Object detection in multi-epoch data

G. Jogesh Babu\textsuperscript{a,\ast}, A. Mahabal\textsuperscript{b}, S.G. Djorgovski\textsuperscript{b}, R. Williams\textsuperscript{c}

\textsuperscript{a} Department of Statistics, 326 Joab L. Thomas Building, The Pennsylvania State University, University Park, PA 16802-2111, USA
\textsuperscript{b} Division of Physics, Mathematics, and Astronomy, California Institute of Technology Pasadena, CA 91125, USA
\textsuperscript{c} Center for Advanced Computing Research, California Institute of Technology Pasadena, CA 91125, USA

Received 13 April 2007; received in revised form 23 January 2008; accepted 12 February 2008

Abstract

In astronomy multiple images are frequently obtained at the same position of the sky for follow-up coaddition as it helps one go deeper and look for fainter objects. With large scale panchromatic synoptic surveys becoming more common, image co-addition has become even more necessary as new observations start to get compared with coadded fiducial sky in real time. The standard coaddition techniques have included straight averages, variance weighted averages, medians etc. A more sophisticated nonlinear response chi-square method is also used when it is known that the data are background noise limited and the point spread function is homogenized in all channels. A more robust object detection technique capable of detecting faint sources, even those not seen at all epochs which will normally be smoothed out in traditional methods, is described. The analysis at each pixel level is based on a formula similar to Mahalanobis distance.

\textsuperscript{\ast} Corresponding author. Tel.: +1 814 863 2837.
E-mail addresses: babu@stat.psu.edu (G. Jogesh Babu), aam@astro.caltech.edu (A. Mahabal),
george@astro.caltech.edu (S.G. Djorgovski), roy@caltech.edu (R. Williams).

\textsuperscript{\ast} Corresponding author. Tel.: +1 814 863 2837.
E-mail addresses: babu@stat.psu.edu (G. Jogesh Babu), aam@astro.caltech.edu (A. Mahabal),
george@astro.caltech.edu (S.G. Djorgovski), roy@caltech.edu (R. Williams).

1572-3127/$ -$ see front matter \textsuperscript{\copyright} 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.stamet.2008.02.003

1. Introduction

Many major projects, ongoing and future synoptic surveys, Palomar-QUEST (http://palquest.org/), MACHO (http://www.macho.mcmaster.ca/), LSST (http://www.lsst.org/), OGLE (http://bulge.astro.princeton.edu/~ogle/) Pan-STARRS (http://pan-starrs.ifa.hawaii.edu/) etc., involve...
repeated scans of large areas of the sky in several wavelength bands. Thus an important area of recent astronomical research has been the investigation of source detection in multi-epoch data. A question frequently asked is: What is the best way to combine image regions with low signal to detect faint objects? Historically, two basic methods have been used to search for faint astronomical objects:

(i) Use a larger telescope to collect a larger number of photons.
(ii) Stack a large number of registered images in order to improve the signal-to-noise ratio.

The former has engineering and monetary limitations while the latter may not work for transients that are seen only once and for very faint objects where the signal remains below detection threshold even after the pixel-to-pixel coadding of several images. Szalay et al. [11] had proposed a method that used chi-square coaddition by treating pixels from different images but at the same location to be uncorrelated. It is an improvement over standard coaddition but has its own limitations, as the pixels in different images at any location are in fact correlated. The procedure described here is a more robust technique to detect faint sources from multi-epoch data. It is designed not only to have high sensitivity, but to detect changes between images.

The analysis at each pixel level is based on a statistic similar to the measure of distance introduced a by P.C. Mahalanobis in 1936 [8,2]. The Mahalanobis distance $D_{\nu, \Sigma}$, of a multivariate row vector $x$ from a group of values with mean vector $\nu$ and covariance matrix $\Sigma$, is defined as

$$ D_{\nu, \Sigma}(x) = \sqrt{(x - \nu)\Sigma^{-1}(x - \nu)^T}. $$

It is used in classical multivariate analysis and differs from Euclidean distance. It is scale-invariant, and is based on correlations between variables by which different patterns can be identified and analysed. If the covariance matrix $\Sigma$ is the identity matrix, then the Mahalanobis distance reduces to the Euclidean distance. If $Y$ is a Gaussian random vector with mean $\nu$ and covariance matrix $\Sigma$, then $D^2_{\nu, \Sigma}(Y)$ is distributed as $\chi^2$ with $p$ degrees of freedom, where $p$ is the rank of $\Sigma$. It is a useful way of determining similarity of an unknown sample set to a known one [3,2]. Mahalanobis distance is available in R Stats package (http://www.r-project.org/). R is a free public domain software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS. Tutorials for R are available at http://astrostatistics.psu.edu/datasets/ and at http://www.sr.bham.ac.uk/~ajrs/R/.

For the methodology described in this paper, it is required that the pixel size for different images is invariant. However, the method does not depend on the point spread function (PSF) since signal from genuine sources spreading to different extent in different images will still lead to significant correlations between different images at the pixels with the source than if there were no source. Of course it is best if the PSFs are fairly similar. In the following sections we provide the details of the method as well as the tests we have run so far and future plans for large-scale implementation.

2. Methodology

We start with $N$ images of a given region of the sky. We first use standard techniques to ensure that all images are of the same size in terms of area covered as well as the pixel dimensions. This involves using a tool like Swarp (available from TERAPIX web site http://terapix.iap.fr/), which can extract only the common overlapping regions from all the images. In addition the objects in the images can be matched to a known catalogue for ensuring that the images have accurate
World Coordinate System (WCS). The images are also background subtracted so the mean value in parts with no obvious sources of each image is close to zero.

2.1. Computation of the statistic

Suppose the dimensions in pixels of each of the images is \((m, n)\). So there are \(R\) pixels in each of the image \((= m \times n)\). We concentrate on a single pixel \(r\). Let \(f^r = (f^r_1, \ldots, f^r_N)\) denote the row vector of photon counts in \(N\) images at pixel \(r\). In the absence of a source, \(f^r\) represents noise, hence it is assumed to be approximately Gaussian, i.e.

\[
f^r \sim \mathcal{MVN}(M, \Sigma),
\]

\(M = (M_1, \ldots, M_N)\) is the vector of means and \(\Sigma\) is the variance–covariance matrix. Then

\[
X_r = (f^r - M)\Sigma^{-1}(f^r - M)^T
\]

has approximately \(\chi^2\) distribution. Note that if the distribution of \(f^r\) is Gaussian, then \(X_r\) has \(\chi^2\) distribution with \(p\) degrees of freedom, where \(p\) is the rank of \(\Sigma\).

If the images are taken using the same wave band but at different epochs under similar conditions, then the background tends to be similar. If the backgrounds are different just due to sky conditions (with the wave bands being the same), then the backgrounds are uniformly correlated. This method takes into account the potential information in different wave bands as well. If \(f^r_i\) are all uncorrelated, then the off-diagonal entries of \(\Sigma\) are all zero. In this case it reduces to the statistic considered by Szalay et al. [11].

To estimate the covariance matrix when the background is similar throughout the image, i.e. \(M_i\) is spatially constant in the \(i\)-th image, let

\[
s_{i,k} = (1/R) \sum_{r=1}^{R} (f^r_i - M_i)(f^r_k - M_k),
\]

where \(R\) is the number of pixels in each image or cutout. Now the \(i, k\)-th entry of the estimated covariance matrix \(\Sigma\) is \(s_{i,k}\).

2.2. Source detection

If \(f^r \sim \mathcal{MVN}(M, \Sigma)\), then by part (c) of Theorem 5–17 of [1] p. 211, \(X_r\) has chi-square distribution. So if there is no source at pixel \(r\), then in either case, when \(\Sigma\) is estimated, \(X_r\) has approximately chi-square distribution. Thus for any \(y\), the probability of detecting a source when \(X_r > y\) can be obtained from \(\chi^2\) tables. This incorporates potential correlations of the background in different images. The critical value (threshold) \(y\) is chosen based on the significance level set for source detection.

Standard programs like sextractor (SourceExtractor, available from TERAPIX web site) look for a connected set of pixels above a threshold resulting in a list of astronomical objects (called catalogue in astronomy jargon) with various flags indicating its validity as an astronomical object based on prior knowledge like that of shapes and profiles of objects built into the program.

While this technique may not do any better than usual methods for bright objects, under lower signal conditions it is expected to strongly outperform the traditional methods. We will extensively test our technique on the already available PQ (Palomar-QUEST) data in readiness for the even larger stream of LSST data.
2.3. False discovery error rate

As we are testing multiple hypotheses, we need to avoid too many false positives. When there are large number of pixels above the threshold, there may as well be too many falsely discovered source pixels. The proportion of falsely discovered source pixels should be kept to a minimum with any source detection algorithm. The method introduced by Benjamini and Hochberg [4] (see also [5,6]), called the False Discovery Rate (FDR) does precisely this, controlling the proportion of incorrectly rejected null hypotheses. Thus FDR depends on the ratio of the number of falsely rejected null hypothesis (in our case declaring a source when actually there is no source) to the total number of rejected null hypothesis (total declared sources in our case). The method allows a priori specification of a proportion of false discoveries to total discoveries, and the procedure is independent of the source distribution. The procedure works as follows: Suppose $0 < \alpha < 1$ is the a priori error rate (similar to significance level) set for the multiple testing problem with $s$ tests of hypotheses, and suppose the $p$-values resulting from these tests are ordered such that $p(1) \leq p(2) \leq \cdots \leq p(s)$. Here we might restrict our attention to those $3\sigma$ or $4\sigma$ detections. Now we find

$$d = \max_{1 \leq k \leq s} \{ k : p(k) \leq \alpha \cdot k/s \},$$

then rejecting the null hypotheses (identify detections) corresponding to $p(1), \ldots, p(d)$ provides FDR $\leq \alpha$. Note that only pixels with $p$-values less than $\alpha$ matter. This procedure is independent of the source detection algorithm used. An implementation of this procedure is discussed in [7].

However, in the examples discussed in the current article, as the detected sources are few and far apart, false discoveries is not a big issue. It is helpful to keep this technique in a toolbox for future surveys like LSST, in case we encounter too many detected sources. Recently some variants of FDR have been investigated in the statistical literature. See [9,10] for details.

3. Application of the method to data sets

We carried out a series of tests using artificial data as well as BRI images from the Palomar-QUEST survey. For the former we did the following tests:

(i) artificial data with zero background and no noise added,
(ii) artificial data with nonzero constant background, but no noise,
(iii) artificial data with non-zero constant background and Poisson noise added. (Due to the discreet nature of the emission received from the sources the noise in the observed astronomical images is expected to be Poisson.)

We found that our method could recover all objects that the standard coadding gets, and then some. It fails if two or more objects are blended such that they overlap by a large fraction.

For real data we obtained cutouts (image subsections) of size $60' \times 2'$ for a few pointings on the sky at several different epochs. Individual CCDs on Palomar-QUEST are about $8'$ wide. The dithering strategy followed for the survey necessitates choosing such narrow strips to ensure that for the different scans we do not straddle different CCDs. (This does not mean that we can not do larger areas. It is just that for current exploratory investigations we wish to keep inter-CCD calibrations out of the picture). The images were coadded using *swarp* as well as using our method. Object catalogs were then obtained for both sets at a series of significance levels. The results there too have been good. We coadded 3, 5 and 7 images using both methods and found
that our method does almost as well as the traditional method in most cases and better in many others.

The method has also been tested using artificial images (Fig. 1) and more stringent tests are planned before large scale use can be done. We created images with 50 artificial objects using the artobs package in IRAF, with and without Poisson noise and artificial background. We found that when no Poisson noise was added all 50 objects were recovered. But when Poisson noise with 10 different seeds was added, 25 objects were recovered once, 26 thrice, 27 four times and 28 twice (as the objects span a wide range of magnitudes including a few close to detection limits, it is expected that several will be “lost” in this fashion). When ten images were coadded using swarp, the number of detections rose to 37. When our method was used on the ten images, the resultant catalogue consisted of 51 objects. All 37 found in the swarped image were found, one more of the original was found, and the remaining 13 were not from the original catalogue. More tests are planned to be able to flag the spurious candidates. Additional planned tests include possibilities like changing the number of neighbouring pixels considered for constructing the covariance matrix.

Some preliminary results based on Palomar-Quest images are shown in Fig. 2. The images on the right are formed by the statistic based on Mahalanobis distance with 4σ sources circled. The images appear to be more grainy in appearance than the traditional coadded images. This is due to the removal of correlation between adjacent pixels (coadded SNR in centroids is far higher than in adjacent pixels as light falls off away from the centroid). We also seem to flag more sources with fewer epochs on the right side panels. This is mainly due the effect of the noise. As our method is designed to flag faint sources, the noise can present itself in the form of statistically significant peaks. When more image layers are added, the graininess decreases. In this case, the noise gets averaged over several images thus reducing false positives. So, either a larger number of images need to be used, or an additional criterion based on number of correlated adjacent pixels be used for qualifying statistically significant peaks as sources. The emergence of 3 (two in the top right panel) objects, in line at the bottom right corner of the lower panels of the three images, absent in the left panels is likely due to the passage of a minor planetary body through the field.
4. Other applications and future work

- In addition to source detection, a related problem that can be addressed with this technique is that of identifying faint objects in a particular region of color–color space. A color indicates the ratio of fluxes in two bands. A colour–colour diagram then is a plot which has such ratios formed from 3 bands: $A/B$ on $x$-axis, and $B/C$ on $y$-axis. An example of colour–colour diagram depicting $B–R$ and $R–I$ colours from Palomar-QUEST is shown in Fig. 3. Objects of a given type occupy particular sections of this plot. Thus for an area of interest e.g. one that is expected to contain hi-z quasars, one could try: $(f_B - f_R)a(f_R - f_I)$, where $f_B, f_R, f_I$ denote the log of fluxes (photon counts) in images with filters $B$, $R$, $I$, and $a$ is a normalizing factor estimated from the covariance matrix. This can be tested by populating different coloured objects and trying detection in single band, combined bands and using the above technique. Standard squaring as is used in calculating Euclidean distances may have to be avoided as sign of the colour is important. For bright normal images this technique may not do well, but under lower signal conditions it is expected to perform better.

- For spatially variable background, we need to estimate mean and variance locally using a close neighbourhood of each pixel. We currently use 9 pixels (the pixel under question and its eight neighbours) to estimate the covariance matrix $\Sigma(r)$ and the mean vector $M(r)$. In this case $X_r$ is taken to be

$$(f' - M(r))\Sigma(r)^{-1}(f' - M(r))^T.$$ 

Note that, 9 pixels may not contain all light, but for most stellar sources taken under good conditions from the ground for images that are not over-sampled, the central pixel plus the 8 neighbouring pixels contain a good fraction of the light (exact number depending on pixel scale and seeing conditions). To generalize the situation, we may have to go to a larger kernel size.

- So far we have modelled based on Gaussian assumptions, where least squares methods and maximum likelihood methods coincide. Further, the dependency of pixel data across...
Fig. 3. An example of high-redshift quasar selection, as outliers in a colour–colour parameter space, from a small area covered in the PQ survey. Normal stars (dots) form a well-defined locus in this parameter space; quasar candidates (solid circles) deviate from this locus, while having exactly the same unresolved image morphology as the stars.

different filters can be modeled and analysed with minimum difficulty through covariances and joint multivariate distributions. For Poisson data, modeling the multi-dimensional data with statistical dependence structure is far from clear. Instead it is more natural to consider modelling functional dependency of Poisson rates across filters. This will be discussed in a later article.

– As we refine the method for large data sets, we are starting to incorporate it in the Palomar-QUEST real-time pipeline. Initially some handholding will be needed as Palomar-QUEST images in each of four filters for a given night is about 1500 times the size of the test images tried so far. The advantage will be to be able to see fainter objects and hence curb the rate of false alerts for transients. The methodology will also set a standard for the forthcoming synoptic surveys like Pan-STARRS and LSST can reap rich benefits thereof.

Acknowledgements

The first author was supported in part by the National Science Foundation grants AST-0434234 and AST-0326524. The second, third and fourth authors were supported in part by the National Science Foundation grants AST-0326524 and AST-0407448.

References


