PROBABILISTIC CROSS-IDENTIFICATION OF COSMIC EVENTS

Tamás Budavári
Department of Physics and Astronomy, The Johns Hopkins University, 3400 North Charles Street, Baltimore, MD 21218, USA; budavari@jhu.edu

Received 2011 February 27; accepted 2011 May 19; published 2011 July 19

ABSTRACT

I discuss a novel approach to identifying cosmic events in separate and independent observations. The focus is on the true events, such as supernova explosions, that happen once and, hence, whose measurements are not repeatable. Their classification and analysis must make the best use of all available data. Bayesian hypothesis testing is used to associate streams of events in space and time. Probabilities are assigned to the matches by studying their rates of occurrence. A case study of Type Ia supernovae illustrates how to use light curves in the cross-identification process. Constraints from realistic light curves happen to be well approximated by Gaussians in time, which makes the matching process very efficient. Model-dependent associations are computationally more demanding but can further boost one’s confidence.

Key words: methods: data analysis – methods: statistical – supernovae: general – surveys – time

1. INTRODUCTION

Advancements in detector technology recently have opened new opportunities for time-domain astronomy. The next-generation surveys will observe most of the sky not once but many times, which enables systematic studies of variable and transient sources. The first such programs are already online. One example is the Catalina Real-Time Transient Survey (CRTS; Drake et al. 2009) that uses three dedicated telescopes and has reported over 2000 new transients to date. Another is the Palomar Transient Factory (PTF; Rau et al. 2009) that has discovered close to 800 supernovae. The Pan-STARRS telescope that started survey operation in 2010 will see thousands of supernovae in a month, not to mention the SkyMapper project (Keller et al. 2007) or the Large Synoptic Survey Telescope (LSST).

Modern surveys implement automated discovery in their pipelines. They provide live streams of events in both machine- and human-readable formats to facilitate rapid follow-up observations and real-time analysis. An interoperable representation is the VOEvent format defined by the International Virtual Observatory Alliance (see http://www.ivoa.net). Capitalizing on this standard, the http://www.VOEventNet.org project promises to deliver alerts to interested subscribers within minutes of discovery. Another important reason for standardized live streams is the joint analysis of events. Users of the http://www.SkyAlert.org site can browse for discoveries in a growing selection of data streams. The project will automatically associate any events within 10^2 and less than 1 hr apart (R. Williams 2011, private communication).

One outstanding issue is the proper statistical foundation of event aggregation. The methods developed for cross-identification of catalogs are not directly applicable. Although positional information is clearly needed for transient events one cannot statistically interpret the spatial thresholds. To make a probabilistic statement, one has to consider the density, but the host objects might not be visible—only the explosions. In fact, events are more abstract entities in spacetime, whose density we must study in 3+1 dimensions. Rigorous statistical assessments are most important in the presence of crowding, i.e., large density compared with uncertainties. Spatially, optical observations are typically much more accurate than X-ray, radio, gamma, or gravitational wave detections. Along the dimension of time, this is especially relevant because the frequency of events depends on the cutoff parameter of the detection. By lowering the threshold, one can study fainter and more distant events that, although of little significance in any one survey, could be interesting when seen in many. In this paper, my goal is to derive the probabilistic matching formalism for transient events, which enables the automated association of detections in space and time.

In Section 2, I use Bayesian hypothesis testing to evaluate the quality of candidate associations. Section 3 describes why it is not impossible to derive probabilities for these matches. In Section 4, I analyze simulated SN Ia light curves to derive time constraints and show an accurate approximation that provides analytic results. Section 5 concludes this study. Throughout the paper, the uppercase letters $P$ and $L$ represent probabilities and likelihoods, respectively, and the lowercase letter $p$ represents probability densities.

2. MATCHING EVENTS

Many things happen in the sky. Asteroids move, stars pulsate, and supernovae and gamma-ray bursts explode everywhere. Such astronomical events are typically discovered in photometric images, based on the change in the apparent brightnesses of different epochs. The imaging cadence is strategically planned to meet scientific goals. Epochs are optimized for the characteristic timescale of the intended targets, which might be weeks, days, or just tens of minutes. Looking at repeated observations, we can spot the significant changes, e.g., using a 5σ threshold. At that point, it is possible to extract the light curve of the detected event from the earlier images. Variable sources are revisited several times over the course of a survey, providing ample data. Cosmic events, however, happen only once, and we need to make the best use of the observations.

Considering that different photometric passbands potentially probe different physical processes, light curves can be quite different. Detections can be delayed from one instrument to another. One thing, however, is common: the start time of the event, which can be estimated in all detections by looking at the changing flux as a function of time.

Let us now consider a single candidate association that consists of $n$ reported events, one from each survey $i$ (where $i = 1$ to $n$). The surveys will have measured $D_i = \{x_i\}$ positions in the sky (unit vectors of the directions) and can constrain $D_i = \{t_i\}$ start times with some uncertainties. The time constraint is not
the precision of the clock, but rather the ability of the model \( M \) of the given survey to estimate the start time \( t \) of a particular event. Formally, it is a likelihood \( L(t|\tau) \equiv p(t|\tau, M) \) that is a function of the model’s \( \tau \) epoch. For now, I assume that \( L(t|\tau) \) is known for every survey. In Section 4, I take a closer look at how to derive such likelihood functions from observed light curves; in particular, I analyze Type Ia supernovae in detail.

I examine whether the detections in the given tuple are from the same event using Bayesian hypothesis testing. I wish to compare two hypotheses. The first one, call it \( H \), claims that all detections are of the same cosmic event, and hence they have a common epoch \( (\tau) \) and a common true position (both unknown). The second hypothesis is the former’s complement, \( K \), that allows for any one of the detections to be from a separate event, which I model using a set of parameters, e.g., the \( [\tau_i] \) epochs. The Bayes factor makes the comparison, which is the ratio of the probabilities of the measurements in the two hypotheses. For independent measurements, one can calculate separate Bayes factors for the position and time measurements, and the combination of the two is their product,

\[
B = \frac{p(D|H)}{p(D|K)} = \frac{p(D|H) \cdot p(D|H)}{p(D|K) \cdot p(D|K)} = B_t B_x.
\]

The \( B_t \) term is the same as that for the static objects, as discussed in detail by Budavári & Szalay (2008). The numerator and denominator of \( B_t \) are integrals of the known likelihood functions multiplied by the prior,

\[
p(D_t|H) = \int d\tau \ p(\tau|H) \prod_i L_i(\tau|i) \quad \text{and} \quad (2)
\]

\[
p(D_t|K) = \prod_i \int d\tau_i \ p(\tau_i|K) \ L_i(\tau_i|i) \quad \text{and} \quad (3)
\]

The choice of the prior naturally affects the Bayes factor, and it requires some more consideration.

Common sense dictates that the prior in the epoch \( \tau \) should be flat, because the event can occur at any time with the same probability density. The calculation of the Bayes factor, however, requires a proper prior, whose integral is unity. Considering that all relevant measurements and the non-zero support of their uncertainties can be constrained to a finite interval without any practical limitations, one can introduce a flat prior on an arbitrary but wide enough \([T, T+\Delta]\) interval and define

\[
p(\tau|\cdot) = \begin{cases} \Delta^{-1}, & \text{if } \tau \in [T, T+\Delta] \\ 0, & \text{otherwise} \end{cases} \quad \text{(4)}
\]

Using this flat prior, the likelihoods of the hypotheses become

\[
p(D_t|H) = \Delta^{-1} \int d\tau \prod_i L_i(\tau|i) \quad \text{and} \quad (5)
\]

\[
p(D_t|K) = \Delta^{-n} \prod_i \int d\tau_i \ L_i(\tau_i|i) \quad \text{and} \quad (6)
\]

where I omitted the integration limits because they are irrelevant to any large-enough interval and arrived at the dependence of

\[B_t \propto \Delta^{n-1}.\]

This means that the wider I make the interval, the larger the Bayes factor becomes, which seems to artificially contradict the expectation of an objective quality measure of the associations.

3. STREAMS OF EVENTS

Cosmic events should not be looked at in isolation. A given telescope provides an entire stream of events, which is typically published online in an automated fashion. My goal is to merge several independent streams of many surveys. The resulting combined stream will carry more information on individual events, which will help with scientific analysis. Along with the stream of associations, I want to include the probabilities.

The rate of events varies from survey to survey. For a given time interval, the number of occurrences can be quite different. Indeed, this is a key piece of information in the calculations. By definition, the posterior probability of an association is computed from the Bayes factor and the prior of the hypothesis. Assuming that the prior probability is uniform, i.e., it takes the same value independent of the position in the sky, one can write it as the ratio

\[P_0 = \frac{N_i}{\prod_{\nu_i}} \Delta^{1-n}.\]

The probability shrinks with increasing time intervals, i.e., it is less probable to randomly select the same event from larger sets. In practice, the number of events is always large; otherwise, their frequencies could not be measured. More formally, \( \Delta \) can be arbitrarily wide to yield large \( N_i \) counts and a small prior. For vanishing priors, the posterior depends only on the product of the \( P_0 \) prior and the Bayes factor:

\[P \simeq \frac{P_0 B_t}{1 + P_0 B_t},\]

and hence the dependence on \( \Delta \) actually cancels in the product of

\[P_0 B_t = \left( \frac{\nu_i}{\prod_{\nu_i}} \left[ \prod_{i} \int d\tau_i \ L_i(\tau_i|i) \right] \right) = \pi_0 \beta_i.\]

The second term of the product, \( \beta_i \), can be directly calculated from the data. The value of \( \nu_i \) in \( \pi_0 \) can also be estimated over time from the ensemble statistics of the streams. Recall the self-consistency argument of Budavári & Szalay (2008): the sum of the priors over all possible \( \prod N_i \) combinations, which is total \( N \), by definition, is also equal to the sum of the posteriors. This equality is an equation for \( \nu_* \), or \( \nu_* \) in this case, that in turn provides well-determined prior and posterior probabilities. In summary, the posterior is well defined and calculated as

\[P \simeq \frac{\pi_0 \beta_i B_t}{1 + \pi_0 \beta_i B_t}.\]

To select candidate associations early, when not enough statistics are available to reliably solve for \( \nu_* \), one can use the upper bound of \( \nu_*^{\text{max}} \) to overestimate the posteriors.
Similarly, if \( \nu_i \) are uncertain, one can use conservative limits instead. By safely overestimating the prior, one ends up with somewhat larger posteriors than the actual ones; thus, a fixed probability cut yields slightly more candidates that can later be pruned. This will not overshoot the values by too much because large posteriors are not particularly sensitive to small variations in the prior (Heinis et al. 2009).

4. CONSTRAINTS FROM LIGHT CURVES

One cannot directly measure the epoch of an event, but only the fluxes as a function of time. If one can model the change, these light curves can provide time constraints. For example, a Type Ia supernova is described by its absolute magnitude \( M \), redshift \( z \), and epoch \( \tau \). For any given set of such parameters, one can derive the fluxes at the times of pre-scheduled observations and compare them to reality. The likelihood function is given by the deviation of the photometric measurements \( f \) and the simulated fluxes \( f(\tau, M, z) \) as

\[
L(\tau, M, z|f) = N(f - f(\tau, M, z), \Sigma),
\]

where \( \Sigma \) represents the photometric errors in the normal distribution \( N \). By integrating over \( M \) and \( z \) with their proper priors, one can derive the \( \tau \) dependence needed for \( \beta_i \).

I use a Gaussian luminosity function (Dahlen et al. 2004) with \( M_* = -19 \) and \( \sigma_M = 0.5 \) for the prior and simulate a typical \( M_* \), SN Ia using the LSST cadence (Ivezic et al. 2008). In this cadence, pairs of observations separated by 1 hr are repeated every 3 days. Throughout this model, I use precision LSST light curve templates from Peter Nugent (via A. Connolly 2011, private communication) but consistently neglect the \( K \)-correction. The error on the \( g \)-band photometry is determined from the \( 5\sigma \) limiting magnitude of 25. The simulated event is set to happen at day 0. The luminosity function determines the kind of SN Ia events that can occur, and the decline at high redshifts is set by the limiting magnitude of the survey. Together, these set the integration domain for the calculations. The peak brightness of the simulated supernova at redshift \( z = 0.8 \) is 24.5 mag in the \( g \) band. This distance is approximately at the peak of the redshift distribution.

Figure 1 shows the likelihood as a function of the epoch for a noiseless set of observations. The results are not sensitive to the end date as long as the integration limit is at the far tail of the decay. The measurements (open circles) follow a Gaussian surprisingly well. The solid line helps to guide the eye: the curve has \( \sigma = 1 \) day and a mean of 0.1 day. Random realizations of light curves with realistic noise yield similar profiles with maxima whose distribution is consistent with the error. To alert the transient community, the convention is to have two independent observations that pass the detection threshold. For LSST-like cadences, these most likely will happen on the same night, 1 hr apart. As part of the notification message, a light curve can be published that consists of the measurements up to the last two \( 5\sigma \) observations. The crosses show results for such truncated light curves, which also closely follow a Gaussian, with \( \sigma = 3 \) days and a 0.5-day mean (dotted line). The shorter time baseline and marginal photometry provide significantly less leverage on the time constraint. I also see more deviation and a slight asymmetry in the points. The offset in the peak at these faint magnitudes is a property of the light curve’s shape and the cadence. When the photometric errors are tighter, the shift (and the width) is (are) significantly smaller. In comparison, for a \( z = 0.5 \) supernova (with the same \( M_* \) luminosity), the brighter light curve yields a negligible offset and an \( \sigma \sim 0.25 \) day. More frequent sampling also decreases the offset, but it is not practical in real life. A better method is to calibrate the effect and correct for the shifts, if needed.

Using Gaussian likelihoods, the integrals in \( \beta_i \) can be analytically calculated. For a two-way association, \( \beta_i \) becomes

\[
\beta_i = \frac{1}{\sqrt{2\pi (\sigma_1^2 + \sigma_2^2)}} \exp \left\{ -\frac{(t_1 - t_2)^2}{2(\sigma_1^2 + \sigma_2^2)} \right\}. \tag{14}
\]

This is not a Bayes factor and is not dimensionless. Substituting \( \sigma = 1 \) day and \( \Delta t = 0 \) yields \( \beta_i = 1/\sqrt{4\pi} \) day\(^{-1} \), which is approximately 0.282 day\(^{-1} \). The numerical integral of the photometric likelihood gives a similar value, 0.287 day\(^{-1} \), which is in good agreement with the analytical estimate. The left panel of Figure 2 illustrates \( \beta_i \) as a function of \( \Delta t \) for Gaussians with the aforementioned \( \sigma \) values. The solid line shows results from cross-matching events with complete light curves, and the dotted line is for alerts whose light curves are truncated. The dashed line illustrates the case of cross-matching an event with a full light curve and another with truncated photometric measurements. The right panel shows the same results when the likelihoods are approximated by top-hat functions with the same standard deviations. The finite support of these functions results in \( \beta_i = 0 \) day\(^{-1} \) at large separations in time, which is cause for concern if the width of the window is not well understood. For a \( z = 0.5 \) redshift \( M_* \), SN Ia with \( \sigma = 0.25 \) day, the value is four times larger, \( \beta_i = 1.128 \) day\(^{-1} \), when seen at the same time.

These values of \( \beta_i \) are not large, especially considering that \( \sigma_0 \) is typically fairly small, because the time constraints are weak. The frequency of all events observed by a telescope can be tens of thousands, say, up to \( v \sim 100,000 \) day\(^{-1} \), when including all transients. Since the true SN Ia rate is expected to be around
finite support of these functions means that is cause for concern if the width of the time window is not well understood.

Figure 2. When cross-matching streams of events, the $\beta_t$ parameter is the time-domain analog of the Bayes factor. Its dependence on the time difference is plotted for the two-way Gaussian case in the left panel, with the uncertainties seen in Figure (1). The line styles are identical to those in Figure 1. The dashed line is for matching one against the other. For comparison, the right panel shows the same plot using top-hat likelihood approximations with the same standard deviations. The finite support of these functions means that $\beta_t = 0$ day$^{-1}$ at large separations in time, which in turn yields 0 posterior probability. The loss of associations, in this case, is cause for concern if the width of the time window is not well understood.

$v_* \sim 1000$ day$^{-1}$ in LSST, our estimate for the lower bound comes to $\pi_0 \sim 10^{-7}$ days. Nevertheless, even for the fainter sources, we have higher $P_i \sim 10^{-8}$ values than in the blind matching of static sources, where $P_0 \sim 10^{-10}$ (Heinis et al. 2009). We can increase our prior by filtering the input streams, but the filtering might be better done on the associations to find the low-significance events in the merged stream. In practice, a certain type of filtering of the streams can happen before merging, while other filters will have to run on the associations.

For completeness, I mention that the general result of the $n$-way Gaussian case is also analytically calculated. The resulting formula is similarly simple and reads

$$
\beta_t = (2\pi)^{1/2} \left( \frac{\prod w_i}{\sum w_i} \right) \exp \left\{ \frac{\left( \sum w_i t_i \right)^2}{2 \sum w_i} - \frac{\sum w_i t_i^2}{2} \right\},
$$

(15)

where $w_i = 1/\sigma_i^2$ are the precision parameters of the $t_i$ epochs.

Photometric measurements also can be directly used to calculate $\beta_t$ using the multivariate likelihood function in Equation (13). In addition to a common epoch, identical events have the same physical properties as well, e.g., the absolute magnitude and the redshift of the supernova. My hypotheses $H$ and $K$ remain the same as before, but I parameterize them with not only $\tau$ but also $M$ and $z$. The calculation of $\beta_t$ involves computing the integrals over all three parameters. Of course, the numerical integral is substantially more expensive computationally than evaluating an analytic formula, but this is the proper use of the model and fully exploits its constraining power. Following the same prescription as before, I generated two light curves using the LSST cadences. The only difference was in their phase: the cycle of one simulated survey was shifted by a day from the other. The result for an $M_*$ supernova at $z = 0.8$ redshift is almost 20 times larger than one based on time coincidence, $\beta_t = 5.3$ day$^{-1}$. If one is willing to live with the extra computation and the explicit dependence on a model, this method can boost the probability, in this case, to almost 1%. Naturally, one would still need spatial constraints for any meaningful match, but one could potentially leverage lower-accuracy positional measurements.

The explicit modeling of the light curve provides an opportunity to simultaneously perform classifications and outlier detection. One can benefit from calculating several Bayes factors or $\beta_t$ values for comparison. Associations that appear to be reliable matches, based on their directions and epochs when using a model-independent approximate time window, might be rejected based on a detailed analysis of their spectral energy distributions (SEDs). In this finding accidental noise or a new discovery? Follow-up observations may help to decide. Beyond the above outlier detection, one also can run several SED models to find out in which class the candidate event might fit best. Naturally, there is also nothing to stop one from creating several merged streams by using different strategies, that target specific scientific experiments and communities.

5. CONCLUSIONS

I introduced a new statistical method to associate cosmic events based on measured constraints in both time and space. Including time-domain information is important for a number of reasons: (1) if one can automatically cross-correlate streams of events from multiple surveys, one can find fainter ones that are not obvious in any one of the observations but become significant together. This enables one to push these studies to uncharted territories. (2) Constraints on the epoch can boost small positional evidence at low accuracies, e.g., gamma-ray bursts. (3) A more fundamental yet extremely important practical matter is the assignment of probabilities. While one can clearly calculate Bayes factors based only on positional information, one cannot define the probabilities without time-domain data. Conceptually, the problem is with the prior. It cannot be determined without counting the occurrences of events in a fixed time interval whose duration is formally arbitrary. I found that the solution is to include the Bayes factors in time, which naturally cancels the artifact.

The light curve measurements constrain the epoch of a given event. For Type Ia supernovae, we found the likelihood to be very close to a Gaussian, which makes the problem analytically tractable. The numerical simulations agree with the
fast approximations. The same method also works for streams of events seen in different wavelengths, as long as the regime is covered by the SED models. To unlock the full potential of the light curves, one must further exploit the SED modeling. The calculation is extended to require matching events to have the same physical properties. For SN Ia, the numerical integrals show that this strategy has the potential to boost the time-domain evidence by a factor of 20, which means that one can get away with less accurate spatial measurements.

Although I illustrated the methodology on supernovae, it is clearly applicable to other types of events. Before gamma-ray bursts were established to be extragalactic, Luo et al. (1996) applied Bayesian hypothesis testing to look for repeating gamma-ray bursts. Since then, researchers have seen the host galaxy of many gamma ray bursts. Using time-domain cross-identification, one can look for other subtle events of related physical processes in hosts. Another example is the detection of gravitational waves. The LIGO Collaboration discusses astrophysically triggered searches (Abbot et al. 2008). The idea is to lower the detection threshold for a short period of time, when other types of cosmic events occur. This is in fact a very similar problem; merging streams of events from LIGO and other projects with my probabilistic approach could boost their significance and help facilitate new discoveries.

The general cross-identification problem in astronomy is rather convoluted and ultimately tightly interwoven with scientific analyses. One obvious example is classification, which in turn affects the matching prior. The circular nature of the problem is not an issue, but one needs to find a consistent solution. Building on explicit assumptions in the presented probabilistic approach, it is straightforward to incorporate different kinds of data in new observations, when available. This is a crucial feature that will prove vital in studying the time domain and understanding the eventful universe.

The author thanks Andy Connolly, Alex Szalay, István Csabai, Matthew Graham, Suvi Gezari, Josh Bloom, Joey Richards, Roy Williams, and Tom Loredo for several stimulating discussions on various aspects of the topic. This study was supported by the Gordon and Betty Moore Foundation via GBMF 554.

REFERENCES

Abbott, B., et al. 2008, Class. Quantum Grav., 25, 114051
Budavári, T., & Szalay, A. S. 2008, ApJ, 679, 301
Dahlen, T., et al. 2004, ApJ, 613, 189
Drake, A. J., et al. 2009, ApJ, 696, 870
Heinis, S., Budavári, T., & Szalay, A. S. 2009, ApJ, 705, 739
Ivezić, Z., et al. 2008, Serbian Astron. J., 176, 1
Keller, S. C., et al. 2007, PASA, 24, 1
Luo, S., Loredo, T., & Wasserman, I. 1996, in AIP Conf. Proc. 384, Gamma-ray Bursts: 3rd Huntsville Symp., ed. C. Kouveliotou, M. F. Briggs, & G. J. Fishman (Melville, NY: AIP), 477
Rau, A., et al. 2009, PASP, 121, 1334