uc3m | Universidad Carlos III de Madrid

Master Degree in Information Health Engineering at UC3M
Academic Year 2019-2020

Master Thesis

Cross-Matching Catalogs in Astronomy and Problem Reformulation in ML

Óscar Manuel Jiménez Rama

Pablo Martínez Olmos
Ricardo Pérez Martínez
Madrid, September 2020

Cross-Matching Catalogs in Astronomy and Problem Reformulation in ML

Óscar Manuel Jiménez Rama *
Tutors:
Pablo Martinez Olmos
Ricardo Pérez Martínez (ISDEFE. S.A., S.M.E., M.P)

Department of Signal Processing and Communications, Universidad Carlos III de Madrid Leganés, 28911 Spain

Abstract

Cross-matching in astronomical catalogs is a first-step and common procedure to study and analyze a variety of issues regarding astrophysics. It consists in the identifications of the same celestial object in multiple catalogs that are captured by different instruments. Multiple methodologies were conceived over the years to improve results in distinctive conditions. In this work we will run two different versions of a Likelihood Ratio Test (*LRT*) [3, 7] algorithm and use the results to extract insights in order to reformulate the problem under a ML frame by means of a mixture model.

1 Introduction

The application of machine learning in the astronomical field is yet immature or not existing, specially regarding the current studied problem: x-matching catalogs. This work, driven and financed by ISDEFE in partnership with the European Space Agency (ESA) (the latter being client to the former), is intended to apply ML techniques as a basis for a preliminary basis in this area, bringing fairly satisfying results. This document portrays the research process carried out by the combined work of ISDEFE and the University Carlos III of Madrid.

Astronomical catalogs are large databases storing the position and intensity magnitude of celestial objects covering a given sky region. Depending on satellite's payload, they can capture electromagnetic excitation produced by stars and other sources, at a given spectral band. In practice, these instruments are limited to a small range of wave lengths, and usually it is needed more than one satellite to capture the excitation of such sky region at wide spectral bands (e.g. infrared, visual, X-ray, etc).

The measures' precision is highly influenced by the instrument's specifications. However, noise is not just a matter of the camera's quality. Images captured are just a projection of the 3D world onto a plane. Thus, any point-measure contained in the image is the sum of all radiation emitted along that specific direction, coming from the background or foreground. In such way, a galaxy laying behind a star of interest, in its direction seen from Earth, could be distorting its flux magnitude measure, completely modifying the star's real spectral properties.

This work is structured as follows. In the first part (Section 3) a Likelihood Ratio Test (*LRT*) is used in order to produce identifications from an optical and infrared catalogs. *LRT* was firstly introduced in [5] and, in the following decades, expanded with [6], [1] and most recently reviewed by [4]. Nowadays is considered a popular methodology for x-matching catalogs. More specifically, two modalities of *LRT* will be implemented based upon [3] (only-magnitude) and [7] (color and magnitude). Best results are obtained with the last, where the addition of color into the likelihood ratio improves the discriminatory behavior of such algorithm. An in-depth study of all parameters and factors influencing the technique is also carried out.

*Corresponding author. Email address: ojimenez@pa.uc3m.es

The second and last section of this document (Section 4) is dedicated to produce experiments over the data to study hidden statistical structure that could shed light and serve as a basis for new ML methodologies trying to solve the counterparts identification problem (x-matching). Results from the previous part will be used in order to compare similarities upon the extracted insights. Statistical grouping is obtained utilizing an unsupervised Gaussian Mixture Model (GMM). The goal is to identify differentiated structures within data, allowing to characterize two populations to be splitted: true vs false identifications.

Results show moderate evidence that needs to be further addressed using more complex techniques and partly limited by data. Nonetheless, a starting point is provided for further research lines.

Finally, conclusions and further work are exposed (Section 5).

2 Data & Preprocessing

In the present work the data used comes from three catalogs covering the same sky region in distinct frequency bands. The goal is to identify, with high confidence, the greatest amount of points simultaneously present in all catalogs.

- **Optical catalog (OPT)**:
  - Number of objects: 19.670
  - Number of bands: 6 from $\lambda \in [450, 1250]nm$ (B, V, R, I, J, K)
- **Mid infrared catalog (MIPS)**:
  - Number of objects: 1.038
  - Number of bands: 1 at $\lambda = 24\mu m$
- **Far infrared catalog (PACS)**:
  - Number of objects: 480
  - Number of bands: 2 from $\lambda \in [100, 160]\mu m$

As a preprocessing step, a position correction was carried out. The issue raised when it was found that only a small number of sources from **PACS** were found near (less that $2''$) from any object in **MIPS**. This fact proofed a positional offset between both catalogs. The position correction was done by minimizing, in an iterative manner, the mean distance of all close object between catalogs. For generalization purposes, the same correction was performed between **MIPS** and **OPT**.

3 Part 1: Likelihood Ratio Test as X-match algorithm

In this section the theoretical background of the *LRT* iterative X-matching algorithm will be developed. Both versions will be explained in detail and their results will be shown for two pairs of catalogs within each *LRT* modality.

Source matching will be performed using a reference and a secondary catalog. In the present document, the secondary catalog will always be set to **OPT** since it holds the greatest number of sources and numerous magnitudes measurements in several bands. Therefore, for each *LRT* version, the algorithm will match **MIPS** and **PACS** (reference) against **OPT** (secondary).

The first modality of the *LRT* will only include brightness and positional information into the computation of the probability ratio, whereas the second version will additionally include the color [1].

3.1 LRT general formulation

When a source is identified in a different catalog, it is referred to as its counterpart. The input of the algorithm consists of both catalogs to be matched, and the output produced is a list of identifications or counterparts.

The iterative algorithm consists in computing the following ratio for each pair of candidates:

$$LR = \frac{q(x_1, x_2, \dots) f(r)}{n(x_1, x_2, \dots)} \tag{1}$$

In equation 1, $q(x_1, x_2, \dots)$ represents the probability distribution that a reference source (**MIPS** or **PACS**) has a counterpart described by parameters $x_1, x_2, \dots$, where such dimensions could hold magnitude, color or other properties that are extracted from sources within the secondary catalog (**OPT**). Denominator $n(x_1, x_2, \dots)$ is the surface density of background sources with parameters $x_1, x_2, \dots$ or, in other words, the probability that the candidate counterpart belongs to the background. Last, $f(r)$ is the angular separation density function and is assumed Gaussian:

$$f(r) = \frac{1}{2\pi\sigma^2} \exp\Big(-\frac{r^2}{2\sigma^2} \Big) \tag{2}$$

, where $\sigma = \sqrt{\sigma_{ref}^2 + \sigma_{sec}^2}$ stands for the standard deviation coming from the $1\ \sigma$ positional errors of the reference and secondary catalogs.

[1] Color is the magnitude difference between the source of interest and the candidate counterpart, i.e. $c = m_{\text{ref}} - m_{\text{sec}}$.

As mentioned at the beginning, variables $x_1, x_2, \ldots$ will differ from one *LRT* modality to other. In the first one, densities $q(m)$ and $n(m)$ will be univariate and solely dependent on the optical magnitudes. On the other hand, the second version will describe the same functions adding the color dimension so that $q(m, c)$ and $n(m, c)$ are now bivariate. $f(r)$ remains as written in equation 2.

For each iterative loop, empirical density functions will improve as a result of a maximization process applied under a given criteria. At each iteration, a set of predefined thresholds will mask the LR (eq. 1) computed using the upgraded estimations q and n to each potential pair of counterparts, producing a final list of identifications that will be used to improve the former densities in the next loop. The algorithm is run until convergence.

3.2 LRT methodology 1: magnitude

In the current version the LR is defined as follows:

$$LR = \frac{q(m)f(r)}{n(m)} \tag{3}$$

Functions q and n depend only on the magnitude intensity of the secondary catalog, **OPT** in our case.

3.2.1 Computation of $n(m)$

The surface density $n(m)$ is computed at the beginning and stays fixed throughout the procedure. The density function will be estimated using a Gaussian Kernel Density Estimator (KDE) using the sources' magnitudes from the secondary catalog (**OPT**) that are far away from every instance contained in the reference catalog (**MIPS** or **PACS**).

The minimum distance considered for a given source to be sufficiently distant depends on the sky area covered by the catalog. In this work, any observation at a distance greater than $5''$ and smaller that $30''$ is considered background.

The resulting estimator is a real function defined over m describing the distribution of magnitudes contained in the background.

3.2.2 Computation of $q(m)$

As opposed to $n(m)$, $q(m)$ changes at each iteration, converging to a more accurate description of the magnitude distribution of counterparts.

As an initial approximation, $q(m)$ is computed using auxiliary functions and other parameters since such distribution is not directly observable. First, a Gaussian KDE estimator is used to compute $\text{total}(m)$, which represents the empirical distribution of sources that are close to each point in the reference catalog, in our case less that $5''$. Then, $\text{real}(m)$ is computed as:

$$\text{real}(m) = \text{total}(m) - D_{\text{ref}}\pi r_0^2 n(m) \tag{4}$$

D_{ref} (sources/arcsec^2) is the sky density of the reference catalog, which is multiplied by the area defined by the circle centered at a given reference source with radio $r_0 = 5''$. Therefore, the empirical distribution of close objects is corrected by the surface density normalized by a constant, in order to respect dimensional analysis. Usually the negative term is smaller than the first one, producing a short adjustment to $\text{total}(m)$.

Once $\text{real}(m)$ is obtained, the initial estimation of $q(m)$ is:

$$q(m) = \frac{\text{real}(m)}{\sum_m \text{real}(m)} Q_0 \tag{5}$$

Due to catalog's limitations, it is only possible to detect a portion of true counterparts. Such portion is represented by Q_0, initially set to $Q_0 = N_1/N_{\text{ref}}$. [2], where N_1 is the sample size used to build $\text{total}(m)$. Q_0 will also be updated at each loop.

3.2.3 Likelihood ratio thresholds

At this point, all three distributions conforming the *LR* (eq. 3) are computed and ready to perform the first iteration. *LR* is obtained for all plausible candidates i.e. close objects from **OPT** that lie proximal ($< r_{\text{cand}} = 1''$) to all sources in the reference catalog. Since $LR(m, r)$ is a function of the candidate's magnitude (m) and its distance to its potential counterpart (r), it can be obtained with ease for each candidate.

By comparing the *LR* calculated with a predefined threshold (L_{th}), a new set of sources is obtained, representing the estimated counterparts for that given L_{th}. Note that different counterparts will be produced depending on the value of L_{th} and one source could have multiple counterparts. Moreover, the value of L_{th} that throws the optimal set of counterparts with high confidence, is initially unknown and needs to be found heuristically by defining a closed range of values.

In the situation when one reference source is estimated to have more than one identification, it is kept the one with highest LR_i.

[2] In [2], it is developed a procedure to estimate Q_0 in an unbiased manner. In this work we stick to a simple initialization as Q_0 rapidly converges to a stable value after a few iterations.

An expected question would be how do we choose the optimal threshold that produces the most reliable results. This is done by defining two measures that quantify the efficiency or quality at the output of the algorithm.

For each set of counterparts associated to a given L_{th}, both quality measures are calculated. The set that maximizes a criteria function depending on both parameters is retrieved for an upgraded expression of $q(m)$ in the subsequent iterations. The process is repeated until convergence.

The two quality measures are Reliability ($\mathbf{R}$, eq. 6) and Completeness ($\mathbf{C}$, eq. 7). $\mathbf{R}(L_{\text{th}})$ represents the fraction of accepted identifications that are correct and it can be thought of as a measure of confidence regarding the quality of the identifications. $\mathbf{C}(L_{\text{th}})$ qualifies the fraction of true identifications that are accepted. In this context, accepted identifications refers to the total number of candidates i.e. those sources close to each observation in the reference catalog, while real counterparts are considered when their computed LR_i is above the threshold.

$$\mathbf{R}(L_{\text{th}}) = 1 - \frac{1}{Q_0 N_{\text{ref}}} \sum_{LR_i \geq L_{\text{th}}} \frac{1 - Q_0}{Q_0 LR_i + (1 - Q_0)} \tag{6}$$

$$\mathbf{C}(L_{\text{th}}) = 1 - \frac{1}{Q_0 N_{\text{ref}}} \sum_{LR_i < L_{\text{th}}} \frac{Q_0 LR_i}{Q_0 LR_i + (1 - Q_0)} \tag{7}$$

Both parameters[3] are in a trade-off relationship, meaning that improving one of them often penalizes the other one. For instance, it is possible to achieve a maximum $\mathbf{C}$ by lowering the threshold (accepting the majority of candidates) at the cost of acquiring less confidence for the identifications i.e. small value of $\mathbf{R}$.

To take the above in account, the most common criteria to select L_{th} is that where $\mathbf{R}(L_{\text{th}})$ and $\mathbf{C}(L_{\text{th}})$ equate or, in other words, where both curves cross with each other. Nevertheless, criteria can be modified to achieve specific requirements.

Returning to our starting point, the update procedure for $q(m)$, once the final set of identifications is obtained for the current iteration, is reused in the consecutive loop to estimate $q(m)$ again with a KDE. In addition, the value of $\mathbf{C}$ associated to the previous set of identifications is used to update the value of Q_0.

3.2.4 Results X-match: MIPS vs OPT

By running the algorithm for each magnitude of **OPT**, the best results where obtained using the **K** band of the secondary catalog ($m = K$).

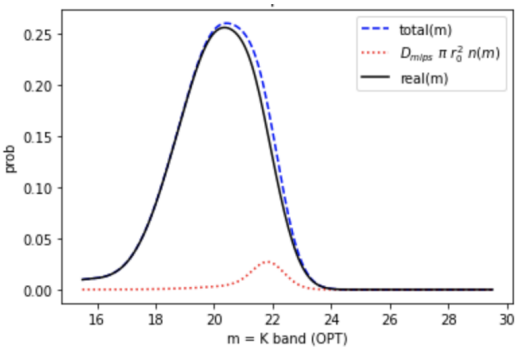

Figure 1: Empirical distributions. Blue dashed line represents total(m), red dotted line depicts the normalized $n(m)$ and the solid black line is the resulting real(m).

In Figure 1 it is shown the three auxiliary functions used to build $q(m)$. As explained, the background factor causes little influence in the computation of real(m).

Before presenting the results obtained, it will be explained the most relevant design decisions and their reasons.

The estimator used for Q_0 (N_1/N_{mips}) produced a small value (0.48) that caused numerical problems. Since Q_0 is updated every iterations and usually converges to a stable value, the initialized value was set to 0.8.

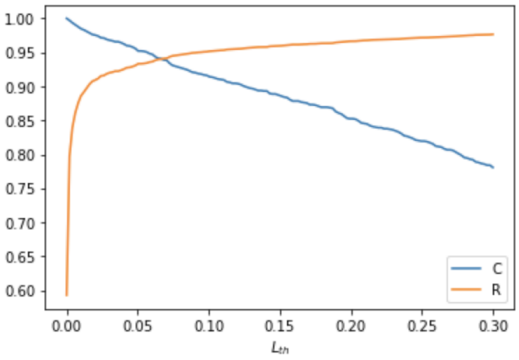

Figure 2: $\mathbf{R}$ and $\mathbf{C}$ as function of L_{th} in last iteration.

For our present purpose, $\mathbf{R}$ and $\mathbf{C}$ were given the same importance. The L_{th} selection criteria or function to be maximized at each run was the sum of both parameters, as opposed to the crossing point of both curves (Figure 3).

[3]The expressions for Completeness and Reliability are modified from the original definition exposed in [6], and the one used in [7] is used instead.

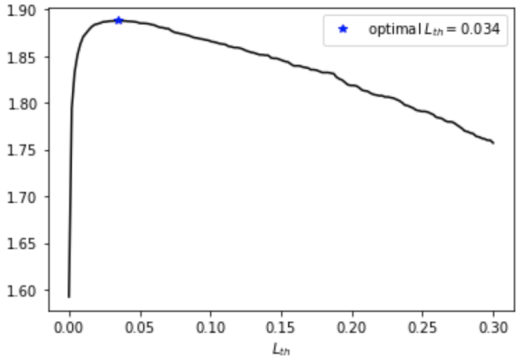

Figure 3: Selection criteria ($\mathbf{R} + \mathbf{C}$) and optimal value of L_{th} in last iteration.

In Figures 4 and 5 it can be seen that parameters $\mathbf{R}$ and $\mathbf{C}$ reach convergence at approximately iteration 10.

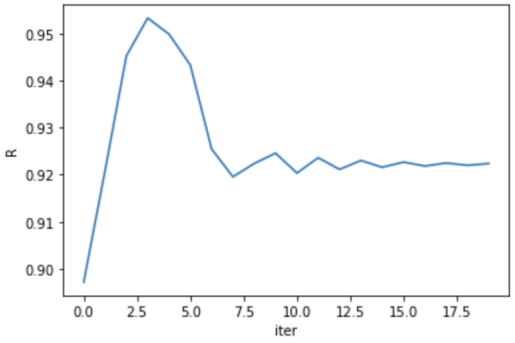

Figure 4: $\mathbf{R}$ evolution.

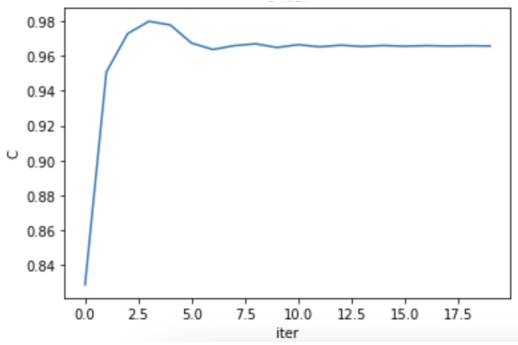

Figure 5: $\mathbf{C}$ evolution.

$iter_{\max}$ (1)	$\mathbf{R}$ (2)	$\mathbf{C}$(3)	L_{th}(4)
20	0.922	0.965	0.034
N_{match}(5)	N_{MIPS}(6)	-	-
630	1138	-	-

Table 1: Final values of the *LRT* (modality 1) for **MIPS** vs **OPT**. (1) Number of iterations. (2) Converged value of $\mathbf{R}$. (3) Converged value of $\mathbf{C}$. (4) Final threshold selected. (5) Number of **MIPS** sources with at least one identification. (6) Total sources in **MIPS**.

Table 1 summarizes the *LRT* results. With a value of 0.922 for $\mathbf{R}$ it can be said that the identifications are reliable. The completeness parameter also achieves a high score in spite of the fact that the number of identifications (column (5)) is slightly above half of the total sources contained in **MIPS** (column (6)). The last is due to the inherent limitation of catalogs described by Q_0, meaning that is actually impossible to identify the totality of sources.

3.2.5 Results X-match: PACS vs OPT

The results for **PACS** achieve higher scores than the previous x-match, mainly caused by the significant decrease of points. This is not always the case since less data implies more deficient estimates regarding the probability densities. Nevertheless, **PACS** contains a sufficiently large sample size to achieve acceptable results.

While in **MIPS** x-match, the best results were obtained using the optical band K, **PACS** obtains optimal results by means of J magnitude in **OPT**.

Moreover, and also influenced by the small catalog size, the initialization of Q_0 had to be adjusted at higher proportion to avoid numerical problems. The final value selected was $Q_0 = 0.9$.

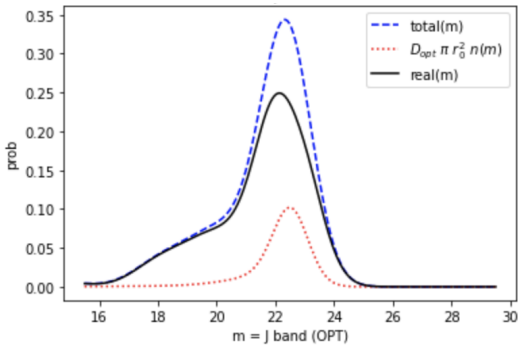

Figure 6: Empirical distributions. Blue dashed line represents total(m), red dotted line depicts the normalized $n(m)$ and the solid black line is the resulting real(m).

As depicted in Figure 6, total(m) is more affected by the (normalized) background density $n(m)$ in contrast with the previous results. This is due to the data sizes used to estimate such densities. In **MIPS**, the number of sources to construct total(m) and $n(m)$ was high w.r.t. the current scenario. Note that the lesser points fed into a **KDE**, the higher the peak (mode) is located within the density. Probability mass tends to concentrate in few points creating abrupt peaks.

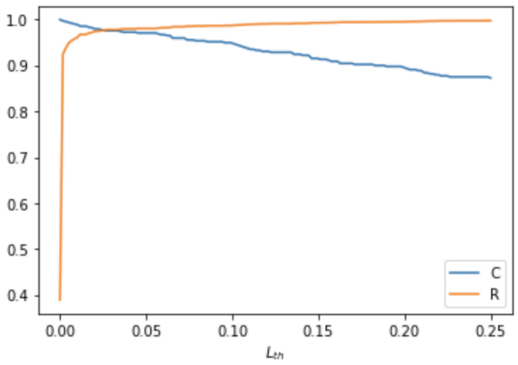

Figure 7: $\mathbf{R}$ and $\mathbf{C}$ as function of L_{th} in last iteration.

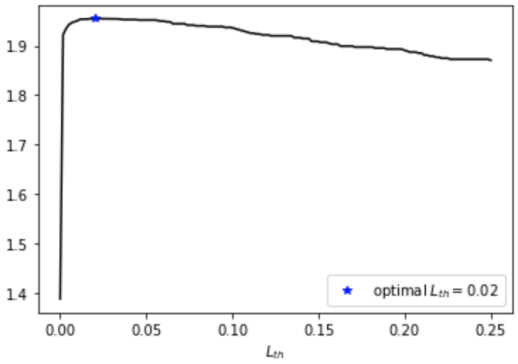

Figure 8: Selection criteria ($\mathbf{R} + \mathbf{C}$) and optimal value of L_{th} in last iteration.

The same L_{th} selection criteria ($\mathbf{R} + \mathbf{C}$) as in **MIPS** was used. Interestingly, as showed in Figures 7 and 8, the optimal threshold is obtained near the intersection of both curves despite the lack of an explicit implementation of such criteria. The reason behind could include many factors that are not straightforward and easy to observe.

Similar to the evolution over iterations of parameters $\mathbf{R}$ and $\mathbf{C}$ showed in the previous section, Figures 9 and 10 exposes the convergence around the 10^{th} iteration.

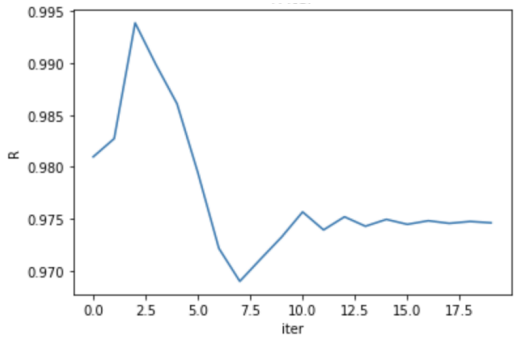

Figure 9: $\mathbf{R}$ evolution.

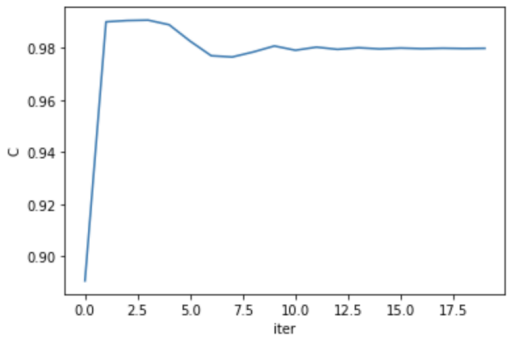

Figure 10: $\mathbf{C}$ evolution.

$iter_{\max}$ (1)	$\mathbf{R}$ (2)	$\mathbf{C}$(3)	L_{th}(4)
20	0.974	0.979	0.02
N_{match}(5)	N_{PACS}(6)	-	-
68	408	-	-

Table 2: Final values of the *LRT* (modality 1) for **PACS** vs **OPT**. (1) Number of iterations. (2) Converged value of $\mathbf{R}$. (3) Converged value of $\mathbf{C}$. (4) Final threshold selected. (5) Number of **PACS** sources with at least one identification. (6) Total sources in **PACS**.

Table 2 summarizes the final parameters. As explained, the reliability and completeness surpassed that obtained with MIPS. Apart from the sample size difference, **PACS** data could hold more statistical coherence when crossed against **OPT**. It is true that the proportion of identifications is significantly lower than in **MIPS** (16% vs 0.55%). That could have been the cost to increase the reliability of the counterparts identifications. This is in line with the actual characteristics of the observations. PACS data are subject to higher uncertainties in both position and brightness estimation, what is made worse by "blending" (i.e: different sources contributing differently to different data points), so it is somehow expected that the algorithm gives lower number of matches, although the R and C of each of them can be higher.

3.3 LRT methodology 2: magnitude and color

In this second modality color is introduced in the *LR* calculation (eq. 8) as a discriminative dimension. Therefore, q and n are bi-dimensional functions. Such change endows both curves with more complexity, allowing them to capture more differentiating characteristics that ultimately improves performance w.r.t the previous *LRT* modality.

$$LR = \frac{q(m,c)f(r)}{n(m,c)} \tag{8}$$

Technically, the color dimension is not continuous, it is a discrete variable with a predefined number of maximum possibilities that has to be adjusted to obtain best results.

In order to define the number of different values that c may have, the range $[c_{\min}, c_{\max}]$[4] is divided in equally spaced color bins, ranging from 2 in advance. Note that c refers to a single color while each bin represents a sub-range within the selected color.

For the sake of generalization, it is possible to consider an only-magnitude *LRT* by introducing a single color bin in such dimension. Nonetheless, the procedures differ since the computation of q varies from one methodology to the other.

3.3.1 Computation of $n(m, c)$

The computation of $n(m, c)$ follows the same procedure as in the only-magnitude *LRT* version, with the difference that a **KDE** is computed for each bin conforming the discrete domain of c.

As a result, each source will only contribute to a single color bin at a time when estimating $n(m, c = b)$, with $b = 2, 3, .., b_{\max}$.

3.3.2 Computation of $q(m, c)$

The addition of the c dimension complicates the procedure described in the only-magnitude *LRT* in order to compute $q(m, c)$. This can be seen when computing the fraction of counterparts with color c, $Q_0(c)$ which now is a biased estimator if $N_1(c)/N_{\text{ref}}$ is considered. Note that, in practice, there is no need to store the value of $Q_0(c)$ for each color bin as parameters $\mathbf{R}$ and $\mathbf{C}$ are computed by using $Q_0 = \sum_c Q_0(c)/N_{ref}$. As a workaround, near sources are first split by color and then, each color bin is also divided by magnitude to obtain the first iteration estimate of $q(m, c)$, again using a **KDE**.

The rest of the iterative algorithm is kept unchanged, including the expressions of $\mathbf{R}$ and $\mathbf{C}$ (eq. 6 & 7 respectively). In the subsequent iterations, $q(m, c)$ is updated by keeping the list of final identifications (from the previous iteration), which will be split by color and magnitude to estimate the new distribution of counterparts $q(m, c)$.

3.3.3 Results X-match: MIPS vs OPT

The improvement in performance obtained by introducing colors can be seen in the following results.

Since K magnitude (**OPT**) achieved best results in the only-magnitude *LRT* shown in 3.2.4, the same band was chosen in this experiment.

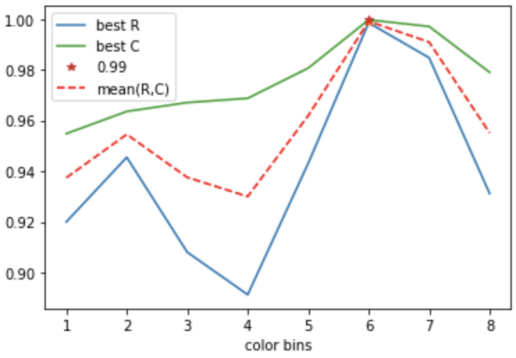

Figure 11: Final values of parameters $\mathbf{R}$ (blue), $\mathbf{C}$ (green) for each run with different color bin separation. The red dotted curve represents the criteria function (mean($\mathbf{R}$,$\mathbf{C}$) to be maximized.

By running the algorithm varying the maximum number of color bins, it was found that a division of 6 ranges in the color dimension (color bins) was the best choice (Figure 11).

The criteria selection for L_{th} at each iteration was changed to the mean between reliability and completeness parameters. It showed to produce a more stable and controlled behavior than in previous criteria, without causing numerical problems.

Regarding the initialization of Q_0, the default option described in [7] did not bear any numerical issue as occurred in only-magnitude *LRT*.

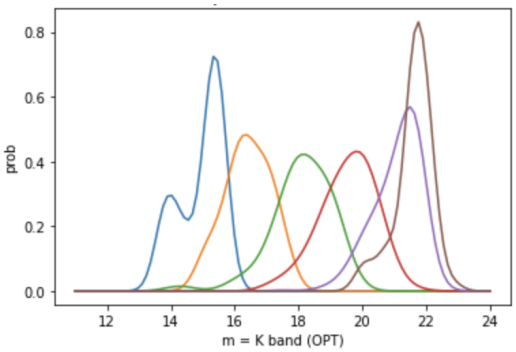

Figure 12: $q(m, c)$ for 6 color bins. Each curve represents $q(m, c = c^*)$. Color bins domain are defined in legend's ranges. Separation between color bins is ≈ 1.6.

Figure 12 along with Table 3 shows the estimated $q(m, c)$ for each color bin, their ranges and the number of sources used to construct them. Despite some $q(m, c = c^*)$ were estimated using a small number of points, 6 color

[4] $c_{\min} = \min(m_{ref}) - \max(m_{sec})$ and $c_{\max} = \max(m_{ref}) - \min(m_{sec})$

bins manifested the best discriminative behavior, most surely the reason behind the performance improvement. Curves although overlapped, are centered in different ranges, accumulating a big amount of the probability mass of each color bin in distinctive areas of the magnitude axis.

It can be seen that most of the sources are located in the green, red and purple color bins (according to the depicted legend in Table 3).

Color bin (1)	$Q_0(c)$ (2)
$-4.21 < K - m_{mips} \leq -2.6$	9
$-2.6 < K - m_{mips} \leq -0.99$	11
$-0.99 < K - m_{mips} \leq 0.61$	59
$0.61 < K - m_{mips} \leq 2.22$	331
$2.22 < K - m_{mips} \leq 3.82$	441
$3.82 < K - m_{mips} \leq 5.43$	61

Table 3: Color bins and their domain. (1) Upper and lower bound of each bin . (2) Number of samples to estimate each $q(m, c = c^*)$.

Color bin (1)	$N_0(c)$ (2)
$-4.21 < K - m_{mips} \leq -2.6$	61
$-2.6 < K - m_{mips} \leq -0.99$	265
$-0.99 < K - m_{mips} \leq 0.61$	2806
$0.61 < K - m_{mips} \leq 2.22$	33133
$2.22 < K - m_{mips} \leq 3.82$	7324
$3.82 < K - m_{mips} \leq 5.43$	22

Table 4: Color bins and their domain. (1) Upper and lower bound of each bin . (2) Number of samples to estimate each $n(m, c = c^*)$.

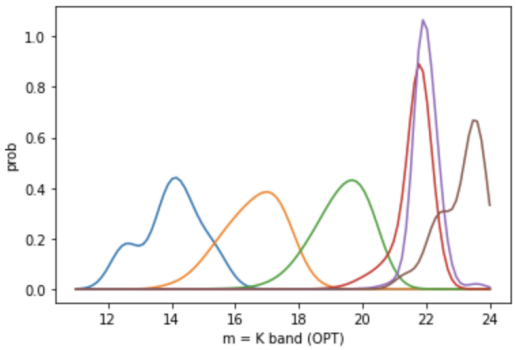

Figure 13: $n(m, c)$ for 6 color bins. Each curve represents $n(m, c = c^*)$.

In the same way as $q(m, c)$, Figure 13 and Table 4 depicts $n(m, c)$. Note that color bins are defined in the same domains.

There are three observations in contrast to the estimation of $q(m, c)$. First, much more sources were fed into the **KDE**, resulting into most reliable estimations. Second, each curve is not defined in the same areas as those in $q(m, c)$. Some appear to be slightly stretched or shifted and overlapping, in some cases, is less present.

Lastly, it is interesting that in spite of being the optimal color divisions 6, the background density does not seem to show the same optimal division as in $q(m, c)$. This is noticeable by looking at the similarity between the red and purple curve, almost trying to explain the same cluster of points. Figure 14 shows the usual behavior of parameter curves within a single iteration (2 in this case). An early iteration was selected for convenience.

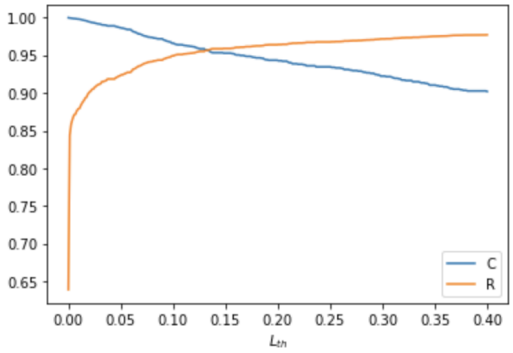

Figure 14: **R** and **C** as a function of L_{th} in iteration 2.

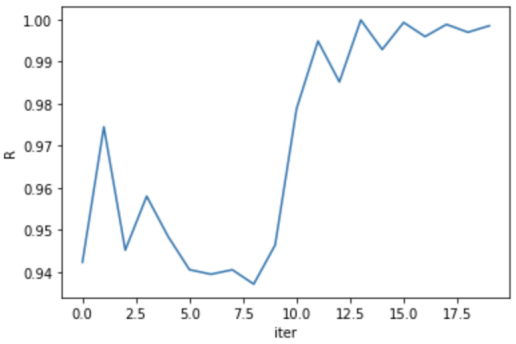

Figure 15: **R** evolution for 6 color bins.

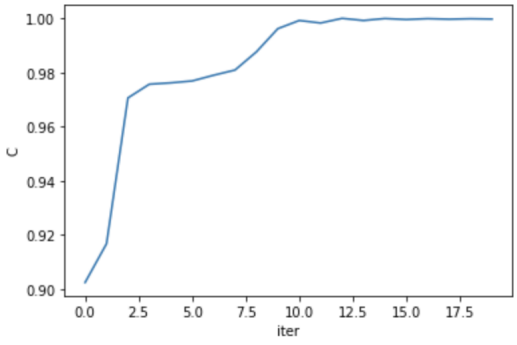

Figure 16: **C** evolution for 6 color bins.

Figures 15 and 16 show the convergence of parameters $\mathbf{R}$ and $\mathbf{C}$ respectively. The main observation in contrast is that convergence seems to be more spiky and late than in other examples.

$it_{\max}$ (1)	$\mathbf{R}$ (2)	$\mathbf{C}$ (3)	L_{th} (4)
20	0.998	0.999	0.001
N_{match} (5)	N_{MIPS} (6)	N_{same} (7)	-
855	1138	570	-

Table 5: Final values of the *LRT* (modality 2) for **MIPS** vs **OPT**. (1) Number of iterations. (2) Converged value of $\mathbf{R}$. (3) Converged value of $\mathbf{C}$. (4) Final threshold selected. (5) Number of **MIPS** sources with at least one identification. (6) Total sources in **MIPS**.(7) Number of sources identified to the same counterpart in both *LRT* modalities.

Table 5 summarizes results. With the same amount of iterations, nearly perfect values for reliability and completeness are achieved, as well as a significant increase in the amount of sources identified (column 5). Also, 570 sources share the same counterpart estimated using both **LRT** modalities (column 7). This exhibits the presence of coherence in the algorithm whether is the only-magnitude version or the color one.

3.3.4 Results X-match: PACS vs OPT

The present experiment ran on **PACS** shows improvement in the parameter scores ($\mathbf{R}$ and $\mathbf{C}$) although some aspects, both in implementation and data, differ from the only-magnitude *LRT* version.

Firstly the selection criteria was changed for each iteration. The best list of identification was that whose L_{th} defined the intersection between both curves [5]. This was the only criteria in lack of numerical issues.

In addition, **PACS** contains many missing values in both bands, reducing even more the total working set of sources fed to the algorithm. Obviously, the intensity band ($\lambda = 160\mu m$) holding more points was used. This is a crucial change that affects directly to the algorithm performance. The first *LRT* modality need not to take into account any magnitude stored in the reference catalog. The positions (along its errors) is the only input to the algorithm needed from **PACS**. In this modality, color computation is required thus **PACS** magnitudes. In other words, the reference catalog fed into both modalities of *LRT* is not the same. Therefore, fair comparison between both techniques under the same catalog is not possible.

Regarding the results and as shown in Figures 17 and 18, the optimal color division is 3 color bins. Also **J** band (**OPT**) for magnitude dimension was considered as its outcomes where optimal in comparison with the rest of the bands.

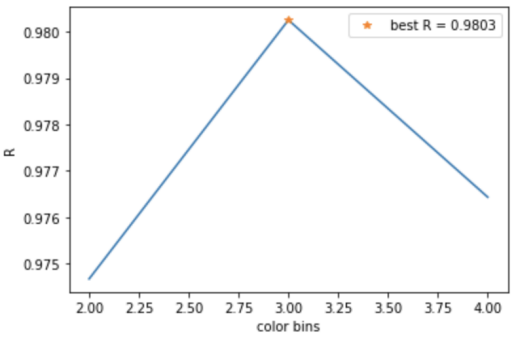

Figure 17: Final values of parameter $\mathbf{R}$ for each run with different color bin separation.

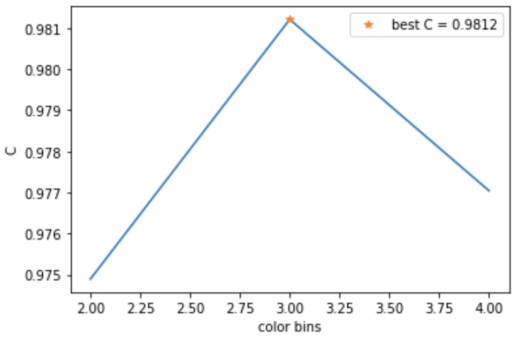

Figure 18: Final values of parameter $\mathbf{C}$ for each run with different color bin separation.

Note that the above plots are almost identical even if they represent different parameters. This is the resulting consequence of the criteria used for optimal L_{th} selection (crossing point of curves).

In the same way than the previous results, $q(m, c)$ is shown in Figure 19 in addition with Table 6. In contrast with **MIPS** results, each curve does not present the same amount of overlapping, mostly due to the reduction of color bins. Moreover, the vast majority of points seem to belong to the right-most curve (green $q(m, c = c^*)$).

[5] Due to (numerical) precision resolution, is not possible to retrieve the same value of $\mathbf{R}$ and $\mathbf{C}$ for the same L_{th}. Nevertheless, the numerical difference is negligible (orders of < 0.001).

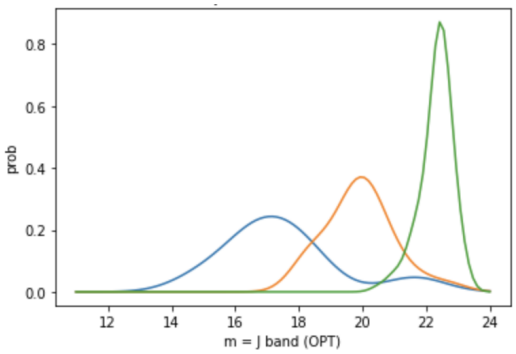

Figure 19: $q(m, c)$ for 3 color bins. Each curve represents $q(m, c = c^*)$.

Color bin (1)	$Q_0(c)$ (2)
$2.46 < J - m_{pacs} \leq 5.26$	9
$5.26 < J - m_{pacs} \leq 8.05$	61
$8.05 < J - m_{pacs} \leq 10.85$	271

Table 6: Color bins and their domain. (1) Upper and lower bound of each bin . (2) Number of samples to estimate each $q(m, c = c^*)$.

For the background density $n(m, c)$ (Figure 20 and Table 7) similar insights can be extracted from those described in reference to Figure 13 from **MIPS** experiment.

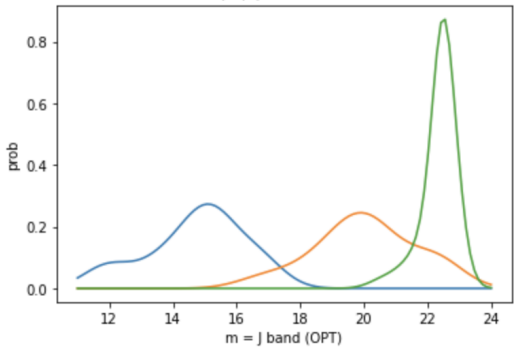

Figure 20: $n(m, c)$ for 3 color bins. Each curve represents $n(m, c = c^*)$.

Color bin (1)	$N_0(c)$ (2)
$2.46 < J - m_{pacs} \leq 5.26$	79
$5.26 < J - m_{pacs} \leq 8.05$	1050
$8.05 < J - m_{pacs} \leq 10.85$	14487

Table 7: Color bins and their domain. (1) Upper and lower bound of each bin . (2) Number of samples to estimate each $n(m, c = c^*)$.

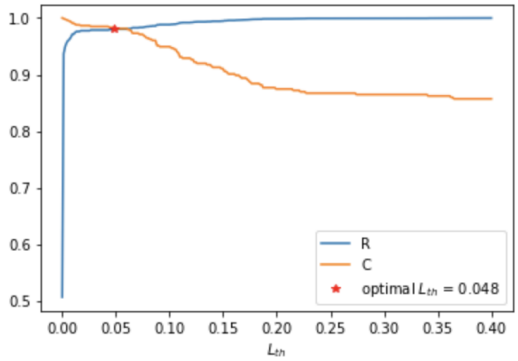

Figure 21: $\mathbf{R}$ and $\mathbf{C}$ as a function of L_{th} in iteration 29.

Figure 21 shows $\mathbf{R}$ and $\mathbf{C}$ as a function of L_{th} in the last iteration. It is clearly seen that the selection criteria is the intersection, outputting the optimal point.

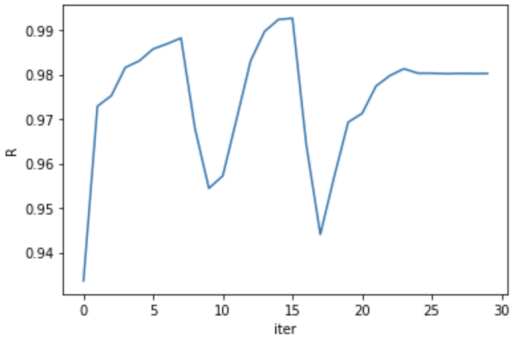

Figure 22: $\mathbf{R}$ evolution for 3 color bins.

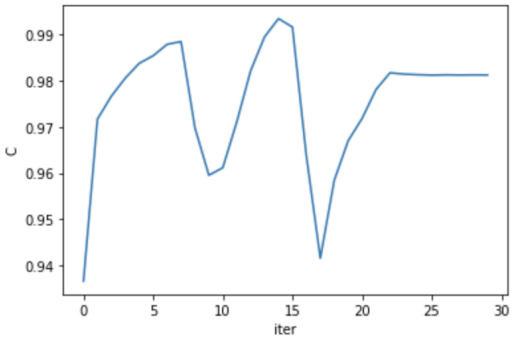

Figure 23: $\mathbf{C}$ evolution for 3 color bins.

Figures 22 and 23 print the evolution of both parameters at each iteration. The first observation is that both curves show almost the exact tendency (due to criteria) and that convergence is reached later in comparison with previous examples. In fact, the maximum number of iterations had to be increased in order to reach convergence. Additionally, the current example shows how the modification of the criteria influences also in the evolution of parameters producing, in this exam-

ple, more abrupt changes w.r.t. other parameters' evolutions already presented.

$it_{\max}$ (1)	$\mathbf{R}$ (2)	$\mathbf{C}$ (3)	L_{th} (4)
30	0.98	0.98	0.048
N_{match} (5)	N_{PACS} (6)	N_{same} (7)	-
59	383	4	-

Table 8: Final values of the *LRT* (modality 2) for **PACS** vs **OPT**. (1) Number of iterations. (2) Converged value of $\mathbf{R}$. (3) Converged value of $\mathbf{C}$. (4) Final threshold selected. (5) Number of **PACS** sources with at least one identification. (6) Total sources in **PACS**.(7) Number of sources identified to the same counterpart in both *LRT* modalities.

Finally, Table 8 stores the final data results. Again, reliability and completeness parameters improve in comparison to the only-magnitude version, however and as mentioned before, addressing differences is not possible due to distinctive scenarios mainly influenced by input data.

In proportion, the number of matches found in both versions w.r.t. the totality of points stays roughly identical, but identifications in the current *LRT* modality are more confident given the resulting values of $\mathbf{R}$ and $\mathbf{C}$.

Despite the acceptable performance, at least in a quantitatively sense, it is oddly negative that only 4 identifications are common in both techniques. This could be explained as a result of the elimination of sources that were fed into the only-magnitude version but absent in the color one.

Last insight precisely proofs a considerable downfall in this family of algorithms: there is not an exact procedure to firmly address the performance of algorithms since its 'unsupervised' nature prevents results to be objectively checked. Quality parameters are directly influenced by data structure and characteristics, not real facts as happen in supervised conditions.

4 Part 2: Problem Reformulation Under ML Frame

The second part of this document is presented as a proposal in the form of a new research line, being its goal the exploration of unique solutions framed into the ML field, in order to approach the studied x-matching problem. This is carried out and based upon the results obtained from the *LRT* algorithm, in addition to well known probabilistic techniques to address and quantify insights and structures that can be extracted from the available data.

A Gaussian mixture model (GMM) is used to obtain a cluster separation between (false or true) identifications, feeding it with pairs of points, each of them belonging to different catalogs. Analyzing similarities on the optimal color bins found in the *LRT* and the results from this section (i.e. clusters obtained), it might be possible to observe common data structures, allowing to deeply understand the x-matching problem from several points of view.

A primal goal is to proof the following hypothesis: the optimal color bin separation found on $q(m, c)$ in *LRT* is related to the cluster's structure found by the mixture model. The last is thoughtfully chosen as a probabilistic unsupervised method in order to satisfy the conditions required for an unbiased study.

4.1 Motivation

One of the main questions that intuitively raise when x-matching astronomical catalogs is if it exists any data structure that can be extracted, independently from both catalogs, similar enough that can provide a discriminatory effect for it to discard a great proportion of wrong identifications. If both catalogs embody a similar statistical distribution within clusters, it might correspond to observations present in both catalogs at the same time. However, there exist many factors that can distort such structures (e.g. number of total samples in both catalogs, among others), producing an increase in the problem's complexity if a parallel analysis is carried out on each catalog.

When looking at the distributions of $q(m, c)$ and $n(m, c)$ (Figures 20 and 19, for example), in spite of their statistical closeness, one can think that there exist subtle differences between identifications that are accepted as correct and those that are not. Detecting such disparities could be useful in order to apply a general discriminatory effect on the candidate's set.

By taking a probabilistic approach (GMM) it may be interesting to see how identifications are distributed among clusters. Utilizing the learned model and evaluating new pairs that are assumed not to be counterparts, intuitively one can expect an statistical difference regarding the assignment of true counterparts versus wrong identification among each cluster.

In order to materialize the above in practice, a set containing the appropriate and relevant information is required. The axiom from which this set is build relies on the fact that true counterparts are necessarily located near each other. Taking close surrounding points from both catalogs should manifest disparate statistical arrangement than those further away.

4.2 Implementation and considerations

Throughout this experiment, **MIPS** and **OPT** catalogs will be used. Therefore, the best results from the mentioned databases, obtained in the color-modality *LRT*, are retrieved for the current purpose.

Firstly, a new training set needs to be defined, which will be used to feed the mixture model. In addition, three more sets will be created to correctly address and detect statistical difference between true and false counterparts. All sets are defined in the same vector space and are standardize with respect the training set. Each entry, no matter the set contained into, does not represent a single point, it is a vector encoding information from pairs of sources, each corresponding to the reference and secondary catalog. Pairs are selected over single points on the grounds that the goal is not to look after structures within each catalog, it is to find that of counterparts across catalogs.

Color is a magnitude that is composed by the combination of sources from both catalogs, thus it fits the desired purpose.

Moreover, it is important to define the correct number of input data dimensions that will be fed into the model. In this scenario where noise is present and both populations (true vs false counterparts) are considerably overlapped, could be useful to consider an expansion regarding the input data dimensionality as much as possible. By the previous, it is easier for the model to extract complex data structures using more dimensions since spacial distance increases exponentially with dimensionality.

The final dimension selected is 7. The first component holds the K band magnitude (**OPT**) and the rest are filled with the colors $m_{opt} - m_{mips}$, where m_{opt} can take brightness belonging to 6 different bands (V, K, B, I, J and R).

Training is unsupervised since true counterparts are not identified. Therefore, the training set is obtained by computing the 7-dimension vector with pairs of sources proximal to each other ($< 2''$). This generated set does not necessarily contain true counterparts but it is expected that all (observable) true identifications are contained in itself.

The identifications thrown by the *LRT* will be considered as ground truth, although the model is not aware of them during training (unsupervised). In fact, color based *LRT* identifications are used to build the 'true' matches set (*LRT* matches set), encoding each pair as the defined 7-dimensional vector.

For comparison purposes it is required a set of points that combines true counterparts and identifications that are most surely not real. In order to produce such set, from now on denoted as test, the same procedure used for the training set will be used but retrieving pairs which are separated less than $6''$ instead of $2''$. By this means, the test set contains all the training elements plus those which lay in the range $2'' < r < 6''$, a subset of false identifications since their distance is high for them to be counterparts.

In order to correctly analyze the populations under study, the test set is divided a second time to produce a new one containing all the elements which are **not** simultaneously present in the *LRT* matches set, i.e. the no-matches set (NM set). Therefore, the total sample size of the test set should be the sum of the NM set and *LRT* matches set.

Set	Sample size	# features
Train	912	7
Test	2538	7
LRT matches	855	7
NM	1683	7

Table 9: Summary of datasets.

Once the four sets are defined (summary in table 9), normalization is carried out. As common practice, training data is transformed to have zero mean and unit variance. Accordingly, the rest of the sets are standardized w.r.t. the training set.

In GMM there is only one free parameter left for validation: number of clusters. In order to select the best model, the Bayesian information criterion (BIC) is used and is defined as:

$$BIC = k \ln(n) - 2 \ln(\hat{L}) \tag{9}$$

,where k is the number of parameters of the model (mean and covariance of each Gaussian component), n is the sample size and $\hat{L}$ is the maximized value of the likelihood of the model M, $p(x|\hat{\theta}, M)$ being $\hat{\theta}$ the parameters in the optimal point.

BIC penalizes the model when the number of parameters is increased to avoid overfitting in spite of greater likelihood values over the model. This criteria has limitations but for the current scenario it is a simple option that serves the purpose. Models with lower BIC are preferred.

Moreover, the covariance matrix type can be chosen to be independent from one component to another, equal, with equal diagonal or to have its own single variance. Fully independent covariance matrices for each cluster usually produces better quality models under the BIC criteria.

GMM learning is achieved by Expectation-maximization (EM) algorithm, thus there is not an unique solution. The model with lowest BIC out of several runs is selected.

Predicted labels of all points from all sets will be obtained using the final model. Ideally, points that are considered false matches would belong to different clusters than those that are considered counterparts.

Finally, to proof our initial hypothesis, the training points from each cluster will be used to feed a **KDE** to address the similarity between $q(m, c)$, obtained in section 3.3.3, and the learned clusters retrieved by the mixture model.

KDE was used in order to simplify the comparison. The learned parameters from each cluster live in a 7-dimensional space, while $q(m, c)$ is a set of curves mapped from a single dimension. Therefore, the intensity from the K band of points within clusters is used to 'project' densities into a single dimension to fairly compare affinities between results.

4.3 Results

Hereafter, two different result will be showed: the GMM model learned using 6 and 5 clusters, both using **MIPS** and **OPT** as matching catalogs.

# of clusters final model	BIC
5	4182.47
6	4297.78

Table 10: Final models and their BIC value.

Spite of the fact that the model with 5 clusters achieves the lowest BIC, it is interesting to compare models with the same number of grouping elements as seen in the color *LRT* modality regarding $q(m, c)$. The BIC value for 6 clusters

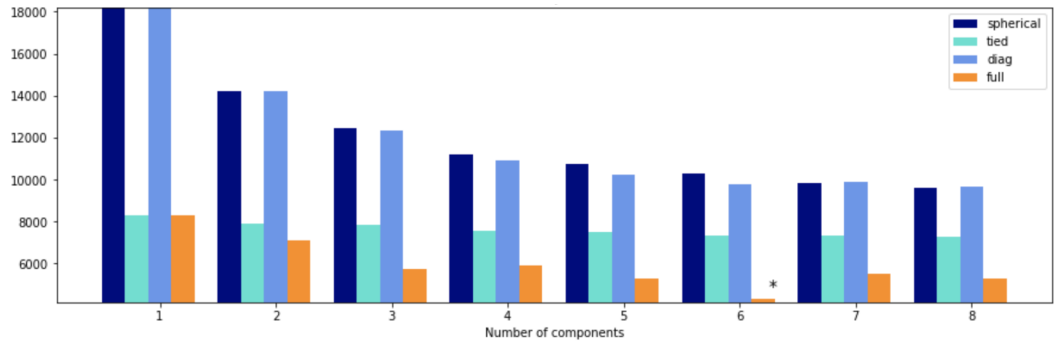

Figure 24: BIC scores per model and covariance matrix type. The starred bar represents the best model (6 clusters).

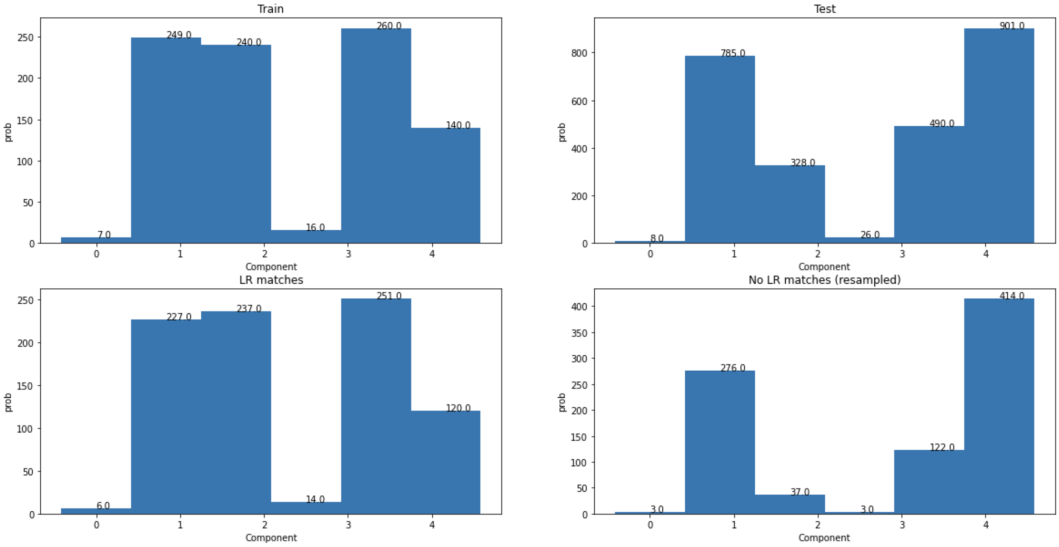

Figure 25: Histograms of predicted labels for all sets (6 clusters). Each bar/cluster represents the number of points from the set that are contained within itself.

still obtains the second lowest from a total of 10 runs (Table 10). As shown in Figure 24, the final model for that execution is 6. It is clear that a full covariance matrix for each cluster produces notable improvements under the BIC criteria.

Figure 25 is composed by four histograms, each of them corresponding to the predicted labels of points conforming the described sets in their corresponding title.

Similarities between histograms from train (upper left corner) and LR matches (bottom left corner) are easily explained since 93.75% of LRT matches are contained in the training set.

The test set (upper right corner) fully contains the train set, in addition to pair of sources that lay between $2'' < r < 6''$. The inclusion of such points increase the population of the four densest clusters depicted in the train histogram.

Given the great difference of sample sizes between the LR matches and the NM sets (855 vs 1683), resampling has been applied to the last set in order to compare equal number of samples. Note that both sets are totally exclusive, therefore no point in any of both sets is present in the other.

Several insights can be extracted when comparing the cluster's distribution of LR matches (bottom left corner) in contrast to false matches (bottom right corner).

In one hand, clusters 0 and 3 are clearly the least representatives in contrast with the rest due to its small population. Nonetheless is observed that only 14.3% (cluster 0) and 6.7% (cluster 3) of

[6]Percentages are computed given the assumption that the cluster's total size is the sum of the LR matches set and the NM set sizes (bottom left and right histograms in Figure 25).

false identifications are located in such clusters. [6] It seems that most of the pairs contained in the mentioned clusters are likely to be true counterparts. However, the discriminative power of both clusters is weak due to the small proportion of points w.r.t. the total.

On the other hand, clusters 2 and 5 show a robust majority from points belonging to a single set. While in cluster 2, 86.5% are LRT matches, 77.5% of pairs are false matches in cluster 5.

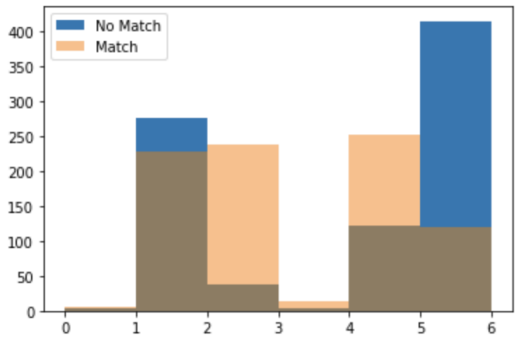

Figure 26: Overlapped histograms of predicted labels from LR matches and NM set.

A deeper analysis should be carried out to confirm the previous conclusions since biases could be hidden. In the best case scenario, the obtained proportions within clusters can only provide a prior probability, not a methodology to separate true identifications from wrong ones.

Continuing with the analysis, it is convenient to inspect the characteristics of each cluster and compare them to the prior knowledge obtained with *LRT*. It may be possible that improvements acquired in the color modality in contrast to the only-magnitude version could be due to discriminative structures found on the current cluster analysis.

In order to address the prior, densities from each cluster are estimated as described in section 4.2.

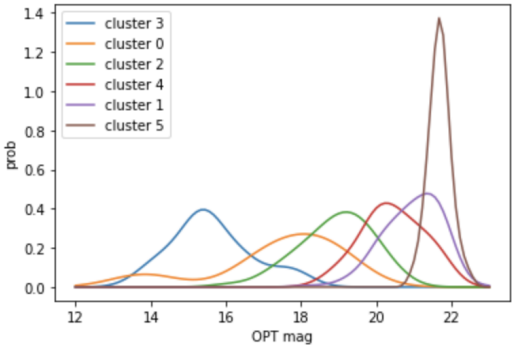

Figure 27: Estimated densities of each cluster by means of a **KDE**.

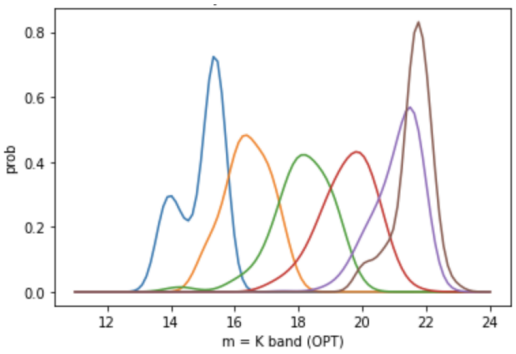

Figure 28: $q(m, c)$ for 6 color bins.

In the above figures it is depicted the estimated densities from each cluster (Figure 27) and $q(m, c)$ for the optimal number of color bins from color *LRT* modality (Figure 28) ran over **MIPS** and **OPT** catalogs.

Before addressing similarities an important fact is that densities from clusters 0 and 2 are estimated with a small sample (8 and 16 respectively), negatively influencing estimations. Note that the model under description (6 bins) is not the best one regarding the BIC criteria.

In spite of the above, it is seen that some cluster's pdfs have similarities with respect some curves depicting some color bins found in $q(m, c)$. For instance, regarding the brown curves (right most graphs in both figures), it can be said that their mode is roughly equal. Purple curves follow the same behavior as well. It is not a coincidence that these densities are precisely the ones corresponding to the densest clusters, therefore estimations are reliable.

For the estimation of $q(m, c)$, the color dimension was divided in equally-spaced ranges to obtain each $q(m, c = c^*)$. Thus, overlapping between color bins is reduced. In contrast to the clusters found on the GMM model, their densities present stronger superposition since clusters' domains were not explicitly defined as in $q(m, c)$, they are learned from inherent structures within data (mainly color).

Modal equality in some curves does not imply a direct relationship. Given the deficient estimates in several curves is impossible to correctly quantify density similarity, e.g. using KL divergence.

Following the same format as in the mixture model for 6 clusters, results are shown for 5 in Figures 29 and 30. Such model achieves the lowest score in BIC among a total of 10 runs.

Again, independent covariance matrices for each cluster produces the best model. Nonetheless, it can be seen that scores are similar for 6 clusters at that execution. Looking at the previous results for

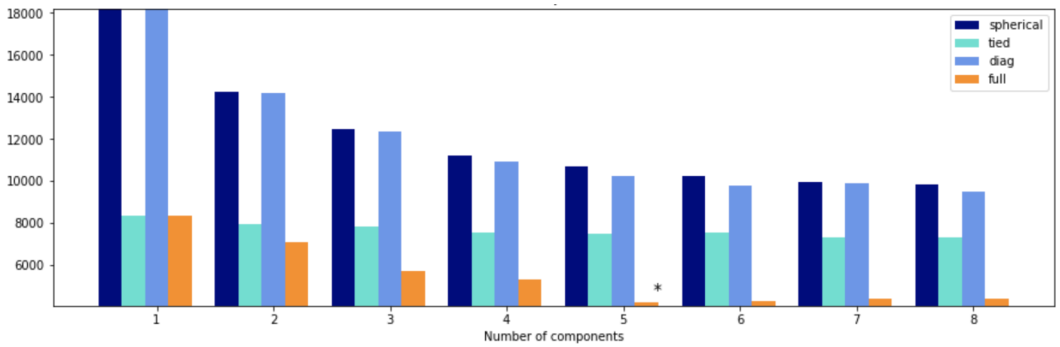

Figure 29: BIC scores per model and covariance matrix type. The starred bar represents the best model (5 clusters).

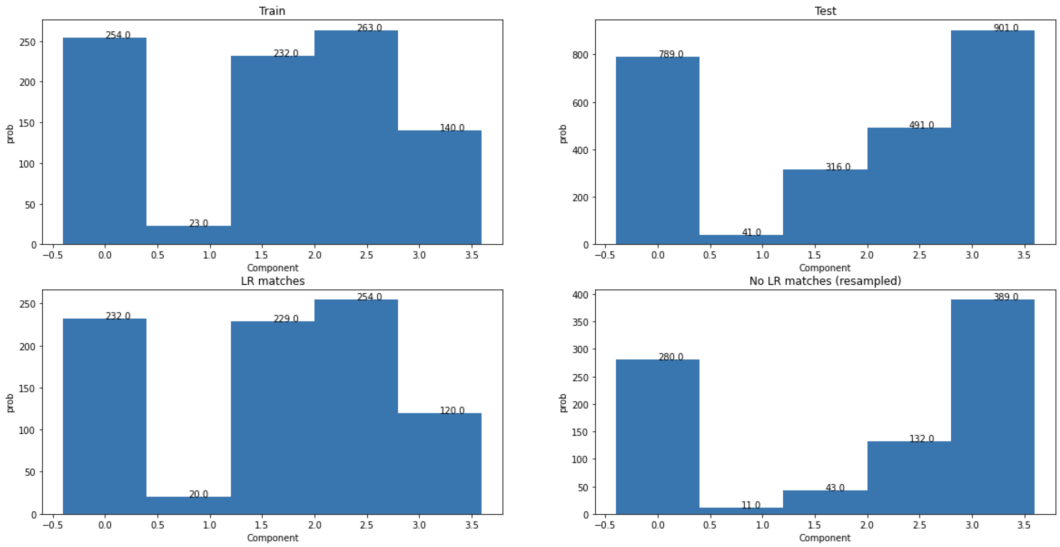

Figure 30: Histograms of predicted labels for all sets (5 clusters). Each bar/cluster represents the number of points from the set that are contained within itself.

6 clusters, two clusters were less representative w.r.t. the others, thus an intuitive claim would be that reducing the number of clusters by one unit would help 'fill' those spaces to maximize the total number of representatives clusters. This is not always the case under our specific circumstances. It was already seen that in the previous run where 6 clusters was the optimal choice, reducing the number of clusters did not improve quantitatively the model (BIC).

As expected, histograms of predicted labels for each set show more relevance and representational power in 4 out of 5 clusters. In comparison to 6 clusters, points have been redistributed rather homogeneously.

After resampling the NM set and comparing it with the LR matches set through the histograms computed upon their predicted labels, there are some interesting insights to mention.

Generally speaking, there are only a couple of clusters (2 and 4) that manifest a majority proportion of points coming from one set (LRT matches vs NM). More specifically, 84.2% of points in cluster 2 are matches and only 23.57% in cluster 4. Therefore, these clusters are (apparently) capturing characteristics that are most discriminative between true and false identifications.

Cluster 3 does show a prevalence regarding the nature of the points within it, however such majority is arguable regarding its significance in terms of discriminative properties. This is due to a weaker

[7] Again, computations regarding proportion follow the same rule as explained in the previous model, i.e. the total population of each cluster is considered the sum of the last attributed to the predicted labels of sets LR matches and NM.

proportion (65.8% of matches) in contrast to that of clusters 2 and 4.[7]

The rest of the cluster do not show robust discriminative statistics (less that 60%).

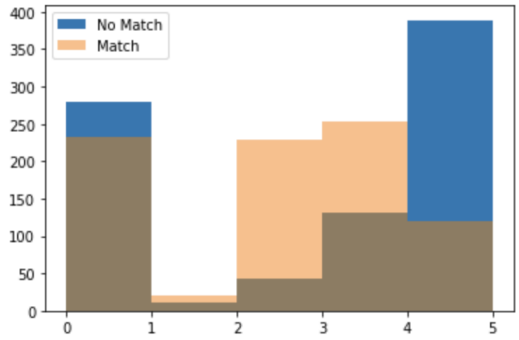

Figure 31: Overlapped histograms of predicted labels from LR matches and NM set.

Similarly as exposed in the previous example, Figure 32 shows densities of each cluster learned by the model, estimated by means of a **KDE**.

As occurred with some clusters in the 6 clusters model, there is one deficient estimation (blue curve representing cluster 1 produced by a small sample size (23).

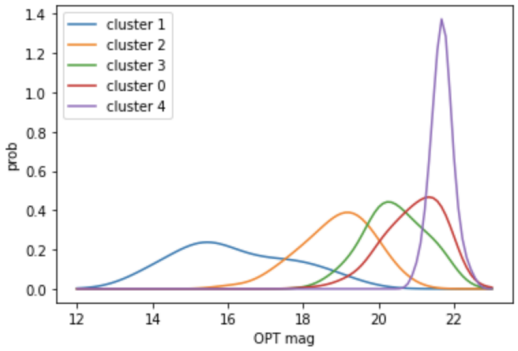

Figure 32: Estimated densities of each cluster by means of a **KDE**.

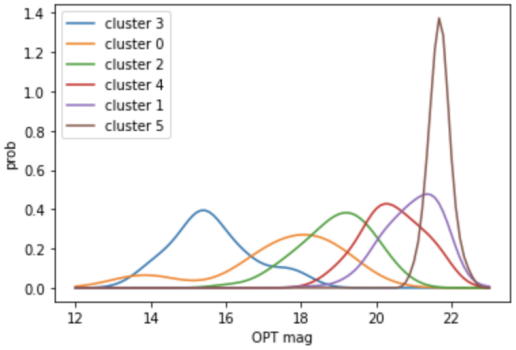

Figure 33: Estimated densities of each cluster by means of a **KDE**.

Looking at Figure 33, depicting the already shown densities for 6 clusters, the four right-most clusters in both graphs have almost identical form (Table 11).

M. 6 clust.	**Color legend**	**M. 5 clust.**	**Color legend**
5	brown	4	purple
1	purple	0	red
4	red	3	green
2	green	2	orange

Table 11: Similarity of densities of clusters between models with 5 and 6 clusters. The 2^{nd} and 4^{th} columns are associated to the color legends in Figures 33 and 32 respectively.

The (apparently) common cluster in both models are also the most representative ones, while fainter clusters (regarding brightness) are both wrongly estimated and less condensed.

Concluding with this experiment, the optimal mixture model selected under the BIC criteria (5 clusters), seems to extract a more reliable distribution within clusters, in spite of sharing statistical structure with the model trained for 6 clusters. Mainly this is due to the fact that points are predicted into denser clusters, leaving just a non-representative one.

5 Conclusions & Further Work

The current work faces a complex problem of unsupervised nature that cannot be defined under any known format within the ML field (i.e. classification, estimation, etc...). The only option is to find reliable structures within data in addition to specific knowledge based tools in order to overcome the x-matching problem.

The results exposed in this document are ambiguous in order to establish a clear and direct relationship between the *LRT* performance in relation to the hidden structures found by the mixture model. Notwithstanding, some similarities between both procedures bring some evidence into the right direction and could help future lines of work willing to inquire deeply in the subject.

Last but not least, it is crucial to mention that all this results are based upon assumptions that could be not accurate enough. Several sets that are used across this experiment rely on the identifications thrown by the *LRT* color modality, considered as ground-truth. This might as well introduce biases that could hassle the current analysis. In any case it is a good starting point to begin with.

Further work could be driven in the direction of astronomical objects classification, ultimately allowing a more accurate process towards a x-

matching algorithm. Class estimation should be carried out individually at each catalog.

Pairs of candidates could also be processed in terms of property similarities, provided that catalogs contain several band measures for each point. Combining the estimated label defining the type of star (i.e. galaxy, star, etc...) and several brightness measures from both candidates, ML could estimate whether both potential counterparts are describing the same Spectral Energy Distribution (SED), which models the magnitude of a single object at any given frequency. A condition for not guaranteed success is the availability of high populated catalogs since SEDs vary dramatically between objects, even of the same class. In such scenario with huge amounts of data, pivoting to deep learning techniques would be a plausible option.

References

[1] Ciliegi, P., Zamorani, G., Hasinger, G., Lehmann, I., Szokoly, G., and Wilson, G. (2003). A deep vla survey at 6 cm in the lockman hole. A&A, 398(3):901–918.

[2] Fleuren, S., Sutherland, W., Dunne, L., Smith, D. J. B., Maddox, S. J., González-Nuevo, J., Findlay, J., Auld, R., Baes, M., Bond, N. A., Bonfield, D. G., Bourne, N., Cooray, A., Buttiglione, S., Cava, A., Dariush, A., De Zotti, G., Driver, S. P., Dye, S., Eales, S., Fritz, J., Gunawardhana, M. L. P., Hopwood, R., Ibar, E., Ivison, R. J., Jarvis, M. J., Kelvin, L., Lapi, A., Liske, J., Michałowski, M. J., Negrello, M., Pascale, E., Pohlen, M., Prescott, M., Rigby, E. E., Robotham, A., Scott, D., Temi, P., Thompson, M. A., Valiante, E., and Werf, P. v. d. (2012). Herschel -ATLAS: VISTA VIKING near-infrared counterparts in the Phase 1 GAMA 9-h data . Monthly Notices of the Royal Astronomical Society, 423(3):2407–2424.

[3] Luo, B., Brandt, W. N., Xue, Y. Q., Brusa, M., Alexander, D. M., Bauer, F. E., Comastri, A., Koekemoer, A., Lehmer, B. D., Mainieri, V., and et al. (2010). Identifications and photometric redshifts of the 2 ms chandra deep field-south sources. The Astrophysical Journal Supplement Series, 187(2):560–580.

[4] Nisbet, D. D. (2018). PhD thesis. PhD thesis, The University of Edinburgh.

[5] Notni, P. and Richter, G. A. (1976). Optical Identifications of Radio Sources in the 5C Areas. Astronomische Nachrichten, 297(6):265.

[6] Sutherland, W. and Saunders, W. (1992). On the likelihood ratio for source identification. , 259:413–420.

[7] Williams, W. L., Hardcastle, M. J., Best, P. N., Sabater, J., Croston, J. H., Duncan, K. J., Shimwell, T. W., Röttgering, H. J. A., Nisbet, D., Gürkan, G., and et al. (2019). The lofar two-metre sky survey. Astronomy Astrophysics, 622:A2.

Acknowledgments and Disclosure of Funding

The work of Óscar Manuel Jiménez Rama has been supported by an ISDEFE scholarship from April 2020 to September 2020.

This work was supported by the project Evolution of Galaxies AYA2017-88007-C3-1-P, within the "Programa estatal de fomento de la investigación científica y técnica de excelencia del Plan Estatal de Investigación Científica y Técnica y de Innovación (2013-2016)"of the "Agencia Estatal de Investigación del Ministerio de Ciencia, Innovación y Universidades"