Faint Object Detection in Multi-Epoch Observations via Catalog Data Fusion

Tamás Budavári¹, Alexander S. Szalay¹, and Thomas J. Loredo²

¹Department of Physics and Astronomy, The Johns Hopkins University, 3400 North Charles Street, Baltimore, MD 21218, USA
²Cornell Center for Astrophysics and Planetary Science, Cornell University, Ithaca, NY 14853, USA

Received 2016 November 10; revised 2017 February 22; accepted 2017 February 22; published 2017 March 23

Abstract

Astronomy in the time-domain era faces several new challenges. One of them is the efficient use of observations obtained at multiple epochs. The work presented here addresses faint object detection and describes an incremental strategy for separating real objects from artifacts in ongoing surveys. The idea is to produce low-threshold single-epoch catalogs and to accumulate information across epochs. This is in contrast to more conventional strategies based on co-added or stacked images. We adopt a Bayesian approach, addressing object detection by calculating the marginal likelihoods for hypotheses asserting that there is no object or one object in a small image patch containing at most one cataloged source at each epoch. The object-present hypothesis interprets the sources in a patch at different epochs as arising from a genuine object; the no-object hypothesis interprets candidate sources as spurious, arising from noise peaks. We study the detection probability for constant-flux objects in a Gaussian noise setting, comparing results based on single and stacked exposures to results based on a series of single-epoch catalog summaries. Our procedure amounts to generalized cross-matching: it is the product of a factor accounting for the matching of the estimated fluxes of the candidate sources and a factor accounting for the matching of their estimated directions. We find that probabilistic fusion of multi-epoch catalogs can detect sources with similar sensitivity and selectivity compared to stacking. The probabilistic cross-matching framework underlying our approach plays an important role in maintaining detection sensitivity and points toward generalizations that could accommodate variability and complex object structure.

Key words: astrometry – catalogs – galaxies: photometry – methods: statistical – surveys

1. Introduction

In the era of time-domain survey astronomy, dedicated telescopes scan the sky every night and strategically revisit the same area several times. The raw data are images, but surveys commonly provide not only image data but also catalogs, which are summaries of the image data that aim to enable a wide variety of studies without requiring users to analyze raw or processed image data. Catalogs typically report object properties based on algorithms that detect sources in images with a measure of statistical significance above some threshold, chosen so that the resulting catalog is likely to be highly pure (i.e., with few or no spurious sources).

The question we address in this paper is how to combine information from a sequence of independent observations to maximize the ability to detect faint objects at or near a chosen detection threshold, which also helps with the data explosion due to false detections at low signal-to-noise thresholds. Focusing on the faint objects that typically dominate the collected data is an important and timely problem for a number of ongoing surveys and is vital for planning the next-generation data-processing pipelines.

There are two different ways to approach the problem. A traditional, resource-intensive approach is to wait until all of the observations are completed and then performing detection by stacking the multi-epoch image data (with potential complications related to registration, resampling, and point-spread function or PSF matching). An optimal procedure for threshold-based detection with image stacks was introduced by Szalay et al. (1999). Once a master object catalog is produced from the stacked images, time series of source measurements are created by forced photometry at the master catalog locations.

An alternative (nonexclusive) approach is to perform source detection independently for each observation, producing a catalog of candidate sources at each epoch (Budavári & Szalay 2014). The detection threshold may be different for each epoch. Interim object catalogs may be produced by analysis of the series of overlapping source detections potentially associated with a single object using any available data; a final catalog would be built using the catalogs from all epochs. Of course, a catalog based on data from many epochs should include many dim sources that escape confident detection in single-epoch or few-epoch catalogs. To enable construction of a deep multi-epoch catalog, the single-epoch catalogs must report information for candidate sources with relatively low statistical significance; i.e., the single-epoch catalogs must have reduced purity. If, on the one hand, we set the single-epoch threshold too high, there will be too few detections; we will not have well-sampled time series for dim sources, and the final catalog will be too small. If, on the other hand, we set the threshold too low, the single-epoch catalogs will be overwhelmed with (mostly false) detections that were seen only once, requiring wasteful expenditure of storage and computing resources for constructing multi-epoch catalogs. An optimal threshold might preserve the size and quality of the final catalog while enabling users to build interim catalogs on the fly that are potentially tailored to specific, evolving needs.

Here we address the second alternative, considering how best to accumulate evidence from possibly marginal detections while the observations are in progress and to prune spurious source detections but keep the sources associated with genuine objects. The study presented here is exploratory, intended to establish the basic ideas and provide initial metrics for studying the feasibility of the incremental approach. To make the analysis analytically tractable and the results easy to interpret,
we restrict ourselves to an idealized setting; we will present a more general and formal treatment in a subsequent paper.

2. Detection Probabilities

We adopt the terminology of Large Synoptic Survey Telescope (LSST) and other time-domain synoptic surveys, using "source" to refer to single-epoch detection and measurement results and "object" to refer to a unique underlying physical system (e.g., a star or galaxy) that may be associated with one or more sources. (Here we limit ourselves to objects that would appear as a single source.) For simplicity, we consider observations in a single band unless stated otherwise.

Consider an object with constant flux \( f \) and direction \( \mathbf{n} \) (a unit-norm vector on the sky corresponding to the celestial coordinates). At each epoch \( \epsilon \), analysis of the image data \( D_\epsilon \) corresponding to a small patch of sky of solid angle \( \Omega \) produces a source likelihood function (SLF) for the basic observables, flux \( f \) and direction \( \mathbf{n} \), of a candidate source in the patch. The SLF is the probability for the data as a function of the (uncertain) values of the observables,

\[
L_\epsilon(f, \mathbf{n}) \equiv p(D_\epsilon|f, \mathbf{n}, \mathcal{C}),
\]

where \( \mathcal{C} \) denotes various contextual assumptions influencing the analysis, e.g., the specification of the photometric model and information about instrumental and sky backgrounds. (Since \( \mathcal{C} \) is common to all subsequent probabilities, we henceforth consider it as implicitly present.) For example, if photometry is done via PSF fitting with weighted least squares, and if the noise level and backgrounds are known, then it may be a good approximation to take \( L_\epsilon(f, \mathbf{n}) \propto \exp(-\chi^2(f, \mathbf{n})/2) \), where \( \chi^2(f, \mathbf{n}) \) is the familiar sum of the squared weighted residuals as a function of the source flux and direction.

We consider a catalog at a given epoch to report summaries of the likelihood functions for candidate sources that have met some detection criteria. The most commonly reported summaries are best-fit fluxes (or magnitudes) with a quantification of flux uncertainty (typically a standard error or the half-width of a 68% confidence region) and, separately, best-fit sky coordinates with an uncertainty for each coordinate. We take these summaries to correspond to a factorized approximation of the SLF,

\[
L_\epsilon(f, \mathbf{n}) \equiv p(D_\epsilon|f, \mathbf{n}) = \ell_\epsilon(f)m_\epsilon(\mathbf{n}),
\]

where the epoch-specific flux factor, \( \ell_\epsilon(f) \), is a Gaussian with mode \( \hat{f}_\epsilon \) (the catalog flux estimate at epoch \( \epsilon \)) and standard deviation \( \sigma_f \), and the direction factor, \( m_\epsilon(\mathbf{n}) \), is an azimuthally symmetric bivariate Gaussian with mode \( \hat{n}_\epsilon \) and standard deviation \( \sigma_n \). This may be a rather crude approximation; we will address it further elsewhere, here merely noting that it is implicitly adopted for most survey catalogs. For simplicity, we take the flux factors to have the same standard deviation at all epochs, \( \sigma_f = \sigma_n \).

We adopt a simple source detection criterion: a candidate source with a flux likelihood mode \( \hat{f}_\epsilon \) larger than a single-epoch threshold value, \( f_{\text{th}} \), is deemed a detection. The probability of detection in a single-epoch catalog is the probability that a source with true flux \( f \) will produce a single-epoch measurement \( \hat{f}_\epsilon \) that falls above the threshold. This probability is just the integral of the Gaussian flux likelihood function above the threshold, which we denote by

\[
P_f = P(\hat{f}_\epsilon > f_{\text{th}}|f) = \frac{1}{2} \text{erfc} \left( \frac{f_{\text{th}} - f}{\sigma_f \sqrt{2}} \right),
\]

where \( \text{erfc}(\cdot) \) is the complementary error function.

For comparison, consider the detection probabilities in the case of stacked exposures from \( k \) observations. We assume that the objects are stationary and have a constant flux and that the dominant source of noise is still the sky, so the relative noise is reduced by \( \sqrt{k} \) after stacking. For a stacked exposure flux threshold \( f'_{\text{th}} \), the probability of detection is

\[
P_f' = \frac{1}{2} \text{erfc} \left( \frac{f'_{\text{th}} - f}{\sigma_f \sqrt{2/k}} \right).
\]

Figure 1 displays the detection probability as a function of the true flux for single-epoch and stacked data for various single-epoch and stacked thresholds. The green dotted lines represent the single-exposure situation \( P_f \) as a function of the true flux in \( \sigma \) units for detection thresholds of \( 2\sigma, 3\sigma, 4\sigma, \) and \( 5\sigma \). Similarly, the yellow solid lines correspond to the stacked detections with \( k = 9 \) exposures.

Consider two curves corresponding to the same threshold, so \( f'_{\text{th}} = f_{\text{th}} \). The probability of detection at \( f = f_{\text{th}} \) is 50% for both a single exposure and stacked exposures. But the curve for stacked exposures is much steeper, with a higher probability of detecting sources brighter than \( f_{\text{th}} \) and a lower probability of detecting sources dimmer than \( f_{\text{th}} \). That is, when constructed with a common threshold, the catalog built from stacked data
will be more complete above threshold and will more effectively exclude sources with a true flux below threshold.

2.1. Multiple Detections and Nondetections

Faint sources will not always be detected. The probability of making $n$ detections among $k$ observations ($k \geq n$) follows the binomial distribution, giving the multi-epoch detection probability

$$P(n|k, f) = \binom{k}{n} P_f^n (1 - P_f)^{k-n}. \quad (5)$$

An interesting quantity is the probability that an object will lead to source detections in $n_0$ or more observations. This is simply the sum

$$P(n \geq n_0|k, f) = \sum_{n=n_0}^k \binom{k}{n} P_f^n (1 - P_f)^{k-n} \quad (6)$$

(this can be expressed in terms of the incomplete beta function). In Figure 2, we plot these probabilities as a function of $f$ in units of $\sigma$ for $k = 9$ observations. From left to right, the solid red curves show the probability of detecting an object of a given flux in exactly 1, 2, etc., up to 9 observations. Similarly, the dashed blue lines correspond to the 1+, —i.e., 1 or more—2+, 3+, etc., cases.

Figure 2. Detection probability in multiple observations as a function of the true flux. Assuming $k=9$ total available observations, the solid red curves show the probability of the object appearing in exactly 1, 2, etc., up to 9 observations (from left to right). Similarly, the dashed blue lines correspond to the 1+, —i.e., 1 or more—2+, 3+, etc., cases.

Figure 3 compares the detection probability curves for the stacked exposure case (solid yellow lines, as in Figure 1) and the multi-epoch, $n \geq n_0$ detection case (dashed blue lines, as in Figure 2). For a particular stacked exposure case, we see that there is a multi-epoch case whose detection probability curve displays a very similar performance. For example, the $3\sigma$ stacked exposure curve is very similar to the multi-epoch $5+$ detection case. This indicates that collecting sources with $5+$ detections from single-epoch $3\sigma$ catalogs is nearly equivalent in terms of catalog completeness and purity to producing a separate, new $3\sigma$ stacked exposure catalog.

Analyzing the single-epoch catalogs has a number of advantages. It can be implemented in an incremental fashion that follows the schedule of the survey, and the time series data are readily available at a given time; there is no need to go back to old images and perform forced photometry at locations that are revealed only in the final stack.

3. Distinguishing Real Sources from Noise Peaks

The previous calculations addressed the detectability of a source of known true flux $f$. In real-life scenarios, the problem is quite the opposite—we are presented with the observations and would like to understand the properties of the sources. In this context, our focus is on how we can reliably distinguish noise peaks from real sources. It is important to emphasize that we have more information than just the fact that a source has been detected; we also have flux measurements at multiple epochs. Our approach is motivated by Bayesian hypothesis testing, where the strength of the evidence for the presence of a source is quantified by the posterior probability for the source-present hypothesis, or, equivalently, by the posterior odds in favor of a source being present versus being absent (the odds are the ratio of probabilities for the rival hypotheses). The posterior odds are the product of prior odds and the data-dependent Bayes factor. The prior odds depend on population properties; they may be specified a priori when there is sufficient knowledge of the population under study or learned adaptively by using hierarchical Bayesian methods (for examples of this in the related context of cross-identification,
We are presuming the use of a fixed source detection algorithm, e.g., fixed apertures. If the algorithm is adaptive, its tuning parameters need to be accounted for in this probability.

The Bayes factor is the ratio of marginal likelihoods for the competing hypotheses, one that claims that the sources are associated with a real object and its complement that assumes there is just noise:

$$B = \frac{\mathcal{M}_{\text{real}}}{\mathcal{M}_{\text{noise}}}.$$  \hspace{1cm} (7)

Each marginal likelihood, $\mathcal{M}_{\text{hyp}}$, is the integral, with respect to all free parameters for the hypothesis, of the product of the likelihood function and the prior probability density for the parameters.

Let us now assume that out of $k$ observations, we measure $n$ detections with measured fluxes $\{\hat{f}_j\}$. We consider the two competing hypotheses separately.

### 3.1. Real-Object Hypothesis

Let $\mathcal{D}$ denote the set of indices for epochs with detections and $\mathcal{N}$ denote the set of indices for epochs with nondetections:

$$\mathcal{D} \equiv \{e: \hat{f}_e > f_{\text{th}}\},$$

$$\mathcal{N} \equiv \{e: \hat{f}_e \leq f_{\text{th}}\}. \hspace{1cm} (8)$$

For a candidate object with $n$ source detections among $k$ catalogs, the likelihood of a candidate true flux $f$ is

$$\mathcal{L}(f) = (1 - P_f)^{k-n} \prod_{e \in \mathcal{D}} \ell_e(f), \hspace{1cm} (9)$$

where $(1 - P_f)$ is the probability of not detecting an object with true flux $f$, which happens $(k - n)$ times, and $\ell_e(\cdot)$ is the flux likelihood function defined above (Gaussians with means equal to $\hat{f}_e$). The marginal likelihood for the real-object hypothesis is obtained by averaging over all possible true flux values, $f$. For an object that is a member of a population with known flux probability density $\pi(f)$, the prior probability for $f$, used for the averaging in the marginal likelihood, is $\pi(f)$, so that

$$\mathcal{M}_{\text{real}} = \int df \, \pi(f) \mathcal{L}(f), \hspace{1cm} (10)$$

which is a one-dimensional integral that can be quickly evaluated analytically or numerically. (When the population distribution is not known a priori, it may be estimated via joint analysis of the catalog data for many candidate objects within a hierarchical model, a significant complication that we will elaborate on elsewhere.)

### 3.2. Noise Peak Hypothesis

The alternative hypothesis is that the detections are simply random noise peaks in the image. The noise hypothesis marginal likelihood, $\mathcal{M}_{\text{noise}}$, is the probability of the catalog data presuming no real object is present.

For epochs with a candidate source reported in the catalog, the datum is the flux measurement, $\hat{f}_e$, and the relevant factor in the marginal likelihood is $\nu(\hat{f}_e) \overset{\text{def}}{=} p(\hat{f}_e|\text{Noise})$, the noise peak distribution, evaluated at $\hat{f}_e$. This distribution will depend on the noise statistics for each catalog.

For epochs with no reported detection, we instead know only that $\hat{f}_e \leq f_{\text{th}}$, so the relevant factor is the fraction of missed noise peaks,$^5$

$$Q_N \equiv \int_{f_{\text{th}}}^{\hat{f}_e} \nu(f). \hspace{1cm} (11)$$

The probability of a false detection is then $P_f = 1 - Q_N$.

To compute these quantities comprising $\mathcal{M}_{\text{noise}}$, we need to know the noise peak distribution, $\nu(f)$. This distribution is not trivial to specify; it will depend both on the noise sources and the source detection algorithm. Typically, a source finder performs a scan, identifying local peaks of the measured fluxes smoothed with a kernel, e.g., corresponding to a specified point source aperture. Under the noise hypothesis, the source finder will find peaks of a smooth random field. The locations and amplitudes of the peaks will form a point process, whose statistical properties can be analytically calculated (Adler 2010; Bardeen et al. 1986; Bond & Efstathiou 1987; Kaiser 2004). The most important consequence is that, even though the underlying noise at the pixel level may be independent and Gaussian, the noise hypothesis output will correspond to sampling from a point process with a more complicated distribution of fluxes. In particular, although the pixel-level noise distributions are symmetric (about the mean background), the distribution for (falsely) detected fluxes is skewed toward positive values. The relevant calculation is presented in the Appendix.

Figure 4 shows the surface density of the noise peaks as a function of the detection statistic $\hat{f}$ (in $\sigma$ units) in the scenario when the sky noise is spatially independent and Gaussian. The surface density is in units of objects per $a^2$, where $a$ is the width of the PSF (see Appendix). The noise peak distribution is the normalized version of this function. The surface density has a mode at $\hat{f} \approx 1.33 \sigma$, and its shape is well approximated with a Gaussian with standard deviation $\approx 0.835 \sigma$ for all positive values of $\hat{f}$. The shaded magenta area highlights the excess density over the Gaussian at negative $\hat{f}$ values.

As noted above, we obtain the probability of detecting a noise peak, $P_N$, by integrating $\nu(f)$ above the flux threshold. Figure 5 shows the results as a function of the flux threshold in $\sigma$ units, see panel (a), and on a scale corresponding to an LSST-like magnitude (panel (b); see Section 4 for a description of the magnitude scale). We see that the fraction of noise peaks above threshold is about 62% at $1 \sigma$, dropping quickly to about a few percent at $3 \sigma$ and becoming negligible at $5 \sigma$. Based on just this figure, it is tempting to set a high detection threshold to reject such "ghost peaks" and keep the catalog of detections nearly pure, but that would mean we lose the opportunity to recover the numerous really faint sources.

Our multi-epoch approach suggests a different strategy: instead of seeking to make the catalogs for each epoch pure, we can adopt a lower single-epoch threshold, relying on the fusion of data across epochs to weed out ghosts. The marginal likelihood and Bayes factor computations accomplish this data fusion.

$^5$ The integral is over possible measured flux values, $\hat{f}$, not true flux; in the Gaussian regime assumed here, the measured value may be negative, albeit with small probability. The estimated flux would be constrained to be positive via the prior density, $\pi(f)$, which would multiply the flux likelihood when computing posterior flux estimates.
The astrometric cross-match Bayes factor, $\mathcal{B}_{\text{flux}, \text{pos}} = B_{\text{flux}} \cdot B_{\text{pos}}$, has been derived in Budavári & Loredo (2015) (see Equation (17) there, as well as Equation (19) for the tangent plane Gaussian limit that holds for high-precision astrometry). That work also discusses the generalizations that account for proper motion and other complications.

In the following section, we assess the discriminative power of multi-epoch source detection by applying it to simulated galaxies and noise peaks, both omitting and including the astrometric data.

4. Simulations

Here we describe the simulations that demonstrate the detection capability of our multi-epoch approach in a setting with a known ground truth. The simulation parameters were chosen to produce data similar to those provided by modern large-scale optical surveys.

4.1. Simulated Galaxies

We assume that the galaxies are brighter than 28 mag and that the 5σ detection limit is 24 mag, corresponding roughly to the parameters of LSST photometry. Panel (b) of Figure 5 shows the noise peak detection probability as a function of magnitude based on these parameters, in contrast to the dimensionless presentation in panel (a). To compute the marginal likelihood for the source-present hypothesis, we must specify a prior for the source flux, $f$. Here we use a standard faint-galaxy number-counts law, with the number counts following the empirical formula of $dn \propto 10^{0.4m - dm}$; see Madau & Thompson (2000). That approximately translates to the properly normalized population distribution of

$$
\pi(f) = \begin{cases} f_{\text{l}} / f^2 & \text{if } f > f_{\text{l}}, \\ 0 & \text{otherwise,} \end{cases}
$$

where $f_{\text{l}}$ is the limiting flux that corresponds to the previously defined magnitude limit.

We generate sets of random detections for 20,000 galaxies with true fluxes between 28 and 23 mag by simply drawing $\hat{f}$ values from a Gaussian centered on the actual fluxes. We also generate 2000 ghost detection $\hat{f}$ values from $\nu(\hat{f})$ by inverting its cumulative distribution (computed numerically on a grid). The number of exposures is set to the previously used $k = 9$ with a single-epoch flux threshold of just 1σ, deep in the noise.

In observations with our specified parameters, the number of ghost detections will greatly outnumber the galaxy detections with this low threshold. The numbers of galaxies and ghosts were chosen to enable display of the distributions of Bayes factors for the two classes of detections (noise and true).

We first analyze the simulated data considering only the photometric information (i.e., ignoring the directional Bayes factors). In Figure 6, the red points represent the resulting Bayes factors for the real sources (right panel) and the noise peaks (left panel). Superficially, the Bayes factors may appear surprisingly large; even for dim sources, the Bayes factors are often $\sim 10^2$, typically considered strong evidence in settings where the competing hypotheses are assigned prior odds of unity. But here the prior odds for a genuine association versus a noise peak match are extremely small, because chance associations are likely due to the high spatial density of the galaxies. Budavári & Szalay (2008), Budavári (2012), and Loredo (2013) discuss how to compute the prior odds in various settings.

Figure 6 shows that, as one would expect, the weight of evidence is strong for the bright galaxies but weakens for the faint galaxies. The smallest Bayes factors arise for galaxies with true magnitudes near 26.5, which corresponds to the mode of the noise peak distribution, $\nu(\hat{f})$. Perhaps surprisingly, sources dimmer than magnitude 26.5 can have larger Bayes factors than those with magnitude 26.5. This happens because $\nu(\hat{f})$ peaks away from $\hat{f} = 0$, i.e., we do not expect noise peaks to have arbitrarily small
measured fluxes; the peak-finding process biases the noise peak distribution away from zero flux. For the weakest detectable sources, the most likely number of detections among the $k = 9$ epochs is one. The flat top of the distribution at the faint end corresponds to very dim sources detected only once, very near threshold. The smaller Bayes factors in that region of the plot correspond to unlikely larger numbers of detections near threshold; the discreteness in the number of detections produces a subtle banding in the distribution.

We now consider the astrometric data by themselves. For simplicity, we assume a constant direction uncertainty of $0.1\degree$ for all detections. We simulate the coordinates for the mock galaxies as follows. Around the true direction of each object, we randomly generate points from a 2D Gaussian. This sky approximation is excellent in this regime; for such tight scatters, the approximation error is below the limit of the numerical representation of double precision floating point numbers.

The coordinates of the noise peaks are generated homogeneously. The surface density of the ghosts is analytically calculated, and the integral above the $1\sigma$ detection threshold yields $\nu = 0.04\Omega^9$. A simple algorithm is to pick a large enough square, with area $\Omega$, and randomly draw the number of peaks from a Poisson distribution with an expectation value $\lambda = \nu\Omega$. Out of these ghosts, we pick a number equal to the number of flux detections, with locations that are closest to the center, where the simulated object is placed.

Figure 7 shows the distribution of astrometric Bayes factors for the real (mock) and noise sources. Note the larger span of the ($\log$) Bayes factor axis. Banding due to discreteness in the number of detections is now clearly apparent among the true-object Bayes factors; the nine levels correspond to the different numbers of detections, with the lowest being one. Comparing to Figure 6, we see that directional cross-matching is a stronger discriminant between real and noise sources in the dim-source regime. The photometric data grow in importance as the sources grow brighter.

Figure 5. Fraction of noise peaks as a function of the detection threshold. At $1\sigma$, the value is about 62%, but at $3\sigma$, it is only a few percent. At $5\sigma$, which is 24 mag in this case, the detected fraction is negligible.

Figure 6. Bayes factor distributions for simulated galaxies as a function of their true brightness (right panel) and of generated random noise peaks (left panel), using simulated photometric data (ignoring astrometric information). The left frame is extended horizontally to match the point density along the axis. Red points indicate the weight of evidence, quantified via (logarithmic) Bayes factors comparing real-source and noise origins for the simulated photometric data. The Bayes factor is high for the bright galaxies but close to unity (logarithm close to zero) for the faint ones. The Bayes factor distribution for the noise peaks (left panel) peaks below the distribution for the faintest real sources (left region of right panel) because the measured fluxes in the noise peaks tend to be in a range that mimics sources between 26 and 25 mag, rather than the dimmest sources.

6 In reality, the direction uncertainty will be roughly constant for bright sources, when the estimation error is dominated by systematic uncertainty, but for dim-source candidates, it will grow with magnitude (i.e., as estimated flux decreases). This will decrease the values of the Bayes factors for dim sources but does not qualitatively impact the findings described below.
The astrometric Bayes factors are essentially constant versus magnitude for a given number of detections among the nine epochs. This is a consequence of our simplifying assumption of a constant direction uncertainty. As noted in footnote 6, in real surveys, the astrometric precision decreases with increasing magnitude (decreasing flux) in the weak-source regime; this leads to some decrease in the Bayes factors with increasing magnitude.

Figure 8 shows the distribution of Bayes factors for the real and noise sources, now combining the photometric and astrometric factors. For the lowest band, corresponding to a single detection, $B_{\text{true}}$ is unity by definition (no constraint coming from a single detection), producing the same Bayes factor distribution as in the flux-only calculation. For multiple detections, the Bayes factors for the true sources are greatly enhanced by including astrometric information (note that the ordinate is logarithmic); just two detections produce quite strong evidence for the true-object hypothesis. The Bayes factors for the noise peaks have moved to much lower values, due to the low likelihood of directional coincidences.

This computation demonstrates that flux and astrometric catalog data, combined across epochs, can efficiently distinguish real objects from spurious detections. We have not addressed what threshold Bayes factor to use for producing a multi-epoch catalog, or, in Bayesian terms, how to convert Bayes factors into posterior probabilities for candidate objects. As noted above, when the object population density (on the sky and in flux) is known a priori, the calculation is straightforward (e.g., the prior odds will be proportional to the ratio of true-object and noise-peak sky densities). When these quantities are unknown, a possibly complicated hierarchical Bayesian calculation can jointly estimate them and the object properties. When many objects are detected, an approximate approach, plugging in empirical estimates of the densities based on the data, is likely to suffice, as described in Budavári (2012) and Loredo (2013).

5. Summary

This paper presents an exploratory study of a new, incremental approach to the analysis of multi-epoch survey data based on the fusion of single-epoch catalogs produced using a source detection algorithm with a modest or low threshold. Although the single-epoch catalogs will include many noise sources (they may even be dominated by them), we show that probabilistic fusion of the single-epoch data can produce interim or final multi-epoch catalogs with properties similar to those expected from catalogs based on image stacking. The approach is essentially a generalization of cross-matching, where object detection corresponds to identifying a set of candidate sources that match in both flux and direction across epochs. Using a probabilistic approach directly provides the required quantities, enabling the fusion of information both across epochs and between flux and direction, by straightforward multiplication of the relevant probabilities. The final quantification of the strength of the evidence is via marginal likelihoods and Bayes factors; these can be used for final thresholding or for producing posterior probabilities for source detections when population properties such as sky densities are known or can be accurately estimated (perhaps as part of the catalog analysis).

The Bayes factor compares predictions of the observed data based on true-object and noise-peak hypotheses for the data and thus requires knowledge of the distribution of the noise peaks. We derive the spatial properties of noise peaks that commonly appear in catalogs. The flux-dependent surface density of ghosts is asymmetric in flux, skewed toward positive flux values. It can be accurately approximated by a shifted Gaussian for most practical purposes.

Based on the Bayes factor, sources with single-epoch measured fluxes exceeding a $3\sigma$ threshold ($m \approx 24.55$ for an LSST-like survey) start to separate out from the noise peaks when the data are combined across just a few epochs. The
evidence for a source becomes very strong once the single-epoch fluxes exceed 5σ (24 mag). When considering only the flux measurements, the faintest sources are hard to distinguish from the noise peaks with measurements at just a few epochs; but astrometric data (celestial coordinate estimates) greatly help to separate genuine and spurious detections.

We have treated only the case of detection of constant-flux sources. Detecting variable and transient sources can be accommodated by introducing one or more time series models into the flux-matching part of the algorithm. Models that accurately describe particular classes of sources will produce optimal catalogs, but flexible models—perhaps simple stochastic processes, or even histogram or other partition-based models, with appropriate priors on variability—may suffice for producing general-purpose multi-epoch catalogs for studying variable sources. This is a potentially complicated generalization of our framework that we plan to explore in future work. Of course, variable source detection using image stacks is also an open research problem; our framework provides an alternative avenue to address it.

Our exploratory study made simplifying assumptions. A strong assumption was that source candidates are isolated enough that the image space can be partitioned into patches that have at most one candidate source. When source candidates are close to each other, the matching across epochs must account for multiple possible source association hypotheses. Similar complications appear when considering classes of objects that may be comprised of multiple sources per epoch, e.g., radio galaxies. In another work, we developed techniques for directional cross-matching in contexts with multiple candidate galaxies. In another work, we developed techniques for directional cross-matching in contexts with multiple candidate associations and with complex object structure and object motion (Budavári & Loredo 2015). These methods can be extended to include flux-matching criteria to generalize the multi-epoch detection framework described here.

The strategy we have described is quite different from conventional approaches to producing survey catalogs. Implementing it will raise new processing and database management challenges; users of the resulting catalogs will need to think about catalogs in a different way. In particular, a low-threshold single-epoch catalog will contain many spurious sources; with a low enough threshold, the spurious sources will greatly outnumber the real sources. However, evidence mounts quickly as catalogs are merged. If interim catalogs are produced consecutively, cumulative culling of early single-epoch catalogs could reduce the storage burden for catalogs subsequent to the first one. Such issues, and the generalizations described above, will be topics for future study.

The authors gratefully acknowledge valuable and inspiring discussions with Andy Connolly and Robert Lupton on various aspects of the topic, as well as the anonymous reviewer for very helpful, detailed feedback that significantly improved the presentation. This study was supported by the NSF via grants AST-1412566 and AST-1312903 and NASA via awards NNG16PJ23C and STScI-49721 under NAS5-26555.

Appendix A

Peaks of Two-Dimensional Random Fields

Consider a two-dimensional Gaussian random field \( F(\mathbf{r}) \) with a known power spectrum, e.g., as discussed at the end of this appendix. Its gradient vector would be \( \boldsymbol{\eta} \), and the second derivative tensor would be \( \zeta \). We would like to find the density of the peaks of this field above a certain height. We will follow the procedure outlined in Bardeen et al. (1986, hereafter BBKS). For simplicity, our notation will also follow that of BBKS as closely as possible.

We will use the fact that the field at our fiducial location \( \mathbf{r}_p \) is a peak, i.e., its gradient is zero, and the second derivative is negative definite. We will Taylor-expand the field and its gradient to second order around a peak at the position \( \mathbf{r}_p \) (equivalent to BBKS Equation (3.1)),

\[
F(\mathbf{r}) \approx F(\mathbf{r}_p) + \frac{1}{2} \sum_{j} \zeta_j (\mathbf{r}_p) (\mathbf{r} - \mathbf{r}_p)_j (\mathbf{r} - \mathbf{r}_p)_j,
\]

(15)

\[
\eta_j (\mathbf{r}) \approx \sum_{j} (\mathbf{r} - \mathbf{r}_p)_j \zeta_j (\mathbf{r}_p),
\]

(16)

where in both equations we use the fact that the gradient of the field at a peak is zero, i.e., \( \eta_j (\mathbf{r}_p) = 0 \). Provided that \( \zeta \) is nonsingular at \( \mathbf{r}_p \), we can express \( \mathbf{r} - \mathbf{r}_p \) from the second equation:

\[
\mathbf{r} - \mathbf{r}_p = \zeta^{-1} (\mathbf{r}_p) \boldsymbol{\eta} (\mathbf{r}).
\]

(17)

We can write a Dirac delta that picks all extremal points of \( F \) as

\[
\delta^{(2)} (\mathbf{r} - \mathbf{r}_p) = [\det \zeta (\mathbf{r}_p)]^{\frac{1}{2}} \delta [\boldsymbol{\eta} (\mathbf{r})].
\]

(18)

We can turn this expression from a continuous random field into the density of a discrete point process by summing over all extremal points in our two-dimensional space divided by the area, see BBKS Equation (3.2). The extremal points are the zeros of \( \boldsymbol{\eta} \). Due to the Dirac delta \( \delta^{(2)} (\mathbf{r} - \mathbf{r}_p) \), in the summation we get \( \zeta (\mathbf{r}_p) = \zeta (\mathbf{r}) \). After averaging, the density of the point process can be written as

\[
n_{\text{ext}} (\mathbf{r}) = [\det \zeta (\mathbf{r})]^{\frac{1}{2}} [\delta^{(2)} [\boldsymbol{\eta} (\mathbf{r})]].
\]

(19)

A.1. Separating Peaks and Minima

In order to pick the peaks of the Gaussian random field, we will also need to have a negative definite \( \zeta \). If we only want peaks of a certain height, we need to calculate the appropriate ensemble average of this density over the constrained range of the variables.

We have six random variables: the field \( F \), the three components of the symmetric tensor \( \zeta \), and the two components of the gradient \( \boldsymbol{\eta} \). The correlations can be computed in a straightforward manner, given the power spectrum of the field, \( P(k) \), where \( k \) is the norm of a 2D wave vector \( \mathbf{k} \). The gradient is uncorrelated with both the field and the second derivatives, due to the parity of the Fourier representation. Let us denote the correlation matrix of the field and the Hessian as \( \mathbf{C} \) and that of the gradient as \( \mathbf{H} \). Furthermore, let us define the different \( k \)-moments of the power spectrum characterizing the field as

\[
\sigma_n^2 = \frac{1}{(2\pi)^2} \int d^2 k k^{2n} P(k).
\]

(20)

Let us also introduce the dimensionless \( \gamma \) and the characteristic scale \( R \) defined as (see BBKS Equation (A5))

\[
\gamma = \frac{\sigma_2^2}{\sigma_0 \sigma_2}, \quad R^2 = \frac{\sigma_2^2}{\sigma_1^2}.
\]

(21)
We can now explicitly write down the autocorrelation matrix $C$ of $v = (F, \zeta_1, \zeta_2, \zeta_3)$ and $H$ for $\eta = (\eta_1, \eta_2)$, as

$$C = \begin{pmatrix}
\sigma_2^2 & -\sigma_1^2/2 & 0 & -\sigma_1^2/2 \\
-\sigma_1^2/2 & 3\sigma_2^2/8 & 0 & \sigma_2^2/8 \\
0 & 0 & \sigma_2^2/8 & 0 \\
-\sigma_1^2/2 & \sigma_2^2/8 & 0 & 3\sigma_2^2/8
\end{pmatrix}, \quad (22)
$$

The determinants of $C$ and $H$ can be written as

$$\det C = \frac{1}{64}(1-\gamma^2)\sigma_0^2\sigma_2^4, \quad \det H = \frac{1}{4}\sigma_1^4. \quad (24)$$

Using the inverse of the correlation matrices, we can express the multivariate Gaussian distribution of $(v, \eta)$ as a product of two independent distributions:

$$dP = dp_1 dp_2 = \left[ \exp \left( -\frac{v^TC^{-1}v}{2} \right) \right] \frac{d^4v}{(2\pi)^2(\det C)^{1/2}} \times \left[ \exp \left( -\frac{\eta^TH^{-1}\eta}{2} \right) \right] \frac{d^2\eta}{2\pi(\det H)^{1/2}}. \quad (25)$$

We will evaluate the two parts of this expression separately. The first term contains the unconstrained probability distribution for $v$, while the second term describes the probability distribution of the components of the gradient $\eta$. The peak density will be given by the expectation value of the density of the extremal point process $n_{\text{ext}}(r)$ as defined in Equation (19) over this distribution, subject to the peak constraints (negative definite $\zeta$).

Before we proceed further, the second derivative tensor can be described more conveniently with the two eigenvalues $\lambda_1, \lambda_2$ and a rotation angle, $\phi$, as follows:

$$\zeta_{11} = \lambda_1 \cos^2 \phi + \lambda_2 \sin^2 \phi, \quad (26)$$

$$\zeta_{12} = (\lambda_1 - \lambda_2) \sin \phi \cos \phi, \quad (27)$$

$$\zeta_{22} = \lambda_1 \sin^2 \phi + \lambda_2 \cos^2 \phi. \quad (28)$$

For simplicity, let us introduce the dimensionless variables $x = (\lambda_1 + \lambda_2)/\sigma_2$ (the trace of the second derivative tensor), $y = (\lambda_1 - \lambda_2)/\sigma_2$, and $z = F/\sigma_0$. The Jacobian of the transformation from $(F, \zeta_{11}, \zeta_{12}, \zeta_{22})$ to $(z, x, y, \phi)$ is

$$J = \sigma_0 \sigma_2^3 y/2. \quad (29)$$

With this notation, we can write the determinant of $\zeta$ in a simple form:

$$|\det \zeta(r)| = \lambda_1 \lambda_2 = \sigma_2^2 \frac{x^2}{4}(x^2 - y^2). \quad (30)$$

The quadratic form containing $v$ in the exponent can be written with the new variables as

$$Q = v^TC^{-1}v = \left( 2y^2 + \frac{x^2 + 2xz + z^2}{1 - \gamma^2} \right). \quad (31)$$

In these variables, the unconstrained probability distribution for $(x, y, z, \phi)$ becomes

$$dP_1 = \frac{4y}{(2\pi)^2(\sqrt{1 - \gamma^2})} \exp \left( -\frac{1}{2}Q \right) dx \ dy \ dz \ d\phi. \quad (32)$$

In order to properly handle the symmetries of the problem, we can assume that $\lambda_1 \geq \lambda_2$. Because any $(\lambda_1, \lambda_2)$ pair can be mapped onto itself by a 180° rotation, the valid range of $\phi$ is $[0, \pi]$. Since none of the terms depend on $\phi$, we can integrate over $\phi$, resulting in

$$dP_1 = \exp \left( -\frac{x^2 + 2yz + z^2}{2(1 - \gamma^2)} \right) \frac{dx \ dz}{2\pi \sqrt{1 - \gamma^2}} \frac{e^{-x^2/2y} dy}{2y}. \quad (33)$$

The constraint $\lambda_1 > \lambda_2$ maps onto $0 < y$. If we perform the integration over $y > 0$, and $x, z \in (-\infty, \infty)$, we get 1, as we should, for the unconstrained probability of a general point.

A.2. Surface Density of the Peaks

In order to compute the surface density of the peaks, we need to multiply the unconstrained differential probability distributions in Equation (25) with the peak density from Equation (19) as

$$dn_{pk} = |\det \zeta| \delta^{(2)}(\eta) dP_1 dP_2, \quad (34)$$

$$dn_{pk} = |\det \zeta| dP_1 \left[ \exp \left( -\frac{\eta^T H^{-1} \eta}{2} \right) \right] \frac{d^2\eta}{2\pi(\det H)^{1/2}}. \quad (35)$$

After trivially integrating over $d^2\eta$ in the second term and substituting the determinants of $\zeta$ and $H$, we get

$$dn_{pk} = \left[ \frac{(x^2 - y^2)y}{2\pi R^2} \frac{e^{-x^2/2y} \ dy}{2\pi R^2} \right] \frac{x^2 + 2yz + z^2}{2(1 - \gamma^2)} \ dx \ dz. \quad (36)$$

For a peak, both eigenvalues of the second derivative tensor must be negative. In the rotated coordinates $(x, y)$, this means that $x < 0$ and $0 < y < -x$. We can easily integrate over the allowed range of $y$ next, yielding

$$\int_0^{-x} dy \ y (x^2 - y^2) e^{-x^2/2y} = \frac{1}{2} \left( x^2 - 1 + e^{-x^2} \right). \quad (37)$$

We are left with

$$dn_{pk} = -\frac{1 + x^2 + e^{-x^2}}{4\pi R^2} \ exp \left( -\frac{x^2 + 2xz + z^2}{2(1 - \gamma^2)} \right) \ dx \ dz \quad (38)$$

Let us introduce the function $B$ as

$$B(s, b) = \sqrt{\frac{\pi}{b}} \ exp \left( \frac{s^2}{2b} \right) \left[ 1 + \text{erf} \left( \frac{s}{\sqrt{2b}} \right) \right]. \quad (39)$$
Evaluating the integral of Equation (38) over \( x \in (-\infty, 0] \) in Mathematica (Wolfram 2016), we obtain

\[
n_{pk}(z) = \frac{e^{-\frac{z^2}{4\sigma^2}}}{8\pi^2\sigma^2} \left[ \left( 1 - \gamma^2 \right) s + \left[ s^2 - \gamma^2 \left( 1 + s^2 \right) \right] \right] \\
\times B(s, 1) + B(s, 3 - 2\gamma^2),
\]

with

\[
s = \frac{\gamma z}{\sqrt{1 - \gamma^2}}.
\]

We get the cumulative surface density \( \lambda(z_0) \) of the noise peaks above the threshold \( z_0 \) by integrating

\[
\lambda(z_0) = \int_{z_0}^{\infty} n_{pk}(z) dz.
\]

A.3. Gaussian Window Function

Finally, we need to evaluate the shape parameter \( \gamma \). Assume that the window function applied to the random field is a Gaussian with a scale \( \alpha \),

\[
w(r) = \frac{1}{2\pi\alpha^2} \exp \left( -\frac{r^2}{2\alpha^2} \right).
\]

Its Fourier transform is also a Gaussian,

\[
W(k) = \exp \left( -\frac{k^2\alpha^2}{2} \right).
\]

We model the sky noise as white noise with a flat spectrum. Thus, the correlations in the measured random field are determined by the window function, i.e.,

\[
P(k) = A |W(k)|^2 = A \exp \left( -k^2\alpha^2 \right).
\]

With this power spectrum, it is straightforward to compute the scale and the shape parameters defined in Equation (21) as

\[
\gamma^2 = \frac{1}{2}, \quad R^2 = \frac{\alpha^2}{2}.
\]

We obtain the cumulative surface density \( \lambda(z_0) \) of the noise peaks above the threshold \( z_0 \) by integrating

\[
\lambda(z_0) = \int_{z_0}^{\infty} n_{pk}(z) dz.
\]

A.4. Nearest Neighbor Statistics

Now we are in a position to compute the probability that a noise peak is within a radius \( r \) of our point of interest located at the origin. The spatial distribution of the noise peaks is described by a Poisson process with the surface density \( \lambda \). The cumulative probability that the nearest peak higher than \( z_0 \) is within a radius \( r \) is given by the well-known expression

\[
P_{pk}(< r | z_0 ) = 1 - \exp \left( -\lambda(z_0) r^2 \pi \right).
\]

The differential probability is given by its derivative with respect to \( r \) as

\[
P_{pk}(r | z_0 ) = 2\pi r \lambda(z_0) \exp \left( -\lambda(z_0) r^2 \pi \right).
\]

Both of these probabilities are shown in Figure 9. The differential probability starts off around the origin scaling with \( r^2 \), due to the available area (phase space for configuration). This, in turn, causes the cumulative function to rise as \( r^2 \), resulting in a very small probability (<0.03) that a noise peak will appear within a PSF scale. Thus, we can safely ignore noise peaks as a major contributor to false detections.

References

Adler, R. J. 2010, The Geometry of Random Fields (Chichester: Wiley)
Bardeen, J. M., Bond, J. R., Kaiser, N., & Szalay, A. S. 1986, ApJ, 304, 15
Bond, J. R., & Efstathiou, G. 1987, MNRAS, 226, 655
Budavári, T. 2012, in Statistical Challenges in Modern Astronomy V, ed. E. D. Feigelson & G. J. Babu (New York: Springer)
Budavári, T., & Loredo, T. J. 2015, Annual Review of Statistics and Its Application, 2, 113
Budavári, T., & Szalay, A. S. 2008, ApJ, 679, 301
Budavári, T., & Szalay, A. S. 2014, adass XXIII, 485, 207
Kaiser, N. 2004, The Likelihood of Point Sources in Pixellated Images, Pan-STARRS Internal Rep. PSDC-002-010-xx
Loredo, T. J. 2013, in Statistical Challenges in Modern Astronomy V, ed. E. D. Feigelson & G. J. Babu (New York: Springer)
Lund, J., & Rudemo, M. 2000, Biometrika, 87, 235
Madau, P., & Thompson, C. 2000, ApJ, 534, 239
Press, W. H. 1997, in Proc. Conf. in Honor of John Bahcall, Unsolved Problems in Astrophysics, 49 (Princeton, NJ: Princeton Univ. Press) arXiv: astro-ph/9604126
Szalay, A. S., Connolly, A. J., & Szokoly, G. P. 1999, AJ, 117, 68
Wolfram Research, Inc. 2016, Mathematica Version 11.0 (Champaign, IL: Wolfram Research, Inc.)