Statistics in astronomy and the SAMSI ASTRO Program

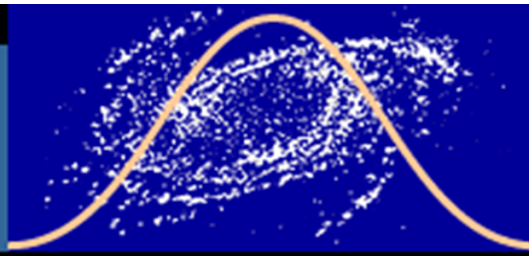
G. Jogesh Babu

(with input from Eric Feigelson)

Penn State

samsi

NSF•Duke•NCSU•UNC



Why astrostatistics?

Astronomers encounter a surprising variety of statistical problems in their research:

- The sky has vast numbers of stars & galaxies and 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 with irregular spacing.
- Spatial distributions in sky (2D), space (3D), and parameter space (pD) is complex (MVN assumption usually inapplicable)

Eric Feigelson and I started collaborating in late 1980s and the term `Astrostatistics' was coined in mid 1990s, when we published a book by the same name.

- 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 (20th 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 (c.1800-1820)
- Prominent astronomers contribute to least-squares theory (c.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.

Astrostatistics at SAMSI

Astrostatistics Program at SAMSI January 2006

- Opening workshop January 18-20, 2006
  - Bayesian astrostatistics
  - Nonparametric inference
  - Astronomy for Statisticians
- Working groups
  - Exoplanets, surveys & population studies
  - Gravitational Lensing
  - Source detection & feature detection
  - Particle physics
- Concluded with SCMA IV at Penn State in June 2006

Astrostatistics at SAMSI

Astrostatistics sub Program Fall 2012

Statistical and Computational Methodology for Massive Datasets

- Workshop September 19-21, 2012
  - The search for transients
  - Missions (Fermi, SDSS. DES, Plank, LSST, LIGO)
  - Sparsity (*high-dimensional data, but low-dim signal*)
  - Data mining
- Working groups
  - Discovery & Classification in Synoptic Surveys;
  - Inference & Simulation in Complex Models,
  - Stochastic Processes & Astrophysical Inference
  - Graphical Models & Graphics Processors

Astrostatistics at SAMSI

Exoplanets Summer 2013

Modern Statistical and Computational Methods

for Analysis of Kepler Data

June 10-28, 2013

Astrostatistics at SAMSI

Statistical, Mathematical and Computational Methods for Astronomy 2016-2017

- Opening workshop August 22-26, 2016
  - Time Domain Astronomy (TDA)
  - Exoplanet data analysis, hierarchical modeling
  - Uncertainty, selection effects for gravitational waves (GW)
  - Pulsar timing arrays and detection of GWs
  - Non-stationary, non-Gaussian, irregularly sampled processes
  - Statistical issues in cosmology

SAMSI ASTRO Program is timely

- Gravitational waves detected 100 years after Einstein's prediction

- First GW (*GW150914)* detected by Laser Interferometer Gravitational-Wave Observatory (LIGO) confirmed Einstein's 1915 general theory of relativity applied to inspiraling binary black holes.

- The planning meeting on September 21, 2015 included LIGO scientists who had only just learned of the candidate detection, and had to keep it secret until confirmed and announced on February 11, 2016.

  A second GW event (*GW151226)* was announced on June 15, 2016. It was recorded on December 26, 2015.

SAMSI 2016-17 Working Groups

- Synoptic Time Domain Surveys
- Multivariate and Irregularly Sampled Time Series
- Uncertainty Quantification and Astrophysical Emulation
- Astrophysical Populations
- Statistics, Computation, and Modeling in Cosmology

Synoptic Time Domain Surveys

- TDA has been getting richer with datasets spanning decades, many spectral bands. Includes both dense and sparse light-curves for hundreds of millions of sources.
- The variety, volume and velocity squarely fall in Big Data regime.
- The light-curves often have large gaps, are heteroskedastic, and the intrinsic variability – often poorly understood – adds an element of uncertainty when multi-band data are not obtained simultaneously.
  - binary black-hole searches from CRT Transient Survey
  - Kepler-type planet search/characterization (also in other WGs)

Group Leaders: Ashish Mahabal (Astro, Caltech), G. Jogesh Babu (Stat, PSU)

Big Questions

- What mathematical and statistical approaches can be used to best characterize and quantify salient features of irregular, heteroscedastic, gapped time series?

- Can we identify specific feature sets, templates and models? How can we identify many weak features or a few strong ones in such high dimensional time series Big Data? Are there specific domain-knowledge based features that can be identified to improve class discrimination?

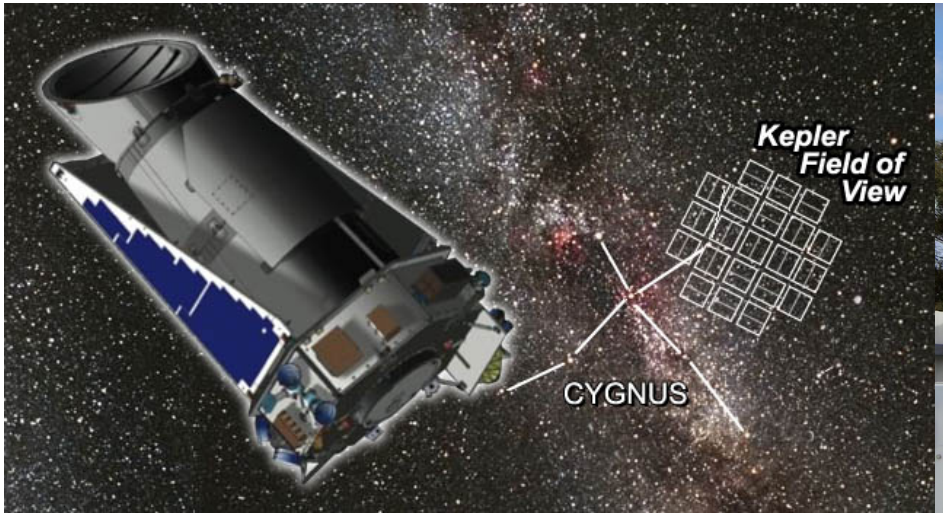
- How can significant outliers/anomalies and subclasses be detected?

Subgroups

- Data challenges
- Creating designer features for classification
- Where should the next point be? Scheduling and optimizing observations.
- Interpolating time-series
- Smart/fast ways of incorporating non-structured ancillary information
- Outlier detection (clustering)
- Domain Adaptation for classification (combining diverse datasets)
- Light-curve decomposition (e.g. ARIMA & ARFIMA)

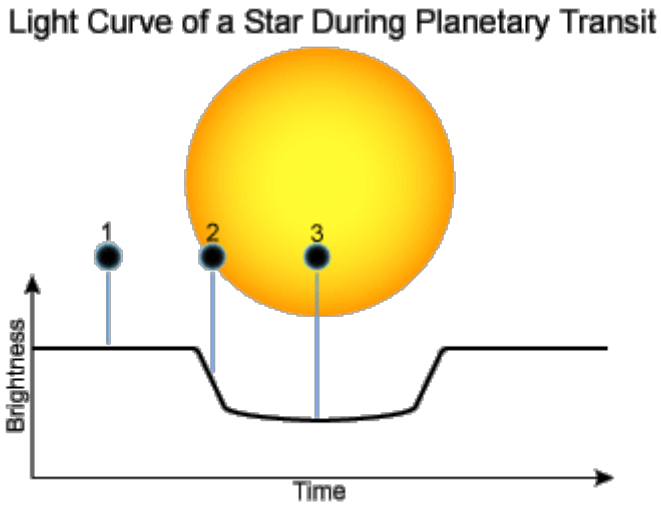
Exoplanets

- A major problem for planet detection is stellar variability
- Kepler AutoRegressive Planet Search (KARPS) uses ARIMA-type models to remove aperiodic stellar variations followed by transit comb filter to find periodic planet transits
- Preliminary KARPS analysis of irregularly spaced time series using cadences from Hungarian Automated Telescope South (HATS) network.

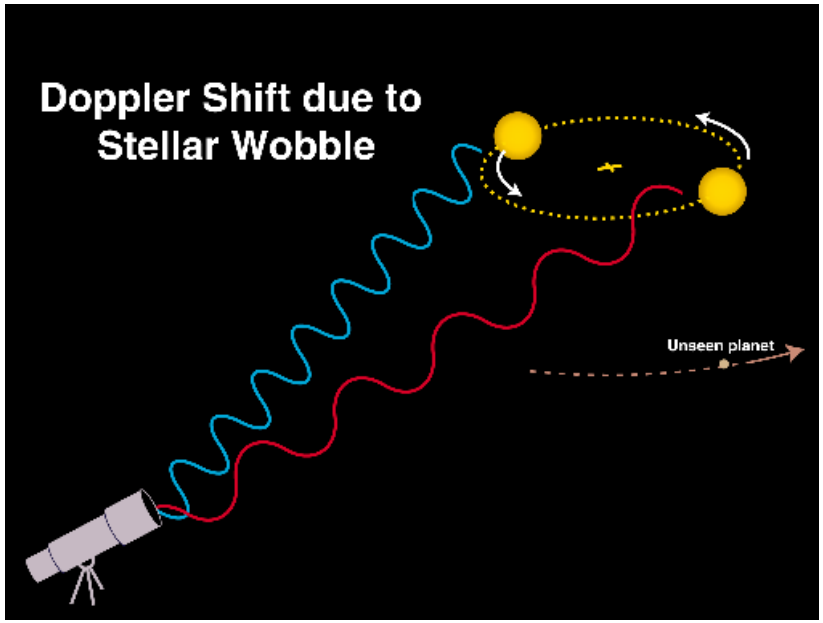


Two methods for detecting planets orbiting stars

Photometric time series with periodic transits when planet passes in front of star



Radial velocity time series with periodic variations when planet pulls star towards and away from us



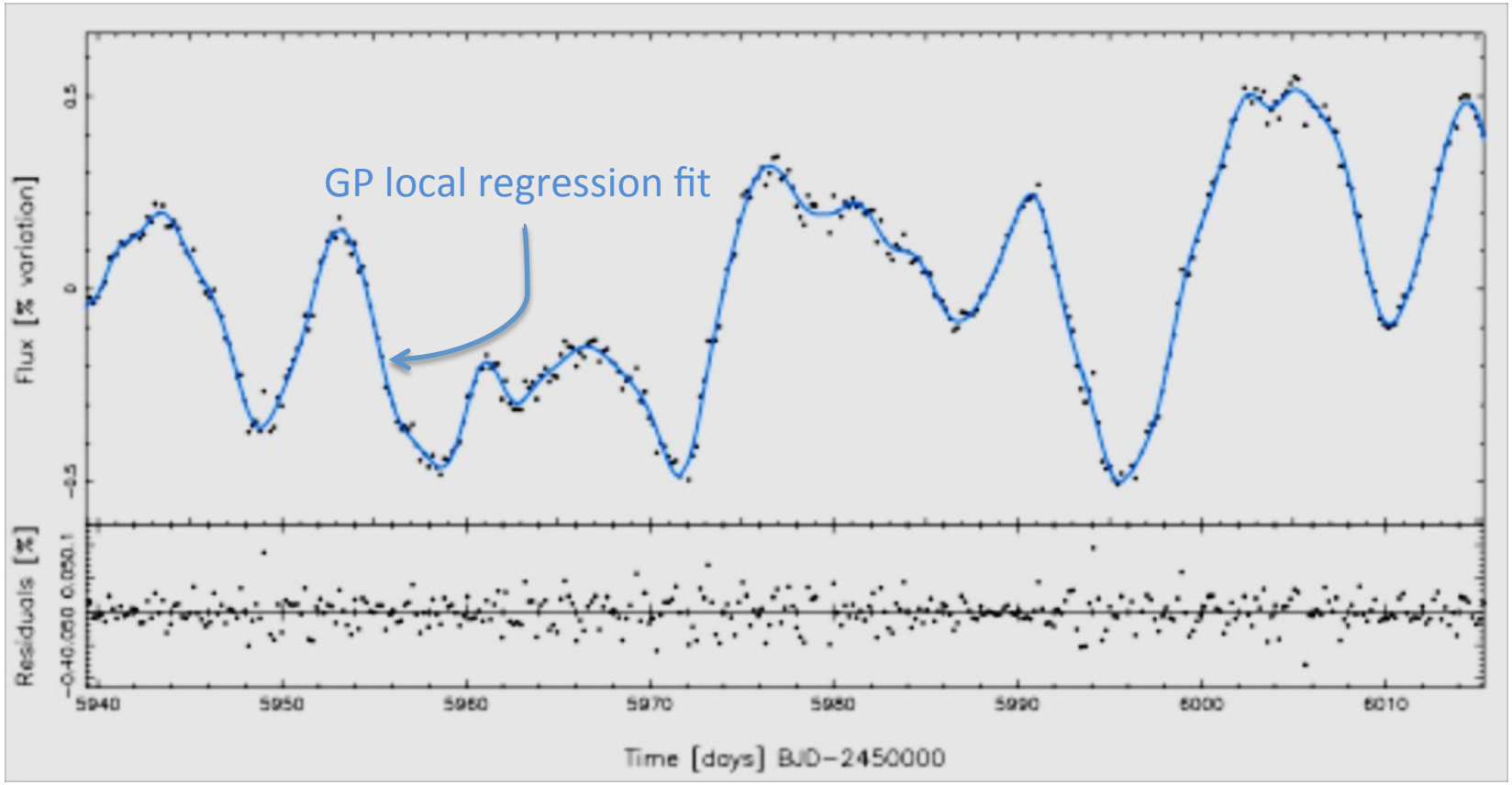
A significant challenge is that astronomical time series are usually acquired with irregularly spaced observations.

Statistical challenges in planet detection

The discovery of planets in astronomical radial velocity or photometric time series (light curves) involves 3 statistical stages:

- Time domain removal/suppression of variations intrinsic to the star
- Frequency domain periodogram to reveal periodic signature of planetary orbit

But the star itself often shows variability in brightness or radial velocity due to magnetic activity. This has been well-studied on our Sun and strongly magnetically active stars. But until the sensitive photometry of the Kepler mission, it was not recognized how stellar magnetic activity is so pervasive.



Planets and stellar activity: Hide and seek in the CoRoT-7 system
Haywood et al. 2014

“[E]xoplanets may still be detected by exploiting differences in timescale, shape and wavelength dependence between the planetary and stellar signals. … However, [stellar] variability, combined with residual instrumental systematics, is still limiting the detection of habitable planets by Kepler. … [A] better understanding of, and ability to mitigate, stellar variability, continues to be crucial for the continued development of exoplanet studies.”

Suzanne Aigrain IAU 2015

Stellar variability reduction methods

Nonparametric modeling

- *Independent Component Analysis* (Waldmann et al. 2012)
- *Wavelet decomposition* (Kepler Team pipeline; Jenkins et al. 2010)
- *Gaussian Processes regression* (Aigrain et al. 2012 & many others)
- *Correntropy, trend entropy, Empirical Mode Decomposition, Singular Spectrum Analysis, …* (Huijse et al. 2012, Roberts et al. 2013, Greco et al. 2016)

Parametric modeling

- *Principal components analysis*
- *Autoregressive modeling*

Periodicity search methods

A common task is the search for periodicity from exoplanetary orbits in stellar time series. Statistical techniques include

- *Fast Fourier Transform* Infinite, stationary, evenly spaced, Gaussian data with sinusoidal signal
- *Lomb-Scargle periodogram* (Scargle 1982) Least-squares fitting of sinusoids to irregularly spaced time series
- *Phase dispersion minimization* (Stellingwerf 1977) Binned variance-based measure designed for non-sinusoidal shapes, irregular data & heteroscedastic measurement errors
- *Minimum strength length* (Dworetzky 2003) Unbinned measure similar to PDM
- *Step-function* (Gregory & Loredo 1992) Bayesian simultaneous modelling of signal shape & strength
- ***Box least squares*** (Kovacs et al. 2002) Matched filter for periodic box shape expected from planetary transit s
- ***Transit Comb Filter*** (Caceres et al. 2017) Matched filter for periodic double-spike expected from transits after differencing operator is applied

Multivariate and Irregularly Sampled Time Series

- The noise from GW detectors changes on long timescales with frequent “glitches”. How can we build robust noise models and exploit information in auxiliary environmental channels?
- Ground-based photometric surveys, with billions of irregular light curves, detect stellar variability from a multitude of sources, from variable stars to supernovae. How do we effectively separate these classes in irregularly sampled data?
- How can we maximize the sensitivity of space-based photometric surveys (Kepler, TESS) for detecting small planets in the presence of stellar activity?
- How do we infer planet masses from observational data affected by stellar activity ‘jitter’?

Group Leaders: Ben Farr (Astro, U. Chicago), Soumen Lahiri (Stat, NCSU)

GW from Pulsar Timing

- GW background produces a fluctuation in the pulse times of arrival (TOAs) from a given pulsar. Study using cross-correlations of pulsar pairs.

- Try partial autocorrelations and multiple correlations to extract the effect of the GWB?

- Demorest et al. (ApJ 2016) use Bayesian methods to compute upper limits on the Isotropic Stochastic GWB from 9 year dataset.

Uncertainty Quantification and Astrophysical Emulation

- Uncertainty Quantification (UQ) and Reduced Order Modeling (ROM) are at the core of many problems in gravitation and cosmology, from direct simulations of the Einstein equations to the inverse problem.

- Leverage expertise from areas such as generalized polynomial chaos and simulator emulation based on stochastic processes to apply to gravitation, astrophysics, and cosmology.

  Group Leaders: Derek Bingham (Simon Fraser), Earl Lawrence (LANL)

Astrophysical Populations

- Improve the statistical methodology to infer the underlying population of exoplanets & gravitational waves (GW) sources.
- The exoplanets community needs to robustly detect and characterize planets in the presence of stellar activity from photometric and Doppler surveys without a first-principles model.
- The GW community is interested in detecting gravitational wave sources for which the details of the primary GW signal and/or backgrounds are unknown.
- Both applications require algorithms to efficiently explore high-dimensional parameter spaces and to establish confidence in detections, despite complex and unknown sources of uninteresting background signals.

Group Leaders: Jessi Cisewski (Stat, Yale); Eric Ford (Astro, Penn State)

Statistics, Computation, and Modeling in Cosmology

- Seek to bring together leading researchers in cosmology, computational spatial and Bayesian statistics, experimental design, and computer modeling to develop methodology necessary for answering fundamental questions about the origin and large scale structure of the universe.

- How can we make inferences for deterministic nonlinear dynamical systems (inference of initial conditions and model parameters)?

Group Leaders: Jeff Jewell (Astro, JPL), Joe Guinness (Stat, NCSU)

Conclusion

Some of the most important scientific problems in astronomy & astrophysics raise challenging issues in statistical methodology: exoplanets, gravitational waves, cosmology

The largest telescope projects produce enormous imaging, time series and tabular datasets requiring sophisticated methods

Over the past decade, the SAMSI organization provides a forum where these issues can be pursued by cross-disciplinary teams of statisticians & astronomers