

STATISTICS IN ASTROPHYSICS AND COSMOLOGY: PHYSTAT05

ANDREW H. JAFFE

Blackett Laboratory, Imperial College London

E-mail: a.jaffe@imperial.ac.uk

In this conference summary on Astrophysics talks at PhyStat05, I will discuss the various philosophical and pragmatic approaches to problems of statistics in astrophysics (and cosmology in particular), and their application to a few modern problems discussed at this meeting. In particular, I will develop a Bayesian formalism for the analysis of data from Cosmic Microwave Background (CMB) experiments.

In his PhyStat talk on the Transit of Venus, Johnston said “Precision astronomy depends on an individual’s judgment” and that remains as true today as in the Enlightenment.

1. Philosophy

One remarkable difference between the members of the astrophysics and particle physics communities present at this meeting was the relative prevalence of Bayesian and Frequentist methods in the two communities. In particle physics, the prevailing methods are strictly frequentist, while astrophysics (and especially cosmology) has become increasingly Bayesian in its outlook in recent years. This philosophical distinction is grounded in the very different practical realities of the two fields: particle physicists can usually run their experiments for longer and longer, “building up statistics”, and making the underlying asymptotic assumptions of a frequentist approach more valid. In cosmology, on the other hand, “there is only one Universe” and there are some experiments that can never be re-run. Moreover, as we will see below in the discussion of CMB data analysis, many of the predictions of cosmological theory are inherently statistical, so we must infer the properties of a correlated multivariate probability distribution from a *single* realization.

2. Case Study: the Cosmic Microwave Background

The Cosmic Microwave Background provides perhaps a rich example of data analysis in a cosmological setting. (Let me emphasize several points at the outset. First, this is a personal view of the CMB data analysis process. Second, this has become quite a large sub-field of cosmology, and I have been quite

spare in my use of references, for which I apologize to my many colleagues whose work has not been cited herein despite important contributions to the field.)

The raw data – voltages output by some sort of antenna or temperature sensor – bear no simple relationship to the ultimate parameters to be measured, cosmological parameters such as the curvature of the Universe and the spectrum of primordial perturbations. Moreover, those raw voltage data are dominated by the noise properties of the measuring instrument. Yet somehow we must find an algorithm to “radically compress”² the millions or billions or more of raw data to just a few cosmological parameters. We start by writing down a simple model for the data from a single detector (the generalization to multiple detectors is straightforward; the data can just be appended as a single very long vector):

$$d_t = A_{tp}s_p + n_t = As + n \quad (1)$$

where d_t is the data taken at time t , s_p is the signal in pixel p (i.e., the CMB *map*), n_t is the noise, and the “pointing matrix”, A_{tp} gives the response of the instrument at time t to pixel $p = 1 \dots \#_p$. Note that *pixel* here refers to a finite area of sky. For simplicity, we can assume that the signal already contains the action of the experimental beam and any pixelization scheme we impose on the sky, in which case $A_{tp} = 1$ when pixel p is being observed at time t , and 0 otherwise. Finally, in the first equality we assume the Einstein summation convention, and in the second use matrix notation, so $As = A_{tp}s_p \equiv \sum_p A_{tp}s_p$. In general there will also be terms representing various other effects that may be present in the data, such as foreground contamination, instrumental systematics, etc. By a suitable generalization of the pixel domain and the pointing matrix, we can in fact estimate (and marginalize over) such effects.¹⁹

To proceed further, we need a model for the noise. We will assume that it can be represented by a stationary zero-mean Gaussian process, with correlations given by

$$\langle n_t n_{t'} \rangle \equiv N_{T,tt'} = N_T(t - t'). \quad (2)$$

More generally, we may subdivide the timestream into individual “stationary periods” within which this equation holds, and between which we assume zero correlation. This assignment is *conservative*, at least in the sense that the Gaussian is the maximum-entropy distribution with a given correlation structure. (This fact will have further implications later on when we discuss the correlations of the underlying signal, that is, the power spectrum, C_ℓ .)

2.1. Mapmaking

The first step, then, is to estimate the map, s_p given Eq. 1. This is a fairly standard inverse problem, but we choose to address it from a Bayesian standpoint. For these purposes, Bayes’ theorem states

$$P(s_p|d_t I) \propto P(s_p|I) \times P(d_t|s_p I), \quad (3)$$

where the left hand side is the *posterior* probability, the first factor on the right is the *prior* probability for the signal, and the final factor is the *likelihood*, the probability of the data given the signal. We write all probabilities as conditional upon some background information, I ; in this case I encodes our knowledge of the noise correlation function, $N(t)$, the fact that we are imposing a Gaussian distribution upon the noise, etc. With this setup, the likelihood is just a multivariate Gaussian:

$$P(d_t|s_p I) = \frac{1}{|2\pi N|^{1/2}} \exp -\frac{1}{2} (d - As)^\dagger N_T^{-1} (d - As), \quad (4)$$

where the superscript \dagger means matrix transpose. Finally, we impose a uniform (albeit improper) prior on the signal $P(s_p|I) \propto \text{const.}$ As we shall see, this prior is actually irrelevant to the ultimate determination of power spectra and cosmological parameters.

By completing the square in the exponential (or taking derivatives, etc.) we see that the likelihood (and the posterior with our constant prior) is proportional to a Gaussian distribution in $\bar{s}_p = s_p + n_p$ with

$$\bar{s}_p = (A^\dagger N_T^{-1} A)^{-1} A^\dagger N_T^{-1} d \quad (5)$$

(the overbar denotes a generalization of the mean over all observations of a single pixel for correlated noise) and variance

$$\langle n_p n_{p'} \rangle = N_{P,pp'} = (A^\dagger N_T^{-1} A)_{pp'}^{-1} \quad (6)$$

which is just the usual Generalized Least Squares (GLS) solution.

[In fact with complex data like that expected from the Planck Surveyor, we cannot always calculate the full Bayesian map, Eq. 5 because of the complicated matrix manipulations involved. However, even in the case of some more general approximation to the map, we can still calculate its full noise correlation structure, replacing Eq. 6, as long as the operations are linear in the data and unbiased — a word not usually associated with Bayesian methods! — with respect to the signal.]

The output of this procedure is represented by the quantities \bar{s}_p and $N_{P,pp'}$, our estimate of the map and its noise correlation structure. Specifically, \bar{s}_p is an estimate of the beam-smoothed and pixelized sky in the pixels labelled by p . We take the beam to be circularly symmetric, with spherical harmonic transform, B_ℓ .

2.2. Power spectrum estimation

Next we must estimate the power spectrum which, by hypothesis, is responsible for realizing the map. Conventionally, we assume a zero-mean Gaussian process with covariance given by

$$\langle s_p s_{p'} \rangle = S_{P,pp'}(C_\ell) = \sum_\ell \frac{2\ell + 1}{4\pi} C_\ell B_\ell^2 P_\ell(\hat{x}_p \cdot \hat{x}_{p'}), \quad (7)$$

where C_ℓ is the cosmological power spectrum, $\hat{x}_p \cdot \hat{x}_{p'}$ is the cosine of the angular distance between the pixels p and p' , and the P_ℓ are the Legendre polynomials. Now, the parameter we wish to estimate is C_ℓ ; we can use the posterior of the previous step as the effective likelihood for the signal, so the model for the data, now just the map \bar{s}_p , is simply

$$\bar{s}_p = s_p + n_p, \quad (8)$$

where pixel noise correlations are given by Eq. 6. Alternately, we can start with the full likelihood, Eq. 4, and jointly estimate s_p and C_ℓ , with prior

$$\begin{aligned} P(s_p, C_\ell|I) &= P(C_\ell|I)P(s_p|C_\ell I) \\ &= P(C_\ell|I) \frac{1}{|2\pi S|^{1/2}} \exp -\frac{1}{2} s^\dagger S^{-1} s \end{aligned} \quad (9)$$

We can then marginalize over s_p giving us the posterior for C_ℓ alone. It turns out that these approaches are mathematically equivalent, showing that indeed the mean and variance of Eqns. 5–6 are *sufficient statistics* for any further calculations. At this point, then, we have the following likelihood function:

$$\begin{aligned} P(d_\ell|C_\ell I) &= P(\bar{s}_p|C_\ell I) \\ &= \frac{\exp -\frac{1}{2}\bar{s}^\dagger(S_P + N_P)^{-1}\bar{s}}{|2\pi(S_P + N_P)|^{1/2}} \end{aligned} \quad (10)$$

where now the parameter of interest, C_ℓ , appears in the covariance matrix, $S_P(C_\ell) + N_P$. Because of this, there is no simple analytic description of the posterior probability, or indeed for the shape of the likelihood considered as a function of C_ℓ . However, we can relatively easily use techniques like Newton-Raphson iteration to find the peak of the likelihood and calculate its curvature about the maximum¹.

Unfortunately, these techniques are prohibitively expensive for data from upcoming experiments such as the Planck Surveyor, scaling as $O(\#_p^3)$ in time and $O(\#_p^2)$ in storage; indeed the latter implies that, for coming megapixel experiments, the covariance matrix is likely too large to store, much less calculate in full generality. There have also been efforts to develop so-called Gibbs Sampler Monte Carlo techniques to calculate the full posterior for the power spectrum.⁶

Thus, we must be practical. Even if we are philosophically disposed to Bayesianism (as are many in the cosmology community), we may need to consider other techniques, although I, in particular, take the rather unorthodox view that these methods are useful as approximations to the Bayesian result. This stands in contrast to many of the Bayesian approaches in particle physics discussed at this meeting, in which the analysts try to find Bayesian techniques which give the same answer as the orthodox frequentist techniques already in use.

The most common of these techniques for estimating C_ℓ are the so-called unbiased pseudo- C_ℓ quadratic estimators, in which some approximation to the spherical harmonic transform of the full sky is calculated, and squared to give the ‘‘pseudo- C_ℓ ’’ spectrum, \hat{C}_ℓ , which is then corrected to give an unbiased estimate of the true spectrum by inverting the relation

$$\langle \hat{C}_\ell \rangle = \sum_{\ell'} M_{\ell\ell'} C_{\ell'} + N_\ell \quad (11)$$

where the ensemble average is taken over Gaussian realizations of the signal and noise with the variances given above. We know that in the limit of a full sky and uniform noise the estimator thus derived is exactly the same as the Bayesian maximum likelihood, with the variance the same as the curvature about the maximum, and indeed the usual relations from the theory of probability and statistics state that these hold ‘‘asymptotically’’, which is usually understood to mean that they will hold for high ℓ , where very many modes contribute to the measurement. (We do know from experience that the results do differ in detail for realistic experiments, such as BOOMERANG¹⁸.)

2.3. Cosmological Parameter Estimation

Finally, we must use these C_ℓ to determine the underlying parameters, θ_i , (e.g., the densities, Ω_i ; the Hubble Constant, H_0 , etc.). Unlike in previous steps, there is a direct relationship between the parameters and the power spectrum, simply $C_\ell = C_\ell(\theta)$, i.e., we have a delta-function prior $P(C_\ell, \theta_i) = P(\theta_i)\delta[C_\ell - C_\ell(\theta)]$. So the likelihood function remains as before, but we wish to determine its parameters as we vary θ . The calculation of $C_\ell(\theta)$ for standard models requires the solution of coupled Einstein-Boltzmann linear differential equations describing the distribution of matter and radiation in the expanding universe, and has been implemented in publicly available codes such as CMBFAST²⁰ and CAMB¹⁴. For realistic models, this is straightforward but relatively time-consuming, and moreover the general exploration of a multi-parameter space is a difficult task. In recent years, the favored technique for this exploration has been Markov Chain Monte Carlo (MCMC).^{3, 15}

In this meeting, MCMC in a cosmological context was discussed in the talks of Leach, Nicholls and Trotta (this volume).

A very different way of exploring CMB power spectra was discussed by Nichol at this meeting: he used a (very non-Bayesian!) non-parametric smoothing technique to examine the overall shape of the spectrum, and answer some very basic questions: does the data in Figure 1 show a series of definite peaks? He then extended these methods to the Cosmological Parameter Estimation problem itself: do the parameters predict the correct overall smoothed

shape? Qualitatively his results agree largely with the consensus discussed in the following, although there are detailed — and not unexpected — differences.

3. Results: Cosmology in 2005

In Figure 1 we show the results of various calculations of C_ℓ along these lines, each from a different dataset; the results are in quite good agreement overall. In general, these results are either the frequentist mean and variance, or the maximum likelihood and curvature; as emphasized above, the prior does not really matter at this point.

Where the prior does enter, however, is in the calculation of the cosmological parameters from these spectra. As was emphasized at this meeting by Cox and Le Diberder, flat priors are dangerous. Indeed, in cosmology, there is no one set of natural parameters on which to impose flat priors. For example, would we want a flat prior on the density relative to the critical density, Ω ? However, that critical density itself depends on the *a priori* unknown Hubble Constant, $H_0 \equiv 100h$ km/s/Mpc, so perhaps a more physical quantity would be Ωh^2 ? There is no “correct” answer to this question; rather we take the advice of Cousins at this meeting: we must perform sensitivity analyses to determine the effect of our priors upon the analysis. In effect, by comparing the work of different authors, we can do just this sort of meta-analysis. If we estimate the cosmological parameters from the data of Figure 1 (from WMAP¹⁰, ACBAR¹¹, BOOMERANG (B03)^{9, 16}, CBI¹⁷, DASI⁷, MAXIMA¹³ and VSA⁵) in various combinations, we see that many features are robust to these changes, and we can highlight a few here:

- The Universe is flat: $\Omega_{\text{tot}} = 1$;
- The primordial perturbations are well described by a nearly scale-invariant power spectrum ($n_s \simeq 1$); and
- The Hubble Constant is approximately $H_0 = 72$ km/s/Mpc.

Perhaps startlingly, the first two are just the predictions of the inflationary theory of the early Universe! So finally, the discussion of statistics leads us, as it should, to the underlying physics.

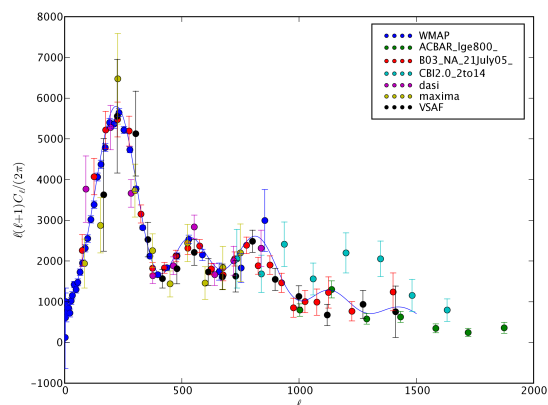


Fig. 1. Recent CMB Power spectrum data; publications are cited in the text.

4. Complications: non-Gaussianity

The model for CMB data in the previous section becomes more restrictive in each step of the process. When making the map, we first assume a model for the *noise*, that it is stationary over some period of time, and a Gaussian of the noise power spectrum. When calculating the power spectrum, we assume that the signal is isotropic on the sky, and a Gaussian realization from the cosmological power spectrum, C_ℓ . Finally, we assume the cosmological parameters directly determine the power spectrum. Much effort has been put into going beyond these assumptions, specifically the Gaussianity of the signal. As the saying goes, “non-Gaussian distributions” are like “non-elephant animals”, and it is very hard to describe an arbitrary distribution without just giving all possible detailed information (e.g., the functional form of the distribution or its moments). Methods must be tuned to find specific “sorts” of non-Gaussianity, such as the existence of higher-order connected moments of the data. Even there, of course, the task immediately becomes very difficult: are the higher moments best described in Fourier (spherical harmonic) space, or spatially on the sky? Which moments are we searching for? What counts as a detection (if you look hard enough, you will always find something!)? For a Bayesian, these problems seem even worse: how do you even write down the distribution in the absence of a very concrete and calculable model?

With the actual WMAP data, we also may be encountering a related problem: There is some ev-

idence that, on the largest scales, the WMAP data do not seem to obey statistical anisotropy. That is, statistical quantities may not just be functions of the distance between points, but may depend on where you are on the sky, as has been discussed in a number of recent works^{4, 8, 12} (and others). This could be evidence of foreground contamination, or, more excitingly, of some underlying misunderstanding of the physics of the universe on large scales. In the absence of a physical model, it is *prima facie* impossible to distinguish a non-Gaussian distribution from an anisotropic distribution or from a combination of the two effects.

Nonetheless, various techniques have been proposed and applied to tease out non-Gaussianities from current data. At this meeting, they were discussed by Jin and Starck (and by Digel and by Bissantz in more traditional astronomical contexts). But a word of caution is in order: most of these methods are derived assuming some sort of independent and identically distributed (iid) random variable is responsible for the non-Gaussianity, but in real-world astrophysics, nothing is ever iid!

5. Other problems

I have concentrated on my speciality, cosmology in general and the CMB in particular, but of course statistics plays a paramount role throughout astrophysics. Indeed, as emphasized in the presentations of Cox and Johnston, astronomy has played a leading role in the development of statistics since the beginning. Other exciting developments discussed at the Conference include:

- Time-series analysis [Clifford];
- Image processing/reconstruction/restoration [Titterton]. For some applications, it is crucial to be able get full error information (i.e., the posterior distribution) of the reconstructed image, and this restricts the possible algorithms;
- Classification problems [e.g., Gray]: finding unusual objects (or usual ones: photometric redshift; galaxy classification from pictures). Note that these tasks usually have vastly different kinds of prior information: physics vs. training sets vs. “experience”.

5.1. Virtual Observatories

As discussed at this meeting by Alex Szalay, we in cosmology and astrophysics are beginning to deal with the massive, heterogeneous datasets covering a variety of instruments, wavebands, areas of the sky, etc. The community is attempting to build tools for uniform and distributed access to and analysis of these data under the rubric of Virtual Observatories. Bob Nichol discussed searching through massive astronomical datasets using KD-trees and the plans to finally move large-scale astronomical data-processing from the desktop to the grid.

6. Conclusions

In his opening talk at this meeting, Sir David Cox said that “We’re eclectic”; this perfectly captures the pragmatism of astronomers and astrophysicists confronting our data, the need to find tools to handle its complexity and volume. Indeed, astronomers have always had to deal with data just beyond the ability of obvious current techniques, and therefore have always been avid consumers — if not developers — of cutting-edge statistical techniques. As we saw throughout this meeting, this fruitful confluence of fields continues to this day.

Acknowledgments

I would like to thank Louis Lyons and the other Phystat05 conference organizers for this great opportunity, and the many wonderful speakers at the meeting on whose work I based this review. I’d like also to thank my collaborators in the MAXIMA, BOOMERANG and COMBAT collaborations who participated in the aspects of my own work that I have discussed, as well as PPARC for their financial support.

References

1. J. R. Bond, A. H. Jaffe, and L. Knox. Estimating the power spectrum of the cosmic microwave background. *Phys. Rev. D*, 57:2117–2137, 1998.
2. J. R. Bond, A. H. Jaffe, and L. Knox. Radical Compression of Cosmic Microwave Background Data. *Astrophys. J.*, 533:19–37, 2000.
3. N. Christensen, R. Meyer, L. Knox, and B. Luey. Bayesian methods for cosmological parameter estimation from cosmic microwave background measurements. *Classical and Quantum Gravity*, 18:2677–2688, July 2001.

4. A. de Oliveira-Costa, M. Tegmark, M. Zaldarriaga, and A. Hamilton. Significance of the largest scale CMB fluctuations in WMAP. *Phys. Rev. D*, 69(6):063516–, Mar. 2004.
5. C. Dickinson et al. High sensitivity measurements of the CMB power spectrum with the extended very small array. astro-ph/0205436.
6. H. K. Eriksen et al. Power Spectrum Estimation from High-Resolution Maps by Gibbs Sampling. *Astrophys. J. Suppl.*, 155:227–241, Dec. 2004.
7. N. W. Halverson et al. Degree Angular Scale Interferometer First Results: A Measurement of the Cosmic Microwave Background Angular Power Spectrum. *Astrophys. J.*, 568:38–45, 2002.
8. F. K. Hansen, A. J. Banday, and K. M. Górski. Testing the cosmological principle of isotropy: local power-spectrum estimates of the WMAP data. *Mon. Not. R. Astr. Soc.*, 354:641–665, Nov. 2004.
9. W. C. Jones et al. A measurement of the angular power spectrum of the CMB temperature anisotropy from the 2003 flight of BOOMERANG. astro-ph/0507494. 2005.
10. A. Kogut et al. Wilkinson Microwave Anisotropy Probe (WMAP) first year observations: TE polarization. *Astrophys. J. Suppl.*, 148:161, 2003.
11. C.-I. Kuo et al. High resolution observations of the CMB power spectrum with ACBAR. *Astrophys. J.*, 600:32–51, 2004.
12. K. Land and J. Magueijo. Examination of Evidence for a Preferred Axis in the Cosmic Radiation Anisotropy. *Physical Review Letters*, 95(7):071301–, Aug. 2005.
13. A. T. Lee et al. A High Spatial Resolution Analysis of the MAXIMA-1 Cosmic Microwave Background Anisotropy Data. *Astrophys. J. Lett.*, 561:L1–L5, 2001.
14. A. Lewis. <http://camb.info/>.
15. A. Lewis and S. Bridle. Cosmological parameters from VSA, CBI and other data: a Monte-Carlo approach. *Phys. Rev. D*, 66:103511, 2002.
16. F. Piacentini et al. A measurement of the polarization-temperature angular cross power spectrum of the cosmic microwave background from the 2003 flight of BOOMERANG. astro-ph/0507507.
17. A. C. S. Readhead et al. Extended mosaic observations with the cosmic background imager. *Astrophys. J.*, 609:498–512, 2004.
18. J. E. Ruhl et al. Improved Measurement of the Angular Power Spectrum of Temperature Anisotropy in the Cosmic Microwave Background from Two New Analyses of BOOMERANG Observations. *Astrophys. J.*, 599:786–805, 2003.
19. R. Stompor et al. Making maps of the cosmic microwave background: The MAXIMA example. *Phys. Rev. D*, 65(2):022003–, 2002.
20. M. Zaldarriaga and U. Seljak. CMBFAST for Spatially Closed Universes. *Astrophys. J. Suppl.*, 129:431–434, 2000.