})} = \frac{\Pr(\boldsymbol{d} \mid \mathsf{H}_0)}{\Pr(\boldsymbol{d} \mid \mathsf{H}_1)} \cdot \frac{\Pr(\mathsf{H}_0)}{\Pr(\mathsf{H}_1)}. \tag{2}$$

The calculable part of Eq. (2) is the evidence ratio

$$\mathrm{R} = \frac{\Pr(\boldsymbol{d} \mid \mathsf{H}_0)}{\Pr(\boldsymbol{d} \mid \mathsf{H}_1)} = \frac{\Pr(\boldsymbol{d} \mid \mathsf{H}_0)}{\prod_i \Pr(\boldsymbol{d}_i \mid \mathsf{H}_1)}, \tag{3}$$

where in the second line we have assumed that the individual datasets $\boldsymbol{d}_i$ under analysis are independent. (The evidence integral factors out since the independent likelihoods do, and also because each likelihood depends only on its own subset of parameters.) Interpretation of this evidence ratio is aided by Eq. (2): for statement $\mathsf{H}_0$ to be more likely to be true than statement $\mathsf{H}_1$, the product of $R$ and the prior probability ratio must be greater than 1. Suppose that an evidence ratio $R$ of 0.1 were found: the dataset combination ($\mathsf{H}_0$) can still be justified, but only if you are willing to take odds of ten to one on there being no significant systematic errors in the system. Blindly multiplying $N$ likelihoods together results, in general and approximately, in factors of improvement in precision of $\sqrt{N}$: the evidence ratio gives an indication of whether or not this improvement is justified, in the form of an odds ratio (which enforces honesty through the threat of bankruptcy).

Other criteria besides evidence have been used to compare different models. Recently [12] have proposed the Akaike and Bayesian information criteria to carry out cosmological model selection. These criteria are approximations to the full Bayesian evidence under rather restrictive assumptions and thus fall under the same framework. The posterior Bayes factors proposed by [13] and also discussed in [14] can be used as an alternative to evidence. This quantity is the Bayesian evidence with the prior set to the posterior and can be readily estimated as an average likelihood of the Markov chains. It has some desirable properties, such as no prior dependence in the limit of prior enclosing the entire volume of posterior. However, it has no simple interpretation within the Bayesian framework and will thus not be discussed in this paper. The use of the evidence itself as a model selection tool has been growing in cosmology (see e.g. [15–18]). Using evidence to check dataset consistency has received much less attention. Application to a particular problem of CMB map contamination can be found in [16]. In this work we construct a much more general approach that can be applied to any setting in which a given model is tested against more than one dataset. The price one has to pay for this generality is that we are relatively insensitive to any *particular* inconsistency. We also aim to provide a short tutorial, establishing the connection with the more conventional $\chi^2$ statistics, followed by a simple analysis of current state-of-the-art experiments.

## II. CONNECTION TO $\chi^2$ ANALYSIS

Consider a general likelihood function of some model parameter vector $\boldsymbol{x}$, which can be (for reasons that will become apparent in a moment) rewritten as

$$L(\boldsymbol{x}) = L_{\max}\hat{L}(\boldsymbol{x}), \tag{4}$$

where $L_{\max}$ is the likelihood at the most likely point in the parameter space and the dimensionless function $\hat{L}$ contains all the likelihood shape information. Assuming a uniform prior spanning between $-p$ and $p$ in each direction, where $p$ is large enough to encompass all regions of high likelihood, gives the approximate evidence

$$\tilde{E} \approx L_{\max}\frac{\int \hat{L} d^M \boldsymbol{x}}{(2p)^M}, \tag{5}$$

where $M$ is the number of parameters in the model. If we identify the numerator of the above fraction with the volume associated with the likelihood $V_L$, and the denominator with the available prior volume $V_\pi$, we have

$$\log E = \log\left(\frac{V_L}{V_\pi}\right) + \log L_{\max}. \tag{6}$$

All the details of the overlap between prior and likelihood is contained within the volume ratio, whereas the maximum likelihood value specifies the goodness of fit. Except when the posterior pdf's take simple analytic forms, this volume factor must be calculated numerically and of course takes up much of the effort in the evidence calculation.

In the case where the measurement errors are Gaussian, we can write the evidence ratio used in this work in terms of the best-fit chi-squared values that may be calculated during an analysis. It can be shown that

$$\log R = \log\left(\frac{V_{12}V_\pi}{V_1 V_2}\right) - \frac{1}{2}\Delta\hat{\chi}^2, \tag{7}$$

where $\Delta\hat{\chi}^2 = \hat{\chi}^2_{12} - (\hat{\chi}^2_1 + \hat{\chi}^2_2)$. Defined this way, $\Delta\hat{\chi}^2$ is

always positive (the goodness of fit cannot decrease with the addition of the extra parameters) and we see that the borderline case of $\log R = 0$ corresponds to the difference in chi squared between the two individual analyses and the joint fit being equal to an effective number of parameters (the difference in the number of degrees of freedom) given by the logarithm of the volume factor.

Returning to the general case, if we retain the assumption of a broad uniform prior, and if the likelihoods are well approximated by multivariate Gaussians, then the volume factor can be calculated analytically: in this case the $i$th likelihood can be written as

$$L_i \approx \hat{L}_i \exp[-\tfrac{1}{2}(\boldsymbol{x} - \hat{\boldsymbol{x}}_i)^T F_i^{-1}(\boldsymbol{x} - \hat{\boldsymbol{x}})], \tag{8}$$

where $F_i$ is the Fisher matrix. This gives, for the likelihood volumes,

$$V_i = (2\pi)^{M/2}|F_i|^{1/2}. \tag{9}$$

In the joint analysis, combining two Gaussian likelihoods results in a new Gaussian, centered at a correctly weighted mean of positions, but whose shape is given simply by

$$\hat{L}_{12} = \exp[-\tfrac{1}{2}(\boldsymbol{x} - \hat{\boldsymbol{x}}_i)^T (F_1 + F_2)^{-1}(\boldsymbol{x} - \hat{\boldsymbol{x}}_i)], \tag{10}$$

and therefore

$$V_{12} = (2\pi)^{M/2}|F_1 + F_2|^{1/2}. \tag{11}$$

Note that in this case, due to the high symmetry of the Gaussian approximation, the overlap integral $V_{12}$ is independent of the distance between the best fitting points. Therefore using $\Delta\chi^2$ as a proxy for the Bayesian evidence change is valid when the Gaussian approximation to the posterior is a good one. In the simple case where $F_1 = F_2$ (a parallel degeneracy) and $V_\pi = (2p)^M$ again, the log evidence becomes

$$\log R = \frac{1}{2}\left(M \log\left[\frac{4}{\pi}\frac{p^2}{|F|^{1/M}}\right] - \Delta\chi^2\right). \tag{12}$$

The log term is typically of the order of unity: $|F|^{1/M}$ is the geometrical average of the principal variances and hence $p^2|F|^{-1/M}$ is the square of the ratio of the prior width to the characteristic likelihood width. Hence we recover the frequentist rule of thumb that the increase in $\chi^2$ is justified if the number of parameters drops by roughly the same number. However the evidence considerations above allow this rule to be calibrated to take into account both the prior information supplied and the (potentially complex) shape of the likelihoods; in general, $V_{12}$ is not independent of the individual peak positions, and so the simple $\Delta\hat{\chi}^2$ procedure does not propagate all the information contained within the likelihood functions.

TABLE I. The priors assumed for the cosmological model considered in this paper. The notation $(a, b)$ for parameter $x$ denotes a top-hat prior in the range $a < x < b$.

| Basic parameter | Prior |
|---|---|
| $\omega_b$ | (0.005, 0.05) |
| $\omega_{\rm dm}$ | (0.01, 0.4) |
| $\Omega_k$ | ( − 0.3, 0.3) |
| $h$ | (0.4, 0.9) |
| $n_s$ | (0.8, 1.2) |
| $\tau$ | (0.01, 0.7) |
| $\log 10^{10} A_s$ | (1, 5) |

## III. COMPARING COSMOLOGICAL DATASETS

### Datasets and method

We use a version of the COSMOMC software package [14], modified to calculate evidence by the thermodynamic integration method. We obtain consistent results using two different methods to calculate the evidence reliably: the error on the log evidence differences is conservatively estimated to be of the order of unity. The details of the evidence calculation method is presented elsewhere [19].

We have chosen three datasets for comparison:

(i) CMB: We use the "standard" selection of CMB experiments: the WMAP data [20] together with the latest VSA [21], CBI [3] and ACBAR data [22]. We also used a modified version of the likelihood code that correctly accounts for the largest WMAP scales [23]

(ii) SN: We use the Riess *et al.* (2004) SN data. We use both "gold" and "silver" datasets. We implemented our likelihood code and checked that it gives results consistent with Riess *et al.*

(iii) SDSS: Finally we use large scale power spectrum measurements from the SDSS experiment [24–26]. We used the likelihood code by Tegmark [7] adapted for COSMOMC by Leach (private communication).

We investigate a 7-parameter cosmological model. In Table I we show the uniform priors assumed for the parameters of our model. We take our priors to be comparatively broad to approximate the state of ignorance we may have been in before any of the three experiments were performed. This has the effect of giving the data as much "freedom" as possible, and correspondingly making the evidence test somewhat conservative.

## IV. RESULTS AND DISCUSSION

In Table II we give the values of $R$ for various combinations of datasets under discussion. We do not detect any discrepancy between datasets: all combinations of the datasets weakly favor $\mathsf{H}_0$. In the last line of Table II we report on the value of $R$ for all experiments combined. In principle, it is possible to have three experiments be pairwise

TABLE II. The logarithm of $R$ for various combinations of datasets. See text for further discussion.

| Dataset combination | log$R$ |
|---|---|
| CMB-SDSS | 0.23 |
| SDSS-SN | 1.5 |
| SN-CMB | 1.6 |
| CMB-SDSS-SN | 4.5 |

consistent with each other, but not when all combined together (imagine, for example, three degeneracy lines forming a triangle). Comfortingly enough, the three-way evidence test also abrogates $\mathsf{H}_1$ and due to a large number of extra parameters (i.e. twice as many as in other datasets) it has also a more positive detection of concordance.

We have illustrated our methodology with application to real cosmological data. As expected, the data are concordant: any obvious conflict in the data would likely have been noticed using the ''chi by eye'' methods employed to date. However, should such discrepancies occur in the future it is imperative to have a method to quantify these discrepancies in the most general settings where Gaussianity cannot be assumed and ever more complex parameter spaces are to be dealt with.

A value of $R$ less than unity ( $\log R < 0$) is a sign that we should investigate the mismatch between datasets further. This can be done by exploring more focused models, either with new cosmological parameters (if the experiments are reckoned to be well understood), or with additional nuisance parameters that quantify the possible systematic errors in the data. Disentangling the degeneracy between new physics and systematic error can only be done if the additional parameters come with fresh information encoded in their prior pdf: this information is then folded into the evidence ratio, providing the crucial difference between this methodology and any method relying on goodness of fit alone.

## ACKNOWLEDGMENTS

We thank Mike Hobson, Sarah Bridle, Jo Dunkley, Andrew Liddle, Uroš Seljak, and Antony Lewis for useful discussions. We also acknowledge Andrew Jaffe for advice on a previous version of this work. N. R. is supported by PPARC. A. S. is supported by the Ministry of Education, Science and Sport of the Republic of Slovenia. This work was supported in part by the U.S. Department of Energy under Contract No. DE-AC02-76SF00515.

[1] R. A. C. Croft, D. H. Weinberg, M. Bolte, S. Burles, L. Hernquist, N. Katz, D. Kirkman, and D. Tytler, Astrophys. J. **581**, 20 (2002).
[2] C.-L. Kuo *et al.*, Bull. Am. Astron. Soc. **34**, 1324 (2002).
[3] A. C. S. Readhead *et al.*, Astrophys. J. **609**, 498 (2004).
[4] R. Rebolo *et al.*, Mon. Not. R. Astron. Soc. **353**, 747 (2004).
[5] A. G. Riess *et al.*, Astrophys. J. **607**, 665 (2004).
[6] D. N. Spergel *et al.*, Astrophys. J. **148**, 175 (2003).
[7] M. Tegmark *et al.*, Phys. Rev. D **69**, 103501 (2004).
[8] L. Verde *et al.*, Mon. Not. R. Astron. Soc. **335**, 432 (2002).
[9] J. J. K. Ó'Ruanaidh and W. J. Fitzgerald, *Numerical Bayesian Methods Applied to Signal Processing* (Springer-Verlag, New York, 1996).
[10] C. M. Bishop, *Neural Networks for Pattern Recognition* (Oxford University Press, New York, 1995).
[11] D. S. Sivia, *Data Analysis: A Bayesian Tutorial* (Oxford University Press, New York, 1996).
[12] A. R. Liddle, Mon. Not. R. Astron. Soc. **351**, L49 (2004).
[13] M. Aitkin, J. Royal Stat. Soc. B **53**, 111 (1991).
[14] A. Lewis and S. Bridle, Phys. Rev. D **66**, 103511 (2002).
[15] A. Jaffe, Astrophys. J. **471**, 24 (1996).
[16] L. Knox, J. R. Bond, A. H. Jaffe, M. Segal, and D. Charbonneau, Phys. Rev. D **58**, 083004 (1998).
[17] M. P. Hobson, S. L. Bridle, and O. Lahav, Mon. Not. R. Astron. Soc. **335**, 377 (2002).
[18] R. Trotta, astro-ph/0504022.
[19] M. Beltrán, J. García-Bellido, J. Lesgourgues, A. R. Liddle, and A. Slosar, Phys. Rev. D **71**, 063532 (2005).
[20] G. Hinshaw *et al.*, Astrophys. J. Suppl. Ser. **148**, 135 (2003).
[21] C. Dickinson *et al.*, Mon. Not. R. Astron. Soc. **353**, 732 (2004).
[22] C. L. Kuo *et al.*, Astrophys. J. **600**, 32 (2004).
[23] A. Slosar, U. Seljak, and A. Makarov, Phys. Rev. D **69**, 123003 (2004).
[24] K. Abazajian *et al.*, Astron. J. **126**, 2081 (2003).
[25] C. Stoughton *et al.*, Astron. J. **123**, 485 (2002).
[26] D. G. York *et al.*, Astron. J. **120**, 1579 (2000).