Anomaly Detection for Multivariate Time Series of Exotic Supernovae

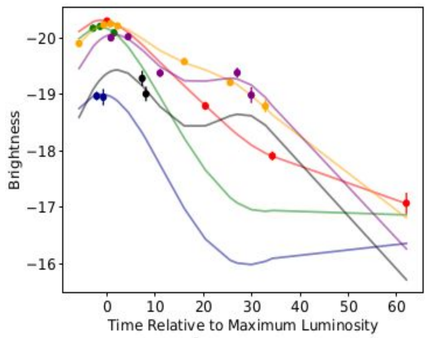
V. Ashley Villar, Miles Cranmer, Gabriella Contardo, Shirley Ho, Joshua Yao-Yu Lin
Corresponding: vav2110@columbia.edu

Abstract Supernovae mark the explosive deaths of stars and enrich the cosmos with heavy elements. Future telescopes will discover thousands of new supernovae nightly, creating a need to flag astrophysically interesting events rapidly for followup study. Ideally, such an anomaly detection pipeline would be independent of our current knowledge and be sensitive to unexpected phenomena. Here we present an unsupervised method to search for anomalous time series in real time for transient, multivariate, and aperiodic signals. We use a RNN-based variational autoencoder to encode supernova time series and an isolation forest to search for anomalous events in the learned encoded space. We apply this method to a simulated dataset of ~12k supernovae, successfully discovering anomalous supernovae and objects with catastrophically incorrect redshift measurements. This work is the first anomaly detection pipeline for supernovae which works with real time, online datastreams.

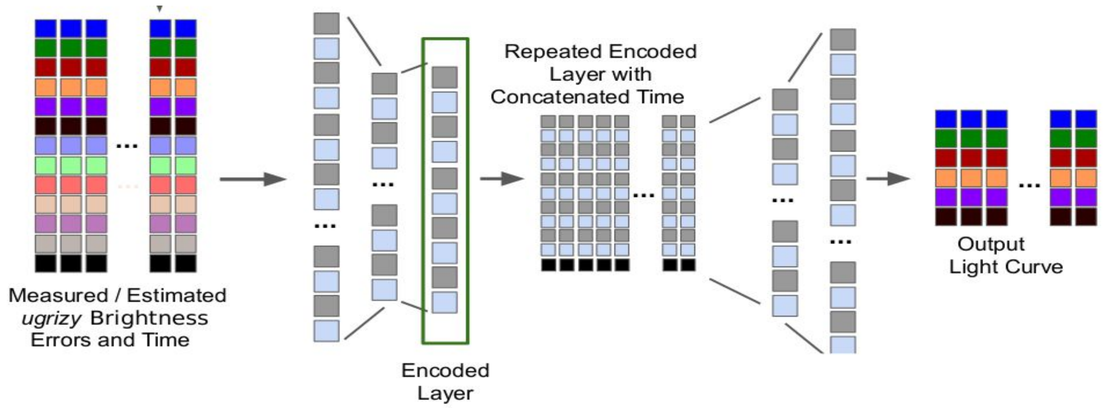
Dataset & Pre-processing

PLAsTiCC is a simulation of 3 years of Vera Rubin Observatory observational data including over 3.5 million transient (including supernovae, SNe) events from eighteen unique physical classes. Each event is a light curve made up of observations across six broadband filters (ugrizY). While the PLAsTiCC data was originally used for a Kaggle competition to classify SNe, we re-purpose this dataset as a training set for anomaly detection. Here, anomalous events will be determined by the metadata (i.e., if the event comes from a rare astrophysical origin). In total, our data set contains ~12k SNe (and SN-like) light curves from twelve extragalactic classes.

We pre-process this dataset by converting to absolute magnitude (using a simulated, noisy photo-z estimate), correcting for cosmological k-corrections and correcting for galactic dust. We then interpolate the multiband light curve using a 2D Gaussian Process over time- and wavelength-space. We use a novel distance metric (based on the Wasserstein-1 distance between filters) to interpolate across wavelength-space.



Encoding light curves



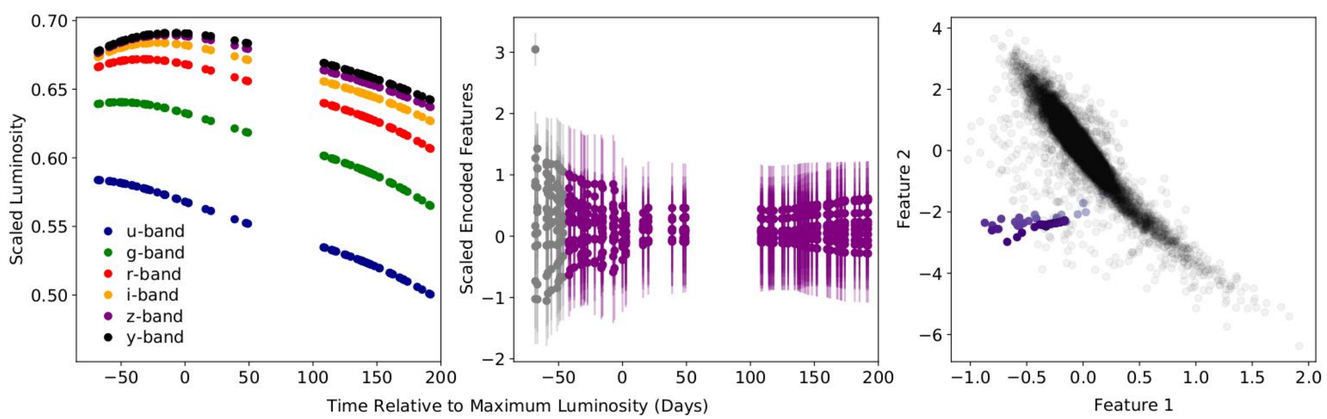
We train a recurrent variational autoencoder (RVAE) on the full dataset. The RVAE uses recurrent neurons to read in the light curve and estimated Gaussian Process errors and encodes this light curve as a vector. Before being passed into the decoder, the encoded layer is repeated N times, each time appended with a phase (defined as time since maximum light) .The decoder then produces the light curve at the specified N times. This architecture is specifically chosen so that the neural network learns the physical meaning of phase relative to maximum light for a SN.

Additionally, the unique repeat layer of our architecture allows us to call for a times not included in the real data, allowing for interpolation and extrapolation of a light curve if desired (although this feature is not used in this work).

Searching for anomalies

We search the RVAE encoded space for anomalies using an isolation forest (IF). The IF algorithm uses a series of decision trees over random attributes to isolate all events in a sample, characterizing each event with an anomaly score based on the number of trees required to isolate an event.

In the left figure, we highlight the evolution of an anomalous event (a Type I SLSN, a member of a class making up just 2% of our training set) detected by our algorithm. We track encoded features as a function of time relative to the peak luminosity of the SLSN (middle). About one month before maximum luminosity, the IF correctly identifies the event as anomalous (shown as a switch from grey to purple). We show the evolution of the anomalous SLSN in a subset of encoded space (left). Black points represent the full sample of events. The encoded values for the anomalous SLSN are shown as squares becoming increasingly purple with time. The SLSN begins in the main distribution of all events (shown in black) and evolves into an anomalous region of feature space.



VAV acknowledges funding through the Simons Foundation.