NWAY - The versatile catalogue matching tool

User Manual, version 1.0

Written by Johannes Buchner

22. August 2016

Contents

1 Why NWAY?

In astrophysics, a common task is to assemble multi-wavelength information about individual sources. This is done by taking the detections of sources in the sky (positions, errors, and fluxes/magnitudes) from a catalogue of one wavelength and matching it to another from another wavelength, or multiple such catalogues. Care has to be taken to consider all possible matches and also the possibility that the source does not have a counterpart in a catalogue of a given depth. For many classes of sources, the Spectral Energy Distribution (SED) provides additional hints, which associations are likely real. For instance, the color distribution of stars in the WISE bands is different than that of quasars or galaxies.

NWAY is a generic solution to these tasks:

1. Matching of N catalogues simultaneously
2. Consideration of all possible matches, partial matches and absence of counterparts in some catalogues.
3. Taking into account the positional uncertainty.
4. Incorporating magnitude, color or other information about the sources of interest.
5. Computation of a probability for each possible match.
6. Computation of a probability that there is no match.

Contributors

- Mara Salvato – idea and leading the science case.
- Tamás Budavári – shared basic implementation of his formulae.
- Sotiria Fotopoulou – initial implementation for matching three catalogues.
- Johannes Buchner – complete code rewrite for the general case, documentation, manual and adding features.

How to read this manual

This manual walks you through the installation and usage. It is best to try out the examples as you read, because they illustrate how NWAY works.

Yellow boxes indicate text about the illustrative examples and their explanations.

Useful explanation for users on how to prepare files or how to run NWAY are found in yellow boxes.

Green boxes indicate examples you can run yourself to learn NWAY.

Sometimes there are `commands` you can copy-paste.

If you are interested in the rigorous mathematical details, go to Chapter 5.

Terminology

Source A detection in a certain wavelength significant enough to be recorded in a catalogue with sky position.

Counterpart The corresponding source in another catalogue.

Association A specific combination of entries from the various catalogues, i.e. a tuple of detections associated with each other.

Match Same as association, used interchangeably.

Object A physical entity in the real universe, emitting radiation.

2 User Manual

2.1 Installation

NWAY is a pure Python program. Install NWAY through

```
$ sudo pip install nway
```

This will give you the tool `nway.py`.

Or if you do not have root access:

```
$ pip install nway -user
```

The tool `nway.py` is installed for you as `~/.local/bin/nway.py`. If that directory is in your `$PATH`, you can run `$ nway.py --help`.

Development version To get the latest development version, fetch NWAY from

https://github.com/JohannesBuchner/nway

and run in the directory

```
$ python setup.py install --user.
```

2.2 Citing NWAY correctly

Please cite Salvato et al. (in prep.).

2.3 Development, questions and issues

Please let us know if you have comments about this manual or problems running/installing NWAY.

For reporting bugs or requesting features, NWAY's issue tracker is at the following location.

https://github.com/JohannesBuchner/nway

NWAY is a small but powerful tool. Most questions or apparent issues arise to understand which information lead to a counterpart being preferred, and therefore understanding the input data well. This manual will guide you through the computation and the pieces of information added together.

2.4 Simple distance-based matching

Before exploring the full power of NWAY, we consider a simple, illustrative case. We have three catalogues, provided as FITS files:

Input:

Primary Catalogue
A
B
...

2nd Catalogue
α
β
...

3rd Catalogue
$\mathbb{A}$
$\mathbb{B}$
...

Output:

Primary Catalogue Entry	2nd Catalogue Entry	3rd Catalogue Entry	Probability	
A	α	$\mathbb{A}$	...	
A	α	$\mathbb{B}$	...	
A	α	(none)	...	
A	β	$\mathbb{A}$	...	
A	β	$\mathbb{B}$	...	A group
A	β	(none)	...	
A	(none)	$\mathbb{A}$	...	
A	(none)	$\mathbb{B}$	...	
A	(none)	(none)	...	
B	...	...	...	B group

In NWAY, only the first catalogue (**the primary catalogue**) plays a special role. For each entry of it, counterparts are sought from the other catalogues.

2.4.1 Example - Preparing input files

Note these points about preparing a catalogue input file:

1. Each catalogue needs to be a FITS file. The second extension should be the table (first extension is a header). TOPCAT writes files in this way.

   Three example catalogues are provided for you in the `doc/` directory: **COSMOS_IRAC.fits, COSMOS_OPTICAL.fits and COSMOS_XMM.fits**.

2. The data table needs to have a extension name and the keyword SKYAREA. The extension name is used as a prefix as all columns are copied to the output catalogue. The SKYAREA keyword tells the area on the sky in square degrees covered by the catalogue. This is important for estimating the chance of random alignments. You can use the tool `python write_header.py mycat.fits mytablename myskyarea` to set the fits header.

   For our example files we have a optical, IRAC and XMM catalogue covering 2 square degrees:

   ```
   python write_header.py COSMOS_OPTICAL.fits OPT 2;
   python write_header.py COSMOS_IRAC.fits IRAC 2
   python write_header.py COSMOS_XMM.fits XMM 2
   ```

3. Each catalogue needs to have a column RA and DEC. To make your life easier, NWAYtries to be a bit fuzzy and detect the columns named RA_something etc. It will print out which columns it found and used.

4. The primary catalogue needs to have a ID column. In our example this is the X-ray catalogue. To make your life easier, NWAYtries to be a bit fuzzy and detect the columns named ID_something etc. It will print out which columns it found and used.

Every possible combination of association is considered. However, in practice you do not want an extremely large output catalogue with extremely distant, unlikely to be physically associated. You can set the largest distance in degrees to consider by setting `--radius`. This speeds up the computation. But use a value that is much larger than the largest positional error.

2.4.2 Example - Matching two catalogues

Lets try the simplest example and match the XMM X-ray catalogue to an optical catalogue. The XMM catalogue has a pos_err column with the positional error in arcseconds. For the optical catalogue we will assume a error of 0.1 arcseconds.

Run this command:

```
python ../nway.py COSMOS_XMM.fits :pos_err COSMOS_OPTICAL.fits 0.1 --out=example1.fits --radius 7 --prior-completeness 0.9
```

Lets understand what we put in:

1. We passed two catalogue files: COSMOS_XMM.fits and COSMOS_OPTICAL.fits. For the first one, we told NWAYto use the column (":") `pos_err` in that catalogue for the positional error (**always in arcsec**). For the second one we specified a fixed error of 0.1 arcsec.

2. We specified where the output should be written (`--out`).

3. The largest XMM error is 5.66 arcsec, so we adopt a cropping radius of 7 arcsec to speed up the matching (`--radius 7`).

4. The parameter `--prior-completeness 0.9` is mentioned below.

Lets understand what NWAY did:

1. ```
   nway arguments:
       catalogues:  COSMOS_XMM.fits, COSMOS_OPTICAL.fits
       position errors/columns:  :pos_err, 0.1
         from catalogue "XMM" (1797), density is 3.706579e+07
         from catalogue "OPT" (560536), density is 1.156188e+10
       magnitude columns:
   ```

   It reads the catalogues and looks at their densities.

2. ```
   matching with 7.000000 arcsec radius
   matching: 1007283192 naive possibilities
   matching: hashing
       using RA  columns: RA, RA
       using DEC columns: DEC, DEC
   100%562333|########################################|Time: 0:00:24
   matching:  32782 matches after hashing
   matching: collecting from 755640 buckets
   100%755640|########################################|Time: 0:00:02
   matching:  14956 unique matches from crossproduct
   matching:  13159 matches
   merging columns ... 10
   100% 10|########################################|Time: 0:00:00
   merging columns: adding angular separation columns
   matching:   5259 matches after filtering
   ```

   Within 30 seconds it created a cross-match of possible associations (1007283192 in principle, 5259 within 7 arcseconds).

3. It found ID, RA, DEC, and positional error columns.

4. ```
   finalizing catalogue
       finding position error columns ...
       Position error for "XMM": found column XMM_pos_err:
           Values are [0.109000..5.660000]
       Position error for "OPT": using fixed value 0.100000
       finding position columns ...
       computing probabilities from separations ...
       grouping by column "XMM_ID" for flagging ...
   ```

   It computed the probability of each association.

5. ```
   writing "example1.fits" (5259 rows, 17 columns)
   ```

   It wrote the output file `example1.fits`. This file contains all columns from the input catalogues and the computed probabilities (see below for their meaning).

So how does NWAY deal with a particular, possible association and compute its probability?

The probability of a given association is computed by comparing the probability of a random chance alignment of unrelated sources (prior) to

the likelihood that the source is the same. The gory mathematical details are laid out in Section 5.1, but from a user point of view the following is important:

1. The chance of a random alignment depends on the source sky density of the various catalogues. **So each catalogue needs to have a FITS header entry `SKYAREA` which tells the area covered by the catalogue in square degrees.** The source density is then computed by the number of entries divided by that area. You can use the tool `python write_header.py mycat.fits mytablename myskyarea` to set the fits header.

2. Varying depths between the catalogues and different coverage can further reduce the fraction of expected matches. This can be adjusted by setting `--prior-completeness=0.9`, if previous experience is that only 90% of sources have a match with the given inputs.

The outputs catalogue then contains three new columns along all columns of the input catalogues:

1. `bf`: logarithm of ratio between prior and posterior from distance matching

2. `bfpost`: Distance posterior probability.

3. `post`: Posterior probability (same as bfpost unless additional information was added).

4. **`post_group_no_match`**: For each entry in the primary catalogue (e.g. A) the probability that no association is the correct one is computed. Because every catalogue is limited by its depth, it is possible that the true counterpart has not been found yet. When building a secure catalogue, **keep only associations where `post_group_no_match<0.`**

5. **`post_group_this_match`**: For each possible association for each entry in the primary catalogue (e.g. A), the relative probability is computed. When building a secure catalogue, **keep only associations where `post_group_this_match>0.6`. Secondary solutions down to 0.1 may be interesting.**

6. **`match_flag`**: The best match is indicated with `1` for each primary catalogue entry. Secondary, almost as good solutions are marked with `2`. The maximum allowed difference is at most 0.005 (`--acceptable-prob` parameter), and also `post>=0.1` is required. All other associations are marked with `0`.

Use the last three columns to identify sources with one solution, possible secondary solutions, and to build final catalogues. Chapter 5.1 explains how these quantities are computed. The `--min-prob` parameter can be used to remove low-probability associations from the catalogue.

2.4.3 Example - Output of matching two catalogues

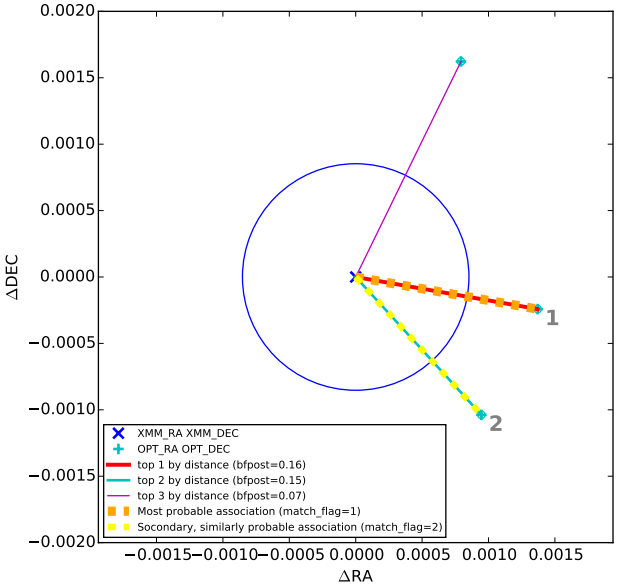
Lets understand the output fits file and the associations found for a particular X-ray source.

Open the fits file and find XMM_ID=60388. As you can see, this is a ambiguous case, where more than one optical counterpart is possible.

Below is an illustration of this ambiguous case (produced with `python explain.py example1.fits 60388`).

Two sources are at a similar distance from the X-ray source (blue, with error circle). Therefore their association probability (`bfpost`) is similar. The slightly higher one is marked as match_flag=1 (orange), the other with 2 (yellow).

Section 2.5 solves this by adding more information (the magnitude distribution). But we can also solve this another way. We know AGN (the X-ray source) emits in the infrared, so you can also match with an IRAC catalogue like so: python ../nway.py COSMOS_XMM.fits :pos_err COSMOS_OPTICAL.fits 0.1 COSMOS_IRAC.fits 0.5 --out=example3.fits --radius 7



2.5 Matching with additional information

For many classes of sources, the Spectral Energy Distribution (SED) provides additional hints, which associations are likely real. For instance, the color distribution of stars in the WISE bands is different than that of quasars or galaxies. A powerful feature of NWAY is to take advantage of this additional information to improve the matching. Section 5.2 has the mathematical details and a comparison to the Likelihood Ratio method.

2.5.1 Example - Using magnitude information

AGN (which we are looking for in our example) have a different optical magnitude distribution than other objects. Lets take advantage of this information.

Run this command: `python ../nway.py COSMOS_XMM.fits :pos_err COSMOS_OPTICAL.fits 0.1 --out=example2.fits --radius 7 --prior-completeness 0.9 -mag OPT:MAG auto -mag-radius 3.5`

The last two parts are new: `--mag OPT:MAG auto --mag-radius 3.5`. Lets look at what they mean:

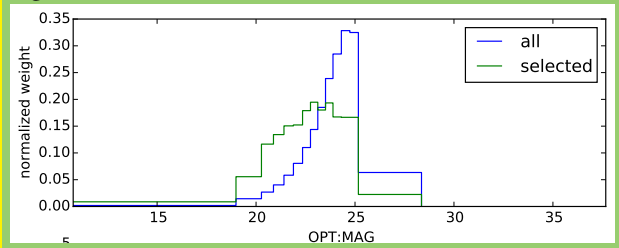
1. You have to tell NWAY in which catalogue and which column to find this information for a individual source. Adding the argument `--mag catalogue-table-name:catalogue-table-column [auto|histogramfile.txt]` does this. Remember that the FITS extension containing the table has a name (OPT in our example).

2. If we know the magnitude distribution of X-ray detected AGN we can provide this prior distribution as a table (histogram). This table contains the color histogram of the sources of interest (X-ray detected AGN) and a histogram of other, field sources.

3. Specifying `auto` instead of a histogram file, as we do in our example, derives the two distributions from the data. For this, all sources inside 3.5 arcseconds of a X-ray source are put into one histogram, and all others into another histogram (`--mag-radius` parameter).

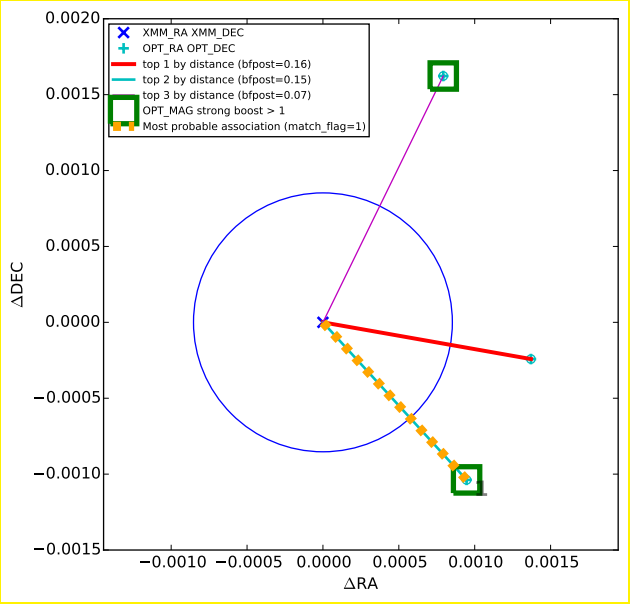
(continued below)

Lets look at the histograms it computed (NWAY created `OPT_MAG_fit.pdf`, and also `OPT_MAG_fit.txt` as a histogram file):



They are clearly different: Lower magnitude (bright) sources are more likely associated to AGN. This will help our matching.

As an example, we show below the ambiguous case from before. The lower association has been selected because it has a better match by magnitude, resolving the ambiguity.



Multiple priors You can specify as many priors as you like, taking advantage of more and more information. Just repeat --mag.

The following example uses one prior from the optical catalogue and another prior from the IRAC catalogue. A three-way match is performed.

```
python ../nway.py COSMOS_XMM.fits :pos_err COSMOS_OPTICAL.fits 0.1 COSMOS_IRAC.fits 0.5 --out=example3.fits --radius 7 --prior-completeness 0.9 --mag OPT:MAG auto --mag IRAC:mag_ch1 auto --mag-radius 3.5
```

Providing a prior as a file In the paper we used the WISE color distribution of X-ray detected AGN from another study. We provided this color prior distribution as a ASCII table (histogram). This table contains the color histogram of the sources of interest (X-ray detected AGN) and a histogram of other, field sources. The file `OPT_MAG_fit.txt` is an example of such a input file, and can be used via `--mag OPT:MAG OPT_MAG_fit.txt`. It contains four columns (lower and upper bin edge, density of selected and non-selected) and looks like this:

```
# OPT_MAG_fit.txt
# lo          hi          selected    others
  10.76000    18.98571    0.00870     0.00183
  18.98571    20.27286    0.05562     0.01448
  ...
```

A general approach Providing priors is not limited to magnitude distributions, you can use colors or any other information you want (e.g. morphology, variability, etc.). The approach is very general, NWAY just looks at the corresponding bin and reweighs the probabilities.

Discovering a prior from distance matching If you set `--mag OPT:MAG auto` and do not set `--mag-radius`, NWAYuses the Bayesian distance matching for discovering the histogram of `OPT:MAG`. It divides association into three groups:

1. Those with `bfpost>0.9` are considered safe matches and used for the "selected" histogram (AGN in our example).
2. Those with `bfpost<0.01` are considered safe non-matches and used for the "all" histogram (non-AGN in our example).
3. All others are considered ambiguous and not used for the histogram.

This is more cautious but if you do not have a large catalogue you may not have enough sources to build a well-sampled histogram.

3 Program Arguments

1. `python ../nway.py --help` ... display help page
2. `python ../nway.py catalogue1 catalogues ...` ... input catalogues as FITS files.
3. `python ../nway.py --out` ... Output file name (also a FITS file).
4. `python ../nway.py --radius` This radius (in degrees) is used to discard distant pairs. Always choose a value that is much larger than the largest positional uncertainty, then this value will not change the results. Smaller values just make the code run faster by reducing the number of combinations to explore.
5. `python ../nway.py --prior-completeness 1` ... set expected matching completeness (default: 1)
6. `python ../nway.py --mag MAGCOLUMN MAGFILE` ... name of <table>:<column> for magnitude biasing, and file name for magnitude histogram (use auto for auto-computation within mag-radius). Example: `--mag GOODS:mag_H auto --mag IRAC:mag_irac1 irac_histogram.txt`
7. `python ../nway.py --mag-radius` ... If set, and a auto prior is defined, then the selected sources are taken from within this radius of the primary sources (in arc seconds). If not set (recommended), the Bayesian posterior from distance matching is used, which incorporates positional errors.
8. `python ../nway.py --min-prob` ... only retain associations in the output catalogue exceeding this `post` value. Recommended: 0.1.
9. `python ../nway.py --acceptable-prob` ... affects the flagging of secondary solutions (`match_flag` column). If the secondary is within this difference (default: 0.005), it is marked as a secondary solution.

4 Input file specifications and units

1. Each catalogue needs to be a FITS file. The second extension should be the table (first extension is a header).
2. The data table needs to have a extension name.
3. The header of the data table needs the keyword SKYAREA, which specifies the area covered by the catalogue in **square degrees**.
4. Each catalogue needs to have a column RA and DEC **in degrees**. To make your life easier, NWAYtries to be a bit fuzzy and detect the columns named RA_something etc.
5. The primary catalogue needs to have a ID column. To make your life easier, NWAYtries to be a bit fuzzy and detect the columns named ID_something etc.
6. Positional error columns need to be in arcseconds.

Example catalogues are provided in the `doc/` directory: COSMOS_IRAC.fits, COSMOS_OPTICAL.fits and COSMOS_XMM.fits.

5 Mathematical details and Implementation

5.1 Distance-based matching

Lets consider the problem of finding counterparts to a primary catalogue ($i = 1$), in our example for the X-ray source position catalogue. Let each N_i denote the number of entries for the catalogues used, and $\nu_i = N_i/\Omega_i$ denote their source surface density on the sky.

If a counterpart is required to exist in each of the k catalogues, there are $\prod_{i=1}^{k} N_i$ possible associations. If we assume that a counterpart might be missing in each of the matching catalogues, there are $N_1 \cdot \prod_{i=2}^{k}(N_i + 1)$ possible associations. This minor modification, negligible for $N_i \gg 1$, is ignored in the following for simplicity, but handled in the code.

If each catalogue covers the same area with some respective, homogeneous source density ν_i, the probability of a chance alignment on the sky of physically unrelated objects can then be written (Budavári & Szalay, 2008, eq. 25) as

$$P(H) = N_1 / \prod_{i=1}^{k} N_i = 1 / \prod_{i=2}^{k} N_i = 1 / \prod_{i=2}^{k} \nu_i \Omega_i. \tag{5.1}$$

Thus $P(H)$ is the prior probability of an association. The posterior should strongly exceed this prior probability, to avoid false positives.

To account for non-uniform coverage, $P(H)$ is modified by a "prior completeness factor" c, which gives the expected fraction of sources with reliable counterpart (due to only partial coverage of the matching catalogues $\Omega_{i>1} \neq \Omega_1$, depth of the catalogues and/or systematic errors in the coordinates). Our prior can thus be written as

$$P(H) = c / \prod_{i=2}^{k} \nu_i \Omega_1. \tag{5.2}$$

Bayes' theorem connects the prior probability $P(H)$ to the posterior probability $P(H|D)$, by incorporating information gained from the observation data D via

$$P(H|D) \propto P(H) \times P(D|H). \tag{5.3}$$

Then, comparing two different hypotheses (H and $\bar{H}$), we can write how probable one is compared to the other as

$$\frac{P(H|D)}{P(\bar{H}|D)} \propto \frac{P(H)}{P(\bar{H})} \times \frac{P(D|H)}{P(D|\bar{H})} \tag{5.4}$$

$$= \frac{P(H)}{P(\bar{H})} \times B \tag{5.5}$$

where the Bayes factor B indicates the strength of hypothesis H based on the observations. In the case considered here, for each association, the model "chance alignment" is compared with the model "real association". The relevant Bayes factor, developed in (Budavári & Szalay, 2008, eq. 18), is dependent on the angular distance ϕ_{ij} between the source positions in catalogues i and j:

$$B = \frac{P(\text{"real association"}|D)}{P(\text{"chance alignment"}|D)}$$

$$= 2^{n-1} \frac{\prod \sigma_i^{-1}}{\sum \sigma_i^{-1}} \exp\left\{ -\frac{\sum_{i=1}^{j} \phi_{ij} \sigma_j^{-1} \sigma_i^{-1}}{\sum \sigma_i^{-1}} \right\} \tag{5.6}$$

Here, σ_i denotes the positional uncertainty in each catalogue. The output column `bf` is $\log B$.

For each association the posterior of the hypothesis "real association" is then

$$P(\text{"real association"}|D) = \left[1 + \frac{1 - P(H)}{B \cdot P(H)} \right]^{-1} \tag{5.7}$$

For each combinatorically possible association across the catalogues, Equations 5.7 and 5.6 describe the probability that this association is real (i.e. not by chance). The output column `bfpost` is the result of equation 5.7.

The task now is to begin with the primary catalogue and find for each source one or more realistic associations to consider.

First, we write the probability that the primary catalogue entry has any association at all, by summing the probability of its possible counterparts:

$$P(\text{"any real association"}|D) = \left[1 + \frac{1 - P(H)}{\sum_k B_k \times P(H)} \right]^{-1} \tag{5.8}$$

This is possible because associations are mutually exclusive. Based on this posterior we can reject or accept the hypothesis that a real association exists for each object in the primary catalogue. The output column

`post_group_no_match` is the result of equation 5.8 through

$$\texttt{post\_group\_no\_match} = 1 - P(\text{"any real association"}|D).$$

If we accept the hypothesis that this source has a counterpart, the probability for any specific association k is

$$P(\text{k}|D, \text{ "any real association"}) = B_k / \sum_k B_k. \tag{5.9}$$

Here, I have used that *a priori* all associations are equally likely. This posterior can be used to reject unlikely counterparts. The output column `post_group_this_match` is the result of equation 5.9. A "very secure" counterpart could be defined by the requirement $P(\text{"any real association"}|D) > 99\%$ and $P(\text{k}|D, \text{ "any real association"}) > 95\%$, for example. However, it is useful to run simulations to understand the rate of false positives. Typically, much lower thresholds are acceptable.

One subtlety of Equation 5.8 is that it only considers the associations enumerated, i.e. the combinations of all detected objects. However, very faint counterparts may have not been detected with the current exposure depth. The possibility that another, undetected counterpart is the correct one is thus not included in Equation 5.8 nor Equation 5.9. Low probabilities in Equation 5.7 even for the most probable association may indicate that no suitable counterpart has been found yet.

5.2 Magnitudes, Colors and other additional information

Astronomical objects of various classes often show distinct color and magnitude distributions. Because most bright X-ray point-sources in deep images are also optically bright compared to generic sources, this information can be exploited. Previous works (Brusa et al., 2005, 2007) have modified the likelihood ratio coming from the angular distance $f(r)$ information (likelihood ratio method, Sutherland & Saunders, 1992) by a factor:

$$LR = \frac{q(m)}{n(m)} \times f(r) \tag{5.10}$$

Here, $q(m)$ and $n(m)$ are associated with the magnitude distributions of source (e.g. AGN) and background objects (e.g. stars, passive galaxies) respectively, but additionally contain sky density contributions.

This idea can be put on solid footing within the Bayesian framework. Here, two Bayes factor are combined, by simply considering two independent observations, namely one for the positions, D_ϕ, and one for the magnitudes D_m. The Bayes factor thus becomes

$$B' = \frac{P(D_\phi|\text{"real association"})}{P(D_\phi|\text{"chance alignment"})} \times \frac{P(D_m|\text{"real association"})}{P(D_m|\text{"chance alignment"})} \quad (5.11)$$

$$= B \times \frac{\bar{q}(m)}{\bar{n}(m)}, \quad (5.12)$$

with $\bar{q}(m)$ and $\bar{n}(m)$ being the probability that a X-ray source or a background object has magnitude m respectively.

For completeness, I mention the fully generalised case. This is attained when an arbitrary number of photometry bands are considered, each consisting of a magnitude measurement m and measurement uncertainty σ_m:

$$B' = B \times \prod \frac{\int_m \bar{q}(m)\, p(m|D_m)\, dm}{\int_m \bar{n}(m)\, p(m|D_m)\, dm} \quad (5.13)$$

Here, $p(m|D_m)$ would refer to a Gaussian error distribution with mean m and standard deviation σ_m. This is convolved with the distribution properties. Alternatively, $p(m|D_m)$ can also consider upper limits. The posterior formulae $P(\cdot|D)$ introduced above (Equations 5.8 and 5.9) remain the same, with B' replacing B.

5.3 Auto-calibration

The probability distributions $\bar{q}(m)$ and $\bar{n}(m)$ can be taken from other observations by computing the magnitude histograms of the target population (e.g. X-ray sources, AGN) and other objects selected in other wavebands.

Under certain approximations and assumptions, these histograms can also be computed during the catalogue matching procedure while also being used for the weighting. One could perform the distance-based matching procedure laid out above, and compute a magnitude histogram of the secure counterparts as an approximation for $\bar{q}(m)$ and a histogram of ruled out counterparts for $\bar{n}(m)$. While the weights $\bar{q}(m)/\bar{n}(m)$ may strongly influence the probabilities of the associations for a single object, the bulk of the associations will be dominated by distance-weighting. One may thus assume that the $\bar{q}(m)$ and $\bar{n}(m)$ are computed with and without applying the magnitude weighting are the same, which is true in practice. When differences are noticed, they will only strengthen $\bar{q}(m)$, and the procedure may be iterated.

5.4 Implementation

My implementation for matching n catalogues is a Python program called NWAY. The input catalogues have to be in FITS format. Information about the (shared) sky coverage has to be provided to the program as well. The program proceeds in four steps.

First, possible associations are found. It is unfeasible to consider all theoretical possibilities (complexity $O(\prod_{i=1}^{k} N_i)$), so the sky is split first to cluster nearby objects. For this, a hashing procedure puts each object into square bins. The bin width w is chosen so that an association of distance w is improbable, i.e. much larger than the largest positional error. An object with coordinates ϕ, θ is thus put into bin $(i,\, j) = (\lfloor \phi/w \rfloor\, , \lfloor \theta/w \rfloor)$, but also into bins $(i{+}1,\, j)$, $(i,\, j{+}1)$ and $(i{+}1,\, j{+}1)$ to avoid boundary effects. This is done for each catalogue separately. Then, in each bin, the Cartesian product across catalogues (every possible combination of sources) is computed. All associations are collected across the bins and filtered to be unique. The hashing procedure adds very low effort $O(\sum_{i=1}^{k} N_i)$ while the Cartesian product is reduced drastically to $O(N_{bins} \cdot \prod_{i=1}^{k} \frac{N_i}{N_{bins}})$. All primary objects that have no associations past this step have $P(\text{"any real association"}|D) = 0$.

The second step is the computation of Bayes factors using the angular distances between counterparts. The prior is also evaluated from the size of the catalogue and the effective coverage, as well as the user-supplied prior incompleteness factor. The posterior for each association based on the distances only is calculated.

In the third step the magnitudes are considered, and the Bayes factors modified. An arbitrary number of magnitude columns in the input catalogues can be specified. It is possible to use external magnitude histograms (e.g. for sparse matching with few objects) as well as computing the histograms from the data itself (see Section 5.3). The breaks of the histogram bins are computed adaptively based on the empirical cumulative distribution found. Because the histogram bins are usually larger than the magnitude measurement uncertainty, the latter is currently not considered. The adaptive binning creates bin edges based on the number of objects, and is thus independent of the chosen scale (magnitudes, flux). Thus the method is not limited to magnitudes, but can be used for virtually any other known object property (colours, morphology, variability, etc.).

In the final step, associations are grouped by the object from the primary catalogue (here: the X-ray source). The posteriors $P(\text{"any real association"}|D)$ and $P(\text{k}|D, \text{"any real association"})$ computed. For the output catalogue a cut on the posterior probability (e.g. above 80%) is applied, and all associations with their posterior probability are written to the output fits

catalogue file.

Bibliography

Brusa, M., Comastri, A., Daddi, E., Pozzetti, L., Zamorani, G., Vignali, C., Cimatti, A., Fiore, F., Mignoli, M., Ciliegi, P., & Röttgering, H. J. A. (2005). XMM-Newton observations of Extremely Red Objects and the link with luminous, X-ray obscured quasars. *A&A*, 432, 69–81.

Brusa, M., Zamorani, G., Comastri, A., Hasinger, G., Cappelluti, N., Civano, F., Finoguenov, A., Mainieri, V., Salvato, M., Vignali, C., Elvis, M., Fiore, F., Gilli, R., Impey, C. D., Lilly, S. J., Mignoli, M., Silverman, J., Trump, J., Urry, C. M., Bender, R., Capak, P., Huchra, J. P., Kneib, J. P., Koekemoer, A., Leauthaud, A., Lehmann, I., Massey, R., Matute, I., McCarthy, P. J., McCracken, H. J., Rhodes, J., Scoville, N. Z., Taniguchi, Y., & Thompson, D. (2007). The XMM-Newton Wide-Field Survey in the COSMOS Field. III. Optical Identification and Multiwavelength Properties of a Large Sample of X-Ray-Selected Sources. *ApJS*, 172, 353–367.

Budavári, T. & Szalay, A. S. (2008). Probabilistic Cross-Identification of Astronomical Sources. *ApJ*, 679, 301–309.

Sutherland, W. & Saunders, W. (1992). On the likelihood ratio for source identification. *MNRAS*, 259, 413–420.