Abstract

The detection and analysis of events within massive collections of time-series has become an extremely important task for time-domain astronomy. In particular, many scientific investigations (e.g. the analysis of microlensing and other transients) begin with the detection of isolated events in irregularly-sampled series with both non-linear trends and non-Gaussian noise. I will discuss a semi-parametric, robust, parallel method for identifying variability and isolated events at multiple scales in the presence of the above complications. This approach harnesses the power of Bayesian modeling while maintaining much of the speed and scalability of more ad-hoc machine learning approaches. I will also contrast this work with event detection methods from other fields, highlighting the unique challenges posed by astronomical surveys. Finally, I will present initial results from the application of this method to 87.2 million EROS sources, where we have obtained a greater than 100-fold reduction in candidates for certain types of phenomena.

Semi-parametric Robust Event Detection for Massive Time-Series Datasets

Alexander W Blocker

25 August, 2010

Outline

What is massive data?

- In short, it's data where our favorite methods stop working
- Orders of magnitude more observations than we are used to dealing with, often combined with high dimensionality (e.g. 40 million time series with thousands observations each)
- Such scale of data is increasingly common in fields such as astronomy, computational biology, ecology, etc.
- Need statistical methods that scale to these quantities of data
- However, need to tradeoff statistical rigor and computational efficiency

Machine Learning vs. Statistics, in broad strokes

Statistical Methods

- Heavy computational burden
- Highly customizable
- Can handle "messy" data
- Internal assessment of uncertainty

Machine Learning Methods

- Computationally efficient
- Generically applicable
- Need clean input
- External assessment of uncertainty

How can we get the best of both worlds?

- Principled statistical methods are best for handling messy, complex data that we can effectively model, but scale poorly to massive datasets
- Machine learning methods handle clean data well, but choke on issues we often confront (outliers, nonlinear trends, irregular sampling, unusual dependence structures, etc.)
- Idea: Inject probability modeling into our analysis in the right places

The problem

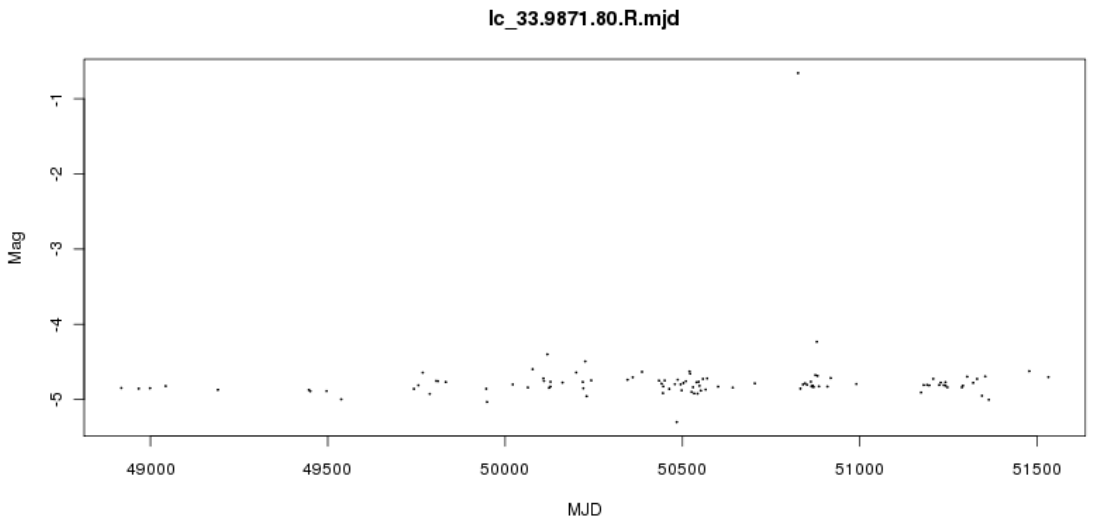
- Have massive (order of 10-100 million) dataset of time series, possibly spanning multiple spectral bands
- Goal is to identify and classify time series containing events
- How do we define an event?
  - Not interested in isolated outliers
  - Looking for groups of observations that differ significantly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

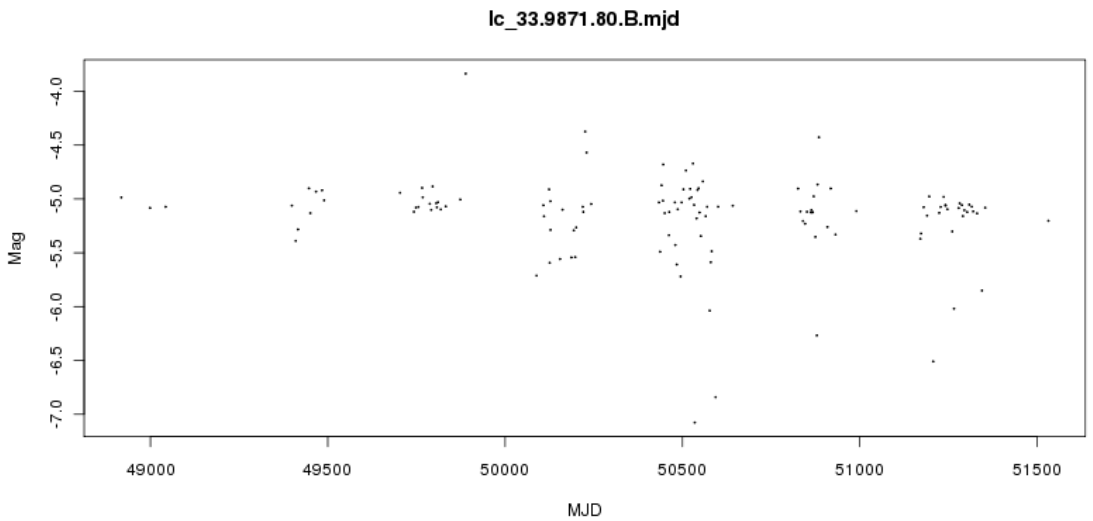
The data

- 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:

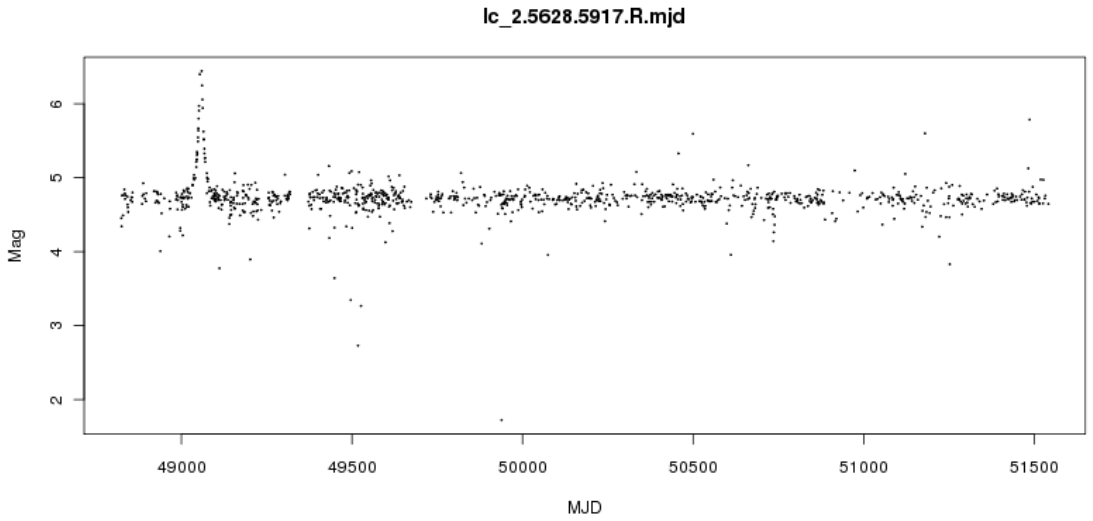
A null time series:

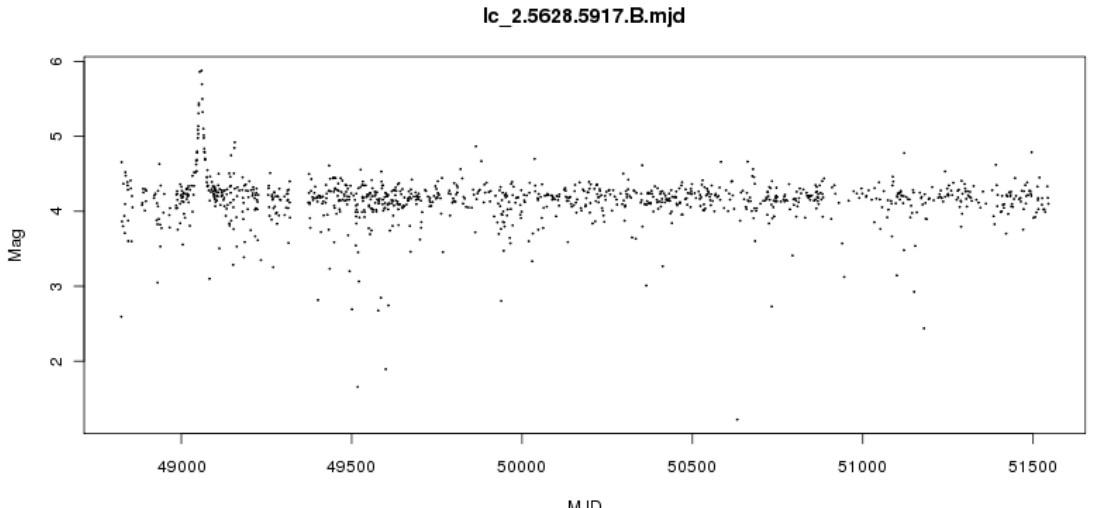




Exemplar time series from the MACHO project:

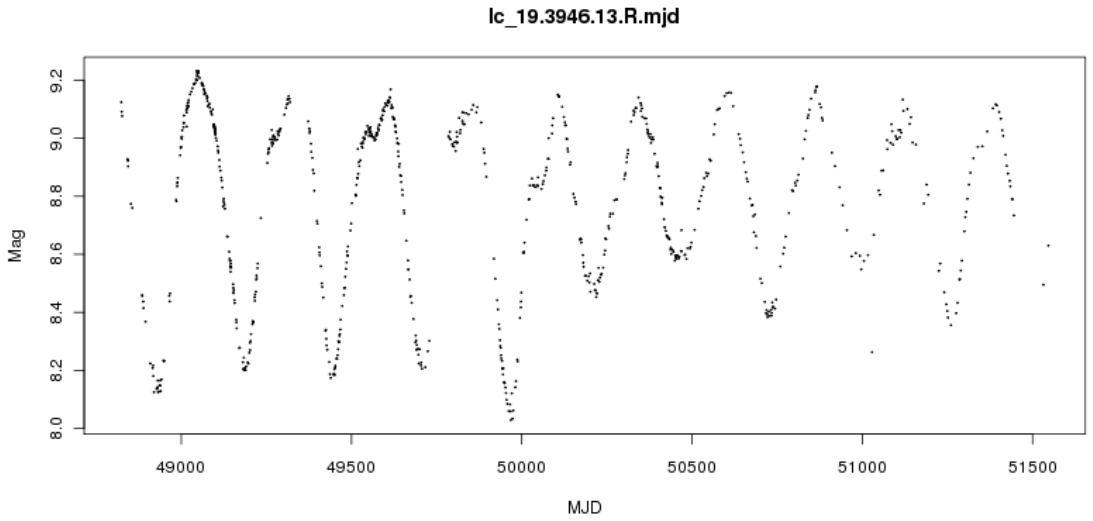
An isolated event (microlensing):

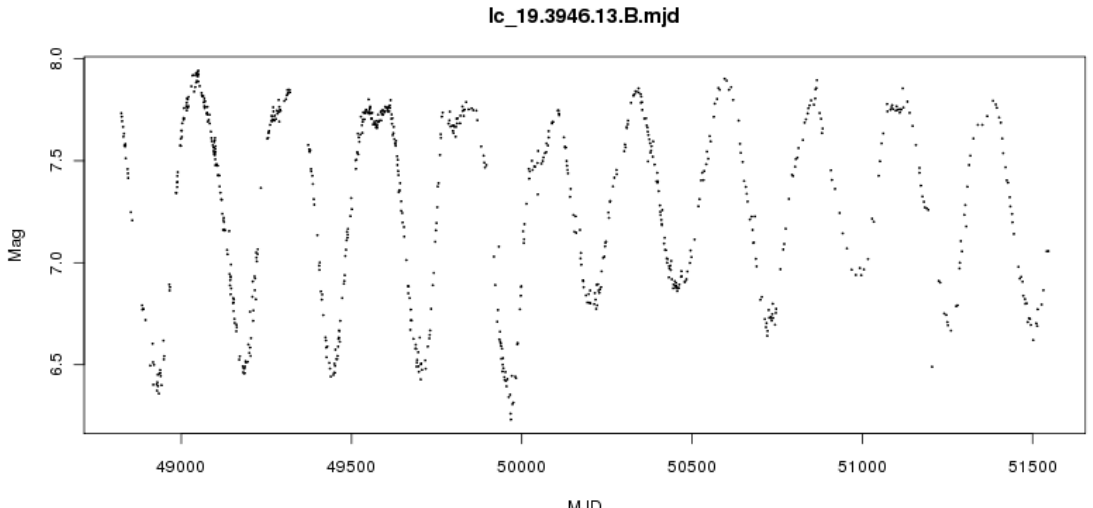




Exemplar time series from the MACHO project:

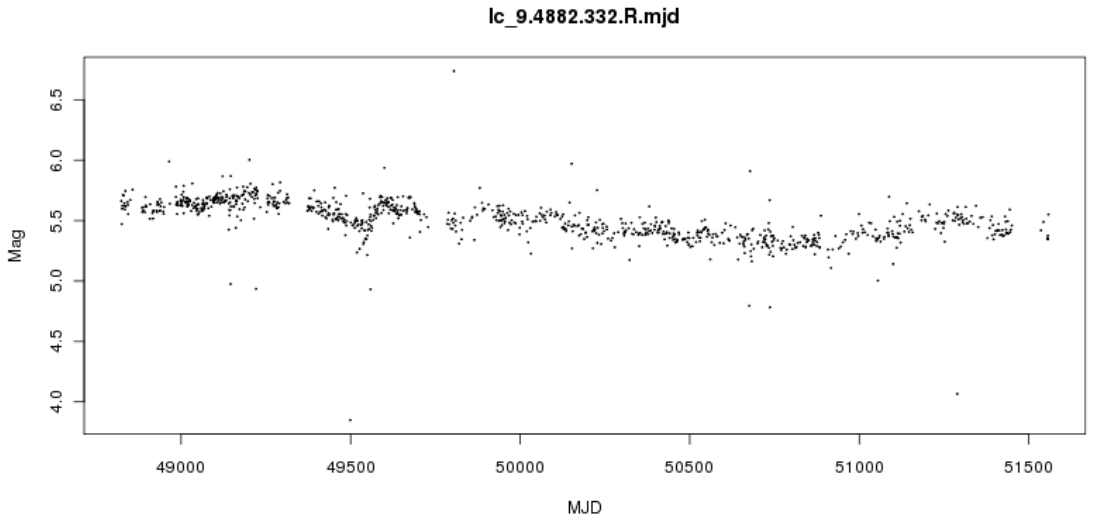
A quasi-periodic time series (LPV):

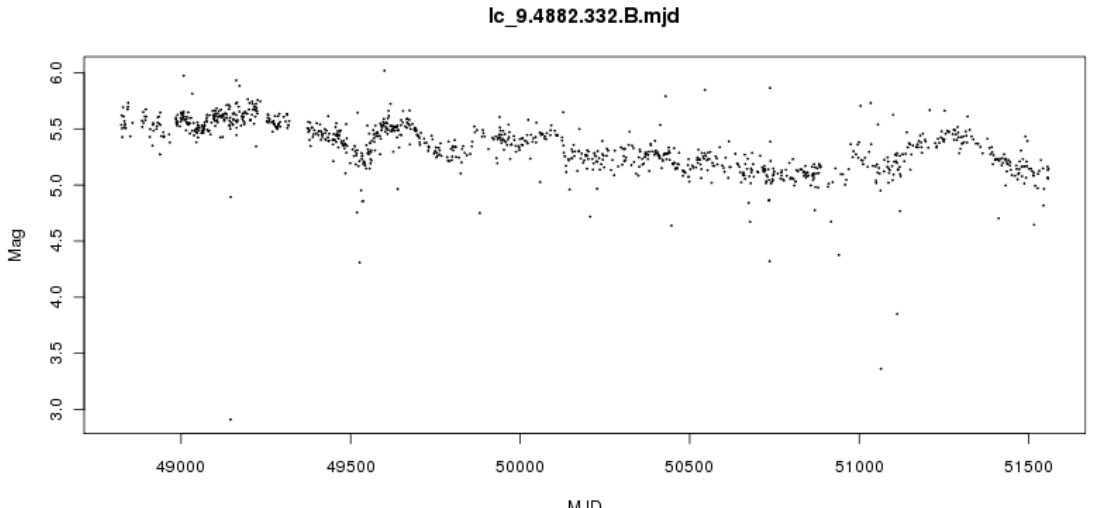




Exemplar time series from the MACHO project:

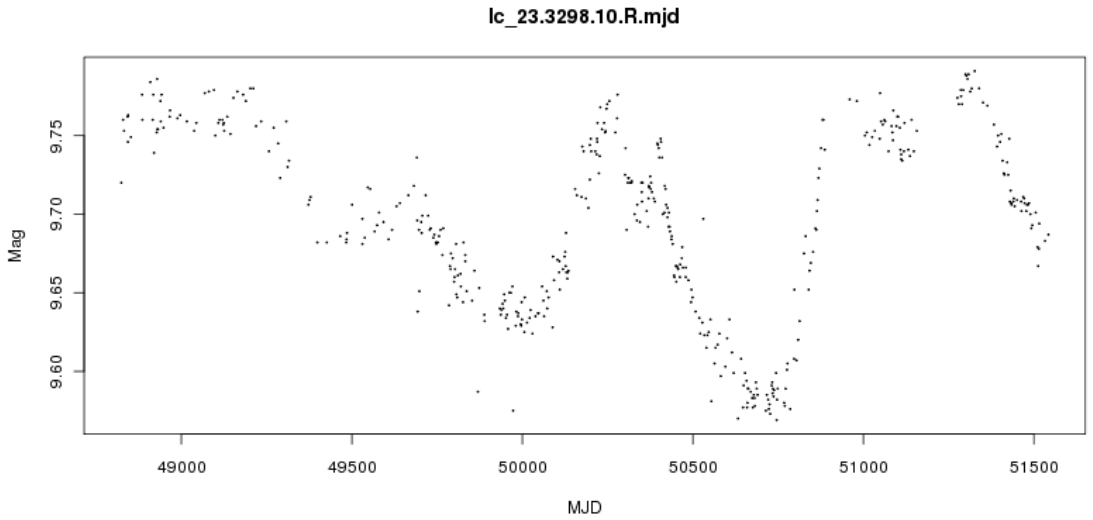
A variable time series (quasar):

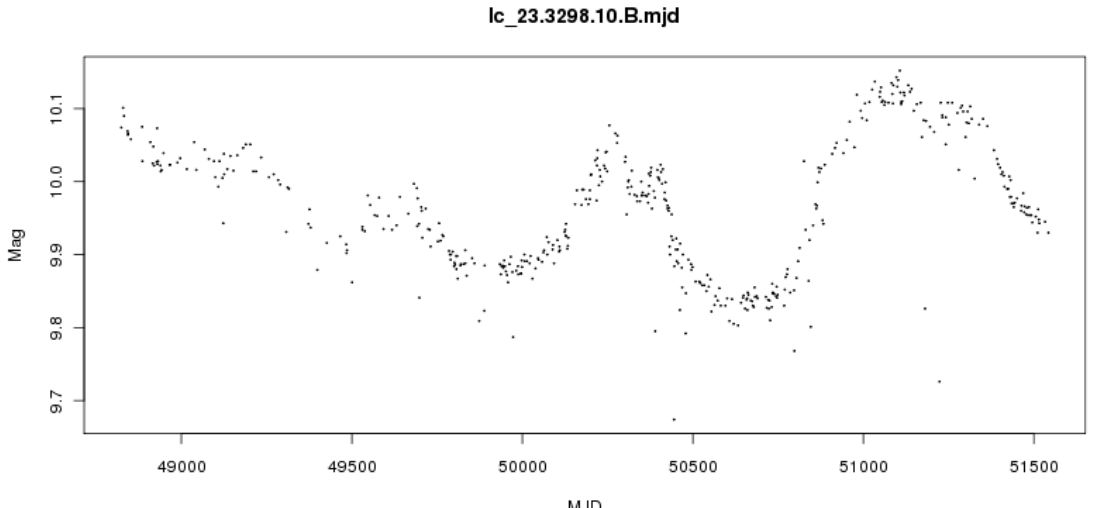




Exemplar time series from the MACHO project:

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 naïve analyses

From astronomy

- Scan statistics are a common approach (Liang et al, 2004; Preston & Protopapas, 2009)
  - However, they often discard data by working with ranks and account for neither trends nor irregular sampling
- Equivalent width methods (a scan statistic based upon local deviations) are common in astrophysics
  - However, these rely upon Gaussian assumptions and crude multiple testing corrections
- Numerous other approaches have been proposed in the literature, but virtually all rely upon Gaussian distributional assumptions, stationarity, and (usually) regular sampling

Statistical/ML approaches

- Symth (2007,2008,2010) has used hidden Markov models to model deviations from learned baselines in sensor count data
- Long history of work in change-point / regime-switching problems within statistics and econometrics
  - For example, Bayesian lines of research going from Smith (1975) through Raftery & Akman (1986) and Carlin, Gelfand, and Smith (1992)
  - On the econometrics side, Andrews (1993) and more recent work by Perron & collaborators (1998, 2003)
- However, our setting is quite distinct from those typically seen in previous work

Differences in astronomical/massive data setting

- Most preceding work has dealt with single time series which provide internal replication for analyzing deviations from "typical" behavior
- In analyzing massive time-domain surveys, we are confronted with large sets of time series that are less informative individually
- We must rely on replication across series and prior scientific knowledge to find deviations from typical behavior

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 a split wavelet basis:

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

- Assume that our residuals $\epsilon(t)$ are distributed as iid $t_\nu(0, \sigma^2)$ random variables to account for extreme residuals ($\nu = 5$)
- Using a Symmlet 4 (aka Least Asymmetric Daubechies 4) wavelet basis
- $(\phi_1, \ldots, \phi_{k_l})$ contains the low-frequency components of a wavelet basis, and $(\phi_{k_l+1}, \ldots, \phi_M)$ contains the mid-frequency components
- For a basis on $(1, 2048)$, we set k_l to 8 and M to 128

Probability model - specification

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

- Idea: $(\phi_1, \ldots, \phi_{k_l})$ will model structure due to trends, and $(\phi_{k_l+1}, \ldots, \phi_M)$ will model structure at the scales of interest for events
- Explicitly accounting for irregular sampling in our time series through this basis formulation
- Placing independent Gaussian priors on all coefficients except for the intercept
  - Next major refinement is to perform empirical Bayesian fitting on a random subsample of time series to set these

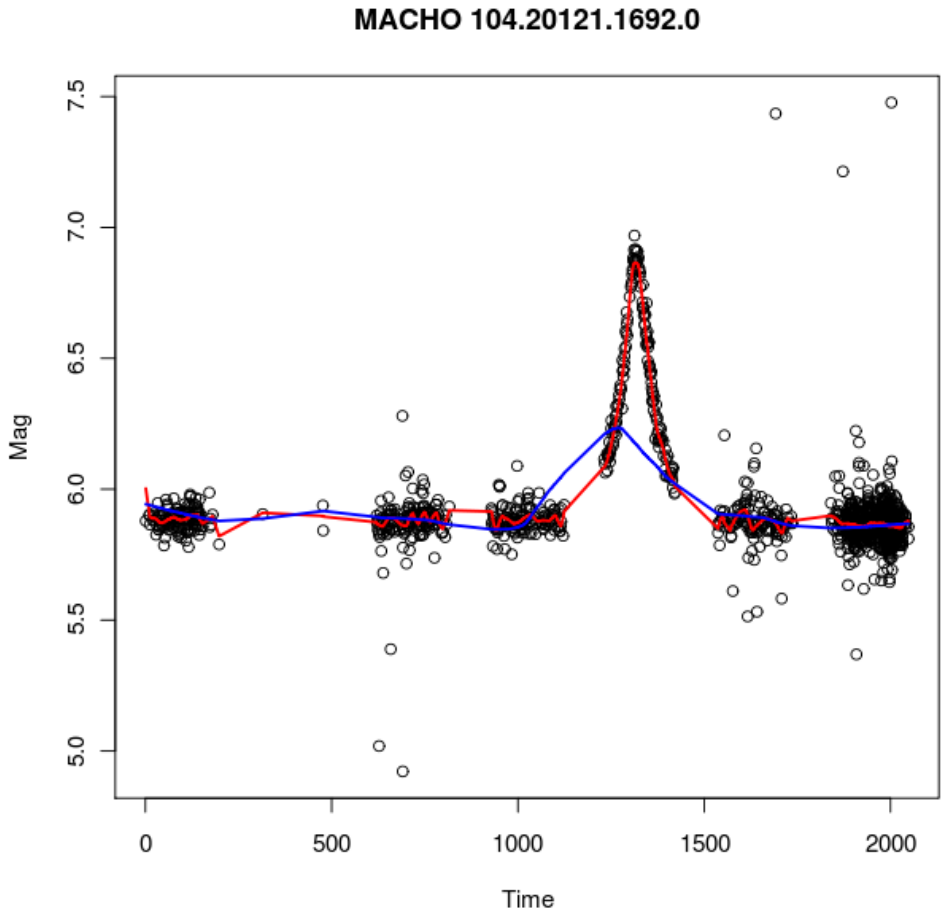
Probability model - estimation

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

- Using EM algorithm with optimal data augmentation scheme of Meng & Van Dyk (1997) to obtain MAP estimates of our parameters
- Implemented procedure in C with direct BLAS/LAPACK interface
- Average time for a full estimation procedure is $\approx 0.15 - 0.2$ seconds including file I/O on the Odyssey cluster

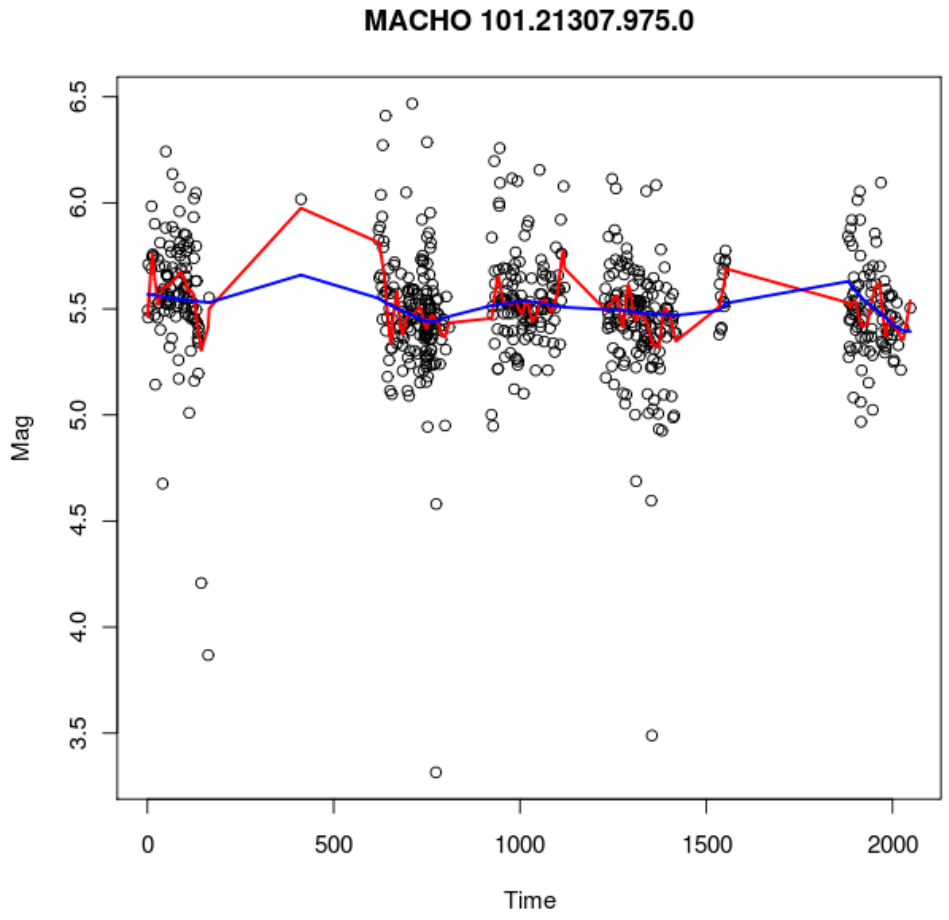
Examples of model fit

Idea is that, if there is an event at the scale of interest, there will be a large discrepancy between fits using $(\phi_1, \ldots, \phi_{k_l})$ vs. entire basis:



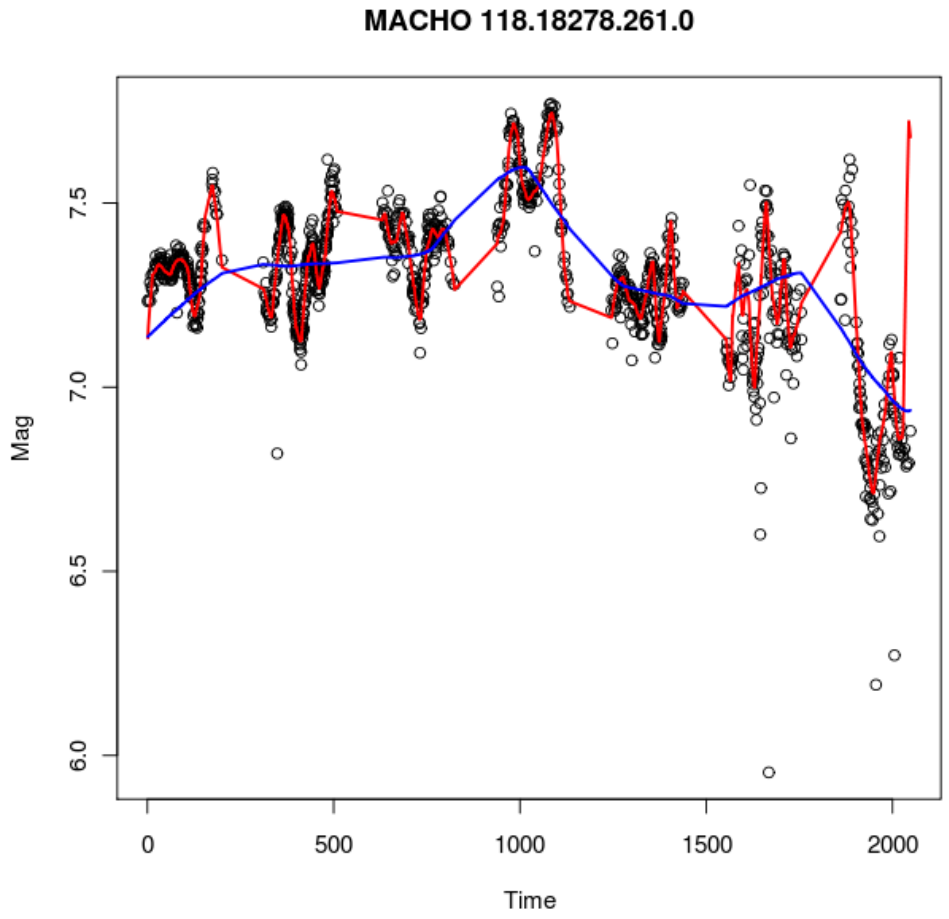
Example of model fit

For null time series, the discrepancy will be small:



Example of model fit

And for quasi-periodic time series, the discrepancy will be huge:



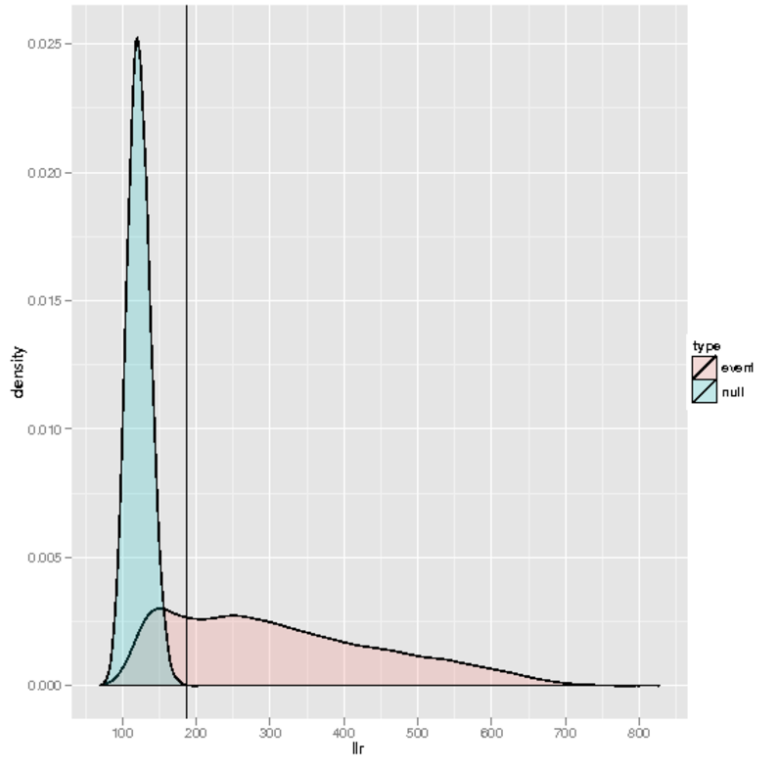
Probability model - testing

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

- We screen time series for further examination by testing $H_0 : \beta_{k_l+1} = \beta_{k_l+2} = \ldots = \beta_M = 0$
- Test statistic is $2(\hat{\ell}_1 - \hat{\ell}_0)$
- Using modified Benjamini-Hochberg FDR procedure with a maximum FDR of 10^{-4} to set the critical region for our test statistic

Distribution of LLR statistic

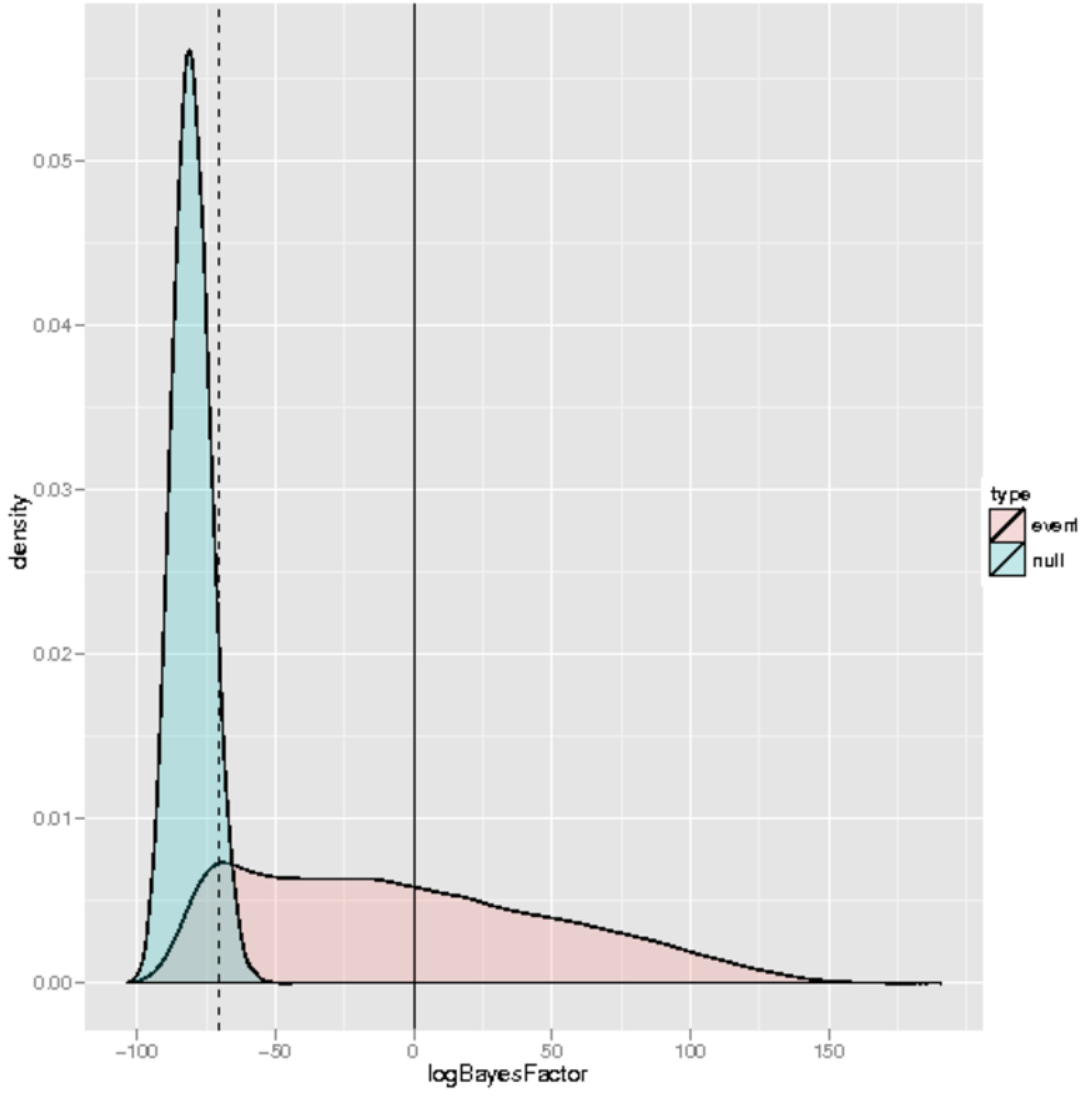
- To assess how well this statistic performs, we simulated $50,000$ events from a physics-based model and $50,000$ null time series
- We obtained approximate power of 80% with the stated FDR based on this simulated data



A sidenote: Why not use a Bayes factor?

- Given our use of Bayesian models, a Bayes factor would appear to be a natural approach for the given testing problem
- Unfortunately, these do not work well with "priors of convenience", such as our independent Gaussian prior on the wavelet coefficients
- Because of these issues, the Bayes factor was extremely conservative in this problem for almost any reasonable prior

Distribution of Bayes factor for simulated data



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:

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

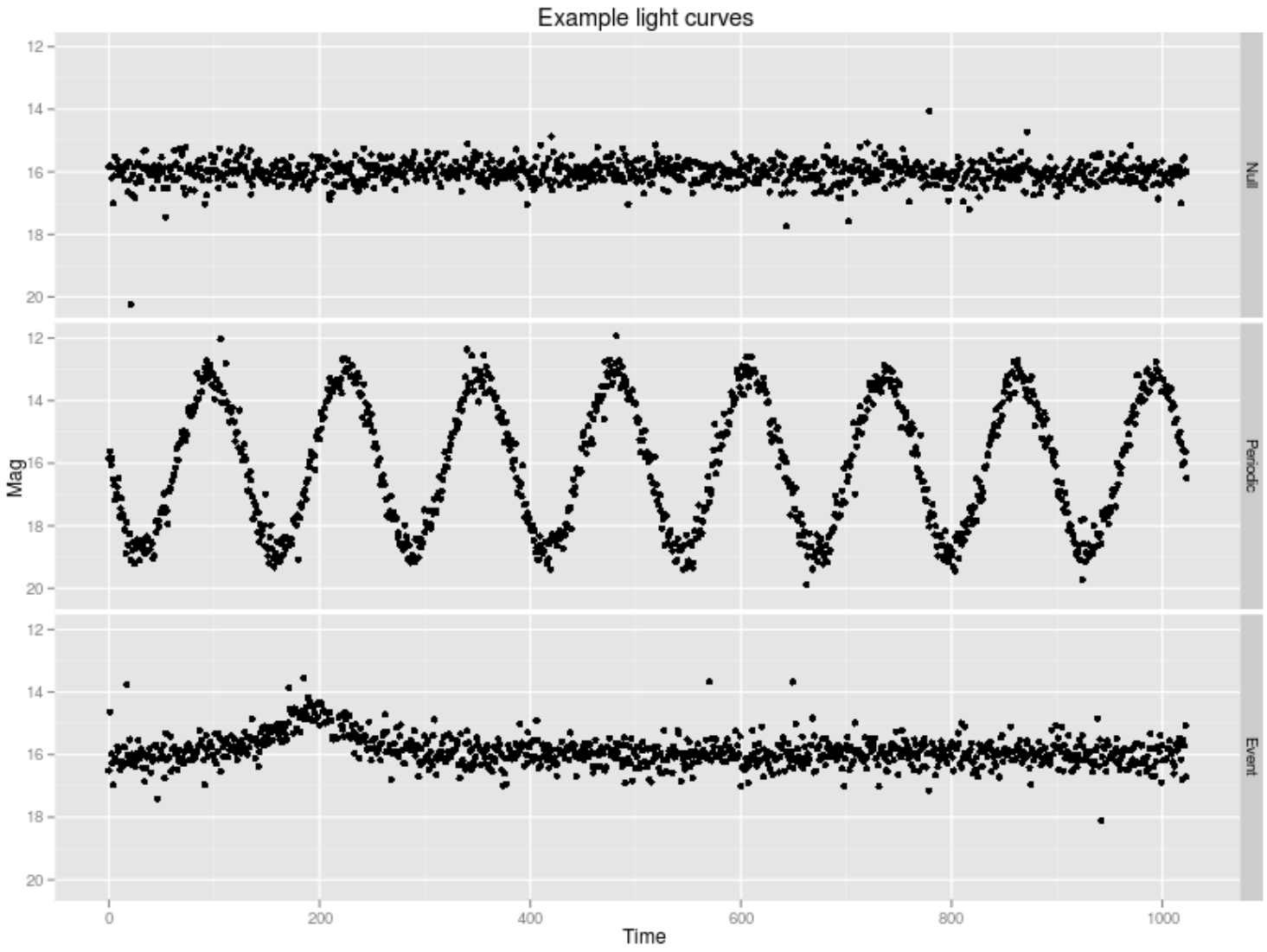
$$CUSUM = \max_t S_t - \min_t S_t$$

Feature Selection II

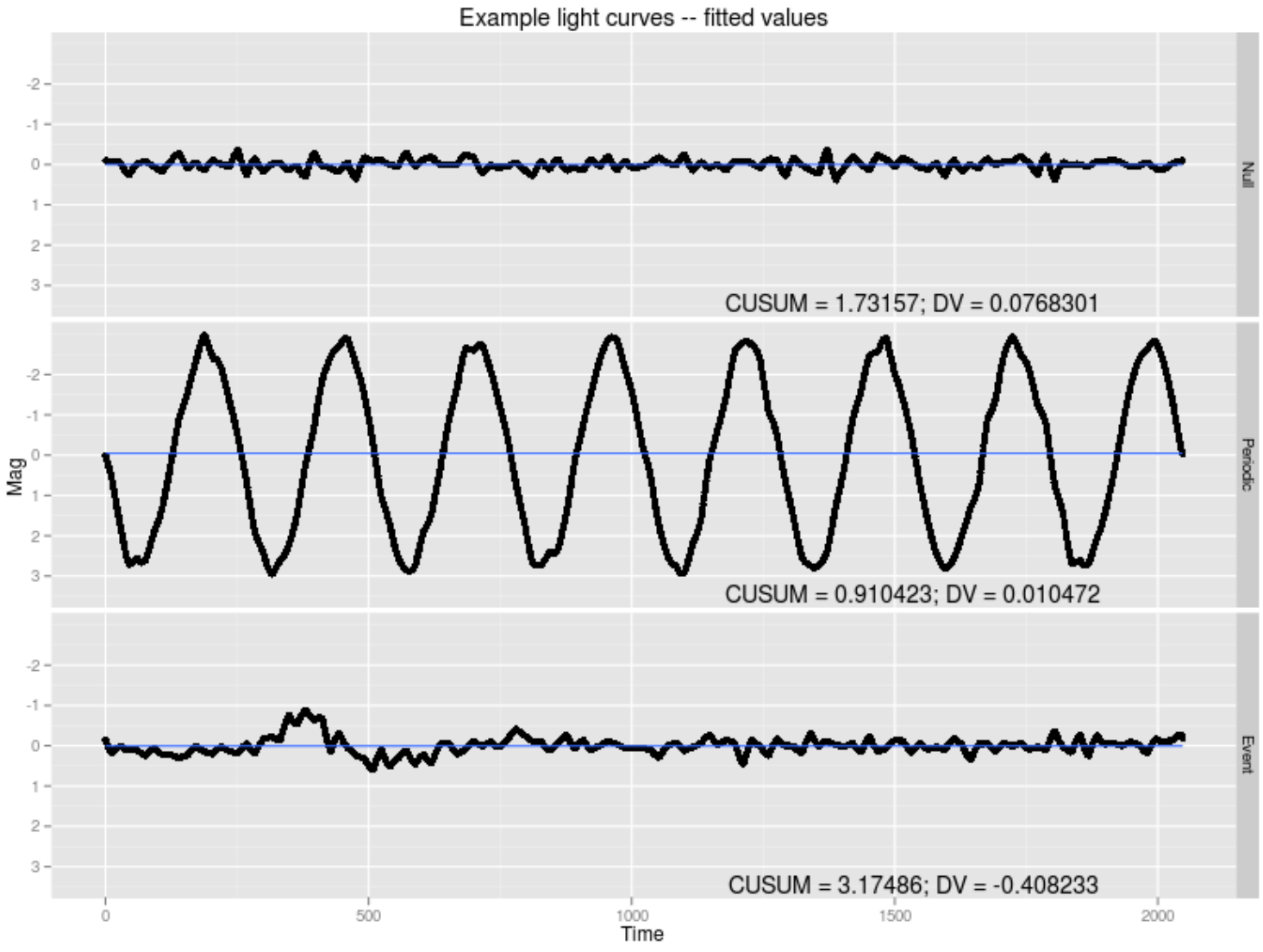
- Second is "directed variation"
  - Idea is to capture deviation from symmetric, variation
  - Defining z_t as before and letting z_{med} be the median of z_t, we define:

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

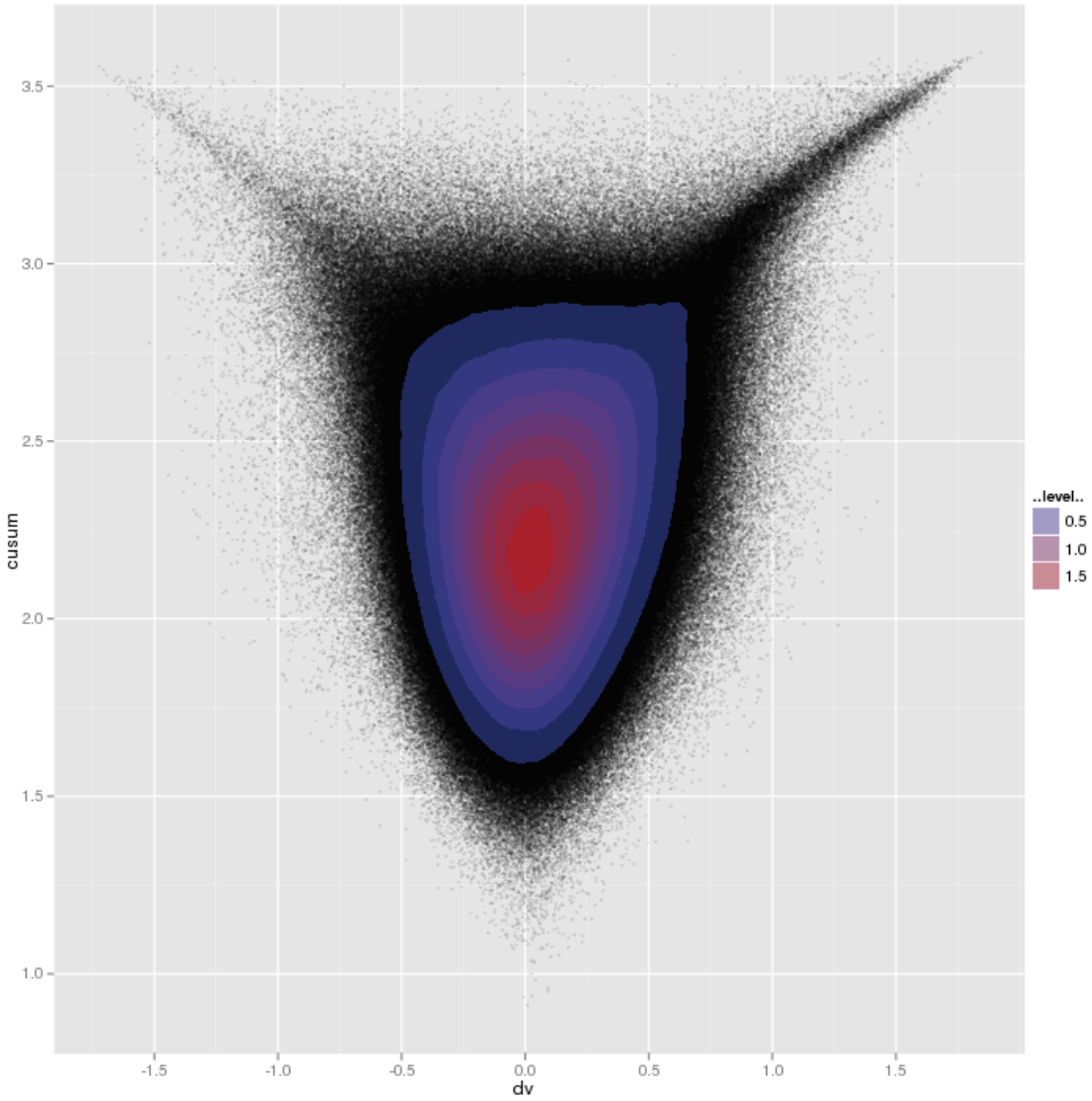
Examples


Example light curves
Mag
Time
Null
Periodic
Event
12
14
16
18
20
0
200
400
600
800
1000

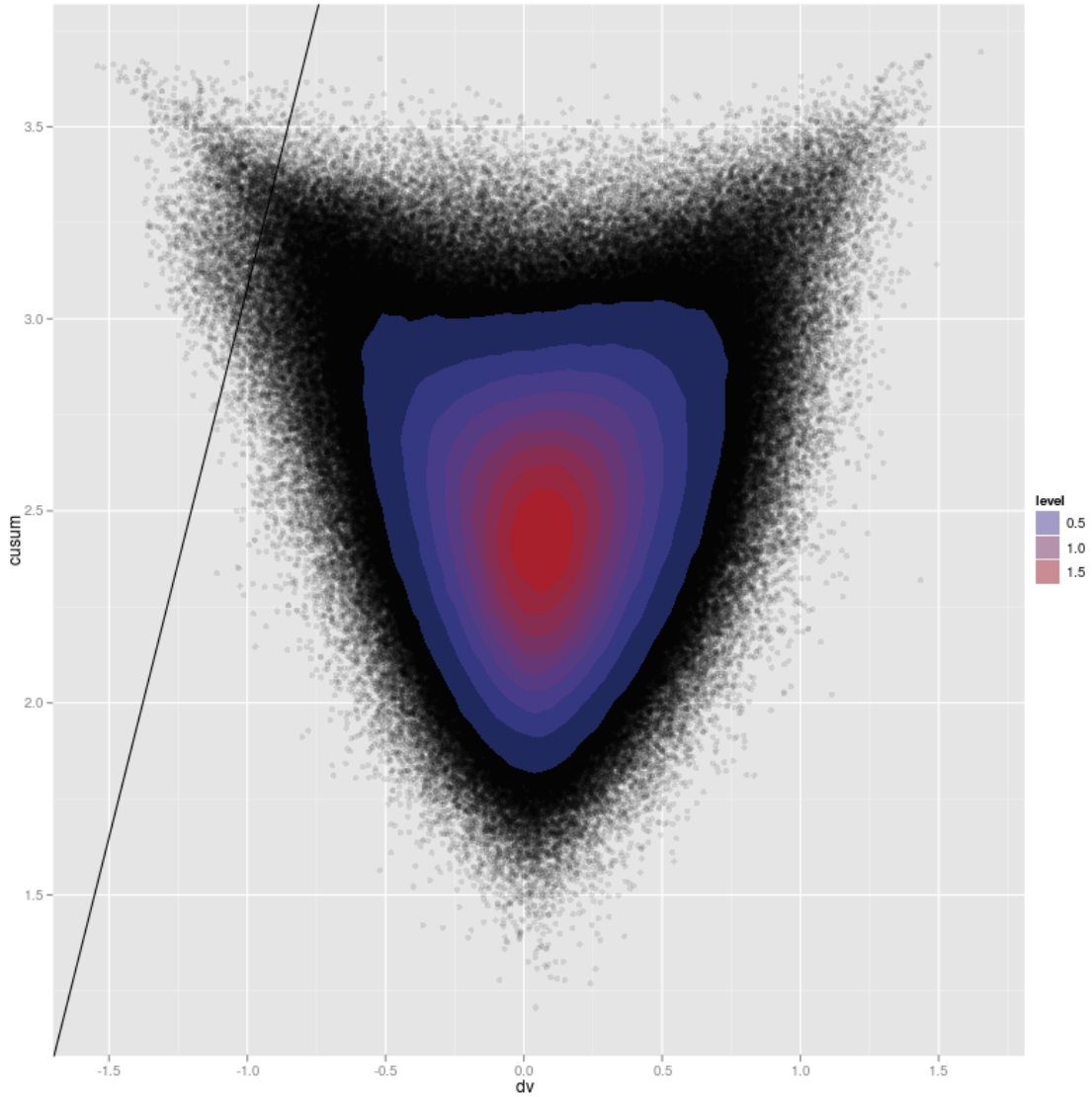
Examples


Example light curves -- fitted values
Null
CUSUM = 1.73157; DV = 0.0768301
Periodic
CUSUM = 0.910423; DV = 0.010472
Event
CUSUM = 3.17486; DV = -0.408233
Mag
Time
-2
-1
0
1
2
3
0
500
1000
1500
2000

Distribution of features on MACHO data



Distribution of features on EROS2 data

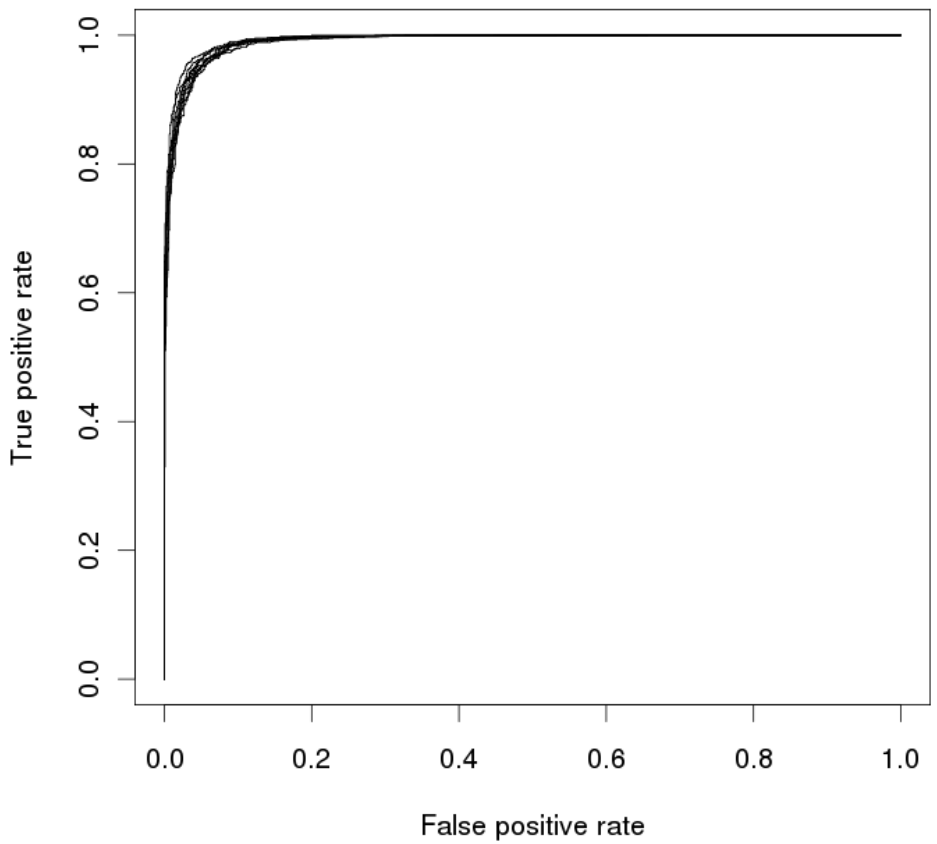


Methods

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

Training

- Obtained excellent performance (cross-validated AUC of 0.991) on MACHO training data (synthetic events and known variables)

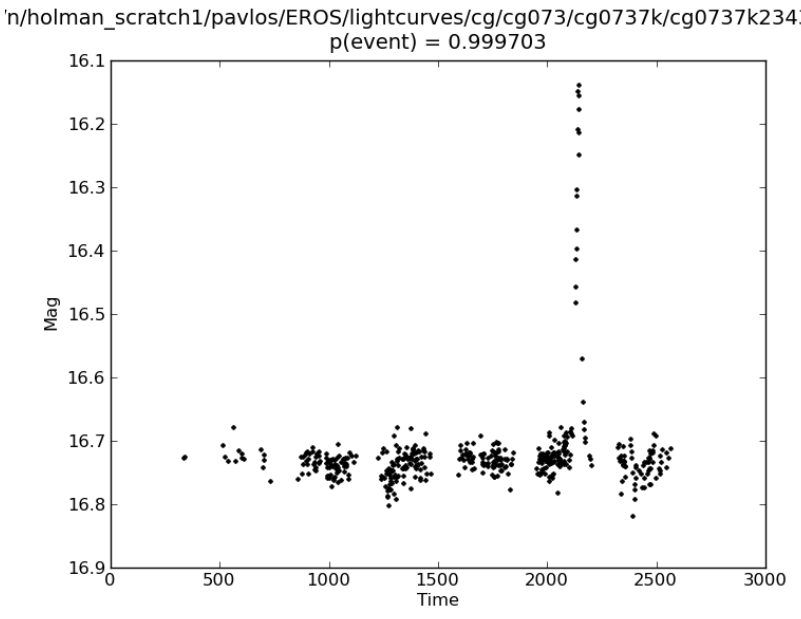


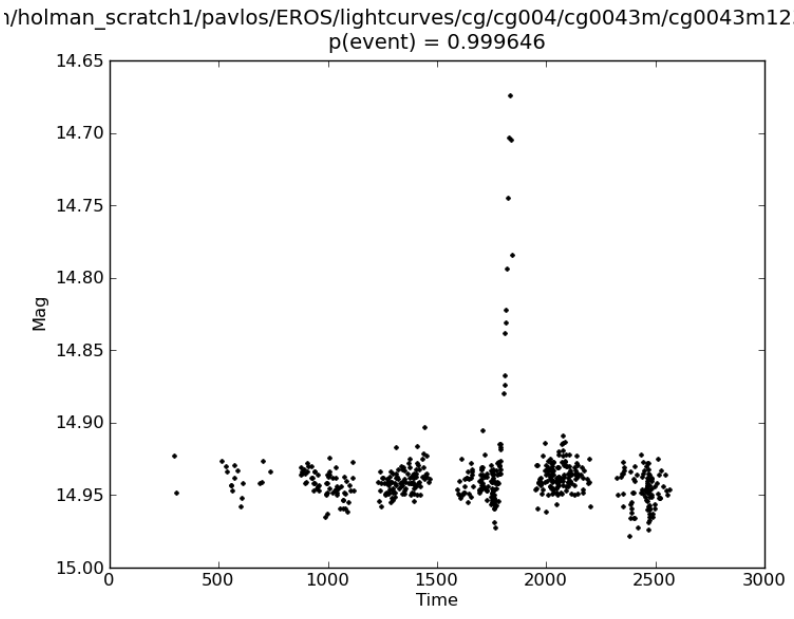
Summary

- Initial results show reduction from 87.2 million candidate light curves by approximately 98% (to approximately 1.5 million) in blue band from likelihood-ratio test
- Approximately 25,000 of the latter group are likely isolated events, based on initial analysis from classification stage
- Currently pursuing scientific follow-up on top candidates

Examples of highly-ranked events

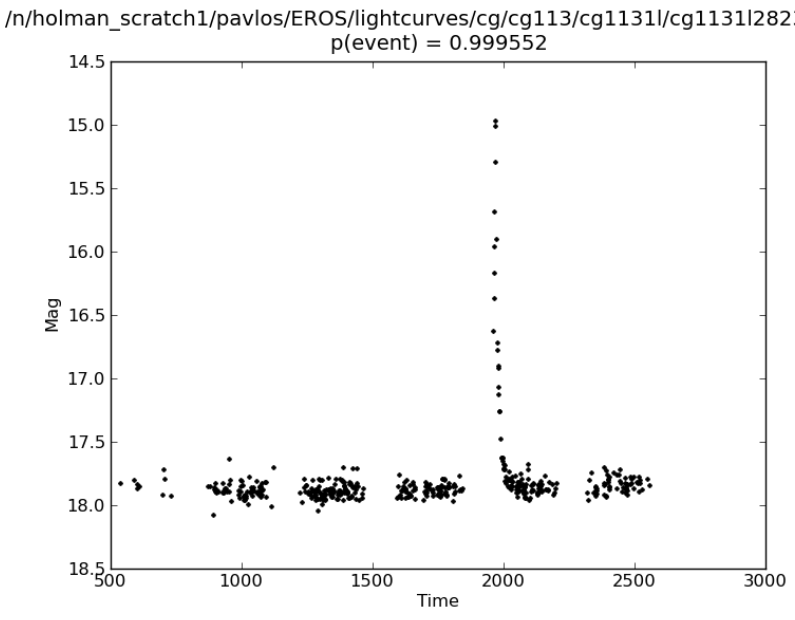
Examples from top 10:

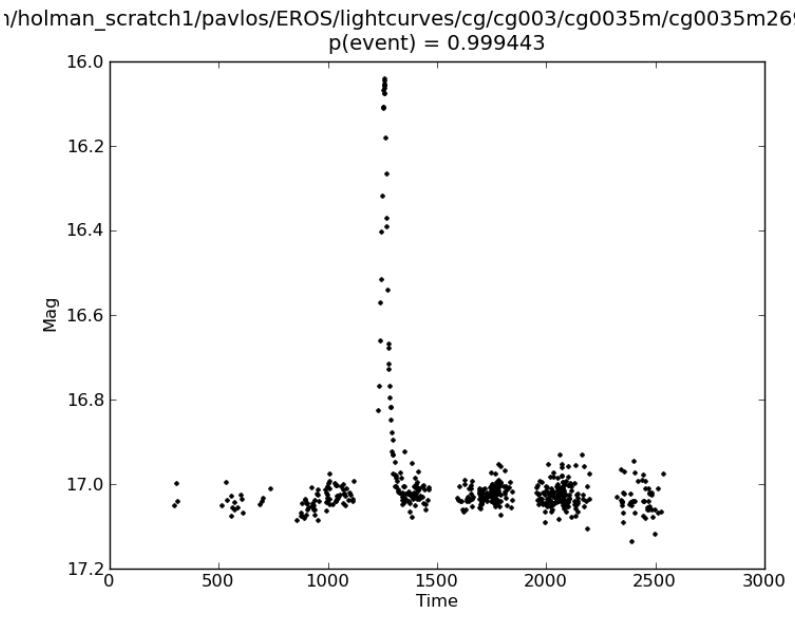




Examples of highly-ranked events

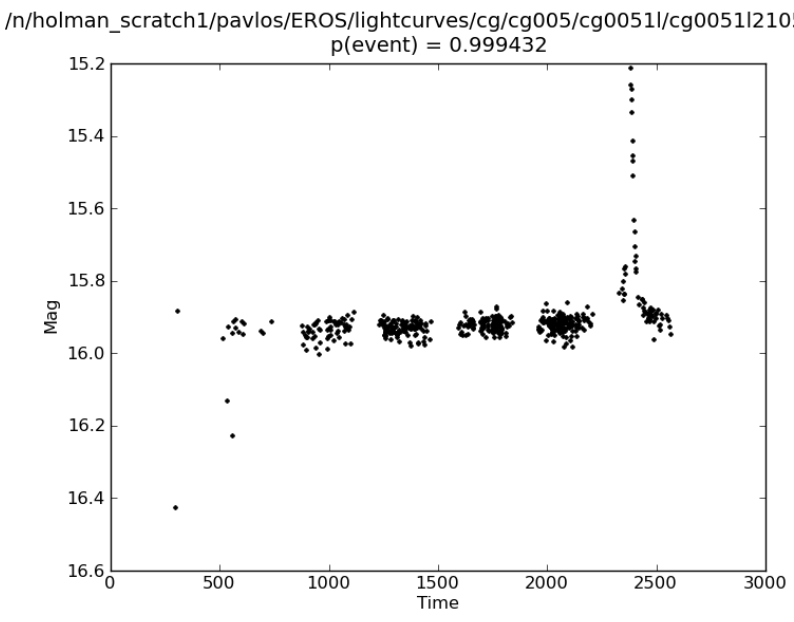
Examples from top 10:

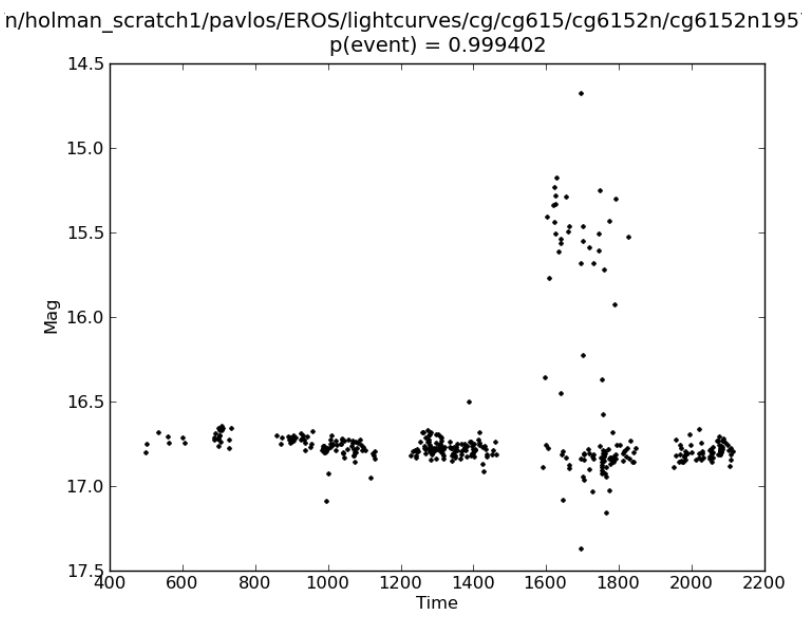




Examples of highly-ranked events

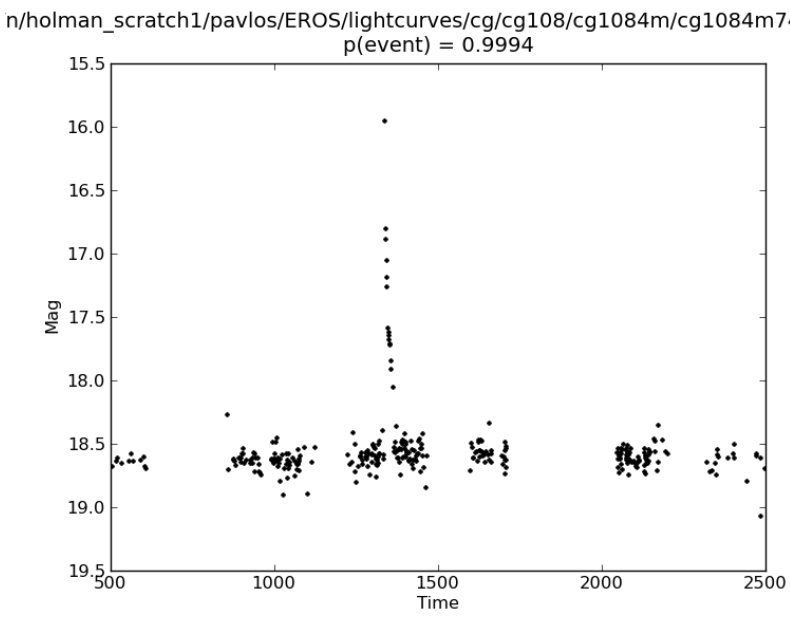
Examples from top 10:

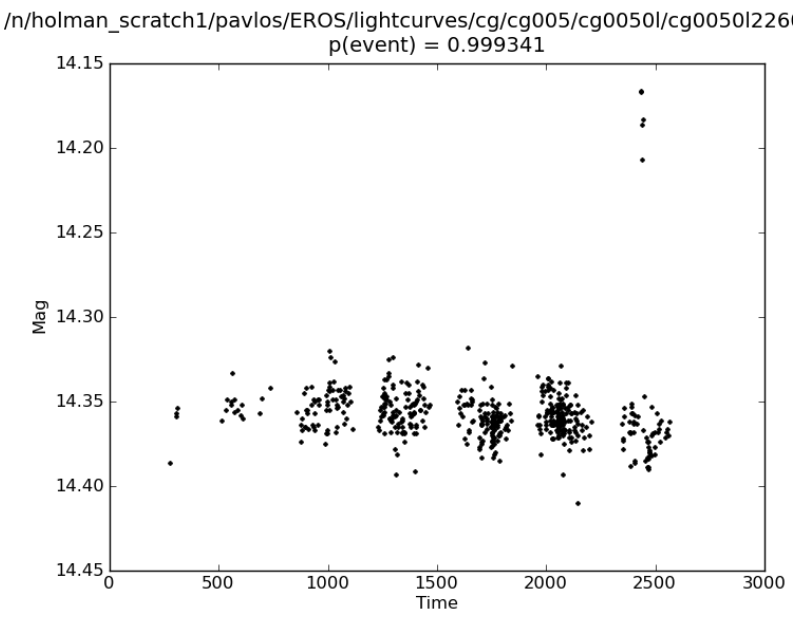




Examples of highly-ranked events

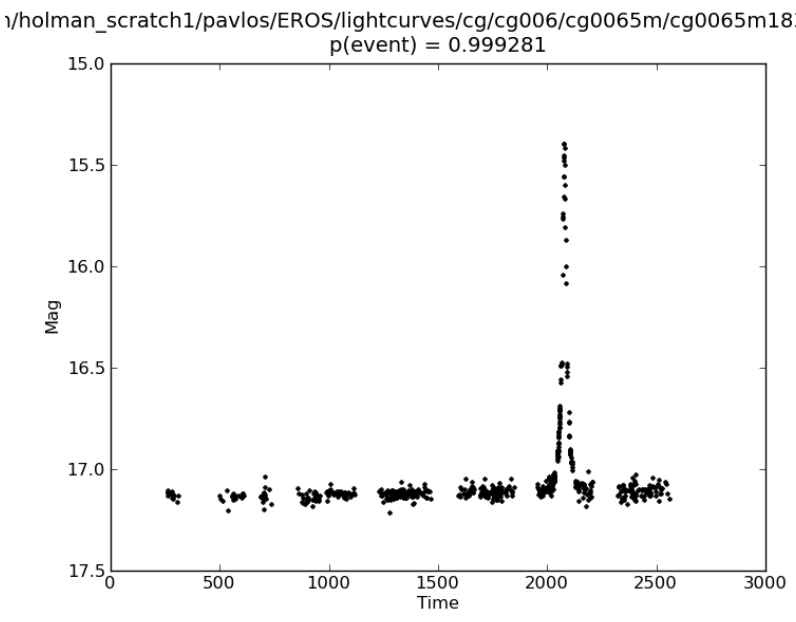
Examples from top 10:

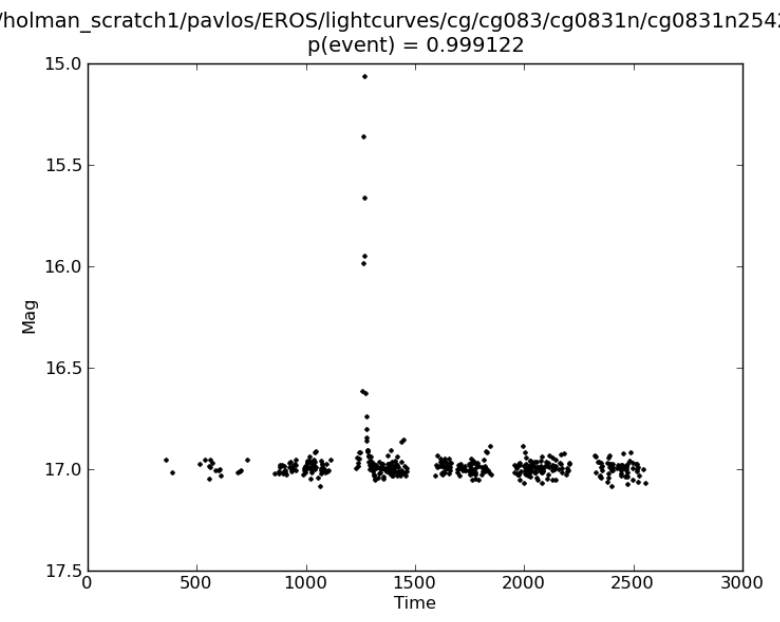




Examples of highly-ranked events

Examples from top 10:





Putting everything in its place: a mental meta-algorithm

- Understand your full (computationally infeasible) statistical model
- Preprocess to remove the "chaff", when possible
  - Be careful! Any prescreening must be extremely conservative to avoid significantly biasing your results
- Use approximations for the critical parts of your models (e.g. empirical Bayes as opposed to full hierarchical modeling) to maintain computational feasibility
  - Hyperparameters can be set based on scientific knowledge (if priors are sufficiently informative) or setup simply for mild regularization (if each observation is sufficiently rich)
  - Otherwise, a random subsample of the data can be used to obtain reasonable estimates

Putting everything in its place: a mental meta-algorithm

- Using estimates from your probability model as inputs, apply machine learning methods for computationally intractable tasks (e.g. for large scale classification or clustering)
  - This maintains computational efficiency and provides these methods with the cleaner input they need to perform well
- Use scale to your advantage when evaluating uncertainty
  - With prescreening, use known nulls
  - Without prescreening, use pseudoreplications or simulated data

Summary

- Massive data presents a new set of challenges to statisticians 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, reasonably sophisticated probability models can be incorporated into the analysis of massive datasets without destroying computational efficiency if appropriate approximations are used
- It is tremendously important to put each tool in its proper place for these types of 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

Acknowledgements

- Many thanks to both Pavlos Protopapas and Xiao-Li Meng for their data and guidance on this project
- Thanks to Jean-Baptiste Marquette for providing the TSC with EROS2 data
- I would also like to thank Edo Airoldi for our discussions on this work and Dae-Won Kim for his incredible work in setting up the MACHO data