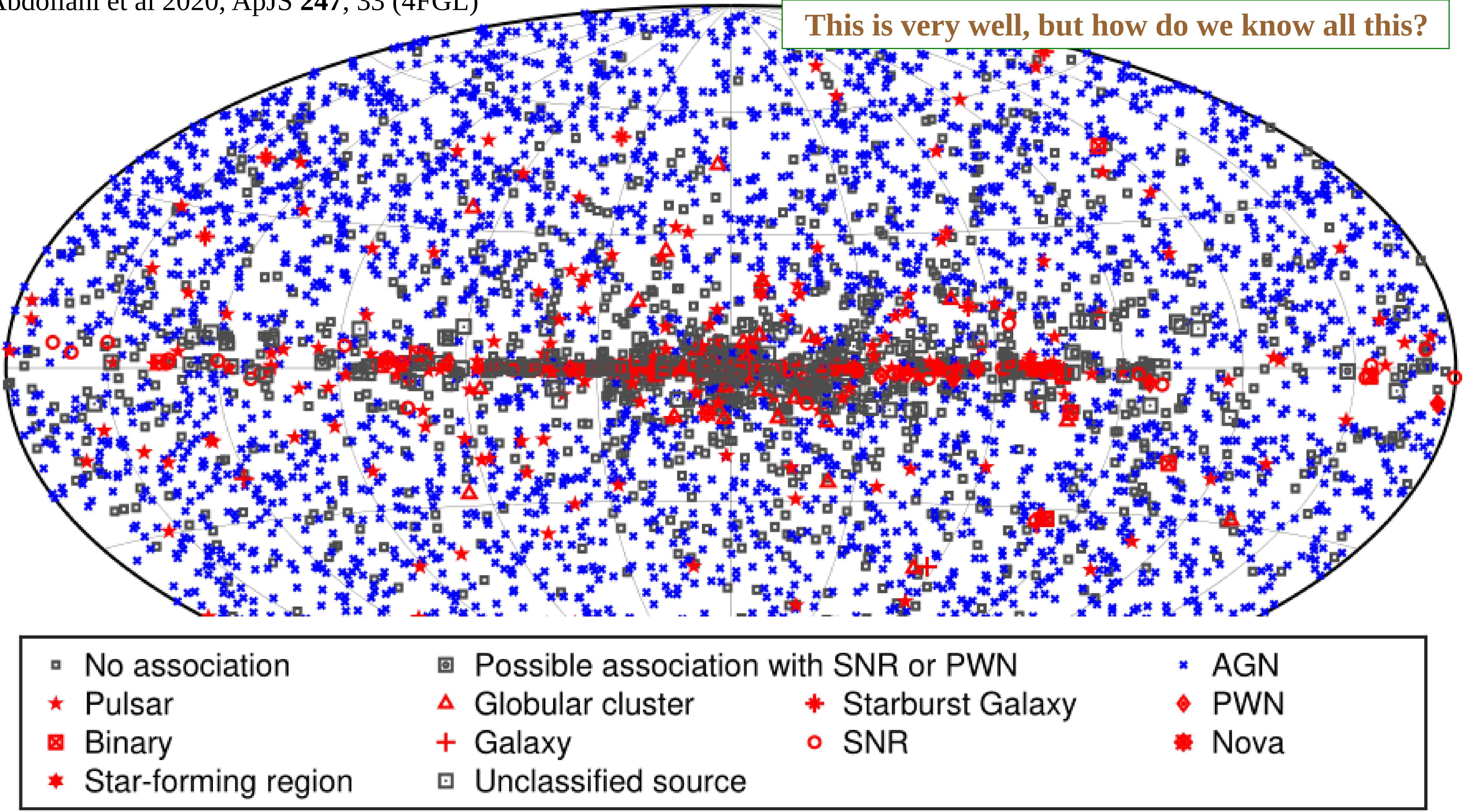
Source association

1. Spatial association
2. Purity vs completeness
3. Chasing systematics
4. Other criteria
5. Galactic complications
6. Extended sources

Abdollahi et al 2020, ApJS **247**, 33 (4FGL)

This is very well, but how do we know all this?



What problem do we want to address?

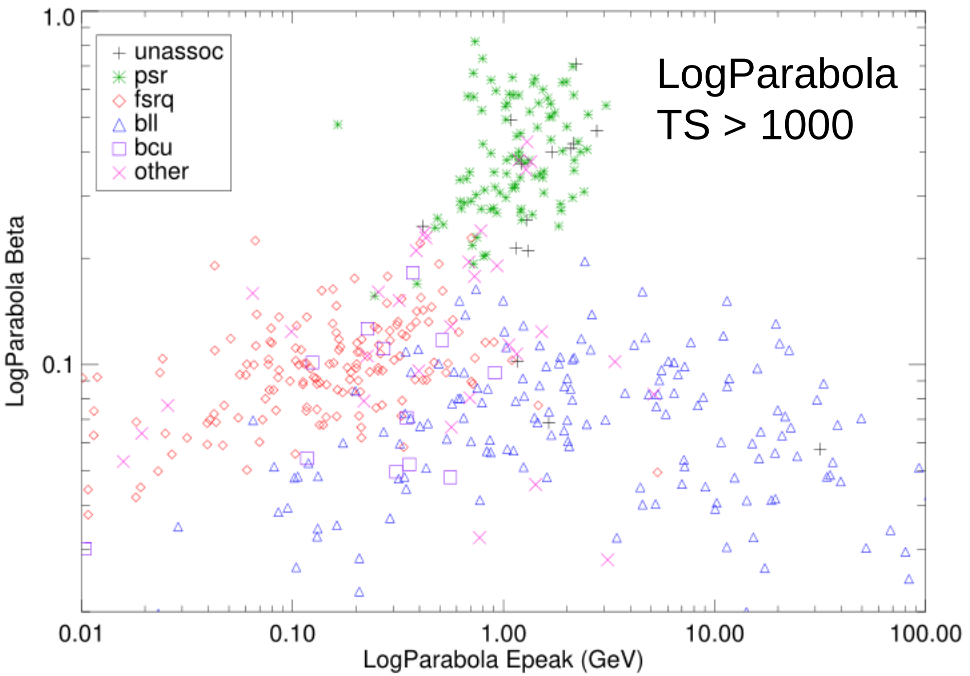
When a γ-ray source is found by chance, **how do we associate it with what we know from other wavelengths**?

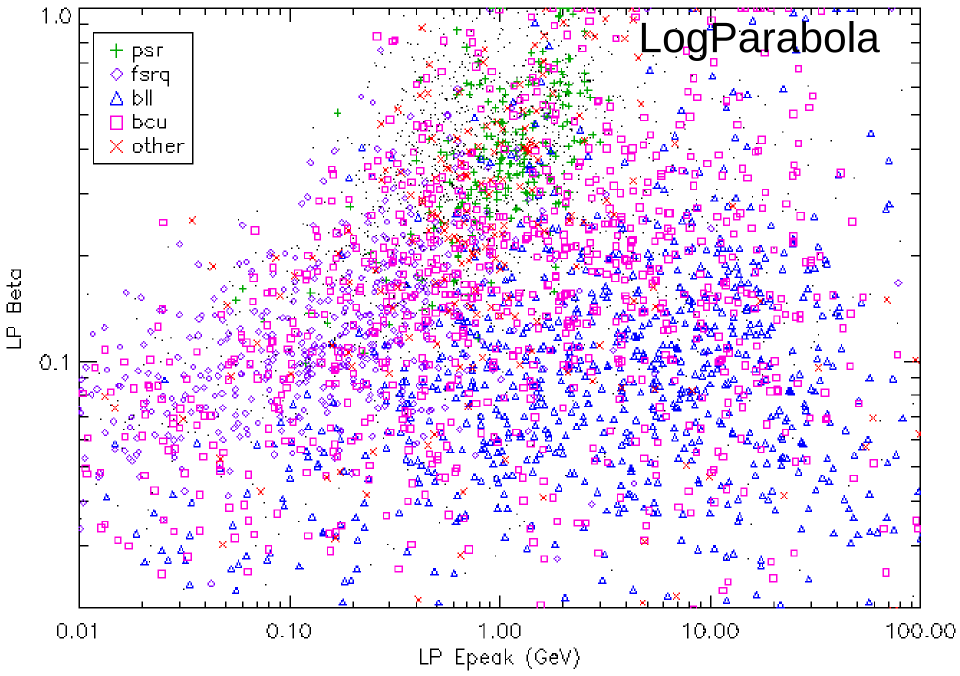
1. Applies to surveys (Fermi-LAT, eROSITA, CTA Galactic and extragalactic surveys)
2. True identifications (correlated variability) are few (except pulsars)
3. Critical for population studies and physical modelling
4. Probabilistic approach
5. Long history, first concepts date back to the 1970s for first radio surveys
6. Often called **cross-match** in the literature
7. Some concepts can apply to pointed observations when you want to assess the probability that what you see comes from a more common source class (*e.g.* blazars) than what you are looking for
8. Of little use for extended sources, unfortunately

Classification

How can you guess what a source is, when it is unassociated or associated to an unknown MWL source?

1. Entirely different problem, which does not use localization
2. Must rely primarily on γ-ray properties (flux, spectrum, timing, position in the sky) and on counterpart properties if any
3. Based on what we know from (hopefully) reliable associations
4. Harness all parameter space globally
5. Typically the kind of thing that machine learning does, lots of developments outside astronomy. Take care of defining proper training and test samples.
6. Specific difficulty with faint sources, not so well characterized, but ML is supposed to know about uncertainties





What information do we need for association?

What quantities do we expect will matter to this problem?

1. How well we localized the γ-ray source (the localization precision). In other contexts the localization precision of the counterparts may matter too (assume negligible here)

2. How many potential counterparts we consider (the counterpart density)

3. The plausibility that those counterparts emit γ-rays (not the same for stars and blazars). If possible, this is handled before, by selecting classes of sources that we **know** collectively emit γ-rays (blazars, pulsars)

4. The individual properties of the counterparts (flux, spectrum, …)

Let us put that into equations

Probabilistic framework

We want to compare two hypotheses:

1. H_0: A putative counterpart is close to a γ-ray peak by chance
2. H_1: The putative counterpart is actually the same as the γ-ray source

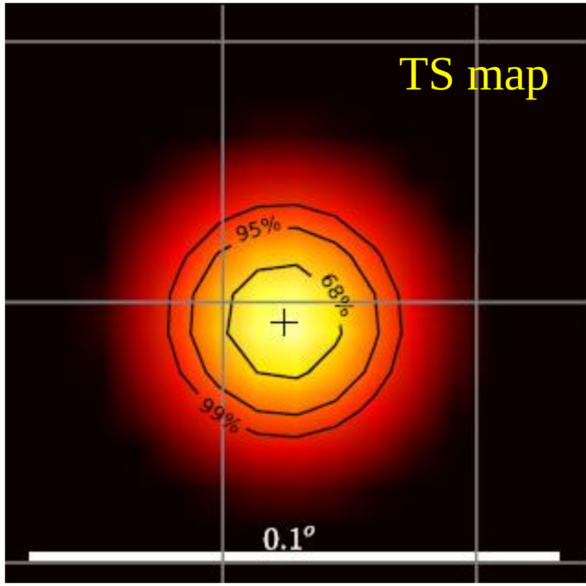
We will adopt a Bayesian approach: **Pr{M|D} = Pr{M} Pr{D|M} / Pr{D}**

where Pr(D) is just a normalization constant

How do we get the localization precision?

The instrument's Point Spread Function (PSF) is the key ingredient

1. If the PSF is the same for all events (not energy dependent, in particular), with dispersion **σ** along one axis, then the dispersion of the average over ***N*** events is **$\sigma / \sqrt{N}$**
2. For many counts, the compound localization will converge to a **Gaussian** (central-limit theorem) of the same dispersion
3. In general, (not the same PSF for all events) the localization precision will be obtained from the logLikelihood using Wilks' theorem. Assuming that the source is truly at position **r_T**, **$\Delta = 2 \ln(L_{max}/L_T)$** is distributed as **$\chi^2$(2 dof)** when the 2D position is fitted to the data. Particularly simple **$F(\Delta) = 1 - \exp(-\Delta/2)$**.
Related to $TS = 2 \ln(L_{max}/L_0)$ used for assessing the significance of a source



How do we get the probability density under H_1?

Definition of the localization error

1. Remember that $\mathbf{F(\Delta) = 1 - \exp(-\Delta/2)}$ with $\mathbf{\Delta = 2 \ln(L_{max}/L_T)}$

2. The likelihood contours are not necessarily symmetrical (either due to instrumental characteristics or to background features such as other nearby sources) but again for enough counts the tip converges to a Gaussian propto $\mathbf{\exp(-(r/\sigma)^2/2)}$, lnL becomes a 2D paraboloid and the contours converge to **ellipses**. 95% confidence contours: $\mathbf{\Delta = -2 \ln(0.05) = -(R_{95}/\sigma)^2}$ **so** $\mathbf{R_{95} = \sqrt{(-2 \ln(0.05))}\, \sigma \approx 2.45\, \sigma}$

3. Under H_1, in the simplest case of an error circle, the probability density of the distance between the γ-ray peak and the counterpart is $\mathbf{f_T(r) = r/\sigma^2 \exp(-(r/\sigma)^2/2)}$
   $\mathbf{r = ||r_P - r_T||}$ is viewed as a random variable in $\mathbf{r_P}$ (γ-ray peak when the counterpart is known) but in the Bayesian approach it can be turned around and viewed as a random variable in $\mathbf{r_T}$ (counterpart position when the γ-ray peak is known)

We neglect here complications related to the sphericity of the sky

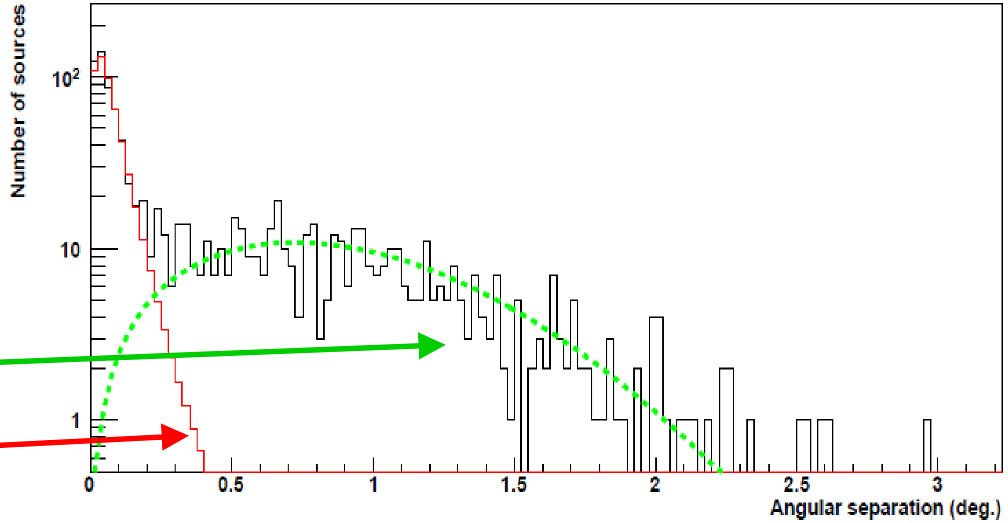
How do we get the probability density under H_0?

In general, we start from a catalog of counterparts

1. Under H_0, if the counterpart density **ρ** is reasonably constant (for example AGN outside the Galactic plane), then, noting **r** the 2D distance to any point in the sky, $dN/dr = 2\pi r\rho$
2. As long as all counterparts are considered equal, we will consider the **nearest one**
3. The probability of finding the nearest neighbor at **x** or beyond is $p(x) = Pr\{N(r<x)=0\}$. We can write $p'(x) = 2\pi x\rho\, p(x)$ so that $p(x) = \exp(-\pi x^2 \rho)$ (reverse CDF or survival function)
4. Under H_0, the probability density of the distance to the closest counterpart is $f_R(r) = -p'(r)$ so that $f_R(r) = 2\pi r\rho \exp(-\pi r^2 \rho)$
5. To get there, we need the catalog to be complete (at a given flux limit). If the detection rate varies over the sky (*e.g.* AGN through the Galactic plane), it must be accounted for.

We assume in the following that the local counterpart density **ρ** can be obtained

Likelihood ratio



We compare the two probability densities (random and true)

1. H_0: $\mathbf{f_R(r) = 2\pi r \rho \exp(-\pi r^2 \rho)}$
2. H_1: $\mathbf{f_T(r) = r/\sigma^2 \exp(-(r/\sigma)^2/2)}$
3. Likelihood ratio $\mathrm{LR}(r) = \dfrac{f_\mathrm{T}(r)}{f_\mathrm{R}(r)} = \dfrac{1}{2\pi\varrho\sigma^2}\exp\left(\pi\rho r^2 - \dfrac{r^2}{2\sigma^2}\right) = \dfrac{1}{K}\exp\left((K-1)\dfrac{r^2}{2\sigma^2}\right)$
4. There is no free parameter here: **σ** comes from the logLikelihood contours (specific to each source), **ρ** is assumed known (but can depend on direction in the sky) and **r** is simply the distance between the γ-ray peak and the counterpart (observed quantity).
5. No hope to ever find a reliable counterpart if $\mathbf{K = 2\pi\rho\sigma^2 > 1}$. In that case (one random counterpart on average in $\mathbf{R_{68}}$) LR does not even decrease with **r**.
6. The likelihood ratio provides a ranking between associations (over a full catalog) but not a probability

Association probability

In the Bayesian approach, we must consider the a priori probabilities

1. A priori $\mathbf{Pr\{H_1\}}$ and $\mathbf{Pr\{H_0\}}$ such that $\mathbf{Pr\{H_1\}+Pr\{H_0\}=1}$ and $\mathbf{\Gamma=Pr\{H_1\}/Pr\{H_0\}}$
2. A posteriori $\Pr\{H_1|r\} = \dfrac{\Pr\{H_1\} f_{\mathrm{T}}(r)}{\Pr\{H_1\} f_{\mathrm{T}}(r) + \Pr\{H_0\} f_{\mathrm{R}}(r)} = \dfrac{1}{1+1/(\Gamma\,\mathrm{LR}(r))}$
3. At this point $\mathbf{Pr\{H_1\}}$ and $\mathbf{Pr\{H_0\}}$ (or $\mathbf{\Gamma}$) are not known yet.

Only spatial characteristics ($\boldsymbol{\rho}$ and $\boldsymbol{\sigma}$) are considered for now.

Thresholding

We want to define a probability threshold

1. A counterpart is considered safe if $\Pr\{H_1|r\} = \frac{1}{1+1/(\Gamma\,\mathrm{LR}(r))} > \beta$
2. Equivalent to $\mathrm{LR}(r) = \frac{1}{K}\exp\left((K-1)\frac{r^2}{2\sigma^2}\right) > \frac{1}{\Gamma(1/\beta-1)} = \alpha$
3. Or to $r < r_{\max} = \sigma\sqrt{-\frac{2\ln(K\alpha)}{(1-K)}}$ where $\mathbf{K = 2\pi\varrho\sigma^2}$
4. No association can be found if **K > 1** (already seen) or **K > 1/α**, if ϱ is too large. Usually α > 1 so the second condition is more stringent.
5. The **r/σ** threshold depends only on **α**, not separately on **Γ** and **β**. It decreases with **K** and **α** so sources will be accepted at larger **r/σ** for smaller ϱ, larger **Γ** and smaller **β**
6. Ex: **β = 0.8**, **Γ = 1/2** ➔ **α = 8**. **ϱ = 1** /sq deg: association possible if **σ < 0.141°**.

Catalog of γ-ray sources

We are now considering entire catalogs

1. We work with a catalog of **M** γ-ray sources indexed by **i**, with localization precision σ_i
2. We note $p_i = \Pr\{H_1|r_i\} = \frac{1}{1+1/(\Gamma\,\mathrm{LR}(r_i))}$. Good association if $p_i > \beta$
3. Remember that the threshold in **r/σ** depends only on **α**, not separately on **Γ** and **β**. So we can decide that we will set **β** to 0.8, say (the same for all counterpart catalogs), and it remains to choose **Γ** (separately for each counterpart catalog).

The localization of counterparts is assumed to be better than the γ-ray localization (in general, a few arcsecs vs a few arcmins) and we consider only the **closest one** in this simple approach

False associations

How many false associations do we expect?

1. The condition $p_i > \beta$ will result in N_{assoc} accepted sources
2. The number of false associations is a random variable F_i whose value is either 0 or 1.
3. In this framework the distance r_i is not a random variable but an observed quantity and the localization error σ_i is known from the logL contours. So F_i is a simple Bernoulli variable with probability $1 - p_i$. Its expectation is $E_T(F_i) = 1 - p_i$ and its variance is $V_T(F_i) = p_i (1 - p_i)$
4. The total number of false associations is $F = \Sigma\{p_i>\beta\} F_i$. Its expectation is $E_T(F) = \Sigma\{p_i>\beta\} (1–p_i)$ and the sources are independent so its variance is $V_T(F) = \Sigma\{p_i>\beta\} p_i(1–p_i)$
5. By construction all $p_i > \beta$ so $\beta E_T(F) < V_T(F) < E_T(F)$, close to the Poisson regime $V(F) = E(F)$
6. p_i depends on the a priori probability ratio Γ, so the expected number of false associations is a function of Γ. When $\Gamma << 1$ (H_1 very unlikely), $p_i < \beta$ for all sources so $E_T(F) = 0$. When $\Gamma >> 1$ (H_1 very likely), $p_i \rightarrow 1$ for all sources so $E_T(F) \rightarrow 0$ too. $E_T(F)$ reaches a maximum for moderate Γ

False associations 2

How many false associations do we expect?

1. We can also estimate the number of false associations in a different manner, either by simulations (move the γ-ray sources, apply the procedure and count) or by a simple surface estimate.

2. In this framework the distance $\mathbf{r_i}$ is again the random variable in H_0 so that the cumulative probability of having one random counterpart within $\mathbf{r_i}$ is $\mathbf{F_R(r_i) = 1 - \exp(-\pi r_i^2 \rho)}$.

3. Accept associations up to $r_i^{\max} = \sigma_i \sqrt{-\frac{2\ln(K_i\,\alpha)}{(1-K_i)}}$ where $\mathbf{K_i = 2\pi\varrho\sigma_i^2}$ and $\boldsymbol{\alpha = 1/(\Gamma(1/\beta-1))}$

4. Expected number of false positives $E_R(F_i) = F_R(r_i^{\max}) = 1 - \exp\left(\frac{K_i \ln(K_i\,\alpha)}{(1-K_i)}\right)$

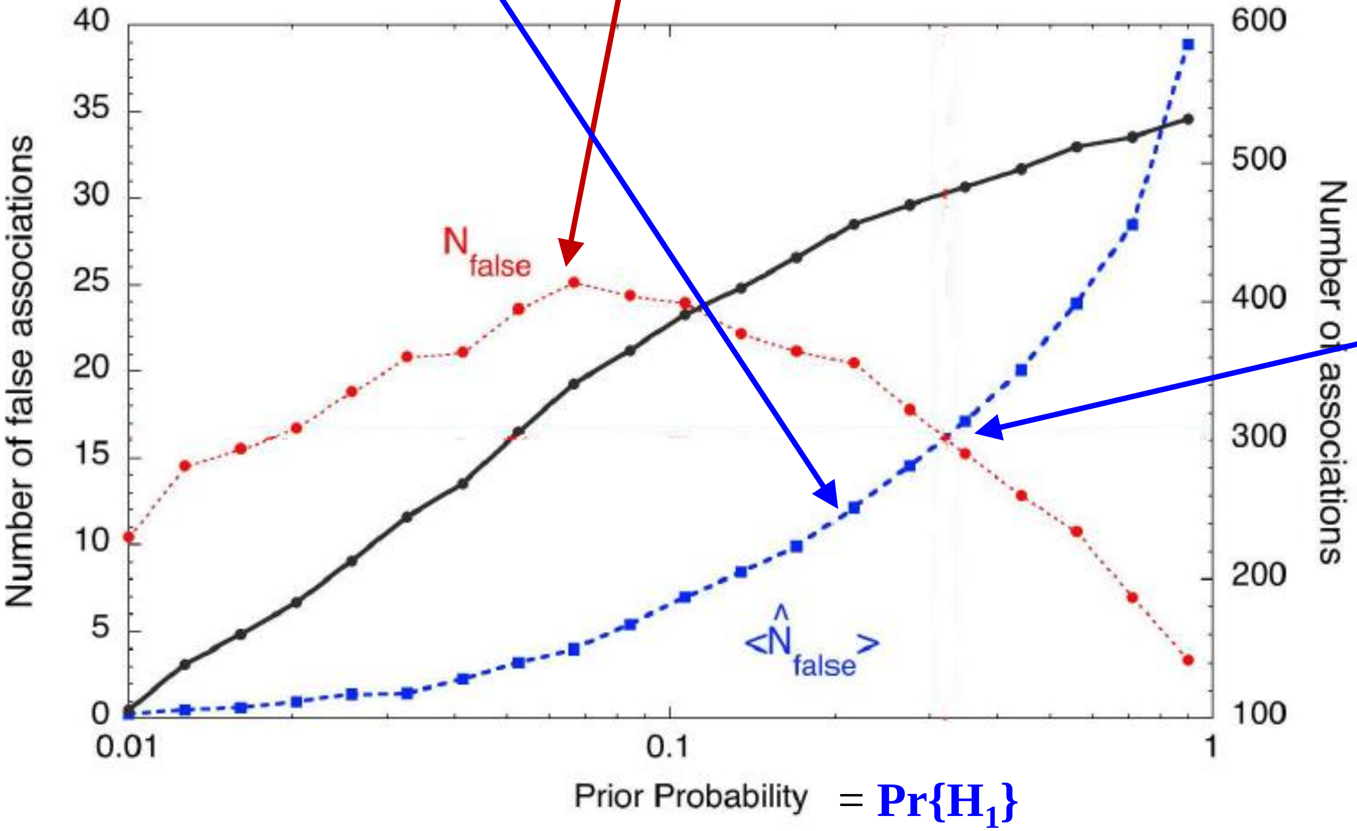
5. Summed over all sources $\mathbf{E_R(F) = \Sigma\, F_R(r_i^{max})}$ restricted to $\mathbf{K_i < \min(1,1/\alpha)}$

6. $\mathbf{K_i}$ does not depend on $\mathbf{\Gamma}$, and $\boldsymbol{\alpha}$ decreases with $\mathbf{\Gamma}$, so $\mathbf{E_R(F)}$ increases with $\mathbf{\Gamma}$ from 0 to **M**

Defining the prior probability

We must reconcile the two estimates of false associations

Writing **$E_R(F) = E_T(F)$** results in an equation over **Γ** or **$Pr\{H_1\}$** that can be solved numerically



Example of such curves

Scale for false associations is at left

Black curve (with scale at right) is the total number of associations

The correct choice of **$Pr\{H_1\}$** is where the red and blue curves intersect, at **$Pr\{H_1\} \approx 1/3$** or **$\Gamma \approx 1/2$**

This means that we expect about 1/3 of the γ-ray sources to be among those counterparts

The reliability **R** (called **precision** in statistics) is the fraction of true associations among accepted ones **$R = 1 - E_T(F)/N_{assoc}$**

In the example **$R \approx 1 - 17/460 = 0.96$**

True associations

How many true associations do we expect?

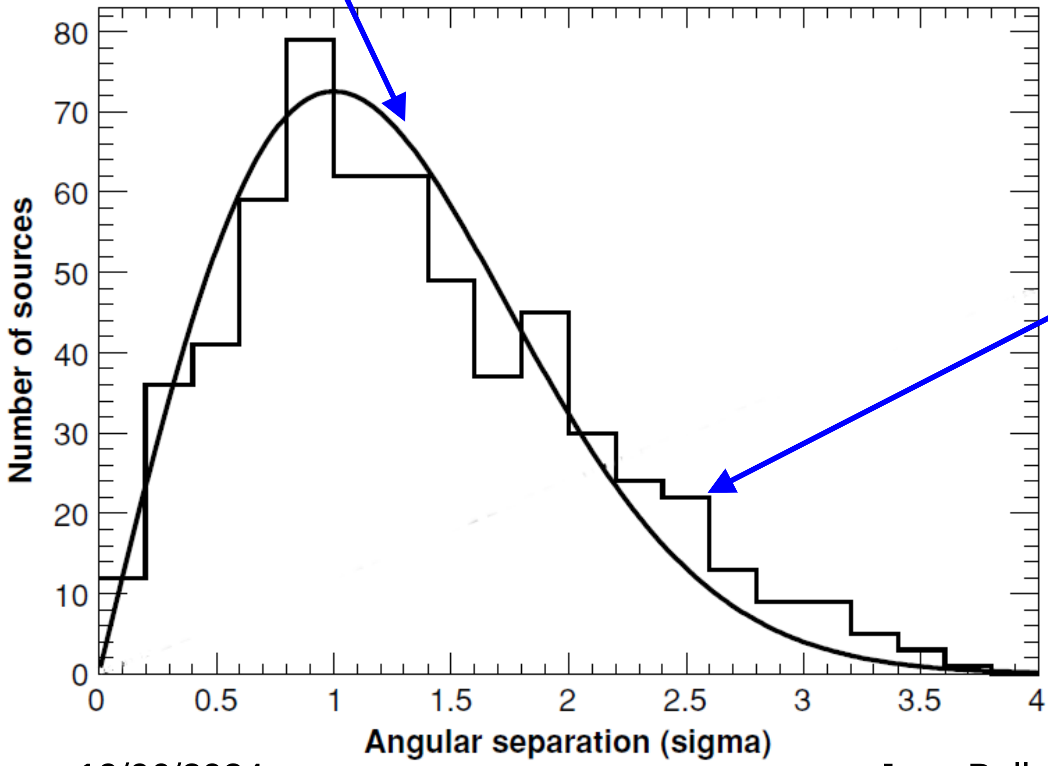
1. The number of true associations is also a random variable $\mathbf{T_i}$ whose value is either 0 or 1.
2. $\mathbf{T_i}$ is again a simple Bernoulli variable with probability $\mathbf{p_i}$. Its expectation is $\mathbf{E(T_i) = p_i}$ and its variance is $\mathbf{V(T_i) = p_i\ (1 - p_i}) = \mathbf{V(F_i)}$
3. The total number of true associations among the accepted sources is $\mathbf{T_{acc} = \Sigma\{p_i>\beta\}\ T_i}$. Its expectation is $\mathbf{E(T_{acc}) = \Sigma\{p_i>\beta\}\ p_i = N_{assoc} - E_T(F)}$ and its variance is $\mathbf{V(T_{acc}) = V_T(F)}$
4. $\mathbf{V(T_{acc}) < (1-\beta)\ E(T_{acc})}$, way below Poisson regime $\mathbf{V(T) = E(T)}$
5. The total number of true associations is $\mathbf{T_{tot} = \Sigma\ T_i}$. Its expectation is $\mathbf{E(T_{tot}) = \Sigma\ p_i}$
6. $\mathbf{E(T_{tot}) - E(T_{acc})}$ is the estimated number of missed associations (we cannot tell which ones)
7. The completeness **C** (called **recall** in statistics) is the fraction of accepted true associations among all true associations and can be estimated as $\mathbf{E(T_{acc}) / E(T_{tot})}$

The Rayleigh distribution

We must quantify the quality of the procedure

The distribution of distances **r** differs for each source but that of **r/σ** is always the same

$f_T(r/\sigma) = x \exp(-x^2/2)$: Rayleigh distribution (2D Gaussian in polar coordinates)



Example on real sources

The black histogram is the observed distribution of **r/σ**

The curve is the Rayleigh distribution normalized to the number of sources

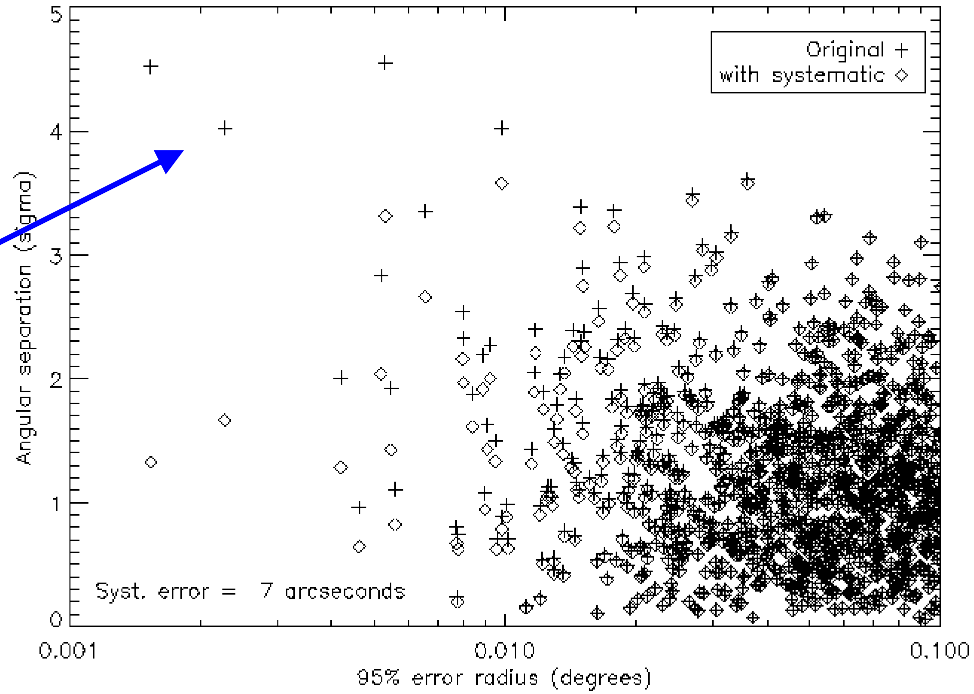
They can be compared by a Kolmogorov-Smirnov test

This example is not perfect. The histogram has a distinct tail, implying that something could be improved

Chasing systematics

What can explain a tail in the observed distribution of r/σ?

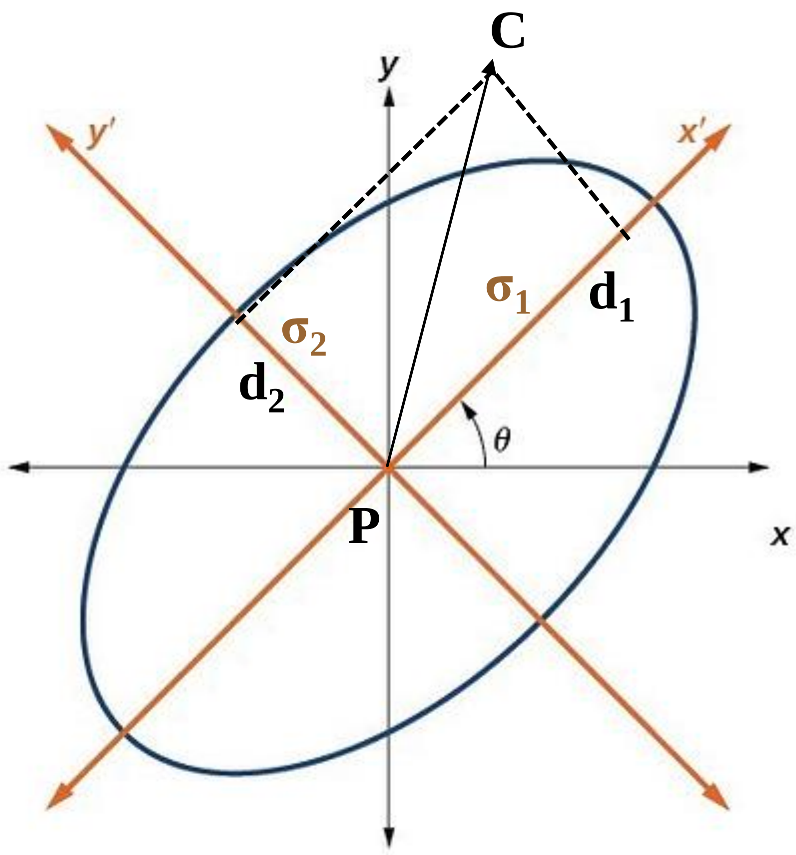
1. In general, the culprit is our estimate of the localization precision **σ**
2. It can be wrong in two ways:
   - An absolute systematic error σ_{abs} (due to imperfect knowledge of the pointing direction) will affect the bright sources and can be checked by looking at bright known sources
   - A relative systematic error f_{rel} (due to confusion or background modeling) will affect all sources. f_{rel} can be fit to optimize the Rayleigh plot
3. Combined as $\sigma_{tot}^2 = (f_{rel}\ \sigma)^2 + \sigma_{abs}^2$



Elliptical errors

How to go beyond a simple error circle?

1. In general, the localization region is an ellipse defined by two errors σ_1 and σ_2 and an angle θ (can be wrt North or West)
2. In that case, the counterpart position **C** wrt the γ-ray source **P** must be expressed in the ellipse axes → **(d1, d2)**
3. The previous formulae can then be used, replacing σ by $\sqrt{(\sigma_1 \sigma_2)}$ and $(r/\sigma)^2$ by $(d_1/\sigma_1)^2 + (d_2/\sigma_2)^2$



Counterparts in large catalogs

How can we do when the counterpart density is too large?

1. An often-used method is to consider the counterpart flux $\mathbf{S_k}$
2. The idea then is to define the source density only from those sources with flux no less than $\mathbf{S_k}$
   $\boldsymbol{\rho_k = N(S \geq S_k) / \Omega}$ (it should ideally be differential at $\mathbf{S_k}$, not so easy to obtain)
3. The likelihood ratio can then be expressed, replacing $\boldsymbol{\rho}$ by $\boldsymbol{\rho_k}$, except that we will now check all possible pairs, $\mathbf{LR_{ik} = exp(-(r_{ik}/\sigma_i)^2/2) / (2\pi\rho_k\sigma_i^2)}$ (no nearest neighbor term in exp)
4. We will then consider for each γ-ray source the counterpart with the largest likelihood ratio instead of the nearest neighbor: $\mathbf{LR_i = max_k\ LR_{ik}}$
5. It is not so easy to define a global a priori probability $\mathbf{Pr\{H_0\}}$

Reliability in likelihood ratio method

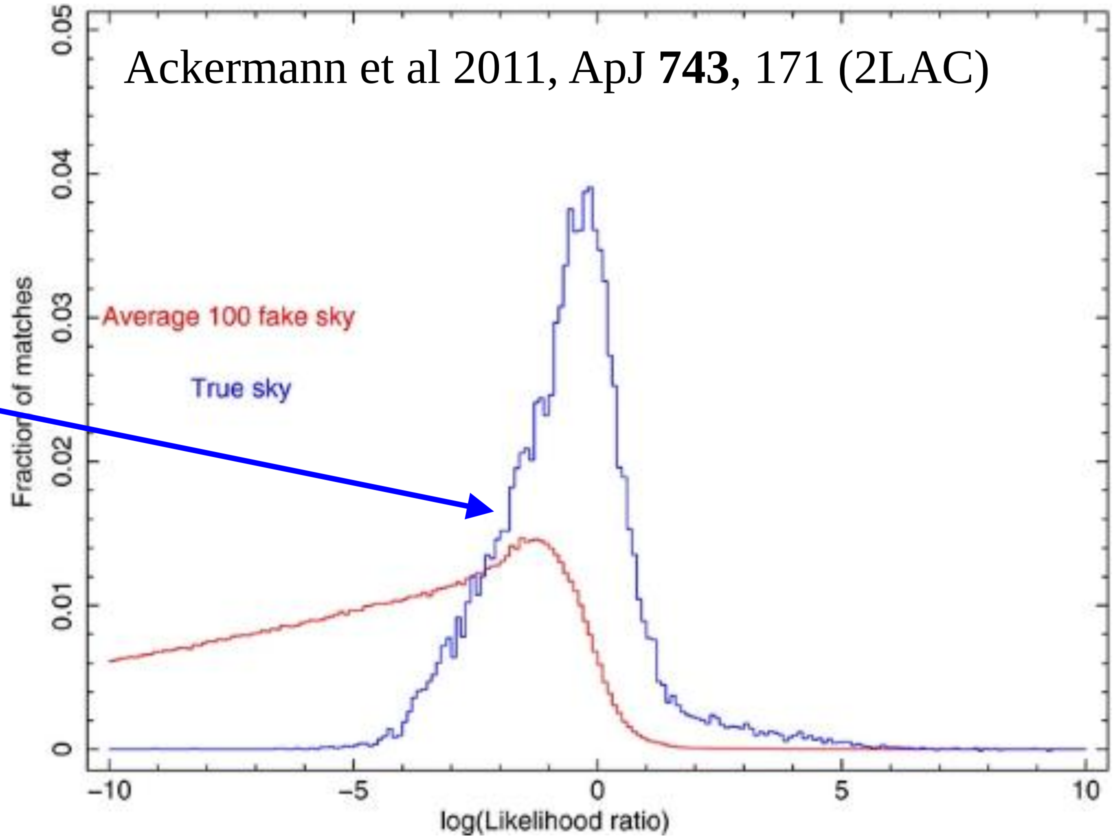
How can we go further?

1. We can try to estimate the distribution of LR under H_0 by simulating γ-ray sources randomly (but with a similar spatial distribution)

2. It is then possible to obtain from the true and the random **LR** distributions a reliability. Noting $\mathbf{N_T}$ and $\mathbf{N_R}$ the numbers of sources in a given **log(LR)** bin we define

$$\mathrm{R}(LR) = \frac{N_T(\mathrm{LR})}{N_T(\mathrm{LR})+N_R(LR)}$$

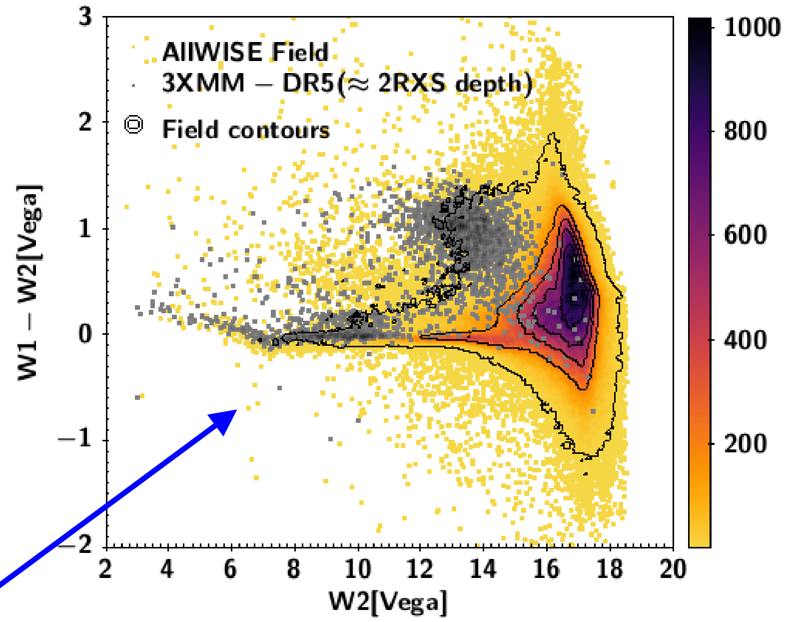
3. This is an approximate probability of association. It is noisy (because $\mathbf{N_T}$ is noisy) so it must be approximated by some analytic function.



Adding other criteria

Can we use other counterpart characteristics?

1. Yes. In the likelihood ratio method we can multiply the spatial term **S** by other terms (other data)
2. In the Bayesian formalism, **Pr{D|M} = Pr{S|M} Pr{C|M}** (actually probability densities)
3. We must know the distributions of the secondary quantity **C** under $\mathbf{H_0}$ and $\mathbf{H_1}$
4. The distribution of **C** under $\mathbf{H_0}$ is taken from that of the full counterpart catalog
5. The distribution of **C** under $\mathbf{H_1}$ is taken from a subset of "sure" identifications (not so easy)
6. Can be easily generalized to multiple characteristics (beware of correlations)
7. Examples: flux, local counterpart spectrum (eg radio spectral index, **optical colors**), hardness ratio wrt γ-rays (eg L_X/L_γ), variability, mass (stars), distance (→ luminosity)



Salvato et al 2018 (ROSAT)

Complications

Modern association tools

1. The localization precision of the counterparts must be accounted for (symmetric formulation)
2. Several counterpart catalogs must be handled together
3. The counterparts must be associated between themselves (in general they are better localized so we know whether an X-ray source is the same as an optical source or not), leading to the possibility of using the **global SED** as another criterion
4. Some sources can be absent at a particular wavelength simply because this source class emits little there (eg pulsars in the optical)

Implemented in the NWAY package by Mara Salvato et al (developed for eROSITA)

Handling Galactic sources

The Galactic plane is much more complex

1. The number of potential classes of emitters is much larger
2. The density of counterparts varies (latitude, distance to GC, spiral arms) so it must be estimated locally, for example via kernel density estimation
3. When the density of counterparts becomes too large locally (as explained before), it can become advantageous to consider them **collectively** (for example, star-forming regions rather than individual young stars, globular clusters rather than individual MSPs)
4. Galactic absorption/extinction biases the counterpart catalogs at many wavelengths (soft X-rays, optical/UV) whereas γ-ray sources are negligibly absorbed
5. Because of all that, care must be taken when simulating fake γ-ray sources to preserve their spatial distribution

Handling extended sources

Extended sources cannot fit into this probabilistic framework

1. Extended here means the radius **R is larger than the localization error** (not only larger than the PSF). Can be the same source even though the centroid is a little off, because the relative weights of emission regions inside the extended source can differ at different wavelengths

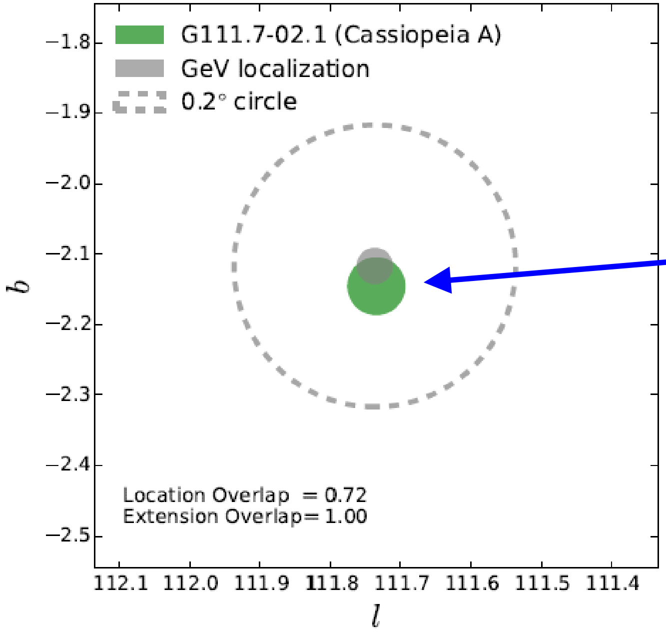
2. Many more parameters come into play: counterpart size (PWN, SNR) $\mathbf{R_{ctpt}}$, the γ-ray size (when it can be measured) $\mathbf{R_{gam}}$, the γ-ray localization (as before) $\mathbf{R_{95}}$ (95%)

3. **Majority** of Galactic TeV sources, unfortunately

4. When you find an extended source, look at images first! You can get images of the sky at many wavelengths from NASA's SkyView.

5. The Manitoba catalog (Ferrand & Safi-Harb 2012, AdSpR **49**, 1313) is a good resource to look into for SNRs and PWNe

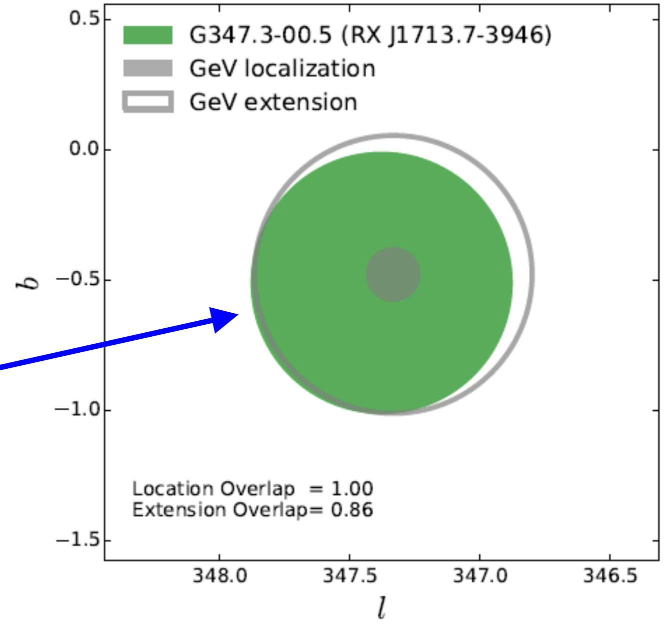
Handling extended sources 2



Can we attempt to quantify this anyway?

1. Yes, sort of, if you assume they are round
2. Location: $O_{\mathrm{loc}} = \frac{S_{\mathrm{ctpt}} \cap S_{95}}{\min(S_{\mathrm{ctpt}}, S_{95})} > O_{\min}$
3. Extension: $O_{\mathrm{ext}} = \frac{S_{\mathrm{ctpt}} \cap S_{\mathrm{gam}}}{\min(S_{\mathrm{ctpt}}, S_{\mathrm{gam}})} > O_{\min}$
4. $\mathbf{O_{min}}$ set to 0.5 or so
5. Estimate reliability from simulations

Acero et al 2016, ApJS **224**, 8 (SNRCat)



References and credits

Credits:

Jürgen Knödlseder (IRAP Toulouse) who created the Bayesian associations machinery for Fermi-LAT
Benoit Lott (LP2i Bordeaux) who took up the task in 2015
Dario Gasparrini (SSDC Roma) who handles the Likelihood Ratio machinery for Fermi-LAT

References:

De Ruiter & Willis 1977 (A&AS **28**, 211) in the radio
Sutherland & Saunders 1992 (MNRAS **259**, 413) for Likelihood Ratio
Mattox et al 1997 (ApJ **481**, 95) for EGRET
Rutledge et al 2000 (ApJS **131**, 335) for ROSAT
Budavari & Szalay 2008 (ApJ **679**, 301) for combining probabilities over multiple catalogs
Pineau et al 2017 (A&A 597, A89) for missing entries in some catalogs
Salvato et al 2018 (MNRAS 473, 4937) introducing NWAY for eROSITA

Hands-on session

Open the folder called Source association under the school's sharepoint directory

Read the README file and do as explained

Conclusions

1. Well established framework for point sources
2. Statistical estimate of false and missed associations
3. Can accommodate source density, flux, other quantities
4. Extended sources more uncertain