Interdisciplinary Science in Astrophysics: A Review

Pavlos Protopapas

*The Time Series Center*
Institute for Applied Computational Science
Harvard Smithsonian Center for Astrophysics

Collaborators: D. Kim (CfA), R. Khardon (Tufts), C. Alcock (CfA), Y. Byun (Yonsei), M. Rowan-Robinson (Imperial), K. Pichara (PUCChile), P. Huijse (UChile), P. Estevez (UChile), M. Trichas (CfA) …
Gargese = Καρυάι

The village was established at the end of the 18th century by the descendents of a group of immigrants from the Mani Peninsula of the Greek Peloponnese who had first settled in Corsica in the 17th century.
Outline

• What is interdisciplinary research?

• Four examples:
  • Parameter estimation, Bayesian inferences, model selection etc
  • Automatic classification
  • Event detection and anomaly detection
  • Designing surveys and future follow ups
Questions/Wish List

• Classification
  Be able to classify objects based on their variability characteristics: quasars, variable stars, supernovae, etc

• Period finding
  For sparse and noisy data, period determination is not easy.
Questions/Wish List

• Novelty detection
  • Classify something as novel serendipitously
    • Pulsars (Jocelyn Bell Burnell and Antony Hewish 1967 while looking for Quasars).
    • CMB (Arno Penzias and Robert Wilson 1967 using a horn antenna designed to relay telephone calls via satellite)
  • Four Jovian moons (Galileo 1609)
Questions/Wish List

- Event detection of rare, low signal-to-noise events
  - Occultation
  - Microlensing
  - Stellar Flares
Questions/Wish List

• Time Series modeling
  • Autoregressive model
  • Gaussian processes
  • …

• Designing observations
  • Use model and observations to design future observations

\[ Y_t = \sum_{i=1}^{p} \alpha_i Y_{t-i} + \epsilon_t \]
Interdisciplinary research

WHAT DOES NOT WORK

• Car mechanic approach. Take your data (car) to the mechanic (statistician or CS) and come back in few days.
  • What you usually get back are trivial solutions or working on the wrong thing.
  • A toy model that works on a small part of your data and will never work in reality. May be a paper if you are lucky.
• “I CAN DO IT ALL” approach. I will read the book and figure it out.
  • Wasteful. There are smart, capable people, outside our field.
  • Re-doing the work that people discovered, understood and explored 50 years ago.
Interdisciplinary research

WHAT WORKS

• Work together, learn the language of the other disciplines

• Learn the techniques but assume you will never become an expert in everything

• Work on real problems and carry them through from data to discovery. The focus should be on the scientific exploration not the methods.

• Methods will improve but most likely you will not be doing new research in the other discipline. It is usually an application or an integration.

• WARNING: There is no good definition of interdisciplinary practitioner so there are no obvious career paths.
Activity

• ADS Keywords
  • Bayes: 1250
  • Machine Learning: 120, Data mining: 900
  • SVM or Random Forest: 130
1. Classification
Motivation

- Today’s astronomical surveys (Pan-STARRS, LSST etc) are/will be producing millions of lightcurves.

- Can we classify them using automatic methods? [YES/maybe]

- Can we use variability as a discriminator? [YES/maybe]

- Can we do this in real time? [Sometimes/No]

- Can we predict/forecast? [No]

- Is it useful? Meaning do these methods work on large datasets? [No]
Variable Objects

GROUP
CLASS
TYPE

I. PULSATING STARS
- Cepheids
- Type I Classical
- RR Lyrae
- Type II W Virginis
- RV Tauri
- Long-period variables
- Mira type
- Semiregular

INTRINSIC

II. ERUPTIVE
- Supernovae
- Novae
- Recurrent novae
- Dwarf novae
- Symbiotic stars
- R Coronae Borealis

VARIABLE STARS

(Cataclysmic stars)

EXTRINSIC

CLASSIFICATION OF VARIABLE STARS

III. ECLIPSING BINARIES

IV. ROTATING VARIABLES

Copyright CSIRO Australia 2004.
Characterize variability

stochastic

periodic

temperamental

noise
## Accuracy/Precision Recall

### Accuracy and Precision

<table>
<thead>
<tr>
<th></th>
<th>Absolute Truth</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AGN</td>
</tr>
<tr>
<td><strong>Our model</strong></td>
<td></td>
</tr>
<tr>
<td>AGN</td>
<td>True-Positive</td>
</tr>
<tr>
<td>Non-AGN</td>
<td>False-Negative</td>
</tr>
</tbody>
</table>

### Recall/Efficiency

<table>
<thead>
<tr>
<th></th>
<th>RECALL/ EFFICIENCY</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Supervised methods

- Machine learning approach.
- A computer program is said to learn given some examples $E$ for a task $T$ and some performance $P$. If we improve $P$ using examples $E$
- For example, discriminate classes given some examples $E$ (T. Mitchell 1999)
- $E$ is the training set
If we do not care the best hyperplane => perceptor

Most methods find an optimal separating hyperplane

• Which points should influence optimality?
  • All points => Linear regression
  • Only “difficult points” close to decision boundary => Support vector machines
SVN (non-linear separability)

• Transformation to separate
SVN (non-linear severable)

The idea is to gain linear separation by mapping the data to a higher dimensional space

– The following set can’t be separated by a linear function, but can be separated by a quadratic one

\[(x-a)(x-b)=x^2-(a+b)x+ab\]

so if we map \(x\rightarrow\{x^2, x\}\) we gain linear separation
Random Forest

• Select a \( n \) subsets of data (with replacement)
  
  usually \( n = N \) (the total number of objects in training set)

• Select \( m \) subset of features

• Create \( k \) decision trees

• Each tree votes and majority wins

• Adapting Boosting: iterating with emphasis on the points that were badly classified in the previous iterations => adaboost
Results CV

• Cross-validation
  1. Divide a dataset into training and test set
  2. Train a model using the training set
  3. Test the model on the test set
Features

• Feature = attribute = summary statistics

• One would like to include features that characterize
  • Variability (SNR, $\chi^2$, var, Q90/Q10, …)
  • Periodicity (best period, second best, goodness of fit)
  • Stochasticity (structure function, AR(q) coeff, …)
  • Shape (cross-correlation with known shapes, …)

• The more the merrier (not always)
Periodic Variable

- Wachman et. al. 2009 ML Journal. Used OGLEII+MACHO periodic variables

<table>
<thead>
<tr>
<th></th>
<th>Cepheid</th>
<th>EB</th>
<th>RRL</th>
<th>Cepheid</th>
<th>EB</th>
<th>RRL</th>
<th>Cepheid</th>
<th>EB</th>
<th>RRL</th>
<th>Cepheid</th>
<th>EB</th>
<th>RRL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cepheid</td>
<td>3416</td>
<td>1</td>
<td>13</td>
<td>3382</td>
<td>1</td>
<td>3</td>
<td>3363</td>
<td>1</td>
<td>3</td>
<td>3352</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>EB</td>
<td>0</td>
<td>3389</td>
<td>0</td>
<td>0</td>
<td>3364</td>
<td>0</td>
<td>0</td>
<td>3342</td>
<td>0</td>
<td>0</td>
<td>3312</td>
<td>0</td>
</tr>
<tr>
<td>RRL</td>
<td>9</td>
<td>0</td>
<td>7259</td>
<td>1</td>
<td>0</td>
<td>7195</td>
<td>0</td>
<td>0</td>
<td>7166</td>
<td>0</td>
<td>0</td>
<td>7138</td>
</tr>
</tbody>
</table>

- P. Dubath 2011, MNRAS.
  
  Used *Hipparcos* periodic variable stars
Non periodic variables/Stochastic

• Kim et al. 2011 and 2012 used SVM and RF to discriminate AGNs and Be stars from the rest in MACHO and EROS and recently to Pan-STARRS

  88% recall and 78% precision.

  Combining with other contextual information (mid-IR, xray) resulted in very high confidence candidates

• Follow up candidates with spectroscopy and retrain.

• Mashala 2011, classification of SN and cataclysmic variables. Extremely high recall/precision (caveat: used cuts and preselected datasets)
Figure 5 from On Machine-learned Classification of Variable Stars with Sparse and Noisy Time-series Data
Issues

• Performance is only given within known classes.
• Most of the sources are not in these classes.
• Training sets are not representative of the actual data
• We need methods that perform well on the whole dataset
• Kim et. al. is the only work that evaluated the performance on the whole dataset with follow up spectroscopy. False positives is the issue when the dataset is predominantly non-variable sources.
• Period discrimination is MAJOR drawback.
• For these methods to be really useful we must deal with missing data/features.
2. Parameter estimation and inferences
Historical perspective

- Statistics, the idea of summarizing observations first was done by astronomers (Greeks of course) Almagest book VII and VIII contains catalogs of stars.

- Least square fit was first demonstrated by Tobias Mayer. *The history of statistics: the measurement of uncertainty before 1900* by Stephen M. Stigler

- Laplace fitting of Saturn’s mass one of the first examples of probability theory.

- Markov Chain Monte Carlo and Metropolis (the second most used algorithm in science) *SIAM News, Volume 33, Number 4*

- Bayes application in scientific data with significant consequences, WMAP *Verde 2003 (>500 citations)*

- Almost every theory has some form of parameter estimation and inferences ($\chi^2$ and beyond) and in this perspective this is by far the most mature, most trusted and most used methodology
Gaussian processes a different regression

• Start with the regression model

\[ y = f_w(x) + \epsilon \]

where \( f_w(x) \) is the regression function, e.g., in linear regression the parameters \( w \) are the slope and intersection (\( f_w(x) = w^T x \)).

• Given data \( D = (x, y) \) one wishes to infer \( w \) and the basic approach is by maximizing the likelihood.

• In Bayesian statistics the concept of probability is introduced through the prior. The posterior probability is

\[ p(w|D) \propto \Pr(D|w) p(w). \]

• Linear regression

\[
\begin{align*}
  w &\sim N(0, \Sigma_w) \\
  y|X, w &\sim N(X^T w, \sigma_n^2 I) \\
  w|X, y &\sim N\left(\frac{1}{\sigma_n^2} A^{-1} XY, A^{-1}\right) \\
  A &= \frac{1}{\sigma_n^2} XX^T + \Sigma_w^{-1}
\end{align*}
\]
Gaussian processes a different regression

- The predictive distribution for a new observation $x^*$ is given by
  
  $$p(f(x^*)|D) = \int p(f(x^*)|w) p(w|D) \, dw$$

- Problem is that linear regression is not very descriptive.

- One can extend this with more complicated $f(x)$ or alternatively instead of fitting model, one fits functions. In particular Gaussian functions (processes) is defined by its mean and a covariance matrix

  $$f(x) = \mathcal{N}(\mu, k(x, x'))$$

- For example:  
  $$k(x, x') = e^{- (x-x')^2 / 2\sigma^2}$$

- The standard Bayesian approach is to identify the hyper-parameters that maximize the marginal likelihood.
Multivariate

• Draw from $y \sim \mathcal{N}(\mu, \sigma)$. Stochastic process = timeseries resulting into a ‘white noise’

• How about if we draw from a multivariate $\vec{y} \sim \mathcal{N}(\vec{\mu}, K)$

• For example $K(t_i, t_j) = e^{-\frac{1}{2\sigma^2}(t_i-t_j)^2}$
DEMO
GP the essence

Figure 1 shows three functions drawn at random from a GP prior; the dots indicate values of $y$ actually generated; the two other functions have (less correctly) been drawn as lines by joining a large number of evaluated points. Panel (b) shows three random functions drawn from the posterior, i.e. the prior conditioned on the five noise-free observations indicated. In both plots the shaded area represents the pointwise mean plus and minus two times the standard deviation for each input value (corresponding to the 95% confidence region), for the prior and posterior respectively.

The characteristic length scale is around one unit. By replacing $|x_p - x_q|$ by $|x_p - x_q|/\gamma$ in Equation 1 for some positive constant $\gamma$ we could change the characteristic length scale of the process. Also, the overall variance of the random function can be controlled by a positive pre-factor before the exp in Equation 1. We will discuss more about how such factors affect the predictions in Section 3.5 and say more about how to set such scale parameters in Chapter 6.

Prediction with Noise-free Observations

We are usually not primarily interested in drawing random functions from the prior, but want to incorporate the knowledge that the training data provides about the function. Initially, we will consider the simple special case where the observations are noise-free, that is we know \{f(x_i), f_i \mid i = 0, ..., n\}. The joint prior distribution of the training outputs, $f$, and the test outputs $f^*$ according to the prior is $\prod_{i=0}^{n} f_i \sim N(0, K_f X, X)$. If there are $n$ training points and $n^*$ test points then $K_f X, X^*$ denotes the $n \times n^*$ matrix of the covariances evaluated at all pairs of training and test points, and similarly for the other entries $K_f X, X^*$.

To get the posterior distribution over functions we need to restrict this joint prior distribution to contain only those functions which agree with the observed data points. Graphically in Figure 1 you may think of generating functions from the prior, and rejecting the ones that disagree with the observations. We could also do this graphically rejection...
Gaussian processes a different regression

- The marginal likelihood is

\[
\log p(y|x; M) = \log \left( \int p(y|f, x; M) p(f|x; M) df \right) \\
= -\frac{1}{2} y^T (K + \sigma^2 I)^{-1} y \\
- \frac{1}{2} \log |K + \sigma^2 I|^{-1} - \frac{n}{2} \log 2\pi
\]

- With the help of the following algebraic property. f: fobserved data and f*: unobserved.

\[
\begin{bmatrix} f \\ f_* \end{bmatrix} = \mathcal{N} \left( \begin{bmatrix} 0 \\ K(X, X) \end{bmatrix}, \begin{bmatrix} K(X, X) & K(X, X_*) \\ K(X_*, X) & K(X_*, X_*) + \sigma^2_n \end{bmatrix} \right)
\]

- Predictive posterior:

\[
f_*|X_*, X, f \sim \mathcal{N}(K(X_*, X)K(X, X)^{-1}f, X(X_*, X_*) - K(X_*, X)K(X, X)^{-1}K(X, X_*))
\]
Gaussian Process Hyperparameters: Example of length scales

(a), \( \ell = 1 \)

(b), \( \ell = 0.3 \)

(c), \( \ell = 3 \)
Gaussian process/Period estimation/discrimination

• In the case of period estimation the underlying function \( f(\cdot) \) is periodic with unknown frequency \( w = 1/p \). To model the periodicity we use a GP with a periodic covariance function

\[
K_\theta(x_i, x_j) = \beta \exp \left\{ - \frac{2 \sin^2(w \pi (x_i - x_j))}{\ell^2} \right\}
\]

• And combine it with another kernel

\[
K_M(y_i, y_j) = \exp \left\{ - \frac{(y_i - y_j)^2}{2\sigma^2_m} \right\}
\]
Parameter fitting

• We have 3 hyperparameters, w, l, \( \sigma \)

• l and \( \sigma \) are easy to estimate, w is very difficult.

• In order to get a good estimation for w we need to sample \( 1/T \) where T is the total time span => 20K trials => too slow
Effect of $l$ and $\sigma$

![Graphs showing the effect of $l$ and $\sigma$ on magnitude over time and phase.]

- Left graph: Time [days] vs. Magnitude for $l = 1$ and $\sigma = 1$
  - Fit: $P = 0.7127154$ CKP: 1
  - Data points: $0090m/lm0090m4818.time 49384$
  - Smoothing: 0.050408 me 0.032 ot 0.36432

- Right graph: Phase vs. Magnitude for $l = 1$ and $\sigma = 1$
  - Fit: $P = 1.54191$ CKP: 50872.49
Effect of $l$ and $\sigma$

FIX $\sigma$ as the median of errors

ADA May 2012
Pavlos Protopapas
1) Separate the light curve in bands containing equal number of points (~50)
2) Select bands with the highest mean first derivative
3) Estimate Fourier transform of times in band with highest derivatives
4) Select ~100 best frequencies in the Fourier space
5) Select the best 500 frequencies
Training set

- Training set:
  - **Periodic**: Use multivariate periodic kernel using parameters selected uniformly from
    - \( P = [0.1, 1000] \) days
    - \( \text{SNR} = [1, 10] \) days
    - \( l = [0.1, 1] \)
  - **Non periodic noise**: Using random stars and shuffle the times
  - **Variables**: Simulate using multivariate GP
  - Optimize recall+precision by adjusting the discriminating threshold.

ADA May 2012
Pavlos Protopapas
Results or EROS-2 and MACHO

• Code is ported in cuda (GPGPU) : **0.1 sec/lightcurve** (each lightcurve has 1000 measurements).

• Analyze 50K lightcurves:
  • found 500 periodic.
  • Inspected all 50K lightcurves.
  • We get the same hit rates (false positives too) as in the synthetic. For the EROS-2 **85% recall and 98% precision**

  [this is for discriminating periodic and finding the correct period]

• We are now analyzing 50 million lightcurves on a cluster of 128 GPUs (XSEDE)

• We can estimate the number of periodic sources as a function of period, color etc. and of course periodic discrimination goes back into classification.
Issues

• Semi-periodics
• Trends
• Combining multi-band data
3. Event detection/outlier detection
Outlier detection/event detection

- Event detection is to find the outlier of a given dataset
  Preston et. al. 2008, Blocker et. al 2011
  - Scan statistics (sequential scan through the data)
  - Rank statistics
  - Found all microlensing in EROS/OGLE and some new
- Outlier detection could also mean to find the outlier within the ensemble.
  Rebragaata et. al. 2007, Keogh et. al. 2008
  - Use metric of similarity and find global outliers
  Pichara 2012
  - Supervised classification to remove known classes
Event detection/issues

• Most methods find too many outliers due to bad measurements
  • Inspection is still necessary

• Large dataset-> Serendipitous discovery is not possible. Discovery of novel phenomena should be guided with these new methods.

• Real time classification and early follow ups
4: Sequential design
Scientific method: the way we know it

• Construct a hypothesis based on theory
• Design experiment to test this hypothesis
• Do the experiment
• Analysis =>
  • New hypothesis if hypothesis fails
  • New experiment if hypothesis succeed
Scientific approach: Sequential approach

- Observation
- Inference
- Intermediate results (STOP)
- Prediction
- Devise new experiment

Other information (priors)

Data flows through the diagram from Observation to Inference, then to Intermediate results, and finally to Prediction. The process is cyclical, allowing for continuous refinement and new experiments based on the results.
Experimental design/decision theory

• The most naïve approach is to divide the probabilities. Or take the two hypothesis and divide the evidence from the Bayes analysis

Playing with dice. Draw: H1=1 or H2=5 :

\[ P(H1|I) = \frac{1}{6} \quad P(H2|I) = \frac{1}{6} \]

Then the two hypothesis are equal.

If the hypothesis is rolling 6 (H1) compare to the rest H2

\[ P(H1|I) = \frac{1}{6} \quad P(H2|I) = \frac{5}{6} \]

The odds of H2 over H1 is \( Q=1/5 \)

• Would you bet money ?
Consequences

• Decision is made based on the consequences.

• If H1 and H2 is about rolling dice then there are no deadly consequences.

• But if H1 is the hypothesis that I will be able to climb this rock and H2 is that I will not make it to the top.
  • Odds 5/1. Good chance but not attractive anymore.

• Utility function
  • Outcomes: die or be cool
Design an experiment

• Questions in astronomy and cosmology
  • What bands shall I use
  • Is it worth following a transient and when is the best time?
  • When shall I stop following up
  • …

• Make the decision based on probability and utility. How expensive is it to follow up. Which instrument is more accessible?
Information theory/Entropy

• Shannon information entropy describes how much knowledge or how uncertain we are about the system expressed by a probability distribution

\[ S = \sum_i p_i \log \frac{1}{p_i} = - \sum_i p_i \log p_i \]

• How much information (in bits) we gain if we acquire \( d \) more data?

\[ \delta S(d) = S[p(H_i|D)] - S[p(H_i|d, D)] = \sum_i p(H_i|d, D) \log p(H_i|d, D) - C \]

• Entropy of a Gaussian

\[ P(x) \propto e^{-(x-\mu)^2/2\sigma^2} \rightarrow S \propto \log (\sigma) \]

\[ P(x) \propto e^{-(x \Sigma^{-1} x)^2} \rightarrow S \propto \log (\det \Sigma) \]

• Which is the Fisher matrix. Asymptotically what we expect.
Mutual Information

• Mutual information measures what information we gain if we knew one of two parameters.

\[ S(X; Y) = \sum_X \sum_Y p(X, Y) \log \left( \frac{p(X, Y)}{p(X)p(Y)} \right) \]

• Mutual information tells us how much information is gained or lost from the dependency of one variable on the other.

• If X is discrete and Y is continuous then

\[ S(X; Y) = \int_Y \sum_X p(X, Y) \log \left( \frac{p(X, Y)}{p(X)p(Y)} \right) \]
Spectral Energy Density (SED)

- Our object is 1D vector
- Where to observe after these 3 observations?
- Models, $M_i$ described by functions $T_i(\nu)$
Probability Density Functions

• Given few observations we estimate the posteriors for each model

\[
\log p(M_i | D) \propto \log p(M_i) - \sum_i \frac{[y_i - T_j(\nu_i)]^2}{\sigma_i^2}
\]

• The probability for each model to be true amongst all other outcomes is:

\[
P(M_i) = \frac{p(M_i | D)}{\sum_i p(M_i | D)}
\]

• The best frequency can be derived as

\[
\nu^* = \operatorname{argmax}_\nu I(Y; M|\nu)
\]

\[
I(Y; M|\nu) = \sum_M \int_Y P(Y, M) \log \left( \frac{P(Y, M)}{P(M) P(Y|\nu)} \right)
\]
SED fitting AGNs/Galaxies/Star Burs

- Given few observations we want to infer the information gained by examining how much mutual information we have for the flux at a given frequency and the model of choice.

\[
I(Y; \mathcal{M}|\nu) = \sum_{\mathcal{M}} \int_Y P(Y, \mathcal{M}) \log \left( \frac{P(Y, \mathcal{M})}{P(\mathcal{M}) P(Y|\nu)} \right)
\]
Utility function

- Some bands are easier than others. Anything that needs satellite needs to be weighted appropriately.
- The choice also depends on the expected observational error. So we run exhaustively on the choices or errors given the resources.
Conclusion

• Little interdisciplinary research in astronomy/astrophysics
• Inferences/modeling matured
• Classification is within our reach
• Event detection and outlier detection is more of a computational problem
• Sequential design is at the infant state