A taste of astrostatistics: problems, opportunities, & connections

Alexander W Blocker

2011 Sep 06

Outline

What is astrostatistics?

- Applying modern statistical tools to the problems of astronomy & astrophysics
- Combining scientific and statistical modeling
- Handling complex instrumental effects
- Linking theory to data

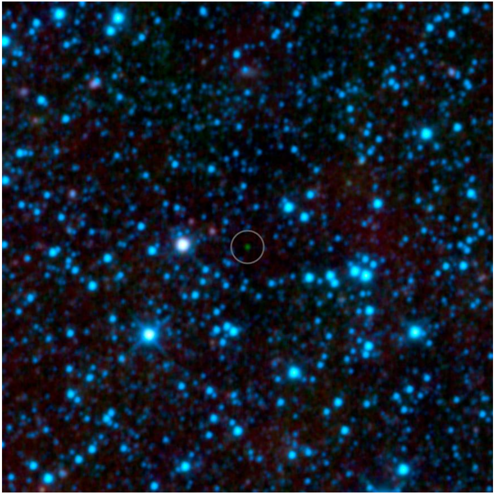
Where the data comes from

- Observations across the entire EM spectrum, from radio to gamma rays
- Optical observations mostly from ground-based telescopes (e.g. Keck); some from Hubble
- High-energy observations (x-rays, gamma, etc.) mostly from space telescopes (e.g. Chandra)
- Each type of data poses distinct challenges

Common types of data

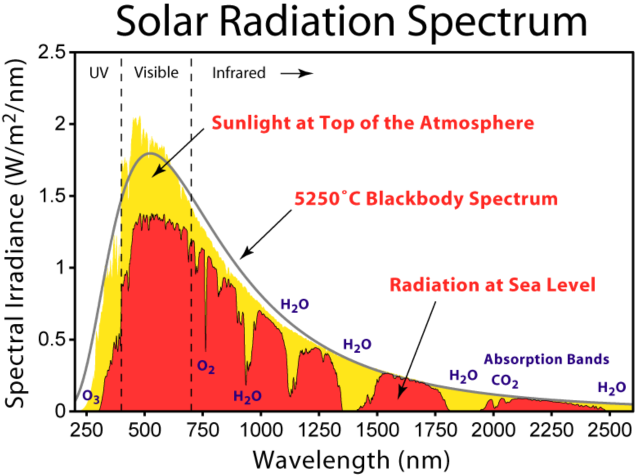
- Images
- Spectra
- Time series
- And all combinations of these

Example — Image

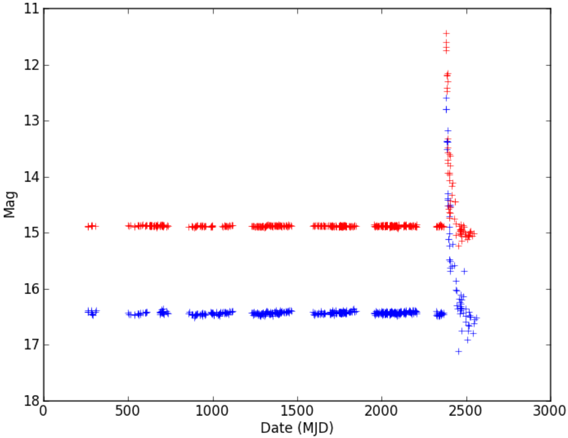


Cold brown-dwarf star from WISE satellite (WISE 1828+2650)

Example — Spectrum



Example — Time Series



Supernova SGR 2002 from EROS2 survey

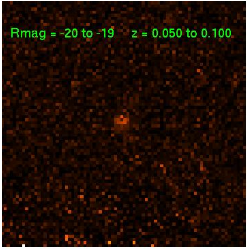
How do we approach problems?

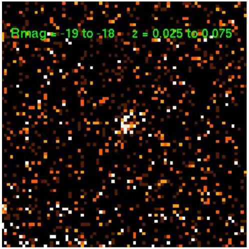
- Probabilistic models for the entire data-generating process
  - Account for instrumental effects, population variation, etc.
  - Framework for inference
- Approximate inference via computation
  - Typically MCMC
  - Can be EM or other methods
- Rigorous model checking & validation
  - Need to establish statistical & scientific validity
  - Value of collaboration — physical plausibility

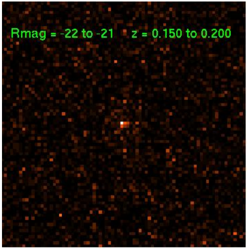
Combining information on faint x-ray sources

- Want to understand properties of a given population of sources (e.g. galaxies at a certain distance)
- For each source, we observe only two counts: one from the background noise & one from a combination of the source & noise
- Also have telescope's sensitivity etc. for given observation
- Goal is to combine information from these faint counts to estimate, e.g., mean intensity and variability in intensity among sources

Example images







Previous approaches in the astrophysics

- Idea: subtract out the background, then average resulting "net" counts
- Use of background subtraction $\Rightarrow$ Gaussian assumption; inappropriate in low count regimes
- Above manifests as negative individual estimates; for sufficiently faint samples, this can lead to negative aggregate estimates
- No clean measure of uncertainties on luminosities
- Solution: model data as Poisson

Assumptions

- Same as those for standard stacking analysis
- For luminosity-based inference, assuming that redshifts are known with no uncertainty
  - Relatively plausible for spectroscopic; not as much for photometric
- Assuming the spectra of sources are know & identical
  - Typically assume power law with photon index ≈ 1.7
- Attempting to make inferences only on selected sample, for now; not dealing with selection effects, etc.

Mathematical framework

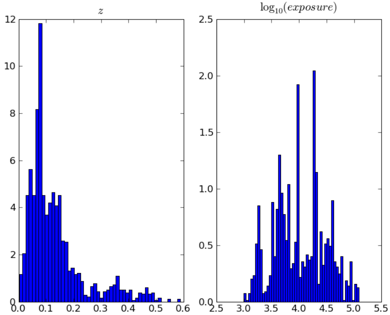
- Modeling source and background counts as Poisson.
- Assuming background count rates follow log-Normal distribution
- Assuming log-luminosities (or log-fluxes) follow a log-t distribution
    - Makes our inferences robust to outliers.
    - More appropriate for modeling distributions with power-law tails.
- Using priors to help regularize estimates; only require informative priors on dispersion parameters
- Need to allow for relationship between distance (redshift) & source intensity; analogous to using a general regression model

MCMC with finesse

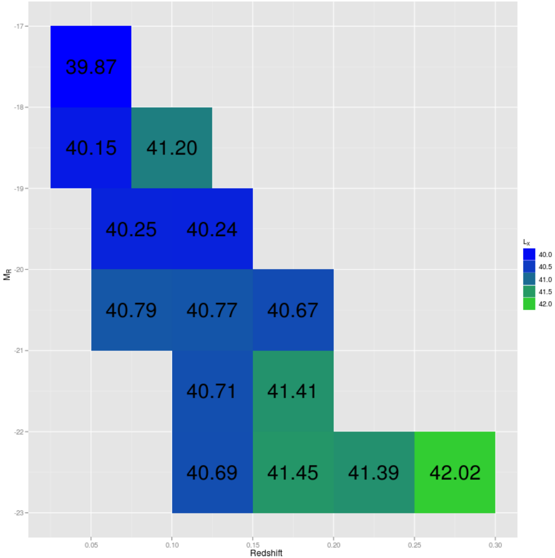
- Using MCMC to simulate from posterior of source intensities given prior & observations; can then extract estimands of interest
- Because our model is Poisson / log-t, we can't use a standard Gibbs sampler
- Combining independence chain MH, parameter expansion, and data augmentation strategies to obtain an efficient sampler
- Using numerical optimization (Halley's method, appropriately) to build good proposal distributions

Description of data

- We worked with 1546 galaxies from SDSS on which we have spectroscopic redshifts.
- The distribution of redshift and exposure for these sources can be seen below.



Summary of results on SDSS data



Finding events in time series — lots of time series

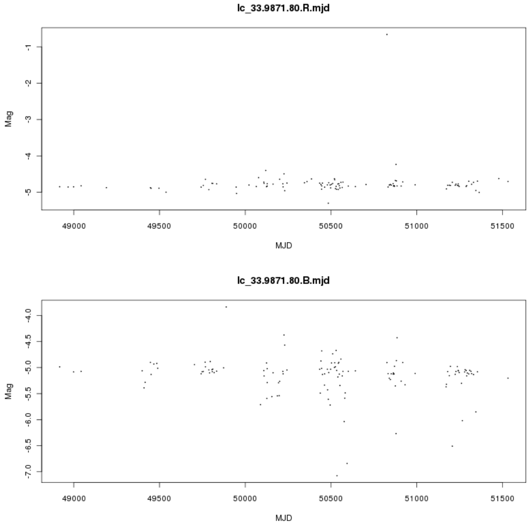
- Have massive (order of 10-100 million) dataset of time series, possibly spanning multiple spectral bands
- Goal is to identify and classify time series containing events
- How do we define an event?
  - Not interested in isolated outliers
  - Looking for groups of observations that differ significantly from those nearby (ie, "bumps" and "spikes")
  - Also attempting to distinguish periodic and quasi-periodic time series from isolated events

The data

- We used data from the MACHO survey for training, and are actively analyzing the EROS2 survey
- MACHO data consists of approx. 38 million LMC sources, each observed in two spectral bands
  - Collected 1992-1999 on 50-inch telescope at Mount Stromlo Observatory, Australia
  - Imaged 94 43x 43 fields in two bands, using eight 2048 x 2048 pixel CCDs
  - Substantial gaps in observations due to seasonality and priorities
- EROS2 data consists on approx. 87.2 million sources, each observed in two spectral bands
  - Imaged with 1m telescope at ESO, La Silla between 1996 and 2003
  - Each camera consisted of mosaic of eight 2K x 2K LORAL CCDs
  - Typically 800-1000 observations per source

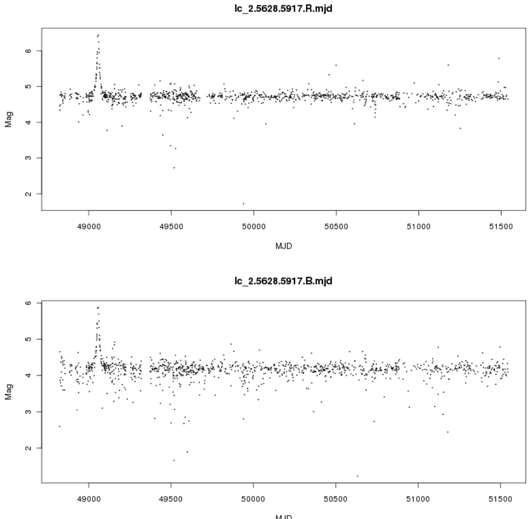
Exemplar time series from the MACHO project:

A null time series:



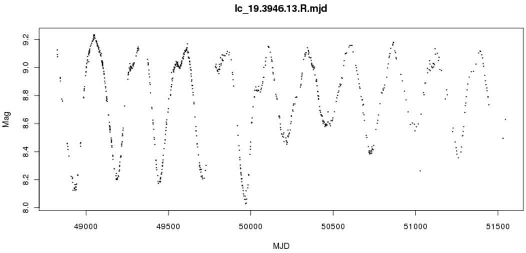
Exemplar time series from the MACHO project:

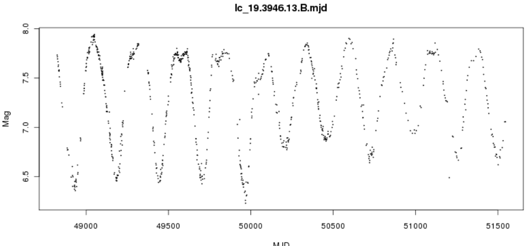
An isolated event (microlensing):



Exemplar time series from the MACHO project:

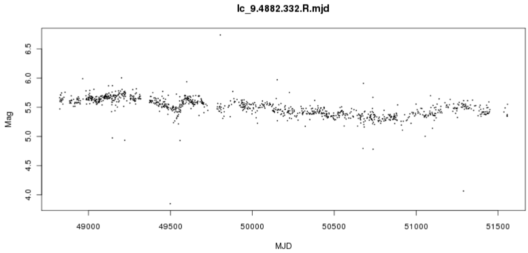
A quasi-periodic time series (LPV):

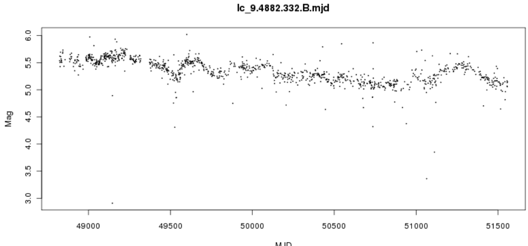




Exemplar time series from the MACHO project:

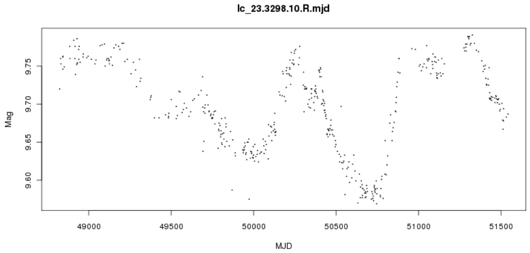
A variable time series (quasar):

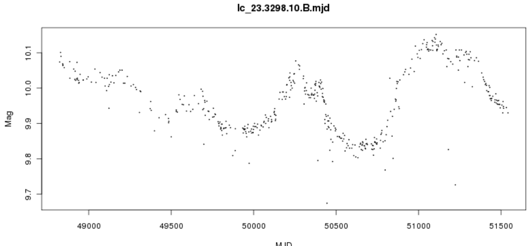




Exemplar time series from the MACHO project:

A variable time series (blue star):





Notable properties of this data

- Fat-tailed measurement errors
  - Common in astronomical data, especially from ground-based telescopes
  - Need more sophisticated models for the data than standard Gaussian approaches
- Quasi-periodic and other variable sources
  - Changes the problem from binary classification (null vs. event) to k-class
  - Need more complex test statistics and classification techniques
- Non-linear, low-frequency trends make less sophisticated approaches far less effective
- Irregular sampling can create artificial events in naive analyses

Our approach

- Use a Bayesian probability model for both initial detection and to reduce the dimensionality of our data (by retaining posterior summaries)
- Using posterior summaries as features for machine learning classification technique to differentiate between events & variables
- Our goal is **not** to perform a final, definitive analysis on these events
  - Objective to predict which time series are most likely to yield phenomena characterized by events (e.g. microlensing, blue stars, flares, etc.)
  - Allows for use of complex, physically-motivated methods on massive datasets by pruning set of inputs to manageable size
  - Provides assessments of uncertainties at each stage of screening and allows for the incorporation of domain knowledge

Summarized mathematically

- Symbolically, let V be the set of all time series with variation at an interesting scale (e.g., the range of lengths for events), and let E be the set of events
- For a given time series Y_i, we are interested in $P(Y_i \in E)$
- We decompose this probability as
  $$P(Y_i \in E) \propto P(Y_i \in V) \cdot P(Y_i \in E | Y_i \in V)$$
  via the above two steps

Probability model - specification

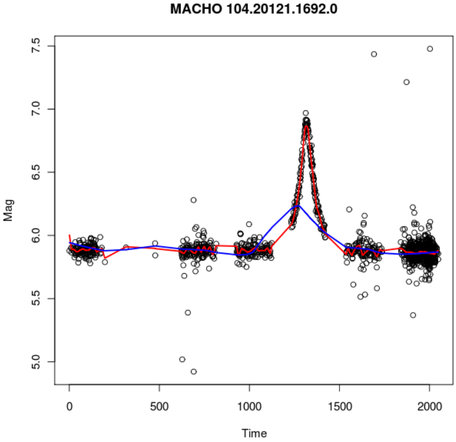
- Linear model for each time series with an incomplete wavelet basis:

$$y(t) = \sum_{i=1}^{k_l} \beta_i \phi_i(t) + \sum_{j=k_l+1}^{M} \beta_j \phi_j(t) + \epsilon(t)$$

- First k_l elements contain low-frequency, “trend” components; remainder contain frequencies of interest; highest frequencies are left as noise
- Idea: compare smooth (trend-only) and complete model fits; if they differ, could have an event
- Assume residuals $\epsilon(t)$ are distributed as iid $t_\nu(0, \sigma^2)$ for robustness ($\nu = 5$) — fat tails
- Address irregular sampling through regularization — informative priors on wavelet coefficients smooth undersampled periods
- Extremely fast estimation via EM — $\approx 0.15 - 0.2$ seconds

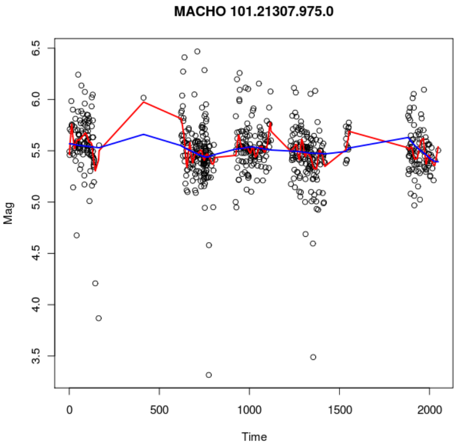
Examples of model fit

Idea is that, if there is an event at the scale of interest, trend-only and complete fits with differ substantially:



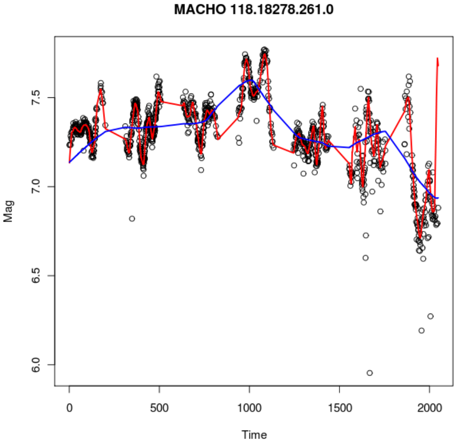
Example of model fit

For null time series, the difference will be small:



Example of model fit

However, for quasi-periodic time series, the difference will be huge:



Probability model - testing

$$y(t) = \sum_{i=1}^{k_l} \beta_i \phi_i(t) + \sum_{j=k_l+1}^{M} \beta_j \phi_j(t) + \epsilon(t)$$

- Using LLR statistic to test if coefficients on all non-trend components are zero ($H_0 : \beta_{k_l+1} = \beta_{k_l+2} = \ldots = \beta_M = 0$)
- Controlling false discovery rate (FDR) to 10^{-4} to set the critical region for our test statistic

Feature Selection I

- Engineered two features based on fitted values for discrimination between diffuse and isolated variability
- First is a relatively conventional CUSUM statistic
- Let $\{z_t\}$ be the normalized fitted values for a given time series, excepting the "trend" components corresponding to $\beta_1, \ldots, \beta_{k_l}$. We then define:

$$S_t = \sum_{k=1}^{t} (z_k^2 - 1)$$

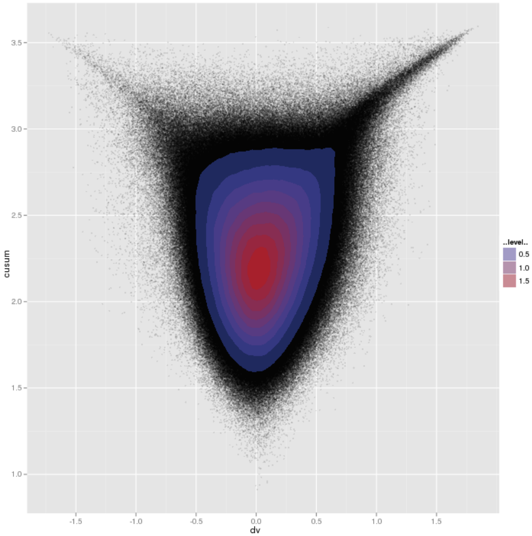
$$CUSUM = \max_t S_t - \min_t S_t$$

Feature Selection II

- Second is "directed variation"
  - Idea is to capture deviation from symmetric, periodic variation
  - Defining z_t as before and letting z_{med} be the median of z_t, we define:

$$DV = \frac{1}{\#\{t : z_t > z_{\mathrm{med}}\}} \sum_{t : z_t > z_{\mathrm{med}}} z_t^2 - \frac{1}{\#\{t : z_t < z_{\mathrm{med}}\}} \sum_{t : z_t < z_{\mathrm{med}}} z_t^2$$

Distribution of features on MACHO data



Methods

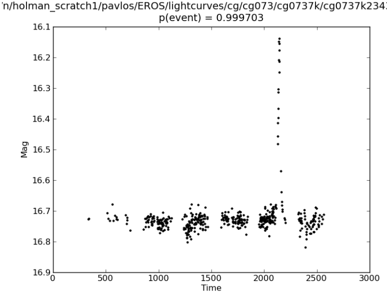
- Tested a wide variety of classifiers on our training data, including kNN, SVM (with radial and linear kernels), LDA, QDA, and others
- Regularized logistic regression performs best
- Using weakly informative (Cauchy) prior for regularization

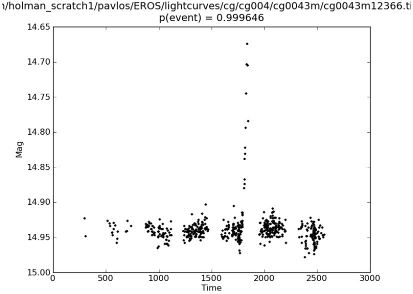
Summary

- First stage shows reduction from 87.2 million candidate light curves by approximately 98% (to approximately 1.5 million) in blue band from likelihood-ratio screen
- Approximately 16,000 of the latter group are likely isolated events, based on analysis from classification stage and filtering for chip-level errors (265 with $P(\text{event}) \geq 0.80$ in both bands)
- Scientific follow-up on candidates yielded identified 126 known gravitational lensings and 42 known supernovae (via Simbad & VizieR)
- Several candidates identified for further analysis in multiple categories

Examples of highly-ranked events

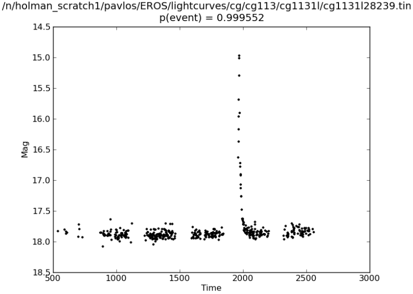
Examples from top 10:

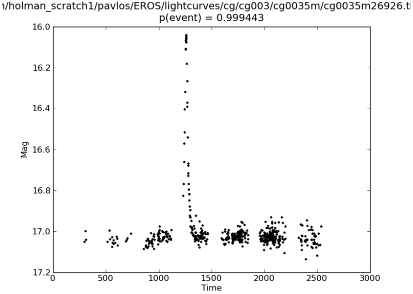




Examples of highly-ranked events

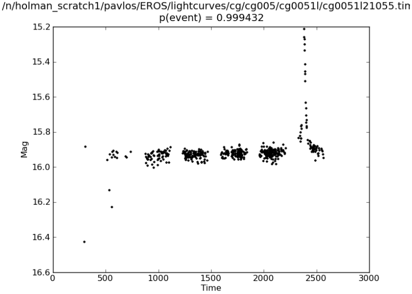
Examples from top 10:

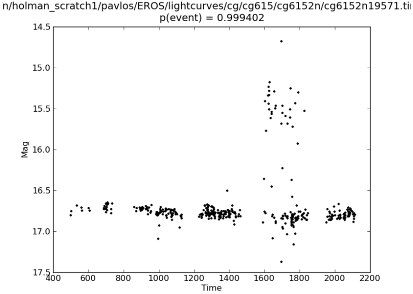




Examples of highly-ranked events

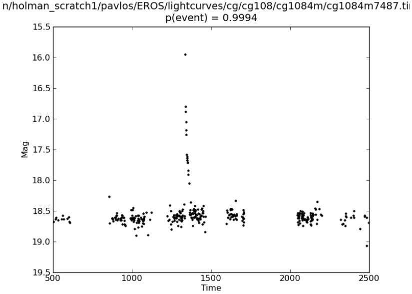
Examples from top 10:

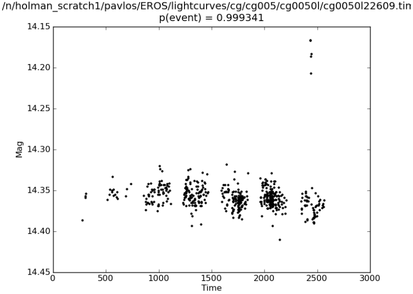




Examples of highly-ranked events

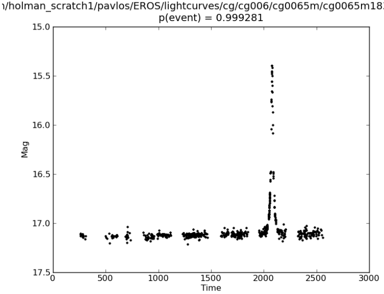
Examples from top 10:

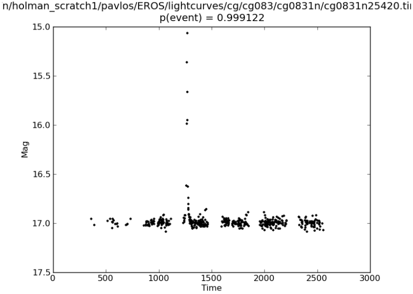




Examples of highly-ranked events

Examples from top 10:





Lessons from event detection

- Massive data presents a new set of challenges to statisticians & astronomers that many of our standard tools are not well-suited to address
- Machine learning has some valuable ideas and methods to offer, but we should not discard the power of probability modeling
- Conversely, we can use reasonable probability models with massive datasets without excessive computational burdens
- It is tremendously important to put each tool in its proper place for these types of analyses
  - Rigorous modeling of observation processes is particularly crucial; mistakes here can destroy information for any later analyses
- Our work on event detection for astronomical data shows the power of this approach by combining both rigorous probability models and standard machine learning approaches

Astrophysics & biology

- These subjects appear extremely dissimilar on the surface
- However, they are following a similar path in terms of data, as both address:
  - An increasing need to address complex instrumental/experimental properties
  - A transition to regimes where non-Gaussian error distributions matter
  - An explosion in the volume of data
- The largest difference is that the astrostatistics community has been facing these problems & building high-quality solutions for longer

Complex instrumental & experimental properties

- Astronomers face extremely complex instrumental & experimental properties:
  - Inhomogeneous sensitivity (sources look dimmer or brighter depending on where the telescope sees them)
  - Blurring due to detector/telescope properties
  - Subtle, non-ignorable patterns of missing data
- All of these are increasingly important in biology as high-throughput methods become more common
- Physical mechanisms differ between fields, but statistical challenges are analogous

Non-Gaussian errors

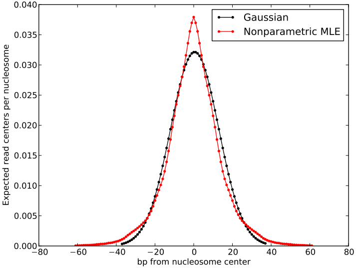
- Optical astronomy dealt almost exclusively with Gaussian errors
- With high-energy observations (x-ray, gamma, etc.), observations are counts, so errors can be extremely non-Gaussian
- We deal with these problems constantly in astrostatistics and have built a methodological foundation to address them
- High-throughput biology must address these problems as, e.g., sequencing (counts) replace micro-arrays (continuous) for analyses of gene expression

Methodological arbitrage

- Both astrophysics & biology have many more problems than statisticians; many opportunities
- Astronomy can be an excellent setting to address these problems
  - Field emphasizes pinning-down & understanding sources of error
  - Have data available to model & analyze complex observation processes
  - Direct physical underpinnings allow us to focus on the core problems

Example — telescopes & enzymes

- In astrophysics, need to account for blurring of observations due to instrument
- Typically handled via PSF (point spread function), which describes distribution of observations given location of source
- Exactly analogous phenomenon occurs with high-throughput sequencing in biology due to enzymatic digestion
- Methods from astrostatistics formed basis for solution to biological problem



Wrapping up

- Astrostatistics is a vibrant, exciting area for research
- Plenty of open problems
- Challenges from foundational theory to computation
- Major opportunity to apply methods across multiple fields
- And, of course, great collaborators

Acknowledgments

Many thanks to

- Xiao-Li Meng & Edo Airoldi, my advisors
- Tom Aldcroft & Paul Green, my collaborators on stacking
- Pavlos Protopapas, Dae-Won Kim, & Jean-Baptiste Marquette, my collaborators on event detection