Cross-matching within the *Chandra* Source Catalog

Arnold Rots, Douglas Burke, Francesca Civano, Roger Hain, and Dan Nguyen

CXC/SAO Harvard-Smithsonian Center for Astrophysics

arots@cfa.harvard.edu

Abstract

Cross-matching among overlapping source detections in the development of the *Chandra* Source Catalog (CSC) presents considerable challenges, since the Point Spread Function (PSF) of the *Chandra* X-ray Observatory varies significantly over the field of view. For the production of the second release of the CSC we have developed a cross-match tool that is based on the Bayesian algorithms by Budavári, Heinis, and Szalay, making use of the error ellipses for the derived positions of the detections.

However, calculating match probabilities only on the basis of error ellipses breaks down when the PSFs are significantly different. This is an issue that is not commonly addressed in cross-match tools. We have applied a satisfactory modification to the algorithm that, although not perfect, ameliorates the issue for the vast majority of such cases.

A separate issue is that as the number of overlapping detections increases, the number of matches to be considered increases at an alarming rate, requiring procedural adjustments to ensure that the cross-matching finishes within a Hubble time.

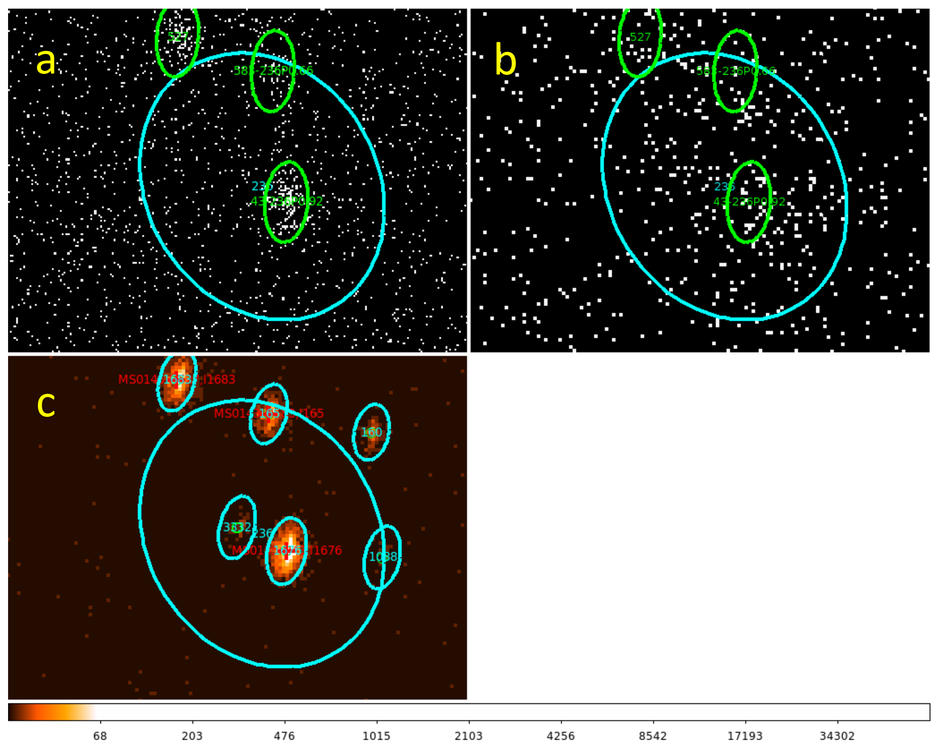
We intend to make the tool available as a general purpose cross-match engine for calculating match probabilities between sources in multiple catalogs simultaneously.

The Chandra Source Catalog

The *Chandra* X-ray Observatory was the third in the line of NASA's Great Observatories, launched in July 1999 into a highly eccentric orbit with a period of 64 hours, and covering the energy range 0.4-10 keV. Its field of view extends to up to 30 arcmin with a spatial resolution varying from 0.5 arcsec on -axis to 30 arcsec at the edge of the field.

The first version of the *Chandra* Source Catalog (CSC) was released in 2010, containing hundreds of source parameters for 95,000 sources, based on 136,000 detections, with 9 full field data product types and 11 per-source data products. Currently, release 2 of the CSC is in production which is expected to be based on 350,000 detections, using more observations and stacking overlapping ones, covering 700 square degrees (or 1.7% of the sky), and including error ellipses for the fitted positions. Because of the large variation in the size of the Point Spread Function (PSF), stacking is restricted to observations pointed within 1-arcmin of each other.

As part of this project, we have been developing a cross-matching tool, *CSCXmatch*, that will be released as a public tool for cross-matching sources from any number of catalogs (including the CSC, of course) and that during production is used to match detections from overlapping stacks.



Detail in Orion in three observation stacks. Stacks a and c are fairly close to on-axis, stack b is significantly off-axis. All ellipses represent 90% encircled energy PSFs. The green ellipses pertain to the data in stack a, the small cyan ellipses to stack c, and the large cyan ellipse to stack b. The three sources from stack a are matched with the corresponding sources in stack c on the basis of their error ellipses. The source from stack b is matched with one source in stack a, but also ambiguously with two sources in stack c through their PSF ellipses; their error ellipses fail to establish any matches with sufficient probability (see ***Resolution***)

Cross-matching

Early cross-match operations were purely visual exercises: if a star on a blue plate coincided with the position of a star on a red plate, they were assumed to represent the same object. This was extrapolated in the first automated catalog cross-match tools to something like: if two objects from different catalogs were located within 1 arcsec (or whatever error circle radius seemed appropriate), they were pronounced a match.

This works quite well for, say, star positions in the visual part of the spectrum, But it becomes problematic when one attempts to match catalogs from different parts of the spectrum and/or derived from observations with significantly different resolution. In both cases one may not assume that a simple proximity confirms a match; the two detected sources may represent physically different objects (either or both not necessarily detectable in the other spectral range) or a source in the catalog with the lower spatial resolution may represent a blend of multiple sources from the higher resolution one.

The reliability of the matches can be significantly improved by calculating rigorous match probabilities based on the detected positions and the detailed uncertainties therein. The approach we have chosen is to use the algorithms developed by Budavári, Heinis, and Szalay, since that method worked well for the cross-matching of the first release of the CSC with SDSS, as reported by Rots and Budavári.

Two Catalogs

For a two-catalog case, a Bayes factor is calculated for each pair sources, one from one catalog, the other from the other catalog:

$$B_{ij} = \frac{2}{\sigma_i^2(j) + \sigma_j^2(i)} \cdot \exp\left(-\frac{\psi_{ij}^2}{2(\sigma_i^2(j) + \sigma_j^2(i))}\right)$$

We assign a flat prior probability:

$$P_0(0) = \frac{\min(N_L, N_M)}{N_L \cdot N_M}$$

and calculate a posterior probability as:

$$P_{ij}(k) = \left(1 + \frac{1 - P_0(k)}{B_{ij} \cdot P_0(k)}\right)^{-1}$$

That posterior is then used as a prior, iterating with:

$$P_0(k+1) = \frac{\sum_{i=1}^{N_L} \sum_{j=1}^{N_M} P_{ij}(k)}{N_L \cdot N_M}$$

More than Two Catalogs

This can in principle be extended to any number of catalogs by modifying the Bayes factor for a tuple of *n* sources to:

$$B = 2^{n-1} \frac{\prod w_i}{\sum w_i} \exp\left(-\frac{\sum_{i<j} w_i w_j \psi_{ij}^2}{2 \sum w_i}\right) \qquad w_i = \frac{1}{a_i b_i}$$

This all seems very straightforward — and it is, except for three issues addressed in the next two sections; these are part of the dark side of cross-matching.

Resolution

Accidents can happen when the spatial resolutions of the catalogs differ significantly. An obvious example is a case where a high-resolution catalog has two or more sources fairly close together which are blended into a single source in a low resolution catalog. The error ellipse of the large blended source may well be small compared to the separations of the high resolution sources, resulting in incomplete or no matches, resulting in match probabilities that are plainly wrong. A similar situation can arise when the two PSFs are similar in size and shape, but have a significant eccentricity and with position angles that differ significantly.

We address this by performing a test on each source pair that is to be considered. The overlap area of the PSFs of the two sources is compared to the area of the largest of the two PSFs. If that ratio falls below a certain threshold, we use the PSF ellipses for calculating the Bayes factors, rather than the error ellipses. There is no strict mathematical basis for this, but the approach makes sense (to us) and it works.

This raises an important issue that has not been commonly recognized in cross-match operations: matches, even those with a high probability, are not necessarily 1-to-1 anymore; they may be one-to-one, one-to-many, or many-to-one. As long as a source has a high probability match with no more than one source per (other) catalog, it can be unambiguously identified. As soon as there are matches with more than one source in a single catalog, the source becomes ambiguously identified — unless those other sources turn out to be ambiguous. What we are thinking of in this last case are high resolution sources being matched with multiple low resolution sources.

See he figure above for illustration.

Practical Considerations

When the number of catalogs gets large, the number of combinations to be considered increases exponentially, not only in the number of catalog combinations that need to be checked, but also in the number of different source combinations (tuples) that need to checked. This is being addressed in two ways.

To reduce the number of source tuples to be considered, we impose a coarse common grid on all the catalogs and consider only pairs of sources that are contained within a single 3 × 3 square of cells — arguing that all other combinations are too far apart to be serious contenders for a match. For matching between *m* catalogs, only *m*-tuples of sources are considered where at least ½(*m*–1)(*m*–2)+1 pairs of the tuple have a pairwise match probability greater than 50%.

In order to address the problem of really large numbers of catalogs, we plan to adopt a hierarchical method of cross-matching. For instance, if we have 100 catalogs, we split them in 10 sets of 10 catalogs. *CSCXmatch* is run on each set of 10 and improved positions and error ellipses, as well as compound PSFs, are derived for the matched tuples. These 10 sets of improved sources are fed into *CSCXmatch* again as 10 pseudo catalogs to derive a final set of source matches over all 100 catalogs.

Proper Motions

One obvious set of matches that will be missed involves sources with significant proper motions in cases where the positions have different epochs. There are two potential reprieves from missing these, but they need to be implemented as post-matching fixes.

If at least one of the catalogs contains proper motion information, one can search the others for source detections in updated positions.

If one has source positions in a certain area from at least three different epochs, one can search for linear position changes in the final match set of sources. This is a viable project for CSC Release 2 since there are a number of locations that have been observed several times.

This work has been supported by NASA under contract NAS 8-03060 to the Smithsonian Astrophysical Observatory for operation of the Chandra X-ray Center

References

Budavári, T., & Loredo, T. J. 2015, Ann. Rev. Stat. Appl. 2015.2, 113
Budavári, T., & Szalay, A. 2008, ApJ 679, 301
Heinis, S., Budavári, T., & Szalay, A. 2009, ApJ 705, 739
Rots, A. H., & Budavári, T. 2011, ApJS 192, 8