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 combinatorically possible matches.
3. Consideration of partial matches across catalogues, i.e. the absence of counterparts in some catalogues.
4. Taking into account the positional uncertainty.
5. Computation of a probability for each possible match.
6. Computation of a probability that there is no match.
7. Incorporating magnitude, color or other information about the sources of interest, refining the match probabilities.

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.

The code is the results of many discussion among colleagues and friends. We thank in particular: Tamás Budavári, Sotiria Fotopoulou, Fabrizia Guglielmetti, Arne Rau, Tom Dwelly, Andrea Merloni and Kirpal Nandra.

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 Best practice matching

To achieve reliable results, we recommend that you run matching with increasing amount of information (distance-based first, §2.5, then adding priors on source properties §2.6) and understand how each influences the results.

Typically, the goal is a compilation of "best matches", i.e. choosing one reliable counterpart for each source. Whatever method used, there are always false selection fractions and false non-selection fractions in play, which should be characterized. To this end, we recommend to shift the source catalogues by a distance much larger than the positional errors to simulate the results for chance alignment. This fake catalogue should not coincide with original positions (tool `nway-create-fake-catalogue.py` may help). For the matching run with this fake catalogue, use NWAY with the same settings and from the output, choose cut-off limits (`p_any`) that correspond to the desired false selection fraction (tool `nway-calibrate-cutoff.py` may help). The output of your first NWAY run advises how to use these tools to characterise false selection rates and to find an appropriate `p_any` threshold.

Then the NWAY output on the real data can be truncated based on this criterion (`p_any>cutoff`), and made into a best match catalogue (`match_flag==1`). It may be worth noting ambiguous cases with multiple solutions and to store these secondary solutions with similar probabilities in another catalogue (`match_flag==2`).

2.5 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.

2.5.1 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**. These are the same files as in Appendix B of Salvato et al (in prep), extracted from Sanders et al. (2007)/McCracken et al. (2007), Ilbert et al. (2010) and Brusa et al. (2010) respectively.

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 providing the coordinates in **degrees**. 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.

5. Otherwise the file can have arbitrary columns which are copied over to the output file.

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.

2.5.2 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 fixed error of 0.1 arcseconds.

Run this command in the doc/ folder:

```
python ../nway.py COSMOS_XMM.fits :pos_err
COSMOS_OPTICAL.fits 0.1 --out=example1.fits
--radius 15 --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 7.3 arcsec, so we adopt a cropping radius of 15 arcsec to speed up the matching (`--radius 15`). A larger radius produces a more complete catalogue. For dense catalogues larger radii can be much slower to compute, as the number of combinations to consider rises exponentially.

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 15.000000 arcsec radius
   matching: 1007283192 naive possibilities
   matching: hashing
       using RA  columns: RA, RA
       using DEC columns: DEC, DEC
   matching: healpix hashing on pixel resolution ~ 18.036304 arcsec (nside=8192)
   100% | 562333|##############################################|Time: 0:00:13
   matching: collecting from 61787 buckets, creating cartesian products ...
   100%|61787|##############################################|Time: 0:00:02
   matching: 462267 unique matches from cartesian product. sorting ...
   merging in 10 columns from input catalogues ...
   100% 10|##############################################|Time: 0:00:00
       adding angular separation columns
   matching:  22435 matches after filtering by search radius
   ```

   Within 20 seconds it created a cross-match of remotely possible associations (1,007,283,192 in principle, 22,435 within 15 arcseconds).

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

4. ```
   Computing distance-based probabilities ...
     finding position error columns ...
       Position error for "XMM": found column XMM_pos_err: Values are [0.109000..7
       Position error for "OPT": using fixed value 0.100000
     finding position columns ...
     building primary_id index ...
     computing probabilities ...
         correcting for unrelated associations ... not necessary

   Computing final probabilities ...
       grouping by column "XMM_ID" and flagging ...
   100%|  1797|####################################|Time: 0:00:00
   ```

   It computed the probability of each association.

5. ```
   creating output FITS file ...
       writing "example1.fits" (37836 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 on the sky 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 six important new columns along with all columns of the input catalogues:

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

2. `dist_post`: Distance posterior probability comparing this association vs. no association, as in Budavári & Szalay (2008).

3. `p_single`: Same as `dist_post` unless additional information was added, see Section 2.6.

4. **`p_any`**: 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. Our testing suggest that the **threshold for a secure catalogue depends on the application.** Section 2.4 explains how to calibrate a threshold.

High `p_any` values by themselves do not necessarily mean that the counterpart is ruled out. It can also mean that there is not enough evidence/information to declare it a counterpart.

5. **p_i**: For each possible association for each entry in the primary catalogue (e.g. A), the relative probability is computed. Our testing suggest that secure, pure catalogue **should keep only associations where p_i>=0.1. Secondary solutions down to 0.1 may be interesting. These thresholds may depend on the application – please report what your testing gives.**

6. **match_flag**: The most probable match is indicated with 1 for each primary catalogue entry. Secondary, almost as good solutions are marked with 2. By default, the maximum allowed ratio is at most 0.5, but the user can modify this threshold via the --acceptable-prob parameter. 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. To filter out low-probability associations (low p_i) from the output catalogue, the --min-prob parameter can be used.

2.5.3 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 from the p_i column, this is a ambiguous case, where more than one optical counterpart is possible.

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

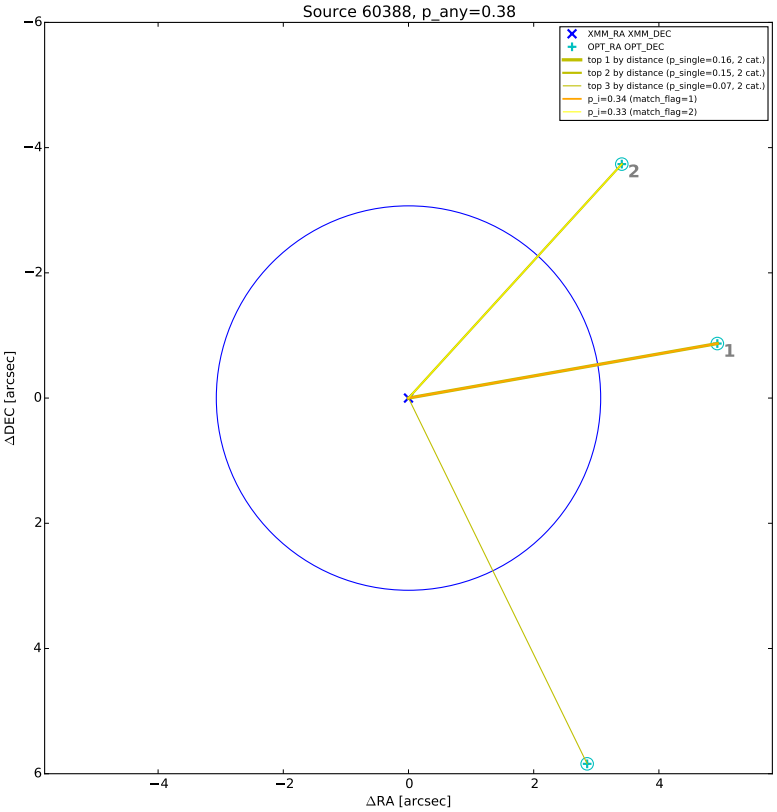
Two sources are at a similar distance from the X-ray source (blue, with error circle). Therefore their association probability (p_i) is similar. The slightly higher one is marked as match_flag=1 (orange), the other with 2 (yellow).

Section 2.6 solves this by adding more information (the magnitude distribution). But we can also solve this another way. We know AGN (the X-ray source) emit in the infrared, so you can also match with an IRAC catalogue.

Make a three-way match like so:

```
python ../nway.py COSMOS_XMM.fits :pos_err
COSMOS_OPTICAL.fits 0.1 COSMOS_IRAC.fits 0.5
--out=example3.fits --radius 15
```

However, overall we should note that p_any is low, indicating that probably neither of the two candidates is the counterpart.


Source 60388, p_any=0.38
XMM_RA XMM_DEC
OPT_RA OPT_DEC
top 1 by distance (p_single=0.16, 2 cat.)
top 2 by distance (p_single=0.15, 2 cat.)
top 3 by distance (p_single=0.07, 2 cat.)
p_i=0.34 (match_flag=1)
p_i=0.33 (match_flag=2)
2
1
ΔDEC [arcsec]
ΔRA [arcsec]
−6
−4
−2
0
2
4
6

2.6 Matching with additional information

For many classes of sources, the Spectral Energy Distribution (SED) provides additional hints, which associations are likely real. For instance, bright X-ray sources have a different color distribution in the WISE bands than non-X-ray emitting objects. 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.

2.6.1 Example - Using magnitude information

X-ray sources (which we are looking for in our example) have a different optical magnitude distribution than non-X-ray emitting 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 15 --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`

We use the column `MAG` from the catalogue `OPT` (FITS table name), therefore `--mag OPT:MAG`. After this follows that the magnitude prior histogram should be generated from the data (mode `auto`), by comparing the `MAG` histogram of sources within 3.5 arcsec of a X-ray source (`--mag-radius`) to that of full histogram.

(example continued below)

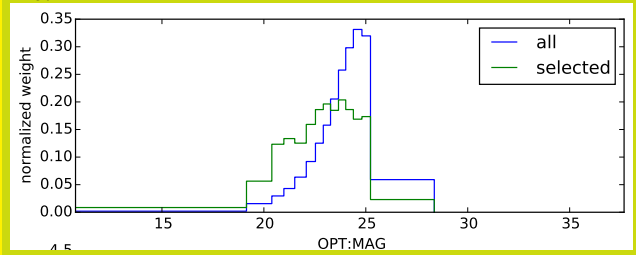
There are three possible ways to specify the prior in NWAY: In all cases you specify `--mag column-name [filename|auto]`. You can use `--mag` several times.

1. "File-mode": 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 (more details below on page 16).

2. "Simple auto-mode": Specifying `auto` instead of a file name derives the two distributions from the data, as we did in our example: All sources inside 3.5 arcseconds (`--mag-radius` parameter) of a X-ray source are put into one histogram, and all others into another histogram.

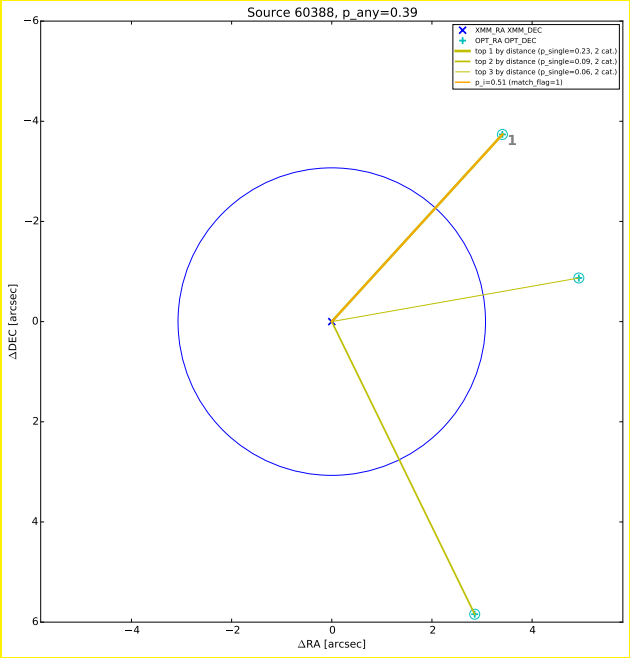
3. "Bayesian auto-mode": Bayesian distance probabilities (`dist_post`) will be used if you leave out `--mag-radius`. This is in general safer and recommended. In small catalogues the histogram may not be sufficiently filled, in which case NWAY will give a warning (more details below on page 17).

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 X-ray sources. This will help our matching.

As an example, we show below the ambiguous case from before. The upper 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 20 --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 demonstrate the use of a WISE magnitude of X-ray sources. If such prior information comes from previous studies, the distributions can be passed to NWAY as a ASCII table histogram. This table contains the histogram of the sources of interest (X-ray sources) and a histogram of other sources (non-X-ray 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 (# indicates comments):

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

Keep in mind that a prior created from a different data set can only be used if it is applicable to the present data set. For example, in the introduction of the paper (Salvato et al., in prep) we stress that a prior from a comparable X-ray exposure depth must be used when deriving color distributions.

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. For example, in Salvato et al. (in prep), the counterparts to ROSAT sources where found using WISE. The prior was build by using the color-magnitude (W1-W2 vs W2) properties of ~3000 secure counterparts to the 3XMM-Bright survey cut at the depth reached by ROSAT.

Discovering a prior from distance matching If you set `--mag OPT:MAG auto` and do not set `--mag-radius`, NWAY uses the Bayesian distance matching for discovering the histogram of `OPT:MAG`, as follows:

1. Those with `dist_post>0.9` are considered safe matches and are used for the "selected" histogram.
2. Those with `dist_post<0.01` are considered safe non-matches and are used for the "others" histogram.
3. Entries of -99 are always ignored. It is usually better to assign -99 where the magnitude error is large, to get cleaner histograms.

This is in general more cautious, and recommended for large catalogues

However, if you only have a small catalogue you may build a poorly sampled histogram, potentially leading to biases. NWAY will warn you when only few sources were selected.

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 make the code run faster and use less memory 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 OPT:MAG 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 `p_i` 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, NWAY tries 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, NWAY tries to be a bit fuzzy and detect the columns named ID_something etc.
6. Positional error columns, if used, 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

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

We now extend the approach of Budavári & Szalay (2008), to allow matches where some catalogues do not participate in a match. Comparing A12 and A14 in Budavári & Szalay (2008), assuming that positions lie on the celestial sphere and adopting the expansions developed in their Appendix B, we can write down likelihoods. For a counterpart across k catalogues, we obtain:

$$P(D|H) = 2^{k-1} \frac{\prod \sigma_i^{-1}}{\sum \sigma_i^{-1}} \exp\left\{ -\frac{\sum_{i<j} \phi_{ij} \sigma_j^{-1} \sigma_i^{-1}}{\sum \sigma_i^{-1}} \right\} \tag{5.4}$$

The likelihood for the hypothesis where some catalogues do not participate in the association has the appropriate terms in the products and sums removed. Therefore, the likelihood is unity for the hypothesis that there is no counterpart in any of the catalogues.

In comparison to our method, the method of Budavári & Szalay (2008) only compares two hypotheses for a association: either all sources belong to the same object (H_1), or they are coincidentally aligned (H_0). In this computation each hypothesis test is run in isolation, and relative match probabilities for a given source are not considered. For completeness, we also compute the posterior of this simpler model comparison:

$$\frac{P(H_1|D)}{P(H_0|D)} \propto \frac{P(H_1)}{P(H_0)} \times \frac{P(D|H_1)}{P(D|H_0)} \tag{5.5}$$

$$B = \frac{P(D|H_1)}{P(D|H_0)} \tag{5.6}$$

$$P(H_1|D) = \left[1 + \frac{1 - P(H_1)}{B \cdot P(H_1)}\right]^{-1} \tag{5.7}$$

The output column `dist_bayesfactor` stores $\log B$, while the output column `dist_post` is the result of equation 5.7. The output column `p_single` gives `dist_post` but modified if any additional information is specified (see Section 5.2). As mentioned several times in the literature, the Budavári & Szalay (2008) approach does not include sources absent in some of the catalogues, while the formulae we develop below incorporate absent sources. This is similar in spirit to Pineau et al., 2016, although the statistical approach is different. We now go further and develop counterpart probabilities.

The first step in catalogue inference is whether the source has any counterpart ($p_{\rm any}$). The posterior probabilities $P(H|D)$ are computed using Bayes theorem (eq. 5.3) with the likelihood (eq. 5.4) and prior (eq. 5.2) appropriately adopted for the number of catalogues the particular association draws from. For each entry in the primary catalogue, the posteriors of all possible associations are normalised to unity, and $P(H_0|D)$, the

posterior probability of the no-counterpart hypothesis, i.e., no catalogue participates, computed. From this we compute:

$$p_{\text{any}} = 1 - P(H_0|D) / \sum_i P(H_i|D) \tag{5.8}$$

If p_{any} is low, this indicates that there is little evidence for any of the considered, combinatorically possible associations, except for the no-association case. The output column `p_any` is the result of equation 5.8.

If $p_{\text{any}} \approx 1$, there is strong evidence for at least one of the associations to another catalogue. To compute the relative posterior probabilities of the options, we re-normalize with the no-counterpart hypothesis, H_0, excluded:

$$p_i = P(H_i|D) / \sum_{i>0} P(H_i|D) \tag{5.9}$$

If a particular association has a high p_i, there is strong evidence that it is the true one, out of all present options. The output column `p_i` is the result of equation 5.9.

A "very secure" counterpart could be defined by the requirement $p_{any} > 95\%$ and $p_i > 95\%$, for example. However, it is useful to run simulations to understand the rate of false positives. Typically, much lower thresholds are acceptable.

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 (e.g. 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. X-ray sources) 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 likelihoods are combined, by simply considering two independent observations, namely one for the positions, D_ϕ, and one for the magnitudes D_m. The likelihood thus becomes

$$P(D|H) = P(D_\phi|H) \times P(D_m|H) \tag{5.11}$$

$$= P(D_\phi|H) \times \frac{\bar{q}(m)}{\bar{n}(m)}, \tag{5.12}$$

with $\bar{q}(m)$ and $\bar{n}(m)$ being the probability that a X-ray (target) source or a generic (field) source has magnitude m respectively. NWAYstores the modifying factor, $P(D_m|H)$, in `bias_*` output columns, one for each column giving a magnitude, color, or other distribution. This modifying factor is however renormalized so that $P(D_m|H) = \frac{\bar{q}(m)}{\bar{n}(m)} / \int \frac{\bar{q}(m')}{\bar{n}(m')} \bar{n}(m') dm'$, which makes $P(D|H) = P(D_\phi|H)$ when m is unknown. In that case, m is marginalised over its distribution in the general population, i.e. $\int P(D_m|H)\, \bar{n}(m')\, dm$. This has the benefit that when m is unknown, the modifying factor is unity and the probabilities remain unmodified.

For completeness, I mention the fully generalized 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:

$$P(D_m|H) = \prod \frac{\int_m \bar{q}(m)\, p(m|D_m)\, dm}{\int_m \bar{n}(m)\, p(m|D_m)\, dm} \tag{5.13}$$

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. However, such options are not yet implemented in NWAY. Instead, we recommend removing magnitude values with large uncertainties (setting them to -99).

5.3 Auto-calibration

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

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 and unnecessary 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 HEALPix bins (Górski et al., 2005). The bin width w is chosen so that any association of distance w are improbable and negligible in practice, i.e. much larger than the largest positional error. An object with coordinates ϕ, θ is placed in the bin corresponding to its coordinate, but also into its neighboring bins 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 $O(N_{\text{bins}} \cdot \prod_{i=1}^{k} \frac{N_i}{N_{bins}})$. 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 posteriors 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. These posteriors have to be modified ("correcting for unrelated associations"), to consider associations unrelated to primary catalogue sources (described in the paper, Salvato et al. in prep, in the appendix section "Computing all possible matches").

In the third step the magnitudes are considered, and the posteriors 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: X-ray source catalogue). The posteriors p_{any} and p_i are computed. For the output catalogue a cut on the posterior probability (e.g. above 80%) can be applied, and all associations with their posterior probability are written to the output fits catalogue file.

Bibliography

Brusa, M., Civano, F., Comastri, A., Miyaji, T., Salvato, M., Zamorani, G., Cappelluti, N., Fiore, F., Hasinger, G., Mainieri, V., Merloni, A., Bongiorno, A., Capak, P., Elvis, M., Gilli, R., Hao, H., Jahnke, K., Koekemoer, A. M., Ilbert, O., Le Floc'h, E., Lusso, E., Mignoli, M., Schinnerer, E., Silverman, J. D., Treister, E., Trump, J. D., Vignali, C., Zamojski, M., Aldcroft, T., Aussel, H., Bardelli, S., Bolzonella, M., Cappi, A., Caputi, K., Contini, T., Finoguenov, A., Fruscione, A., Garilli, B., Impey, C. D., Iovino, A., Iwasawa, K., Kampczyk, P., Kartaltepe, J., Kneib, J. P., Knobel, C., Kovac, K., Lamareille, F., Leborgne, J.-F., Le Brun, V., Le Fevre, O., Lilly, S. J., Maier, C., McCracken, H. J., Pello, R., Peng, Y.-J., Perez-Montero, E., de Ravel, L., Sanders, D., Scodeggio, M., Scoville, N. Z., Tanaka, M., Taniguchi, Y., Tasca, L., de la Torre, S., Tresse, L., Vergani, D., & Zucca, E. (2010). The XMM-Newton Wide-field Survey in the Cosmos Field (XMM-COSMOS): Demography and Multiwavelength Properties of Obscured and Unobscured Luminous Active Galactic Nuclei. *ApJ*, 716, 348–369.

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.

Górski, K. M., Hivon, E., Banday, A. J., Wandelt, B. D., Hansen, F. K., Reinecke, M., & Bartelmann, M. (2005). HEALPix: A Framework for High-Resolution Discretization and Fast Analysis of Data Distributed on the Sphere. *ApJ*, 622, 759–771.

Ilbert, O., Salvato, M., Le Floc'h, E., Aussel, H., Capak, P., McCracken, H. J., Mobasher, B., Kartaltepe, J., Scoville, N., Sanders, D. B., Arnouts, S., Bundy, K., Cassata, P., Kneib, J.-P., Koekemoer, A., Le Fèvre, O., Lilly, S., Surace, J., Taniguchi, Y., Tasca, L., Thompson, D., Tresse, L., Zamojski, M., Zamorani, G., & Zucca, E. (2010). Galaxy Stellar Mass Assembly Between $0.2 < z < 2$ from the S-COSMOS Survey. *ApJ*, 709, 644–663.

McCracken, H. J., Peacock, J. A., Guzzo, L., Capak, P., Porciani, C., Scoville, N., Aussel, H., Finoguenov, A., James, J. B., Kitzbichler, M. G., Koekemoer, A., Leauthaud, A., Le Fèvre, O., Massey, R., Mellier, Y., Mobasher, B., Norberg, P., Rhodes, J., Sanders, D. B., Sasaki, S. S., Taniguchi, Y., Thompson, D. J., White, S. D. M., & El-Zant, A. (2007). The Angular Correlations of Galaxies in the COSMOS Field. *ApJS*, 172, 314–319.

Pineau, F.-X., Derriere, S., Motch, C., Carrera, F. J., Genova, F., Michel, L., Mingo, B., Mints, A., Nebot Gómez-Morán, A., Rosen, S. R., & Ruiz Camuñas, A. (2016). Probabilistic multi-catalogue positional cross-match. *ArXiv e-prints*.

Sanders, D. B., Salvato, M., Aussel, H., Ilbert, O., Scoville, N., Surace, J. A., Frayer, D. T., Sheth, K., Helou, G., Brooke, T., Bhattacharya, B., Yan, L., Kartaltepe, J. S., Barnes, J. E., Blain, A. W., Calzetti, D., Capak, P., Carilli, C., Carollo, C. M., Comastri, A., Daddi, E., Ellis, R. S., Elvis, M., Fall, S. M., Franceschini, A., Giavalisco, M., Hasinger, G., Impey, C., Koekemoer, A., Le Fèvre, O., Lilly, S., Liu, M. C., McCracken, H. J., Mobasher, B., Renzini, A., Rich, M., Schinnerer, E., Shopbell, P. L., Taniguchi, Y., Thompson, D. J., Urry, C. M., & Williams, J. P. (2007). S-COSMOS: The Spitzer Legacy Survey of the Hubble Space Telescope ACS 2 $\deg^2$ COSMOS Field I: Survey Strategy and First Analysis. *ApJS*, 172, 86–98.

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