Discovering Celestial Objects with Machine Learning

Pavlos Protopapas (CfA and Harvard SEAS)
The Time Series Group
time series center:

short overview what is it about and what we do.

science

what are the general questions

periodic variable stars:

classification using kernels

results

search engine:

morphological searches

quasars

discoveries of Quasars using light variability

event detection

outer solar system questions
time series everywhere

cadence design  stock prices

pavlos heart rate while cycling
focus: astronomy (light curves = time series). We have other data too such as labor data, real estate data, heart monitor data, archeological data, brain activity etc.
<table>
<thead>
<tr>
<th>Dataset</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>MACHO</td>
<td>66 million objects. 1000 flux observations per object in 2 bands (wavelengths)</td>
</tr>
<tr>
<td>TAOS</td>
<td>100000 objects. 100K flux observations per object. 4 telescopes.</td>
</tr>
<tr>
<td>ESSENCE</td>
<td>thousands objects, hundred observations.</td>
</tr>
<tr>
<td>Minor Planet Center</td>
<td>Light curves - few hundred objects. Few hundred observations</td>
</tr>
<tr>
<td>Pan-STARRS</td>
<td>billion of objects.</td>
</tr>
<tr>
<td>OGLE</td>
<td>few million objects.</td>
</tr>
<tr>
<td>MMT</td>
<td>occultation studies, variability studies,</td>
</tr>
<tr>
<td>some HAT-NET</td>
<td></td>
</tr>
<tr>
<td>SDSS 82</td>
<td></td>
</tr>
<tr>
<td>EROS</td>
<td>~100 million objects. 1000 flux observations per object in 2 bands (wavelengths)</td>
</tr>
<tr>
<td>DASCH</td>
<td>Harvard plates</td>
</tr>
</tbody>
</table>

disk: ~100 TB of disk LUSTER, NFS
computing nodes: Odyssey cluster at Harvard (~8000 cores).
db server: dual core with 16 GB of memory and 2 TB of disk.
few servers for development
exotic: GPGPU dedicate machine Nvidia GTX285
        GPU cluster with 16 machines with Nvidia Tesla T10 GPU’s attached to each node
        4 dedicated machines with Tesla T10
astronomy:
  - eclipsing binaries
  - extra-solar planets
  - supernovae
  - asteroids
  - TNO via occultation
  - AGNs
  - variable stars
  - microlensing

  and many many more

computer science and statistics:
  - outlier/anomaly detection
    - clustering, classification
    - motif detection
    - scalability, the feature space
    - representation of the time series
    - distance metric
    - event detection at low SNR

computational challenges
  - size
  - interplay/accessibility
  - distributed computing and disbursement of data:
    - standards (VO),
    - subscription query
    ...

research opportunities
Charles Alcock (Astro-F)
Roni Khardon (CS-F)
Carla Brodley (CS-F)
P. Estevez (EE, Uchile, F)

★Doug Alan (SoftEng)
Rahul Dave (SoftEng/Astro)

★Sio Ao (EE-PostDoc)

Dae-Won Kim (Astro-GS)
★Federica Bianco (Astro-GS)
★Gabriel Wachman (CS-GS)

Alex Blocker (Stat-GS)
Zhan Li (Stat-GS)
Pablo Huijse (Uchile-GS)

Umaa Rebbapragada (CS-GS)
Andrew Wang (Astro)

Dan Preston (CS-MSc)
Patrick Ohiomoba (Math-Msc)

Matthias Lee (CS-Un)
Devin Pleuer (CS-Un)

Tom Buckley (CS-Un)

David Smalling (GS-Economics)
Jean-Baptiste Margue (AIP-France)
Rosaline Reid
variable stars

Right Now:
- Periodic stars only
- Cepheids, RRL's, EB's
  
  Input: data (time series and other features) from survey
  
  Output: list of periodic variable stars (Cepheids, RRLs, Eclipsing Binaries)
  
  Lets do something real.

- Start with MACHO, EROS, MMT variable survey

Later:
- QSOs

Working on:
- Early prediction
typical light curves
Other features: brightness, color, period
folding
typical light curves
we have millions of light curves
triangular inequality
tree structures

Euclidean distance between light curves Q and C z-normalized

\[ r_{QC} = \sum_{n} (Q_n - C_n)^2 \]

or

\[ r_{QC} = \sum_{n} Q_n C_n \]
Kernel for Time Series

\[ \max_s \langle x, y + s \rangle \]

**Pros:**
- Does exactly what we want
- Can compute using FFT in \( O(n \log n) \)

**Cons:**
- Is not positive semidefinite
Kernel for Time Series

Theorem 1: S1 satisfies the Cauchy Schwartz inequality.

Theorem 2: Can construct a distance measure using S1 that satisfies triangle inequality.

Theorem 3: Any 3x3 Gram matrix of S1 is positive semidefinite.

Theorem 4: S1 is NOT positive semidefinite.
Kernel for Time Series

$K_1:\quad \sum_{s=1}^{n} e^{\lambda\langle x, y+s \rangle}$

**Pros:**
- Positive semidefinite
- Intuitively approximates maximum alignment
- Works as well
- $O(n \log n)$
Classification Stage Overview

SVM

Kernel K1:
Similarity measure of “shape”

Kernel K2:
magnitude (brightness), color, period

Final kernel: K1 + K2
Approach Overview

Train on OGLE
Multi-stage processing of MACHO
  Eliminate non-variables
  Eliminate non-periodic variables
  Eliminate non-Cepheid, RRL, EB periodic variables
Test on MACHO
  Rank classifications by confidence
  Set aside low-confidence predictions
Support Vector Machine (SVM)

• **Supervised** machine learning algorithm for classification (Meyer et al. 2003, Neurocomputing)
  - **Training** model
    - using known types of samples
  - **Predicting** candidates
    - using constructed model
• Example usages in astronomy
  - Classification of galaxy types using SED (Tsalmantza et al. 2009, A&A)
  - Estimating photometric redshift (Wadadekar 2005, PASP)
Support Vector Machine

\[ f(x) = w^T x + b. \]
Training Set: OGLE

14087 periodic variables of type Cepheid, EB, RRL
Periods given
Training Set: OGLE

- 14087 periodic variables of type Cepheid, EB, RRL
- Periods given
- 99.8% accuracy on cross-validation … so we're done, right?
- We know we can classify well given a group of Cepheids, EBs, and RRLs and their periods.
Current Results

- Cross-validation over OGLEII:

<table>
<thead>
<tr>
<th>Category</th>
<th>Count</th>
<th>False Positives</th>
<th>True Positives</th>
</tr>
</thead>
<tbody>
<tr>
<td>CEPH</td>
<td>3413</td>
<td>0</td>
<td>13</td>
</tr>
<tr>
<td>EB</td>
<td>0</td>
<td>3388</td>
<td>0</td>
</tr>
<tr>
<td>RRL</td>
<td>12</td>
<td>2</td>
<td>7259</td>
</tr>
</tbody>
</table>
Test Set: MACHO

- ~25 million stars (LMC and SMC)
- ~50,000 are periodic variables
- Two primary issues:
  - Finding those ~50,000 periodic variables or eliminating just the other 24,950,000 stars
  - Finding the periods
Eliminate Non-Periodic Variables

- For each time series (9 million of them):
  - Find the period
  - Fold time series to the period
- Finding the period is *hard*
Period Estimation

- Use Lomb-Scargle Periodogram to generate period candidates
- If period candidate \( \sim 1 \text{d} \), check for aliasing
- Check for asymmetry
- Find all candidate periods and put them back into the model
But Is It Periodic?

- Period finder works on any (non-)periodic data
- Check the shape
  - Need to formalize notion of “shape”
- Variance ratio
  - Move a sliding window along the folded time series
  - Compare local variances to global variances
  - For this application, this is a reliable estimate of “having shape”
  - Roll this back into the model
Eliminate Non-Cepheid, RRL, EBs

- Model learned from OGLE cannot make useful predictions on other kinds of data
- It is critical to remove as many data points as possible that are not in one of the three classes
Eliminate Non-Cepheids, RRLs, EBs

Table 4: Confusion Matrices for classification on MACHO using abstention thresholds of 1 (none), 0.99, (0.9 going left to right, up to down.)
Shift-invariant Grouped Multi-task Learning for Gaussian Processes

Work with R. Khardon and Y. Wang
Future Work

- Period finding [Y. Wang and P. Huise]
  - Using
- Estimation of non-Cepheid, RRL, EB classes or subclasses
- Ranking of predictions according to confidence [come with a correct probabilistic model]
Automatic Classification of Quasars using lightcurves

*MACHO Database*

Dae-Won Kim,
Yong-Ik Byun, Charles Alcock, Roni Khardon
QSO

- **QSO variability** research
  - But, due to the small number of **well-sampled** QSO times series, variability for individual QSOs are poorly known
    - Only ~70 well-sampled QSOs from MACHO and OGLE
Quasars

Quasar is a shortening of "quasi-stellar radio source", and they've also been called quasi-stellar objects or QSOs

Quasar is a compact region in the centre of a massive galaxy surrounding the central supermassive black hole.

The quasar is powered by an accretion disc around the black hole.

Quasars show a very high redshift, which is an effect of the expansion of the universe between the quasar and the Earth.

Quasars are a subset of Active Galactic Nuclei.
QSO Variability

- Due to the in-falling material into black hole at the center of Active Galactic Nuclei (AGN)
MACHO QSOs

- Only **60** known QSOs in total
  - 48 of them detected by time series analysis in LMC and SMC (Geha et al. 2003, ApJ)
- **51** QSOs from 30 LMC fields; 15 degree² (~**3** QSOs/degree²)
- There should be **a lot more** QSOs not yet discovered
  - MACHO LMC has total of **80** fields; 40 square degrees
  - ~**13** QSOs/degree² (SDSS Data Release 5; Schneider et al. 2007, AJ)
MACHO QSOs

• Previous time series work shows very low efficiency
  o Geha et al. 2003 selected total ~2,500 QSOs candidates
  o Manually removed 2,140 false positives and observed only 260 targets spectroscopically
  o 47 of them were confirmed as QSOs; ~2%

• What is the nature of false positives?
  o Majority of them considered to be Be stars
    ▪ B-type stars showing emission line originated from circumstellar disk and variability as well. Typically close to main sequence (Porter 2003, PASP)
  o long-period variables and even RR Lyraes
Example Light-Curves of MACHO QSOs and Be Stars

Be stars
Time Series Features

• How to separate variables from non-variables?
• How to separate QSOs from other variables?
  o n time series features
    ▪ Extracted from each MACHO time series
  o 1 color index; MACHO B-R
Time Series Features

- Stetson L (Stetson 1996, PASP) vs. color index (MACHO B-R)

\[
J = \frac{\sum_{k=1}^{n} w_k \text{sgn}(P_k) \sqrt{|P_k|}}{\sum_{k=1}^{n} w_k}
\]

\[
P_k = \begin{cases} 
\delta_{i(k)} \delta_{j(k)}, & \text{if } i(k) \neq j(k) \\
\delta_{i(k)}^2 - 1, & \text{if } i(k) = j(k)
\end{cases}
\]

\[
\delta = \sqrt{\frac{n}{n-1}} \frac{v - \bar{v}}{\sigma_v}
\]
Time Series Features

- Autocorrelation function
  
  \[ R(\tau) = \frac{\mathbb{E}[(X_t - \mu)(X_{t+\tau} - \mu)]}{\sigma^2} \]

- Number of data points above (below) the empirical line of:
Example of Autocorrelation Function

- Cepheids
- Eclipsing Binaries
- RR Lyraes
- Non-variable stars
- Quasars
- Be stars
Time Series Features

- Sigma (Shin 2008, MNRAS)
- Eta (von Neumann 1941)
- Con (Wozniak 2000, AcA)
- Range of cumulative sum; Max(S) – Min(S); Si = Si-1 + (xi – mean(x))
Support Vector Machine (SVM)

• **Supervised** machine learning algorithm for classification (Meyer et al. 2003, Nerocomputing)
  o **Training** model
    ▪ using known types of samples
  o **Predicting** candidates
    ▪ using constructed model

• Example usages in astronomy
  o Classification of galaxy types using SED (Tsalmantza et al. 2009, A&A)
  o Estimating photometric redshift (Wadadekar 2005, PASP)
Support Vector Machine

\[ f(x) = w^T x + b. \]
Support Vector Machine

- Known variables we are using to train model:
  - 58 QSOs
  - 128 Be stars
  - 582 Microlensings
  - 193 Eclipsing Binaries
  - 288 RR Lyraes
  - 73 Cepheids
  - 365 Long period variables
  - 4,288 Non-variable stars
Training QSOs SVM Model

• We’re using 2-class SVM model
  o QSOs vs. others (e.g., Be stars, Cepheids, RR Lyraes, etc)
  o 10 time series features + 1 color index
  o ~6,000 total light-curves to train model
  o 10-fold cross-validation
  o 10x10 grid search
Training QSOs SVM Model

- Recall: ~80%
- Precision: ~75% c.f. Geha's work ~2%, OGLE ~5% (6 out of 111 candidates) (Dobrzycki et al. 2005 A&A)
- Almost all false positives are Be stars
- None of the periodic variables and microlensing are misclassified as QSOs
- One misclassified Be star, which is actually a QSO, is correctly diagnosed as a QSO by our SVM model

*For RR Lyraes, Eclipsing Binaries and Cepheids, recall is ~99%.*
QSO Candidates

• ~20million MACHO LMC stars
• We're using Odyssey Cluster at Harvard
  o ~8000 CPUs (Intel Xeon 2.30GHz)
  o 16,384 Gbytes memory
  o 32,410 GFlops* (61th in the world; June 2008)
• We used 150 cores for the analysis
• It takes about 2 days to analysis the whole 20million time series
  o Without the cluster, it would take more than several months

*FLoating point Operation Per Seconds
QSO Candidates

- Total ≈1200 QSO candidates
  - 45 known QSOs are successfully recovered (total 51 previously known QSOs in LMC); 45/51 = ~88% recall rate
  - 10 known Blue Variables (among 1,300 known Blue Variables; Keller et al. 2002, AJ*) are selected as QSOs
    - None of the other types of known variables (e.g. RR Lyraes) are selected as QSOs

* Spectroscopic results for randomly selected 100 samples shows ~90% of them are Be stars.
QSO Candidates

- Example of QSO candidates (x-axis : MJD, y-axis : MACHO B magnitude)
QSO Candidates

1. Not uniformly distributed QSOs <- false positives
2. Intercept of red line <- number density of QSOs?
3. Different distribution of red (outer region of LMC) and yellow line (center region of LMC). <-Number of points in LC?
X-ray matching XMM and Chandra

- From the known MACHO QSOs 6 are in the existing footprints of Chandra.
  - We found 4 x-ray counter-parts

From all the candidates 46 are in the Chandra footprint
We found 18 to have x-ray counter-parts
Crossmatching with Kozlowski's QSOs Catalog

  - Spitzer Deep Wide Field Survey (SDWFS)
  - Spitzer Surveying the Agents of a Galaxy's Evolution (SAGE) survey
- They are obtaining spectra of 1,000 candidates

![Image of data diagrams]
Crossmatching with Kozlowski's QSOs Catalog

- Total 436 crossmatched QSOs
(Kozlowski & Kochanek 2009).
Spectroscopic validation

- Active learning approach.

- Time at 6dF
  - Multi-fiber spectrograph using V,R grating
  - Ask for 2 nights to observe 400 candidates. {First run was unsuccessful due to bad weather}.

- Magellan, July 2010, 1/2 hour.
  - Take spectra of 15 candidates in the SMC
  - Analysis not complete yet. Indications of few QSOs
  - Retrain the model and got for two more runs in winter. (problem is TAC and cost)
Figure 3. Composite SDSS QSO spectra according to redshift. The blue lines are the spectra, and dashed lines are emission lines practical for QSO confirmation. Colored regions indicate the wavelength coverage.
Expected Outcome

• New QSO Detection Algorithm based on Optical Variability and Color
• Variability Characterization for Spectroscopically Confirmed QSOs

• Studies on crossmatched X-ray sources
• LMC Internal Dust Extinction from QSO distribution
• Subclasses based on the variability

• Active learning approach.
breath
Event detection

- Motivation: Solar system, occultation by outer solar system objects.
- Surveys: TAOS, Pan-STARRS, Whipple
- First approach: Rank statistics.
- Second approach: Alex Blocker’s talk
The Solar System comprises the Sun and the retinue of celestial objects gravitationally bound to it:

- the eight planets
- 162 known moons
- three currently identified dwarf planets and their known moons (Pluto, Ceres and Eris)
- thousands of small bodies. This last category includes asteroids, meteoroids, comets, and interplanetary dust.

**Mercury, Venus, Earth, Mars, Jupiter, Saturn, Uranus, Neptune.**
Use occultations of background stars:
A tantalizing example has been reported by Schlichting et al: the HST-FGS event:
Whipple

• Indirect exploration of the outer solar system
• The primary goals of the Whipple mission are to detect and characterize objects in all of the major solar system populations beyond Neptune:
  – Kuiper Belt (and scattered disk …)
  – “Sedna region” (100 – 2,000 AU)
  – Oort Cloud (3,000 AU - ?)
• Achieve these goals by monitoring >10,000 stars to look for occultations by small objects
The Whipple mission:

- Science team includes: Charles Alcock, Gerbs Bauer, Mike Brown, Matt Holman, Scott Kenyon, Hal Levison, Steve Murray, Pavlos Protopapas, Ruth Murray-Clay, Hilke Schlichting, Paul Weissman, & Mike Werner
- SAO, JPL, and Ball Aerospace
- Whipple is a Discovery Class mission that will be proposed to NASA in response to the recent Discovery Announcement of Opportunity
*Whipple*: a Discovery Class mission to search for occultations

- Original concept inspired by *Kepler*
- Schmidt-Cassegrain telescope design
- 37 square degree field of view
- Earth leading orbit:
  - opposite direction from *Spitzer* & *Kepler*
Whipple: a proposed spacecraft to search for occultations.
Whipple will survey all of the small body populations in the solar system:
Whipple will survey all of the small body populations in the solar system:
Whipple will survey all of the small body populations in the solar system:
Whipple will survey all of the small body populations in the solar system.
*Whipple*: a proposed spacecraft to search for occultations

- Hybrid CMOS focal plane array
- 10,000 (80,000) stars, 40 (5) Hz readout
Anticipated Event Rates:

• Oort Cloud:
  – $10^{12}$ objects ($D>3$ km) each in Inner and Outer Oort Clouds
  – $N(D) \sim D^{-1.8}$; randomized eccentricities
  – 10 – 100 events per year

• “Sedna” population:
  – Take guidance from the Caltech survey (Meg Schwamb’s talk)
  – $100 \text{ AU} < a < 1000 \text{ AU}; q > 30 \text{ AU}$
  – 1-1000 per year. Very uncertain!

• Kuiper Belt:
  – $\sim 5,000$ events per year (Schlichting et al 2009 event)
Comparison between TAOS, Pan-STARRS, IMACS and Whipple:

<table>
<thead>
<tr>
<th></th>
<th>TAOS</th>
<th>Magellan/IMACS</th>
<th>Pan_STARRS (video guide stars)</th>
<th>Whipple</th>
</tr>
</thead>
<tbody>
<tr>
<td>Duration</td>
<td>5 years</td>
<td>Few days per year</td>
<td>3 years</td>
<td>3 years</td>
</tr>
<tr>
<td>Number of targets followed</td>
<td>200</td>
<td>1000</td>
<td>70</td>
<td>~10,000</td>
</tr>
<tr>
<td>Efficiency of 100% at 50 AU</td>
<td>3KM</td>
<td>0.5km</td>
<td>1km</td>
<td>0.5km</td>
</tr>
<tr>
<td>StarHours at SNR&gt;25</td>
<td>5000</td>
<td>20000</td>
<td>100000/year</td>
<td>10000/hour</td>
</tr>
<tr>
<td>Events</td>
<td>&lt;1/year</td>
<td>1-20/year</td>
<td>2-40/year</td>
<td>&gt;4000 year</td>
</tr>
</tbody>
</table>
Scan statistics

Idea: Given a time series find the region [sub sequence of measurements] that are not consistent with noise.

Original idea originates from epidemiology

time series: \( f_i \) the value at time \( t_i \)

\[
S(r,w) = \sum_{i=r}^{r+w} f_i
\]

work with Dan Preston and Iara Cury
Scan statistics

Second part is to find the p-value. Build the distribution $P(S(r,w) > S_0)$

We could simulate time series with the same noise characteristics and build distribution of $P(S)$.

**Problem:**

1) too expensive for large datasets
2) Modeling the noise is not as simple (systematics, etc)

**Solution 1:**

Reshuffle the sequence (the noise model is taken care)

Reshuffling:

Basically get all $f_i$'s and put the in a random order. Do this many times and each time calculate all $S(r,w)$. Create $S_0$ thus $P(S(r,w) > S_0)$

No good. I need to do this for each time series since the noise could be very different (data are taken different times, different filters etc)
Scan statistics

We need something that eliminates the need for modeling the noise.

Solution: Rank the y-values (flux)

Consider a Time Series T with n points. We then create a Time Series $T_R$ by converting each point in T into a ranked value. Thus, the highest point in $T_R$ will be n, the second highest $n - 1$, third highest $n - 2$, etc.

$$Q(r, w) = \sum_{i=r}^{r+w} R_i$$

where R’s are the rank values and Q is the new statistics (equivalent to S)

Advantage: All time series of n points in rank space have exactly the same $P(Q_0)$. We need to calculate this only once. Then for real data we calculate $Q(r, w)$ and we know the p-value for each point.

To calculate this simply select w numbers out of 1..n and calculate Q. From that build the distribution $P(Q)$
This distribution can be found analytically

It is the same problem as finding the number of partitions of $S$ with $w$ distinct parts, each part between 1 and $n$, inclusive. Consider the values $e_1, e_2, \ldots, e_w$.

If we can find all possible solutions to $0 < e_1 < e_2 < \ldots < e_w \leq n$, we can simply multiply by $w!$ (all possible permutations) and obtain our result.

This will be the same as the following: Subtract 1 from the smallest part, 2 from the second, etc. to get

$$0 \leq e_1 - 1 \leq e_2 - 2 \leq \ldots \leq e_w - w \leq n - w.$$ 

We have subtracted a total of

$$1 + 2 + \ldots + w = w(w + 1)/2,$$

so we are now looking for the number of partitions of $- w(w + 1)/2$ with at most $w$ parts, and with largest part at most $n - w$.

To find the number of partitions, we consider a specific application of q-binomial coefficients.
A similar problem is finding the number of distinct partitions of \( k \) elements which fit inside an \( m \) by \( n \) rectangle. This can be found by finding the coefficient of \( q^k \) in the q-binomial coefficient:

\[
q^k = \binom{m+n}{m}_q
\]

This is the same problem as the number of partitions with at most \( m \) parts and largest part at most \( n \). Applying this to our problem, we find that we are looking for the coefficient of \( q^k \)

Then, to find the number of possible sums of \( Q \), we find the coefficient of \( m - w(w+1)/2 \) in the previous equation and multiply by \( w! \).

An iterative solution does exist. To do so, we create the following recurrence:

\[
P(w, n, Q) = \sum_{j=1}^{n} P(w - 1, j - 1, Q - j)
\]

Dynamic programming. Memory management.
tets1: probability distribution build “red” by simulation and “green” with the analytical formula
test3: Real data. MACHO looking for microlensing.
Thank you
QSO

- Very energetic galaxy
  - e.g. 3C 273: about $2 \times 10^{12}$ times that of our sun; about 100 times that of average giant galaxies
- Very distant galaxy
  - $0.05 < z < 6.5$
  - Gunn-Peterson Trough
Quasar

- Very ancient galaxy
Support Vector Machine

- e.g., 7-fold cross validation

![Diagram of Support Vector Machine process]

1. Train Model
2. Predict
3. Repeat 7 times
Appendix
Appendix

Spectra of AGN
Time Series Features

- Stetson L (Stetson 1996, PASP)

\[ J = \frac{\sum_{k=1}^{n} w_k \text{sgn}(P_k) \sqrt{|P_k|}}{\sum_{k=1}^{n} w_k} \]

\[ P_k = \begin{cases} \delta_{i(k)} \delta_{j(k)}, & \text{if } i(k) \neq j(k) \\ \delta_{i(k)}^2 - 1, & \text{if } i(k) = j(k) \end{cases} \]

\[ \delta = \sqrt{\frac{n}{n-1}} \frac{v - \bar{v}}{\sigma_v} \]
Time Series Features

- Sigma (Shin 2008, MNRAS)

\[
\frac{\sigma}{\mu} = \sqrt{\frac{\sum_{n=1}^{N} (x_n - \mu)^2 / (N - 1)}{\sum_{n=1}^{N} x_n / N}}
\]
Time Series Features

- Eta (von Neumann 1941)

\[ \eta = \frac{\delta^2}{\sigma^2} = \frac{\sum_{n=1}^{N-1} (x_{n+1} - x_n)^2 / (N - 1)}{\sigma^2} \]
Time Series Features

- Con (Wozniak 2000, AcA) : number of consecutive data points above (below) 3σ
Time Series Features

• Range of cumulative sum

Max(S) – Min(S); Si = Si-1 + (xi – mean(x))
Time Series Features

Support Vector Machine

\[
\begin{align*}
\min_{w, b, \xi} & \quad \frac{1}{2} w^T w + C \sum_{i=1}^{l} \xi_i \\
\text{subject to} & \quad y_i (w^T \phi(x_i) + b) \geq 1 - \xi_i, \\
& \quad \xi_i \geq 0.
\end{align*}
\]

\[x_i \in \mathbb{R}^n \quad y \in \{1, -1\}^l,\]
Support Vector Machine

Effect of Kernel
(Ben-Hur & Weston, PyML HOWTO)

linear: $K(x_i, x_j) = x_i^T x_j$.

polynomial: $K(x_i, x_j) = (\gamma x_i^T x_j + r)^d$, $\gamma > 0$.

radial basis function (RBF): $K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2)$, $\gamma > 0$.

sigmoid: $K(x_i, x_j) = \tanh(\gamma x_i^T x_j + r)$. 
Support Vector Machine

Effect of $\gamma$ parameter
(Ben-Hur & Weston, PyML HOWTO)
Support Vector Machine

- How to decide the best C and γ parameters?
  - N-fold cross validation
    1. All labeled samples are randomly partitioned into N subsamples.
    2. Select (N-1) subsamples and train SVM model using specific C and γ values.
    3. Predict remained one subsamples and check if the predicted labels are same with original labels.
    4. Repeat these processes N times.
    5. Calculated recall rate = # of correctly predicted samples / # of total samples = # of true positives / (# of true positives + # of false negatives)

<table>
<thead>
<tr>
<th>Test result</th>
<th>Actual condition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test shows &quot;infected&quot;</td>
<td>Infected</td>
</tr>
<tr>
<td></td>
<td>True Positive</td>
</tr>
<tr>
<td>Test shows &quot;not infected&quot;</td>
<td>False Negative (i.e. infection not detected)</td>
</tr>
</tbody>
</table>
Support Vector Machine

• How to decide the best C and $\gamma$ parameters?
  o **Grid search (=Brute force search)**
    1. Set certain range for each parameter (i.e. $C_{\text{max}}$, $C_{\text{min}}$, $\gamma_{\text{max}}$ and $\gamma_{\text{min}}$).
    2. Divide the region into $N \times N$ grid.
    3. For each grid, calculate recall rate.
    4. Select $C$ and $\gamma$ which give the best recall rate.
    5. Make finer $N \times N$ grid and repeat above processes until recall rate converges.
Predicting Quasar Candidates

- Two types of false positives
  - can manually remove very easily but we want to **automate** whole prediction processes because we want apply our algorithm to another dataset in the future (e.g. MACHO SMC, MACHO Galactic bulge, or another survey data.)
- We're expecting to have **1,000~1,500** quasar candidates in the end.
Crossmatching with Chandra Catalog

- Total ~1,100 X-ray sources around LMC.
- ~40 fields, total ~0.7 degree^2
- We found 19 crossmatched ones.
- One is a previously known Seyfert 1 galaxy.

![Graph showing MJD vs MACHO B magnitude]

- x-axis: MJD
- y-axis: MACHO B magnitude

Avg. Phot. Error: 0.06, Color: 0.03, Period: 848.3