Astrostatistics:

Review of the emerging cross-disciplinary field

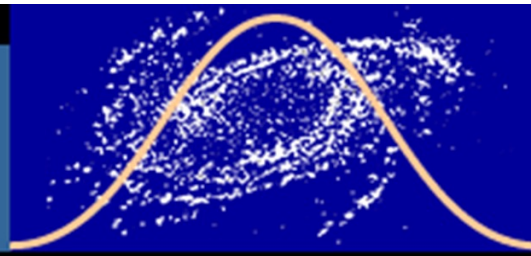
G. Jogesh Babu

The Pennsylvania State University

Department of Statistics

and

Department of Astronomy & Astrophysics



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 w/ 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 statistics in astronomy….

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.

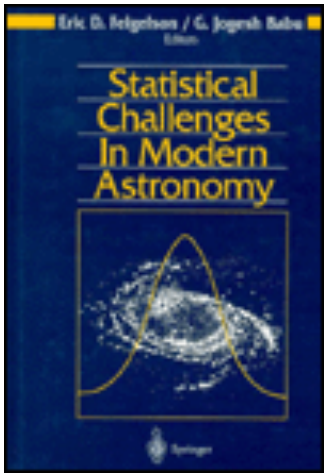
Statistical Problems in Astronomy

Surprising variety of statistical problems in astronomical research:

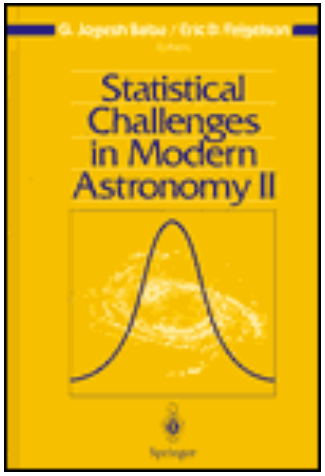
- The sky has vast numbers of stars & galaxies and defuse gas on all scales.
- Most stars have orbiting planets, most galaxies have a massive black hole
- Astronomers acquire huge datasets of images, spectra & time series of planets, stars, galaxies, quasars, supernovae, etc.
- Various properties of cosmic populations observed and empirically studied with all kinds of telescopes (n>>p)
- Properties are measured repeatedly but often with irregular spacing.
- Spatial distributions in sky (2D), space (3D), and parameter space (pD) are complex (MVN assumption usually inapplicable)

Papers in astronomical literature tripled to ~1300/yr in past decade ("Methods: statistical" or "machine learning" papers in *NASA-Smithsonian Astrophysics Data System*)

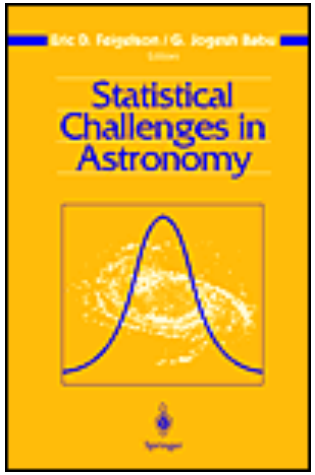
Statistical Challenges in Modern Astronomy:
Cross-disciplinary conferences



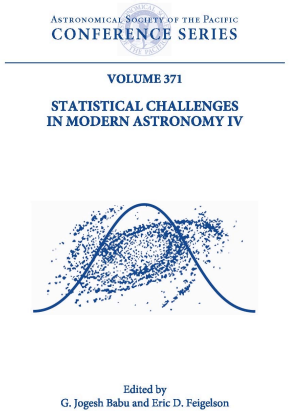
1991



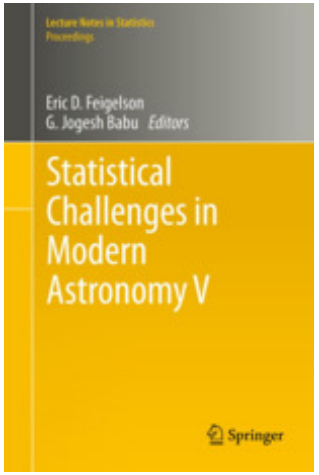
1996



2001



2006

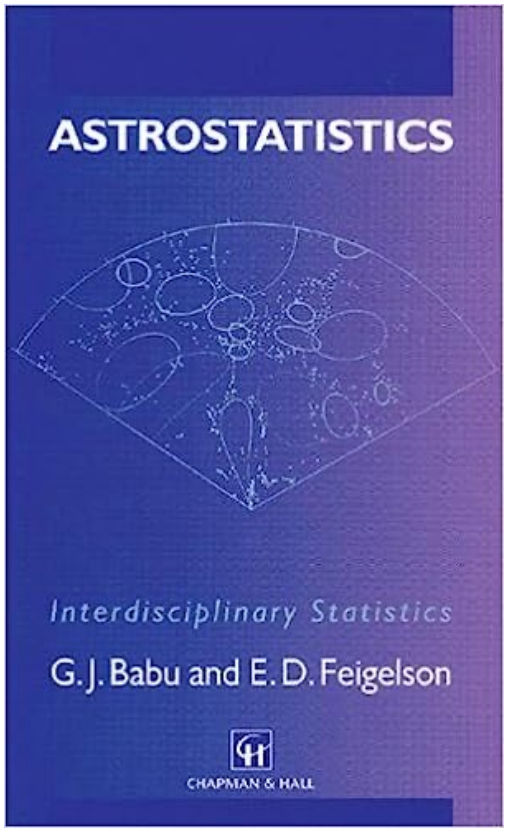


2011

SCMA VI, CMU 2016

SCMA VII, PSU / Virtual 2021

SCMA VIII, PSU 2023



Astronomer Eric Feigelson and I started collaborating in late 1980s.

The term ***Astrostatistics*** was coined in mid 1990s, when we published a book by the same name.

- The Center for Astrostatistics was created at Penn State in 2003 to facilitate development and promulgation of statistical expertise for astronomy and astrophysics.

- The activities of the Center are multi-faceted:
  - Promote research on forefront problems
  - Provide forums where active astrostatistical researchers can interact
  - Foster new cross-disciplinary collaborations
  - Liaise with other organizations oriented towards statistical applications in physical sciences.

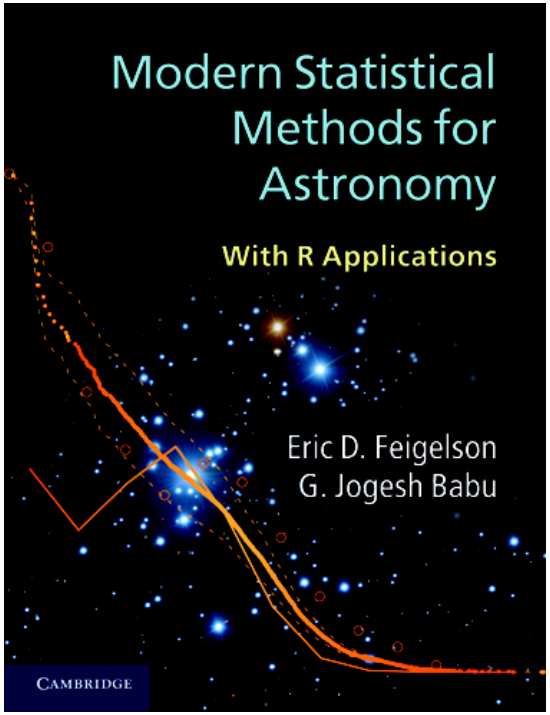
Under-utilized methodology:

- 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

Growing field

- Short training courses (Penn State, India, Brazil, Greece, China, Italy, France, Germany, Spain, Sweden, Chile, Netherlands, Thailand, Indonesia, Japan, Taiwan)

- IAU/AAS/CASCA/… meetings

- Cross-disciplinary research collaborations:
    - Harvard/ICHASC
    - Carnegie-Mellon
    - Penn State
    - NASA-Ames/Stanford
    - CEA-Saclay/Stanford
    - Cornell
    - Imperial College London
    - Swinburne/Melbourne
    - Univ Tokyo/NAOJ
    - Univ Toronto
    - Simon Fraser…

A new imperative: Large-scale surveys & megadatasets

- Huge imaging, spectroscopic & multivariate datasets are emerging from specialized survey projects & telescopes:
  - $10^{9\text{-}10}$-object photometric catalogs x 10^{0}-10^{3} epochs from 2MASS, SDSS, VISTA, CRTS, Pan-STARRS, DES, LSST …
  - $10^{6\text{-}8}$- galaxy redshift catalogs from SDSS, LAMOST, …
  - Spectral-image datacubes (VLA, ALMA, IFUs)
  - Radio interferometer data streams (e.g., 30 Tflops processor for LOFAR)
- *The Virtual Observatory is an international effort to federate many distributed on-line astronomical databases.*

Powerful statistical tools are needed to derive scientific insights from TBy-PBy-EBy databases

Broad Timeline of Astrostatistics

- LyndenBell - Woodroofe estimator (1971)
- Lomb--Scargle periodogram (1982)
- Statistical Challenges in Modern Astronomy (1991)
- Astrostatistics (1996)
- ADA meetings started (2001)
- First Astrostatistics Program at SAMSI (2006)
- International Astrostatistics Association (2010)
- Modern Statistical Methods for Astronomy with R Applications (2012)
- IAU, AAS, ASA, IEEE, ASAIP (2012--5)
- ISI- Astrostatistics Special Interest Group (2017)

Astrostatistics at SAMSI

- Several long-term research programs on Astrostatistics were organized at SAMSI since 2006.
- These programs brought together many statisticians, astronomers, physicists, and computer scientists together for fruitful collaborations.
- The programs also launched careers of many scientists. The programs and their outcomes will be presented.
- The first SAMSI Astrostatistics Program (Spring 2006), focused on: Bayesian statistics, Exoplanets, Particle physics.

- Astrostatistics sub–program (Fall 2012) of *Statistical and Computational Methodology for Massive Datasets*, focused on: The search for transients, Sparsity, Discovery & Classification in Synoptic Surveys, Inference & Simulation in Complex Models, Graphical Models & Graphics Processors.

- Exoplanets: Modern Statistical and Computational Methods for Analysis of Kepler Data. June 10-28, 2013.

- Statistical, Mathematical and Computational Methods for Astronomy 2016-2017. It focused on: Time Domain Astronomy (TDA), Exoplanet data analysis, hierarchical modeling, Uncertainty, selection effects for gravitational waves (GW), Pulsar timing arrays and detection of GWs.

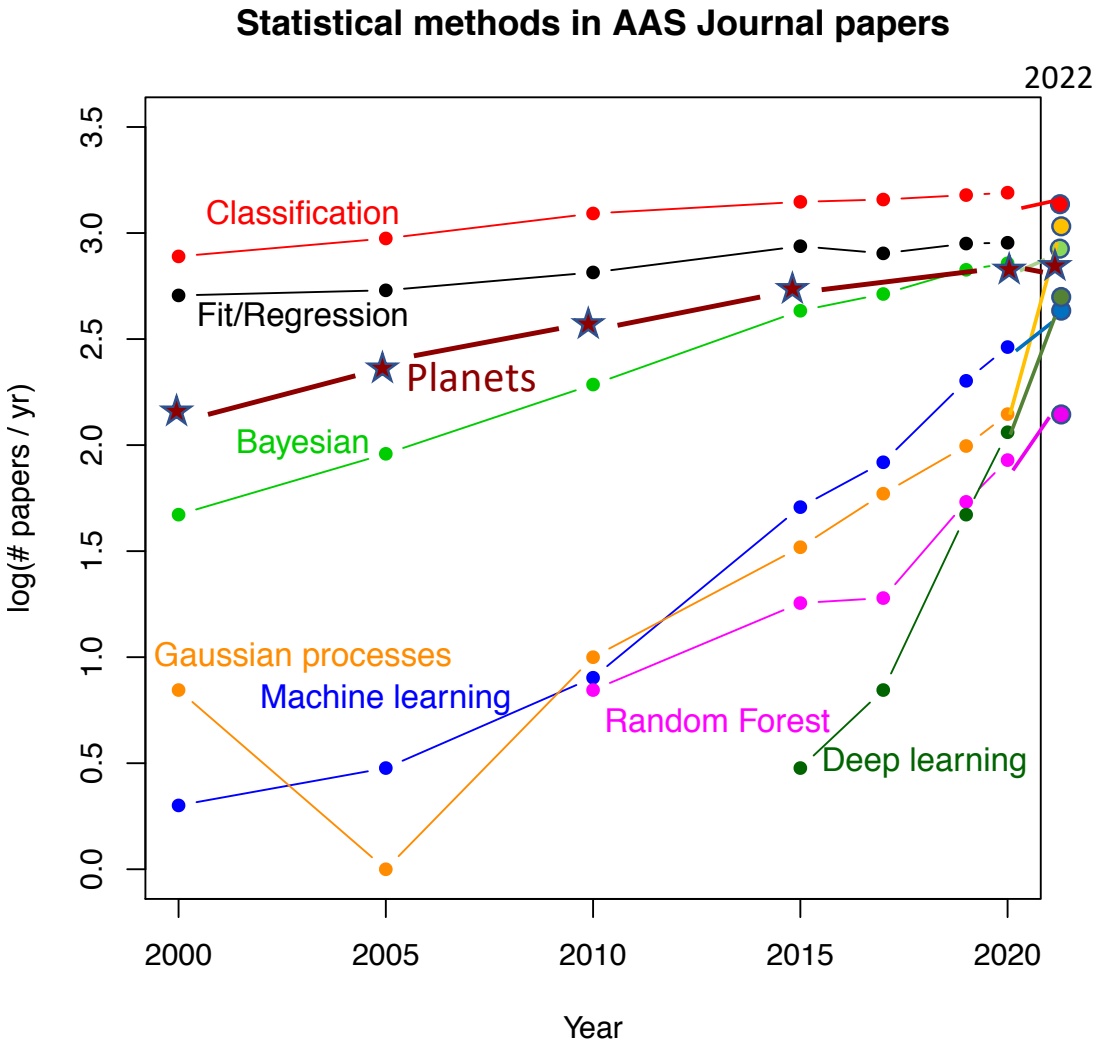
  - The program was timely. The first GW (*GW150914)* detected (100 years after Einstein's prediction) by Laser Interferometer Gravitational-Wave Observatory (LIGO) confirmed Einstein's 1915 general theory of relativity.

  - *Time Series Analysis for Synoptic Surveys and Gravitational Wave Astronomy* (March 20-23, 2017) held at (and in collaboration with) the *International Center for Theoretical Sciences,* Bangaluru, India.

Recent resurgence in Astrostatistics

Machine Learning is rising much faster than Exoplanets or any other major topical (stars, galaxies, etc) field in astronomy.

High usage of Bayesian modeling in astronomy/astrophyics, probably more than other fields. Often complex hierarchical models.



Astrostatistics: Future

Improving statistical practice across the astronomical community

- Some of the most important scientific problems in astronomy & astrophysics raise challenging problems in statistical methodology & computational issues: exoplanets, gravitational waves, cosmology
- The largest telescope projects produce enormous imaging, time series and tabular datasets requiring sophisticated methods
- Tremendous need for continued progress driven by both 'data analysis' (Big Data) and `science analysis' (complex modeling)
- Considerable progress in astronomers' usage of modern sophisticated statistical methods … but inadequately promulgated to the full community
- Considerable progress among astrostatisticians in developing innovative methods for challenging problems, but involvement by more statisticians is needed.