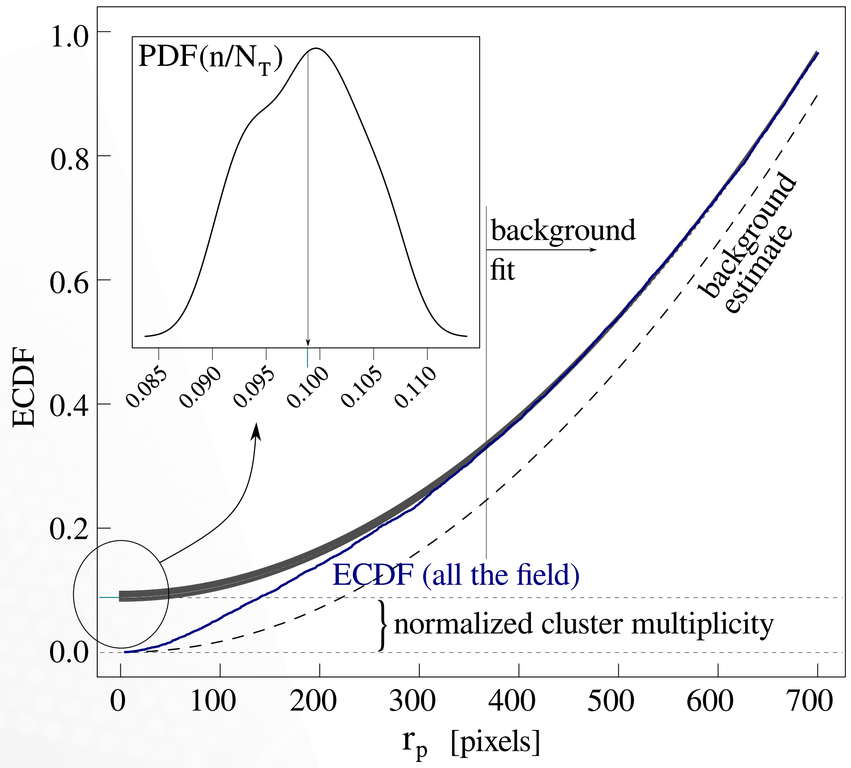
A formal approach to compute density profiles and isochrone fitting in stellar associations

Marcelo Lares (IATE), Luciana Gramajo (OAC) & Bruno Sánchez (IATE)

A star cluster can be characterized by a set of fundamental parameters (distance, reddening, age, and metallicity), whose estimation is valuable for a variety of studies related to their formation and evolution. The main diagnostic tool to obtain them is the Color-Magnitude Diagram (hereafter CMD), but a carefull field cleaning must be performed previously to isolate the contribution of stars in the clusters from stars in the background. Several procedures have been proposed to reduce this contamination. In general, they use data discretization, implementing binning schemes to estimate membership probabilities. However, the choice of the bin sizes is arbitrary, and the uncertainties in magnitudes and colours are usually not taken into account. Moreover, the membership are assigned informally on a star-by-star basis. Then, the fitting of isochrones to the scatter plot of selected stars is made by eye. This approach is therefore subjective and does not complain with the principle of reproducible science. Here we propose a formal approach to:

1. Estimate the projected density profile
2. Compute the statistical distribution in the CMD
3. Find the best fit isochrones to the CMD



Scheme of the statistical computation of the density profile.

1. Estimating the projected radial density profile

The radial profile of a star cluster is characterized by a set of measurements of the projected radial distances to the cluster center estimate. If we assume that the radial profile follows an unknown probability distribution function, the set of N c measured radii is in fact the set $\{R_1, R_2, \ldots, R_{N_C}\}$ of identically distributed random variables, whose parent distribution is to be determined. Often in practice, for example in a photometric field, the radial distances of stars are contaminated by background stars, following an unknown signal-to-background ratio. Let S be the set of all projected radii relative to the cluster center estimate in a photometric field. This set includes the stars in the cluster and the stars in the background, without distinction between them. Each value can be considered as a realization of a random variable that follows the same underlying distribution function, with contributions from the cluster (i.e., the signal), and from the background. Let N be the number of objects in S up to a given maximum radius Rmax . The star cluster under study which is in the center of the field comprise a subset *C* of N_C unknown stars. The number of stars up to a given radius from the center can be computed from the data as: $n(r) = |\{r' \in S \,/\, r' < r\}|$, which allows to compute the empirical cumulative distribution function (ECDF, $\hat{F}_R(r)$) of the random variable r. This is an estimator of the underlying cumulative distribution function $F_R(r)$, defined as the probability of randomly drawing a value of the variable R less or equal than a given value r:

$$F_R(r) = P(r' \in S \,/\, r' \leq r) \longleftarrow \hat{F}_R(r) = n(r)/N$$

By the law of large numbers, the ECDF approaches the cumulative distribution when the sample size becomes large (see the figure).

The estimated cumulative distribution function is related by definition to the density distribution function by:

$$\int_a^b \hat{f}_R(r)dr = \hat{F}_R(b) - \hat{F}_R(a) = P_R(a < r < b) \longrightarrow f_R(r) = \frac{dF_R(r)}{dr}$$

The ECDF(r) of all the stars in a field centered at a cluster is the sum of the contributions of the cluster and background ECDFs, whose difference as a function of r needs to be inferred. Assuming a uniform background, the background cumulative distribution can be modeled by $\hat{F}_R^{\,b}(r) = f_N + \alpha\, r^2$, where f_N and α are the coefficients resulting from a least squares fit to the background, at sufficiently large distances to the cluster. The first term is due to the excess of stars with respect to the background, given by the presence of the star cluster, and is indeed an estimate of the fraction of stars in the field that belong to the cluster. Finally, we estimate the ECDF of cluster stars radii as the difference between the full measured cumulative profile and the modeled background profile:

$$\hat{F}_R^s(r) = \hat{F}_R(r) - \hat{F}_R^b(r)$$

Since this is affected by a noise component, we expand it as a sum of orthogonal functions. A suitable basis is that composed by a set of orthogonal harmonic functions plus a linear term. The maximum number of terms is chosen so that the mean behaviour of the distribution is reproduced, but the high frequency component given by the finite sampling is filtered. Then, the probability density distribution estimate is obtained by a simple differentiation:

$$\hat{F}_R(r) = r + \sum_1^k a_k sin(k\,\lambda\, r) \longrightarrow \hat{f}_R(r) = 1 + \sum_1^k a_k\, k\, \lambda cos(k\,\lambda\, r)$$

which gives the probability of having a star (in the observed field) at a distance r from the cluster center.

From the previous equations, the cumulative profile of the cluster is estimated by:

$$\hat{F}_R^s(r) = \hat{F}_R(r) - (f_N + \alpha\, r^2)$$

According to this, the probability of having a star belonging to the star cluster at a distance $[r, r+\delta r]$ from the cluster center estimate, is given by:

$$\Delta(r, r+\delta r) = \frac{\hat{F}_R^s(r+\delta r) - \hat{F}_R^s(r)}{A(r+\delta r) - A(r)}$$

where *A(r)* is the area enclosed in a circle of radius *r*. The curvature of the sky is neglected since the angular coverage of the region in study is small, so the formula of the area of a circle is that of a circle in a plane. This expression can be written without change as follows:

$$\Delta(r, r+\delta r) = \frac{\frac{\hat{F}_R^s(r+\delta r) - \hat{F}_R^s(r)}{\delta r}}{\frac{A(r+\delta r) - A(r)}{\delta r}}$$

and then, taking the limit:

$$\rho(r) = \lim_{\delta r \to 0} \frac{\frac{\hat{F}_R^s(r+\delta r) - \hat{F}_R^s(r)}{\delta r}}{\frac{A(r+\delta r) - A(r)}{\delta r}} \longrightarrow \boxed{\rho(r) = \frac{\hat{f}_R(r)}{2\pi r}}$$

2. Estimating the statistical CMD

We construct a smooth function that is an estimate of the color-magnitude distribution of the stars in the star cluster. A single star in the field is an outcome (radius, color, magnitude) from the vector random variable *X=(R,C,M)*. The total sample of stars in the field is then the collection of N identically distributed random variables X. If a star is either in the cluster or in the background, then the set of all stars in the sampsle, S, can be expressed as *S = C + B*, where *C* is the set of cluster stars and *B* is the set of background stars.

Let $R_{\vec{X}}$ be a region within the range of values of *X*, and $A_R = \{\vec{x}/\vec{x} \in R_{\vec{X}}\}$. The probability of finding a star in this region is given by

$$P(A_R) = \int_{R_{\vec{X}}} \hat{f}_X(\vec{x})d\vec{x}$$

The probability of finding a cluster star that is not in the background is:

$$P(\{\vec{x} \in S/\vec{x} \notin B\}) = \int_{R_{\vec{X}}} \left(1 - f_{\vec{X}}^b\right)\, d\vec{x}$$

and the probability of finding a star in that belongs to the cluster in a given region R_{CM} of the CMD:

$$P\left((c,m) \in R_{CM}^s\right) = P\left((c,m) \in R_{CM} \,\wedge\, (c,m) \notin R_{CM}^b\right)$$

which in terms of the density funcions is equivalent to:

$$\hat{f}_{CM}^s = \hat{f}_{CM}\left(1 - \hat{f}_{CM}^b\right)$$

3. Formal fitting of isochrones to the CMD

We define a parametric model in order to describe the physical parameters of an observed cluster. The model parameters can be fitted by comparing the predicted structures in the color-magnitude diagram density distribution to the estimated distribution resulting from the background subtraction procedure. The main features on the CMD can be described by a simple stellar population, i.e., a set of stars with a common age and metalicity. A simple stellar population can be generated from a theoretical isochrone, thus using the parameters of the isochrone as the model parameters. To a first approximation, a plausible model can be set by fixing the distance modulus, the age of the isochrone, the metallicity and the reddening. The distance modulus is an estimate of the distance to the cluster, and his impact on the CMD is just a vertical shift with respect to a zero-distance theoretical isochrone. Although the reddening also affects the distance determination, it is simpler to disentangle the contribution of the reddening and treat it separately. Therefore, there are components of reddening both in color and magnitude.

The goodness-of-fit beteen data and models is a key ingredient of the fitting process. Although there is not a unique way to define this function, it must satisfy several conditions to be a useful indicator of the degree to which data is expected to be a random realization of the model distribution, i.e., the probability of the data given the model. It is worth mentioning, for example, that it must be defined in the interval [0, 1], with greater values for model predictions that resemble the observables. In addition, the surface should not be shallow, since that would give rise to overestimated confidence intervals. It should also be noticed that the Likelihood function is different to the probability of the model given the data, which is formaly the posterior probability, $P(M|D;\theta)$, in the Bayes theorem.

The Likelihood function is usually defined, following the idea of a χ^2 formulation, in terms of the distances of points to a given curve representing the model. Since we have constructed estimates of the multivariate density distributions, such an approach can not be used. Instead, we define a goodness-of-fit measure on the basis on the similarities between the reduced distributions and the synthetic distribution of a given model. Since both are smooth functions normalized to the total volume under the likelihood hypersurface, we use the differences between model and observed bin heights as a measure of likelihood. If we assume that the measures of all bins are independent, then the joint probability of obtaining an observed matrix like the matrix resulting from the data is the product of all individual probabilities:

$$P(Data|Model) = \prod_{i=1}^{N_{bins}} N(\mu = B_i^{obs} - B_i^{teo}, \sigma)$$

where *N* is the Normal function. This value is chosen so that the Likelihood surface has a conspicuous peack around the best model. Once this function is defined, it can be used to make the fitting process by Monte Carlo Markov chains.