Statistical Searches in Astrophysics and Cosmology

Ofer Lahav
University College London, UK

Abstract

We illustrate some statistical challenges in Astrophysics and Cosmology, in particular noting the application of Bayesian methods and model selection criteria. We describe two examples where Bayesian methods have improved our inference: (i) photometric redshift estimation and (ii) orbital parameters of extra-solar planets. While sub-communities in Astrophysics, High Energy Physics and Statistics develop separately their specific techniques, it is beneficial to 'compare notes' and to exchange methods.

1 Introduction

The dramatic increase of data in Astronomy has renewed interest in the principles and applications of statistical inference methods. These methods can be viewed as a bridge between the data and the models. Common statistical problems in Astronomy fall broadly into the following tasks:

- Data compression (e.g. of galaxy images or spectra).
- Classification (e.g. of stars, galaxies or Gamma Ray Bursts).
- Reconstruction (e.g. of blurred galaxy images or mass distribution from gravitational lensing).
- Feature extraction (e.g. signatures feature of stars, galaxies or quasars).
- Parameter estimation (e.g. orbital parameters of extra-solar planets or cosmological parameters).
- Model selection (e.g. Are there 0,1,2,... planets around a star? Is a cosmological model with non-zero neutrino mass more favourable?).

It is possible for these tasks to be related. For example, estimation of cosmological parameters from the Cosmic Microwave Background (CMB) or galaxy redshift surveys are commonly deduced from a compressed information, usually in the form of the angular and 3D power spectra, respectively. A further example is classification of galaxy spectra. It can be achieved in a compressed space of the spectra, or in the space of astrophysical parameters estimated from the spectra.

The Astro-statistics community is fortunate to have these days excellent textbooks, among them (in chronological order): Lyons (1986), Lupton (1993), Babu & Feigelson (1996), Sivia (1996), Cowan (1998), Starck & Murtagh (2002), Martinez & Saar (2002), Press et al. (1992), Wall & Jenkins (2003), Saha (2003) and Gregory (2005). Useful reviews on Bayesian methods in Cosmology can be found in the book edited by Hobson et al. (2009) and in Trotta (2008).

2 Inference Methods

There is an ongoing debate between the 'Frequentist' approach and the 'Bayesian' methodology. The 'Frequentist' approach interprets probability as the frequency of the outcome of a repeatable experiment. In contrast, the 'Bayesian' methodology (first published in 1764) views the interpretation of probability more generally and it includes a degree of belief, formulated as:

$$P(model|data) = P(data|model)P(model)/P(data),$$

where on the right hand side the first term is the *likelihood*, the second is the *prior* and the third is the *evidence*.

In the Bayesian approach the choice of *priors* may strongly affect the inference. However it is an 'honest' approach in the sense that all the assumptions are explicitly spelled out in a logical manner.

2.1 Sources of Systematics

A major part of research in Astronomy is devoted to the effect of systematic errors. Consider the example of estimating a specific parameter, e.g. the Dark Energy equation of state parameter w from Baryon Acoustic Oscillations observed in galaxy clustering (e.g. Eisenstein et al. 2005; for review of Cosmological parameters see e.g. Lahav & Liddle 2010). We can distinguish three types of systematics:

- Cosmological uncertainty (due to the assumptions on the other $N - 1$ cosmological parameters' associated priors).
- Astrophysical uncertainty (e.g. what is the relation between the clustering of luminous galaxies and the matter fluctuations?).
- Observational uncertainty (e.g. selection effects in the galaxy sample).

Each of these contributes to the error budget of w in a different way and should be incorporated in the statistical analysis accordingly.

2.2 Justifying Priors

The choice of prior is crucial in the Bayesian framework, yet the justification of each prior is not always spelled out in research articles. To give an example, a prior on the curvature of the universe can be justified in a number of ways, some theoretical, some empirical:

- Theoretical prejudice (e.g. 'according to Inflation, the universe must be flat').
- Previous observations (e.g. 'we know from the CMB WMAP experiment the universe is flat to within 2%, under the assumption of other priors').
- Parameterized ignorance (e.g. 'a uniform prior' or 'a Jeffreys prior').

2.3 Recent Trends in Astro-statistics

Trends noted in recent conferences incude the following:

- Astro-statistics has become a 'respectable' discipline of its own.
- 'Bayesian' approaches are more commonly used, and in better co-existence with 'Frequentist' methods.
- There is more awareness of model selection methods, e.g. the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC), see e.g. Liddle et al. (2006).
- Computer intensive methods, e.g. Markov Chain Monte Carlo (MCMC) are more popular.
- Free software packages are more widely used.

It is beyond the scope of this short review to cover every topic. I shall focus on two examples; photometric redshifts and extrasolar planets. Both cases illustrate how Bayesian approaches have improved our inference on the science questions of interest.

3 An Example From Cosmology: Photometric Redshifts

Mapping the galaxy distribution in 3D requires the galaxy redshifts. In the absence of spectroscopic data, redshifts of galaxies may be estimated using multi-band photometry, which may be thought of as very low-resolution spectroscopy. While the redshift error per galaxy is relatively large, having a great number of galaxies could reduce the errors on measures of the galaxy clustering. Photometric surveys over large areas of the sky may compete well with spectroscopic surveys. Photo-z methods proved very useful e.g. in recent analyses of the GOODS, COMBO-17 and SDSS Luminous Red Galaxies. Several wide-field photometric redshifts are planned, e.g. the Dark Energy Survey, PanSTARRS, LSST, Euclid and WFIRST. Understanding the photometric redshift errors is crucial for quantifying the errors on e.g. the Dark Energy equation of state parameter w from galaxy clustering and weak lensing.

In more detail, photometric redshift methods rely on measuring the signal in the photometric data arising from prominent "break" features present in galaxy spectra e.g. the 4000 A break in red, early-type galaxies, or the Lyman break at 912 A in blue, star-forming galaxies. There are two basic approaches to measuring a galaxy photometric redshift z (e.g. Csabai et al. 2003 and references therein). The first, template matching, relies on fitting model galaxy spectral energy distributions (SEDs) to the photometric data, where the models span a range of expected galaxy redshifts and spectral types. This is done via a simple χ^2 statistic, i.e. via the likelihood $P(colours|z)$, but it may lead to catastrophic errors. Benitez (2000) generalized the method by incorporating Bayesian priors. The prior $P(z|magnitude)$ for the redshift of a galaxy given its magnitude (apparent luminosity) then multiplies the likelihood to give the posterior

$$P(z|colours, magnitude) \propto P(colours|z, magnitudes) \times P(z|magnitude) \; .$$

This Bayesian chain, which can also be generalized to include galaxy type, greatly reduces the number of outliers.

Another approach utlilises an existing spectroscopic redshift sample as a training set to derive an empirical photometric redshift fitting relation. An example of a training-based method, ANNz, which is also Bayesian, utilizes Artificial Neural Networks (Collister & Lahav 2004). When applied to SDSS galaxies the rms error using ANNz is $\sigma_z = 0.02$, compared with $\sigma_z = 0.07$ using a template method.

4 An Example From Extra-solar Planets: Orbital Parameter Estimation

Astronomers have faced a growing number of free parameters in modelling astrophysical systems, for example cosmological parameters or extra-solar planet orbital parameters. In the case of a model with N free parameters marginalizing over N-1 parameters, it proves to be computationally expensive if the parameter space is mapped into a grid. An alternative method, the Markov Chain Monte Carlo (MCMC), has been known since the 1950's and a wide range of methods exists in the literature to implement it, e.g. the Metropolis-Hasting algorithm.

The key idea is to turn a probability distribution function in N dimensions into a cloud of points which represents the probability distribution function. The probability distribution function could incorporate the probabilities for the priors, in the Bayesian spirit. The MCMC algorithm constructs a random walk in the model parameter space with steps drawn from a multi-dimensional proposal distribution (e.g. a Gaussian). It is crucial to apply tests for convergence, i.e. to ensure that the parameter space is properly sampled, in particular if there are several peaks in a high dimensional space.

MCMC algorithms have been applied widely to parameter estimation from the CMB and other cosmological data sets (e.g. Lewis & Bridle 2003; Verde et al. 2003) and to both detecting and characterizing orbits of extrasolar planets (e.g. Gregory 2005; Ford 2005; Balan & Lahav 2009).

Nearly 2000 extrasolar planets have been discovered so far. Most of those were discovered using measurements of the radial velocity of the host star. The radial velocity curve can be modelled by approximately a dozen parameters, depending on the complexity of the assumed model. It is also important to allow for more than one planet around the star, hence for more free parameters. This leads to the challenging problem of non-linear minimization in a high dimensional parameter space. Deriving these parameters accurately is very important as this can then influence the interpretation for an individual object, as well as the statistics of orbital parameters for an ensemble of extra-solar planets.

For example, in many of the discovery papers the approach taken is to estimate first the period P and then for that fixed P to solve later for the orbital parameters. As there is degeneracy of parameters and dependence on their priors this could lead to the wrong value of P. This was pointed out by Gregory (2005), who developed an MCMC Bayesian approach to cope with the multi-parameter estimation. He illustrated the method for the data for HD73526, where he found three possible solutions for P. In fact the previously reported one turned out to be the least probable orbit (but apparently the data for this system somewhat changed since the publication of the paper).

5 Future Work in Astro-statistics

The following topics represent current and further work in Astro-statistics:

- Model selection methodology (e.g. which criteria and the role of priors).
- MCMC machinery and extensions (e.g. nested sampling).
- Detection of non-Gaussianity and shape finders (e.g. for galaxy survey and CMB maps).
- Blind de-convolution (e.g. for recovering galaxy shapes from blurred images).
- Object classification (e.g. stars, galaxies and quasars).
- Comparing simulations with data (e.g. large galaxy surveys with N-body and hydrodynamic simulations).
- Visualization (of e.g. 3D galaxy surveys or multi-parameter space).
- Virtual Observatories (including both Real Data and Mock data).

Astronomy, High Energy Physics and Statistics independent communities and meetings like this provide great opportunities to 'compare notes' and exchange ideas. Fundamental issues in statistical inference from data will not go away. With the exponential growth of data in Astronomy there is a great need for further interaction of astronomers with experts in other fields.

Acknowledgements

I would like to thank my collaborators for numerous dicussions over the years on the topics mentioned briefly here, and the organizers for a stimulating meeting.

References

[1] Babu, G.J., Feigelson, E.D., 1996, *Astrostatistics*, Chapman & Hall

[2] Balan, S. & Lahav, O., 2009, MNRAS, 394, 1936

[3] Benitez, N., 2000, ApJ, 536, 571

[4] Collister, A. & Lahav, O., 2004, PASP 116, 345

[5] Cowan, G.,1998 *Statistical Data Analysis*, Oxford University Press

[6] Csabai, I., et al. 2003, AJ, 125, 580

[7] Eisenstein, D.J. et al., 2005, ApJ, 633, 560

[8] Ford, E.B., 2006, ApJ, 642, 505

[9] Gregory, P.C., 2005, *Bayesian Logical Data Analysis for the Physical Sciences*, Cambridge University Press

[10] Hobson et al. (eds.) , 2009, *Bayesian Methods in Cosmology*, Cambridge University Press

[11] Lahav, O.. & Liddle, A. R., 2010, in *Reviews of Particle Physics*, arXiv:1002.3488

[12] Lewis, A. & Bridle, S.L., 2002, PRD, 66, 103511

[13] Liddle, A.R., Mukherjee, P. & Parkinson, D., 2006, astro-ph/0608184

[14] Lupton, R., 1993, *Statistics in Theory and in Practice*, Princeton University Press

[15] Lyons, L., 1986, *Statistics for Nuclear and Particle Physicist*, Cambridge University Press.

[16] Martinez, V.J., Saar, E., 2002, *Statistics of the Galaxy Distribution*, Chapman & Hall/CRC

[17] Press, W.H. et al., 1992, *Numerical Recipes*, Cambridge University Press

[18] Saha, P., *Principles of Data Analysis*, 2006 Available free on the WWW.

[19] Sivia, D., 1996, *Data Analysis: A Bayesian Tutorial*, Oxford University Press

[20] Starck, J-L, Murtagh, F., 2002, *Astronomical Image and Data Analysis*,

[21] Trotta, R., 2008, Contem. Physics, 49 (2), 71

[22] Verde, L. & the WMAP Team, 2003, ApJS, 148, 196

[23] Wall, J.V, Jenkins, C.R., 2003, *Practical Statistics for Astronomers*, Cambridge University Press