Monthly Notices of the ROYAL ASTRONOMICAL SOCIETY

Mon. Not. R. Astron. Soc. **414,** 2337–2344 (2011) doi:10.1111/j.1365-2966.2011.18552.x

Designing decisive detections

Roberto Trotta,[1,2★] Martin Kunz[3,4] and Andrew R. Liddle[4]

[1] *Astrophysics Group, Imperial College London, Blackett Laboratory, Prince Consort Road, London SW7 2AZ*
[2] *Astrophysics Department, Oxford University, Denys Wilkinson Building, Keble Road, Oxford OX1 3RH*
[3] *Départment de Physique Théorique, University of Geneva, Quai E. Ansermet 24, 1205 Geneva, Switzerland*
[4] *Department of Physics and Astronomy, University of Sussex, Brighton BN1 9QH*

Accepted 2011 February 15. Received 2011 February 15; in original form 2010 December 19

ABSTRACT

We present a general Bayesian formalism for the definition of figures of merit (FoMs) quantifying the scientific return of a future experiment. We introduce two new FoMs for future experiments based on their model selection capabilities, called the *decisiveness* of the experiment and the *expected strength of evidence*. We illustrate these by considering dark energy probes and compare the relative merits of stages II, III and IV dark energy probes. We find that probes based on supernovae and on weak lensing perform rather better on model selection tasks than is indicated by their Fisher matrix FoM as defined by the Dark Energy Task Force. We argue that our ability to optimize future experiments for dark energy model selection goals is limited by our current uncertainty over the models and their parameters, which is ignored in the usual Fisher matrix forecasts. Our approach gives a more realistic assessment of the capabilities of future probes and can be applied in a variety of situations.

Key words: methods: statistical – cosmological parameters.

1 INTRODUCTION

As cosmology becomes increasingly dominated by results emerging from large-scale observational programmes, it is imperative to be able to justify that resources are being deployed as effectively as possible. In recent years, it has become standard to quantify the expected outcome of cosmological surveys to enable comparison, a procedure exemplified by the figure of merit (FoM) introduced by Huterer & Turner (2001) and later used by the Dark Energy Task Force (DETF) for dark energy surveys (Albrecht et al. 2006, 2009). Still in its infancy, however, is the topic of survey design, where an experiment is optimized, within design or cost constraints, to generate the best scientific outcome (Bassett 2005; Bassett, Parkinson & Nichol 2005; Parkinson et al. 2007, 2010).

Both in quantifying and in optimizing survey capability, it is important to identify the scientific questions one hopes to answer. The DETF FoM measures the expected parameter constraints on a two-parameter dark energy model, using a Fisher matrix approach; this is an example of a parameter estimation FoM, in which the correct cosmological model is assumed to be known and the task is to estimate its parameter values (see also e.g. Mortonson, Huterer & Hu 2010). However, many of the most pressing questions in cosmology concern not parameters but models, i.e. the identification of the correct set of parameters to describe our Universe. Examples are whether cosmic acceleration is due to a cosmological constant, quintessence or modified gravity, and whether or not the Universe has zero spatial curvature. These are model selection questions; hence, forecasts of the capabilities of future probes should be assessed by their power to answer such questions, rather than the more limited question of the error they will be able to achieve assuming that a given model is true (i.e. the usual Fisher matrix forecast). Alternative FoMs, which quantify the ability of experiments to answer model selection problems, have been previously discussed by Mukherjee et al. (2006), Trotta (2007b) and Trotta et al. (2010).[1]

In this paper we present a comprehensive formalism for the construction of survey FoMs, incorporating both model and parameter uncertainty in light of the present observational situation. In order to do so, we build on the methodology introduced in Trotta et al. (2010). We construct two new model selection FoMs, the *decisiveness* and the *expected strength of evidence*, which quantify the expected capability of an experiment to perform model comparison tests. For illustration we focus on the case of dark energy observations, though our formalism is broadly applicable.

2 BAYESIAN FRAMEWORK FOR PERFORMANCE FORECASTING

2.1 Expected utility of an experiment

In order to build up towards the definition of our FoMs, we need to consider the different levels of uncertainty that are relevant when

★E-mail: r.trotta@imperial.ac.uk

[1] For an alternative, essentially frequentist, perspective on this issue, see Amara & Kitching (2010).

predicting the probability of a certain model selection outcome from a future probe. These can be summarized as follows.

(i) *Level 1*: current uncertainty about the correct model (e.g. is it a cosmological constant or a dark energy model?).

(ii) *Level 2*: present-day uncertainty in the value of the cosmological parameters for a given model (e.g. present error on the dark energy equation-of-state parameters assuming an evolving dark energy model).

(iii) *Level 3*: realization noise, which will be present in future data even when assuming a model and a fiducial choice for its parameters.

The commonly used Fisher matrix forecast (see e.g. Tegmark, Taylor & Heavens 1996) ignores the uncertainty arising from levels 1 and 2, as it assumes a fiducial model (level 1) and fiducial parameter values (level 2). It averages over realization noise (level 3) in the limit of an infinite number of realizations. Furthermore, in the Fisher matrix formalism the likelihood is approximated by construction as a Gaussian, which might be inaccurate for parameter spaces exhibiting curving degeneracies and/or multimodal distributions. Clearly, the Fisher matrix procedure provides a very limited assessment of what we can expect for the scientific return of a future probe, as it ignores the uncertainty associated with the choice of model and parameter values.

The Bayesian framework allows improvement on the usual Fisher matrix error forecast, thanks to a general procedure which fully accounts for all three levels of uncertainty given above. This will allow us to define a new type of FoM which represents in a more realistic way the uncertainties involved in making predictions.

Following Loredo (2003), we think of future data D_f as *outcomes*, which arise as consequence of our choice of experimental parameters e (*actions*). For each action and each outcome, we define a utility function $\mathcal{U}(D_\mathrm{f}, e)$. Formally, the utility only depends on the future data realization D_f. However, as will become clear below, the data D_f are realized from a fiducial model and model parameter values. Therefore, the utility function implicitly depends on the assumed model and parameters from which the data D_f are generated. The best action is the one that maximizes the expected utility, i.e. the utility averaged over possible outcomes:

$$\mathcal{EU}(e) \equiv \int \mathrm{d}D_\mathrm{f}\, p(D_\mathrm{f}|e, d)\mathcal{U}(D_\mathrm{f}, e). \quad (1)$$

Here, $p(D_\mathrm{f}|e, d)$ is the predictive distribution for the future data, conditional on the experimental set-up (e) and on current data (d). For a single fixed model, the predictive distribution is given by

$$\begin{aligned} p(D_\mathrm{f}|e, d) &= \int \mathrm{d}\theta\; p(D_\mathrm{f}, \theta|e, d) \\ &= \int \mathrm{d}\theta\; p(D_\mathrm{f}|\theta, e, d)p(\theta|e, d) \\ &= \int \mathrm{d}\theta\; p(D_\mathrm{f}|\theta, e)p(\theta|d), \end{aligned} \quad (2)$$

where the last line follows because $p(D_\mathrm{f}|\theta, e, d) = p(D_\mathrm{f}|\theta, e)$ (conditioning on current data is irrelevant once the parameters are given) and $p(\theta|e, d) = p(\theta|d)$ (conditioning on future experimental parameters is irrelevant for the present-day posterior). So we can predict the probability distribution for future data D_f by averaging the likelihood function for the future measurement (level 3 uncertainty) over the current posterior on the parameters (level 2 uncertainty). The expected utility then becomes

$$\mathcal{EU}(e) = \int \mathrm{d}\theta\, p(\theta|d) \int \mathrm{d}D_\mathrm{f}\, p(D_\mathrm{f}|\theta, e)\mathcal{U}(D_\mathrm{f}, e). \quad (3)$$

So far, we have tacitly assumed that only one model was being considered for the data. In practice, there will be several models that one is interested in testing (level 1 uncertainty), and typically there is uncertainty over which one is best. This is in fact one of the main motivations for designing a new dark energy probe. If N models $\{\mathcal{M}_1, \ldots, \mathcal{M}_N\}$ are being considered, each one with parameter vector θ_i ($i = 1, \ldots, N$), the current posterior can be further extended in terms of model averaging (level 1), weighting each model by its current model posterior probability, $p(\mathcal{M}_i|d)$, given by

$$p(\mathcal{M}_i|d) = \frac{p(d|\mathcal{M}_i)p(\mathcal{M}_i)}{p(d)}, \quad (4)$$

where $p(d|\mathcal{M}_i)$ is the Bayesian evidence for model $\mathcal{M}_i$, $p(\mathcal{M}_i)$ is the model's prior and $p(d)$ a normalizing constant. Using equation (3), this gives the model-averaged expected utility

$$\begin{aligned} \mathcal{EU}(e) = \sum_{i=1}^{N} p(\mathcal{M}_i|d) \int \mathrm{d}\theta_i\, p(\theta_i|d, \mathcal{M}_i) \\ \times \int \mathrm{d}D_\mathrm{f}\, p(D_\mathrm{f}|\theta_i, e, \mathcal{M}_i)\mathcal{U}(D_\mathrm{f}, e, \mathcal{M}_i). \end{aligned} \quad (5)$$

This expected utility is the most general definition of an FoM for a future experiment characterized by experimental parameters e. As we show below, the usual Fisher matrix forecast is recovered as a special case of equation (5), as are other FoMs that have been defined in the literature (e.g. Bassett 2005; Wang 2008; Amara & Kitching 2010). Therefore, equation (5) gives us a formalism to define in all generality the scientific return of a future experiment. This result clearly accounts for all three levels of uncertainty in making our predictions: the utility function $\mathcal{U}(D_\mathrm{f}, e, \mathcal{M}_i)$ (to be specified below) depends on the future data realization, D_f (level 3), which in turn is a function of the fiducial parameter value, θ_i (level 2), and is averaged over present-day model probabilities (level 1).

2.2 Figures of merit from expected utility

The expected utility of equation (5) provides the most general formalism for the evaluation of the scientific return of an experiment. It reduces to previously used FoMs for specific choices of priors and utility functions. For example, the DETF advocated using the inverse of the area of the future probe covariance matrix on the dark energy parameters as an FoM quantifying the strength of the statistical constraints from the experiment. This FoM can be recovered by setting $N = 1$ in equation (5) (only one fiducial model is considered), taking a Dirac delta function for the current posterior, $p(\theta|d, \mathcal{M}) = \delta(\theta - \theta_\star)$ (only the fiducial parameter vector $\theta_\star$ is considered), assuming no realization noise or, equivalently, averaging over many future data realizations, so that $p(D_\mathrm{f}|\theta, e, \mathcal{M}) = \delta[D_\mathrm{f} - D(\theta_\star)]$, where $D(\theta_\star)$ describes a no-noise data realization around the fiducial parameter values, and defining the utility function as the determinant of the future Fisher matrix, evaluated at the fiducial parameter values, $\theta_\star$.

Another example is the Gaussian linear model considered by Trotta et al. (2010), where the utility function was chosen to be the inverse of the marginal error on the parameters of interest. It is a property of the Gaussian linear model that the error ellipse does not depend on the fiducial model nor data realization, but only on the design matrix (Kunz, Trotta & Parkinson 2006). Therefore, in this case the integration over future data D_f gives unity in equation (5), and the same expression is recovered as in Trotta et al. (2010).

Mukherjee et al. (2006) defined two model selection FoMs, each of which considers two models, a cosmological constant model and a two-parameter dark energy model. One FoM quantifies the strength with which the dark energy model will be excluded *if* the cosmological constant is correct; the current posterior is therefore taken to be that model and the FoM is the Bayes factor (defined below) in favour of the cosmological constant. The other FoM is the opposite, quantifying whether the cosmological constant can be ruled out if the dark energy model is correct. The current posterior is now the dark energy model space, and the FoM measures in how much of that space the cosmological constant model could be excluded (e.g. the inverse parameter area above a certain Bayes factor threshold, by analogy to the DETF FoM above).

Trotta (2007b) introduced a methodology to compute the predicted posterior odds distribution (PPOD) for a model comparison from a future experiment. A PPOD-based FoM is another special case of our general formalism: it is obtained from equation (5) by assuming no realization noise, $p(D_{\mathrm{f}}|\theta, e, \mathcal{M}) = \delta[D_{\mathrm{f}} - D(\theta_\star)]$, and adopting as utility function the tail probability of the Bayes factor obtainable by a future probe.

For a given experimental configuration e, the expected utility can be evaluated as follows.

(i) Draw a uniformly weighted sample for the fiducial value for the parameters, $\theta_\star$, from a Monte Carlo Markov chain distributed according to the present, model-averaged, posterior $p(\theta|d) = \sum_i p(\mathcal{M}_i|d)p(\theta_i|\mathcal{M}_i, d)$ (levels 1 and 2).

(ii) Generate pseudo-data D_{f} for the future probe, assuming $\theta_\star$ as fiducial parameter values.

(iii) Evaluate the utility function from the future data (to be defined below).

(iv) Loop back to (i) and average the utility function over the so-obtained samples.

In general, the above procedure is computationally very expensive, as it involves two nested averages, one over the fiducial parameters (step i) and one over future pseudo-data realizations (step ii). Furthermore, in the context of model-selection-oriented FoMs to be introduced below, the evaluation of the utility (step iii) requires the computation of Bayes factors from the pseudo-data, which again are costly. If one wanted to use Markov chain Monte Carlo techniques, one would typically need $\sim 10^4$ samples in step (i) and another $\sim 10^5$ samples to obtain a reliable estimate of the utility function in steps (ii) and (iii). Therefore, the typical number of likelihood evaluations required would be of the order of $\sim 10^9$, which is at the limit of what can be achieved today unless one adopts highly accelerated inference methods (Fendt & Wandelt 2006; Auld, Bridges & Hobson 2008; Frommert et al. 2010; Bridges et al. 2011). Therefore, we shall make some simplifying assumptions that reduce this computational burden very considerably.

First, we will consider only $N = 2$ competing models. Secondly, we will work in the Gaussian likelihood approximation, i.e. we will assume that both the present-day and the future likelihood are well approximated by Gaussian distributions. This is the same kind of approximation involved in the usual Fisher matrix forecast. The assumption of Gaussianity further allows us to side-step the pseudo-data generation step: for a given value of the fiducial parameters, $\theta_\star$, the maximum likelihood estimate $\hat{\theta}_{\mathrm{f}}$ from future data D_{f} generated from $\theta_\star$ is distributed as a Gaussian with mean $\theta_\star$ and covariance matrix given by the inverse of the likelihood Fisher matrix for the future probe. As a consequence, we do not need to generate pseudo-data at all in step (ii), and we can instead work directly in parameter space, by drawing $\hat{\theta}_{\mathrm{f}}$ directly from a Gaussian distribution centred on $\theta_\star$.

Having made the above simplifications, we now turn to using the expected utility to define two new FoMs based on model selection.

3 FIGURES OF MERIT FOR MODEL SELECTION

To assess the science return of proposed missions in terms of their model selection capabilities, we propose to adopt the expected utility of equation (5) as an FoM for experiment e, after defining an appropriate utility function $\mathcal{U}(D_{\mathrm{f}}, e, \mathcal{M}_i)$. There are many ways to do this, and we introduce here two proposals. The first one is named *decisiveness*, and it gives the probability that the proposed experiment will achieve a decisive outcome for model selection. A good experiment should be as decisive as possible. A complementary approach, named *expected strength of evidence*, is to compute by how much the experiment is expected to prefer one or the other model on average. Again, a good experiment will be able to prefer one of the models strongly.

In a two-way Bayesian model comparison, the key Bayesian statistic is the Bayes factor B_{01}, which is formed from the ratio of the Bayesian evidences of the two models being considered:

$$B_{01} = \frac{p(d|\mathcal{M}_0)}{p(d|\mathcal{M}_1)}, \tag{6}$$

where the Bayesian evidence is the average of the likelihood under the prior in each model:

$$p(d|\mathcal{M}_i) = \int \mathrm{d}\theta_i \, p(d|\theta_i, \mathcal{M}_i)p(\theta_i|\mathcal{M}_i). \tag{7}$$

The Bayes factor updates the prior probability ratio of the models to the posterior one, indicating the extent to which the data have modified one's original view on the relative probabilities of the two models. The Bayes factor can be evaluated by a general numerical method such as nested sampling (Bassett, Corasaniti & Kunz 2004; Skilling 2004; Parkinson, Mukherjee & Liddle 2006; Feroz, Hobson & Bridges 2009) or, if one model is nested within the other, by the Savage–Dickey density ratio (SDDR; Trotta 2007a, 2008). The Bayes factor is usually interpreted on Jeffreys' scale shown in Table 1 (Jeffreys 1961; Gordon & Trotta 2007).

3.1 'Decisiveness' figure of merit

A 'decisive' experiment is one that is able to gather strong evidence in favour of one of the competing models. Therefore, its utility function is 0 (1) if the Bayes factor it will obtain is below (above) the 'strong' threshold for the evidence, $\ln B = 5$ (see Table 1) (this

Table 1. Empirical scale for evaluating the strength of evidence when comparing two models, $\mathcal{M}_0$ versus $\mathcal{M}_1$ (Jeffreys' scale). The rightmost column gives our convention for denoting the different levels of evidence above these thresholds.

$\lvert\ln \mathcal{B}_{01}\rvert$	Odds	Strength of evidence
<1.0	$\lesssim 3:1$	Inconclusive
1.0	$\sim 3:1$	Weak evidence
2.5	$\sim 12:1$	Moderate evidence
5.0	$\sim 150:1$	Strong evidence

level of evidence is sometimes called 'decisive', hence the name of the FoM). Therefore, we are led to the following utility function:

$$\mathcal{U}(D_{\rm f}, e, \mathcal{M}_i) = \begin{cases} 1 & \text{if } |\ln \mathcal{B}_{01}| > 5 \\ 0 & \text{otherwise,} \end{cases} \quad (8)$$

where $\mathcal{B}_{01}$ is the Bayes factor between the two models, obtained by the future experiment *e*. The best experiment is the one that maximizes this quantity, i.e. the one whose probability of obtaining a strong model selection outcome for either of the models is maximized. We thus define the *decisiveness* $\mathscr{D}$ of an experiment *e* as its expected utility, equation (5), with the utility function (8). We note that $\mathscr{D}$ as an FoM is especially resilient to the scatter in the Bayes factor coming from averaging over data realizations and the unknown fiducial parameter values (Jenkins & Peacock 2011). In fact, our formalism takes this scatter into full account, and if too many realizations are scattered out of the 'decisive' region (e.g. due to large noise on the measurements from the future probe), then this will lead to a lower FoM. Therefore, using $\mathscr{D}$ to optimize the design of an experiment is particularly useful to guard against this effect.

3.2 'Expected strength of evidence' figure of merit

Instead of the discrete utility function above given above, we can adopt one that is more gradual in assessing the merit of the future probe. Such a utility function is

$$\mathcal{U}(D_{\rm f}, e, \mathcal{M}_i) = (-1)^i \ln \mathcal{B}_{01}, \quad (9)$$

which describes the strength of the model selection result from the future probe. By plugging this utility function into equation (5), we obtain an FoM that we call the 'expected strength of evidence' and denote by $\mathscr{E}$. The rationale is that for every given fiducial value of the parameters and for every data realization, the best experiment is the one that maximizes the support to the true model (i.e. the model out of which the data actually come from), even though it might be that the experiment in question is not strong enough to achieve decisiveness.

The factor $(-1)^i$ in equation (9) is to ensure that the utility only rewards support for the *correct* model; for example, under the more complex model ($\mathcal{M}_1$), we want to maximize $-\ln \mathcal{B}_{01}$, the odds in favour of $\mathcal{M}_1$. Bayes factors can occasionally favour the wrong model; for example, if the true model were a dark energy model with $w = -0.999$, anything other than an extraordinarily precise experiment is likely to favour the more predictive cosmological constant model. Nevertheless, support for the wrong model will happen only in a small parameter space region and will be overwhelmed when the average over the current posterior is carried out, making the above nearly equivalent to the simpler choice $\mathcal{U}(D_{\rm f}, e, \mathcal{M}_i) = |\ln \mathcal{B}_{01}|$. We have found in the dark energy application presented below that for all future dark energy probes the difference in the FoM between these two choices is less than about 5 per cent, so in practice it is almost negligible.

It might seem at first glance that an experiment that maximizes the expected strength of evidence is also the one that minimizes the error ellipse in the parameter space of interest. If this was true, then the ranking of probes obtained with the expected strength of evidence would be the same as the one from the DETF FoM. However, consider the SDDR expression for nested models (Trotta 2007a):

$$\mathcal{B}_{01} = \frac{p(\phi|D_{\rm f}, e, d, \mathcal{M}_1)}{p(\phi|\mathcal{M}_1)}|_{\phi=\phi_0}, \quad (10)$$

where ϕ are the extra parameters of interest for the more complicated model, which reduces to the simpler model for $\phi = \phi_0$. The odds against $\mathcal{M}_0$ are maximized when the marginal posterior on the extra parameters is as small as possible at the location in parameter space predicted by the simpler model. This means that maximizing $-\ln \mathcal{B}_{01}$ requires minimizing the posterior error *along the direction connecting the fiducial value of* (w_0, w_a) *to* $(-1, 0)$ [if we restrict our consideration to the dark energy example, where $\phi = (w_0, w_a)$]. In other words, the expected strength of evidence FoM favours experiments that deliver error ellipses whose most tightly constrained principal direction points towards the location of the simpler model in parameter space, hence minimizing model confusion. If instead the data come from $\mathcal{M}_0$, then the utility function requires that the height of the posterior at the location of the true model be as large as possible. Since the posterior is normalized, this requires the posterior to be as tightly constrained around the true value as possible, which is obviously desirable.

To summarize, the decisiveness FoM $0 \le \mathscr{D} \le 1$ can be understood as an absolute scale measuring the model selection capabilities of an experiment, with $\mathscr{D} = 1$ denoting the maximum possible performance in terms of model comparison utility (i.e. an experiment that is guaranteed to achieve a decisive model selection result). On the other hand, many probes might still be interesting to build but may fall short of the achieving strong evidence anywhere in parameter space; hence, such experiments would all have $\mathscr{D} = 0$. Yet it is still a relevant question to try and rank them according to their merits. This can be done by looking at the expected strength of evidence, which always returns a non-zero value. Therefore, the expected strength of evidence $\mathscr{E}$ can be regarded as a relative scale of the capabilities of the probes.

4 APPLICATION TO FUTURE DARK ENERGY PROBES

We now apply our newly defined model selection FoMs to a set of representative proposals for future dark energy probes. We consider a Λ cold dark matter (ΛCDM) model with dark energy in the form of a cosmological constant versus an evolving dark energy model where the equation of state is $w(z) = w_0 + w_a z/(1 + z)$, described by the two parameters (w_0, w_a). This is a case of nested models, i.e. where the simpler model (the cosmological constant) is obtained as a special case of the evolving dark energy model by setting $w_0 = -1$, $w_a = 0$. The other cosmological parameters (common to both models) are the baryonic density, the dark matter density, the spatial curvature, the amplitude of scalar adiabatic fluctuations and the spectral index of perturbations. We include curvature in our analysis as this impacts strongly on the constraints on evolving dark energy models (Clarkson, Cortes & Bassett 2007; Wang & Mukherjee 2007).

The current posterior is obtained using the following data sets: *Wilkinson Microwave Anisotropy Probe* 5 (Dunkley et al. 2009), Acbar07 (Kuo et al. 2007), Cosmic Background Imager (CBI) (Sievers et al. 2007), *BOOMERanG* 03 (Jones et al. 2006) for the cosmic microwave background, Sloan Digital Sky Survey (SDSS) Luminous Red Galaxies (LRG) Data Release 4 (DR4) (Tegmark et al. 2006) for $P(k)$, the Hubble Key Project determination of H_0 (Freedman et al. 2001), big bang nucleosynthesis limits on $\Omega_{\rm b} h^2$ (Kirkman et al. 2003) and the Union supernova (SN) Ia compilation (Kowalski et al. 2008). The priors on the common parameters are irrelevant as they cancel from the Bayes factor between the two models [as long as these priors are sufficiently wide to include the maximum likelihood and uncorrelated with the dark energy priors; see Trotta (2007a)], so the only important prior is the one on (w_0, w_a). We choose a Gaussian prior centred on $w_0 = -1$, $w_a = 0$ with Fisher matrix $\Pi = \text{diag}(1, 1/2)$. With this prior and the above data sets, we obtain a Bayes factor $B_{01} = 13.7$ in favour of the ΛCDM model (representing moderate evidence against an evolving dark energy). This means that 93 per cent of samples from the current posterior will be drawn from a ΛCDM model and 7 per cent from a model with evolving dark energy.

4.1 Future dark energy probes

We use a selection of future missions based on the DETF classification (Albrecht et al. 2006), using Fisher matrices provided by the DETFAST package (Dick & Knox 2006). This package provides only the Fisher matrices evaluated at a fixed fiducial ΛCDM cosmology, so we have to assume that the Fisher matrices do not vary significantly for different fiducial parameters drawn from the current posterior. In other words, we take the Fisher matrix for the future experiment at a fiducial ΛCDM point and translate it in parameter space, without recomputing it for each new sample of $\theta_\star$. This is clearly an oversimplification, but since the dark energy parameters are the most important ones for this application, and since 93 per cent of points drawn from the current posterior belong to the ΛCDM case, we expect that the results are not too strongly biased. We intend to study the impact of this assumption and to provide a more comprehensive study of the power of future dark energy probes in future work, while using the simplified approach as an illustration of our new FoMs here.

The DETF has classified the dark energy probes in stages, with stage II being those that are currently ongoing or completed, stage III being medium-term projects and stage IV future large projects (optical Large Survey Telescopes, 'LST'; space-based missions, 'S'; and the Square Kilometre Array, 'SKA'). The probes that we consider here include weak lensing (WL), Type Ia SN, baryon acoustic oscillations (BAOs), cluster counts (CL) and combinations of several probes (ALL). A suffix '-o' and '-p' denotes optimistic and pessimistic assumptions about systematic errors, respectively. The 'p' in the names of the stage III experiments signals the use of photometric redshifts while an 's' is used for spectroscopic surveys (that tend to cover a much smaller area). For further, detailed information consult the DETF report.

The utility function computation proceeds as follows. In order to evaluate the decisiveness, equation (8), and the expected strength of evidence, equation (9), we need the Bayes factor $\ln \mathcal{B}_{01}$ for the future experiment. This is obtained analytically via the SDDR formula (10):

$$\ln \mathcal{B}_{01} = \frac{1}{2} \ln \frac{|\Pi|}{|F_\phi|} - \frac{1}{2} (\bar{\phi} - \phi_0)^t F_\phi (\bar{\phi} - \phi_0), \tag{11}$$

where $\phi = (w_0, w_a)$ are the dark energy parameters of interest, Π is their prior Fisher matrix and F_ϕ is the marginal posterior Fisher matrix for ϕ. We have defined $\phi_0 = (-1, 0)$, and $\bar{\phi}$ is the posterior mean from both current and future data. This can be obtained as the ϕ components of the posterior mean vector in the full parameter space:

$$\bar{\theta} = F^{-1} \left(L^{\mathrm{f}} \hat{\theta}_{\mathrm{f}} + L \hat{\theta} + \Pi \theta_0 \right). \tag{12}$$

In the above, L^{f} is the future probe likelihood Fisher matrix, L is the current constraints' Fisher matrix, θ_0 is the prior mean, $\hat{\theta}_{\mathrm{f}}$ is the future maximum likelihood location while $\hat{\theta}$ is the present constraints' maximum likelihood point. The Fisher matrix from the future and present data, F, is given by

$$F = L^{\mathrm{f}} + L + \Pi. \tag{13}$$

The prior used in equation (11) is the same as the one adopted for the analysis of the present-day data. This is because the prior in the context of Bayesian model selection should be understood as representing the a priori plausible parameter values under the model. Therefore, we do not update the prior to the posterior from the present-day inference step when evaluating the future Bayes factor. The likelihood is obtained from the Fisher matrix formalism, with the above-mentioned additional assumption that the future likelihood Fisher matrix is independent of the fiducial parameter value adopted.

Some of the dark energy probes can achieve a very strong model selection in favour of an evolving dark energy model in parts of the parameter space, often obtaining $\ln \mathcal{B}_{01} \ll -100$. This would correspond to a detection of a non-constant equation of state at many-sigma confidence level. However, we do not expect our Gaussian approximation to the likelihood to hold true so far into the tails of

Table 2. Results for FoMs of various dark energy probes. $\mathscr{D}$ is the decisiveness given in equation (8) and $\mathscr{E}$ is the expected strength of evidence given in equation (9).

Experiment	DETF FoM	$\mathscr{D}$	$\mathscr{E}$
CL-II	0.13	0	2.3
SN-II	1.4×10^{-2}	2.0×10^{-3}	2.7
WL-II	0.7	4.3×10^{-3}	2.8
BAO-IIIp-p	7.1×10^{-5}	2.0×10^{-4}	2.5
BAO-IIIs-p	0.87	1.2×10^{-3}	2.7
BAO-IIIs-o	1.0	1.5×10^{-3}	2.7
CL-IIIp-p	0.56	2.9×10^{-3}	2.7
CL-IIIp-o	8.5	1.5×10^{-2}	3.4
SN-IIIp-p	4.9×10^{-3}	1.5×10^{-3}	2.7
SN-IIIp-o	2.0×10^{-3}	8.5×10^{-3}	3.1
SN-IIIs	4.2×10^{-2}	5.9×10^{-3}	2.9
WL-IIIp-p	6.4	9.9×10^{-3}	3.2
WL-IIIp-o	17	1.6×10^{-2}	3.5
ALL-IIIp-p	59	2.7×10^{-2}	4.1
ALL-IIIp-o	150	0.53	5.1
ALL-IIIs-p	130	0.38	4.8
ALL-IIIs-o	200	0.58	5.3
BAO-IVLST-p	1.8×10^{-2}	1.2×10^{-3}	2.6
BAO-IVLST-o	4.0×10^{-2}	1.6×10^{-3}	2.7
BAO-IVSKA-p	1.3	4.3×10^{-3}	3.0
BAO-IVSKA-o	3.4	9.0×10^{-3}	3.3
BAO-IVS-p	1.4	3.7×10^{-3}	3.0
BAO-IVS-o	3.4	7.0×10^{-3}	3.2
CL-IVS-p	0.50	3.1×10^{-3}	2.8
CL-IVS-o	9.5	1.6×10^{-2}	3.5
SN-IVLST-o	0.32	1.4×10^{-2}	3.4
SN-IVS-p	0.65	9.9×10^{-3}	3.2
SN-IVS-o	0.76	1.6×10^{-2}	3.5
WL-IVLST-p	15	1.3×10^{-2}	3.5
WL-IVLST-o	170	0.77	6.0
WL-IVSKA-p	4.6	1.9×10^{-2}	3.6
WL-IVSKA-o	280	0.81	6.3
WL-IVS-p	83	0.37	4.7
WL-IVS-o	140	0.67	5.5
ALL-LST-p	180	0.54	5.1
ALL-LST-o	900	0.89	6.9
ALL-SKA-p	160	0.49	5.0
ALL-SKA-o	950	0.90	7.0
ALL-IVS-p	480	0.81	6.2
ALL-IVS-o	900	0.90	6.9

the distribution. Therefore, in order to be conservative, we impose a floor at $\ln \mathcal{B}_{01} = -20$ when computing the expected strength of evidence from equation (9): any value of $\ln \mathcal{B}_{01}$ below this floor is remapped to the floor value.

4.2 Results

The results for the future probes are presented in Table 2 and plotted in Fig. 1, where we compare the DETF FoM with our new model selection FoMs. We notice that the decisiveness FoM separates the sample into two distinct groups, those with $\mathscr{D} \lesssim 0.1$ (single probes up to level III and several pessimistic single probes at level IV, together with BAO-IVS-o) that are unlikely to provide a decisive answer to the question whether dark energy is dynamical or not and the rest with $\mathscr{D} \gtrsim 0.1$. This division is not apparent in $\mathscr{E}$ and the DETF FoM, and it leads to critical values of $\mathscr{E} \approx 4$ and DETF FoM ≈ 70 below which an experiment is unlikely to obtain a strong model selection result.

The ranking of the experiments between $\mathscr{D}$ and $\mathscr{E}$ is almost the same, while the DETF FoM gives a similar but not always identical ranking. Looking at the right-hand panels in Fig. 1, we notice that the WL and SN probes tend to lie above the trend line (are more likely to provide a decisive model selection result than would be expected from the DETF FoM) while spectroscopic BAO probes lie below.

In Fig. 2, we show the distribution of $\ln(B_{01})$ for 10^6 outcomes for the ALL-SKA-o probe (the most powerful probe considered here). The red bars on the right hand side are for data drawn from a ΛCDM model, for which this probe often but not always achieves a decisive outcome. The blue bar on the left shows that the probe

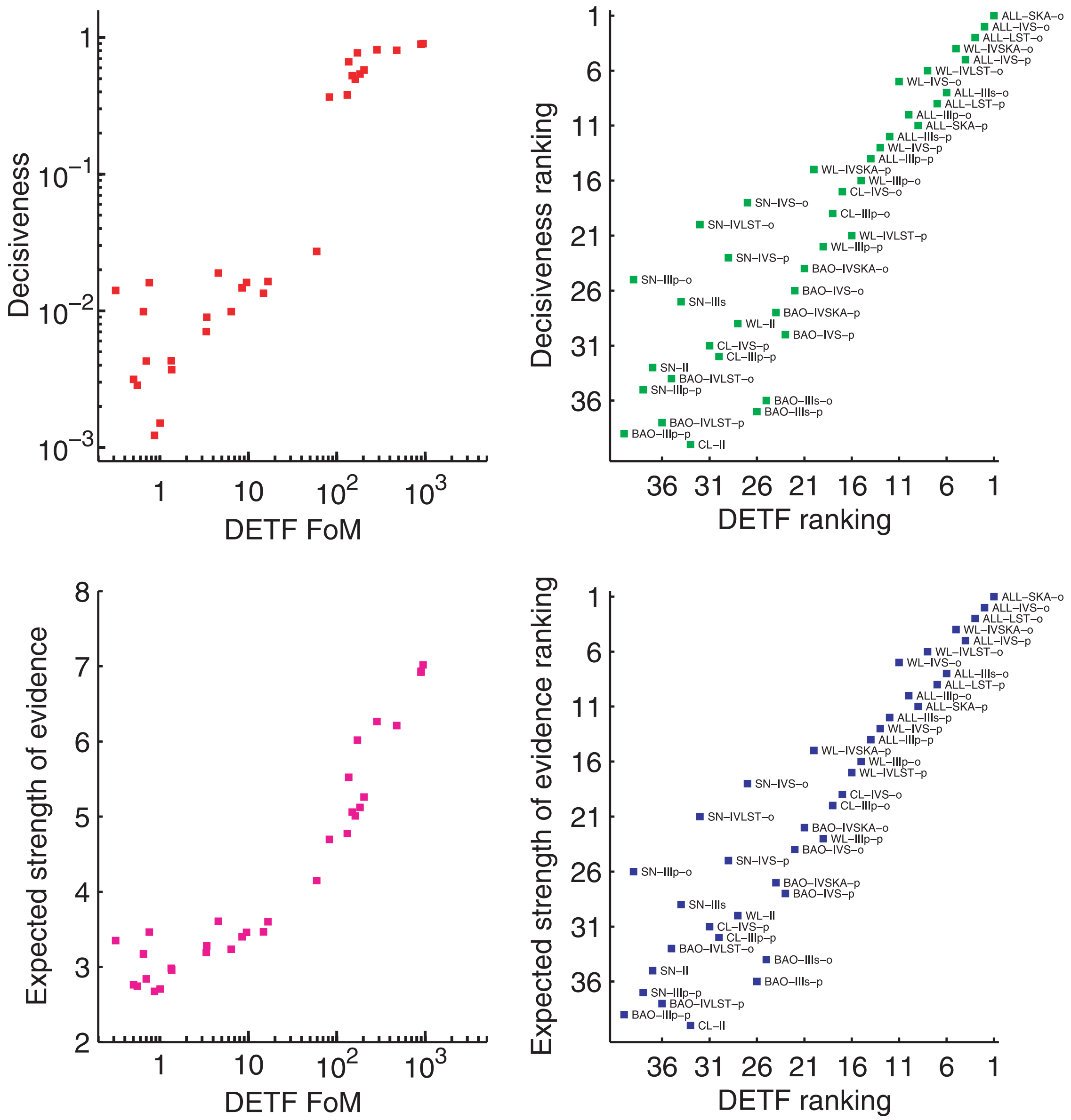

Figure 1. Comparison of our model selection FoMs to the DETF FoM (left-hand panels) and the ranking of dark energy probes derived from them (right-hand panels).

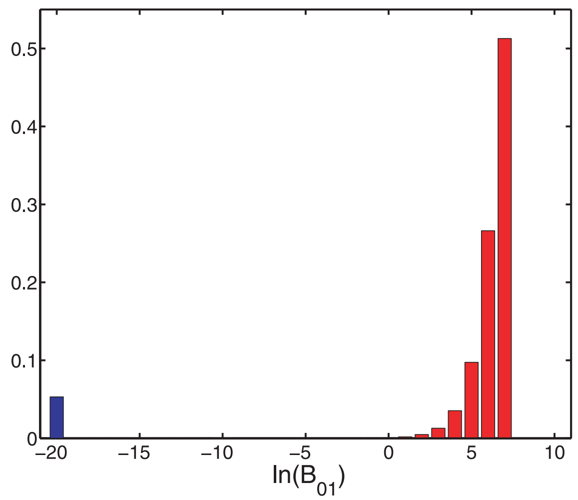

Figure 2. Histogram of $\ln(\mathcal{B}_{01})$ values for the ALL-SKA-o DETF case. Red bars (on the right, values above 0) show cases drawn from a ΛCDM model (93 per cent of cases according to the current posterior) and the blue bar (on the far left) show cases with an evolving dark energy fiducial model [7 per cent of cases, capped at $\ln(\mathcal{B}_{01}) = -20$ as described in the text].

will deliver very powerful results if the dark energy is actually evolving (given the priors adopted and the current knowledge on dark energy parameters). It is not surprising that the model selection outcomes against ΛCDM tend be stronger than those that support it; it is always more difficult to strongly support a nested model, as the simpler model only 'profits' from its predictiveness (thanks to Occam's razor effect), but can never provide a better fit.

5 CONCLUSIONS

We have presented a general Bayesian formalism for the definition of FoMs encapsulating the expected scientific return of future experiments. Our method fully accounts for all sources of uncertainties involved in the prediction, including present-day model and parameter uncertainties, and realization noise. It thus improves on the usual Fisher matrix methods by producing more realistic forecasts for the possible distribution of future experimental outcomes.

We used this framework to define two FoMs for probes that measure the dark energy equation of state in order to test the ΛCDM paradigm: the *decisiveness* $\mathscr{D}$, which quantifies the probability that a probe will deliver a decisive result in favour or against the cosmological constant, and the *expected strength of evidence* $\mathscr{E}$, which returns a measure of the expected power of a probe for model selection. We compared these quantities to the widely used DETF FoM for a range of probes and found that the rankings agree reasonably well, but that WL and SN probes have a higher than expected model selection power relative to their DETF FoM ranking. We also found, for our choice of prior, that there is a critical DETF FoM of around 70 below which probes are very unlikely to obtain a strong model selection result.

An additional advantage of the formalism presented in this paper, and of any FoMs that use it, is the possibility to include further observations, for example those that constrain the growth history or the presence of effective anisotropic stresses. One just extends the likelihood based on the predictions of the underlying models, but the procedure is unchanged, and the interpretation of the results is unchanged as well. There is therefore no need to define new FoMs as data analysis goals for future probes evolve.

The methodology presented here is widely applicable to a variety of forecasting and optimization problems. Our application to the model selection capabilities of future dark energy missions is but a first step towards a fully Bayesian approach to performance forecast.

ACKNOWLEDGMENTS

RT was partially supported by the Royal Astronomical Society through the Sir Norman Lockyer Fellowship and by St Anne's College, Oxford. MK and ARL acknowledge support from the Science and Technology Facilities Council (grant number ST/F002858/1), and MK also acknowledges support from the Swiss NSF.

REFERENCES

Albrecht A. et al., 2006, preprint (arXiv:astro-ph/0609591)
Albrecht A. et al., 2009, preprint (arXiv:0901.0721)
Amara A., Kitching T., 2010, preprint (arXiv:1009.3274)
Auld T., Bridges M., Hobson M. P., 2008, MNRAS, 387, 1575
Bassett B. A., 2005, Phys. Rev. D, 71, 083517
Bassett B. A., Corasaniti P. S., Kunz M., 2004, ApJ, 617, L1
Bassett B. A., Parkinson D., Nichol R. C., 2005, ApJ, 626, L1
Bridges M., Cranmer K., Feroz F., Hobson M., de Austri R. R., Trotta R., 2011, JHEP, 3, 12
Clarkson C., Cortes M., Bassett B. A., 2007, J. Cosmology Astropart. Phys., 0708, 011
Dick J., Knox L., 2006, Code available at http://www.physics.ucdavis.edu/DETFast/
Dunkley J. et al., 2009, ApJS, 180, 306
Fendt W. A., Wandelt B. D., 2006, ApJ, 654, 2
Feroz F., Hobson M. P., Bridges M., 2009, MNRAS, 398, 1601
Freedman W. et al., 2001, ApJ, 553, 47
Frommert M., Pflüger D., Riller T., Reinecke M., Bungartz H. J., Ensslin T. A., 2010, MNRAS, 406, 1177
Gordon C., Trotta R., 2007, MNRAS, 382, 1859
Huterer D., Turner M. S., 2001, Phys. Rev. D, 64, 123527
Jeffreys H., 1961, Theory of Probability, 3rd edn. Oxford Univ. Press, Oxford
Jenkins C. R., Peacock J. A., 2011, MNRAS, in press (doi:10.1111/j.1365-2966.2011.18361.x) (arXiv:1101.4822)
Jones W. C. et al., 2006, ApJ, 647, 823
Kirkman D., Tytler D., Suzuki N., O'Meara J. M., Lubin D., 2003, ApJS, 149, 1
Kowalski M. et al., 2008, ApJ, 686, 749
Kunz M., Trotta R., Parkinson D., 2006, Phys. Rev. D, 74, 023503
Kuo C.-L. et al., 2007, ApJ, 664, 687
Loredo T. J., 2003, in Erickson G. J., Zhai Y., eds, AIP Conf. Proc. 707, Bayesian Inference and Maximum Entropy Methods in Science and Engineering: 23rd International Workshop. Am. Inst. Phys., New York, p. 330
Mortonson M. J., Huterer D., Hu W., 2010, Phys. Rev. D, 82, 063004
Mukherjee P., Parkinson D., Corasaniti P. S., Liddle A. R., Kunz M., 2006, MNRAS, 369, 1725
Parkinson D., Mukherjee P., Liddle A. R., 2006, Phys. Rev. D, 73, 123523
Parkinson D., Blake C., Kunz M., Bassett B. A., Nichol R. C., Glazebrook K., 2007, MNRAS, 377, 185
Parkinson D., Kunz M., Liddle A. R., Bassett B. A., Nichol R. C., Vardanyan M., 2010, MNRAS, 401, 2169
Sievers J. L. et al., 2007, ApJ, 660, 976
Skilling J., 2004, in Fischer R., von Toussaint U., Preuss R., eds, AIP Conf. Proc. 735, Bayesian Inference and Maximum Entropy Methods in Science and Engineering. Am. Inst. Phys., New York, p. 395, available at http://www.inference.phy.cam.ac.uk/bayesys/
Tegmark M., Taylor A., Heavens A., 1997, ApJ, 480, 22
Tegmark M. et al., 2006, Phys.Rev. D, 74, 123507
Trotta R., 2007a, MNRAS, 378, 72

Trotta R., 2007b, MNRAS, 378, 819
Trotta R., 2008, Contemporary Phys., 49, 71
Trotta R., Kunz M., Mukherjee P., Parkinson D., 2010, in Hobson M., Jaffe A., Liddle A., Mukherjee P., Parkinson D. R., eds, Bayesian Methods in Cosmology. Cambridge Univ. Press, Cambridge, p. 99
Wang Y., 2008, Phys. Rev., D77, 123525
Wang Y., Mukherjee P., 2007, Phys. Rev. D, 76, 103533

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