Statistical Challenges in 21st Century Cosmology
Proceedings IAU Symposium No. 306, 2014
A. F. Heavens, J. -L. Starck & A. Krone-Martins, eds.

doi:10.1017/S1743921314011090

Concluding Remarks from a Cosmologist

Andrew H. Jaffe

Astrophysics, Blackett Laboratory
Prince Consort Road
Imperial College London
UK
email: a.jaffe@imperial.ac.uk

Abstract. I conclude the SCCC21 conference highlighting some of the contrasts we heard about, some specific topics (the statistics of random fields), and discuss how some of these play out in the analysis of cosmic microwave background data. I conclude with a hopeful look at the efficacy of blind analyses in the CMB and elsewhere in cosmology.

Keywords. cosmology: cosmological parameters, methods: statistical

1. (False?) Dichotomies

Throughout this meeting, we saw contrasts between different groups of researchers, different philosophies, and different techniques to attack the same problem. Some of these distinctions are real, but we often find that they are not as sharp as we might initially think.

Cosmologists & Statisticians. By their very presence, pretty much everyone here straddles this line, no matter what our formal training might be.

Inference & Algorithms. Perhaps this is a better distinction than the previous: do you come with science questions to answer, looking for the right tool? Or do you start with the tools, and look for applications?

Theorists & Observers. Astronomy has always been a discipline straddling observation and theory. Although there have always been observers who put a lot of theory into the interpretations of their data, astronomy has traditionally been led by data, from the purely observational beginnings of the Hertzsprung-Russell diagram and the Morgan-Keenan-Kellman system of stellar classification, both of which only latterly gained theoretical underpinnings. Nowadays, that tendency has only got stronger, with the rise of a group of theoretically trained astrophysicists who spend much of their time wondering about how to analyze data. They (we) are perhaps the equivalent of the *phenomenologists* of particle physics, charged with making predictions for what will be seen by experimenters.

Bayesians & Frequentists. Historically, the interpretation of probability has been a contentious point. This remains so in the statistics community as well as amongst practicing scientists. There is no need to rehearse the controversies here; rather, it's important to point out that we practicing scientists cannot really afford idealogical purity (see also Section 3) — it is more important to get some sort of answer out even if lacking a principled statistical basis. It is also important to remind ourselves that "Bayesian" and "frequentist" are themselves catch-all descriptions which can hide vast philosophical and practical differences — see, for example, Good (1971) and his "46,656 varieties of Bayesians".

Laggards & Leaders. See astronomer-turned-statistician Ewan Cameron's talk and his contribution to these proceedings, "What we talk about when we talk about fields,"

Cameron (2014) in which he gently berates the cosmology community for lagging behind the statisticians.

Humans & Machines; Supervised & Unsupervised. Many algorithms discussed here require human intervention: we come with particular parameterizations of the phenomena we are studying. But in the coming "big data" era, that may not always be possible: we want the data to teach us what's in it. Of course, these dichotomies are particularly false: any unsupervised algorithms is written by humans, able to make certain kinds of classifications more easily than other.

Real & Idealized. It is heartening to see even the statisticians working with real data, eschewing the caricature of simplifying our real scientific problems to ones that are mathematically tractable but unrecognizable and scientifically uninteresting. (When I first starting learning more formal statistics than I had managed as an undergraduate, I was amused to learn the polysyllabic term "heteroscedastic", so complicated it was clearly seen as a special case — we just call them "errors".)

Statistics & Systematics. If we think we can describe the distribution of some quantity contributing to our data, we call it a "statistical error"; otherwise, we call it a "systematic". As CMB experimentalist Paul Richards from UC Berkeley once said, systematic errors are what can make a good experiment get worse instead of better when you get more data.

Sparse & Dense. The first thing a physicist will do with almost data she can get her hands on is to Fourier transform it. We do this because in many cases the properties of the data are simpler — more sparse — in the harmonic domain. In recent years, the idea of sparsity has taken hold: can we find general methods for finding transformations of the data into a sparse basis?

Blind & Open. At a higher methodological level, cosmologists are beginning to realize that we are not without our own unacknowledged prejudices: sometime we find the answer we are looking for. To combat this, blind analysis, a tool that has a long tradition in the particle physics community, is finding its place in cosmology (see Section 3.1).

2. Random Fields

Although pioneering cosmologists like Peebles (1980) had clearly been thinking of the distribution of matter in the Universe as realizations of a random field, it took the work of Adler (1981) and its subsequent use in a cosmological context by the immense and much-cited (if less often read) Bardeen *et al.* (1986) to bring this formulation into broader use, specifically for the calculation of the distribution of extrema, as proxies for the location of galaxies and clusters as well as voids.

Those works were largely concerned with Gaussian fields — luckily for us, this seems like a good description of the distribution of matter on large scales. We may well ask ourselves why this is the case. There are two somewhat obvious, but not necessarily equivalent, routes. The first is through the *central limit theorem*: if we can describe the distribution of matter by a suitable combination of individual processes, we can expect the result to be well-approximated by a Gaussian distribution (at least far from the tails...). But in fact we know more about the distribution than this: it seems to be an isotropic field, described by an approximately scale-invariant Harrison-Zeldovich-Peebles-Yu power spectrum (Peebles & Yu 1970; Harrison 1970; Zeldovich 1972). To get this, we need more physics. A free (non-interacting) quantum field will have an isotropic Gaussian correlation function, and cosmic inflation is one way to embed such a field in an expanding Universe and get a (nearly) scale-invariant power spectrum (see, e.g., Planck Collaboration XXII 2014, and references therein).

https://doi.org/10.1017/S1743921314011090 Published online by Cambridge University Press

Furthermore, we know that the Gaussian description isn't enough: late-time effects such as nonlinear gravitational evolution and lensing certainly modify the statistical properties of the field, as do any departures from a free quantum field. The former effects have already been observed (Planck Collaboration XXIV 2014); the latter, which would give us a hint about the nature of the interactions experienced by the field, were the subject of much discussion at the conference, and the target of ongoing observations.

The statistical challenge lies in what infinitesimal subset of the ways in which a distribution may depart from an isotropic Gaussian are of cosmological interest. Some of those may be driven by particular theoretical ideas, such as the weak non-Gaussianity induced by the coupling of the inflationary potential (Planck Collaboration XXIV 2014), while others may just be convenient phenomenological descriptions of non-Gaussian distributions (Planck Collaboration XXIII 2014). During this meeting, we heard about many different versions of these:

- simple phenomenological parameterizations, driven by even simpler models ($f_{\rm NL}$);
- going beyond point estimators to various ansätze for the distribution $Pr(f_{\rm NL}|{\rm data})$;
- going beyond single numbers to full n-point polyspectra;
- Minkowski functionals and the beautiful Gaussian Kinematic Formula (Taylor 2006);
- physical reconstruction vs nonlinear transformations and clipping; and
- going beyond the CMB — today: QSOs, tomorrow: Euclid, 21cm, and the coming age of photo-z.

3. Case Study: The CMB

Almost all of these topics show up in the study of the Cosmic Microwave Background. Indeed, several of the first theorists to investigate the details of the distribution of the CMB on the sky (Bond & Efstathiou 1984) went on to pioneer the analysis of CMB data from real experiments (Bond *et al.* 1998; Efstathiou 2006).

A simple flowchart for CMB data analysis is shown in Fig. 1: we start with raw time-ordered data, average these into a map of the sky, evaluate the angular power spectrum, C_ℓ of this map, and determine the set of cosmological parameters responsible for this spectrum. Each step represents significant data compression. For *Planck*, there are trillions of initial samples, tens of millions of pixels in a map, 2500 multipoles in the power spectrum which is finally represented by six parameters (Planck Collaboration XVI 2014). At each step, the intermediate result (i.e., the map or the power spectrum) can be thought of as a *sufficient statistic* — in a Bayesian sense we don't really need to even specify a prior for them as they are just convenient representations of our original data. For the map, we only need to assume that the signal is fixed on the sky (so any time evolution violates the model); for the power spectrum, we need to further assume isotropy and Gaussianity.

Even without further complications, this is a computationally daunting task. If the noise is not white, the time for the mapmaking step (see, e.g., Cantalupo *et al.* 2010) naively scales as the cube of the number of pixels, as does the power spectrum step (Bond *et al.* 1998). Both of these can be simplified by clever numerics (Ashdown *et al.* 2007) or by using frequentist methods (Hivon *et al.* 2002) but these result in approximate or incomplete solutions. For mapmaking, it is impossible to calculate the full noise covariance of the resulting Gaussian distribution, while for power spectra even the form of the distribution is not well determined.

And even this picture is quite a bit simpler than what is needed in practice. A raft of "systematic effects" need to be accounted for, usually only in an approximate or parameterized way. The most important is probably the effect of astrophysical foregrounds which contaminate the primordial signal. Much of this can be taken care of by

https://doi.org/10.1017/S1743921314011090 Published online by Cambridge University Press

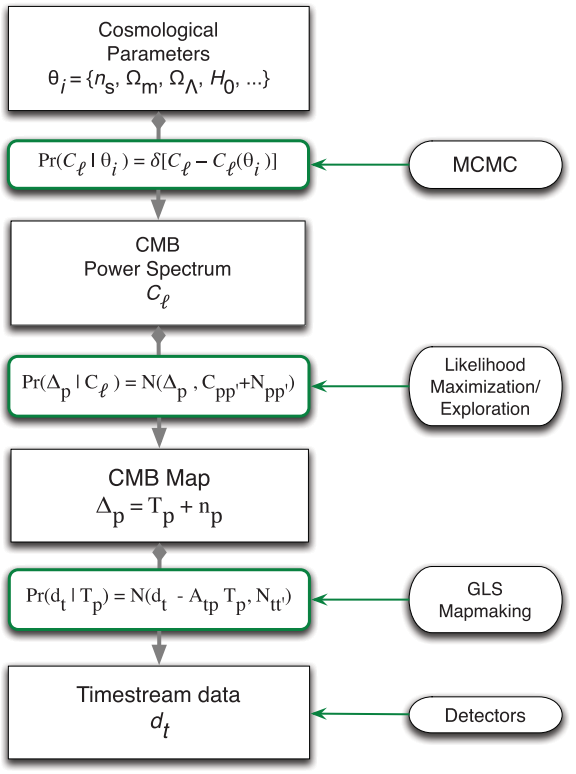

Figure 1. The flow of data in the analysis of CMB experiments — the simplified version.

conservative masking of the sky but the low-level remnants of foreground emission remain throughout the sky and must be accounted for (e.g., Planck Collaboration XII 2014). Similarly, the data-flow described above assumes axisymmetric beam-shapes (i.e., the point-spread function in the usual astronomical parlance) and dense, uniform sampling of the sky in the time domain. Except in special cases (Armitage-Caplan & Wandelt 2004), this cannot be accounted for exactly.

In practice, then, we cannot afford idealogical purity — the cosmological community is largely Bayesian in its outlook, but in fact many of our algorithms are essentially frequentist, assuming that the asymptotic correspondence between Bayesian and frequentist results holds at least at small scales (high ℓ). We further remove at least some systematic effects by measuring some appropriate set of instrumental parameters and fixing their values, or in some cases use forward Monte-Carlo modeling to approximately parameterize the posterior distribution (e.g., Planck Collaboration VII 2014).

3.1. *Blind Analyses*

We never have enough information to give a complete statistical description of a realistic experiment: there are *always* systematic effects for which we have not accounted. When do we stop looking? At this meeting, in talks from Hiranya Peiris and others, we were admonished to do something other than just stopping when we thought we had the right answer. One way to ensure this, used very often in the particle physics community, is to use a blind analysis: perfect the tools on simulations and on subsets or combinations of the data which are expected to have none of the signal being sought after.

https://doi.org/10.1017/S1743921314011090 Published online by Cambridge University Press

We attempted such an analysis for the recent publications from Polarbear (The Polarbear Collaboration 2014, 2013), an experiment measuring the polarization of the CMB, in particular searching for the so-called B-mode pattern which results from a background of inflationary gravitational waves and/or lensing of polarized CMB photons along the line of sight. The Polarbear Collaboration (2014) gives the first direct detection of the B-mode power spectrum due to the latter lensing effect, analyzed using largely blind techniques. These are especially useful during the initial discovery phase during which it would be very easy to confuse an instrumental or foreground systematic effect for the sought-after signal.

References

Adler, R. J. 1981, The Geometry of Random Fields (Society for Industrial and Applied Mathematics (SIAM, 3600 Market Street, Floor 6, Philadelphia, PA 19104))
Armitage-Caplan, C. & Wandelt, B. D. 2004, *PRD*, 70, 123007
Ashdown, M. A. J., Balbi, A., Bartlett, J. G., *et al.* 2007, *A&A*, 467, 761
Bardeen, J. M., Bond, J. R., Kaiser, N., & Szalay, A. S. 1986
Bond, J. R. & Efstathiou, G. 1984, 285, L45
Bond, J. R., Jaffe, A. H., & Knox, L. E. 1998, *PRD*, 57, 2117
Cameron, E. 2014, arXiv:1406.6371v1, 6371
Cantalupo, C. M., Borrill, J. D., Jaffe, A. H., Kisner, T. S., & Stompor, R. 2010, *ApJ*, 187, 212
Efstathiou, G. 2006, *MNRAS*,370, 343
Good, I. J. 1971, *American Statistician*, 25, 62
Harrison, E. 1970, *PRD*, 1, 2726
Hivon, E., Górski, K. M., Netterfield, C. B., *et al.* 2002, *ApJ*, 567, 2
Peebles, P. J. E. 1980, The large-scale structure of the universe
Peebles, P. J. E. & Yu, J. T. 1970, *ApJ*, 162, 815
Planck Collaboration VII. 2014, *A&A*, in press, arXiv:1303.5068
Planck Collaboration XII. 2014, *A&A*, in press, arXiv:1303.5072
Planck Collaboration XVI. 2014, *A&A*, in press, arXiv:1303.5076
Planck Collaboration XXII. 2014, *A&A*, in press, arXiv:1303.5082
Planck Collaboration XXIII. 2014, *A&A*, in press, arXiv:1303.5083
Planck Collaboration XXIV. 2014, *A&A*, in press, arXiv:1303.5084
Taylor, J. 2006, *The Annals of Probability*, 34, 122
The Polarbear Collaboration. 2013, arXiv:1312.6645
The Polarbear Collaboration. 2014, arxiv:1403.2369
Zeldovich, Y. B. 1972, *MNRAS*,160, 1P