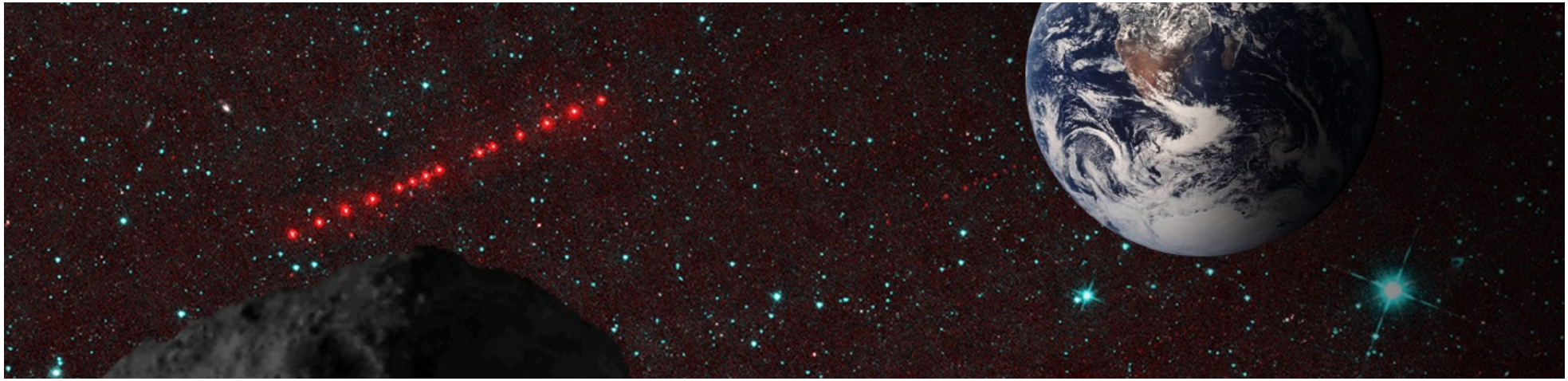


Probabilistic Source Matching: a parametric approach

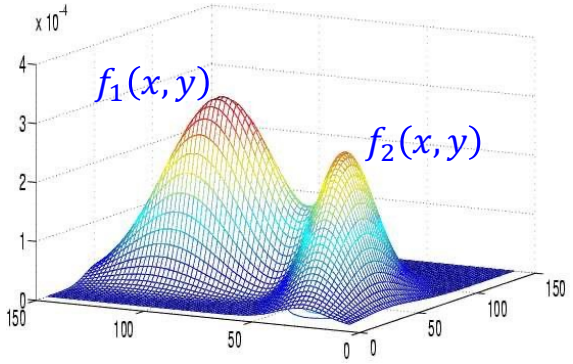
Frank Masci

May 22, 2023

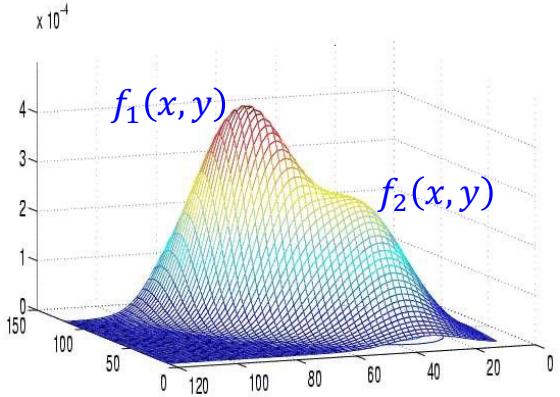
Caltech – IPAC

Probabilistic Source (centroid) Matching: general framework

- The probability of finding a source at some 2D position *x*, *y* within an area Δ*x*Δ*y* is given by $f(x,y)\Delta x\Delta y$ where $f(x,y)$ is its bivariate probability density function (PDF).
- Suppose we have two generic bivariate PDFs corresponding to two nearby sources: $f_1(x,y)$ and $f_2(x,y)$, each separated by different amounts. These can be pictured as follows:



$D \approx 0.3$



$D \approx 0.8$

- Likelihood of finding both sources at some *x*, *y* within a region Δ*x*Δ*y* is $\propto f_1\ (x,y)f_2(x,y)\Delta x\Delta y$.
- We can define a quantitative measure for both sources to have the same PDF (i.e., occupy the same region):

$$D = \iint [f_1(x,y)f_2(x,y)]^{1/2}\,dx\,dy$$

Probabilistic Source (centroid) Matching: assuming normality

- This metric satisfies $0 \leqslant D \leqslant 1$ where smaller values of D imply the PDFs are more dissimilar.
- Sometimes referred to as the *dissimilarity* metric (Bhattacharyya, 1946).

- In particular, if $f_1(x, y)$ and $f_2(x, y)$ are two *bivariate normal* distributions:

$$f_1(x, y) \sim N(\mu_1, C_1) \quad \text{and} \quad f_2(x, y) \sim N(\mu_2, C_2) \quad \text{where}$$

$$\mu_1 = \begin{pmatrix} x_1 \\ y_1 \end{pmatrix}, \quad C_1 = \begin{pmatrix} \sigma_{x1}^2 & \sigma_{x1y1}^2 \\ \sigma_{x1y1}^2 & \sigma_{y1}^2 \end{pmatrix} \quad \text{and} \quad \mu_2 = \begin{pmatrix} x_2 \\ y_2 \end{pmatrix}, \quad C_2 = \begin{pmatrix} \sigma_{x2}^2 & \sigma_{x2y2}^2 \\ \sigma_{x2y2}^2 & \sigma_{y2}^2 \end{pmatrix}$$

- Their [dis]similarity metric can be written:

$$D = \underbrace{\frac{2[det(C_1)det(C_2)]^{1/4}}{[det(C_1 + C_2)]^{1/2}}}_{(1)} \underbrace{\exp\left[-\frac{1}{4}(x_2 - x_1 \quad y_2 - y_1)(C_1 + C_2)^{-1}\begin{pmatrix} x_2 - x_1 \\ y_2 - y_1 \end{pmatrix}\right]}_{(2)}$$

- **Two terms:** (1) measure of [dis]similarity in "size" or spread
  (2) measure of [dis]similarity in separation of means (source centroids)

Probabilistic Source (centroid) Matching: test statistic

- For our application, we are exclusively interested in the second term of *D*.
- This term is maximal when the argument of the exponential is minimal.
- Defining $\Delta x = x_2 - x_1$ and $\Delta y = y_2 - y_1$, the test *statistic* to minimize is:

$$S = (\Delta x \quad \Delta y)(C_1 + C_2)^{-1}\begin{pmatrix}\Delta x \\ \Delta y\end{pmatrix}$$

- Expanding the covariance matrices and centroid-difference vectors,

$$S = \frac{\left(\sigma_{y1}^2 + \sigma_{y2}^2\right)\Delta x^2 + \left(\sigma_{x1}^2 + \sigma_{x2}^2\right)\Delta y^2 - 2\left(\sigma_{x1y1}^2 + \sigma_{x2y2}^2\right)\Delta x \Delta y}{\left(\sigma_{x1}^2 + \sigma_{x2}^2\right)\left(\sigma_{y1}^2 + \sigma_{y2}^2\right) - \left(\sigma_{x1y1}^2 + \sigma_{x2y2}^2\right)^2}$$

- If all covariances = 0:

$$S = \frac{\Delta x^2}{\left(\sigma_{x1}^2 + \sigma_{x2}^2\right)} + \frac{\Delta y^2}{\left(\sigma_{y1}^2 + \sigma_{y2}^2\right)}$$

Principal axes representation and a quiz

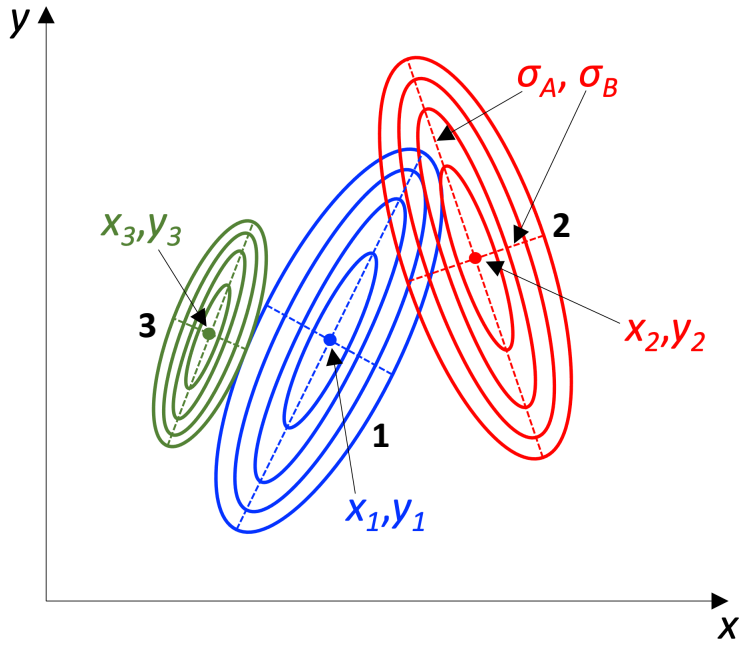
- Below is a schematic of three centroids and their error ellipses, i.e., projections of their iso-probability-density contours in the *x, y* plane.
- The variances along the principal axes of ellipse #2 are given by solving for eigenvalues of covariance matrix:

$$\sigma_A^2 = \frac{1}{2}\left(\sigma_{x2}^2 + \sigma_{y2}^2 + \left[\left(\sigma_{x2}^2 - \sigma_{y2}^2\right)^2 + 4\sigma_{x2y2}^4\right]^{1/2}\right)$$

$$\sigma_B^2 = \frac{1}{2}\left(\sigma_{x2}^2 + \sigma_{y2}^2 - \left[\left(\sigma_{x2}^2 - \sigma_{y2}^2\right)^2 + 4\sigma_{x2y2}^4\right]^{1/2}\right)$$

- The *S* metric can then be recast in terms of σ_A and σ_B.
- In practice, it's easier to work with projected σ_x, σ_y values.

Quiz: which position (**centroid 2** or **centroid 3**) is the "best" or most likely match to **centroid 1**?



Null hypothesis and distribution for *S*

- If the errors in the centroids are normally distributed, the test statistic *S* will follow the *null probability density distribution* (defining the null hypothesis **H0**):

$$PDF(S) \sim \chi^2_\nu = \frac{1}{2}\exp\left[-\frac{S}{2}\right] \text{ for } \nu = 2 \text{ degrees of freedom}$$

- **H0:** two centroids are associated with the same source OR are consistent within random measurement errors.

- Probability of obtaining at least the value *S* under **H0** by chance:

$$Prob(> S) = \int_S^\infty \chi^2_{\nu=2} dS = \exp\left[-\frac{S}{2}\right].$$

- If $Prob(> S) < P_{crit}$, we reject **H0** and declare the pair of centroids *probably unassociated*.
  => Their separation is very unlikely to be due to random measurement errors alone.

- Alternatively, we can invert the above and find the maximum tolerable value S_{max} above which to reject **H0**:

$$S_{max} = -2\log(P_{crit})$$

- For example, if $P_{crit} = 0.05$, $S_{max} \simeq 6$.

Overview of process

- We first match source positions between two samples within a coarse radius *R* tuned using some prior knowledge of their uncertainties, e.g.,

  $R \simeq 5\left(\left\langle \sigma_{x1}^2 + \sigma_{y1}^2 \right\rangle + \left\langle \sigma_{x2}^2 + \sigma_{y2}^2 \right\rangle\right)^{1/2}$

  where angled brackets denote averages.

- For each input seed position (e.g., blue centroid on slide 5), compute statistic *S* for each matching candidate position and its significance *Prob*(> *S*). Then threshold against some P_{crit} (slide 6).

- Alternatively (simpler), threshold the *S* values against some S_{max} value corresponding to P_{crit} (slide 6).

- For multiple candidates, most likely match is the one with smallest *S* (largest *Prob*). However, see next slide.

Checks and other considerations

1. For multiple matching candidates with $S < S_{max}$ and little dynamic range between their S values, we can complement the matching with a flux-match, however, beware of intrinsic source variability.

2. Are the prior positional uncertainties σ_x , σ_y and their covariances plausible? It is advised to cross-check them by computing sample (co)variances, then recalibrating if necessary.

3. Can normality of the sample positional errors be justified? I.e., are the marginal PDFs along each axis approximately normal (aka Gaussian)? Repeated measurements of the same source can be used here.

4. Does a simple Euclidean distance match using $S = \Delta x^2 + \Delta y^2$ perform better or give similar results (in terms of completeness & reliability) than the S defined on slide 4 that includes uncertainty priors?
   - If so, either the priors are implausible (point 2 above), normality is not justified (point 3 above), or prior uncertainties are indeed plausible but negligible for both samples.

5. **Homework:** how would one modify the test statistic S and null PDF on slide 6 to account for chance (hence false) associations due to high source density? There's more to it than just measurement error alone.