Introduction to Astrostatistics

Eric Feigelson

Center for Astrostatistics

Penn State University

Outline

I Role of statistics in astronomy

II History of statistics in astronomy

III Astrostatistics today and tomorrow

What is astronomy?

Astronomy is the observational study of matter beyond Earth: planets in the Solar System, stars in the Milky Way Galaxy, galaxies in the Universe, and diffuse matter between these concentrations.

Astrophysics is the study of the intrinsic nature of astronomical bodies and the processes by which they interact and evolve. This is an indirect, inferential intellectual effort based on the assumption that physics – gravity, electromagnetism, quantum mechanics, etc – apply universally to distant cosmic phenomena.

What is statistics? *(No consensus !!)*

- “… briefly, and in its most concrete form, the object of statistical methods is the reduction of data”

  (R. A. Fisher, 1922)

- “Statistics is the mathematical body of science that pertains to the collection, analysis, interpretation or explanation, and presentation of data.”

  (Wikipedia, 2014)

- “A statistical inference carries us from observations to conclusions about the populations sampled”

  (D. R. Cox, 1958)

Does statistics relate to scientific models?

The pessimists …

“Essentially, all models are wrong, but some are useful.”

(Box & Draper 1987)

“There is no need for these hypotheses to be true, or even to be at all like the truth; rather … they should yield calculations which agree with observations” (Osiander’ s Preface to Copernicus’ *De Revolutionibus*, quoted by C. R. Rao in *Statistics and Truth*)

The positivists …

“The goal of science is to unlock nature’s secrets. … Our understanding comes through the development of theoretical models which are capable of explaining the existing observations as well as making testable predictions. …

“Fortunately, a variety of sophisticated mathematical and computational approaches have been developed to help us through this interface, these go under the general heading of statistical inference.”

(P. C. Gregory, *Bayesian Logical Data Analysis for the Physical Sciences,* 2005)

Recommended steps in the statistical analysis of scientific data

The application of statistics can reliably quantify information embedded in scientific data and help adjudicate the relevance of theoretical models. But this is not a straightforward, mechanical enterprise. It requires:

- exploration of the data
- careful statement of the scientific problem
- model formulation in mathematical form
- choice of statistical method(s)
- calculation of statistical quantities ← *easiest step with R*
- judicious scientific evaluation of the results

Astronomers often do not adequately pursue each step

- Modern statistics is vast in its scope and methodology. It is difficult to find what may be useful (jargon problem!), and there are usually several ways to proceed. Very confusing.

- Some statistical procedures are based on mathematical proofs which determine the applicability of established results. It is perilous to violate mathematical truths! Some issues are debated among statisticians, or have no known solution.

- Scientific inferences should not depend on arbitrary choices in methodology & variable scale. Prefer nonparametric & scale-invariant methods. Try multiple methods.

- It can be difficult to interpret the meaning of a statistical result with respect to the scientific goal. Statistics is only a tool towards understanding nature from incomplete information.

We should be knowledgeable in our use of statistics and judicious in its interpretation

Astronomy & Statistics: A glorious past

For most of western history,
the astronomers were the statisticians!

Ancient Greeks to 18th century

Best estimate of the length of a year from discrepant data?

- Middle of range: Hipparcos (4th century B.C.)
- Observe only once! (medieval)
- Mean: Brahe (16th c), Galileo (17th c), Simpson (18th c)
- Median with bootstrap (21th c)

19th century

Discrepant observations of planets/moons/comets used to estimate orbital parameters using Newtonian celestial mechanics

- Legendre, Laplace & Gauss develop least-squares regression and normal error theory (~1800-1820)
- Prominent astronomers contribute to least-squares theory (~1850-1900)

The lost century of astrostatistics….

In the late-19th and 20th centuries, statistics moved towards **human sciences** (demography, economics, psychology, medicine, politics) and **industrial applications** (agriculture, mining, manufacturing).

During this time, astronomy recognized the power of **modern physics**: electromagnetism, thermodynamics, quantum mechanics, relativity. Astronomy & physics were wedded into **astrophysics**.

Thus, **astronomers and statisticians substantially broke contact**; e.g. the curriculum of astronomers heavily involved physics but little statistics. Statisticians today know little modern astronomy.

The state of astrostatistics today

(not so good but rapidly improving)

Many astronomical studies are confined to a narrow suite of familiar statistical methods:

- Fourier transform for temporal analysis (Fourier 1807)
- Least squares regression (Legendre 1805, Pearson 1901)
- Kolmogorov-Smirnov goodness-of-fit test (Kolmogorov, 1933)
- Principal components analysis for tables (Hotelling 1936)

Even traditional methods are sometimes misused!

- *Kolmogorov-Smirnov test has three limitations*
- *Likelihood ratio test can't be used for parameters near zero*
- *Bayesian priors should not be improper*

https://asaip.psu.edu/Articles/beware-the-kolmogorov-smirnov-test/

Protassov et al. 2002

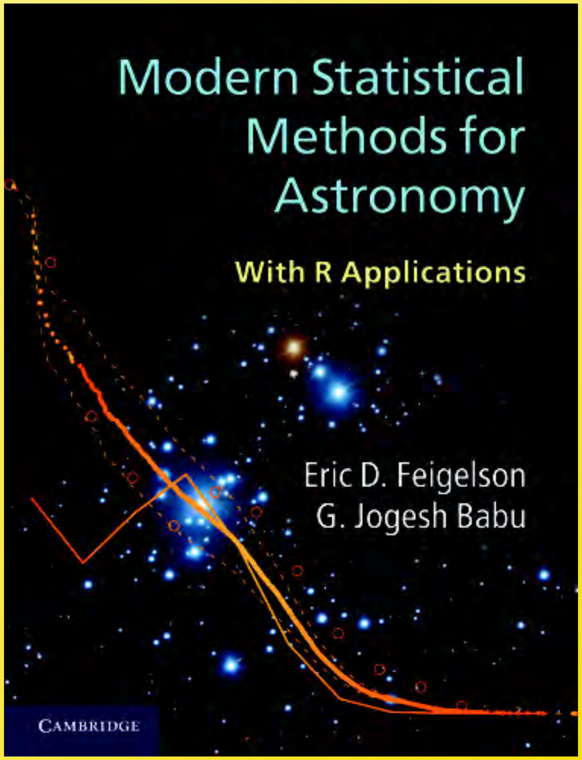
Tak et al. 2018

Under-utilized methodology from the 20th century:

- **modeling** (MLE, EM Algorithm, BIC, bootstrap)
- **multivariate classification** (LDA, SVM, CART, RFs)
- **time series** (autoregressive models, state space models)
- **spatial point processes** (Ripley's K, kriging)
- **nondetections** (survival analysis)
- **image analysis** (computer vision methods, False Detection Rate)
- **statistical computing** (R)

Advertisement …

Modern Statistical Methods for Astronomy with R Applications
E. D. Feigelson & G. J. Babu,
Cambridge Univ Press, 2012



Winner 2012 PROSE Award for Best Astronomy & Cosmology Book

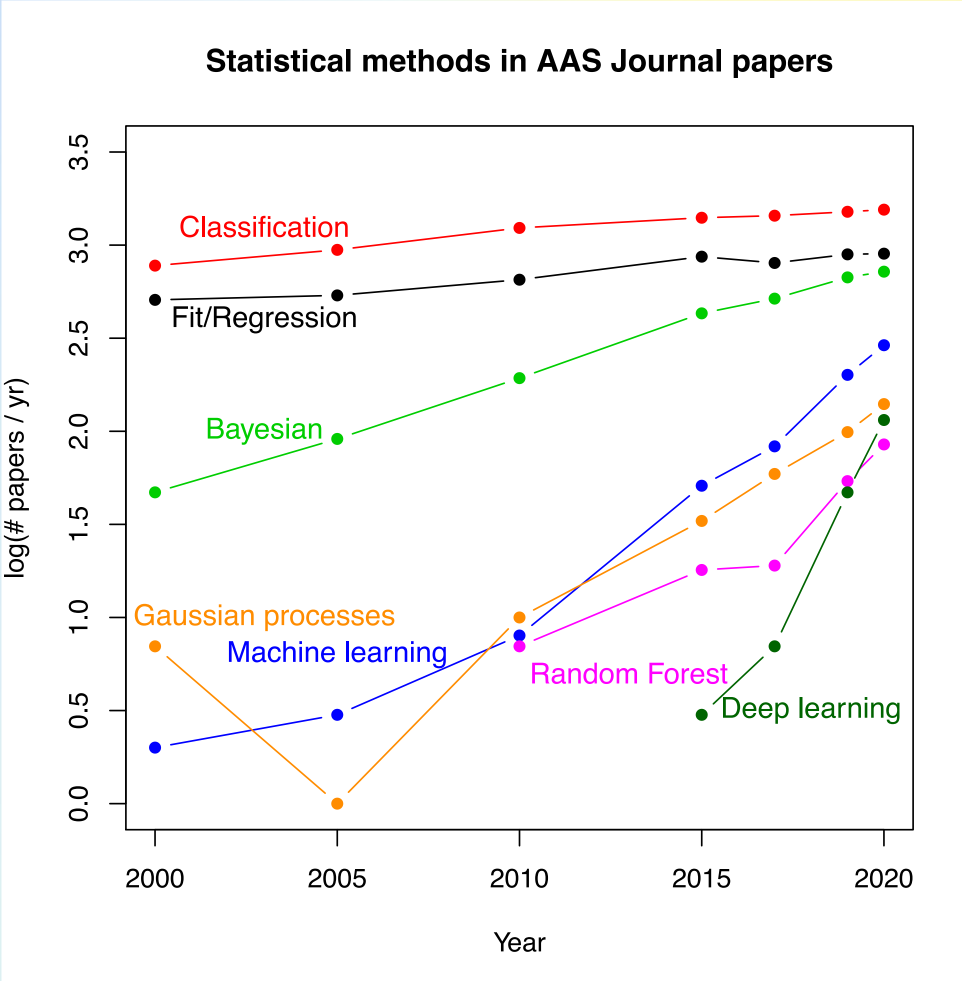
Astrostatistics is difficult: it involves many fields of statistics

Cosmology	↔	***Statistics***
Galaxy clustering	↔	Spatial point processes, clustering
Galaxy morphology	↔	Regression, mixture models
Galaxy luminosity fn	↔	Gamma distribution
Power law relationships	↔	Pareto distribution
Weak lensing morphology	↔	Geostatistics, density estimation
Strong lensing morphology	↔	Shape statistics
Strong lensing timing	↔	Time series with lag
Faint source detection	↔	False Discovery Rate
Multiepoch survey lightcurves	↔	Multivariate classification
CMB spatial analysis	↔	Markov fields, ICA, etc
ΛCDM parameters	↔	Bayesian inference & model selection
Comparing data & simulation	↔	Uncertainty Quantification

Recent resurgence in astrostatistics

- Improved access to statistical software: R/CRAN, Matlab & Python

- A significant fraction of papers in the astronomical literature use modern methodology and is growing exponentially



- **Short training courses** (Penn State, India, Brazil, Greece, Italy, France, Germany, Spain, Sweden, Japan, China, Taiwan, Thailand, Indonesia, IAU/AAS/CASCA/… meetings)

- **Cross-disciplinary research collaborations** (Harvard, Carnegie-Mellon, Penn State, CEA-Saclay/Stanford, Cornell, Imperial College London, Swinburne, …)

- **Cross-disciplinary conferences** (*Statistical Challenges in Modern Astronomy 1991-2021, Astronomical Data Analysis 1991-2016,* SAMSI 2006/2012/2016, *Astroinformatics 2012-2020*)

- **Scholarly societies** (Internationl Stat Institute SIGAstro, International Astrostatistical Assn, International Astro Union Commission B3, American Astro Soc Working Group, American Stat Assn Interest Group, LSST Info/Stat Science Collaboration, IEEE Astro Data Miner Task Force)

To treat massive data streams and databases …

Rapid rise of astroinformatics

Methodology: Computationally intensive astronomy, data mining, multivariate regression & classification, machine learning, Monte Carlo methods, NlogN algorithms, etc.

Software & hardware: Parallel processing on multi-processors machines, cloud computing, CUDA & GPU computing, database management & promulgation, etc.

Workshops & training schools emerging. IAU Symposium 2016, IEEE Symposium 2018. Growing perception that more community training is needed.

In my opinion …

Astronomers often confuse software with methods:

- ➢ Statistical methodology guides data reduction, scientific analysis and interpretation
- ➢ Software implements statistical methods with algorithms & code

Examples:

- ○ Bayesian inference requires careful formulation of the likelihood and choice of the priors. There are many MCMC algorithms, and the more efficient INLA, to compute the posterior.
- ○ Machine learning requires careful formulation of the scientific problem into a classification problem. Performance of classification algorithms that can be compared with ROC curves.

Statistics guides the scientist on what to compute
Informatics helps the scientist perform the computation

Several textbooks in astrostatistics

Bayesian Logical Data Analysis for the Physical Sciences: A Comparative Approach with Mathematica Support
Gregory, 2005

Practical Statistics for Astronomers
Wall & Jenkins, 2nd ed, 2012

Modern Statistical Methods for Astronomy with R Application,
Feigelson & Babu, 2012

Statistics, Data Mining, and Machine Learning in Astronomy: A Practical Python Guide for the Analysis of Survey Data,
Ivecic, Connolly, VanderPlas & Gray, 2014 (2nd edition in preparation)

+ many texts written by statisticians to teach specific fields of methodology, often with R code. A new volume with “R” in the title is published every ~10 days.

A vision of astrostatistics by 2030 …

- Astronomy graduate curriculum has 1 year of statistical and computational methodology
- Some astronomers have M.S. in statistics and computer science
- Astrostatistics and astroinformatics is a well-funded, cross-disciplinary research field involving a few percent of astronomers (cf. astrochemists) pushing the frontiers of methodology.
- Astronomers regularly use advanced methods coded in R.
- *Statistical Challenges in Modern Astronomy* meetings are held biannually with hundreds of participants