SLAC-PUB-10921

Mon. Not. R. Astron. Soc. **000**, 1–6 (2004) Printed 15 December 2004 (MN LATEX style file v2.2)

Bayesian evidence as a tool for comparing datasets

Phil Marshall[1], Nutan Rajguru[2], Anže Slosar[3]
[1] *Kavli Institute for Particle Astrophysics and Cosmology, Stanford University PO Box 20450, MS29, Stanford, CA 94309, USA*
[2] *Astrophysics Group, Cavendish Laboratory, Madingley Road, Cambridge CB3 0HE, UK.*
[3] *Faculty of Mathematics and Physics, University of Ljubljana, Jadranska 19, 1000 Ljubljana, Slovenia*

Accepted —; received —; in original form 15 December 2004

ABSTRACT
We introduce a 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?". The behaviour of the resulting odds ratio is demonstrated with a two-dimensional toy example. We make explicit the connection between evidence ratios and the differences in peak-chisquared values, the later of which is more widely used and more cheaply calculated. Calculating evidence ratios for three cosmological datasets (recent CMB data (WMAP, ACBAR, CBI, VSA), SDSS and the most recent SNe Type 1A data) we find that concordance is favoured and the tightening of constraints on cosmological parameters is indeed justified.

Key words: methods: statistical, cosmological parameters

1 INTRODUCTION

The apparent mutual agreement of a wide range of cosmological observations has led to the current climate of "concordance" in cosmology (Spergel et al. 2003; Rebolo et al. 2004; Readhead et al. 2004; Kuo et al. 2002; Tegmark et al. 2004; Riess et al. 2004; Croft et al. 2002; Verde et al. 2002) The practice of combining datasets to increase the precision of the parameters of the world model is now standard, but quantitative consistency checking is emphasised 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 modelling 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.

Independent datasets may be combined by the multiplication of their associated likelihood functions; this is the foundation of the joint analyses that that have led to the precision inference of cosmological parameters. If the "second level" inference of the most appropriate model for the data is undertaken at all, it is usually done in a somewhat qualitative way. 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 H from data $\boldsymbol{d}$ is just the probability $\Pr(\boldsymbol{d}|\mathsf{H})$, and can be calculated in principle by marginalising the unnormalised posterior probability distribution function over all M parameters $\boldsymbol{\theta}$ in the model:

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

In practice calculating the integral is rarely feasible, but other techniques exist to provide estimates of the evidence (see, for example, Ó'Ruanaidh & Fitzgerald 1996). More detailed introductions to the evidence and its central role in the problem of model selection are available elsewhere (Sivia 1996; Bishop 1995) – 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 ($\Pr(\boldsymbol{d}|\boldsymbol{\theta},\mathsf{H})$), 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 ($\Pr(\boldsymbol{\theta}|\mathsf{H})$): 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

Work supported in part by the Department of Energy Contract DE-AC02-76SF00515
Stanford Linear Accelerator Center, Kavli Institute for Particle Astrophysics and Cosmology, Stanford University, Stanford, CA 94309

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 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 H_1 that "there is a different set of parameters for modelling each dataset." The conservatism of such a model comparison exercise is readily apparent: the large increase in model complexity incurred when moving from H_0 to H_1 means that the joint analysis is intrinsically more favourable. This means that any result in favour of 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:

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

The calculable part of Equation 2 is the evidence ratio

$$\begin{aligned} \mathrm{R} &= \frac{\Pr(\boldsymbol{d}|\mathsf{H}_0)}{\Pr(\boldsymbol{d}|\mathsf{H}_1)} \\ &= \frac{\Pr(\boldsymbol{d}|\mathsf{H}_0)}{\prod_i \Pr(\boldsymbol{d}_i|\mathsf{H}_1)} \end{aligned} \tag{3}$$

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 Equation 2: for statement H_0 to be more likely to be true than statement H_1, the product of R and the prior probability ratio must be greater than one. Values of $\mathrm{R} \sim 1$ indicates a satisfactory compliance of the data, while $\mathrm{R} \ll 1.0$ indicated that H_1 is favoured. Suppose that an evidence ratio R of 0.1 were found: the dataset combination (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 is enforces honesty through the threat of bankruptcy).

The outline of the rest of this paper is as follows. In section 2 we illustrate the above theory with some toy examples. In the folllowing section (section 3) we establish the connection with the more conventional χ^2 statistics. Section 4 contains our analysis of current state-of-the-art data with the outlined method and we then conclude with a brief discussion (section 5).

2 ILLUSTRATION: 2-DIMENSIONAL GAUSSIAN LIKELIHOODS

In order to illustrate the behaviour of the evidence ratio in an easily visualised situation, we imagine a pair of likelihood functions (modelling the input of two independent experiments) who happen to have the same Gaussian form when plotted as a function of the two model parameters, x and y. Applying uniform priors broad enough to contain the high likelihood regions results in familiar elliptical error contours of constant posterior probability. The normalisation N_i of the likelihood functions in the data space is unimportant: for clarity we include it for it to be cancelled explicitly later. Denoting the width of both distributions in the two principal directions by σ_a and σ_b, the relation between the two likelihoods is specified by the separation of the two maxima (with components of the position vector denoted by hats) and the orientations of the ellipse axes θ_i. Without loss of generality we can align the principal directions of the first ellipse with our coordinate system and put it at the origin. The two likelihoods can then be written as:

$$L_1 = N_1 G(x, y; \sigma_a, \sigma_b, 0, 0, 0), \tag{4}$$

$$L_2 = N_2 G(x, y; \sigma_a, \sigma_b, \hat{x}, \hat{y}, \theta), \tag{5}$$

where

$$G(x, y; \sigma_a, \sigma_b, \hat{x}, \hat{y}, \theta) = \exp\left[-\frac{u^2}{2\sigma_a^2} - \frac{v^2}{2\sigma_b^2}\right], \tag{6}$$

with

$$\begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} x - \hat{x} \\ y - \hat{y} \end{pmatrix}. \tag{7}$$

Suppose we were to analyse each dataset individually first: the evidence associated with each model is given by the integral of likelihood over prior. Assuming a flat prior between $-p$ and p for both x and y, we have

$$E_i = \int_{-p}^{p}\int_{-p}^{p} \frac{L_i(x,y)}{4p^2} \mathrm{d}x\mathrm{d}y. \tag{8}$$

If the prior extends far enough to encompass the entire region of high likelihood, one can approximate

$$E_i \approx \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{L_i(x,y)}{4p^2} \mathrm{d}x\mathrm{d}y \approx \frac{2\pi\sigma_a\sigma_b}{4p^2} N_i. \tag{9}$$

The evidence for the joint analysis of the two datasets is the integral of the product of the likelihoods over the prior:

$$E_{12} \approx \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{L_1 L_2}{4p^2} \mathrm{d}x\mathrm{d}y. \tag{10}$$

The evidence ratio introduced above is then given by

$$\mathrm{R} \sim \frac{E_{12}}{E_1 E_2} = \frac{4p^2}{(2\pi\sigma_a\sigma_b)^2} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{L_1}{N_1}\frac{L_2}{N_2} \mathrm{d}x\mathrm{d}y, \tag{11}$$

and is analytically calculable. We now investigate three interesting cases.

2.1 Parallel degeneracies

Suppose our two datasets have parallel degeneracies. We thus set $\theta = 0$ and $\hat{y} = 0$, and investigate the evidence ratio as a function of $\hat{x}$ and p. The logarithm of the evidence ratio in this case is given by

$$\log \mathrm{R} = \log \frac{p^2}{\pi \sigma_a \sigma_b} - \frac{\hat{x}^2}{4\sigma_a^2}. \qquad (12)$$

Note that the evidence ratio is a function of the ratio of p^2 to σ_a^2: this quantity represents the fraction of the prior volume occupied by the likelihood, and provides the quantification of Occam's razor. In the limit where our approximation is valid (i.e. $p^2 \gg \sigma_a \sigma_b$), the first term is always positive and the log evidence is negative only if the mismatch is several "sigma", $\hat{x}/\sigma_a \gg 1$. The dependence of the evidence ratio on the Occam factor is linear, and provides a gradual increase in conservatism as the prior is widened. In contrast, the dependence on the mismatch between the positions of the likelihood is exponential – it is here that the evidence ratio is most sensitive, offering very long odds against the dataset combination.

2.2 Perpendicular degeneracy

Perpendicular degeneracies are a much more common occurrence; for example the supernovae and CMB are nearly orthogonally degenerate. We do the same as in the previous sub-section, but set $\theta_2 = \pi/2$. The following expression of the $\log \mathrm{R}$ is then obtained:

$$\log \mathrm{R} = \log \frac{2p^2}{\pi\,(\sigma_a^2 + \sigma_b^2)} - \frac{\hat{x}^2}{2(\sigma_a^2 + \sigma_b^2)}. \qquad (13)$$

We see a similar behaviour as before, but now the evidence falls slowly if the width σ_a is large: if we have two elongated Gaussians they are much more likely to find a common intersection.

We plot the value of R for several values of θ in the Figure 1. Note that for plausible values of the prior (even of the order 10 times bigger than the size of high likelihood), the R never prefers H_0 with a large value of significance. On the other hand it is very strong at disfavouring H_0 in case of poor data mismatch. This plot also shows how complementary degeneracies decrease the ability to find discrepancies by this statistical method. We conclude that the complementary degeneracy is thus beneficial from the point of view of parameter estimation, while a parallel degeneracy is beneficial from the point of view of model selection.

The Figure 2 shows the arrangement of likelihood ellipses for the two models discussed in the Figure 1 at a particular value of $\hat{x} = 7$. We see how the evidence test claims incompatibility between datasets that have no overlap of likelihood contours.

2.3 Extension to more complex likelihoods

It is often the case that the likelihoods are functions of many more than the two variables used in this illustration: marginalising the ensuing posterior provides two-dimensional projections of the combined likelihood which may hide the failure of the likelihoods to overlap in the higher dimension space. Conditioning on all but two parameters allows slices through the posterior to be plotted, but these suggest underestimated parameter uncertainties (and so can give a false impression of dataset mismatch).

We stress that the illustration of the previous subsections used a Gaussian simply because it is a convenient functional form with which to develop intuition about the test we discuss here. However, the method of calculating evidence ratios is completely general and does not depend on any assumptions about the shape of likelihood.

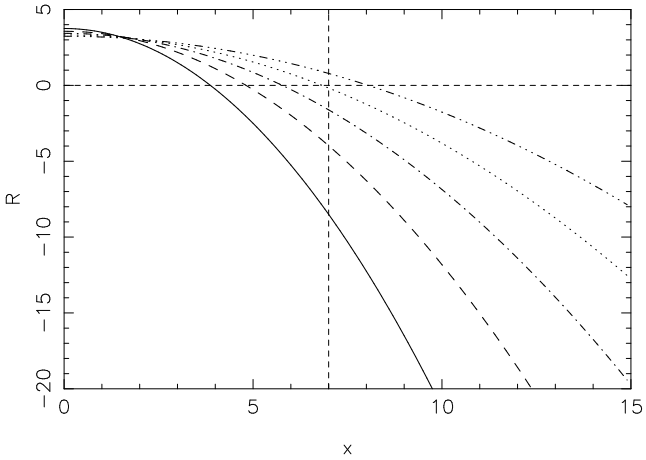

Figure 1. In this figure we plot the value of R as a function of $\hat{x}$. Values of other parameters are $\sigma_a = 1$, $\sigma_b = 3$ and $p = 20$. Different linestyles corresponds to different values of θ_2: 0 (solid), $\pi/6$ (dashed), $\pi/4$ (dot dashed), $\pi/3$ (dotted) and $\pi/2$ (triple dot dashed). The vertical line correspond to $\hat{x} = 7$ - for this particular value of $\hat{x}$ we plot the likelihood ellipses in the Figure 2. See text for further discussion.

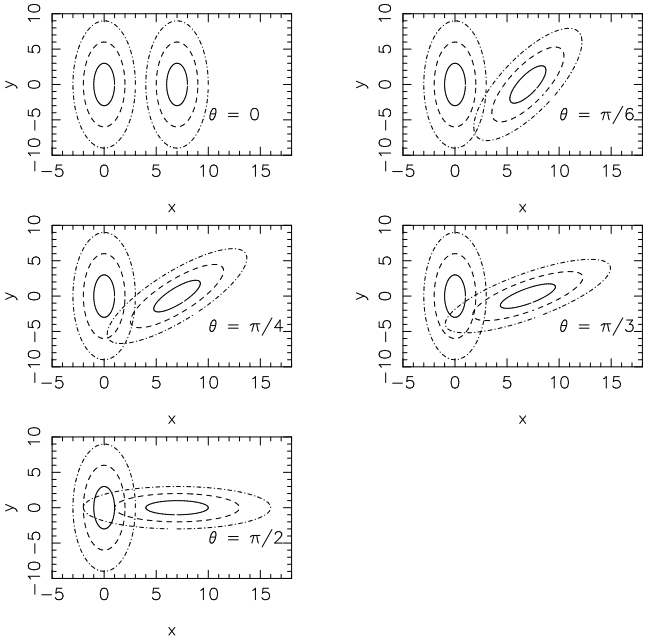

Figure 2. Likelihood ellipses for the example plotted in Figure 1. The value of theta is printed inset and the value of $\hat{x} = 7$ (dashed vertical line in the Figure 1). Filled, dashed and dot-dashed linestyles correspond to one, two and three sigma respectively.

2.4 Datasets with different amounts of constraining power

It is natural to ask how the evidence ratio responds to datasets with different amounts of constraining power. (In our toy model this would correspond to allowing the widths σ_a and σ_b to differ bewteen likelihoods.) As an extreme example of this we may consider a second dataset with no constraining power. Assuming L_2 =constant leads to a cancellation in the evidence ratio such that the evidence for two datasets taken together equals that of the first dataset alone. The result R = 1 is not unexpected – the second dataset is completely uninformative and therefore it is impossible to judge either how well or how bad it fits the model. The safest bet is therefore to assume that hypothesises H_0 and H_1 are equally likely. Likewise, we can expect the evidence ratio to be close to unity whenever one dataset provides much less information than the other.

3 CONNECTION TO χ^2 ANALYSIS

Following the evidence analysis of the previous section we might now wonder if this result can be (and indeed has been) arrived at through other means. The answer is of course yes, in the sense that all good statistics can be derived from reasonable questions posed in the language of probabilities. The steep dependence of the evidence ratio at large likelihood centroid offsets is due to the decreasing goodness of fit incurred at these points: this is visible at some level in almost any statistic that depends on this goodness-of-fit. More quantitatively, the decrease in fit quality is measured (exactly in the case of Gaussian measurement errors) by the minimum value of the chi-squared statistic, and more generally by the maximum log likelihood. 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

$$L(\boldsymbol{x}) = L_{\max}\hat{L}(\boldsymbol{x}), \tag{14}$$

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.

As before, assuming a flat prior spanning between $-p$ and p in each direction, where p is large enough to encompass all region of high likelihood, gives the approximate evidence

$$\tilde{E} = L_{\max}\frac{\int \hat{L}\mathrm{d}^M\boldsymbol{x}}{(2p)^M}, \tag{15}$$

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_π, we have

$$\log E = \log\left(\frac{V_L}{V_\pi}\right) + \log L_{\max}. \tag{16}$$

All the details of the overlap between prior and likelihood is contained with in 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. From equation (16) we can read off that

$$\log R = \log\left(\frac{E_{12}}{E_1E_2}\right) \tag{17}$$

$$= \log\left(\frac{V_{12}V_\pi}{V_1V_2}\right) + \log\hat{L}_{12} - (\log\hat{L}_1 + \log\hat{L}_2), \tag{18}$$

and, cancelling the likelihood normalisations, get that

$$\log R = \log\left(\frac{V_{12}V_\pi}{V_1V_2}\right) - \frac{1}{2}\Delta\hat{\chi}^2, \tag{19}$$

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 (difference in 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

$$L_i \approx \hat{L}_i \exp\left[-\frac{1}{2}(\boldsymbol{x} - \hat{\boldsymbol{x}}_i)^T F_i^{-1}(\boldsymbol{x} - \hat{\boldsymbol{x}}_i)\right], \tag{20}$$

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

$$V_i = (2\pi)^{M/2}\,|F_i|^{1/2}\,. \tag{21}$$

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

$$\hat{L_{12}} = \exp\left[-\frac{1}{2}(\boldsymbol{x} - \hat{\boldsymbol{x}}_i)^T\,(F_1 + F_2)^{-1}\,(\boldsymbol{x} - \hat{\boldsymbol{x}}_i)\right], \tag{22}$$

and therefore

$$V_{12} = (2\pi)^{M/2}\,|F_1 + F_2|^{1/2}\,. \tag{23}$$

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 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:

$$\log R = \frac{1}{2}\left(M\log\left[\frac{4}{\pi}\frac{p^2}{|F|^{1/M}}\right] - \Delta\chi^2\right). \tag{24}$$

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 χ^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

Table 1. 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
$\omega_{\rm b}$	$(0.005, 0.05)$
$\omega_{\rm dm}$	$(0.01, 0.4)$
$\Omega_{\rm k}$	$(-0.3, 0.3)$
h	$(0.4, 0.9)$
n_s	$(0.8, 1.2)$
τ	$(0.01, 0.7)$
$\log 10^{10} A_s$	$(1, 5)$

does not propagate all the information contained within the likelihood functions.

4 COMPARING COSMOLOGICAL DATASETS

4.1 Datasets and method

We use a version of the `CosmoMC` software package (Lewis & Bridle 2002), modified to calculate evidence by the thermodynamic integration method. We used two different methods to calculate the evidence reliably and got consistent results. The error on log evidence differences is conservatively estimated to be of the order of unity. The details of the evidence calculation method will be presented elsewhere (Beltran 2004).

We have chosen three datasets for comparison:

- CMB. We use the "standard" selection of CMB experiments: the WMAP data (Hinshaw et al. 2003) together with latest VSA (Dickinson et al. 2004), CBI (Readhead et al. 2004) and ACBAR data Kuo et al. (2004). We also used a modified version of the likelihood code that correctly accounts for the largest WMAP scales (Slosar et al. 2004)
- 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.
- SDSS. Finally we use large scale power spectrum measurements from the SDSS experiment (York et al. 2000; Stoughton et al. 2002; Abazajian et al. 2003). We used the likelihood code by Tegmark (Tegmark et al. 2004) adapted for `CosmoMC` by Samuel Leach (private communication).

We investigate a 7-parameter concordant cosmological model. In table 1 we show the flat priors taken on the parameters of our model, where symbols have their usual meaning. We take our priors to be comparatively wide 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.

4.2 Results

In table 2 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 favour H_0. As discussed in section 2 the R is never very large, unless we use an extremely wide prior range, but the fact that no ratios less than unity are observed is in agreement with the many claims of concordance in the cosmological parameters. It seems that there are no large mismatches hidden in the multi-dimensional parameter space.

Table 2. In this table we show the logarithm of R for various combinations of datasets. The error on evidence difference values is estimated to be of the order of unity. See text for further discussion.

Dataset combination	$\log \mathrm{R}$
CMB – SDSS	0.23
SDSS – SN	1.5
SN – CMB	1.6
CMB – SDSS – SN	4.5

In the last line of the table 2 we report on the value of R for all experiments combined. In principle, it is possible to have three experiments consistent with each other but not when combined together (imagine, for example three degeneracy lines forming a triangle). Comfortingly enough, this test also gave a null result 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.

5 DISCUSSION

In this paper we have investigated the evidence as a tool to compare datasets. Evidence is usually used as a tool to compare models: all that we have done here is asked "Can a given model explain all the data at once, or is a more complex model required?" The particular alternative we use here is that where each dataset has its own set of parameters to model it, and as such we are performing the least sensitive test of dataset concordance possible: consequently any departure from concordance is to be taken very seriously.

One might ask whether a more narrow approach might be better suited: a calibration mismatch in one of the CMB experiments would affect mainly just one parameter (the amplitude of primordial fluctuations) and thus a specialised test in which just that particular parameter would vary would be better suited. However, what we propose here is a very general test: in our ignorance we parametrise any possible discrepancies between data in terms of parameters of the standard model. If the observer has concrete physical models of what could have gone wrong, these models can be compared using the evidence values in the standard way.

We have illustrated our method 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. Unfortunately, the evidence calculation is computationally a rather expensive operation: further work on approximating the evidence (and understanding these approximations) in the cosmological setting is perhaps war-

ranted. We have shown how the evidence ratio used in this work is closely related to the difference in chi-squared statistic that is often used when working within the assumptions of Gaussian measurement errors and the Gaussian approximation to the posterior; the effective number of parameters to which this difference should be compared is a function of both the shape of the full likelihood surface and the prior distributions for the parameters. The latter are somewhat less important when comparing datasets than when doing standard model selection, simply because we may apply highly uninformative priors on the understanding that this increases the conservatism of the test. The decrease in goodness-of-fit incurred when enforcing the constraints due to more than one experiment is less significant when viewed against a backdrop of total ignorance. Moreover, if we restrict our attention to informative (or physically-motivated) priors, than we are failing to allow the data to tell us about possible systematic errors in the experiments.

A value of R less than unity is a sign that we should investigate the mismatch bewteen datasets further. This can be done by exploring more focussed 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. Again, such models are both more complex and more flexible, and the evidence ratios between them and H_0 give the relative probability of each proposed systematic effect being important. This approach is somewhat different to that involving the hyperparameters (Lahav et al. 2000; Hobson et al. 2002): these authors used the evidence to calculate the appropriateness of introducing hyperparameters, which then, through the modified posterior pdf, provide more realistic parameter uncertainties and a tool for visualising the effect of the systematic errors. The test suggested here is in the first instance simpler, and is readily applied using the results of the individual dataset analyses. However, its more powerful extension is the investigation of more physically-motivated parameters, using the evidence to check their suitability for modelling the data in hand. 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, Joanna X Dunkley, Andrew Liddle, Uroš Seljak and Antony Lewis for useful discussions. NR is supported by a PPARC studentship. AS is supported by Ministry of Education and Sport of Republic of Slovenia.

REFERENCES

Abazajian K., Adelman-McCarthy J. K., Agüeros M. A., Allam S. S., Anderson S. F., Annis J., Bahcall N. A., Baldry I. K., et al., 2003, AJ, 126, 2081

Beltran M. e. a., 2004, in preparation

Bishop C. M., 1995, *Neural Networks for Pattern Recognition*. Oxford University Press

Croft R. A. C., Weinberg D. H., Bolte M., Burles S., Hernquist L., Katz N., Kirkman D., Tytler D., 2002, ApJ, 581, 20

Dickinson C., Battye R. A., Carreira P., Cleary K., Davies R. D., Davis R. J., Genova-Santos R., Grainge K., et al., 2004, MNRAS, 353, 732

Hinshaw G., Spergel D. N., Verde L., Hill R. S., Meyer S. S., Barnes C., Bennett C. L., Halpern M., et al., 2003, ApJS, 148, 135

Hobson M. P., Bridle S. L., Lahav O., 2002, MNRAS, 335, 377

Kuo C.-L., Ade P., Bock J. J., Daub M. D., Goldstein J., Holzapfel W. L., Lange A. E., Newcomb M., Peterson J. B., Ruhl J., Runyan M. C., Torbet E., 2002, Bulletin of the American Astronomical Society, 34, 1324

Kuo C. L., Ade P. A. R., Bock J. J., Cantalupo C., Daub M. D., Goldstein J., Holzapfel W. L., Lange A. E., Lueker M., Newcomb M., Peterson J. B., Ruhl J., Runyan M. C., Torbet E., 2004, ApJ, 600, 32

Lahav O., Bridle S. L., Hobson M. P., Lasenby A. N., Sodré L., 2000, MNRAS, 315, L45

Lewis A., Bridle S., 2002, Phys.Rev.D, 66, 103511

Ó'Ruanaidh J. J. K., Fitzgerald W. J., 1996, *Numerical Bayesian Methods Applied to Signal Processing*. Springer-Verlag, New-York

Readhead A. C. S., Mason B. S., Contaldi C. R., Pearson T. J., Bond J. R., Myers S. T., Padin S., Sievers J. L., Cartwright J. K., et al., 2004, ApJ, 609, 498

Rebolo R., Battye R. A., Carreira P., Cleary K., Davies R. D., Davis R. J., Dickinson C., Genova-Santos R., Grainge K., et al., 2004, MNRAS, 353, 747

Riess A. G., Strolger L., Tonry J., Casertano S., Ferguson H. C., Mobasher B., Challis P., Filippenko A. V., et al., 2004, ApJ, 607, 665

Sivia D. S., 1996, *Data Analysis: A Bayesian Tutorial*. Oxford University Press

Slosar A., Seljak U., Makarov A., 2004, Phys.Rev.D, 69, 123003

Spergel D. N., Verde L., Peiris H. V., Komatsu E., Nolta M. R., Bennett C. L., Halpern M., Hinshaw G., Jarosik N., Kogut A., Limon M., Meyer S. S., Page L., Tucker G. S., Weiland J. L., Wollack E., Wright E. L., 2003, ApJS, 148, 175

Stoughton C., Lupton R. H., Bernardi M., Blanton M. R., Burles S., Castander F. J., Connolly A. J., Eisenstein D. J., et al., 2002, AJ, 123, 485

Tegmark M., Strauss M. A., Blanton M. R., Abazajian K., Dodelson S., Sandvik H., Wang X., Weinberg D. H., et al., 2004, Phys.Rev.D, 69, 103501

Verde L., Heavens A. F., Percival W. J., Matarrese S., Baugh C. M., Bland-Hawthorn J., Bridges T., Cannon R., et al., 2002, MNRAS, 335, 432

York D. G., Adelman J., Anderson J. E., Anderson S. F., Annis J., Bahcall N. A., Bakken J. A., Barkhouser R., et al., 2000, AJ, 120, 1579

This paper has been typeset from a TEX/ LATEX file prepared by the author.