Large Synoptic Survey Telescope

www.lsst.org

The Spectrum of LSST Data Analysis Challenges: Kiloscale to Petascale

Thomas J. Loredo[1], G. J. Babu[2], K. D. Borne[3], E. D. Feigelson[2], A. G. Gray[4], LSST Informatics and Statistics Science Collaboration

[1]Cornell Univ., [2]Penn State Univ., [3]George Mason Univ., [4]Georgia Institute of Technology

The unprecedented science opportunities enabled by LSST's wide-fast-deep mode of operation are accompanied by equally unprecedented data analysis challenges, due to the huge size and synoptic scope of LSST data products. The most obvious challenges are those associated with processing the petabyte-scale fundamental LSST image data. But the challenges will not end with the production of official LSST catalogs and databases. Science with LSST data will present new data analysis challenges spanning a broad range of sizes, types, and complexity, requiring innovative methodological research across this full range. We present representative examples of LSST data analysis problems of various scales, displaying some of the diversity of astroinformatics/astrostatistics research astronomers using LSST data must undertake.

Kiloscale Problems

- "Kiloscale" = 10^2 to 10^4 objects or samples
- Storage requirements: Modest
- *Challenges are essentially statistical:* New methods for variability, periodicity, outlier & changepoint detection; upper limits; population modeling using catalogs with heteroscedastic source measurement errors
- Example kiloscale datasets:
  - Multicolor, multi-epoch photometry for a single Object Catalog object
  - Catalogs for modest-sized populations (TNOs, GRBs, microlensing events, rare stellar or galaxy types...)

Sample Problem: Population modeling accounting for source uncertainties

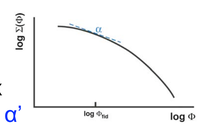
Catalogs tabulate source properties that are uncertain due to noise, counting uncertainties, or "scatter" in properties inferred from correlations. E.g.:

- Flux (magnitude) and color uncertainties
- Photo-z uncertainties
- Uncertainties in luminosities based on fundamental plane or Tully-Fisher relations

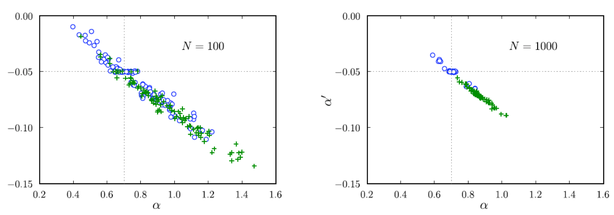
Ignoring uncertainties can corrupt inferences about population properties. The statistics literature on *measurement error* or *errors-in-the-variables models* provides tools for taking source uncertainties into account.

Example: TNO number counts

- Simulate TNOs from a "rolling power law" flux distribution, with power law slope α at fiducial flux Φ_{fid}, and rate of change of slope α'



- Simulate observations from a photon-counting instrument, with ~15% flux errors at threshold
- Estimate α and α' two ways:
  - Maximum likelihood (ML, plugs in best-fit fluxes)
  - Maximum *marginal* likelihood (MML, Bayesian approach averaging over flux errors)



- Results for 100 simulated surveys, green=ML, blue=MML; crosshair shows true values
  - Left: Surveys of 100 objects; ML noticeably biased but parameter uncertainties are larger than bias
  - Right: Surveys of 1000 objects; ML converges away from truth ("inconsistent"); MML is accurate

Lessons/issues for LSST

- Population modeling must explicitly account for source uncertainties
- Small uncertainties merely postpone the inevitable; need algorithms that scale to megascale/gigascale

Megascale Problems

- "Megascale" = 10^5 to 10^7 objects or samples
- Storage requirements: Datasets of several *megabytes to gigabytes*, corresponding to "low-volume" queries against LSST data products
- *Challenges are both statistical and computational:* Methods must be computationally efficient; advanced data visualization techniques are needed
- Example megascale datasets:
  - Multi-epoch image data for extended objects (sample=pixel)
  - Catalogs for large populations (quasars; variable Galactic stars; low redshift galaxies...)
  - LSST follow-up measurements for objects in previously-compiled catalogs

Sample Problem: Source detection with multi-epoch/multi-band imaging data

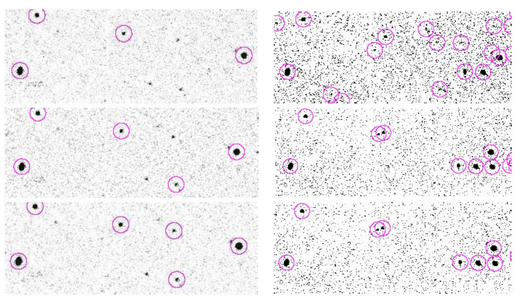
What is the best way to combine multiple images to detect very faint objects? The simplest approach—image registration & stacking—can behave poorly, e.g., for transients.

Mahalanobis distance

- Standard tool from classical multivariate analysis
- Distance D of a sample vector $\boldsymbol{x}$ from a population with mean location $\boldsymbol{\mu}$ and covariance matrix Σ:

$$D^2 = (\boldsymbol{x} - \boldsymbol{\mu}) \cdot \Sigma^{-1} \cdot (\boldsymbol{x} - \boldsymbol{\mu})$$

- Locally, the noise is assumed to be homogeneous and Σ is estimated using the neighboring pixels
- Uncorrelated case (diagonal Σ) reduces to χ^2 co-addition approach of Szalay et al. (1999)
- Incorporating correlations between epochs/bands can improve power (e.g., takes advantage of background correlations)
- Sensitive to transients as well as persistent sources
- Computationally efficient



Top-bottom: 3-, 5-, 7-epoch co-added *B* band images from Palomar-QUEST with "4σ" detections circled. *Left:* Traditional pixelwise averaging (w. outlier removal). *Right:* Mahalanobis distance maps. Multiple new detections in lower right are likely due to a minor planet passing through the field.

Lessons/issues for LSST

- Basic D^2 makes no use of PSF; this exaggerates noise if few epochs/bands used; explore use of PSF or simpler adjacency criteria
- Use false discovery rate (FDR) control or other multiple testing methods to adaptively set detection thresholds to meet survey criteria

Gigascale Problems... and Beyond

- "Gigascale" = 10^8 to 10^{10} objects or samples
- Large-scale pixel-based computation will be *tera-* or *peta*-scale
- Storage requirements: *Multiple-terabyte* datasets, corresponding to "high-volume" queries against LSST data products; out-of-core processing needed; extreme examples can go to *petabytes*
- Example gigascale and beyond datasets:
  - Calibrated image data for a small number of fields (~10^9 pixels)
  - Catalogs of very large populations (10^8 to 10^{10} stars/galaxies)
  - Large parts of the Level-One Source Catalog, for serendipitous discovery (up to 10^3 attributes for 10^6 to 10^{10} objects)
  - Images for pixel-based calculations of new object attributes (petascale)
- *Challenges are fundamentally computational:* Even basic statistical methods will require innovative algorithms; high-performance parallel/distributed/cloud computing essential

Sample Problem #1: LSST Temporal Behavior Classifier

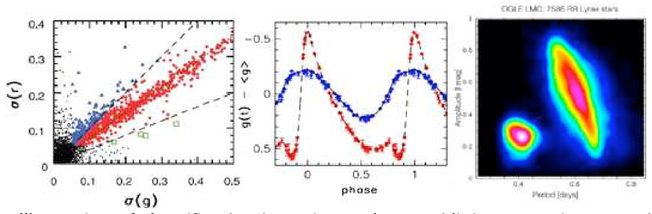


Illustration of classification based on colors and light curve shape, period and amplitude. *Left:* Relationship between the photometric root-mean-square scatter measured in SDSS *g* and *r* bands for point sources with colors typical of RR Lyrae stars. Dashed lines mark the region consistent with RR Lyrae light curves. *Middle:* Examples of RR Lyrae light curves (blue: c type, red: ab type) measured by SDSS. *Right:* Distribution of RR Lyrae stars of both types in the amplitude vs. period diagram (measured by the OGLE survey).

- Extract a major portion of the LSST Source Catalog (e.g., the time series for all photometrically variable objects)
- Sort the known variability classes via cuts or bins, with (potentially) hundreds of thousands of training examples of each class (perhaps 1000's of classes)
- Characterize the temporal behavior of each variability class (e.g., using 10-100 Fourier or wavelet coefficients) using a portion of the Source Catalog (= the training set)
- Use the remainder of the Source Catalog (= the test set) to test the accuracy of each classifier
- Iterate the above steps until a complete set of high-accuracy classifiers are discovered for all classes of optical variability
- Provide classifications for all new optical transients discovered

Sample Problem #2: LSST Correlation Engine

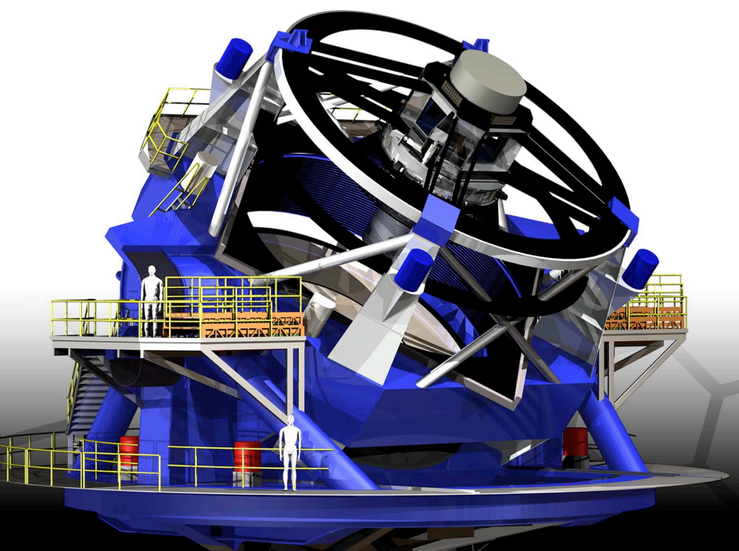
- Extract a major portion of the LSST Object Catalog (e.g., all stars, or all galaxies, or all Solar System objects)
- Develop fast algorithms to examine correlations using the 200+ science catalog database attributes—*N*-point, or in combinations of 2-at-a-time, 3-at-a-time, etc.—to discover new physically significant correlations
- Example: *N*-point correlation functions via *kd*-trees (Moore et al. 2001)
- Discover new classes of objects and/or new classes of behavior within a class of objects

Some goals of LSST gigascale data mining

- Provide rapid probabilistic classifications for 100,000 events each night
- Find new "fundamental planes" of correlated astrophysical parameters
- Compute multi-point multi-dimensional correlation functions over the full panoply of astrophysical parameter spaces
- Discover voids in interesting parameter spaces (*e.g.,* period gaps)
- Discover new properties of known classes
- Discover improved rules for classifying known classes of objects
- Identify novel, unexpected temporal behavior
- Hypothesis testing – verify existing (or generate new) astronomical hypotheses with high confidence, using millions of training samples
- Serendipity – discover rare one-in-a-billion objects or classes of objects

References

- *Source uncertainties:* Loredo, T. J. (2004), in *Maximum Entropy and Bayesian Methods*, AIP Conf. Series vol. 735
- *Multi-epoch object detection:* Babu et al. (2008), *Statistical Methodology*, vol. 5 (special issue on astrostatistics)
- *Temporal classification:* Ivezic et al. (2008), in *Classification and Discovery in Large Astronomical Surveys*, AIP Conf. Series vol. 1082
- *N-point correlations:* Moore et al. (2001), in *Mining the Sky*, astro-ph/0012333



January 2010

LSST is a public-private partnership. Design and development activity is supported in part by the National Science Foundation. Additional funding comes from private gifts, grants to universities, and in-kind support at Department of Energy laboratories and other LSSTC Institutional Members.