Astrostatistics

Eric D. Feigelson
Pennsylvania State University

1. Introduction

The term "astronomy" is best understood as short-hand for "astronomy and astrophysics". Astronomy is the observational study of matter beyond Earth: planets and other bodies in the Solar System, stars in the Milky Way Galaxy, galaxies in the Universe, and diffuse matter between these concentrations of mass. The perspective is rooted from our viewpoint on or near Earth, typically using telescopes or robotic satellites. Astrophysics is the study of the intrinsic nature of astronomical bodies and the processes by which they interact and evolve. This is an inferential intellectual effort based on the well-confirmed assumption that physical processes established to rule terrestrial phenomena – gravity, thermodynamics, electromagnetism, quantum mechanics, plasma physics, chemistry, and so forth – also apply to distant cosmic phenomena.

Statistical techniques play an important role in analyzing astronomical data and at the interface between astronomy and astrophysics. Astronomy encounters a huge range of statistical problems: samples selected with truncation; variables subject to censoring and heteroscedastic measurement errors; parameter estimation of complex models derived from astrophysical theory; spatial point processes of galaxies in space; time series of periodic, stochastic, and explosive phenomena; image processing of both grey-scale and Poissonian images; data mining of terabyte-petabyte datasets; and much more. Thus, astrostatistics is not focused on a narrow suite of methods, but rather brings the insights from many fields of statistics to bear on problems arising in astronomical research.

2. History

As the oldest observational science, astronomy was the driver for statistical innovations over many centuries (Stigler 1986; Hald 1998). Hipparchus, Ptolemy, al-Biruni, and Galileo Galilei were among those who discussed methods for averaging discrepant astronomical measurements. The least squares method, and its understanding in the context of the normal error distribution, was developed to address problems in Newtonian celestial mechanics during the early 19th century by Pierre-Simon Laplace, Adrian Legendre, and Carl Friedrich Gauss. The links between astronomy and statistics considerably weakened during the first decades of the 20th century as statistics turned its attention social and biological sciences while astronomy focused on astrophysics. Maximum likelihood methods emerged slowly starting in the 1970s, and Bayesian methods are now gaining considerably popularity.

Modern astrostatistics has grown rapidly since the 1990s. Several cross-disciplinary research groups emerged to develop advanced methods and critique common practices[1]. Monographs were written on astrostatistics (Babu & Feigelson (1996), galaxy clustering (Martinez & Saar 2001), image processing (Starck & Murtagh 2006), Bayesian analysis (Gregory 2005), and Bayesian cosmology (Hobson et al. 2009). The *Statistical Challenges in Modern Astronomy* (Babu & Feigelson 2007) conferences bring astronomers and statisticians together to discuss methodological issues.

The astronomical community is devoting considerable resources to the construction and promulgation of large archival datasets, often based on well-designed surveys of large areas of the sky. These datasets have terabytes to petabytes of images, spectra and time series. Reduced data products include tabular data with ~ 10 variables measured for billions of astronomical objects. Major projects include the Sloan Digital Sky Survey, International Virtual Observatory, and planned Large Synoptic Survey Telescope[2]. Too large for traditional treatments, these datasets are spawning increased interest in computationally efficient data visualization, data mining, and statistical analysis. A nascent field of astroinformatics allied to astrostatistics is emerging.

3. Topics in contemporary astrostatistics

Given the vast range of astrostatistics, only a very small portion of relevant issues can be presented here. We outline three topics of contemporary interest.

Heteroscedastic measurement errors

Astronomical measurements at telescopes are made with carefully designed and calibrated instruments, and 'background' levels in dark areas of the sky are examined to quantitatively determine the noise levels. Thus, unlike in social and biological science studies, heteroscedastic measurement error are directly obtained for each astronomical measurement. This produces unusual dataset structures. For example, a multivariate table of brightness of quasars in 6 photometric bands will have 12 columns of numbers giving the measured brightness and the associated measurement error in each band.

Unfortunately, few statistical techniques are available for this class of non-identically distributed data. Most errors-in-variables methods are designed to treat situations where the heteroscedasticity is not measured, and instead becomes part of the statistical model (Carroll et al. 2006). Methods are needed for density estimation, regression, multivariate analysis and classification, spatial processes, and time series analysis. Common estimation procedures in the astronomical literature weight each measurement by its associated error. For instance, in a functional regression model, the parameters $\hat{\theta}$ in model M are estimated by minimizing the weighted sum of squared residuals $\sum_i(O_i - M_i(\hat{\theta})^{\mathbf{2}}/\sigma_{\mathbf{i}}^{\mathbf{2}}$ of the observed data O_i where σ_i^2 are the known variances of the measurement errors.

More sophisticated methods are being developed, but have not yet entered

[1] http://hea-www.harvard.edu/AstroStat; http://www.incagroup.org; http://astrostatistics.psu.edu

[2] http://www.sdss.org, http://www.ivoa.net, http://www.lsst.org

into common usage. Kelly (2007)[3] treats structural regression as an extension of a normal mixture model, writing a likelihood which can either be maximized with the EM Algorithm or used in Bayes' theorem. The Bayesian approach is more powerful, as it also can simultaneous incorporate censoring and truncation into the measurement error model. Delaigle & Meister (2008) describe a nonparametric kernel density estimator that takes into account the heteroscedastic errors. More methods (e.g., for multivariate clustering and time series modeling) are needed.

Censoring and truncation

In the telescopic measurement of quasar brightnesses outlined above, some targeted quasars may be too faint to be seen above the background noise level in some photometric bands. These nondetections lead to censored data points. The situation is similar in some ways to censoring treated by standard survival analysis, but differs in other ways: the data are left-censored rather than right-censored; censoring can occur in any variable, not just a single response variable; and censoring levels are linked to measurement error levels. Survival techniques have come into common usage in astronomy since their introduction (Isobe et al. 1986). They treat some problems such as density estimation (with the Kaplan-Meier product-limit estimator), two-sample tests (such as the Gehan, logrank and Peto-Prentice tests), correlation (using a generalization of Kendall's τ), and linear regression (using the Buckley-James line).

Consider a survey of quasars at a telescope with limited sensitivity where the quasar sample is not provided in advance, but is derived from the photometric colors of objects in the survey. Now quasars which are too faint for detection are missing entirely from the dataset. Recovery from this form of truncation is more difficult than recovery from censoring with a previously established sample. A major advance was the derivation of the nonparametric estimator for a randomly truncated dataset, analogous to the Kaplan-Meier estimator for censored data, by astrophysicist Lynden-Bell (1971). This solution was later recovered by statistician Woodroofe (1985), and bivariate extensions were developed by Efron & Petrosian (1992).

Periodicity detection in difficult data

Stars exhibit a variety of periodic behaviors: binary star or planetary orbits; stellar rotation; and stellar oscillations. While Fourier analysis is often used to find and characterize such periodicities, the data often present problems such as non-sinusoidal repeating patterns, observations of limited duration, and unevenly-spaced observations. Non-sinusoidal periodicities occur in elliptical orbits, eclipses, and rotational modulation of surface features. Unevenly-spaced data arise from bad weather at the telescope, diurnal cycles for ground-based telescopes, Earth orbit cycles for satellite observatories, and inadequate observing time provided by telescope allocation committees.

[3] The astronomical research literature can be accessed online through the SAO/NASA Astrophysics Data System, http://adsabs.harvard.edu.

Astronomers have developed a number of statistics to locate periodicities under these conditions. Stellingwerf (1972) presents a widely used least-squared technique where the data are folded modulo trial periods, grouped into phase bins, and intra-bin variance is compared to inter-bin variance using χ^2. The method treats unevenly spaced data, measurement errors, and non-sinusoidal shapes. Dworetsky (1983) gives a similar method without binning suitable for sparse datasets. Gregory & Loredo (1992) develop a Bayesian approach for locating non-sinusoidal periodic behaviors from Poisson distributed event data. Research is now concentrating on methods for computationally efficient discovery of planets orbiting stars as they eclipse a small fraction during repeated transits across the stellar surface. These methods involve matched filters, Bayesian estimation, least-squares box-fitting, maximum likelihood, analysis of variance, and other approaches (e.g. Pontopappas et al. 2005).

References

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

[2] Babu, G. J. & Feigelson, E. D. (2007), *Statistical Challenges in Modern Astronomy IV*, Astro. Soc. Pacific

[3] Carroll, R. J., Ruppert, D., Stefanski, L. A. & Crainiceanu, C. M. (2006), *Measurement Errors in Nonlinear Models*, Chapman & Hall/CRC

[4] Delaigle, A. & Meister, A. (2008). Density estimation with heteroscedastic error. *Bernoulli*, 14, 562-579.

[5] Efron, B. & Petrosian, V. (1992), A simple test of independence for truncated data with applications to redshift surveys, *Astrophys. J.*, 399, 345-352

[6] Gregory, P. C. (2005) *Bayesian Logical Data Analysis for the Physical Sciences*, Cambridge Univ. Press

[7] Hald, A. (1998) *A History of Mathematical Statistics from 1750 to 1930*, Wiley

[8] Hobson, M. P., et al., editors (2009), *Bayesian Methods in Cosmology*, Cambridge Univ. Press

[9] Isobe, T., Feigelson, E. D., Nelson, P. I. (1986), Statistical methods for astronomical data with upper limits. II - Correlation and regression, *Astrophys. J.*, 306, 490-507

[10] Kelly, B. C. (2007), Some Aspects of Measurement Error in Linear Regression of Astronomical Data, *Astrophys. J.*, 665, 1489-1506

[11] Lynden-Bell, D. (1971), A method of allowing for known observational selection in small samples applied to 3CR quasars, *Mon. Not. Royal Astro. Soc.*, 155, 95-118

[12] Martinez, V. J. & Saar, E. (2002), *Statistics of the Galaxy Distribution*, CRC Press

[13] Protopapas, P., Jimenez, R., Alcock, C. (2005), Fast identification of transits from light-curves, *Mon. Not. Royal Astro. Soc.*, 362, 460-468

[14] Starck, J.-L. & Murtagh, F. (2006), *Astronomical Image and Data Analysis*, Springer

[15] Stigler, S. M. (1986) *The History of Statistics: The Measurement of Uncertainty before 1900*, Harvard Univ. Press

[16] Woodroofe, M. B. (1985), Estimating a distribution function with truncated data, *Ann. Statist.*, 13, 163-177