Cross-Matching with the Chandra Source Catalog

Arnold Rots *2020-01-23*

Table of Contents

1. Scope

This note provides the recipe for calculating cross-match probabilities for CSC sources, using the algorithm from Budavári & Szalay (2008) and Heinis, Budavári, & Szalay (2009) that was also used for the CSC-SDSS cross-match, which, in turn, was used for the absolute astrometric error determination for Release 1 by Rots & Budavári (2011). See also Budavári & Loredo (2015). I will discuss multi (i.e., more than two) catalog matching and provide recommendations for situations where the spatial resolutions are significantly different.

2. Context and Parameters

Let us assume we have two sets of sources, $\mathscr{L}$ and $\mathscr{M}$, containing N_L sources from $\mathscr{L}$ that are contained in the intersection with the coverage of $\mathscr{M}$ and N_M sources from $\mathscr{M}$ that are contained in the intersection with the coverage of $\mathscr{L}$.

Each source n has an elliptical (95% confidence) position error associated with it, characterized by semi-major axis a_n (in radians), semi-minor axis b_n (in radians), and position angle ϕ_n. We assume that the error distribution is Gaussian.

3. Pair-wise Cross-Match Procedure

3.1. Initialization

First some basic spherical trigonometry. If we have two sources (α_1,δ_1) and (α_2,δ_2), the distance between the two, ψ, using the polar triangle including the two positions, is:

$$\psi = \arccos(\sin\delta_1 \cdot \sin\delta_2 + \cos\delta_1 \cdot \cos\delta_2 \cdot \cos(\alpha_1 - \alpha_2))$$

And, assuming that the positions are close enough that the angles at both of them are each other's supplement, the complement of the smaller of the two (the position angle of the vector **ψ**, counted counter-clockwise from the negative right ascension axis), φ, is:

$$\varphi = 0.5\,\pi - \arcsin\left(\frac{\sin(\alpha_1 - \alpha_2)}{\sin\psi}\cos\big(0.5(\delta_1 + \delta_2)\big)\right)$$

Each source i from $\mathscr{L}$ is matched against each source j from $\mathscr{M}$. and the Bayes factor B_{ij} for each of those matches is calculated as:

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

If $\overline{\boldsymbol{\psi}_{ij}}$ is the vector between sources i and j, let Ψ_{ij} be the angular distance in radians and ϕ_{ij} the vector's position angle. $\sigma_i(j)$ is the (Gaussian) standard deviation of the position error of source i along the vector $\overline{\boldsymbol{\psi}_{ij}}$, i.e., the distance, in radians, along that vector from the position of i to where it intersects with its error ellipse, scaled to Gaussian standard deviation. This means:

$$\sigma_i^2(j) = S_C \cdot \frac{a_i^2 b_i^2}{a_i^2 \sin^2\big(\phi_i - \phi_{ij}\big) + b_i^2 \cos^2\big(\phi_i - \phi_{ij}\big)}$$

$S_C = S_{95} = 0.1669041$ for conversion from 95% confidence level to standard deviation;
$S_C = S_{90} = 0.217146$ for 90% confidence.

The expression is the same for $\sigma_j(i)$, after switching i and j. Note that the expression is insensitive to the sign of the angle difference and allows its value to be taken modulo 180˚.

We start out with the prior $P_0(0)$:

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

The numbers of sources in the catalogs (N_L, N_M), as well as the expected number of matches, need to be scaled to the whole sky.

3.2. Iteration

Here the iteration starts.

Calculate for each pair (i, j) its posterior match probability P_{ij}:

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

Now update P_0:

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

Iterate until:

$$\frac{P_0(k+1) - P_0(k)}{P_0(k+1)} < 10^{-3}$$

Again, the numbers of sources in the catalogs (N_L, N_M), as well as the expected number of matches, need to be scaled to the whole sky. It is prudent to limit the maximum number of iterations to something like 20: if all Bayes Factors are very small this will not converge and no harm will be done if the iterations are terminated.

3.3.Probability Thresholds

Budavári & Szalay (2008), in Section 5.3, propose a self-consistent mechanism for determining the threshold that match probabilities have to meet in order to be considered accepted. On that basis we have adopted the following considerations and criterion for acceptance

For a given set of n m-tuples the iteration on the probabilities for the individual m-tuples is derived from the Bayes Factors and a prior that involves the sum of probabilities and the source densities. The issue here is that the source densities introduce a scaling of the probabilities as derived from the BFs (that's why the P versus $\log(BF)$ curves are never identical) and that for assigning matches we want to apply a uniform thresholding criterion. Here is the recipe:

Assume we have a list of n source m-tuples with probability $p[i]$ and a $S_P = \sum_{i=1}^{n} p[i]$.
Note that I count array elements as 1-relative for clarity.

1. If $S_P < 0.2$ reject all matches; else:
2. Sort the list according to decreasing $p[i]$
3. Set $k = S_P$ (truncate)
4. Set $P^* = p[k] + (k - S_P) \cdot (p[k] - p[k+1])$
   This is a simple linear interpolation
5. Set the threshold for these n m-tuples to $P = s \cdot P^*$
6. Accept all m-tuples in the list with $p[i] > P$

s should probably still be an input parameter. I have tested this extensively on the test sets Orion, M31, Sim-4, and Cosmos-4. The best results are obtained with $s = 0.90$. For increased transparency we decided to scale all probabilities to a common threshold value, (arbitrarily) set at 0.70, by setting $p^* = \frac{0.70}{P} p$, resulting in tuple matches with $p^* > s \cdot 0.70$ to be accepted as true matches.

4. Multi-Set Matching

Cross-matching sources between two sets is pretty straightforward, as demonstrated in Section 3. Extending this to more than two sets is not quite trivial.

The first problem here is that the number of n-tuples increases dramatically as the number of catalogs increases. To alleviate that and bring the number of n-tuples to be considered down to a manageable level, Budavári & Szalay (2008) propose a filtering algorithm that is based on the distance between the individual elements of the tuple. What I propose is to base consideration on the pair-wise probability between the elements. Suppose we have n catalogs

1. Determine the match probabilities between all sources in each pair of catalogs
2. Designate all sources as ambiguous that have more than one match probability above the limiting probability with each of the other catalogs, excepting matches with sources that have a PSF size 4.0 or more times greater
3. For $m = 3 \ldots n$ find all tuples for which at least ½$(m-1)(m-2)+1$ pairs have a match probability greater than the threshold specified in Section 3.3 and where each member is an unambiguous source
4. Determine the Bayes Factors as specified in Sections 4.1 and the probabilities as in Section 3
5. Accept tuples with probabilities exceeding the threshold in Section 3.3

Calculate the Bayes Factors for $n > 2$ by using their logarithms, to guard against overflows, following the expressions in one of the following two subsections.

4.1. Bayes Factors for Elliptical Errors

Tamás Budavári (private communication) provided the equations to be used for n-tuple matching using elliptical errors.

The covariance matrix $\bar{\bar{\boldsymbol{C}}}_i$ representing the error ellipse $(a_i, b_i, \phi_i = 0)$, as defined in Section 2, and its inverse, $\bar{\bar{\boldsymbol{C}}}_i^{-1}$ are:

$$\bar{\bar{\boldsymbol{C}}}_i = \begin{bmatrix} a_i^2 & 0 \\ 0 & b_i^2 \end{bmatrix} \qquad \bar{\bar{\boldsymbol{C}}}_i^{-1} = \begin{bmatrix} a_i^{-2} & 0 \\ 0 & b_i^{-2} \end{bmatrix}$$

For arbitrary ϕ_i the inverse covariance matrix becomes:

$$\bar{\bar{\boldsymbol{C}}}_i^{-1} = \frac{1}{a_i^2 b_i^2} \begin{bmatrix} a_i^2 \sin^2\phi_i + b_i^2 \cos^2\phi_i & (b_i^2 - a_i^2)\sin\phi_i \cos\phi_i \\ (b_i^2 - a_i^2)\sin\phi_i \cos\phi_i & a_i^2 \cos^2\phi_i + b_i^2 \sin^2\phi_i \end{bmatrix}$$

For reference, the inverse of the symmetric 2×2 matrix $\begin{bmatrix} p & r \\ r & q \end{bmatrix}$ is: $\frac{1}{pq - r^2}\begin{bmatrix} q & -r \\ -r & p \end{bmatrix}$.

Note that the denominators in the scaling coefficients are the determinants of the matrices to be inverted and that $\left|\bar{\bar{\boldsymbol{C}}}_i^{-1}\right| = \left|\bar{\bar{\boldsymbol{C}}}\right|^{-1}$.

We define for an *n*-tuple of sources the matrix $\bar{\bar{\boldsymbol{K}}}$ (effectively the covariance matrix of the combined tuple) through its inverse, as the sum of the inverse covariance matrices:

$$\bar{\bar{\boldsymbol{K}}}^{-1} = \sum_{i=1}^{n} \bar{\bar{\boldsymbol{C}}}_i^{-1}$$

And the vectors $\bar{\boldsymbol{y}}$ (the mean position of the tuple) and $\bar{\boldsymbol{u}}$ by:

$$\bar{\bar{\boldsymbol{K}}}^{-1} \cdot \bar{\boldsymbol{y}} = \sum_{i=1}^{n} \bar{\bar{\boldsymbol{C}}}_i^{-1} \cdot \bar{\boldsymbol{x}}_i = \bar{\boldsymbol{u}}$$

If the k'_{ij} are the elements of matrix $\bar{\bar{\boldsymbol{K}}}^{-1}$, then:

$$\bar{\boldsymbol{y}}^T \cdot \bar{\bar{\boldsymbol{K}}}^{-1} \cdot \bar{\boldsymbol{y}} = \frac{k'_{22}u_1^2 - 2k'_{12}u_1u_2 + k'_{11}u_2^2}{k'_{11}k'_{22} - {k'_{12}}^2} = \bar{\boldsymbol{u}}^T \cdot \bar{\bar{\boldsymbol{K}}} \cdot \bar{\boldsymbol{u}}$$

The position vectors $\bar{\boldsymbol{x}} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$ are defined in the tangent plane with tangent point ideally at $\bar{\boldsymbol{y}}$. In practice, one may choose to use, (α_0, δ_0), the average values of the source positions in the tuple. Then:

$$x_1 = \frac{-\cos\delta \sin(\alpha - \alpha_0)}{\sin\delta \sin\delta_0 + \cos\delta \cos\delta_0 \cos(\alpha - \alpha_0)}$$

$$x_2 = \frac{\sin\delta \cos\delta_0 - \cos\delta \sin\delta_0 \cos(\alpha - \alpha_0)}{\sin\delta \sin\delta_0 + \cos\delta \cos\delta_0 \cos(\alpha - \alpha_0)}$$

Finally, the Bayes Factor for the tuple becomes:

$$B = 2^{n-1} \frac{\sqrt{|\overline{\overline{\boldsymbol{K}}}|}}{\prod_{i=1}^{n} \sqrt{|\overline{\overline{\boldsymbol{C}}}_i|}} \exp\left\{\frac{1}{2}\left(\overline{\boldsymbol{y}}^T \cdot \overline{\overline{\boldsymbol{K}}}^{-1} \cdot \overline{\boldsymbol{y}} - \sum_{i=1}^{n} \overline{\boldsymbol{x}_i}^T \cdot \overline{\overline{\boldsymbol{C}}_i}^{-1} \cdot \overline{\boldsymbol{x}_i}\right)\right\}$$

Note that, since $\overline{\overline{\boldsymbol{K}}}$ is the covariance matrix of the complete tuple, the spatial properties of the master source are provided by the vector $\overline{\boldsymbol{y}}$, representing the resulting position, and the eigenvectors of $\overline{\overline{\boldsymbol{K}}}$, providing the parameters of its error ellipse; see also Section 9.

5. Spatial Resolution Troubles

Cross-matching has a dark side to it, as Tom Loredo commented recently, and this is particularly apparent in situations where the spatial resolution varies widely between the catalogs to be matched – and this is unfortunately applicable *par excellance* to the *Chandra* data. Consider two sources separated by a few arcseconds observed on-axis (where they are easily resolved) and 10 or 15 arcminutes off-axis (where there is no chance of resolving them). In such a situation the position errors may all be fairly small and essentially irrelevant: the deciding factor for matches is really the size of the PSF.

Although the use of error ellipses works well for matching catalogs with comparable spatial resolution and usually yields match probabilities of order 0.99, for general purpose matching I recommend using the PSF standard deviation ellipses. The implementation is further discussed in Section 7.

6. PSF Parameters

In principle, the parameters of the ECF 90% PSF ellipses are available from the Region files. For single-Obi stacks this does not present a problem and one can just take the values and scale the semi axes down by $\sqrt{0.217146}$. For multi-Obi stacks one needs to calculate a composite PSF where each Obi is assigned a weight proportional to the number of counts it contributes to the total number detected in the source. If there are p_i photons detected in Obi i of n Obis:

$$w_i = \frac{p_i}{\sum_{j=1}^{n} p_j}$$

One can take two approaches to deriving the parameters of the composite PSF:

6.1. Parameter Average **(current solution)**

Convert the individual Obi-based PSF ellipse parameters to coefficients of the canonical ellipse equation, calculate a weighted average, and convert the result back to ellipse parameters.

If a and b are the semi major and minor axes of an ellipse, with the major axis aligned along the x axis, the canonical equation is

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$$

Or:

$$(x\ y) \cdot \begin{bmatrix} \frac{1}{a^2} & 0 \\ 0 & \frac{1}{b^2} \end{bmatrix} \cdot \begin{bmatrix} x \\ y \end{bmatrix} = \overline{\boldsymbol{x}}^T \cdot \overline{\overline{\boldsymbol{Q}}} \cdot \overline{\boldsymbol{x}} = 1$$

For an ellipse rotated counter-clockwise by an angle ϕ the elements q_{ij} of matrix $\boldsymbol{Q}$ are:

$$q_{11} = \frac{\cos^2 \phi}{a^2} + \frac{\sin^2 \phi}{b^2}$$

$$q_{22} = \frac{\sin^2 \phi}{a^2} + \frac{\cos^2 \phi}{b^2}$$

$$q_{12} = q_{21} = \sin \phi \, \cos \phi \left(\frac{1}{a^2} - \frac{1}{b^2} \right)$$

Calculate the individual a_{ij} elements for each Obi, calculate their weighted averages, and derive the composite ellipse parameters through the reverse formulae:

$$a = \sqrt{\frac{2}{(q_{11} + q_{22}) - \sqrt{(q_{11} - q_{22})^2 + 4q_{12}^2}}}$$

$$b = \sqrt{\frac{2}{(q + q_{22}) + \sqrt{(q_{11} - q_{22})^2 + 4q_{12}^2}}}$$

$$\phi = \begin{cases} 0 & \text{for } q_{12} = 0 \text{ and } q_{11} < q_{22} \\ \frac{1}{2}\pi & \text{for } q_{12} = 0 \text{ and } q_{11} > q_{22} \\ \frac{1}{2}\arctan\left(\frac{2q_{12}}{q_{11} - q_{22}}\right) & \text{for } q_{12} \neq 0 \text{ and } q_{11} < q_{22} \\ \frac{1}{2}\pi + \frac{1}{2}\arctan\left(\frac{2q_{12}}{q_{11} - q_{22}}\right) & \text{for } q_{12} \neq 0 \text{ and } q_{11} > q_{22} \end{cases}$$

Note that for $q_{11} = q_{22}$ the ellipse becomes a circle.

6.2. PSF Average (preferred solution)

Construct a composite PSF by taking the weighted average of the PSF images and determine the ECF 90% ellipse or, alternatively, the 1-σ ellipse of this function directly.

7. Mixing Error Ellipses and PSFs

As noted above, in some cases calculating probabilities on the basis of error ellipses provides better results, in other cases calculating them on the basis of the PSF size. Qualitatively, it is easy to see why this would be the case. However, translating that into quantitative criteria is one of the knottiest problems. To make matters worse, it cannot be a global choice for certain catalog or stack sets, either:

even when Chandra stacks are pointed well apart from each other, there will always be areas where the PSF are similar, in size and rotation.

The decision on which ellipse to use will have to be made individually for each source pair. After some experimentation, two ways to approach this problem have emerged as feasible candidates, one simply based on which pair of ellipses provides the highest Bayes Factor, the other based on the fraction of overlap area between the two PSFs.

7.1. Value of the Bayes Factor (preferred solution)

For each source tuple that is to be considered calculate the Bayes Factor based on the error ellipses as well as on the PSFs. Use the largest value and proceed. This is a very simple prescription that was originally rejected, but after the introduction of additional improvements it turns out to work well and avoids the necessity of the PSF overlap ratio parameter which presents by itself the knotty problem referred to above.

7.2. Ellipse Overlap Fraction (acceptable solution)

The premise is that if the PSFs are very similar one should use the error ellipses; otherwise the PSF ellipses should be used. The criterion as to what “very similar” means quantitatively is to be based on the fractional area of the larger of the two PSFs in which the two overlap. Two concentric ellipses can have zero, two, or four points of intersection. If there are zero or two such points, the fractional overlap area is the ratio of the area of the smaller to that of the larger ellipse. If there are four intersection points the area of overlap may be approximated by that of the parallelogram formed by those four points. If the fraction is greater than P_{65}, one should use the error ellipses, otherwise use the PSFs. P_{65} is another tweaking parameter; we shall set it to 0.65 for now. This has been implemented in CSCxmatchMS2mix6.

The area of an ellipse with semi axes a and b is πab.

The points of intersection may be determined using the algorithm provided by Hughes and Chraibi (2012). If we have two ellipses (a_1, b_1, ϕ_1) and (a_2, b_2, ϕ_2), rotate them by $-\phi_1$, and set $\phi = \phi_2 - \phi_1$, we can define coefficients for canonical equations $px^2 + qy^2 + rxy = 1$:

$$p_1 = \frac{1}{a_1^2}, \qquad q_1 = \frac{1}{b_1^2}, \qquad r_1 = 0$$

$$p_2 = \frac{\cos^2\phi}{a_2^2} + \frac{\sin^2\phi}{b_2^2}, \qquad q_2 = \frac{\sin^2\phi}{a_2^2} + \frac{\cos^2\phi}{b_2^2}, \qquad r_2 = 2\sin\phi\cos\phi\left(\frac{1}{a_2^2} - \frac{1}{b_2^2}\right)$$

Then we derive the coefficients of the quadratic equation for the squares of the y-coordinates of the intersections:

$$py^4 + qy^2 + r = 0$$

$$p = -p_1 q_1 r_2^2 - (p_1 q_2 - q_1 p_2)^2$$

$$q = 2(p_1 - p_2)(p_1 q_2 - q_1 p_2) + a_1 r_2^2$$

$$r = -(p_1 - p_2)^2$$

The determinant:

$$D = q^2 - 4pr \begin{cases} < 0: & \text{no intersecting points} - \text{smaller contained in larger} \\ = 0: & \text{two intersecting (tangent)points} - \text{smaller contained in larger} \\ > 0: & \text{four intersecting points} - \text{proper intersection} \end{cases}$$

The coordinates of the four points of intersection are then easily derived:

$$y^2 = \frac{-q \pm \sqrt{D}}{2p}$$

$$x^2 = a_1^2 \left(1 - \frac{y^2}{b_1^2}\right)$$

The area A of the overlap parallelogram is most easily calculated as half the sum of the outer products of the neighboring vertex vectors, taking care that they are followed counter-clockwise:

$$A = 0.5\big((x_1 y_2 + x_2 y_3 + x_3 y_4 + x_4 y_1) - (x_2 y_1 + x_3 y_2 + x_4 y_3 + x_1 y_4)\big)$$

A better solution is to calculate the ellipse sectors exactly. The area of an ellipse sector between angles θ_0 and θ_1, counted counter-clockwise from the major axis, is (David Eberly, Geometric Tools, LLC):

$$A(\theta_0, \theta_1) = F(\theta_1) - F(\theta_0)$$

Where:

$$F(\theta) = \frac{ab}{2} \cdot \left\{\theta - \arctan\left(\frac{(b-a)\ \sin 2\theta}{(b+a) + (b-a)\cos 2\theta}\right)\right\}$$

8. Very Large Numbers of Catalogs

Section 4 provided the logic for matching any number of catalogs. However, beyond 10-15 catalogs the number of combinations to be considered becomes impractically large and we need to apply additional measures to keep the problem manageable.

8.1. Connected Components and Exclusion of Ambiguous Sources

Dan Nguyen proposed to use the Connected Components algorithm as implemented in the Boost Graphic Library to identify the candidate source tuples that need to be taken into consideration, based on the pair-wise match probabilities, without having to inspect all possible tuples. This is extremely effective, but it requires the prior exclusion of pair-wise matches with sources that have ambiguous matches. These are identified by inspecting all source pairs that satisfy the match probability threshold and excluding sources that have more than one qualified match with any source in any single other catalog. A qualified match is defined as a match exceeding the probability threshold where the ratio of the areas of the PSF ellipses of the source under consideration and the matched source is less than P_{25}.

I.e., this is an asymmetric criterion where matches with sources that have much larger PSFs (and hence are more prone to an ambiguous match) are excluded for the source with the smaller PSF. $P_{25} = 0.25$ has been found to be a good value to adopt.

8.2. Recursive Matching

A second option is to adopt a hierarchical scheme of recursive matching:

- Group the stacks in into manageable sub-ensembles; this grouping may be done intelligently (i.e., not at random)
- Run these through the pipeline
- Determine the Master Source properties (see below) for the established Master Sources, using only unambiguously matched sources, and put them into pseudo stacks
- Run the pipeline again on these pseudo stacks

This method is consistent with the recommendations made by Budavári & Szalay (2008).

9. Spatial Master Source Properties

The spatial Master Source properties are determined on the basis of the following two equations note that this is similar to the matter of determining elliptical errors; see Section 4.1 for details):

$$\bar{\bar{\boldsymbol{K}}}^{-1} = \sum_{i=1}^{n} \bar{\bar{\boldsymbol{C}}}_i^{-1}$$

$$\bar{\bar{\boldsymbol{K}}}^{-1} \cdot \bar{\boldsymbol{y}} = \sum_{i=1}^{n} \bar{\bar{\boldsymbol{C}}}_i^{-1} \cdot \bar{\boldsymbol{x}}_i$$

Their positions are determined by evaluating the vector $\bar{\boldsymbol{y}}$, while error ellipses are determined by their inverse covariance matrices $\bar{\bar{\boldsymbol{K}}}^{-1}$, calculated as the sum of the inverse covariance matrices of the contributing unambiguous matches. Of course, the (tangent plane) vector $\bar{\boldsymbol{y}}$ needs to be transformed to (α, δ); the plane's tangent point is (α_0, δ_0). The Master Source PSFs are determined by their inverse covariance matrices $\bar{\bar{\boldsymbol{K}}}^{-1}$, calculated as the sum of the inverse covariance matrices of the contributing unambiguous matches weighted by the inverse area of their error ellipses (i.e., the inverse of the product of the error ellipse semi-axes for the source they refer to); the weights are to be scaled so their sum is unity.

10. 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

Hughes, G. B., & Chraibi, M. 2012, Comput Vis Sci 15, 291

Rots, A. H., & Budavári, T. 2011, ApJS 192, 8