Interpreting the Dark Energy Figure of Merit

Christopher Hirata (Caltech)
Daniel Eisenstein (University of Arizona)

Point of Contact:
Christopher Hirata
Caltech M/C 130-33
Pasadena, CA 91125

(626) 395-4200

chirata@tapir.caltech.edu

February 2009

1. Introduction

One of the most remarkable discoveries of modern cosmology is that the expansion of the Universe is accelerating (Riess et al. 1998; Perlmutter et al. 1999). The oldest explanation for this phenomenon is the cosmological constant Λ. However the theoretical difficulty of explaining its magnitude, as well as evidence for an early inflationary epoch that was evidently not powered by Λ, has spurred many theorists to consider alternatives. These have usually fallen into two categories: "dark energy" models that invoke a smoothly distributed substance with negative pressure to drive cosmic acceleration in accordance with the Friedmann equation, and "modified gravity" models in which cosmic acceleration results from the breakdown of general relativity (GR) and hence of the Friedmann equation.

At the same time, advances in observational techniques have enabled large surveys that would test some of these models by making precision measurements of the expansion history of the Universe and the growth of cosmic structures. Measurements of e.g. the luminosity distance-redshift relation using supernovae (SN) or the angular diameter distance from baryon acoustic oscillations (BAO) in the galaxy distribution could detect the change in energy density of a (non-Λ) dark energy component, or confirm that Λ correctly describes the expansion history of the Universe to some desired degree of precision. Likewise, for any measured expansion history, and under the assumption of smoothly distributed dark energy, GR makes a prediction for the rate of growth of density perturbations. This could be tested using the correlation function of galaxy shapes distorted via weak lensing (WL).

The need to assess, compare, and optimize observational programs has driven the need for a quantitative measure of the dark energy performance of an experiment – that is, a Figure of Merit (FoM). Several FoMs have been proposed, based on the anticipated error bars on parameterized dark energy or modified gravity models. The FoM concept has grown in popularity since the report of the Dark Energy Task Force (DETF; Albrecht et al. 2006), and has most recently played a major role on the Joint Dark Energy Mission (JDEM) FoM Science Working Group (FoMSWG; Albrecht et al. 2009). Observers have used the FoM as an optimization tool, to assess their systematics control needs, and to advertise the power of their experiments. Brainstorming theorists have used it to illustrate the enhancing power of new measurement techniques, and the destructive power of hitherto unrecognized sources of error.

The FoM is a necessary evil. It is necessary in that experiments must be optimized somehow, and some quantitative measure is preferable to the alternatives, e.g. allocating resources to the largest ego in the conference room. It can be used to show how combinations of particular experiments are more powerful than any one individually. The judicious use of FoMs can also assess the robustness of an experiment, by showing how much the FoM degrades if some aspect of the experiment fails. The FoM is evil in that it is not unique, since we do not know *a priori* the correct form of any possible deviation from Λ; it may be difficult to compute in advance, especially if there are astrophysical sources of uncertainty; and its careless application can give misleading relative assessments of robustness. The purpose of this White Paper is to highlight the usage of the FoM and its potential pitfalls. Most of these pitfalls were discussed by the FoMSWG and mentioned in its report (Albrecht et al. 2009), but we emphasize them here too since there is a human tendency to focus on the numbers and recipes, and skip the caveats. Concrete examples are used wherever possible, drawn from the cosmic microwave background (CMB), SN, WL, and BAO techniques; the purpose here is to illustrate generic dangers, rather than to disparage or advocate for any specific dark energy probe.

2. The Figures of Merit

In assigning a FoM to a particular experiment (or combinations of experiments), there are four necessary steps (Albrecht et al. 2009). First, a parameter space is defined, which includes the dark energy parameters as well as the conventional cosmological parameters (e.g. Ω_m, Ω_k, n_s) and any nuisance parameters (e.g. the supernova absolute magnitude). Second, a fiducial point in this parameter space is chosen; usually the dark energy is set to Λ – as appropriate if we want to assess the ability of an experiment to detect small deviations from Λ – and the conventional parameters are set to the most recent determinations.[1] Third, one must forecast the covariance matrix **C** of the dark energy parameters for an experiment under consideration. Finally, one must choose a merit function FoM(**C**) that outputs a single number describing the desirability of that experiment. The last two steps are usually the most controversial: calculating **C** usually involves measurement systematics that are not known at an early stage in the design, or (worse) astrophysical parameters that cannot be measured before the observational program is underway. The merit function involves a choice of which parameters are "more interesting," which is inherently subjective.

The most commonly used FoM is that introduced by the DETF (Albrecht et al. 2006). The DETF parameterized the equation of state $w(z)$ of dark energy using two parameters w_0 and w_a:

$$w(z) = w_0 + \frac{z}{1+z} w_a \tag{1}$$

They then defined their FoM to be proportional to the inverse area of the error ellipse in the w_0-w_a plane:

$$\mathrm{FoM}_{\mathrm{DETF}} = \left[\det \mathbf{C}(w_0, w_a)\right]^{-1/2}, \tag{2}$$

where $\mathbf{C}(w_0,w_a)$ is the covariance matrix of the two dark energy parameters after marginalizing out all of the other cosmological parameters. Larger $\mathrm{FoM}_{\mathrm{DETF}}$ is "better" since it corresponds to a smaller error ellipse.

The $\mathrm{FoM}_{\mathrm{DETF}}$ assigns credit to experiments not just for measuring a constant w, but also for measuring how (and if) it varies with redshift. It does not however assign credit for constraining alternative types of dark energy evolution. Albrecht & Bernstein (2007) proposed an alternative formulation that does assign such credit, allowing $w(z)$ to take on a different value in 9 intervals equally spaced in scale factor between z=4 and z=0. These 9 values $w_0 \ldots w_8$ are their dark energy parameters, and they construct a $\mathrm{FoM}_{\mathrm{AB}}$ in analogy to Eq. (2). They found that in most practical cases $\mathrm{FoM}_{\mathrm{AB}}$ for an experiment is tightly correlated with $\mathrm{FoM}_{\mathrm{DETF}}$; we agree, and mostly use $\mathrm{FoM}_{\mathrm{DETF}}$ in this White Paper, but we caution that significant exceptions are at least mathematically possible. A project that measures $w(z)$ very accurately but only over a narrow redshift range will score well on $\mathrm{FoM}_{\mathrm{DETF}}$ because the parameters w_0 and w_a can be measured,

[1] Some authors, e.g. Mukherjee et al. (2005) and Liddle et al. (2006), have advocated Bayesian rather than Fisher matrix techniques, which do not select a particular fiducial point but explore the entire space of presently allowed dark energy models. These have not come into general use and were not adopted by the FoMSWG, largely because of their computational expense and because of concerns that implementation-dependence could make them even harder to compare than Fisher-based FoMs.

whereas in FoM_{AB} it will get penalized for not exploring the behavior of dark energy at other redshifts.

The DETF generated a variety of Fisher matrices for hypothetical experiments. The FoMSWG went further and made their Fisher matrices publically available, including forecasts for the CMB observations from the Planck satellite, the Dark Energy Survey (DES) WL project, and for the combined SN and BAO surveys anticipated to be completed by ~2016.[2]

The DETF emphasized that one should measure the growth of structure, but did not define a growth of structure FoM. The FoMSWG did introduce such a FoM: the inverse-variance of the parameter γ defined by $d \ln G / d \ln a = \Omega_m^{\gamma}$, where G is the growth function and a is the scale factor (Albrecht et al. 2009).

3. Combining Experiments

Measuring dark energy parameters usually requires not just a measurement of the low-redshift expansion rate of the Universe, but also requires breaking degeneracies with "conventional" cosmological parameters such as Ω_m and "nuisance" parameters such as the absolute magnitude of a Type Ia SN. In almost every case, different cosmological probes must be combined to break these degeneracies. This is appropriate, but the consumer of the FoM numbers must be aware of precisely which combinations of experiments have been used in order to do apples-to-apples comparisons.

As an example, consider a hypothetical SN Experiment A that measures Type Ia SN to 1% in flux (1σ) in each of 12 redshift bins equally spaced in the range $0<z<1.2$. Combining with the FoMSWG Planck and WL(DES) Fisher matrices, this SN experiment achieves $FoM_{DETF}=208$. However, this SN experiment achieves only $FoM_{DETF}=8.3$ when combined with Planck (i.e. excluding DES), due to a very weakly broken degeneracy of w_a with Ω_m and Ω_k. An alternative SN Experiment B that reaches only 2% in flux (i.e. a factor of 2 worse) achieves $FoM_{DETF}=79$ when combined with Planck and WL(DES), and achieves $FoM_{DETF}=61$ when combined with Planck and a prior assumption that the Universe be flat ($\Omega_k=0$). It is very important that the numbers "61" or "79" for B not be compared directly against the 8.3 for A. While this is a trivial example, FoM numbers stated by different projects rarely come with uniform assumptions.

Moreover, even an external prior referring to a specific experiment, e.g. "Planck," does not completely specify the Fisher matrix. The FoMSWG Fisher matrix assumed an uncertainty on the optical depth due to reionization τ of $\sigma(\tau)=0.01$ after accounting for the possibility of a complicated reionization scenario. Assuming a single-step reionization instead leads to a smaller uncertainty in τ, and improves the FoM_{DETF} from 208 to 243 – a 17% change. There are many "knobs" of this nature that can be tuned in Fisher forecasts, especially for WL and large scale structure where the fiducial galaxy population and the generality of the galaxy biasing or intrinsic alignment model can be varied. *One should always read the fine print before comparing FoM numbers.* Comparisons at the tens of percents level without standardized assumptions are generally not appropriate. As we shall see below, for particularly complicated data models dominated by systematics marginalization, even factor of ~2 differences in FoMs forecast by different groups must be treated with caution.

[2] http://jdem.gsfc.nasa.gov/fomswg.html

4. Systematic Errors

Systematic errors can be incorporated in FoM calculations by adding systematic uncertainties to the covariance matrix of the cosmological observables (e.g. Albrecht et al. 2006, 2009). However, systematic errors are often much more difficult to forecast than statistical errors. Moreover, the specific form of the assumed systematic error can be just as important as the assigned "amount" of systematic error. In some cases this may be physical. For example, in WL observations the spurious contribution to the shear power spectrum from the correlation of intrinsic galaxy ellipticities with the cosmic shear field has a specific redshift dependence that can be used to separate it from the true cosmic shear signal (Hirata & Seljak 2004). However, if the functional form of a systematic is unknown and one assumes a simplified parameterization, the error bars on dark energy parameters can be underestimated – sometimes severely – because the Fisher matrix will find a combination of observables that cancels a systematic error of the assumed form. Some recent forecasts have gone to great lengths to avoid this problem (Bernstein 2008).

Many systematic errors can be corrected (at least partially) with sufficient data. In this case, one should add to the covariance matrix the expected residual from the corrections (if it is significant in comparison to statistical errors), rather than the raw magnitude of the systematic. For example, galaxy bias can shift the BAO scale; the relevant systematic error is not the amount of the shift, but rather how well it can be determined using other information about the galaxies (e.g. their clustering amplitude). Light from supernovae suffers extinction in the host galaxy, but it would be overly conservative to assign a systematic error equal to typical galactic extinction – rather one should focus on the accuracy to which the extinction can be determined from the reddening of observed colors. A SN project with broader wavelength coverage or higher signal-to-noise ratio may be able to estimate the extinction more accurately and (if this improvement can be quantified) claim the associated reduction in systematics and increase in FoM.

The situation becomes more complicated when external data sets are brought in to constrain systematics. A common example is the need to constrain the redshift distribution of source galaxies used for WL, for which only photometric redshifts ("photo-z's") are available. The distribution can be constrained by cross-correlating the WL source galaxies within a particular range ("slice") of photo-z with an overlapping spectroscopic survey of galaxies: the photo-z sample will correlate only with the spectro-z sample at redshifts that are represented in the true redshift distribution (Newman 2008). This introduces additional nuisance parameters associated with galaxy biasing, since these tests measure not the redshift probability distribution $P(z)$ but its product with the galaxy bias, $b(z)P(z)$. If galaxy bias can be assumed to vary slowly in redshift, it is possible to measure the mean redshift $\langle z \rangle$ and σ_z width of a photo-z slice at the $\leq 10^{-3}$ level by this method (Newman 2008). However, if galaxy sub-populations with different bias also have different photo-z error – e.g. the highly biased red galaxies have better photo-zs than blue galaxies – then $\langle z \rangle$ and σ_z are degenerate with bias parameters. If one is truly agnostic about the relation between bias and photo-z error, the uncertainty in $\langle z \rangle$ is comparable to the width of the photo-z error distribution, typically $\sim 4\times10^{-2}(1+z)$. Since we are not completely ignorant about the behavior of photo-zs and galaxy bias, it may be appropriate in a case like this to assign a smaller uncertainty than $4\times10^{-2}(1+z)$ to $\langle z \rangle$, especially if this is to be interpreted as a "1σ"

uncertainty. In the case of an ambitious[3] WL survey, assigning an error of $10^{-3}(1+z)$ to $\langle z \rangle$ and combining with the FoMSWG Planck and SN data models leads to $\mathrm{FoM}_{\mathrm{DETF}}=210$. This degrades to 90 if the error on $\langle z \rangle$ increases to $5\times10^{-3}(1+z)$, and 59 if the error is $10^{-2}(1+z)$. In this example, a forecaster using an oversimplified bias model might report $\mathrm{FoM}_{\mathrm{DETF}}=210$. A forecaster trying to take into account uncertainties associated with galaxy bias (at least in some crude way) might report 59 or 90. This illustrates the danger that a quick look at the FoM may reward the experiment with the most simplified forecasting tools. More seriously, the 59 and 90 numbers could come out of two different pipelines with similar levels of sophistication with reasonable inputs. If Experiment A claimed $\mathrm{FoM}_{\mathrm{DETF}}=59$ and Experiment B claimed 90, it is unlikely that an external reviewer would be able to assess the significance of this difference.

5. Robustness

Some systematic errors are sufficiently hard to predict that they are more properly considered "risks" for a particular dark energy probe rather than additional contributions to the forecast covariance matrix. One often-discussed example is the possibility of SN luminosity evolution. Examples from other dark energy probes would include the possibility that galactic outflows have substantially altered the matter power spectrum in the quasilinear regime, thereby compromising WL measurements[4]; or the possibility that very large scale feedback, e.g. from reionization (Wyithe & Loeb 2007), could affect large scale structure or even BAO measurements in ways that cannot be captured by painting halo occupation models onto *N*-body simulations. There is additionally the possibility that new exotic physics could break a supposedly robust cosmological probe. There is no especially good way to incorporate these risks into the computation of a single FoM. One could introduce an *ad hoc* systematic model (as the DETF did for SN evolution) or otherwise penalize the FoM to take account of the risks, and Bayesians could argue for a probability distribution $P(\mathrm{FoM})$ based on prior expectations about the systematics; but such procedures can all too easily obscure risk and subjective choices behind the numbers. Ultimately, the nastiest astrophysical risks must be considered as part of the robustness of an overall dark energy program.

Astrophysical risks in future projects can be mitigated to some extent by the use of multiple techniques of measuring dark energy. Here again the FoM can play a role, by showing how much degradation occurs with the loss (or partial loss) of any of the techniques used. For example, using the combined FoMSWG Fisher matrix forecasts for the CMB, SN, WL, and BAO, one finds that in 2016 we should achieve a $\mathrm{FoM}_{\mathrm{DETF}}=116$. This degrades with the use of only some combinations of the data: one finds $\mathrm{FoM}_{\mathrm{DETF}}=28$ (CMB+SN+WL), 31 (SN+WL+BAO), 84 (CMB+WL+BAO), or 87 (CMB+SN+BAO). Thus for the measurement of w_0 and w_a, the loss of the CMB or BAO techniques would be most damaging, but some capability remains with any 3 of the 4 techniques. For the growth of structure, we find $\mathrm{FoM}_{\gamma}=0$

[3] The actual parameters used for this example were a galaxy density of 30 arcmin^{-2}, median redshift $z_{\mathrm{med}}=1.1$, sky coverage of 10^4 deg^2, 14 redshift slices, the FoMSWG intrinsic alignment model, and errors on $\langle z \rangle$ uncorrelated between redshift slices. As in FoMSWG, the shear power and cross-spectra, and the shear ratio test (Jain & Taylor 2003) were used.

[4] See Levine & Gnedin (2006) for an extreme example invoking AGN feedback. While this model is probably unrealistic as it allows AGNs to evacuate all baryons from their host haloes, a smaller but still significant effect may be possible.

if WL is dropped since none of the other techniques measures low-redshift growth of structure. In order to measure the growth of structure by multiple techniques, an additional technique would have to be added to the mix.

The formalism can also model partial losses of specific techniques, e.g. using the FoMSWG Fisher matrices, $\mathrm{FoM}_{\mathrm{DETF}}$ degrades to 59 if the BAO error bars are inflated by a factor of 2 (e.g. if density field reconstruction techniques were unavailable).

These "degraded" FoMs provide a measure of robustness against loss or degradation of one technique. However they are not by themselves the last word on robustness, for at least three reasons.

First, *in order to claim even a degraded FoM in the event of a failure of one technique, one must be able to identify which technique failed*. In an experiment that does SN+WL+BAO, a "failure" of one technique means that SN+WL, WL+BAO, and SN+BAO all give different answers. One of these three constraints is correct, but which one? It is much better if each technique has sufficient internal checks to diagnose its own health and veto itself if necessary. For example, BAO measurements with spectroscopic redshifts can test whether the radial and tangential distance scales are consistent, and the WL shear ratio or "cosmography" method (Jain & Taylor 2003) can test whether all of the shear ratios (which depend nontrivially on two redshifts, z_1 and z_2) are consistent with a single distance function $D(z)$. These internal checks are arguably as necessary as the use of multiple "techniques."

Second, *the degraded FoM as a "robustness" measure is contingent on the independence of the techniques*. This is not as obvious as it seems. For example, the physics of SN and WL are so different that their independence is a safe bet. On the other hand, a project that observes supernovae in one small field, and uses the same field to calibrate photo-*z*s for WL, could run into serious trouble if that one small field turns out to exhibit unusual Galactic dust properties. Since it is unlikely that the SN and WL techniques would be affected in the same way, cross-checks of these methods would still bolster confidence in the final result if both techniques worked – but the probability of both techniques failing or falling short may be greater than the product of their individual risks.

Third, *the degraded FoM is not a substitute for making each technique as robust as possible*. A project that does 4 techniques sloppily runs the risk that all 4 will fail for different reasons.

6. Summary

The FoM as a benchmark for progress in measuring dark energy has proven to be remarkably popular and versatile. It can be used to rapidly assess the dark energy performance of many combinations of experiments, and quantify the impact of systematic errors. It can also be used to assess the robustness of the overall dark energy program if a particular technique fails to achieve its full potential, or if a hardware failure or programmatic necessity degrades or deletes certain experiments. Its popularity has swept throughout the dark energy community, and it is now in use by theorists and observers, and has been applied to projects on the ground and in space, and across wavelengths from X-rays to radio. But as with any forecasting tool, it is open to abuse and misinterpretation if not used carefully. In particular:

- FoM forecasts often depend critically on external data sets (e.g. Planck) to break degeneracies.

- The way in which systematic errors are included in a FoM can have a substantial effect on the answer. This includes not just the tagline – "we allow for 1% systematics in the supernova flux" – but also the specific parameterization and the correlation matrix of systematics parameters.
- The commonly used FoM_{DETF} has units of inverse variance: the difference between FoM_{DETF}=500 and 350 is equivalent to a 30% reduction in inverse variance or a 20% increase in standard deviation (σ). It is rare even for experienced observers to be able to predict error bars to this accuracy years before their hardware is built. It is also unlikely that an external reviewer could assess the relative level of conservatism of different experiments at this level, especially if the data model is complicated or the error bars contain a substantial contribution from systematics.
- Degraded FoMs computed with a subset of the dark energy techniques can quantify performance in the event of the loss or degradation of one technique. They tell part of the story on the robustness of an experiment or program. However, other factors less amenable to quantitative assessment are also critical: the robustness of the techniques and their implementation, risk factors common to multiple techniques, and the ability to diagnose at the end of the day which techniques have worked as hoped and which have not.

References

Albrecht A. et al. 2006, "Report of the Dark Energy Task Force" (astro-ph/0609591)

Albrecht A. et al. 2009, "Findings of the Joint Dark Energy Mission Figure of Merit Science Working Group" (arXiv:0901.0721)

Albrecht A., Bernstein G. 2007, "Evaluating dark energy probes using multidimensional dark energy parameters" *Phys. Rev. D* **75**:103003

Bernstein G. 2008, "Comprehensive Two-point analyses of weak gravitational lensing surveys" *Astrophys. J.* submitted (arXiv:0808.3400)

Hirata C., Seljak U. 2004, "Intrinsic alignment-lensing interference as a contaminant of cosmic shear" *Phys. Rev. D* **70**:063526

Jain B., Taylor A. 2003, "Cross-correlation tomography: Measuring dark energy evolution with weak lensing" *Phys. Rev. Lett.* **91**:141302

Levine R., Gnedin N. 2006, "Active galactic nucleus outflows and the matter power spectrum" *Astrophys. J.* **649**:L57

Liddle A. et al. 2006, "Present and future evidence for evolving dark energy" *Phys. Rev. D* **74**:123506

Mukherjee P. et al. 2006, "Model selection as a science driver for dark energy surveys" *Mon. Not. R. Astron. Soc.* **369**:1725

Newman J. 2008, "Calibrating redshift distributions beyond spectroscopic limits with cross-correlations" *Astrophys. J.* **684**:88

Perlmutter S. et al. 1999, "Measurements of Omega and Lambda from 42 high-redshift supernovae" *Astrophys. J.* **517**:565

Riess A. et al. 1998, "Observational evidence from supernovae for an accelerating Universe and a cosmological constant" *Astron. J.* **116**:1009

Wyithe J., Loeb A. 2007, "The imprint of cosmic reionization on galaxy clustering" *Mon. Not. R. Astron. Soc.* **382**:921