Pedagogical Seminar: The Mathematical Methods of Source Association

Emma Taylor

11th March, 2004

1 Introduction

Due to different spatial resolutions of different instruments it is very difficult to match up objects from different surveys. For example, there can up to hundreds of optical objects within the error ellipse of a far-infrared source, so how can we know which object is it's true optical counterpart?

I am going to outline three methods of source associations which return associations along with a figure of merit for the match.

2 The Poisson Model

Consider a catalogue with a magnitude limit m_T. If we search to a radius r_s from our source then the simple expected number of events as extreme as an association with an object of magnitude m_0 a distance r_0 from the source will be $E_0 = \pi r_0^2 N(\leq m_0)$ and the poisson probability of this event occurring would be $P_0 = 1 - exp(-E_0)$.

However, a posteriori probabilities as low as P_0 could have also been produced by brighter objects at a greater distance from the source or fainter objects closer to the source. Therefore, the true number of expected events is $E = E_0 + E_1 + E_2$ where E_1 and E_2 account for brighter/farther and fainter/closer objects respectively,

$$
\begin{aligned}
E_1 &= \int_{r_0}^{r_s} N[\leq m_{max}(r)] 2\pi r dr \\
E_2 &= \int_{m_0}^{m_T} n(m) dm \int_0^{r_{max}(m)} 2\pi r dr
\end{aligned}
$$

Where

- $N[\leq m_{max}(r)]$ is the cumulative number density of objects brighter than a limiting magnitude m_{max} which is a function of r such that $\pi r^2 N[\leq m_{max}(r)] = E_0$.

- $n(m)$ is the differential magnitude function i.e. $\frac{dN(\leq m)}{dm} = n(m)$

- $r_{max}(m)$ is the distance from the source at which an object of magnitude m will have an a posteriori probability of P_0 therefore $\pi r_{max}^2(m) N(\leq m) = E_0$

Therefore,

$$
\begin{aligned}
E_1 &= \int_{r_0}^{r_s} 2E_0 \frac{dr}{r} = E_0 ln(r_s^2/r_0^2) \\
E_2 &= \int_{N(\leq m_0)}^{N(\leq m_T)} E_0 \frac{dN(\leq m)}{N(\leq m)} = E_0 ln\Big[\frac{N(\leq m_T)}{N(\leq m_0)}\Big] \\
E &= E_0 + E_0 ln(r_s^2/r_0^2) + E_0 ln\Big[\frac{N(\leq m_T)}{N(\leq m_0)}\Big] \\
&= E_0[1 + ln(E_c/E_0)]
\end{aligned}
$$

where $E_c = \pi r_s^2 N(\leq m_T)$. The a priori probability is $P = 1 - exp(-E)$.

3 The Likelihood Ratio Method

Consider a candidate object of magnitude m, positional offset x and y from the estimated source position. The probability that the true counterpart to the source lies in an infinitesimal box $x \pm dx/2$, $y \pm dy/2$ with magnitude $m \pm dm/2$ is:

$$p = q(m)e(x, y)dmdxdy$$

where $e(x, y)$ is the probability distribution function of the positional errors and $q(m)$ is the magnitude distribution function.

The equivalent probability for a background object is:

$$p = n(m)dmdxdy$$

where $n(m)$ is the surface density per magnitude of background objects with magnitude m.

The likelihood ratio is then the ratio of these two infinitesimal probabilities.

$$LR = \frac{q(m)e(x,y)}{n(m)}$$

This is the probability of finding the true ID at this position with this magnitude relative to the probability of finding a similar chance background object.

We also need a measure of the reliability of this association. From Bayes theorem:

$$\begin{aligned} R = P(genuine|m,x,y) &= \frac{P(genuine,m,x,y)}{P(genuine,m,x,y) + P(chance,m,x,y)} \\ &= \frac{q(m)e(x,y)dmdxdy}{[q(m)e(x,y)+n(m)]dmdxdy} \\ &= \frac{LR}{1+LR} \end{aligned}$$

However, this does not allow for there being more than one candidate for the source. The Sutherland and Saunders (1992) corrected form of the reliability is:

$$R_j = \frac{LR_j}{\sum_i LR_i + (1-Q)}$$

where $Q = \langle \sum_{chance} LR_i \rangle$ the mean total likelihood for the source for all the chance candidates.

3.1 The Rutledge et al. (2000) Method

The true counterpart to the source is assumed to be bright and the distribution of positional offsets assumed to be Gaussian giving an LR of the form:

$$LR = \frac{exp(-r_2/\sigma_i^2)}{\sigma_i^2 N(< B_i)}$$

The reliability is then estimated to be:

$$\widetilde{R}(LR) = \frac{N_{source}(LR) - N_{background}(LR)}{N_{source}(LR)}$$

However, this does not allow for the fact that there may be other objects in the field just as likely to be the true association. To overcome this a set of M+1 hypotheses are imposed and the probability of these being true calcuated using the estimated reliability $\widetilde{R}$.

Hypotheses i...M - The ith object is uniquely associated with the source (P_{id_i}).

Hypothesis M+1 - none of the M objects are uniquely associated with the source (P_{no-id}).

Since $\widetilde{R}(LR_i)$ is the binomial probability that the ith object is associated with the source, the probability that none of the M objects are associated with the source is just:

$$P_{no\_id} = \frac{\prod_{j=1}^{M}(1 - R_j)}{S}$$

where S is just a normalisation factor to ensure that all the probabilities sum to 1.

The probability that the source is associated with the ith object is simply:

$$P_{id-i} = \frac{R_i \prod_{j \neq i}^{M}(1 - R_j)}{S}$$

4 The EM Algorithm

4.1 Set up the Distributions

We have a number of sources labelled, r, with which we seek associations within a search radius R_r. Objects within R_r are labelled i, each having an associated position relative to the source, x_i^r, and a magnitude m_i^r. An indicator function, z_i^r, is used to label whether an object is the true ID, $z_i^r = 1$, or it is a background object, $z_i^r = 0$. We also allow for the possibility that there is no counterpart for the source, $z_0^r = 1$.

Suppose there are N_r objects within the search radius R_r, then we restrict $\sum_{i=0}^{N_r} z_i^r = 1$, this allows at most only one counterpart per source.

- $P(x_i^r|z_i^r = 1)$ is the probability of an object, i, having a position x_i^r given that it is associated with the source r.
- $P(x_i^r|z_i^r = 0)$, is the probability of an object i having a position x_i^r given that it is not associated with the source, i.e. it is just a background object, which is estimated as $P(x_i^r|z_i^r = 0) = \mu_r = 1/(\pi R_r^2)$
- $P(m_i^r|z_i^r = 0)$ is the magnitude distribution for the background objects.
- $P(m_i^r|z_i^r = 1)$ is the magnitude distribution of the counterparts for the sources. This is unknown and must be estimated. We assume it may be represented by a histogram with B bins of width w.

$$P(m_i^r|z_i^r = 1) = \sum_{t=1}^{B} h_t T_t(m_i^r)$$

T_t is a top hat function that is 1 if m_i^r lies within the tth bin and 0 otherwise. We set $w\sum_t h_t = 1$ such that h_t is the normalised bin height.

- The probability that a source has no counterpart is unknown and set to γ, $P(z_0^r = 1) = \gamma$.
- Initially assume that if there is an counterpart, each of the candidates has an equal probability of being the true ID. Therefore, $P(z_i^r = 1) = \frac{(1-\gamma)}{N_r}$.

4.2 Set up the EM Algorithm

The algorithm utilises the KL (Kullbeck-Leibler) divergence which measures the 'distance' between two probability densities. This is given by

$$KL(W(x_i)||Y(x_i)) = \sum_i W(x_i) ln \frac{W(x_i)}{Y(x_i)}$$

i.e. it is the expectation with respect to $W(x)$ of the difference between $lnW(x)$ and $lnY(x)$. Note that the KL divergence cannot take a negative value and is at its minimum value of 0 when $W = Y$.

To set this up we take the KL divergence of the posterior distribution, $P(Z|D,\Theta)$ and some arbitrary distribution Q(Z).

$$\begin{aligned} KL(Q(Z)||P(Z|D,\Theta)) &= \langle ln \frac{Q(Z)}{P(Z|D,\Theta)} \rangle_{Q(Z)} \\ &= \langle \left(lnQ(Z) - ln\left(\frac{P(Z,D|\Theta)}{P(D|\Theta)} \right) \right) \rangle_{Q(Z)} \end{aligned}$$

since $P(X,Y|T) = P(X|Y,T)P(Y|T)$.

We wish to know the likelihood $P(D|\Theta)$, the probability of the data, D, given the parameters, Θ. In our case the parameters are $\Theta = \{h_t, \gamma\}$, these are what we need to learn using the EM algorithm.

$$\begin{aligned} KL(Q(Z)||P(Z|D,\Theta)) &= \langle (lnQ(Z) - lnP(Z,D|\theta) + lnP(D|\Theta)) \rangle_{Q(Z)} \\ &= \langle (lnQ(Z) - lnP(Z,D|\Theta)) \rangle_{Q(Z)} + lnP(D|\Theta) \end{aligned}$$

since the likelihood $P(D|\Theta)$ is independent of Z.

$$lnP(D|\Theta) = \langle (lnP(Z,D|\Theta) - lnQ(Z)) \rangle_{Q(Z)} + KL(Q(Z)||P(Z|D,\Theta))$$

Since Q(Z) is independent of Θ we set $Q(Z) = P(Z|D,\Theta)$, the posterior distribution, hence reducing the KL divergence to its minimum value,

0. We then alter the parameters to maximise $\langle(lnP(Z, D|\Theta)\rangle_{Q(Z)}$ and hence maximise the log likelihood, $lnP(D|\Theta)$. Having set the KL divergence to zero ensures that when the parameters are adjusted we do not inadvertently reduce the KL divergence and consequently the log likelihood, since the divergence can only increase from 0. Since Q does not depend on the parameters, Θ, it can be reset to the new value of the posterior given the new parameters without affecting the log likelihood. This therefore gives an iterative procedure, summarised below, for maximising the log likelihood by finding the best values for the parameters, Θ.

1. Initialise the parameters, Θ.

2. Set Q to the current estimate of the posterior distribution for Z, i.e. $Q(Z) = P(Z|D, \Theta)$.

3. Maximise $\langle(lnP(D, Z|\Theta))\rangle_Q(z)$ given this value of Q(Z) by altering the values of Θ.

4. Repeat steps 2 and 3 until suitably converged.

The expectation of an indicator variable is its probability of being 1. Therefore,

$$< z_i^r >= P(z_i^r = 1|D^r, \Theta)$$

but $P(z_i^r|D^r, \Theta) \propto P(z_i^r = 1, D^r|\Theta) = P(Z^r, D^r|\Theta)$. Therefore we can calculate the expectation values from the distribution we are 'learning'.

4.3 Final Procedure

1. Initialise the parameters h_t and γ.

2. Calculate the marginal posteriors $P(z_i^r = 1|D^r, \Theta) =< z_i^r = 1 >$ and $P(z_0^r = 1|D^r, \Theta) =< z_0^r = 1 >$. This is equivalent to setting Q(Z) to the current estimate of the posterior distribution.

3. Use these values of $< z_i^r = 1 >$ and $< z_0^r = 1 >$ to set h_t and γ.

4. Repeat steps 2 and 3 until the values of h_t and γ are suitably converged.

5 Conclusions

Associating astronomical sources can be very tricky!! The method chosen should depend on any prior knowledge regarding the source population and the ease with which the parameters can be estimated in the case in question.

References

Downes, A. J. B., Peacock, J. A., Savage, A., and Carrie, D. R.: 1986, *MNRAS* **218**, 31

Feigelson, E. D. and Babu, G. J. (eds.): 2003, *Statistical challenges in astronomy*

Lucy, L. B.: 1974, *AJ* **79**, 745

Mann, R. G., Oliver, S., Carballo, R., Franceschini, A., Rowan-Robinson, M., Heavens, A. F., Kontizas, M., Elbaz, D., Dapergolas, A., Kontizas, E., Granato, G. L., Silva, L., Rigopoulou, D., Gonzalez-Serrano, J. I., Verma, A., Serjeant, S., Efstathiou, A., and van der Werf, P. P.: 2002, *MNRAS* **332**, 549

Mann, R. G., Oliver, S. J., Serjeant, S. B. G., Rowan-Robinson, M., Baker, A., Eaton, N., Efstathiou, A., Goldschmidt, P., Mobasher, B., Sumner, T. J., Danese, L., Elbaz, D., Franceschini, A., Egami, E., Kontizas, M., Lawrence, A., McMahon, R., Norgaard-Nielsen, H. U., Perez-Fournon, I., and Gonzalez-Serrano, J. I.: 1997, *MNRAS* **289**, 482

Rutledge, R. E., Brunner, R. J., Prince, T. A., and Lonsdale, C.: 2000, *ApJS* **131**, 335

Serjeant, S., Dunlop, J. S., Mann, R. G., Rowan-Robinson, M., Hughes, D., Efstathiou, A., Blain, A., Fox, M., Ivison, R. J., Jenness, T., Lawrence, A., Longair, M., Oliver, S., and Peacock, J. A.: 2003, *MNRAS* **344**, 887

Sutherland, W. and Saunders, W.: 1992, *MNRAS* **259**, 413