Monthly Notices of the ROYAL ASTRONOMICAL SOCIETY LETTERS

Mon. Not. R. Astron. Soc. **401,** L15–L18 (2010) doi:10.1111/j.1745-3933.2009.00776.x

The virtues of frugality – why cosmological observers should release their data slowly

Glenn D. Starkman,[1]* Roberto Trotta[2] and Pascal M. Vaudrevange[1]

[1] *CERCA & Department of Physics, Case Western Reserve University, 10900 Euclid Ave, Cleveland, OH 44106, USA*
[2] *Astrophysics Group, Imperial College London, Blackett Laboratory, Prince Consort Road, London SW7 2AZ*

Accepted 2009 October 9. Received 2009 October 8; in original form 2009 August 26

ABSTRACT

Cosmologists will soon be in a unique position. Observational noise will gradually be replaced by cosmic variance as the dominant source of uncertainty in an increasing number of observations. We reflect on the ramifications for the discovery and verification of new models. If there are features in the full data set that call for a new model, there will be no subsequent observations to test that model's predictions. We give specific examples of the problem by discussing the pitfalls of model discovery by prior adjustment in the context of dark energy models and inflationary theories. We show how the gradual release of data can mitigate this difficulty, allowing anomalies to be identified and new models to be proposed and tested. We advocate that observers plan for the frugal release of data from future cosmic-variance-limited observations.

Key words: methods: data analysis – methods: statistical – cosmic microwave background – cosmology: observations.

1 INTRODUCTION

Cosmologists can only make observations on (or occasionally within) our past light cone. Whatever the reality of the multiverse, we Earth-bound humans of the $T_{\rm CMB} = 2.725\,{\rm K}$ era have access to only a finite volume of space, containing finite energy and information. The exciting period in which we find ourselves learning more and more about this volume of accessible space and its contents cannot last forever. While we are unlikely to gather *all* the existing information content of the observable Universe, we are already making substantial inroads on the information of cosmic significance.

The most notable example of confronting the finite information content of the Universe is our measurements of the power in the lowest multipoles C_ℓ of the cosmic microwave background (CMB) temperature anisotropies. Their statistical error bars are now smaller than the 'cosmic variance' errors – the expected difference between what we measure for these multipoles and what we would measure if we could average over many independent horizon volumes. The range of ℓ for which this is true is increasing as the *Wilkinson Microwave Anisotropy Probe* (*WMAP*) continues to report new results. This trend will accelerate as new experiments join the fray. (Though we could wait a few hundred million years to gain access to a mostly independent last scattering surface.)

The CMB temperature–temperature power spectrum is unlikely to be the last place where the finite Universe limits cosmology. Astronomical surveys are already cataloguing an increasing fraction of all the structures within our past light cone. Redshifted hydrogen hyperfine instruments will eventually extend the volume over which we map the structure of matter nearly out to the horizon.

There are consequences to becoming a data-limited science. We upset the balance between applying the brain's remarkable pattern-finding abilities and testing the robustness of the patterns we discover. We may see patterns in finite data, but, unable to collect new data, we have no way to confirm their reality, missing out on potentially significant discoveries. We risk falling for what particle physicists call 'the look elsewhere effect', i.e. the spurious 'discovery' of statistically significant anomalies which are merely the consequence of performing a large number of tests on the same data. A small fraction of these are bound to report significant 'evidence' for unexpected features due to random noise. Unlike experimental scientists, we may no longer be able to collect data, form a new hypothesis and test its predictions. Our ability to distinguish between statistical fluctuations and real effects becomes limited.

Given that the challenge of finite data is upon us, our best hope is to devise strategies to minimize its effects. The approach that we shall explore and advocate is to simulate the cycle of data acquisition and analysis by being frugal. By allowing colleagues to see only subsets of the data, construct hypotheses based on them and then test those hypotheses on larger subsets, we can aim to avoid unexplained anomalies with untestable explanations.

The benefits of frugality arise not from some magical improvements in the statistical power of the data, but from acknowledging and mitigating a basic human failing: over-confidence. Specifically, by assigning all probability to the set of physical models that we have thought about and consequently zero probability to all other models, we ignore that we may not have considered the correct model. Frugality allows us to redress those wrongs by admitting such

*E-mail: gds6@case.edu

models and testing their predictions on our remaining data. We examine the effects of (and several strategies for) dividing cosmological data into several pieces so that new models can be consistently explored.

2 MODEL DISCOVERY

2.1 Bayesian model selection and prior updating

We take a Bayesian outlook on hypothesis testing, as we believe (and show below) that this closely reflects the way we think about models. Another reason for being wary of the usual (frequentist) practice of reporting p values is that the latter are *not* probabilities for hypotheses, despite being commonly misinterpreted as such (Sellke, Bayarri & Berger 2001; Gordon & Trotta 2007). Suppose we have a model M_0 with parameters θ_0, that we wish to evaluate in light of data d. Our updated state of belief in the model's parameters is given by the posterior probability distribution function (pdf) on θ_0, obtained via the Bayes theorem:

$$p(\theta_0|d, M_0) = p(d|\theta_0, M_0)\frac{p(\theta_0|M_0)}{p(d|M_0)}, \tag{1}$$

where $p(d|\theta_0, M_0)$ is the likelihood, $p(\theta_0|M_0)$ the prior on the parameters θ_0 and $p(d|M_0)$ is the marginal likelihood for M_0. Now suppose we note a feature in the data that is not reproduced by model M_0 (e.g. by computing the doubt, as in Starkman, Trotta & Vaudrevange 2008). We invent a model M_1 with parameters θ_1 as an explanation for the said feature and compute the evidence for both models ($i = 0, 1$):

$$p(d|M_i) = \int \mathrm{d}\theta_i\, p(d|\theta_i, M_i)p(\theta_i|M_i)\,. \tag{2}$$

Each model's posterior probability in light of d is given by $p(M_i|d) = p(d|M_i)p(M_i)/p(d)$. The *ratio* of our degrees of belief in the models, the Bayes factor $B_{10} = p(d|M_1)/p(d|M_0)$, penalizes models that are unnecessarily complex, e.g. because of an excessive number of free parameters, automatically encapsulating Occam's razor (see e.g. Trotta 2007a, 2008). In order to increase confidence in the new model M_1, all that is required is $B_{10} > 1$, i.e. that M_1 be a more 'effective' description of the *presently* available data. There is no dependence on the model's predictivity for *future* observations.

In practice, a new model probably would not (and arguably should not) be accepted until it produces a correct prediction for future data d' that differs from the old model's, thus enabling the models to be distinguished. Formally, the models' relative posterior odds after seeing both sets of data are given by

$$\frac{p(M_1|d, d')}{p(M_0|d, d')} = \frac{p(d'|M_1)}{p(d'|M_0)}\frac{p(d|M_1)}{p(d|M_0)}\frac{p(M_1)}{p(M_0)}\,. \tag{3}$$

Before the data set d came along, model M_1 was not even on the table: $p(M_1) = 0$. The step of introducing M_1, *while absolutely crucial*, formally requires the injection of an infinite amount of information to raise $p(M_1)$ from 0 to a finite value. This prior adjustment is on top of the change in degree of belief coming from d. It amounts to using the data d twice, first to introduce M_1 by adjusting its prior and then to evaluate the evidence from d.

The duplicate use of the data d leads to posterior odds which can seriously overstate the statistical significance of a new effect. We suggest to 'forget' about the details of d, compress its information into a new non-zero (and still subjective) prior $p(M_1)$ and then compute the posterior odds arising solely from d', i.e.

$$\frac{p(M_1|d, d')}{p(M_0|d, d')} = \frac{p(d'|M_1)}{p(d'|M_0)}\frac{p(M_1)}{p(M_0)}. \tag{4}$$

If an unlimited amount of data is accessible and the anomaly is correctly modelled by M_1, it is *guaranteed* to become eventually favoured by the Bayes factor, independent of the exact choices of priors. Using a finite, cosmic-variance-limited data set only *increases the likelihood* that M_1 is confirmed before the data are exhausted, the more the bigger the fraction of unused data in d'.

2.2 Examples of prior adjustments in cosmology

Two notable examples in cosmology of devising new models and then adjusting their priors are the discovery of dark energy and the realization that inflation can easily accommodate $\Omega < 1$.

The discovery of a non-zero, yet tiny cosmological constant Λ was in stark contradiction to prior expectations. Particle-physics considerations suggested that Λ should either be 0 (model $\mathcal{M}_1$) or have a uniform prior between $\pm M_\mathrm{p}^4$ (model $\mathcal{M}_2$), $p(\Lambda|\mathcal{M}_1) = \delta(\Lambda)$, $p(\Lambda|\mathcal{M}_2) = \Theta(|\Lambda| - M_\mathrm{p}^4)/2M_\mathrm{p}^4$, where M_p is the reduced Planck mass, $\Theta(x)$ is a step function and $\delta(x)$ is a Dirac delta distribution. Oversimplifying history, let us assume that these were the only theories at hand and had equal priors:[1] $p(\mathcal{M}_1) = p(\mathcal{M}_2) = \frac{1}{2}$.

Along came supernova (SN) redshift measurements (Perlmutter et al. 1999), suggesting a late-time acceleration of the Universe driven by (in the simplest models) a small $\frac{\Lambda_0}{M_\mathrm{p}^4} \approx 10^{-120}$. To simplify, let us assume that the available SN data presented a 5σ deviation from $\Lambda = 0$. Computing the Bayes factor using the Savage–Dickey density ratio (Trotta 2007b) gives

$$B_{12} = \frac{p(\Lambda = 0|d, \mathcal{M}_2)}{p(\Lambda = 0|\mathcal{M}_2)} = \frac{10^{121}}{\sqrt{2\pi}}\mathrm{e}^{-25/2} \approx 10^{115}\,. \tag{5}$$

Due to strong Occam's razor effect of the prior on $\mathcal{M}_2$, a vanishing cosmological constant should have still been vastly preferred, with odds of the order of 10^{115}:1, over a model including a hugely fine-tuned Λ. An $\sim$23σ detection of a non-zero cosmological constant would have been required to over-ride Occam's razor of the prior.

However, the particle-physics community started reconsidering priors and developed a new model $\mathcal{M}_3$ involving anthropic reasoning which gave more weight to small values of Λ, $p(\Lambda|\mathcal{M}_3) = \Theta(10\Lambda_0 - \Lambda)/10\Lambda_0$, with model priors now being $p(\mathcal{M}_1) = p(\mathcal{M}_2) = p(\mathcal{M}_3) = \frac{1}{3}$. Under the new anthropic prior, the effect of Occam's razor is vastly reduced, giving a Bayes factor $B_{13} \approx 10^{-4}$, now favouring model $\mathcal{M}_3$. The parameter value that was a priori considered unnatural under the original model for a cosmological constant (small non-zero Λ) described the data better than the prevailing model of $\Lambda = 0$, but not sufficiently well to be preferred. Introducing an anthropic model based on the landscape picture in string theory (Bousso & Polchinski 2000; Giddings, Kachru & Polchinski 2002; Douglas 2003; Susskind 2003; Starkman & Trotta 2006) allowed a small, non-zero cosmological constant to become the preferred description of the data which has since been supported by other observations such as CMB and baryon acoustic oscillations.

It is interesting that *ex post facto* one might argue that perhaps Λ is restricted to be a positive quantity, in which case the appropriate prior would be uniform in $\ln \Lambda$ rather than in Λ (Evrard & Coles 1995; Kirchner & Ellis 2003). Under this model $\mathcal{M}_4$, and assuming a cut-off $\Lambda > \Lambda_\mathrm{min} = 10^{-500} M_\mathrm{p}^4$ (see Starkman & Trotta 2006), one

[1] An interesting suggestion for choosing the model's priors based on a maximum entropy argument has been put forward by Brewer & Francis (2009).

obtains a Bayes factor $B_{14} \approx 10^{-2}$, i.e. moderate support for Λ, analogously to what can be obtained by anthropic arguments.

An earlier example of discovering a new model through adjusting priors happened in the mid- to late-90s. The overwhelming evidence for $\Omega_{\rm tot} \approx 0.3 < 1$ posed a problem for inflation, as it had been viewed to generically predict a flat universe with $\Omega \approx 1$ to high accuracy – this generally accepted model could not describe observations. Different models (mostly using multiple stages of inflation) were devised that produced open universes (Bucher, Goldhaber & Turok 1995). In other words, after observing that $\Omega \approx 0.3$, the priors for single-stage inflation, $p(M_0)$, and for multistage inflation, $p(M_1)$, were adjusted from $p(M_0) \gg p(M_1)$ to $p(M_1) \approx p(M_0)$. The prediction for future observations – corroborating evidence for $\Omega \approx 0.3$ – was proven wrong by measurements of $\Omega \approx 1$ (Netterfield et al. 2002). The priors were reverted back to $p(M_0) \gg p(M_1)$, making multistage models all but obsolete.

Note that in both the above examples, it was crucial that predictions of the new model could be tested by follow-up *independent* observations which either confirmed or rejected the new model.

3 THE NEED FOR FRUGALITY

With the launch of the *Planck* satellite, the power spectrum of the temperature fluctuations, $C_\ell^{\rm TT}$, will be limited by cosmic variance all the way up to $\ell > 2000$. No future observation will ever obtain more precise measurements of the CMB temperature fluctuations in this ℓ range (barring problems with unanticipated systematics), and higher ℓ ranges begin to be dominated by foreground sources. If there are features in the *Planck* data that cannot be adequately explained by Λ cold dark matter (ΛCDM; such as a strong correlation between different multipoles), we could and should devise a revised concordance model. But we would be unable to test its predictions with future CMB temperature measurements!

After the *Cosmic Background Explorer* (*COBE*) experiment (Smoot et al. 1992) observed hints of a low quadrupole, it took subsequent confirming measurements by *WMAP* to establish this (Spergel et al. 2003, 2007; Komatsu et al. 2009) and to detect the planarity of the quadrupole and octopole and their alignment with each other, perpendicular to the ecliptic, with an axis towards the CMB dipole (de Oliveira-Costa et al. 2004; Schwarz et al. 2004; Land & Magueijo 2005), where cosmic variance is already the limiting factor. Thus, possible new models explaining the low ℓ multipole alignments cannot be tested on their predictions for future measurements of these multipoles. Instead, one has to look for different predictions from the new models, e.g. by looking for circles in the sky as a signature of a topologically non-trivial universe (Cornish et al. 2004). If only parts of the *WMAP* data had been released, tantalizing enough to induce people to look for new models, there would have been room to test the predictions of these models for the low ℓs.

In the (perhaps not so distant) future, a similar situation will arise with other cosmological experiments. Large-scale structure observations by way of galaxy counts will eventually measure the positions and redshifts of all galaxies in our Hubble patch with high precision (neglecting uncertainties due to non-linearities). The distribution of hydrogen will be mapped with observations or the Lyα forest. Eventually all observations on cosmological scales will reach the cosmic variance limit, as we only have this one Universe from which to sample.

In light of this, it seems imperative to reflect on ways to extract an optimal amount of information from complete finite data sets. They should be used not only to better constrain the parameters of the concordance model, but also to discover and test new models. We need to devise schemes for incremental data release as cosmological analogues of blind analysis, a procedure often used in particle physics, where the need to avoid the (possibly unconscious) influence of the statistical methodology adopted on the significance of the results is a well-recognized problem (see e.g. Lyons 2008). For example, one wants to avoid (unwillingly) biasing the significance of a signal when designing the ‘cuts’ on the number of observed events. Several strategies have been devised to this end. For example, a random number can be added to the data and subtracted only after all corrections and other data manipulations have been performed, or just a fraction of the data is employed to define the statistical procedure, while the remainder of the data are only revealed in a subsequent phase. After that point, no further adjustments of the methodology are allowed. The split of data into subsets can either happen in time (an obvious solution for many particle-physics experiments) or in data space. In the latter case, a ‘signal box’ of data is left closed until potential anomalies in the first chunk of observations have been identified and statistical tests for their confirmation designed, at which point the box is opened and the analysis unblinded. An example of such a procedure is the miniBooNE neutrino oscillation experiment (Bazarko 2001). Another method is sometimes adopted by precision measurements where the analysis team is allowed to see the full data sets, but with arbitrary units. The resulting parameter constraints are rescaled to the actual units only at the very end of the analysis.

All of these strategies are designed with the common aim of keeping a part of the information hidden from the first stage of the analysis, so as to be able to exploit the full statistical power of the hidden data upon unblinding. We now turn to the discussion of possible ways of applying this idea in the cosmological context.

4 STRATEGIES FOR THE RELEASE OF PARTIAL DATA

There is always a random element involved in choosing a good way to split data, where the definition of ‘good’ often depends on the unknown anomalies one is hoping to be able to test. Suppose we throw a single coin $2N$ times after which it is lost. The first N throws include an equal number of heads and tails, while the last N tosses are all tails. Splitting this data set into these two chunks, the first set points towards the model of a fair coin. The second set (all tails) raises serious doubt about this model. But we have no way of verifying the predictions of a new model (e.g. the coin was exchanged for an all-tails coin) as the coin was lost. Had we split the data into four equal chunks, then after examining the third chunk we would likely have proposed a new model of an unfair all-tails coin. The predictions of this new model would have been tested (and confirmed) by the fourth chunk of data.

Two opposing forces are at play when considering ways to release partial data. On the one hand, releasing individual data points will lead to many statistical flukes that can be mistaken for features in the data. On the other hand, releasing all data at once will only allow us to determine the parameters of the existing models and not to check the predictions of potential new models. It seems hard to find an optimal number of chunks, even more so as it is not even clear how data should be split.

The most natural way to release partial data is often by time ordering, such as is employed by many experiments, e.g. *WMAP*. A natural cut-off between data sets is the point in time when (if) the doubt (Starkman et al. 2008) on a concordance or reference model reaches a critical threshold, after which an alternative model should be devised. Using only data that were not used to compute the doubt on the original model, compute the doubt on the new model. Iterate this process until all data have been taken or funding runs out. This method does not detect all features as the likelihood function typically does not incorporate all predictions of the original model. For example, the riddles of why the two-point correlation function of the temperature fluctuations vanishes at separation angles larger than 60° (Copi et al. 2006) and of the alignments of the quadrupole and octopole (de Oliveira-Costa et al. 2004; Schwarz et al. 2004; Land & Magueijo 2005) would escape detection as the likelihood function is insensitive to these features.

Summary statistics for CMB measurements are often presented in the form of (binned) C_ℓs building on isotropy and Gaussianity of a_{lm}s. Other quantities, such as $C(\theta)$, would work as well. A possible course of action would be to exclusively release binned C_ℓs in the first data release. Then a search for deviations from the concordance model – new features – could be conducted. If any unexpected features are noted in the data, new models would be devised and their predictions for the unbinned C_ℓs could be compared against the second, unbinned data release. One might envision performing a finer graining of the binning process, going from e.g. $\Delta\ell = 10$ bins in the first year to $\Delta\ell = 5$ bins in the second year to $\Delta\ell = 1$ bins in the third year, or in terms of the two-point function $C(\theta)$ using averaged values over $\Delta\theta = 10^\circ, 1^\circ, 0\overset{\circ}{.}1, \ldots$ for each release cycle. A possible complication is the fact that the successive data releases include the previous data and hence are correlated.

However, there is a way to split data guaranteeing uncorrelated data chunks: principal component analysis (PCA; Huterer & Starkman 2003). Each principal component, i.e. eigenvector and eigenvalue of the covariance matrix of the data, is released separately, giving as many attempts at finding new models as there are well-constrained PCAs. Their order seems to be a matter of taste. Releasing the best-constrained component first would make it easiest to detect any features and then using the less-well-constrained modes to verify any new model. Not producing any hints at a new model, this procedure – as any splitting of data – would not have any negative impact on parameter estimation (as Bayesian updating of posterior pdfs does not care about the order of the information being added).

Independent of how the data are split, sizing the individual chunks also seems to be rather an art. They should neither be too small, i.e. not so noisy as to induce spurious features, nor too large, or new models will not be testable. It may prove beneficial to release data chunks with the same information content, as measured e.g. by the mean square error or an information-theory-based measure such as the Kullback–Leibler divergence.

5 CONCLUSIONS

Cosmologists are in a paradoxical situation. They strive to acquire data of the highest possible quality to constrain parameters of their models as quickly as possible. But they should be open to new features in the data that are not predicted by current models, and hence to the possibility of having to devise new models and test their predictions. We have argued that for the latter step, availability of fresh data is crucial, which for cosmic-variance-limited data sets is simply not possible. We therefore propose that such *ultimate* data sets be treated as the precious resources they are and released slowly and carefully.

We have discussed various strategies for parsing such data sets. It remains an art to find the optimal way to split data and release it, involving inevitably a certain degree of luck to detect unexpected features. It seems to us from this first overview that the most promising way of 'dividing the plunder' is to employ a PCA decomposition of the data and release data parts of equal information content. This is a compromise between being able to find new features and having enough data left to reliably test possible new models. However, the best strategy is likely to depend heavily on the particular data set and on the taste of the individual investigators. Wishing to avoid that basic human failing of over-confidence, we acknowledge that there is a reasonable chance that we have overlooked the optimal strategy.

We urge our observational colleagues to be frugal with their data. Slicing the data and doling it out slowly is in all of our long-term best interests.

ACKNOWLEDGMENTS

We would like to thank Andrew Jaffe, Louis Lyons, Irit Maor and John Ruhl for stimulating discussions. GDS and PMV are supported by a grant from the DOE to the particle-astrophysics group at CWRU. PMV is supported by the office of the Dean of the College of Arts and Sciences at CWRU. RT would like to thank CWRU for hospitality. We thank E. Scrooge for motivation.

REFERENCES

Bazarko A., 2001, Nucl. Phys. Proc. Suppl., 91, 210
Bousso R., Polchinski J., 2000, JHEP, 06, 006
Brewer B., Francis M., 2009, preprint (arXiv:0906.5609)
Bucher M., Goldhaber A. S., Turok N., 1995, Phys. Rev. D, 52, 3314
Copi C. J., Huterer D., Schwarz D. J., Starkman G. D., 2006, MNRAS, 367, 79
Cornish N. J., Spergel D. N., Starkman G. D., Komatsu E., 2004, Phys. Rev. Lett., 92, 201302
de Oliveira-Costa A., Tegmark M., Zaldarriaga M., Hamilton A., 2004, Phys. Rev. D, 69, 063516
Douglas M. R., 2003, JHEP, 05, 046
Evrard G., Coles P., 1995, Class. Quant. Grav., 12, L93
Giddings S. B., Kachru S., Polchinski J., 2002, Phys. Rev. D, 66, 106006
Gordon C., Trotta R., 2007, MNRAS, 382, 1859
Huterer D., Starkman G., 2003, Phys. Rev. Lett., 90, 031301
Kirchner U., Ellis G., 2003, Class. Quant. Grav., 20, 1199
Komatsu E. et al., 2009, ApJS, 180, 330
Land K., Magueijo J., 2005, Phys. Rev. Lett., 95, 071301
Lyons L., 2008, Ann. Appl. Stat., 2, 887
Netterfield C. B. et al., 2002, ApJ, 571, 604
Perlmutter S. et al., 1999, ApJ, 517, 565
Schwarz D. J., Starkman G. D., Huterer D., Copi C. J., 2004, Phys. Rev. Lett., 93, 221301
Sellke T., Bayarri M., Berger J. O., 2001, American Statistician, 55, 62
Smoot G. F. et al., 1992, ApJ, 396, L1
Spergel D. N. et al., 2003, ApJS, 148, 175
Spergel D. N. et al., 2007, ApJS, 170, 377
Starkman G. D., Trotta R., 2006, Phys. Rev. Lett., 97, 201301
Starkman G. D., Trotta R., Vaudrevange P. M., 2008, preprint (arXiv:0811.2415)
Susskind L., 2003, preprint (arXiv:hep-th/0302219)
Trotta R., 2007a, MNRAS, 378, 72
Trotta R., 2007b, MNRAS, 378, 819
Trotta R., 2008, Contemporary Phys., 49, 71

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