PROCEEDINGS OF SCIENCE

The Virtues of Frugality - Why Cosmological Observers should Release their Data Slowly

Pascal M. Vaudrevange[*a], **Glenn D. Starkman**[a], **Roberto Trotta**[b]

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

The increasing quality of cosmological observations is both a blessing and a curse. On the one hand, better data means tighter constraints on parameters and a higher chance of discovering deviations from predictions of the concordance model. On the other hand, it poses a problem for model discovery: data that is limited by cosmic variance cannot be improved upon by future measurements. We elaborate on the consequences of this condundrum and present some strategies to mitigate its consequences.

International Workshop on Cosmic Structure and Evolution - Cosmology2009,
September 23-25, 2009
Bielefeld, Germany

[*] Speaker.

1. Introduction

Our knowledge of the universe comes from data that is confined in a spatially and temporally finite box: most measurements are taken of events on our past light cone, limited in the distant past by the surface of last scattering. According to the cosmological concordance model, all inhomogeneities in the energy density of the universe – be it in matter, radiation or gravitational waves – were seeded by fluctuations of a random (possibly slightly non-)Gaussian field. In other words, we are dealing with a single realization of a non-repeatable random process. Thus, all measurements are subject to sample variance, also called cosmic variance in this context, which represents a fundamental limit on the size of observational/ theoretical error bars.

Based on a variety of observations within the past decade – from measurements of inhomogeneities in the cosmic microwave background radiation (CMB) and the distribution of large scale structure (LSS) to redshift measurements of distant supernovae – the cosmological concordance model has emerged. After going a series of iterations of model discoveries, e.g. finding unexpectedly flat rotation curves of galaxies and devising dark matter to explain the missing visible matter, it has become clear that the total energy density of the universe is approximately the critical energy density, $\Omega_{\text{tot}} \approx 1$: baryonic and dark matter contribute roughly a quarter to the total energy content, and dark energy provides the missing three quarters to make the universe spatially flat, or nearly so.

The very same procedure of model discovery – finding some unexpected observational feature, building a model to describe it, and then verifying the model's prediction by taking more data – that led to the construction of the concordance model is now in danger of becoming inapplicable. After the arrival of cosmic-variance-limited experiments, there will soon be no way to test predictions of new models as the measurements cannot get any better. For example, the recently launched Planck satellite will measure the angular power spectrum C_ℓ of the temperature anisotropies in the CMB with cosmic variance-limited error bars up to $\ell > 2000$. Any potential unexpected feature in its measurements of the C_ℓ would still be attempted to be described by new models. However, these models' predictions could not be verified by future measurements of the temperature anisotropies – we will know everything there is to know about them (barring unexpected systematics). In this work, we describe the process of Bayesian model discovery and present different ways to mitigate the implications of cosmic variance limited data sets. A more detailed analysis can be found in [1].

2. Bayesian Model Discovery

In standard Bayesian statistics, the posterior belief in the parameters θ_0 of a model M_0 given data d is computed according to

$$p(\theta_0|d,M_0) = p(d|\theta_0,M_0)\frac{p(\theta_0|M_0)}{p(d|M_0)}, \tag{2.1}$$

where $p(d|\theta_0,M_0)$ is the likelihood, $p(\theta_0|M_0)$ is the prior on the parameters θ_0 and $p(d|M_0)$ is the marginal likelihood for M_0. Let us assume that we notice a feature in the data d that is not well described by M_0. Thus we invent a new model M_1 in the hope of better fitting the data, compute both models' evidences

$$p(d|M_i) = \int d\theta_i \, p(d|\theta_i, M_i) p(\theta_i|M_i) \,, \tag{2.2}$$

and evaluate their ratio $B_{01} = p(d|M_0)/p(d|M_1)$. This Bayes factor automatically applies Occam's razor to penalize models with too large a number of parameters (see e.g. [2, 3]). Notably, in Bayesian model comparison, the relative degree of belief in two competing models is computed taking into account only the data presently at hand.

On an intuitive level, one would expect that in addition to the relative goodness-of-fit to existing data, the predictivity of the models should also taken into account. In other words, a new model would only be accepted after it correctly predicted features in future data sets d'. Formally, real life model discovery is performed by computing the models' relative posteriors

$$\frac{p(M_1|d,d')}{p(M_0|d,d')} = \frac{p(d'|M_1)}{p(d'|M_0)} \frac{p(d|M_1)}{p(d|M_0)} \frac{p(M_1)}{p(M_0)} \,, \tag{2.3}$$

where the prior on model M_1 vanishes $p(M_1) = 0$ as it was not invented when the first data set d was collected. In order to make sense of this expression, this prior needs to be set to a non-zero value. However, blindly plugging in a non-zero $p(M_1)$ into the above expression implies using the data d twice: once to make the prior non-zero, and then to compute the evidence for M_1. Clearly, this duplicate use of data can have serious side-effects on the statistical significance of the new model.

Thus we suggest to encode all the information contained in d in the prior $p(M_1)$ and compute the relative degree of belief in the two models as

$$\frac{p(M_1|d,d')}{p(M_0|d,d')} = \frac{p(d'|M_1)}{p(d'|M_0)} \frac{p(M_1)}{p(M_0)} \,. \tag{2.4}$$

It does not matter which exact value is chosen for the non-zero prior if there is an unlimited amount of future data d': if M_1 correctly models the feature in the data, its evidence will become bigger than the old model's after enough data d' has been collected. However, for a finite amount of future data sets, such as we are facing due to cosmic variance, it is not clear whether model M_1 will ever become the preferred one even if it is correct. Turning the argument around: we know that data sets will (soon) be limited by cosmic variance. In order to have as many iterations as possible of the procedure outlined above, we need to be very careful to not "waste" any data.

3. Frugality

Nowadays, we are already facing the consequence of cosmic variance limited measurements, e.g. when discussing the anomalies surrounding the low-ℓ multipoles of the CMB. The low-ℓ part of the temperature autocorrelation function shows somewhat systematically low values (but still within the cosmic variance error bars). More curiously, the first couple of multipoles are aligned along a common axis [4, 5, 6]. Any model that attempts to describe this feature, like non-trivial topologies [7], cannot be tested by its predictivity for the low-ℓ C_ℓs as no future measurements will be able to obtain better data. Instead, different signatures must be found that can be experimentally verified, e. g. circles in the sky as signs of non-trivial topology.

The situation will get worse in the near future. Once Planck data is released, we will know everything there is to know about the temperature fluctuations. Any models based on anomalies discovered in the angular TT power spectrum must be verified by their predictions for other observables.

But not only CMB experiments are affected. A similar situation will sooner or later arise for other cosmological observations as well. There is just one universe to sample from, so all measurement will eventually be limited by sample variance. Observations of large scale structure will at some point map out the positions of all galaxies in our Hubble patch (although this might be in the somewhat distant future) and measurements of the Ly-α forest will produce maps of the hydrogen distribuion.

Thus it seems highly beneficial to stretch the available data so that no opportunity at discovering new models is squandered. Releasing data in chunks at a time – ideally just large enough to instill doubt –, would allow model builders to work devise new models and compute their predictions for the next chunk of data. Of course, some anomalies might turn out to be caused by systematics and thus would disappear in subsequent data releases. But iterating this procedure until all data is exhausted would allow for several attempts at discovering new models.

It should be noted that splitting the data into different chunks does not impact parameter estimation. After all data is released, the parameter limits will be identical and independent of whether they were derived from a sequence of partial data or from the full data release – if the concordance model prevails. The big question remaining is how to exactly split the data into chunks.

4. Strategies for Releasing Partial Data

There seems to be no objectively best way to determine how to divide the data. The optimal strategy obviously depends on the (unknown!) anomaly that might hide in the data. Suppose for example that we toss a coin $2N$ times, and obtain an equal number of heads and tails in the first N tosses, and all tails in the second half. If we happen to split the coin toss data just along those lines, the first data release would be perfectly compatible with a fair coin. However, upon releasing the second half of the data, it would become clear that something is odd with the coin. But now suppose that we are limited by cosmic variance: after we took all $2N$ data points, the coin is destroyed. Even though we suspect that the coin was not fair (or maybe even switched with an all-tails coin), there is no way for us to verify our suspicion. Had we released the data in four equal chunks, we would have become suspicious after seeing the third chunk (all tails), and our suspicions would be confirmed by the final data release.

In general, the chunks should be neither too small nor too big. Splitting the data into too small chunks will lead to many "detections" of spurious features (i.e. statistical fluctuations) that will disappear in subsequent data releases, leading to much wasted effort by model builders. Releasing chunks that are too large on the other hand severely limits the number of iterations during which one can find and verify new models. In the extreme case of just a single data release, like Planck's cosmic variance limited TT spectrum, there will be no chance of verifying models.

A time ordered strategy is employed by many current experiments, e.g. WMAP. However, instead of releasing data in yearly cycles, we propose to use the concept of doubt [8] to determine the time between data releases. Only when the data in the current data chunk raises doubt about the concordance model to above a predetermined threshold, should this chunk be released. However, some anomalies might still escape detection if the likelihood function is insensitive to them, c. f. the alignment of the low ℓ multipoles or the disappearing two point angular correlation function $C(\theta)$ for large angles [9].

Another method would be to wait until the end of the experiment, and release binned data. The first data release would have the largest bins and the final data release would contain the unbinned data. For example, binning C_ℓs is already standard practice in CMB data analysis (but for different reasons). After releasing bins of size e.g. $\Delta\ell = 10$, a search for anomalies could be performed. If deviations from the expected shape are found, models could be devised, their predictions computed, and finally compared to the next data release with smaller bin size, say $\Delta\ell = 5$. Alternatively, the binning could happen in angular space for $C(\theta)$, with bins $\Delta\theta = 10^\circ, 5^\circ, \ldots$.

Both methods mentioned so far suffer from the problem that subsequent data releases are correlated, which implies issues of using parts of the data twice and biasing the results. However, using principal component analysis (PCA) [10] would guarantee uncorrelated chunks of data. After all data has been collected, the data is split into principal components (i.e. eigenvectors of the covariance matrix), and each PCA is released individually. This would allow for a number of discovery iterations determined by the number of well-constrained PCAs. However, the order of data release is not clear. If the best-constrained PCAs are released first, the chance of finding an anomaly are largest. However it would be very difficult to verify it as subsequent releases would be of poorer quality. Conversely, releasing poorly constrained PCAs first would make it unlikely to find an anomaly. But if one was found, it would be easily confirmed by the well-constrained next data releases. Besides releasing data chunks that raise sufficient doubt, another strategy is to release data with the same information content such as the Kullback-Leibler divergence.

5. Conclusions

We argued that cosmic variance limited data be treated as the precious resource that it is. Instead of releasing all collected data at once in order to just provide better constraints of cosmological parameters, our experimental colleagues should restrict the data releases to amounts just large enough to instill doubt on the concordance model. This would allow model builders to formulate new models to describe unexpected features and test their predictions against the next data release. If all data is released at once, there is no way to test a new model's predictivity.

Acknowledgments

We would like to thank the organizers of the International Workshop on Cosmic Structure and Evolution, Bielefeld, Germany for a wonderful and stimulating workshop. We are grateful to Andrew Jaffe, Louis Lyons, Irit Maor and John Ruhl for stimulating discussions. G.D.S. and P.M.V. are supported by a grant from the DOE to the particle-astrophysics group at CWRU. P.M.V. is supported by the office of the Dean of the College of Arts and Sciences at CWRU. R.T. would like to thank CWRU for hospitality. We thank E. Scrooge for motivation.

References

[1] G. D. Starkman, R. Trotta, and P. M. Vaudrevange, (2009), 0909.2649.

[2] R. Trotta, Mon. Not. Roy. Astron. Soc. **378**, 72 (2007), astro-ph/0504022.

[3] R. Trotta, Contemporary Physics. **49**, 71 (2008), arXiv:0803.4089 [astro-ph].

[4] A. de Oliveira-Costa, M. Tegmark, M. Zaldarriaga, and A. Hamilton, Phys. Rev. **D69**, 063516 (2004), astro-ph/0307282.

[5] D. J. Schwarz, G. D. Starkman, D. Huterer, and C. J. Copi, Phys. Rev. Lett. **93**, 221301 (2004), astro-ph/0403353.

[6] K. Land and J. Magueijo, Phys. Rev. Lett. **95**, 071301 (2005), astro-ph/0502237.

[7] N. J. Cornish, D. N. Spergel, G. D. Starkman, and E. Komatsu, Phys. Rev. Lett. **92**, 201302 (2004), astro-ph/0310233.

[8] G. D. Starkman, R. Trotta, and P. M. Vaudrevange, (2008), 0811.2415.

[9] C. J. Copi, D. Huterer, D. J. Schwarz, and G. D. Starkman, Mon. Not. Roy. Astron. Soc. **367**, 79 (2006), astro-ph/0508047.

[10] D. Huterer and G. Starkman, Phys. Rev. Lett. **90**, 031301 (2003), astro-ph/0207517.