Challenges and Methods for Massive Astronomical Data

mini-Workshop on

Computational AstroStatistics

The workshop is sponsored by CHASC/C-BAS, NSF grants DMS 09-07185 (HU) and DMS 09-07522 (UCI), and the Chandra X-Ray Center

Small Challenges

- CHASC

  Chandra AstroStatistics Collaboration

- C-BAS

  California-Boston-Smithsonian AstroStatistics Collaboration

Past Challenges

1997 Meeting on "Statistical Challenges in Modern Astronomy"

Chandra*

AXAF Data Analysis Challenges

Aneta Siemiginowska[1], Martin Elvis[1], Alanna Connors[2], Peter Freeman[3], Vinay Kashyap[3], and Eric Feigelson[4]

ABSTRACT The high quality of the AXAF X-ray data provides new challenges for the X-ray data analysis. It is clear that an "old" approach is not enough to fully exploit the capabilities of the AXAF instruments. We describe a few of the statistical and computational problems that we have so far identified. Some of them appear to be theoretically solvable but computationally challenging, while others state problems for theoretical statistics which, so far as we know, are unsolved.The problems divide, from an astronomical point of view, into: Modeling the Data (e.g. nonlinear parameter estimation, uncertainties in the model, weighting the data, correlated residuals), Source Detection (events in N-space, use of wavelets, significance of detected structures) and Instrument Related Issues (pile-up in AXAF ACIS, overlapping orders in grating spectra).

Here are a few examples from many of workshops that have been organized by CHASC in the past:

- Data Analysis Challenges in Solar & Stellar Coronal Astrophysics at AAS 199 (2002, Washington, DC)
- Data Analysis Challenges in Solar and Stellar Astrophysics at AAS/SPD 2003 (2003, College Park, MD)
- Current Challenges in Poisson Multi-Scale Deconvolution Methods (2003, Cambridge, MA)

* Chandra was launched on July 23, 1999

Current Challenges

- Astrostatistical efforts have focused on principled analysis of individual observations, on one or a few sources.
- The new data forces us to consider combining multiple datasets and infer parameters that are common to entire populations.
- Many astronomers really want to use every data point and even non-detections, but this becomes problematic for many statistical techniques.
- ….

Workshop Goals

- Explore new problems in data analysis that arise from data complexity.
- Focus on problems generally considered intractable due to insufficient computational power or inefficient algorithms.
    - Accounting for uncertainties in instrument calibration.
    - Classification, regression, and density estimations of massive data sets that may be truncated and contaminated with measurement errors and outliers.
    - Designing statistical emulators to efficiently approximate the output from complex astrophysical computer models and simulations, thus making statistical inference on them tractable.
- Define a path for development of new algorithms that target specific issues.

Workshop Day 1

9.30 - Noon Session 1A Moderator: Andreas Zezas (Crete)

Kirk Borne (George Mason) - LSST: Informatics and Statistics Research Challenges
Keith Arnaud (GSFC) - LISA: A Big Problem on a Small Data Set
Brandon Kelly (SAO) - Constraining astronomical populations with truncated data sets

Noon - 1:30pm : Lunch

1:30pm - 4pm : Session 1B Moderator: Paul Baines (UC Davis)

Peter Freeman (CMU) - Nonlinear Data Reparametrization with Diffusion Map
Joey Richards (UC Berkeley) - Real-time Classification for The Palomar Transient Factory
Daryl Geller (Stony Brook) - Spherical wavelets for CMB temperature and polarization data analysis

4pm - 4:20pm : Coffee break

4:20pm - 5:30pm : Open Discussion Moderator: Vinay Kashyap (SAO)

Workshop Day 2

9:30am - Noon : Session 2A Moderator: Jeremy Drake (SAO)

Paola Testa/Alisdair Davey (SAO) - Challenges in Data Distribution and Analysis with the Solar Dynamics Observatory
Ashish Mahabal (CalTech) - Where statistical methods can help with Transients classification from surveys
Pavlos Protopapas (SAO) : Stellar Variability Classification Using Machine Learning

Noon - 1:30pm - Lunch

1:30pm - 4pm : Session 2B Moderator: Brandon Kelly (CfA)

Alexander Gray (Georgia Tech) - Beyond RAM: Fast Statistical Analysis in Databases.
Alex Blocker (Harvard) - Semi-parametric Robust Event Detection for Massive Time-Series Datasets
Lukasz Wyrzykowski (Cambridge) - Transient classification with Gaia

4pm - 4:20pm : Coffee break

4:20pm - 5:15pm : Open Discussion Moderator: David van Dyk (UC Irvine)

5:15pm - 5:30pm : Alanna Connors (Eureka Sci) : Workshop Wrap-up

Discussion Time