Education and Heritage in the era of Big Data in Astronomy
Proceedings IAU Symposium No. 367, 2020
R. M. Ros, B. Garcia, S. R. Gullberg, J. Moldon & P. Rojo, eds.
doi:10.1017/S174392132100003X

Penn State Summer Schools

Gutti Jogesh Babu

Dept. of Statistics
Dept. of Astronomy & Astrophysics, Penn State University,
University Park, USA
email: babu@psu.edu

Abstract. Intensive week-long Summer Schools in Statistics for Astronomers were initiated at Penn State in 2005 and have been continued annually. Due to their popularity and high demand, additional full summer schools have been organized in India, Brazil, Space Telescope Science Institute.

The Summer Schools seek to give a broad exposure to fundamental concepts and a wide range of resulting methods across many fields of statistics. The Summer Schools in statistics and data analysis for young astronomers present concepts and methodologies with hands on tutorials using the data from astronomical surveys.

Keywords. statistical inference, classification, clustering, high-performance computing.

1. Introduction

Astronomical research today often involves organizing vast surveys of imaging, photometric and spectroscopy data producing tera/petabyte databases and billion-object catalogs. Many time-domain surveys in visible light are underway: DES, SDSS IV-V, ZTF, SNF, Pan-STARRS, and others leading to ever increasing databases. Radio interferometric telescopes like the NRAO Jansky Very Large Array and Atacama Large Millimeter Array (ALMA) are producing enormous datasets. The forthcoming Vera C. Rubin Observatory project's 10-year Legacy Survey of Space and Time (LSST) is expected to deliver 500 petabyte set of images and data products. While the scientific promise of these surveys is great, the scientific goals cannot be achieved with a narrow suite of statistical methods and old-fashioned labor-intensive approaches in common use by the astronomical community.

Astronomers can be enormously assisted by improved knowledge of fundamental statistical inference and of modern application methods. The procedures may be buried in the standard reduction of data from a detector, in the exploration of a megascale multivariate sample, in the analysis of a stochastic time series, and in the fitting of a nonlinear astrophysical model to observational results. Model fitting methods such as maximum likelihood estimation, Bayesian inference, and nonparametric goodness-of-fit tests are needed. Due to the structure of undergraduate and graduate curricula, astronomers are not well trained in statistics or informatics, receive inadequate conceptual foundations in mathematical statistics, and little training in advanced algorithms and computational techniques.

2. Week-long summer schools

Intensive week-long Summer Schools in Statistics for Astronomers were initiated at Penn State in 2005 and have been continued annually. Additional full summer schools have been organized in India, Brazil, Taiwan, and Space Telescope Science Institute, with shorter roving tutorials in a dozen countries on four continents. These schools are organized by the Center for Astrostatistics at the Pennsylvania State University.

The curriculum emphasizes principles and methods of statistical inference. Lectures are accompanied by hands-on software tutorials with applications to contemporary astronomical datasets. Ambitious in scope for a short course with about 40 contact hours, it is taught roughly at the level of senior classes for undergraduate statistics majors (for example at the level of the textbook by Ross (2019), but requires coverage of material normally distributed over many courses. A typical School has 8-10 different instructors, most experienced Professors of Statistics. Some lectures are given by statistically-expert astronomers.

The topics covered during the very popular Penn State schools include: probability, inference, regression, model selection & validation, maximum likelihood methods, non-parametric methods, multivariate methods, clustering & classification, Bayesian inference, MCMC, spatial statistics, Time series, Fundamentals of scientific computing, high-performance computing, Bayesian computation, Machine Learning algorithms, optimization, Gaussian process regression, neural networks & deep learning.

3. Statistical issues arising in astronomical research

The following are some of the questions often arise in astronomical research.

How can one reliably extract a population of supernovae from a huge multi-epoch sky survey? How do we model the vast range of variable objects arising from accretion onto compact objects? How can we find faint emission line galaxies or diffuse structures in from large spectro-imaging datacubes? How do we address data limitations such as heteroscedastic measurement errors, censoring and truncation (nondetections), sparse data, and irregular time sampling? How can uncertainties in astrophysical models be incorporated into the inferential process? How do we classify time series expected to be generated by LSST type surveys, that are generally sparse, irregular, heteroscedastic, and describe a large set of different transient or variable objects (AGNs, binary stars, supernovae, etc.)? The summer school curriculum is designed to train the participants to address these issues.

4. Conclusion

Penn State Summer Schools at the Center for Astrostatistics have trained about 950 young astronomical researchers in statistical inference and in the use of the R statistical computing environment. They consistently attract large number of women who are underrepresented in the physical sciences. In addition to US, participants arrived from Australia, Belgium, Canada, Chile, China, Czech Republic, Denmark, France, Germany, Hungary, Indonesia, Italy, Iran, Israel, Japan, Mexico, Morocco, Netherlands, Poland, Russia, South Africa, Spain, Sweden, Switzerland, Taiwan, UK. In 2018, a full week of astroinformatics is added. This new component gave training in high performance computing with an emphasis on Bayesian computing. Participants used the hybrid CPU/GPU cluster at Penn State. The lectures and tutorials of these summer schools are archived at the Center for Astrostatistics's website https://astrostatistics.psu.edu/.

References

Ross, S. 2019, *A First Course in Probability* (10th Edition), New York: Pearson.
Center for Astrostatistics's Website. *Lecture notes and tutorials of recent summer schools.*