The ARCHES cross-correlation tool

François-Xavier Pineau[1]

[1]Observatoire Astronomique de Strasbourg, Université de Strasbourg, CNRS

Paris, 1th December, 2015

INTRODUCTION

- This talk: "*Cross-correlation tool development & catalogue creation*" (WP4)
- Aims of ARCHES's WP4:
  - Create a public *n*-catalogues cross-correlation tool:
    - ★ No magic BUT a flexible/multi-purpose/scriptable multi-catalogue xmatch engine
    - ★ Usable as a building block from you own specific code
  - Use/develop statistical methods to compute probabilities of associations:
    - ★ Astrometry based probabilities only!
    - ★ Can be combined with photometry based probabilities (in a further step)
  - Use the tool to build ARCHES catalogue(s)
- Beyond the ARCHES project:
  - tool will be part of the CDS XMatch Service
  - $\Rightarrow$ will be maintained, will keep evolving

INTRODUCTION

- This talk: "*Cross-correlation tool development & catalogue creation*" (WP4)
- Aims of ARCHES's WP4:
  - Create a public *n*-catalogues cross-correlation tool:
    - ★ No magic BUT a flexible/multi-purpose/scriptable multi-catalogue xmatch engine
    - ★ Usable as a building block from you own specific code
  - Use/develop statistical methods to compute probabilities of associations:
    - ★ Astrometry based probabilities only!
    - ★ Can be combined with photometry based probabilities (in a further step)
  - Use the tool to build ARCHES catalogue(s)
- Beyond the ARCHES project:
  - tool will be part of the CDS XMatch Service
  - $\Rightarrow$ will be maintained, will keep evolving

INTRODUCTION

- This talk is mainly focused on the probabilistic part
- More details on the tool during the *Hands on* session

METHOD

- Steps to probabilistic positional xmatch
  - Make simplifying assumptions
  - Select candidates: select and group together sources possibly being various detections of a same real source
    - ⋆ Need for a selection criterion
  - Make hypothesis: are the sources really from a same real sources or from different real sources?
  - For each hypothesis:
    - ⋆ derive the associated *likelihood*
    - ⋆ derive the associated *prior*
  - Compute astrometry based probabilities

SIMPLIFYING ASSUMPTIONS

- Radical simplifying assumptions:
  - No proper motions
  - No blending
  - No clustering (density of sources = Poisson law)
  - No systematic offsets
  - You can trust positional uncertainties provided in catalogues

CANDIDATE SELECTION

Candidate selection criterion

How to select a group of n sources from n distinct catalogues as possibly being various observations of a same actual source?

- *Statistical hypothesis testing*
  - H_0 (null hypothesis): all n sources are from the same real source
  - $H_1 = \bar{H}_0$ (alternative hypothesis): at least one source (out of n) is spurious
- User input: γ, the probability to accept H_0 while it is true
  - γ (I call it completeness) is called *true negative* rate
  - we usually fix $\gamma = 0.9973$ (99.73%, value of the 3σ rule in 1 dimensional pb)
  - $\Leftrightarrow$ fixing the *type I error* $= 0.027\%$ = proba to reject null hypothesis while it is true
  - we (theoretically) miss 27/10 000 real association
- The criterion used is based on a χ^2 test of $2(n-1)$ degrees of freedom
- Now, a few slides to explain it since it plays a role in probabilities

CANDIDATE SELECTION

Candidate selection criterion

How to select a group of n sources from n distinct catalogues as possibly being various observations of a same actual source?

- *Statistical hypothesis testing*
  - H_0 (null hypothesis): all n sources are from the same real source
  - $H_1 = \bar{H}_0$ (alternative hypothesis): at least one source (out of n) is spurious
- User input: γ, the probability to accept H_0 while it is true
  - γ (I call it completeness) is called *true negative* rate
  - we usually fix $\gamma = 0.9973$ (99.73%, value of the 3σ rule in 1 dimensional pb)
  - $\Leftrightarrow$ fixing the *type I error* $= 0.027\% =$ proba to reject null hypothesis while it is true
  - we (theoretically) miss 27/10 000 real association
- The criterion used is based on a χ^2 test of $2(n-1)$ degrees of freedom
- Now, a few slides to explain it since it plays a role in probabilities

CANDIDATE SELECTION

Candidate selection criterion

How to select a group of n sources from n distinct catalogues as possibly being various observations of a same actual source?

- *Statistical hypothesis testing*
  - H_0 (null hypothesis): all n sources are from the same real source
  - $H_1 = \bar{H}_0$ (alternative hypothesis): at least one source (out of n) is spurious
- User input: γ, the probability to accept H_0 while it is true
  - γ (I call it completeness) is called *true negative* rate
  - we usually fix $\gamma = 0.9973$ (99.73%, value of the 3σ rule in 1 dimensional pb)
  - $\Leftrightarrow$ fixing the *type I error* $= 0.027\% =$ proba to reject null hypothesis while it is true
  - we (theoretically) miss 27/10 000 real association
- The criterion used is based on a χ^2 test of $2(n-1)$ degrees of freedom
- Now, a few slides to explain it since it plays a role in probabilities

CANDIDATE SELECTION

Classical 2 catalogues case

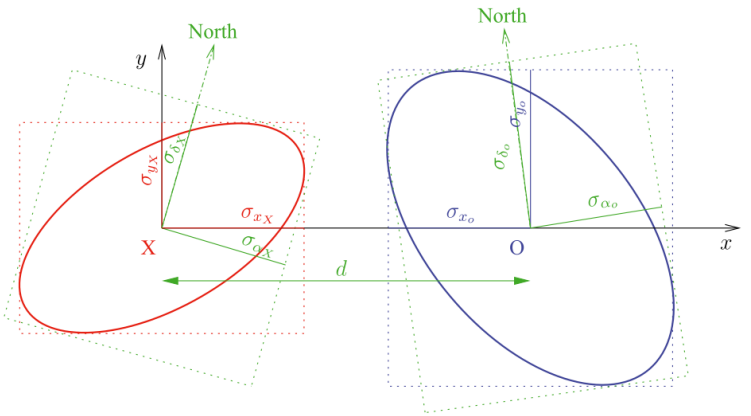
- In the classical case (e.g. De Ruiter et al. 1977):
  - Errors are independant on α and δ
  - Source 1 has errors σ_{α_1} and σ_{δ_1} on α and δ respectively
  - Source 2 has errors σ_{α_2} and σ_{δ_2} on α and δ respectively
  - The normalized distance (or σ-distance) is defined by:

$$r = \left[\frac{\Delta\alpha^2}{\sigma_{\alpha_1}^2 + \sigma_{\alpha_2}^2} + \frac{\Delta\delta^2}{\sigma_{\delta_1}^2 + \sigma_{\delta_2}^2} \right]^{1/2}$$

CANDIDATE SELECTION

Classical 2 catalogues case

- More generally (see e.g. Pineau et al. 2011)
  - We assimilate locally the surface of the sphere to the Euclidian plane
  - The positions of the 2 sources are 2 dimentional vectors: $\vec{\mu_1}$ and $\vec{\mu_2}$.
  - Errors on $\vec{\mu_1}$ and $\vec{\mu_2}$ are oriented ellipses defined by covariance matrices $\mathbf{V_1}$ and $\mathbf{V_2}$ respectively:



  - The normallized distance becomes (vectorial form):

$$r = \left[(\vec{\mu_1} - \vec{\mu_2})^T (\mathbf{V_1} + \mathbf{V_2})^{-1} (\vec{\mu_1} - \vec{\mu_2}) \right]^{1/2}$$

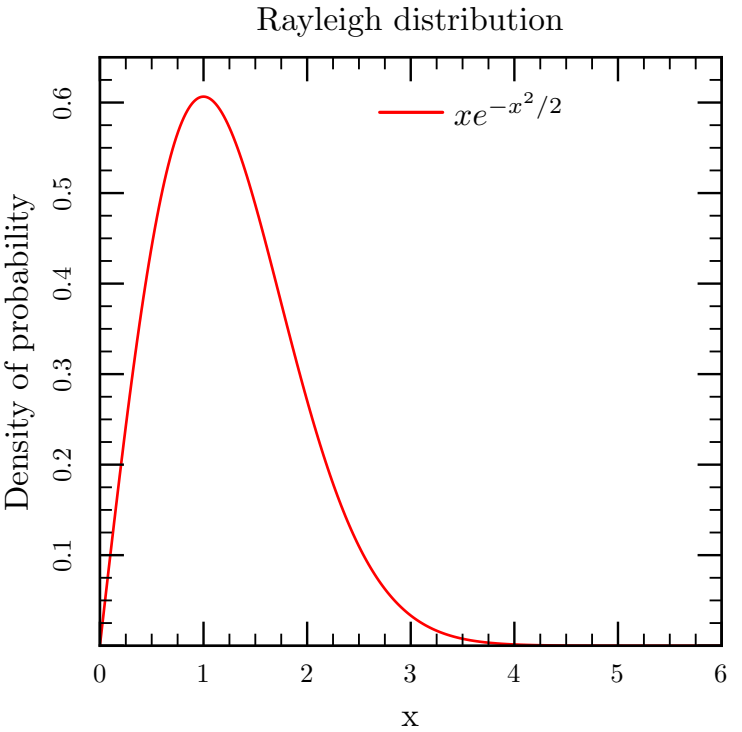
  - $\Rightarrow$ equation of an ellipse of radius r and covariance matrix $\mathbf{V_1} + \mathbf{V_2}$

CANDIDATE SELECTION

Classical 2 catalogues case

- For real associations, i.e. when H_0 is true
  - The distribution of normalized distances is a Rayleigh distribution of scale $\sigma = 1$

$$r \overset{H_0}{\sim} Rayleigh$$



CANDIDATE SELECTION

Classical 2 catalogues case

- Fixing the completeness $\gamma \Leftrightarrow$ fixing a normalized distance threshold k_γ:

$$\int_0^{k_\gamma} \mathit{Rayleigh}(r)\mathrm{d}r = \gamma$$

- For $\gamma = 99.73\%$ (the 1D 3σ rule) $\Rightarrow k_\gamma = 3.4395$ (not 3!)
- So, for 2 sources from 2 distinct catalogues, the selection criterion is

$$\left[(\vec{\mu_1} - \vec{\mu_2})^T (\mathbf{V_1} + \mathbf{V_2})^{-1} (\vec{\mu_1} - \vec{\mu_2})\right]^{1/2} \leq k_\gamma$$

- I.e. source 2 kept as candidate if it is inside an error ellipse of covariance matrix $\mathbf{V} = \mathbf{V_1} + \mathbf{V_2}$ and of radius k_γ, centered around source 1.
- $\Rightarrow$ the surface area of the acceptance region is $|\mathbf{V_1} + \mathbf{V_2}|^{1/2} \pi k_\gamma^2$

CANDIDATE SELECTION

Classical 2 catalogues case

- Fixing the completeness $\gamma \Leftrightarrow$ fixing a normalized distance threshold k_γ:

$$\int_0^{k_\gamma} \mathit{Rayleigh}(r)\mathrm{d}r = \gamma$$

- For $\gamma = 99.73\%$ (the 1D 3σ rule) $\Rightarrow k_\gamma = 3.4395$ (not 3!)
- So, for 2 sources from 2 distinct catalogues, the selection criterion is

$$\left[(\vec{\mu_1} - \vec{\mu_2})^T (\mathbf{V_1} + \mathbf{V_2})^{-1} (\vec{\mu_1} - \vec{\mu_2})\right]^{1/2} \leq k_\gamma$$

- I.e. source 2 kept as candidate if it is inside an error ellipse of covariance matrix $\mathbf{V} = \mathbf{V_1} + \mathbf{V_2}$ and of radius k_γ, centered around source 1.
- $\Rightarrow$ the surface area of the acceptance region is $|\mathbf{V_1} + \mathbf{V_2}|^{1/2} \pi k_\gamma^2$

CANDIDATE SELECTION

Classical 2 catalogues case

- Fixing the completeness $\gamma \Leftrightarrow$ fixing a normalized distance threshold k_γ:

$$\int_0^{k_\gamma} \mathit{Rayleigh}(r)\mathrm{d}r = \gamma$$

- For $\gamma = 99.73\%$ (the 1D 3σ rule) $\Rightarrow k_\gamma = 3.4395$ (not 3!)
- So, for 2 sources from 2 distinct catalogues, the selection criterion is

$$\left[(\vec{\mu_1} - \vec{\mu_2})^T (\mathbf{V_1} + \mathbf{V_2})^{-1} (\vec{\mu_1} - \vec{\mu_2})\right]^{1/2} \leq k_\gamma$$

- I.e. source 2 kept as candidate if it is inside an error ellipse of covariance matrix $\mathbf{V} = \mathbf{V_1} + \mathbf{V_2}$ and of radius k_γ, centered around source 1.
- $\Rightarrow$ the surface area of the acceptance region is $|\mathbf{V_1} + \mathbf{V_2}|^{1/2} \pi k_\gamma^2$

CANDIDATE SELECTION

Now, a different version of the same story more easily generalisable to *n*-catalogues.

CANDIDATE SELECTION

Revisited 2 catalogues case

- I have 2 sources from 2 distinct catalogues, I suppose H_0 is *true*
- Maximum Likelihood Estimate (MLE) of the position of the real source $\rightsquigarrow$ the weighted mean position

$$\vec{\mu_\Sigma} = \mathbf{V_\Sigma}(\mathbf{V_1}^{-1}\vec{\mu_1} + \mathbf{V_2}^{-1}\vec{\mu_2})$$

in which

$$\mathbf{V_\Sigma} = (\mathbf{V_1}^{-1} + \mathbf{V_2}^{-1})^{-1}$$

- The error on this MLE is ... $\mathbf{V_\Sigma}$
- The result is the same with a (by block) Weighted Least Squares method
- We can now define the Mahalanobis distance:

$$D_M = \left[\sum_{i=1}^{2}(\vec{\mu_i} - \vec{\mu_\Sigma})^T \mathbf{V_i}^{-1}(\vec{\mu_i} - \vec{\mu_\Sigma})\right]^{1/2} \overset{H_0}{\sim} \chi_{dof=2}$$

CANDIDATE SELECTION

Revisited 2 catalogues case

- I have 2 sources from 2 distinct catalogues, I suppose H_0 is *true*
- Maximum Likelihood Estimate (MLE) of the position of the real source $\rightsquigarrow$ the weighted mean position

$$\vec{\mu_\Sigma} = \mathbf{V_\Sigma}(\mathbf{V_1}^{-1}\vec{\mu_1} + \mathbf{V_2}^{-1}\vec{\mu_2})$$

in which

$$\mathbf{V_\Sigma} = (\mathbf{V_1}^{-1} + \mathbf{V_2}^{-1})^{-1}$$

- The error on this MLE is ... $\mathbf{V_\Sigma}$
- The result is the same with a (by block) Weighted Least Squares method
- We can now define the Mahalanobis distance:

$$D_M = \left[\sum_{i=1}^{2}(\vec{\mu_i} - \vec{\mu_\Sigma})^T \mathbf{V_i}^{-1}(\vec{\mu_i} - \vec{\mu_\Sigma})\right]^{1/2} \overset{H_0}{\sim} \chi_{dof=2}$$

CANDIDATE SELECTION

Revisited 2 catalogues case

- I have 2 sources from 2 distinct catalogues, I suppose H_0 is *true*
- Maximum Likelihood Estimate (MLE) of the position of the real source $\rightsquigarrow$ the weighted mean position

$$\vec{\mu_\Sigma} = \mathbf{V_\Sigma}(\mathbf{V_1}^{-1}\vec{\mu_1} + \mathbf{V_2}^{-1}\vec{\mu_2})$$

in which

$$\mathbf{V_\Sigma} = (\mathbf{V_1}^{-1} + \mathbf{V_2}^{-1})^{-1}$$

- The error on this MLE is ... $\mathbf{V_\Sigma}$
- The result is the same with a (by block) Weighted Least Squares method
- We can now define the Mahalanobis distance:

$$D_M = \left[\sum_{i=1}^{2}(\vec{\mu_i} - \vec{\mu_\Sigma})^T \mathbf{V_i}^{-1}(\vec{\mu_i} - \vec{\mu_\Sigma})\right]^{1/2} \overset{H_0}{\sim} \chi_{dof=2}$$

CANDIDATE SELECTION

Let's merge the two approaches.

CANDIDATE SELECTION

Classical versus Revisited 2 catalogues case

Doing the math, we find

- the equality

$$|\mathbf{V_1} + \mathbf{V_2}| = \frac{|\mathbf{V_1}||\mathbf{V_2}|}{|\mathbf{V_\Sigma}|}$$

- the normalized (or σ) distance = Mahalanobis distance

$$(\vec{\mu_1} - \vec{\mu_2})^T(\mathbf{V_1} + \mathbf{V_2})^{-1}(\vec{\mu_1} - \vec{\mu_2}) = \sum_{i=1}^{2}(\vec{\mu_i} - \vec{\mu_\Sigma})^T\mathbf{V_i}^{-1}(\vec{\mu_i} - \vec{\mu_\Sigma})$$

- Remark: not surprising when we known that *Rayleigh* = $\chi_{dof=2}$

CANDIDATE SELECTION

Classical versus Revisited 2 catalogues case

Conlusion on the 2 catalogues case

- Selection criterion

$$\boxed{D_M \leq k_\gamma}$$

- k_γ defined such as

$$\int_0^{k_\gamma} \chi_{dof=2}(x)\mathrm{d}x = \gamma$$

- $\Rightarrow$ Region of acceptance of surface

$$S_{1,2} = \left[\frac{|\mathbf{V_1}||\mathbf{V_2}|}{|\mathbf{V_\Sigma}|}\right]^{1/2} \pi k_\gamma^2$$

CANDIDATE SELECTION

Generalization to n catalogues

We easily generalize to n catalogues:

- Selection criterion

$$D_M = \left[\sum_{i=1}^{n} (\vec{\mu_i} - \vec{\mu_\Sigma})^T \mathbf{V_i^{-1}} (\vec{\mu_i} - \vec{\mu_\Sigma})\right]^{1/2} < k_\gamma$$

- Now

$$D_M \overset{H_0}{\sim} \chi_{dof=2(n-1)}$$

- or, equivalently $D_M^2 \overset{H_0}{\sim} \chi^2_{dof=2(n-1)}$
- So k_γ now defined such as

$$\int_0^{k_\gamma} \chi_{dof=2(n-1)}(x)\mathrm{d}x = \gamma$$

CANDIDATE SELECTION

Generalization to n catalogues

- Region of acceptance of volume

$$S_{1,2} = \left[\prod_{i=1}^{n} |\mathbf{V_i}| / |\mathbf{V_\Sigma}| \right]^{1/2} V_{2(n+1)}(k)$$

with $V_{2(n+1)}(k)$: volume of the $2(n-1)$-sphere of radius k

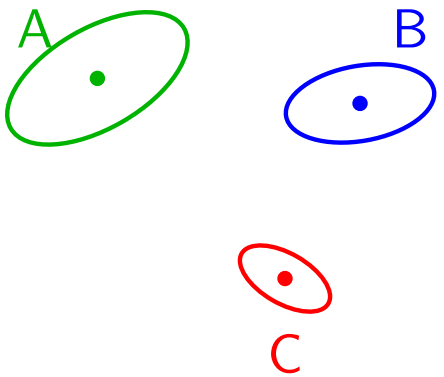
CANDIDATE SELECTION

Generalization to *n* catalogues: iterative form

- $\chi^2_{dof=2(n-1)} \Leftrightarrow (n-1)\chi^2_{dof=2}$
- $\Rightarrow$ we can perform (n-1) successives and iteratives xmatches:

$$D_M = \left[\sum_{i=2}^{n} (\mu_{\vec{\Sigma}_{i-1}} - \vec{\mu_i})^T (\mathbf{V_{\Sigma_{i-1}}} + \mathbf{V_i})^{-1} (\mu_{\vec{\Sigma}_{i-1}} - \vec{\mu_i}) \right]^{1/2}$$

  - $\mu_{\vec{\Sigma}_{i-1}}$: the weighted mean position of the previous xmatch
  - $\mathbf{V_{\Sigma_{i-1}}}$: the error on the weighted mean position of the previous xmatch
- Result independant of the xmatch order (for INNER JOIN!)



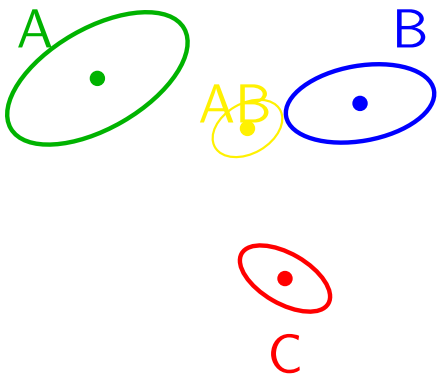
CANDIDATE SELECTION

Generalization to *n* catalogues: iterative form

- $\chi^2_{dof=2(n-1)} \Leftrightarrow (n-1)\chi^2_{dof=2}$
- $\Rightarrow$ we can perform (n-1) successives and iteratives xmatches:

$$D_M = \left[\sum_{i=2}^{n} (\mu\vec{\Sigma}_{i-1} - \vec{\mu_i})^T (\mathbf{V_{\Sigma_{i-1}}} + \mathbf{V_i})^{-1} (\mu\vec{\Sigma}_{i-1} - \vec{\mu_i}) \right]^{1/2}$$

  - $\mu\vec{\Sigma}_{i-1}$: the weighted mean position of the previous xmatch
  - $\mathbf{V_{\Sigma_{i-1}}}$: the error on the weighted mean position of the previous xmatch
- Result independant of the xmatch order (for INNER JOIN!)



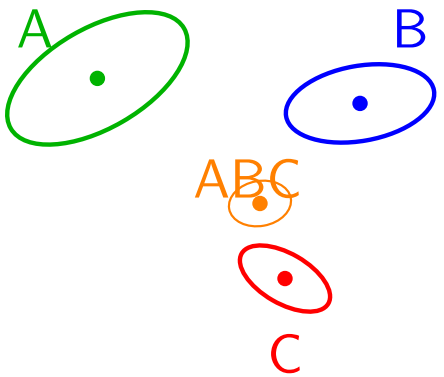
CANDIDATE SELECTION

Generalization to *n* catalogues: iterative form

- $\chi^2_{dof=2(n-1)} \Leftrightarrow (n-1)\chi^2_{dof=2}$
- $\Rightarrow$ we can perform (n-1) successives and iteratives xmatches:

$$D_M = \left[\sum_{i=2}^{n} (\mu_{\vec{\Sigma}_{i-1}} - \vec{\mu_i})^T (\mathbf{V_{\Sigma_{i-1}}} + \mathbf{V_i})^{-1} (\mu_{\vec{\Sigma}_{i-1}} - \vec{\mu_i}) \right]^{1/2}$$

  - $\mu_{\vec{\Sigma}_{i-1}}$: the weighted mean position of the previous xmatch
  - $\mathbf{V_{\Sigma_{i-1}}}$: the error on the weighted mean position of the previous xmatch
- Result independant of the xmatch order (for INNER JOIN!)



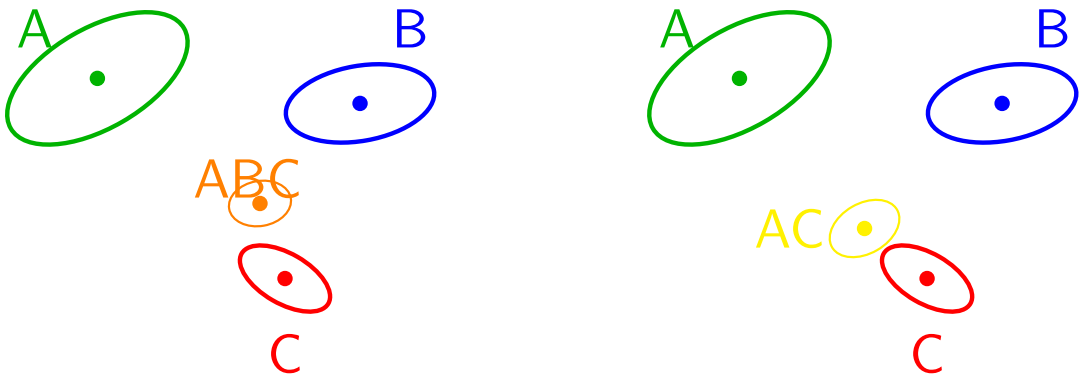
CANDIDATE SELECTION

Generalization to *n* catalogues: iterative form

- $\chi_{dof=2(n-1)} \Leftrightarrow (n-1)\chi_{dof=2}$
- $\Rightarrow$ we can perform (n-1) successives and iteratives xmatches:

$$D_M = \left[\sum_{i=2}^{n} (\mu\vec{\Sigma}_{i-1} - \vec{\mu_i})^T (\mathbf{V_{\Sigma_{i-1}}} + \mathbf{V_i})^{-1} (\mu\vec{\Sigma}_{i-1} - \vec{\mu_i})\right]^{1/2}$$

  - $\mu\vec{\Sigma}_{i-1}$: the weighted mean position of the previous xmatch
  - $\mathbf{V_{\Sigma_{i-1}}}$: the error on the weighted mean position of the previous xmatch
- Result independant of the xmatch order (for INNER JOIN!)



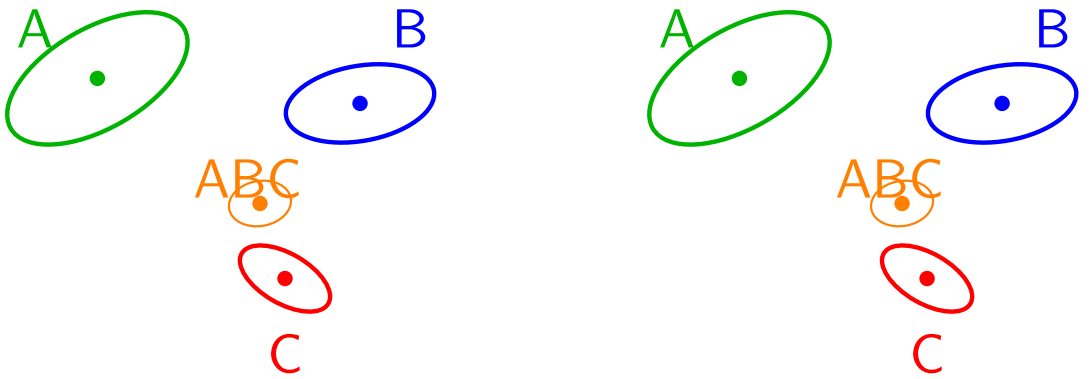
CANDIDATE SELECTION

Generalization to *n* catalogues: iterative form

- $\chi^2_{dof=2(n-1)} \Leftrightarrow (n-1)\chi^2_{dof=2}$
- $\Rightarrow$ we can perform (n-1) successives and iteratives xmatches:

$$D_M = \left[\sum_{i=2}^{n} (\mu\vec{\Sigma}_{i-1} - \vec{\mu_i})^T (\mathbf{V_{\Sigma_{i-1}}} + \mathbf{V_i})^{-1} (\mu\vec{\Sigma}_{i-1} - \vec{\mu_i}) \right]^{1/2}$$

  - $\mu\vec{\Sigma}_{i-1}$: the weighted mean position of the previous xmatch
  - $\mathbf{V_{\Sigma_{i-1}}}$: the error on the weighted mean position of the previous xmatch
- Result independant of the xmatch order (for INNER JOIN!)



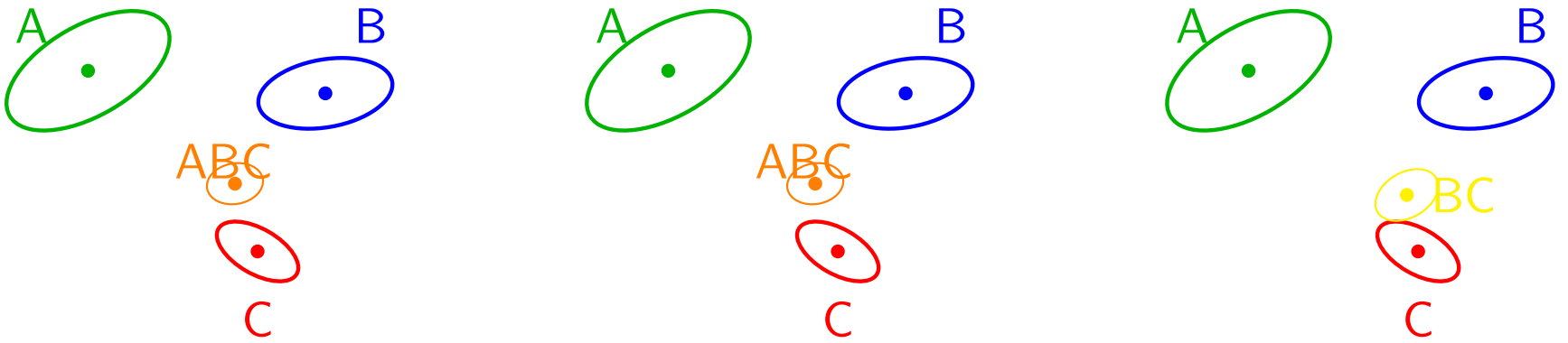
CANDIDATE SELECTION

Generalization to *n* catalogues: iterative form

- $\chi_{dof=2(n-1)} \Leftrightarrow (n-1)\chi_{dof=2}$
- $\Rightarrow$ we can perform (n-1) successives and iteratives xmatches:

$$D_M = \left[\sum_{i=2}^{n} (\mu_{\vec{\Sigma}_{i-1}} - \vec{\mu_i})^T (\mathbf{V_{\Sigma_{i-1}}} + \mathbf{V_i})^{-1} (\mu_{\vec{\Sigma}_{i-1}} - \vec{\mu_i})\right]^{1/2}$$

  - $\mu_{\vec{\Sigma}_{i-1}}$: the weighted mean position of the previous xmatch
  - $\mathbf{V_{\Sigma_{i-1}}}$: the error on the weighted mean position of the previous xmatch
- Result independant of the xmatch order (for INNER JOIN!)



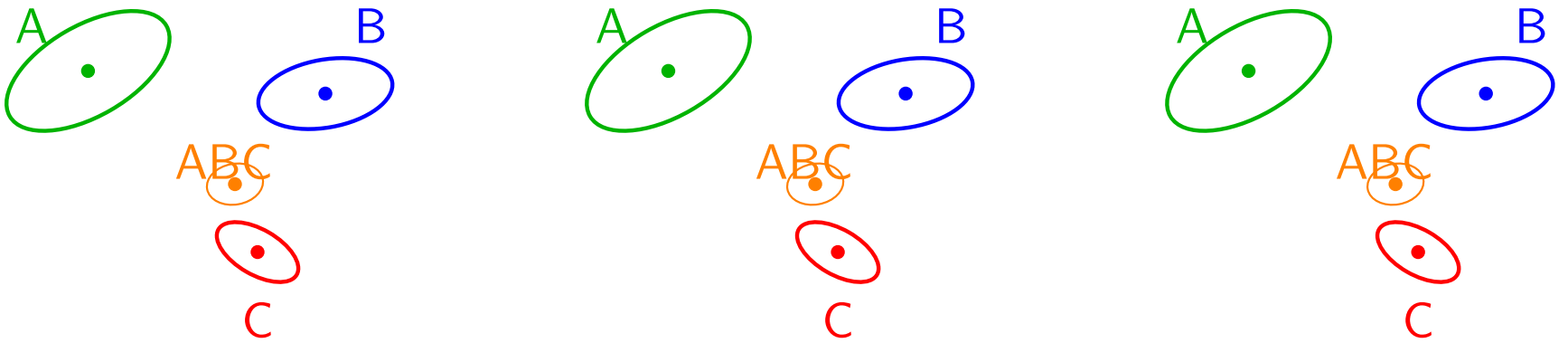
CANDIDATE SELECTION

Generalization to *n* catalogues: iterative form

- $\chi_{dof=2(n-1)} \Leftrightarrow (n-1)\chi_{dof=2}$
- $\Rightarrow$ we can perform (n-1) successives and iteratives xmatches:

$$D_M = \left[\sum_{i=2}^{n} (\mu_{\vec{\Sigma}_{i-1}} - \vec{\mu_i})^T (\mathbf{V_{\Sigma_{i-1}}} + \mathbf{V_i})^{-1} (\mu_{\vec{\Sigma}_{i-1}} - \vec{\mu_i}) \right]^{1/2}$$

  - $\mu_{\vec{\Sigma}_{i-1}}$: the weighted mean position of the previous xmatch
  - $\mathbf{V_{\Sigma_{i-1}}}$: the error on the weighted mean position of the previous xmatch
- Result independant of the xmatch order (for INNER JOIN!)



HYPOTHESES

For 2 and 3 catalogues

To compute Bayes probabilities, we MUST consider all possible hypothesis.

- Law of total probabilities:

$$\sum_{i=1}^{n} p(H_i) = 1$$

- For 2 catalogues
  - 2 hypothesis
    - AB (H_0)
    - A_B

- For 3 catalogues
  - 5 hypothesis

HYPOTHESES

For 2 and 3 catalogues

To compute Bayes probabilities, we MUST consider all possible hypothesis.

- Law of total probabilities:

$$\sum_{i=1}^{n} p(H_i) = 1$$

- For 2 catalogues
  - 2 hypothesis
    - AB (H_0)
    - A_B

- For 3 catalogues
  - 5 hypothesis

HYPOTHESES

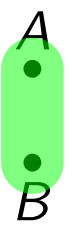
For 2 and 3 catalogues

To compute Bayes probabilities, we MUST consider all possible hypothesis.

- Law of total probabilities:

$$\sum_{i=1}^{n} p(H_i) = 1$$

- For 2 catalogues
  - 2 hypothesis
    - AB (H_0)
    - A_B



- For 3 catalogues
  - 5 hypothesis

HYPOTHESES

For 2 and 3 catalogues

To compute Bayes probabilities, we MUST consider all possible hypothesis.

- Law of total probabilities:

$$\sum_{i=1}^{n} p(H_i) = 1$$

- For 2 catalogues
  - 2 hypothesis
    - ⋆ AB (H_0)
    - ⋆ A_B

- For 3 catalogues
  - 5 hypothesis
    - ⋆ ABC (H_0)
    - ⋆ AB_C
    - ⋆ A_BC
    - ⋆ AC_B
    - ⋆ A_B_C

HYPOTHESES

For 2 and 3 catalogues

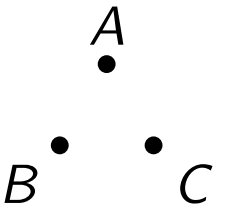
To compute Bayes probabilities, we MUST consider all possible hypothesis.

- Law of total probabilities:

$$\sum_{i=1}^{n} p(H_i) = 1$$

- For 2 catalogues
  - 2 hypothesis
    - AB (H_0)
    - A_B

- For 3 catalogues
  - 5 hypothesis
    - ABC (H_0)
    - AB_C
    - AC_B
    - A_BC
    - A_B_C



HYPOTHESES

For 2 and 3 catalogues

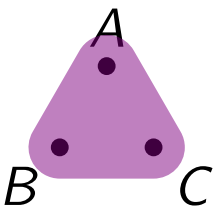
To compute Bayes probabilities, we MUST consider all possible hypothesis.

- Law of total probabilities:

$$\sum_{i=1}^{n} p(H_i) = 1$$

- For 2 catalogues
  - 2 hypothesis
    - AB (H_0)
    - A_B

- For 3 catalogues
  - 5 hypothesis
    - ABC (H_0)
    - AB_C
    - AC_B
    - A_BC
    - A_B_C



HYPOTHESES

For 2 and 3 catalogues

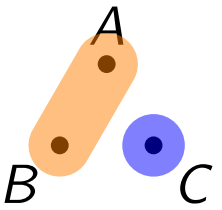
To compute Bayes probabilities, we MUST consider all possible hypothesis.

- Law of total probabilities:

$$\sum_{i=1}^{n} p(H_i) = 1$$

- For 2 catalogues
  - 2 hypothesis
    - ★ AB (H_0)
    - ★ A_B

- For 3 catalogues
  - 5 hypothesis
    - ★ ABC (H_0)
    - ★ AB_C
    - ★ AC_B
    - ★ A_BC
    - ★ A_B_C



HYPOTHESES

For 2 and 3 catalogues

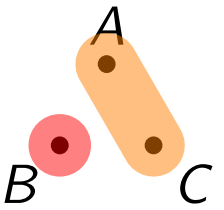
To compute Bayes probabilities, we MUST consider all possible hypothesis.

- Law of total probabilities:

$$\sum_{i=1}^{n} p(H_i) = 1$$

- For 2 catalogues
  - 2 hypothesis
    - ★ AB (H_0)
    - ★ A_B

- For 3 catalogues
  - 5 hypothesis
    - ★ ABC (H_0)
    - ★ AB_C
    - ★ AC_B
    - ★ A_BC
    - ★ A_B_C



HYPOTHESES

For 2 and 3 catalogues

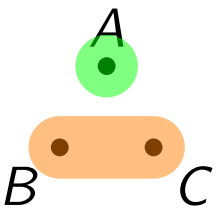
To compute Bayes probabilities, we MUST consider all possible hypothesis.

- Law of total probabilities:

$$\sum_{i=1}^{n} p(H_i) = 1$$

- For 2 catalogues
  - 2 hypothesis
    - AB (H_0)
    - A_B

- For 3 catalogues
  - 5 hypothesis
    - ABC (H_0)
    - AB_C
    - AC_B
    - A_BC
    - A_B_C



HYPOTHESES

For 2 and 3 catalogues

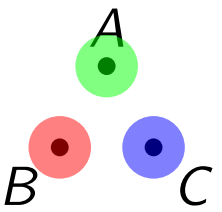
To compute Bayes probabilities, we MUST consider all possible hypothesis.

- Law of total probabilities:

$$\sum_{i=1}^{n} p(H_i) = 1$$

- For 2 catalogues
  - 2 hypothesis
    - ★ AB (H_0)
    - ★ A_B

- For 3 catalogues
  - 5 hypothesis
    - ★ ABC (H_0)
    - ★ AB_C
    - ★ AC_B
    - ★ A_BC
    - ★ A_B_C



HYPOTHESES

For n-catalogues

- We generalised for n catalogues
- The number of hypothesis to be tested is given by the BELL number

Table : Values of the seven first BELL numbers

n	2	3	4	5	6	7
B_n	2	5	15	52	203	877

  - n number of catalogues
  - n=5 catalogues $\rightsquigarrow$ 52 probabilities to be computed

- $\Rightarrow$ Combinatorial explosion!

HYPOTHESES

For n-catalogues

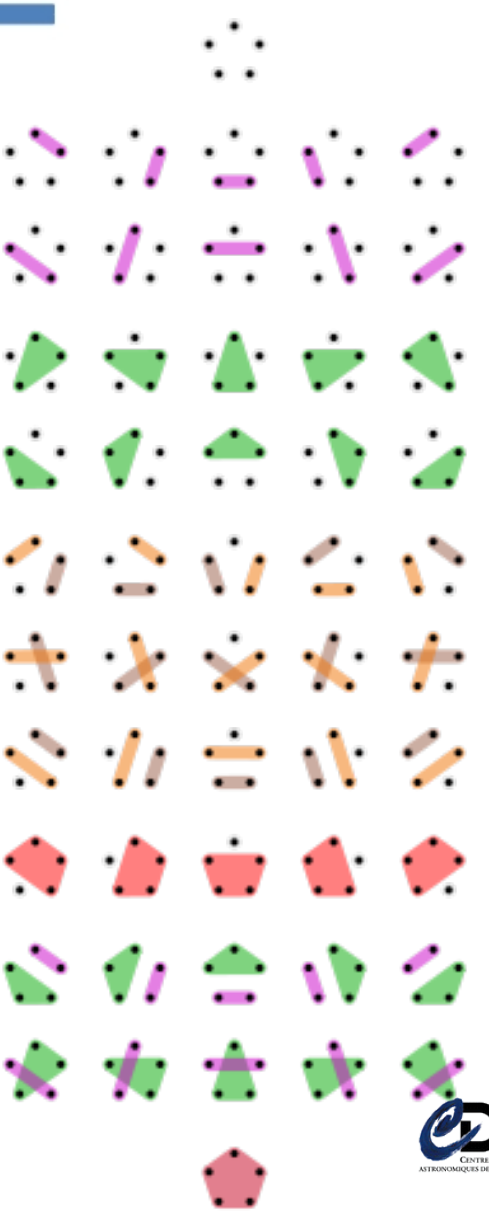
- We generalised for n catalogues
- The number of hypothesis to be tested is given by the BELL number

Table : Values of the seven first BELL numbers

n	2	3	4	5	6	7
B_n	2	5	15	52	203	877

  - n number of catalogues
  - n=5 catalogues $\rightsquigarrow$ 52 probabilities to be computed

- $\Rightarrow$ Combinatorial explosion!



Credits: wikipedia

HYPOTHESES

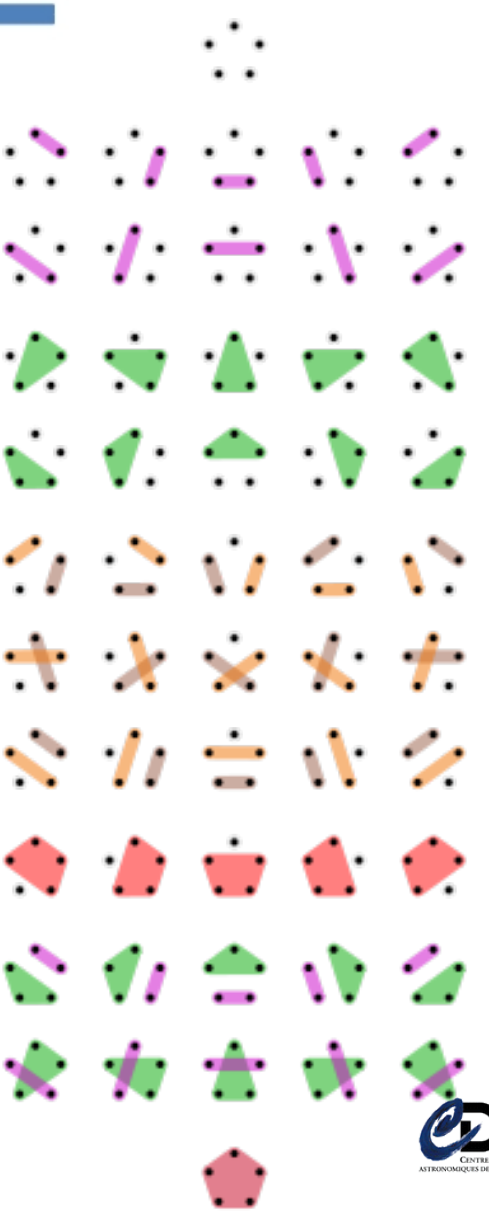
For n-catalogues

- We generalised for n catalogues
- The number of hypothesis to be tested is given by the BELL number

Table : Values of the seven first BELL numbers

n	2	3	4	5	6	7
B_n	2	5	15	52	203	877

  - n number of catalogues
  - n=5 catalogues $\rightsquigarrow$ 52 probabilities to be computed
- $\Rightarrow$ Combinatorial explosion!



Credits: wikipedia

BAYES PROBA

For 2 catalogues

- Let's call x the Mahalanobis distance D_M
- Imagine that
  - we xmatch 2 tables and we obtain n_{tot} associations
  - we know the number of spurious associations: n_{H_1}
  - $\Rightarrow$ we know the number of real associations: $n_{H_0} = n_{tot} - n_{H_1}$
- Distribution of x of real associations (likelihood $p(x|H_0)$)
  - almost a $\chi_{dof=2}(x)$, because...
  - ... its integrate over the acceptance domain must be $= 1$
  - $\Rightarrow p(x|H_0) = \chi_{dof=2}(x)/\gamma$
- Distribution of x of spurious associations (likelihood $p(x|H_1)$)
  - Poisson $\propto x$
  - again it must integrates to 1 over the domain of acceptance
  - $\Rightarrow p(x|H_0) = 2x/k_\gamma^2$

BAYES PROBA

For 2 catalogues

- Let's call x the Mahalanobis distance D_M
- Imagine that
  - we xmatch 2 tables and we obtain n_{tot} associations
  - we know the number of spurious associations: n_{H_1}
  - $\Rightarrow$ we know the number of real associations: $n_{H_0} = n_{tot} - n_{H_1}$
- Distribution of x of real associations (likelihood $p(x|H_0)$)
  - almost a $\chi_{dof=2}(x)$, because...
  - ... its integrate over the acceptance domain must be $= 1$
  - $\Rightarrow p(x|H_0) = \chi_{dof=2}(x)/\gamma$
- Distribution of x of spurious associations (likelihood $p(x|H_1)$)
  - Poisson $\propto x$
  - again it must integrates to 1 over the domain of acceptance
  - $\Rightarrow p(x|H_0) = 2x/k_\gamma^2$

BAYES PROBA

For 2 catalogues

- Let's call x the Mahalanobis distance D_M
- Imagine that
  - we xmatch 2 tables and we obtain n_{tot} associations
  - we know the number of spurious associations: n_{H_1}
  - $\Rightarrow$ we know the number of real associations: $n_{H_0} = n_{tot} - n_{H_1}$
- Distribution of x of real associations (likelihood $p(x|H_0)$)
  - almost a $\chi_{dof=2}(x)$, because...
  - ... its integrate over the acceptance domain must be $= 1$
  - $\Rightarrow p(x|H_0) = \chi_{dof=2}(x)/\gamma$
- Distribution of x of spurious associations (likelihood $p(x|H_1)$)
  - Poisson $\propto x$
  - again it must integrates to 1 over the domain of acceptance
  - $\Rightarrow p(x|H_0) = 2x/k_\gamma^2$

BAYES PROBA

For 2 catalogues

- Let's call x the Mahalanobis distance D_M
- Imagine that
  - we xmatch 2 tables and we obtain n_{tot} associations
  - we know the number of spurious associations: n_{H_1}
  - $\Rightarrow$ we know the number of real associations: $n_{H_0} = n_{tot} - n_{H_1}$
- Distribution of x of real associations (likelihood $p(x|H_0)$)
  - almost a $\chi_{dof=2}(x)$, because...
  - ... its integrate over the acceptance domain must be $= 1$
  - $\Rightarrow p(x|H_0) = \chi_{dof=2}(x)/\gamma$
- Distribution of x of spurious associations (likelihood $p(x|H_1)$)
  - Poisson $\propto x$
  - again it must integrates to 1 over the domain of acceptance
  - $\Rightarrow p(x|H_0) = 2x/k_\gamma^2$

BAYES PROBA

For 2 catalogues

- Let's call x the Mahalanobis distance D_M
- Imagine that
  - we xmatch 2 tables and we obtain n_{tot} associations
  - we know the number of spurious associations: n_{H_1}
  - $\Rightarrow$ we know the number of real associations: $n_{H_0} = n_{tot} - n_{H_1}$
- Distribution of x of real associations (likelihood $p(x|H_0)$)
  - almost a $\chi_{dof=2}(x)$, because...
  - ... its integrate over the acceptance domain must be $= 1$
  - $\Rightarrow p(x|H_0) = \chi_{dof=2}(x)/\gamma$
- Distribution of x of spurious associations (likelihood $p(x|H_1)$)
  - Poisson $\propto x$
  - again it must integrates to 1 over the domain of acceptance
  - $\Rightarrow p(x|H_0) = 2x/k_\gamma^2$

BAYES PROBA

For 2 catalogues

- Let's call x the Mahalanobis distance D_M
- Imagine that
  - we xmatch 2 tables and we obtain n_{tot} associations
  - we know the number of spurious associations: n_{H_1}
  - $\Rightarrow$ we know the number of real associations: $n_{H_0} = n_{tot} - n_{H_1}$
- Distribution of x of real associations (likelihood $p(x|H_0)$)
  - almost a $\chi_{dof=2}(x)$, because...
  - ... its integrate over the acceptance domain must be $= 1$
  - $\Rightarrow p(x|H_0) = \chi_{dof=2}(x)/\gamma$
- Distribution of x of spurious associations (likelihood $p(x|H_1)$)
  - Poisson $\propto x$
  - again it must integrates to 1 over the domain of acceptance
  - $\Rightarrow p(x|H_0) = 2x/k_\gamma^2$

BAYES PROBA

For 2 catalogues

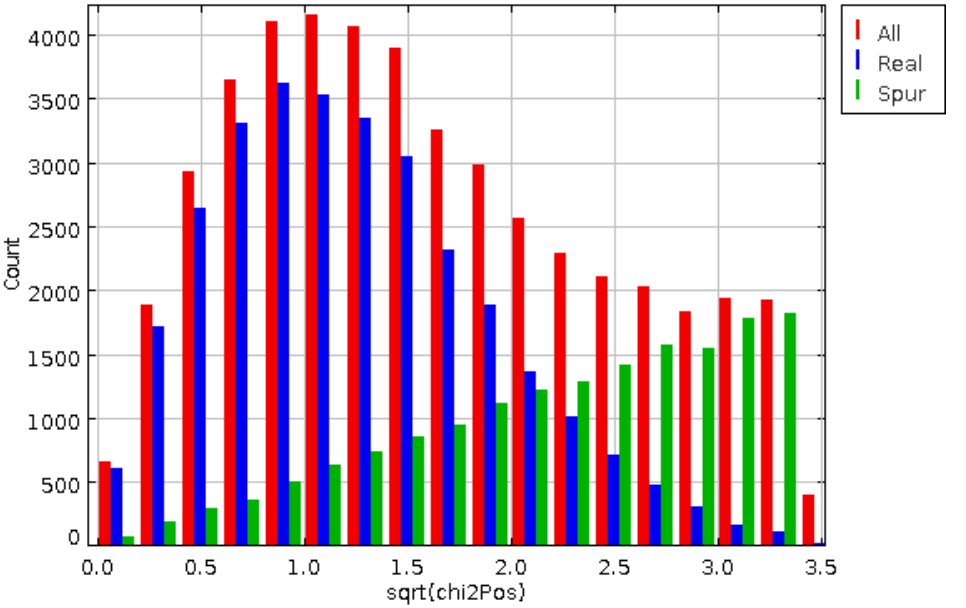
- Let's call x the Mahalanobis distance D_M
- Imagine that
  - we xmatch 2 tables and we obtain n_{tot} associations
  - we know the number of spurious associations: n_{H_1}
  - $\Rightarrow$ we know the number of real associations: $n_{H_0} = n_{tot} - n_{H_1}$
- Distribution of x of real associations (likelihood $p(x|H_0)$)
  - almost a $\chi_{dof=2}(x)$, because...
  - ... its integrate over the acceptance domain must be $= 1$
  - $\Rightarrow p(x|H_0) = \chi_{dof=2}(x)/\gamma$
- Distribution of x of spurious associations (likelihood $p(x|H_1)$)
  - Poisson $\propto x$
  - again it must integrates to 1 over the domain of acceptance
  - $\Rightarrow p(x|H_0) = 2x/k_\gamma^2$

BAYES PROBA

For 2 catalogues

- Blue curve: $n_{H_0} \times p(x|H_0)$
- Green curve: $n_{H_1} \times p(x|H_1)$
- Red curve = blue + green
- For an association of given x:

$$p(H_0|x) = \frac{\text{Blue curve}(x)}{\text{Red curve}(x)}$$



- Bayes formula:

$$p(H_0|x) = \frac{p(H_0)p(x|H_0)}{p(H_0)p(x|H_0) + p(H_1)p(x|H_1)}$$

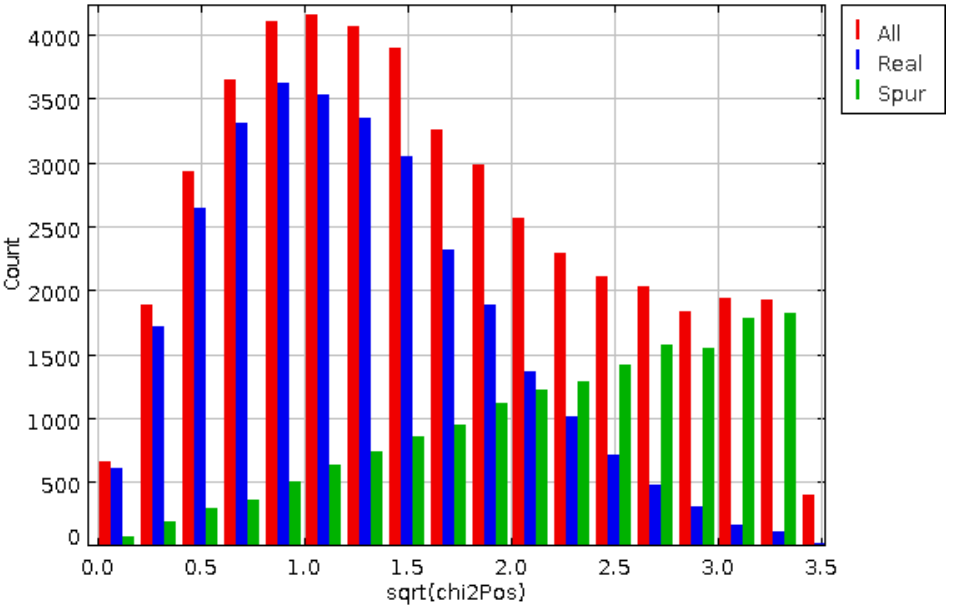
- Here priors $p(H_0) = n_{H_0}/n_{tot}$ and $p(H_1) = n_{H_1}/n_{tot}$

BAYES PROBA

For 2 catalogues

- Blue curve: $n_{H_0} \times p(x|H_0)$
- Green curve: $n_{H_1} \times p(x|H_1)$
- Red curve = blue + green
- For an association of given x:

$$p(H_0|x) = \frac{\text{Blue curve}(x)}{\text{Red curve}(x)}$$



- Bayes formula:

$$p(H_0|x) = \frac{p(H_0)p(x|H_0)}{p(H_0)p(x|H_0) + p(H_1)p(x|H_1)}$$

- Here priors $p(H_0) = n_{H_0}/n_{tot}$ and $p(H_1) = n_{H_1}/n_{tot}$

BAYES PROBA

For 2 catalogues

- We know n_{tot}: number of associations found by the candidate selection criterion
- How to estimate n_{H_0} or n_{H_1}?
  - one solution is to fit the previous histogram
  - an analytical solution exists!

BAYES PROBA

Raw priors estimate for 2 catalogues

- For one random source of catalogue A + one random source of catalogue B distributed on a common surface area S:
  - $S_{1,2}$: surface area of the acceptance region
  - Proba of spurious match: $S_{1,2}/S$
- For n_A sources in catalogue A and n_B sources in catalogue B

$$n_{H_1} = \sum_{i=1}^{n_A} \sum_{j=1}^{n_B} S_{i,j}/S$$

- For circular errors:

$$n_{H_1} = \frac{\pi k \gamma^2}{S} (E\{|V_A|^{1/2}\} + E\{|V_B|^{1/2}\})$$

  - $E\{|V_A|^{1/2}\}$, $E\{|V_B|^{1/2}\}$: means over cat A and cat B sources respectively
  - Super fast to compute!!

BAYES PROBA

Raw priors estimate for 2 catalogues

- For one random source of catalogue A + one random source of catalogue B distributed on a common surface area S:
  - $S_{1,2}$: surface area of the acceptance region
  - Proba of spurious match: $S_{1,2}/S$
- For n_A sources in catalogue A and n_B sources in catalogue B

$$n_{H_1} = \sum_{i=1}^{n_A} \sum_{j=1}^{n_B} S_{i,j}/S$$

- For circular errors:

$$n_{H_1} = \frac{\pi k \gamma^2}{S} (E\{|V_A|^{1/2}\} + E\{|V_B|^{1/2}\})$$

  - $E\{|V_A|^{1/2}\}$, $E\{|V_B|^{1/2}\}$: means over cat A and cat B sources respectively
  - Super fast to compute!!

BAYES PROBA

Raw priors estimate for 2 catalogues

- For one random source of catalogue A + one random source of catalogue B distributed on a common surface area S:
  - $S_{1,2}$: surface area of the acceptance region
  - Proba of spurious match: $S_{1,2}/S$
- For n_A sources in catalogue A and n_B sources in catalogue B

$$n_{H_1} = \sum_{i=1}^{n_A} \sum_{j=1}^{n_B} S_{i,j}/S$$

- For circular errors:

$$n_{H_1} = \frac{\pi k \gamma^2}{S} (E\{|V_A|^{1/2}\} + E\{|V_B|^{1/2}\})$$

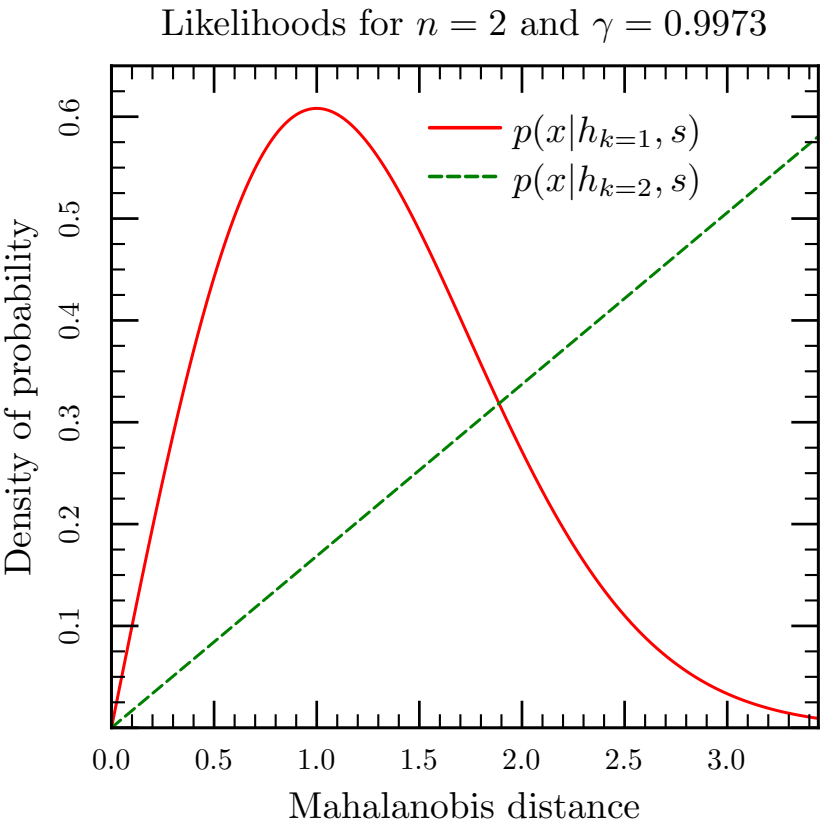
  - $E\{|V_A|^{1/2}\}$, $E\{|V_B|^{1/2}\}$: means over cat A and cat B sources respectively
  - Super fast to compute!!

BAYES PROBA

For 2 catalogues

Summary for 2 catalogues

- 2 hypotheses
- 2 likelihoods
  - $H_0 = AB$: $p(x|H_{AB})$, Chi of 2 dof
  - $H_1 = A\_B$: $p(x|H_{A\_B})$, Poisson 2D
- 2 priors based on (super fast) geometrical estimates



BAYES PROBA

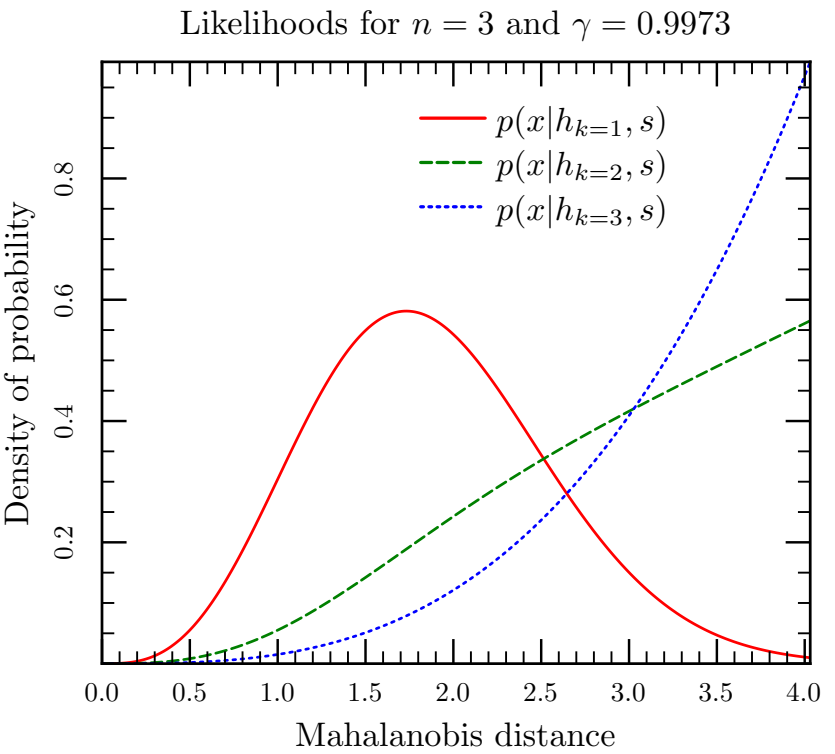
Transition

So now, for 3 catalogues

BAYES PROBA

For 3 catalogues

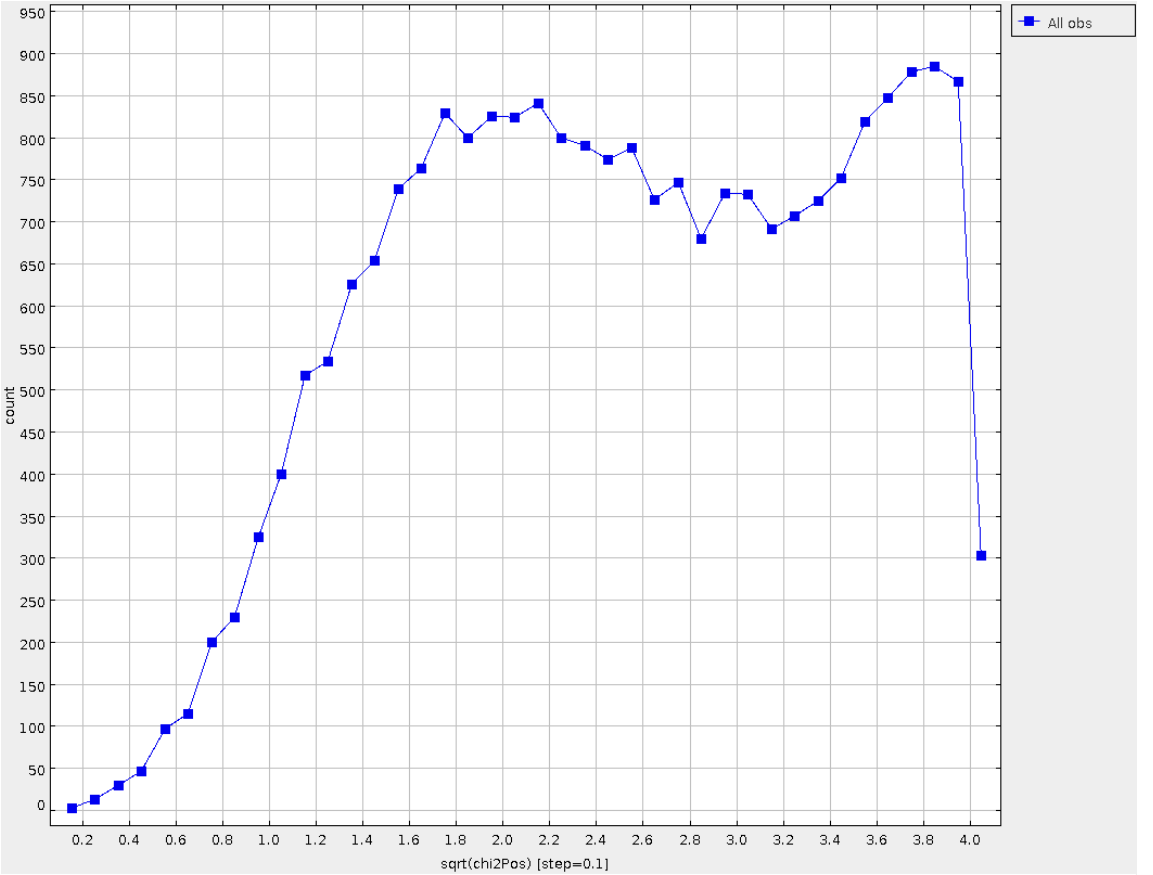
- 5 hypotheses
  - ABC, 1 real source
  - AB_C, 2 real sources
  - AC_B, 2 real sources
  - A_BC, 2 real sources
  - A_B_C, 3 real sources
- 3 likelihoods
  - 1 likelihood by number of real source
  - $p(x|H_{ABC}) = \chi_{dof=4}(x)/\gamma$: Chi 4 dof
  - $p(x|H_{AB\_C}) = p(x|H_{AB\_C}) = p(x|H_{AB\_C})$
  - $p(x|H_{A\_B\_C}) = 4x^3/k_\gamma^4$: Poisson 4D
- And priors?



BAYES PROBA

For 3 catalogues

- 5 hypotheses
- $\Rightarrow$ we need 5 priors
  - $p(H_{ABC})$, $p(H_{AB\_C})$, ...
- Need 4 xmatches:
  - A and B $\rightsquigarrow n_{H_0 AB}$
  - A and C $\rightsquigarrow n_{H_0 AC}$
  - B and C $\rightsquigarrow n_{H_0 BC}$
  - A, B and C $\rightsquigarrow n_{H_0 ABC}$
- Having all this, problem solved!

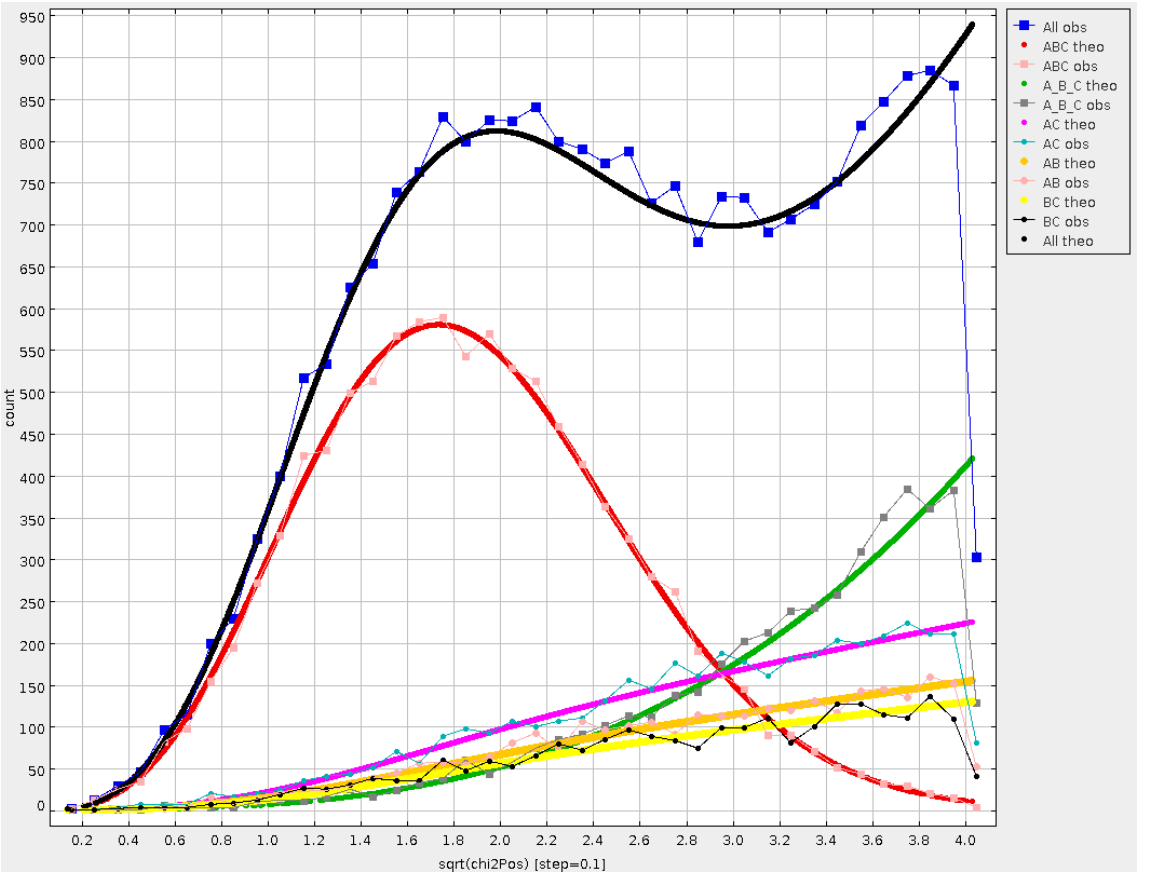


x: Mahalanobis distance
y: count

BAYES PROBA

For 3 catalogues

- 5 hypotheses
- $\Rightarrow$ we need 5 priors
  - $p(H_{ABC})$, $p(H_{AB\_C})$, ...
- Need 4 xmatches:
  - A and B $\rightsquigarrow n_{H_0 AB}$
  - A and C $\rightsquigarrow n_{H_0 AC}$
  - B and C $\rightsquigarrow n_{H_0 BC}$
  - A, B and C $\rightsquigarrow n_{H_0 ABC}$
- Having all this, problem solved!



x: Mahalanobis distance
y: count

BAYES PROBA

For n catalogues

- This generalises easily for n catalogues, BUT
  - Number of hypothesis increases dramatically
  - Number of priors increases dramatically
  - Number of xmatches to be performed increases dramatically
- Remark:
  - instead of computing $p(H|x)$ one can compute $p(H|\{\vec{\mu}\}, \{\mathbf{V}\})$
  - but $\mathbf{V}$ and magnitudes are NOT independant
  - $\Rightarrow$ we cannot deals with SEDs separatly (to be investigated)

We have been building a general purpose xmatch software implementing all this, but not only.

A VERSATILE XMATCH TOOL

Flexible, scalable, efficient

- Flexible and evolutive
  - dedicated script language
  - easy to add functionalities
- Efficient
  - tree datastructures
  - multithreading
- Scalable
  - web services (several machines)
  - parallel submission
  - allsky xmatch done cell by cell

Example of xmatch script

```
# Get XMM sources from a file
get File file=3xmm.fits
where SC_DET_ML<4
set pos ra=SC_RA dec=SC_DEC
set cols *
# Get SDSS DR9 sources from VizieR
get VizieR tabname=V/139/sdss9 mode=cone ...
set pos ra=RAJ2000 dec=DEJ2000
set cols ObjID,/(e_)?[ugriz]mag/,u-g as ug
addmeta ug datatype=float unit=mag ucd=...
# Perform a simple 3" xmatch
xmatch cone dMax=3 join=inner nThreads=48
merge dist mec
# Save intermediary result
save result1.vot votable
# Chain xmatches
get ...
xmatch ...
...
```

A VERSATILE XMATCH TOOL

Flexible, scalable, efficient

- Flexible and evolutive
  - dedicated script language
  - easy to add functionalities
- Efficient
  - tree datastructures
  - multithreading
- Scalable
  - web services (several machines)
  - parallel submission
  - allsky xmatch done cell by cell

Example of xmatch script

```
# Get XMM sources from a file
get File file=3xmm.fits
where SC_DET_ML<4
set pos ra=SC_RA dec=SC_DEC
set cols *
# Get SDSS DR9 sources from VizieR
get VizieR tabname=V/139/sdss9 mode=cone ...
set pos ra=RAJ2000 dec=DEJ2000
set cols ObjID,/(e_)?[ugriz]mag/,u-g as ug
addmeta ug datatype=float unit=mag ucd=...
# Perform a simple 3" xmatch
xmatch cone dMax=3 join=inner nThreads=48
merge dist mec
# Save intermediary result
save result1.vot votable
# Chain xmatches
get ...
xmatch ...
...
```

A VERSATILE XMATCH TOOL

Flexible, scalable, efficient

- Flexible and evolutive
  - dedicated script language
  - easy to add functionalities
- Efficient
  - tree datastructures
  - multithreading
- Scalable
  - web services (several machines)
  - parallel submission
  - allsky xmatch done cell by cell
    - e.g. HEALPix cell
    - one generic script

Example of xmatch script

```
# Get XMM sources from a file
get File file=3xmm.fits
where SC_DET_ML<4
set pos ra=SC_RA dec=SC_DEC
set cols *
# Get SDSS DR9 sources from VizieR
get VizieR tabname=V/139/sdss9 mode=cone ...
set pos ra=RAJ2000 dec=DEJ2000
set cols ObjID,/(e_)?[ugriz]mag/,u-g as ug
addmeta ug datatype=float unit=mag ucd=...
# Perform a simple 3" xmatch
xmatch cone dMax=3 join=inner nThreads=48
merge dist mec
# Save intermediary result
save result1.vot votable
# Chain xmatches
get ...
xmatch ...
...
```

A VERSATILE XMATCH TOOL

Features: various xmatch algorithms

Algorithm	param	#tbl	prop.mot.	index struct.
chi2 (χ^2)	proba	2	l[1], r[2], b[3]	M/TM-tree

XMatches are chainable: 1 χ^2 xmatch of 4 tables = 3 χ^2 xmatches of 2 tables!

4 to 11 joins ($LIRF\bar{L}\bar{I}\bar{R}L'I'R'F'$) are supported according to the algorithm.

- [1] l: left table contains extended objects or proper motions;
- [2] r: right table contains extended objects or proper motions;
- [3] b: both left and right tables contain extended objects or proper motion.

A VERSATILE XMATCH TOOL

Features: various xmatch algorithms

WP4

Algorithm	param	#tbl	prop.mot.	index struct.
chi2 (χ^2)	proba	2	l[1], r[2], b[3]	M/TM-tree
proba2_v*x*	proba	2	no (?)	M-tree
proba3_v*x*	proba	3	no (?)	M-tree
probaN_v*x*	proba	n	no (?)	M-tree

XMatches are chainable: 1 χ^2 xmatch of 4 tables = 3 χ^2 xmatches of 2 tables!

4 to 11 joins ($LIRF\bar{L}\bar{I}\bar{R}L'I'R'F'$) are supported according to the algorithm.

- [1] l: left table contains extended objects or proper motions;
- [2] r: right table contains extended objects or proper motions;
- [3] b: both left and right tables contain extended objects or proper motion.

A VERSATILE XMATCH TOOL

Features: various xmatch algorithms

WP4

Algorithm	param	#tbl	prop.mot.	index struct.
chi2 (χ^2)	proba	2	l[1], r[2], b[3]	M/TM-tree
proba2_v*x*	proba	2	no (?)	M-tree
proba3_v*x*	proba	3	no (?)	M-tree
probaN_v*x*	proba	n	no (?)	M-tree
knn	k+dist	2	r, b	kd/M/TM-tree
cone	dist	2	l, r, b	kd/M/TM-tree

XMatches are chainable: 1 χ^2 xmatch of 4 tables = 3 χ^2 xmatches of 2 tables!

4 to 11 joins ($LIRF\bar{L}\bar{I}\bar{R}L'I'R'F'$) are supported according to the algorithm.

- [1] l: left table contains extended objects or proper motions;
- [2] r: right table contains extended objects or proper motions;
- [3] b: both left and right tables contain extended objects or proper motion.

A VERSATILE XMATCH TOOL

Features: various xmatch algorithms

WP4

Algorithm	param	#tbl	prop.mot.	index struct.
chi2 (χ^2)	proba	2	l[1], r[2], b[3]	M/TM-tree
proba2_v*x*	proba	2	no (?)	M-tree
proba3_v*x*	proba	3	no (?)	M-tree
probaN_v*x*	proba	n	no (?)	M-tree
knn	k+dist	2	r, b	kd/M/TM-tree
cone	dist	2	l, r, b	kd/M/TM-tree
ext_l[1]	r	2	no	M-tree
ext_r[2]	r	2	no	M-tree
ext_b[3]	r	2	no	M-tree

XMatches are chainable: 1 χ^2 xmatch of 4 tables = 3 χ^2 xmatches of 2 tables!

4 to 11 joins ($LIRF\bar{L}\bar{I}\bar{R}L'I'R'F'$) are supported according to the algorithm.

- [1] l: left table contains extended objects or proper motions;
- [2] r: right table contains extended objects or proper motions;
- [3] b: both left and right tables contain extended objects or proper motion.

A VERSATILE XMATCH TOOL

Features: various xmatch algorithms

WP4

Algorithm	param	#tbl	prop.mot.	index struct.
chi2 (χ^2)	proba	2	l[1], r[2], b[3]	M/TM-tree
proba2_vx	proba	2	no (?)	M-tree
proba3_vx	proba	3	no (?)	M-tree
probaN_vx	proba	n	no (?)	M-tree
knn	k+dist	2	r, b	kd/M/TM-tree
cone	dist	2	l, r, b	kd/M/TM-tree
ext_l[1]	r	2	no	M-tree
ext_r[2]	r	2	no	M-tree
ext_b[3]	r	2	no	M-tree

XMatches are chainable: 1 χ^2 xmatch of 4 tables $=$ 3 χ^2 xmatches of 2 tables!
4 to 11 joins ($LIRF\bar{L}\bar{I}\bar{R}L'I'R'F'$) are supported according to the algorithm.

- [1] l: left table contains extended objects or proper motions;
- [2] r: right table contains extended objects or proper motions;
- [3] b: both left and right tables contain extended objects or proper motion.

Building ARCHES catalogue(s)

WP4

- Group XMM FOVs having similar properties (F. Carrera)
  - tool (STILTS script) dedicated to ARCHES
  - $\approx$ 200 groups in output
    - surface area covered by each group (UNION of FOVs)
- For each group of FOVs
  - (automatically) write a (quite complex) script
    - another tool dedicated to ARCHES
    - one ARCHES script $\approx$ 340-800 lines
    - example available during Hands on session
  - submit the script to the xmatch tool
- Simply concatenates the results

COMPLICATION

Successive xmatches

A trap prooving (if necessary) that *n*-xmatches are complex.

- I want all sources from A having candidates in B or C (or both)
- $\Leftrightarrow$ A inner join (B full join C)
- but imagine, 3 source *a*, *b* and *c*:
  - *a* is χ^2 compatible with *b* but NOT with *c*
  - *b* and *c* are χ^2 compatible
  - then (B full join C) ouput contains 1 row:
    - ⋆ row containing *b* and *c*
  - then A inner join (B full join C) does not contains any row!
  - BUT I WANTED the row with *a* and *b*
  - Solution *ffull join*: result of A ffull join B:
    - ⋆ row containing *b* and *c*
    - ⋆ row containing *b* alone
    - ⋆ row containing *c* alone
  - then A inner join (B ffull join C) contains 1 row:
    - ⋆ row containing *a* and *b* :)

COMPLICATION

Successive xmatches (continuation)

- Now imagine that *a* and *c* are χ^2 compatible and *a*, *b* and *c* are also χ^2 compatible
- Then *ffull join* still contains
  - row containing *b* and *c*
  - row containing *b* alone
  - row containing *c* alone
- But now A inner join (B ffull join C) contains 3 rows:
  - row containing *a*, *b* and *c* :)
  - row containing *a* and *b* :(
  - row containing *a* and *c* :(
- Need a specific post filter to remove 2 out of 3 rows

CONCLUSION

- In the framework of ARCHES we have been developping
  - a flexible, multi-purpose tool
  - able to xmatch several catalogues in various ways
  - able to compute probabilities assuming an ideal world
- Provides the basis for an identification process:
  - needs a layer on top for your particular problem
  - photometry based proba to be added for proper identification
- Not the end of story:
  - will be integrated to the CDS XMatch service
  - opens the road for more complex multi-catalogues proba?
- Script + open tool: good for reproductibility, large parts of the process out of the black box (allows for criticisms), ...