Bayesian Time-Series Selection of AGN Using Multi-Band Difference-Imaging in the Pan-STARRS1 Medium-Deep Survey

Sidharth Kumar
Department of Astronomy, University of Maryland, College Park

1 Introduction

With the advent of the LSST era, the requirement of automated object identification in large volumes of data in existing catalogs, as well as in real-time data, is necessitated. With the knowledge of prior event types in telescopic surveys, it is possible to look for specific events in the data with a high degree of completeness and efficiency using variability [2, 3, 4], color based selection [3], and host-galaxy properties [5]. Although, a consummate and complete method would warrant the use of all these parameters in conjunction, time-series methods by themselves can contribute significantly to the selection process. Many time-series methods have been applied in the past to the identification of a broad spectrum of objects, as well as specific types; [4] discuss the identification of AGN via damped-random walk parameterization of difference-imaging light-curves, [6] on the applicability of single and multiple Ornstein-Uhlenbeck processes (hereafter OU process) to the identification of AGN, [3] on the separation of AGN from variable stars in photometric surveys through damped-random walk parameterization, and [7] on the photometric identification of specific supernovae (SNe) types. Particularly ubiquitous is the application of robust Bayesian methods [2] to the selection of sources using analytical deterministic and stochastic models for the light-curves. However, the applicability of these class of methods have been limited to single-band detections [4], or predominantly limited to using magnitude time-series data [3]. For the first time, we present multi-band difference-image selection of AGN and SNe, in the Pan-STARRS1 medium-deep fields in the g, r, i and z bands. Using Bayesian analysis, we estimate the likelihoods of a diverse range of SN models as compared to the that of the OU process. We then combine the model comparisons filter-wise using a K-means clustering algorithm [12] to provide a robust classification in each filter. The classifications are then combined across the g,r,i and z filters, to give a final classification including two measures to estimate the quality of the classification (§3).

The Pan-STARRS1 survey [9] has two operating modes, 1. The 3π survey which covers 3π square degrees at $\delta > -30$ degrees in 5 bands with a cadence of 2 observations per filter in a 6 month period, 2. Deeper multi-epoch images of 7 square degree fields in 5 bands, the so-called Medium Deep Field (MDF) Survey, for both extensive temporal coverage and depth. Depending on the weather, the accessible fields are observed with a staggered 3-day cadence in each of bands during dark and gray time ($g_{\mathrm{P1}}, r_{\mathrm{P1}}$ on the first day, i_{P1} on the second day, z_{P1} on the third day, and then repeat with $g_{\mathrm{P1}}, r_{\mathrm{P1}}$), and in the y_{P1} band during bright time. On average, the cadence is 6 detections per filter in a 1 month period in $g_{\mathrm{P1}}, r_{\mathrm{P1}}, i_{\mathrm{P1}}$, and z_{P1}, with a 1 week gap during bright time when the MDFs are exclusively observed in y_{P1}. While the 3π may detect millions of sources, in our studies we will exclusively use sources detected in the MDFs ($\approx 10^4$) since source classification entails dense time-series.

An exhaustive list of Pan-STARRS alerts are available in an online alerts database located in Harvard [9]. To derive the list of extragalactic transients, we cross-matched 18058 detected in the first 2.5years of the Pan-STARRS1 medium-deep survey to within 3" of host galaxies detected in the deep-stack star-galaxy catalogs (Heinis 2014, In preparation), resulting in 8565 distinct extragalactic transients. These transients can be categorized broadly into stochastically varying, like AGN, or explosive-transient, like SNe. In the next section we discuss time-series models which can be used to assess their variability, and hence their categorization.

2 Time-Series Models

We assess the general shapes of the light-curves by comparing their similarities to SN-like bursting behavior, or with damped-random-walk type behavior like those of AGN. While an attempt to an exact SN or AGN fit is more tedious, the general shape of a SN light-curve could be approximated to certain analytical functional forms (Gaussian, Gamma, and generic analytic SN model (Analytic-SN)), and that of an AGN light-curve approximated by an Ornstein Uhlenbeck process [1] (OU process).

Model	Type	Equation
Gaussian	SN	$Flux(t) = \alpha + \beta e^{-(t-\mu)^2/\sigma^2}$
Gamma distribution	SN	$Flux(t) = \alpha + \beta \frac{(t-\mu)^{k-1} e^{-(t-\mu)/\theta}}{\theta^k \Gamma(k)}$
Analytic SN model	SN	$Flux(t) = \alpha + \beta \frac{e^{-(t-t_o)/t_{fall}}}{1+e^{-(t-t_o)/t_{rise}}}$
OU process	AGN	$dz(t) = -\frac{1}{\tau} z(t) dt + c^{1/2} N(t; \mathit{0}, dt)$
No-Model	ALL	$Flux(t) = \frac{\sum_i \left(1/{\delta_i}^2\right) y_i}{\sum_i \left(1/{\delta_i}^2\right)}$

Table 1: Difference flux models used in the characterization of AGN and SN.

The models are described in Table 1. For SNe, although the asymmetry in the rising and falling limbs of the light-curve, cosmological redshift corrections, and extinction are not factored in, the attempt is to classify the source as a burst, or as stochastically varying, thereby not necessitating fitting specific inflections in the light-curves. Note, that since the models are compared with each other, only their relative fitnesses are important. Also, should the necessity arise of classifying the objects into particular sub-types of the major classifications, or that of extracting particular details about the parameters of the SNe, exact models [7] must be included in the comparisons, which although is beyond the scope of this paper, is a likely extension.

To assess the aptness of the models, we derive both the corrected Akaike information criterion (AICC) [13] and the leave one out cross-validation likelihood (LOOCV) [2].

$$AICC = 2k - 2log\mathcal{L} + \frac{2k(k+1)}{n-k-1} \tag{1}$$

The AICC Eq.(1) measures the over-parameterization of a dataset by a model by correcting the maximum likelihood $\mathcal{L}$ with the number of parameters k, as well as the finite size of the dataset n as compared to k. Although the $AICC$ by itself is a good indicator of model fitness in the event that the data is representative of the general distribution of the source time-series, it may not take into account the variations in likelihood resulting from noisy data, thereby misrepresenting the actual goodness of fit. The LOOCV on the other hand is more robust to such variations which are especially common in difference-imaging. Since the AICC and the LOOCV are independent of each other, they they can be used simultaneously to assess model fitness. This establishes a balance between the overall model fit via the AICC, and the robustness of the model to noise via the LOOCV. In this paper we use uniform priors for the SN and AGN model parameters, thereby not biasing the models toward particular regions of parameter space. We also use a Gaussian error model for all time-series models including the OU process, since our aim is to assess how well the mean time-series of each model fits the lightcurves.

3 Classification Method

To quantify the model fits to the data, we assess the leave one out cross-validation likelihood (LOOCV) and the corrected Akaike information criterion (AICC) for each model, filter-wise for each source. The LOOCV for each model, in each filter, are evaluated using a standard Metropolis-Hastings algorithm, and the AICC is computed from the sampled maximum likelihood. Sources which are best fit by the No-model are filtered out based on the No-Model having the highest LOOCV and the lowest AICC amongst the 5 models. We then construct the relative sign vector $RV_{i,f}$ Eq.(2)

for each object, in each filter, which is a measure of how well the data is described by the SN models as compared to the AGN model.

$$\begin{aligned} RV_{i,f} &= \{sgn(LOOCV_{Gauss} - LOOCV_{OU}), sgn(LOOCV_{Gamma} - LOOCV_{OU}), \\ &\quad sgn(LOOCV_{Analytic\text{-}SN} - LOOCV_{OU}), sgn(AICC_{Gauss} - AICC_{OU}), \\ &\quad sgn(AICC_{Gamma} - AICC_{OU}), sgn(AICC_{Analytic\text{-}SN} - AICC_{OU}\} \qquad (2) \end{aligned}$$

where i is the object id, f is the filter, and sgn denotes the sign function, defined to be +1 for positive values and −1 for negative values. Ideally, for an SN the above will be $RV_{SN} = \{+1 +1 +1 -1 -1 -1\}$ since the SN models should have a larger LOOCV, and a smaller AICC as compared to the OU process, while for AGN the signs should be reversed. However, it is possible that inherent biases in the data or the model cause one or more of the models to perform consistently worse as compared to the OU process in fitting the SN light-curves, due to noisy difference imaging resulting from astrometric errors. However, we demonstrate that our method is robust to such biases, due to redundancies in the use of multiple models in multiple filters, to describe the SN light-curves.

To test our classification we chose a diverse set of examples to reflect the spectrum of photometric properties of the dataset covering the entire gamut of SNe and AGN lightcurves. As examples for AGN, we considered 255 AGN selected from the GALEX Time Domain Survey [11]. For SNe, we identified 3300 extragalactic candidates based on their offsets (0.4” − 1”) from their respective galactic hosts. We derived the offset limits by fitting a bi-modal distribution to the distribution of extragalactic alerts; a Gaussian distribution for the AGN parameterized by μ_{AGN}, σ_{AGN}, and for SN by μ_{SN}, σ_{SN}. All sources with offsets greater than $\mu_{AGN} + 2\sigma_{AGN}$ were designated as SNe. From this set we selected 100 SN by eye-balling their light-curves. We then constructed the vector of $RV_{i,f}$ for all verification set sources filter-wise, and use a K-means clustering supervised-machine-learning algorithm [12] that partitions the source vector into two distributions corresponding to SN and AGN. The algorithm obtains the centers of the two distributions or clusters by attempting to minimize the sum of squares of the distances of points x_j within each distribution or cluster S_i from the mean of the cluster μ_i.

$$\sum_{i=1}^{k} \sum_{x_j \epsilon S_i} ||x_j - \mu_i||^2 \qquad (3)$$

Each source is then assigned a class $C_{i,f}$ as a SN, or an AGN, depending on the center μ_i it is clustered around. The squared-distance of the source point $d_{i,f}$ from the clustering center μ_i in filter f, is a measure of how reliably it is classified as the particular type, with a distance of $x_j - \mu_i = 0$ being the best, and larger distances indicating less reliable classifications. This process is repeated for each source, in each of the g, r, i, and z bands independently. The advantage of classifying the sources filter-wise is that, a. the behavior of each source types in each filter could be very different, b. some filters may be more noisy than others thereby being less suited for classification, and c. the filters can reinforce the type of classification observed in other filters thereby leading to a more robust final classification. The final source classification C_i is decided by

$$C_i = \frac{\sum_f C_{i,f}}{4} \;\;\&\;\; d_i = \frac{\sum_f d_{i,f}}{N_{filters}} \tag{4}$$

where the sign of C_i indicates the type of the source (+1 for SN, −1 for AGN), and $|C_i|$ measures the quality of the classification. d_i is the average of squared-distances over all filters. Fig.1 and Fig.2 show the plots of C_i vs d_i for AGN and SNe, respectively. We classify AGN as having $C_i < -0.2$ and $d_i < 20$ or $C_i > -0.2$ and $d_i > 20$, while the SNe as having $C_i > -0.2$ and $d_i < 20$ or $C_i < -0.2$ and $d_i > 20$. A large value of d_i typically indicates that the source belongs to a class which is different from the one it is clustered around, *i.e.*, in this case it is the other class. The efficiency of our classification scheme on the training set is just over 90%, with 100% completeness.

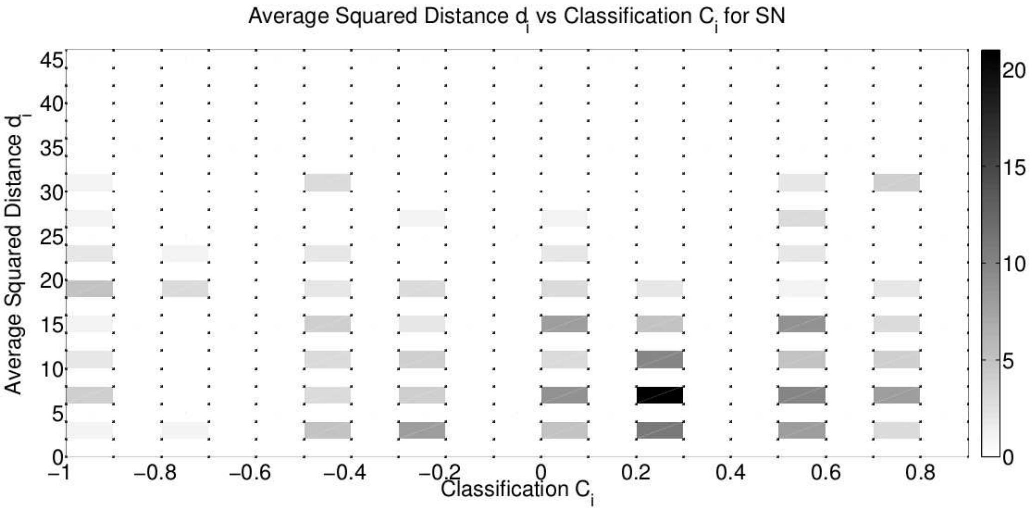

Figure 1: Distribution of 255 GALEX-TDS AGN in d_i vs C_i.

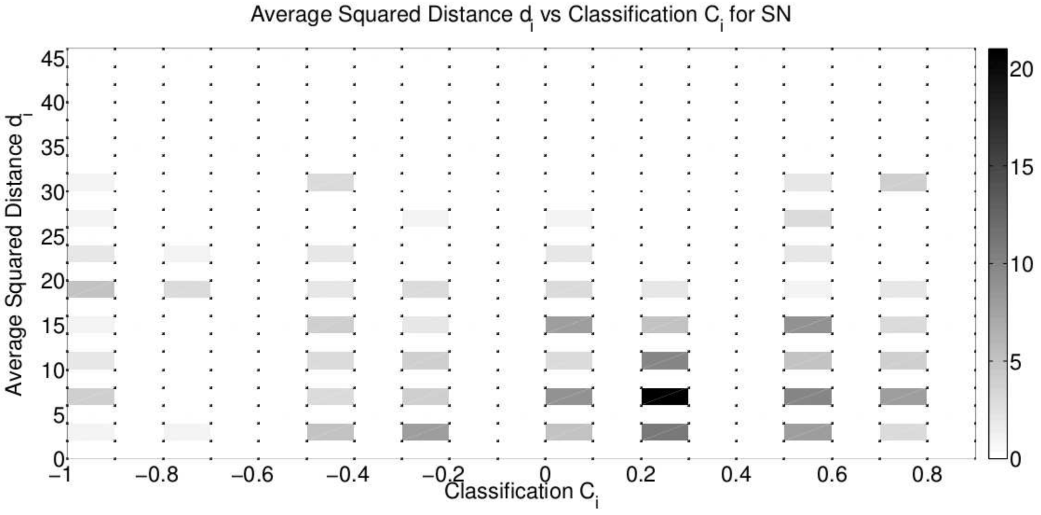

Figure 2: Distribution of 100 offset-selected SN in d_i vs C_i.

4 Conclusions and Future Work

We have discussed a Bayesian classification method to classify Pan-STARRS1 medium-deep transients identified with galactic hosts, using difference-image multi-band photometry, into SNe and AGN with 90% efficiency and 100% completeness. The methods herein can be applied to identifying AGN and SNe in existing catalogs, as well as providing real-time identification of sources in the era of Pan-STARRS2 and LSST. In addition, the method can be simply extended to the identification of particular sub-types of the broad source classes, provided their respective specific time-series models.

References

[1] G. E. Uhlenbeck and L. S. Ornstein, Phys. Rev. **36**, 823 (1930).

[2] C. A. L. Bailer-Jones, Astron. Astrophys. **546**, A89 (2012) [arXiv:1209.3730 [astro-ph.IM]].

[3] N. R. Butler and J. S. Bloom, arXiv:1008.3143 [astro-ph.CO].

[4] Y. Choi *et al.*, arXiv:1312.4957 [astro-ph.CO].

[5] R. J. Foley and K. Mandel, Astrophys. J. **778**, 167 (2013) [arXiv:1309.2630 [astro-ph.CO]].

[6] R. Andrae, D. -W. Kim and C. A. L. Bailer-Jones, arXiv:1304.2863 [astro-ph.CO].

[7] R. Kessler *et al.*, Publ. Astron. Soc. Pac. **122**, 1415 (2010) [arXiv:1008.1024 [astro-ph.CO]].

[8] M. Ganeshalingam *et al.*, arXiv:1107.2404 [astro-ph.CO].

[9] R. Chornock *et al.*, Astrophys. J. **780**, 44 (2014) [arXiv:1309.3009 [astro-ph.CO]].

[10] B. C. Kelly *et al.*, Astrophys. J. **730**, 52 (2011) [arXiv:1009.6011 [astro-ph.HE]].

[11] S. Gezari *et al.*, Astrophys. J. **766**, 60 (2013) [arXiv:1302.1581 [astro-ph.CO]].

[12] T. Kanungo *et al.*, IEEE Trans. Pat. Ana. Mach. Intel. **7**, 24 (2002)

[13] K. P. Burnham and D. R. Anderson Statistical Theory and Methods (2002)