Cross-correlation and automated classification: methods and tools

F.-X. Pineau[1]

[1] CDS, Observatoire Astronomique de Strasbourg

Treasures Hidden in High Energy Catalogues
IRAP, Toulouse, 23rd May, 2018

Cross-match tools

Cross-match tools and web-services (not exhaustive)

- Standalone tools
  - **TOPCAT** / STILTS: powerful general purpose tool, no probabilic cross-matches; Open Source (GPL), Java
  - **NWAY**: probabilistic cross-matches, able to account for photometry; Open Source, Python
  - C3, catsHTM: Python
- Web Services
  - SQL based: CasJobs (SDSS, Galex, ...), TAP (IVOA standard) services, SkyQuery, ...
  - Not SQL: **CDS Cross-match service** (asynchronous)
  - HTTP API: CDS Cross-match service (TOPCAT / STILTS, wget/curl, astroquery)
  - **ARCHES tool**: HTTP API + dedicated language, complexe cross-matches, probabilities

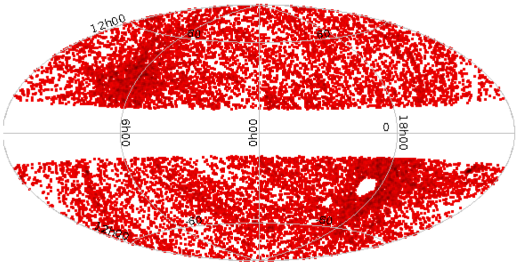
Main focus

- Roughly reproducing Salvato et al. (2018) results (J/MNRAS/473/4937/xmmslew2) and comparing them with the ARCHES tool + CDS classification (prototype) service
  - Goal: are two independent tools with similar methods provide coherent results?

NWAY / ARCHES: input data

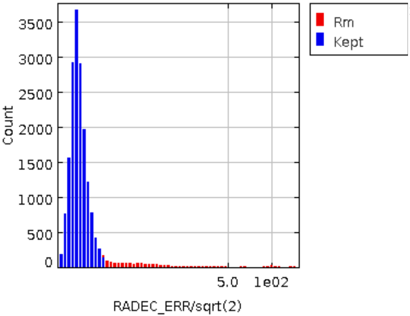
Input data provide by Mara Salvato:

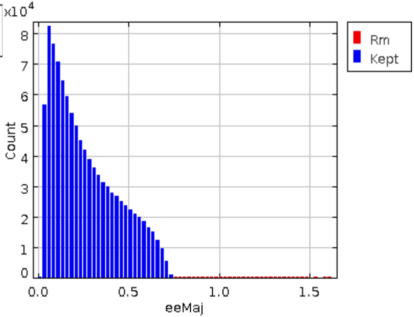
- XMM Slew survey Release 2: 17 672 / 29 393 sources
  - $|b| > 15°$, no SMC, noLMC
- 2’ extraction in AllWISE: 1 009 830 sources
  - Made using the CDS Xmatch service through TOPCAT?
- Surface area of both datasets



NWAY / ARCHES: input data

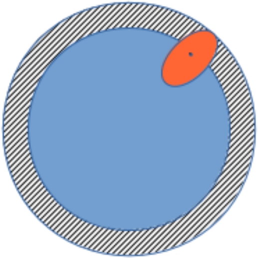
- Remove large positional errors:
  - Less noise in the normalised distance histograms
  - Finer prior estimation in the ARCHES tool
  - XMM: 958 sources removed (5%)
  - AllWISE: 834 sources removed ($<$ 0.1%)





NWAY / ARCHES: input params

- Adjust common surface area to account for border effects in priors computation
  - Not negligible for numerous 2 arcmin cones
  - $n_{spur} \propto \rho_X \rho_{IR} \frac{\Omega_{common} \Omega_{common-}}{\Omega_{common}} \chi_{\mathrm{ell}}$
  - χ-ellipse must be in the common surface area
  - $\Rightarrow$ for a cone search area, the center of the χ-ellipse must be in a smaller cone



- Legend of the figure:
  - Hatched area: Ω_{common}
  - Uniformly filled area: $\Omega_{common-}$
  - Orange: χ association ellipse

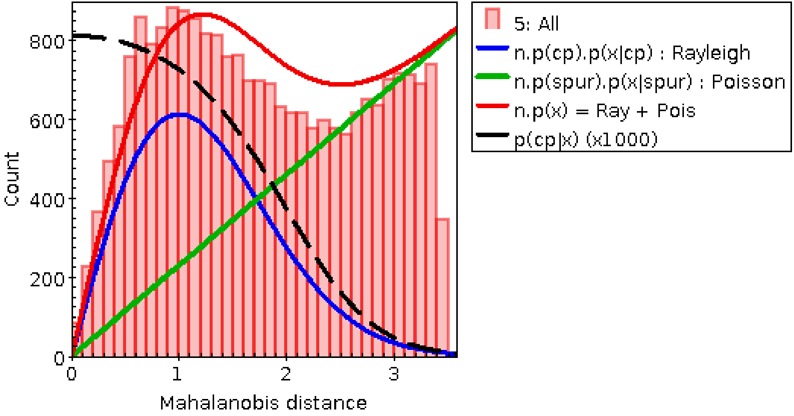
NWAY / ARCHES: input script

- Write and ARCHES cross-match script

```
# Load and set the XMM data to be cross-matched
get FileLoader file=XMMSL2_exgal_fewcol_2017APR12.fits
where RADEC_ERR < 10.0
set pos ra=RA dec=DEC
set poserr type=CIRCLE param1=RADEC_ERR/sqrt(2)
set cols *
prefix x
# Load and set the AllWISE data to be cross-match
get FileLoader file=candidate_ALLWISE_counterparts_unique_2017APR12.fits.gz
where eeMaj < 0.75
set pos ra=RA dec=DEC
set poserr type=ELLIPSE param1=eeMaj param2=eeMin param3=eePA
set cols *
prefix w
# Perform the cross-match, add the angular distance and save the result
xmatch probaN_v1 joins=I completeness=0.9973 area_w=0.0185176929488340157
               area_x=0.018517692948834015 area_xw=0.01388
merge dist mec
save xmmslew2_vs_allwise.fits fits
```

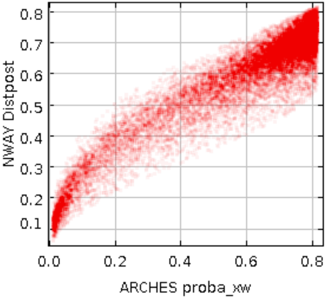
NWAY / ARCHES: result

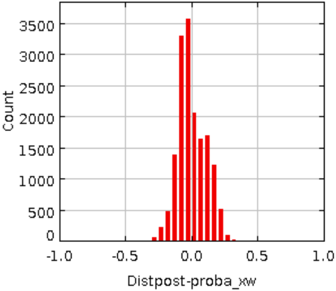
- Cross-match result: 23 813 associations
- Number of spurious matches ($n.p(spur)$) overestimated (?)



NWAY / ARCHES: position only

- NWAY vs ARCHES purely positional probabilities
  - scatter ($\sigma = 0.1$): priors, bi-normal ($\mathrm{d}x\mathrm{d}y$) vs Rayleigh ($2\pi r\mathrm{d}r$) (Eq. 150 vs Eq. 149 in Pineau et al. 2017).





NWAY / ARCHES: photometry

- Use photometric information to help separating good/spurious matches
- Purely positional cross-match

$$p(real|x) = \frac{p(real)p(x|real)}{p(real)p(x|real) + p(spur)p(x|spur)}$$

  - x: Mahalanobis distance; $p(x|real)$: Rayleigh; $p(x|spur)$: Poisson.
- Adding photometric likelihoods

$$p(real|x,\vec{m}) = \frac{p(real)p(x|real)p(\vec{m}|real)}{p(real)p(x|real)p(\vec{m}|real) + p(spur)p(x|spur)p(\vec{m}|spur)}$$

  - $\vec{m}$: position of the match in a photometric parameter space

NWAY / ARCHES: photometry

- Purely photometric probabilities

$$p(real|\vec{m}) = \frac{p(real)p(\vec{m}|real)}{p(real)p(\vec{m}|real) + p(spur)p(\vec{m}|spur)}$$

- $\sim$ few supervised automated classification methods
  - Linear Discriminant Analysis (LDA)
  - Kernel Density Classification (see Richards et al. 2004)
    - Stars/QSO photometric separation
- Goal of the classification: separate good and spurious matches

Automated classification

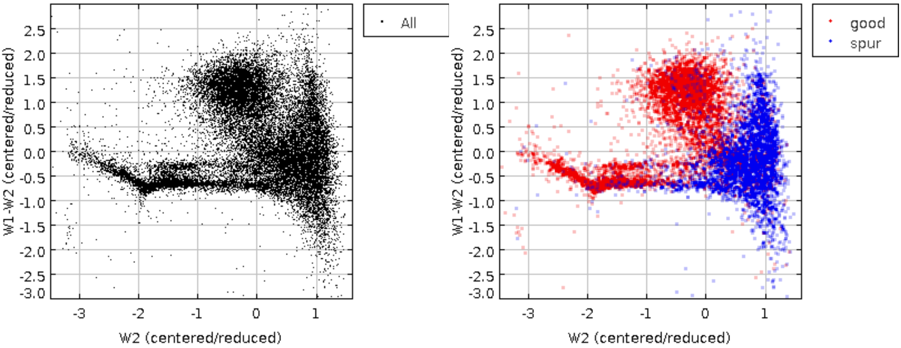
- Supervised / unsupervised (or clustering)
- Full/Reduced set of parameters (curse of dimensionality in k-NN like approaches)
- Supervised methods:
  - decision trees: OC1, Random forest, ...;
  - neural networks: SOM, LVQ, MLP, ...;
  - SVM
  - Bayes based: k-NN, KDC, LDA, ...
- Tools: R, Python (scikit -learn), ...
- Recurrent problem: tune input parameters to avoid over-fitting/under-fitting the LS

Automated classification

- Most important than the algorithm
  - Separability of classes in the parameter space
  - Learning samples quality / representativity
- Choosing a classif algo:
  - Easy to understand and to interpret
  - Naturally provide probabilities
  - Fast, easily reproducible (no random aspects)
- Personal choice: Kernel Density Classification

NWAY / ARCHES: photometry

- From Salvato et al. (2018): W2 vs W1-W2
- Learning samples arbitrary defined:
  - Good: d<6" && RADEC_ERR<8 && proba_xw>0.75
  - Spurious: d>13" && proba_xw<0.05



Classification service

- 3 CSV files: all matches, good matches, spurious matches
- Each file contains: $id, w2, w1 - w2$
- Using the CDS prototype service:

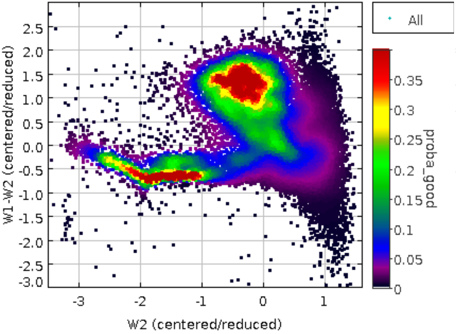
```
# Put the data files into the distant server
./classif.bash put good slew_vs_allwise.good.csv # 4565 rows
./classif.bash put spur slew_vs_allwise.spur.csv # 3082 rows
./classif.bash put data slew_vs_allwise.all.csv  # 23799 rows
# Performs the classification of the data and save the result
./classif.bash kdc samplepoint -k 75 -p good:0.425\;spur:0.575 -ho > result.csv
# Ask for the confusion matrix by self-classifying the LS
./classif.bash kdc samplepoint -k 75 -p good:0.425\;spur:0.575 -cr
```

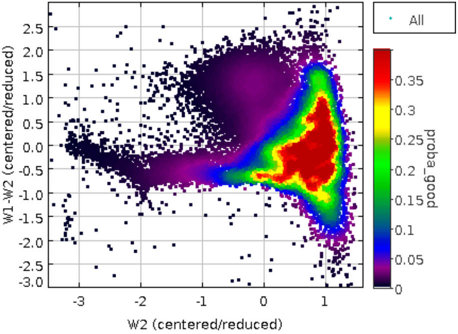
- Confusion matrix:

actual \predicted	good	spurious
good	85.96%	14.04%
spurious	9.73%	90.27%

NWAY / ARCHES: photometry

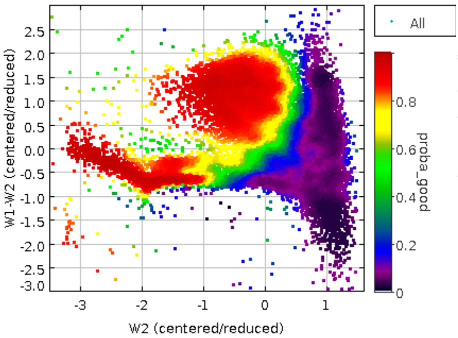
- Likelihoods (distributions) computed by kernel smoothing (sample point estimator, k=75)
  - left: $p(\vec{m}|good)$
  - right: $p(\vec{m}|spur)$

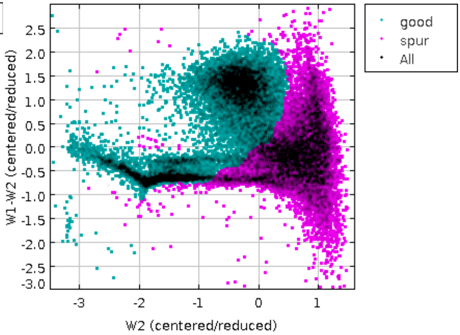




NWAY / ARCHES: photometry

- Left: classification result $p(good)$
- Right: binary classification Good/Spurious ($p(good) > 0.5$, $p(good) < 0.5$)



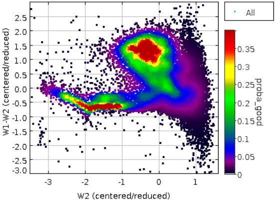


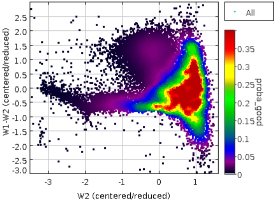
Merging with positional proba

- Accounting for photometric likelihoods (simplified Eq. 154 of Pineau et al. 2017)

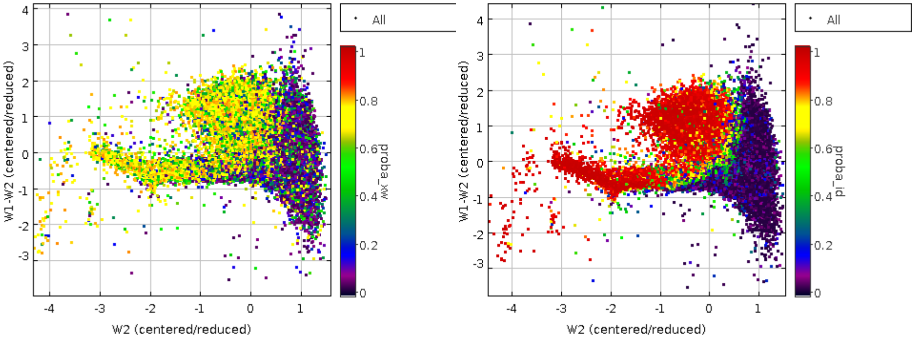
$$p(real|x,\vec{m}) = \frac{p(real|x)p(\vec{m}|real)}{p(real|x)p(\vec{m}|real) + (1 - p(real|x))p(\vec{m}|spur)}$$

$p(real|x) =$ purely positional probability

$p(\vec{m}|real) =$ 

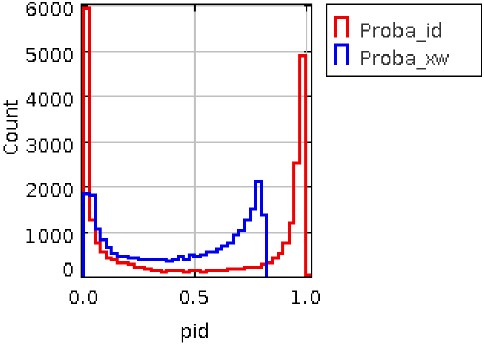
$p(\vec{m}|spur) =$ 

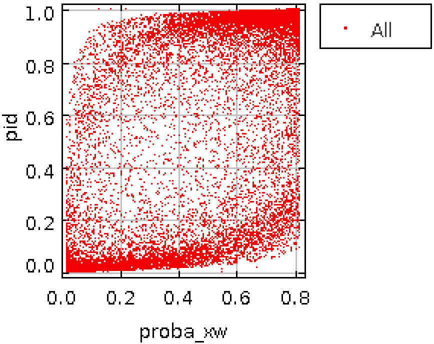
NWAY / ARCHES: photometry



- Left: positional probabilities
- Right: probabilities accounting for photometric likelihoods

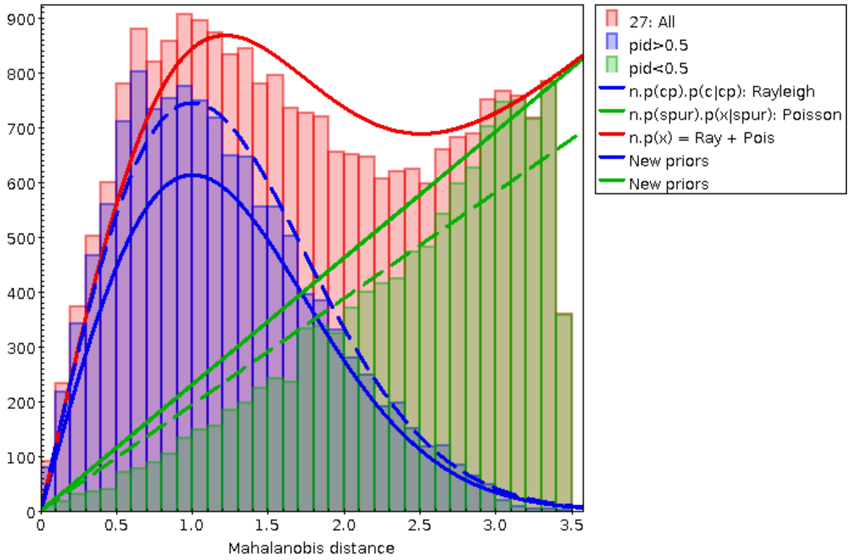
NWAY / ARCHES: proba id





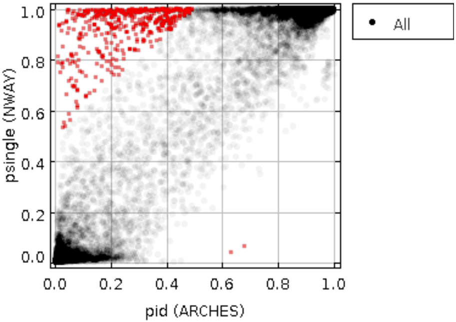
- Clearer separation of low and high probabilities

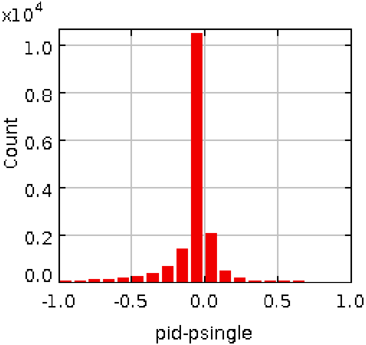
NWAY / ARCHES: proba id



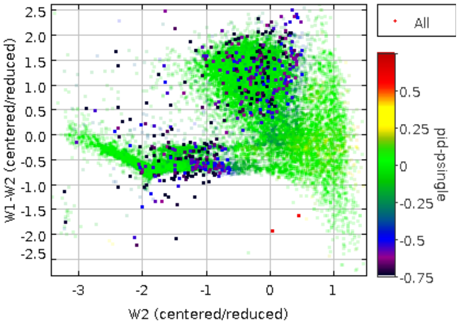
NWAY / ARCHES

Comparing NWAY and ARCHES outputs

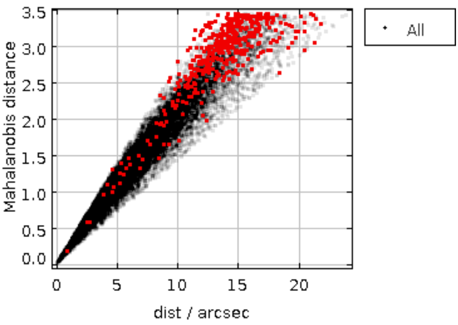




NWAY / ARCHES



All
W1-W2 (centered/reduced)
W2 (centered/reduced)
pid-psingle
2.5
2.0
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
-2.0
-2.5
-3
-2
-1
0
1
0.5
0.25
0
-0.25
-0.5
-0.75

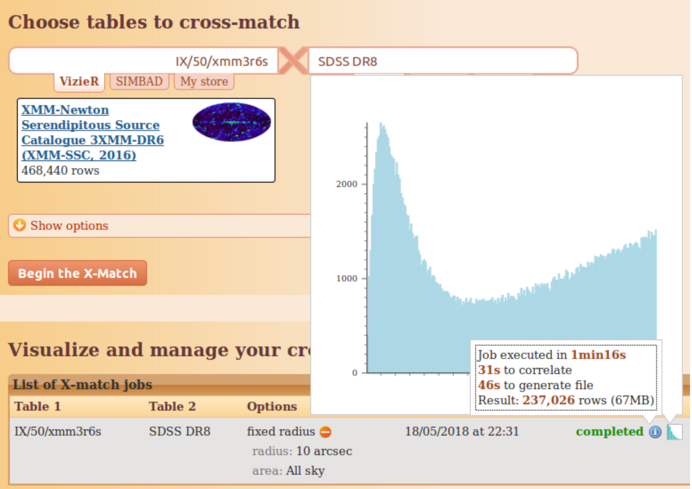


All
Mahalanobis distance
dist / arcsec
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
0
5
10
15
20

XMM vs SDSS DR8

Testing the same method to cross-match XMM and SDSS

XMM vs SDSS DR8



http://cdsxmatch.u-strasbg.fr

XMM vs SDSS DR8

- Quick-and-dirty test in 4 dimensions
  - Simple 10 arcsec cross-match
  - Keep only *primary unresolved* objects having a *clean photometry*
  - Mahalanobis distance: $d_\sigma \approx \frac{d}{\sqrt{\left(\frac{SC\_POSERR}{\sqrt{2}}\right)^2 + RA\_ERR \times DE\_ERR}}$
  - Lazy learning samples definition:
    - 19 676 "real" associations: $d < 1''$ && $d_\sigma < 1.5$
    - 7 784 "spurious" associations: $d > 8''$ && $d_\sigma > 6$
  - User defined prior $p(cp)$ going to $d_{\sigma,max} = 5$
  - From Eq. 149 of Pineau et al. (2017):

$$p(cp|d_\sigma) = \frac{1}{1 + \frac{1-p(cp)}{p(cp)} \frac{2}{d^2_{\sigma,max} e^{-\frac{d^2_\sigma}{2}}}}$$

XMM vs SDSS DR8

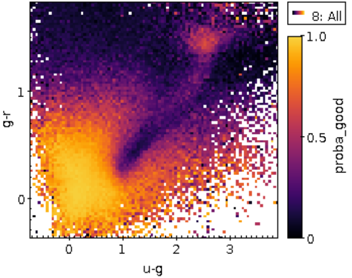
- Using the CDS prototype service:

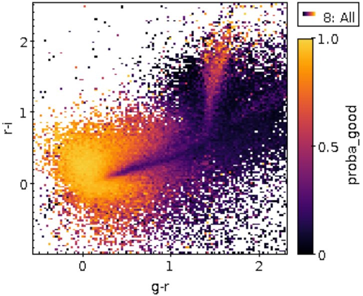
```
# Put the data files into the distant server
./classif.bash put good xmm_sdss8.unres.good.csv # 19676 rows
./classif.bash put spur xmm_sdss8.unres.spur.csv # 7784 rows
./classif.bash put data xmm_sdss8.unres.all.csv
# Performs the classification of the data and save the result
./classif.bash kdc samplepoint -k 25 -p good:0.55\;spur:0.45 -ho > result.csv
# Ask for the confusion matrix by self-classifying the LS
./classif.bash kdc samplepoint -k 25 -p good:0.55\;spur:0.45 -cr
```

- Confusion matrix:

actual \predicted	good	spurious
good	86.91%	13.09%
spurious	12.28%	87.72%

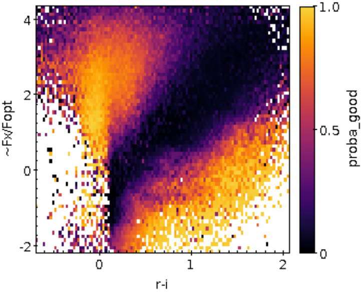
XMM vs SDSS DR8





Mean of the 4D KDC output probabilities in

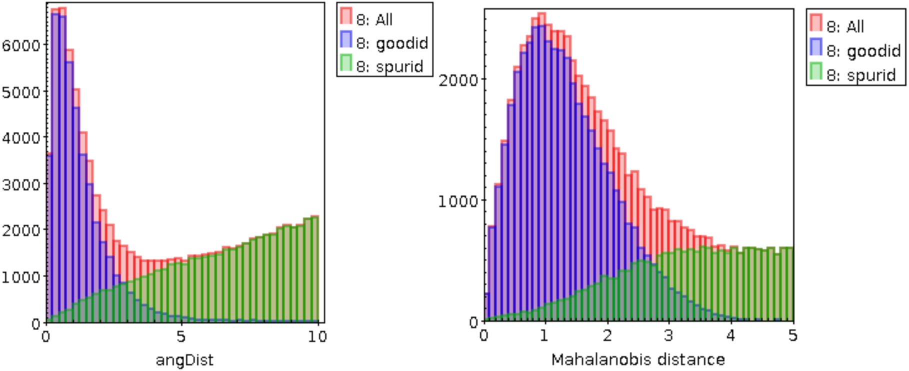
- $u - g$ vs $g - r$
- $g - r$ vs $r - i$
- $r - i$ vs $\propto F_X/F_r$



XMM DR7 vs SDSS DR8

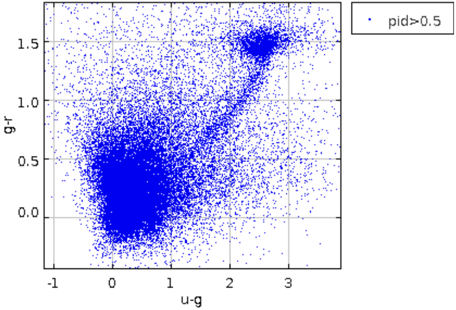
Using 4D photometric likelihoods to compute final proba $p(cp|d_\sigma, \vec{m})$, and considering:

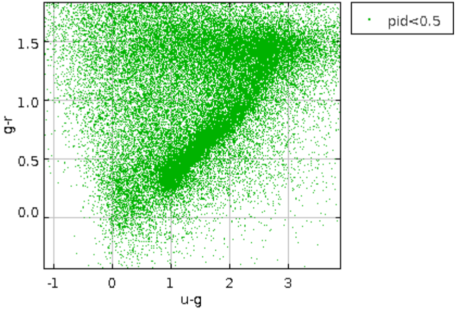
- real matches as $p(cp|d_\sigma, \vec{m}) > 0.5$
- spurious matches as $p(cp|d_\sigma, \vec{m}) < 0.5$



XMM DR7 vs SDSS DR8

- $u - g$ vs $g - r$ diagrams of estimated as real and estimated as spurious associations.





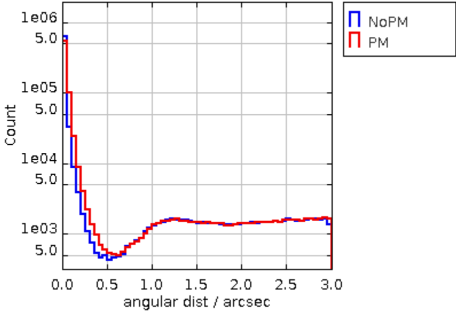
Outliers

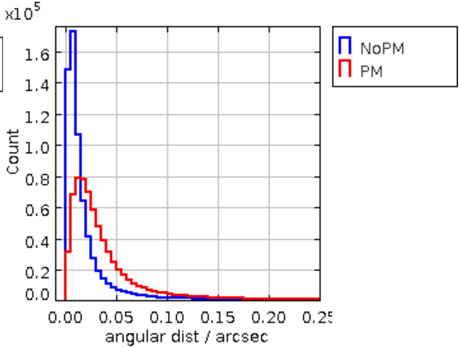
- Outliers automatic selection:
  - Select low likelihoods $p(\vec{m}|cp)$ and $p(\vec{m}|spur)$ and high $p(cp|x)$;
  - Check sources having a high $p(\vec{m}|spur)$ and a high $p(cp|x)$;
  - ...

Gaia DR2 vs PS1

- Checking PanSTARRS (STSCI / VizieR version) versus Gaia DR2 astrometric compatibility.
- ARCHES tool used to cross-match Gaia DR2 and PanSTARRS DR1 in 1400 XMM FOVs
- Simple 3 arcsec cross-match
  - Without taking into accounts PMs
  - Gaia DR2 positions computed at PS1 epoch

Gaia DR2 vs PS1

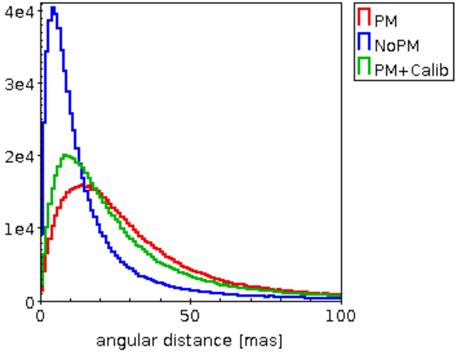


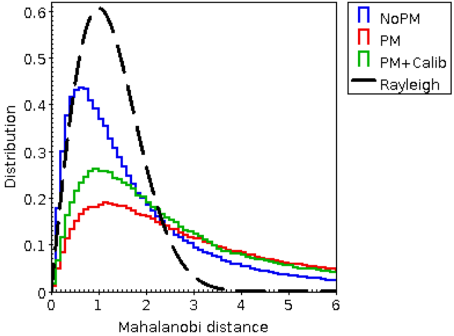


Results are better **NOT** taking into account Gaia DR2 PMs!!

Gaia DR2 vs PS1

Attempt to re-calibrate PS1 from Gaia DR2 positions (at PS1 epochs)



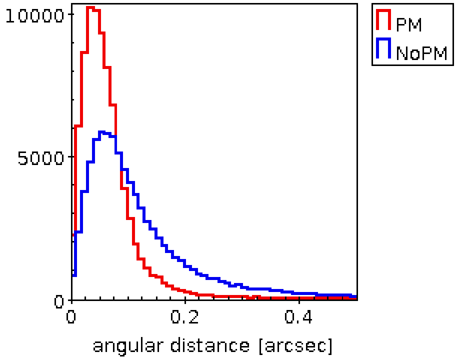


Improve results, but still not enough!!

- Rayleigh distribution assumption not satisfied!

Gaia DR2 vs SDSS DR12

- Gaia DR2 PMs do improve the cross-match with SDSS DR12



Complications with 3 cats

- 3 catalogues (X, S, W)
- 5 possibilities (5 priors, 5 likelihoods)
  - XSP: one actual source;
  - XS_P, XP_S, X_SP: 2 actual sources
  - X_S_P: 3 actual sources.
- Photometry: need to build 5 learning samples, perform a 5 classes classification
  - On-going tests with XMM-SDSS-AllWISE

Complication with outer joins

- I want X (XMM) sources χ-compatible with S (SDSS) **AND** P (PanSTARRS)
  - Can be done iteratively: $X \to XS \to XSP$
- I want X (XMM) sources χ-compatible with S (SDSS) **OR** P (PanSTARRS)
  - Can't be done iteratively!!
    - X χ-compatible with S
    - X χ-compatible with P
    - S (and XS) not χ-compatible with P
    - $\Rightarrow$ first step, XS, then nothing (XP missed!).
  - The ARCHES tool selects XS and XP, and remove them if XSP is also found.

Conclusion

- NWAY and ARCHES provides coherents results ($\sigma \approx 0.1$)...
  - ... but probabilities have to be used with care
- One can use Bayes based supervised classification techniques to compute photometric likelihoods, idependently from the positional part
  - dimensionality reduction
  - confusion matrix minimisation to compute the “best” KS bandwidth
- Complexity increase dramatically with the number of catalogues
- Do not forget the positional cross-match assumptions: Rayleigh, Poisson.
  - they are not so often satisfied!

The Kernel Density Classif

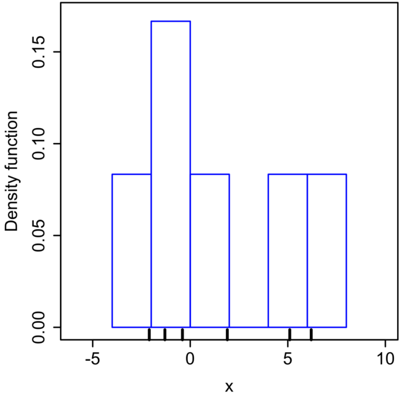
- Original paper: Richards et al (2004)
  - star/quasar (c_1/c_2) classification from $\vec{x}$ =(u-g, g-r, r-i, i-z)
- Supervised method: requires a learning sample for each class c_i
- Direct application of the Bayes' formula

$$p(c_i|\vec{x}) = \frac{p(c_i)p(\vec{x}|c_i)}{\sum_{j=1}^{n} p(c_j)p(\vec{x}|c_j)} \qquad (1)$$

  - c_i: object class
  - $\vec{x}$: vector in the parameter space
  - $p(c_i)$: user defined priors
    - iterate while priors $\neq$ posteriors means
  - $p(\vec{x}|c_i)$: likelihoods (p.d.f) computed by kernel smoothings (KS)
    - one KS by learning sample class

Histogramming vs KS in 1D

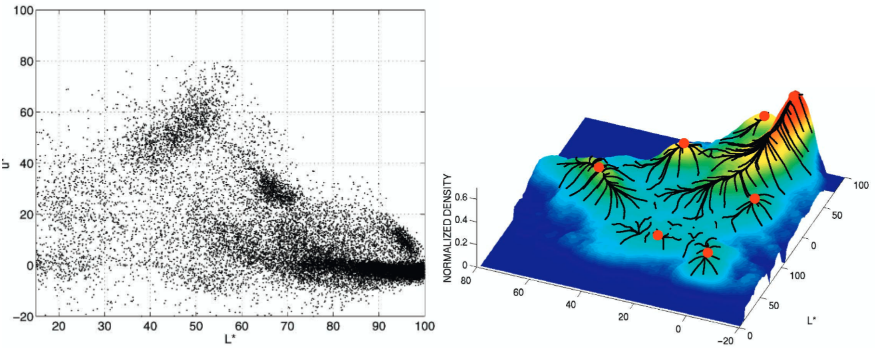
- KS: density = sum of kernels centered around each data point
- Normalised density = probability density function (p.d.f)



Credits: https://en.wikipedia.org/wiki/File:Comparison_of_1D_histogram_and_KDE.png

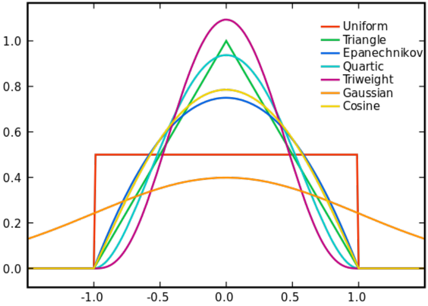
Kernel smoothing in 2D

- KS: density = sum of 2D kernels (e.g. 2D Gaussians) centered around each data point
- Normalised density = probability density function (p.d.f)



Credits: Comaniciu, D. and Meer, P. (1997)

Kernels



Credits: `https://en.wikipedia.org/wiki/File:Kernels.svg`

- We use only the multivariate **Epanechnikov** kernel
  - finite support (unlike Gaussian kernels)
  - theoretically the best (even if it is not that important)

Various Kernel Smoothing

- Fixed bandwidth: all kernels have the same bandwidth
- Variable/Adaptative bandwidth
  - balloon estimator:
    - 1 fixed bandwidth per density estimation
    - bandwidth = distance to the measurement point's k^{th}-NN
  - knn averaging: balloon estimator with a uniform kernel
  - sample-point estimator:
    - 1 bandwidth per data point in the LS
    - data point bandwidth = distance to the data point's k^{th}-NN