Probabilistic description of stellar ensembles

Miguel Cerviño

Abstract I describe the modelization of stellar ensembles in terms of probability distributions. This modelization is primary characterized by the number of stars included in the considered resolution element whatever its physical (stellar cluster) or artificial (pixel/IFU) nature. It provides a solution of the *direct problem* of characterize probabilistically the observables of stellar ensembles as a function of their physical properties. In addition, this characterization implies that intensive properties (like color indices) are intrinsically biased observables, although the bias decreases when the number of stars in the resolution element increases. In the case of a low number of stars in the resolution element ($N < 10^5$), the distributions of intensive and extensive observables following non trivial probability distributions. Such situation a can be computed by means of Monte Carlo simulations where data mining techniques would be applied.

Regarding the *inverse problem* of obtain physical parameters from observational data, I show how some of the scatter in the data provides valuable physical information, since related with the system size (and the number of star in the resolution element). However, to make use of such an information it is needed to follow iterative procedures in the data analysis.

1 Introduction

We know for sure that galaxies are formed by stars. We also know that not all stars are equal, but they have different characteristics depending on some physical parameters, like their mass, metallicity and evolutionary stage. Observationally, in a first approximation neglecting the particular peculiarities of each individual star, we can classify stars according their position in a color-magnitude diagram (maybe one of the greatest success in the application of pre-computational data mining to as-

Miguel Cerviño
IAA-CSIC, Placeta de la Astronomía s/n, 18008 Granada, Spain, e-mail: mcs@iaa.es

trophysics). I show in Fig. 1 the color-magnitude diagram obtained from *Hipparcos* data[1]. Such a diagram shows, at least, two relevant features:

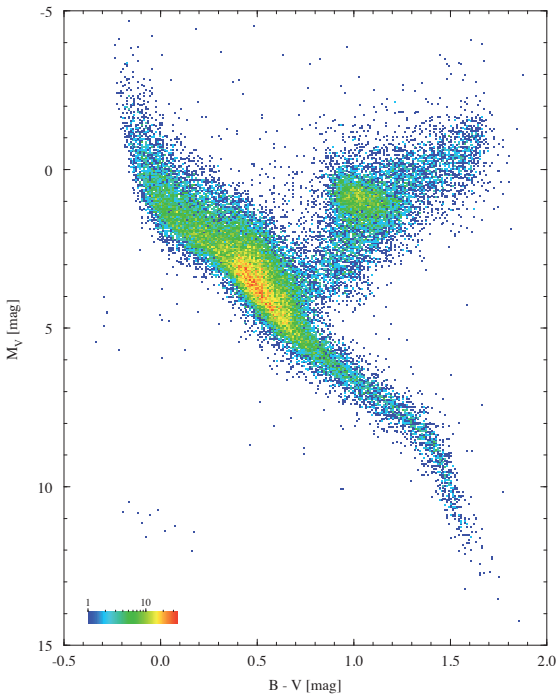

Fig. 1 Hertzsprung-Russell (M_V, $B-V$) diagram for the 41704 single stars from the *Hipparcos Catalogue* with relative distance precision $\sigma_\pi/\pi < 0.2$ and $\sigma_{(B-V)}$ less than or equal to 0.05 mag. Colours indicate number of stars in a cell of 0.01 mag in $(B-V)$ and 0.05 mag in V magnitude (M_V).

- Stars are located in particular regions of the diagram. Currently we know that such regions are the solutions of stellar evolution theory when collapsed in particular observable axes, so only particular regions of the color magnitude diagram are allowed. We can easily identify different areas according the *evolutionary state* of the stars in the sample. As an example, the Main Sequence (MS, nuclear Hydrogen burning phase) runs from top-left to bottom-right the figure, and the Red Giant (RG) phase lies in the middle-right area. Each evolutionary stage is characterized by the internal structure of the star, which is defined by the mass and metallicity of the star at birth and the age of the star.
- Not all regions containing stars have similar density. We also know that it is due to two different reasons:

  1. The density of the area is proportional to the time spend in each evolutionary phase, so the MS, where stars last 90% of their live, are more populated that RG phases. Also, the lifetimes of different Post-MS phases explain the relative stellar densities in the color-magnitude diagram for Post-MS regions.

[1] Caption and figure taken from the *Hipparcos* site at `http://www.rssd.esa.int/index.php?project=HIPPARCOS`

2. However, the variation of density along the MS cannot be explained just by the fact than the more massive the star, the more luminous and the faster consumption of their nuclear Hydrogen fuel; neither by the different ages of the stars in the sample. Massive stars are *intrinsically* less common than low mass stars: Stars of different masses are not formed with equal probability, but the mass distribution of stars *at birth*, $m_{t=0}$ follows a probability distribution called the Initial Mass Function (IMF, $\phi(m_{t=0})$), which, at least in its upper mass range ($m_{t=0} > 2\mathrm{M}_{\odot}$), can be approximated by a power law, $\phi(m_{t=0}) \propto m_{t=0}^{-\alpha}$, with exponent $\alpha \sim 2.35$ obtained by Salpeter [8].

In the case of color-magnitude diagrams, making use of stellar evolution theory, we can obtain the physical properties of each star in the sample: ages, stellar masses (e.g. VOSA by Bayo et al. [1]), and from this information, we can obtain properties of the ensemble as an entity (age of a cluster, IMF, star formation processes in a region, amount of gas transformed into stars, etc...).

Of course, we can obtain the maximum information about an stellar ensemble when we know all the components in the ensemble. However, it is not the common case. Even in deep observations of resolved stellar clusters there are stars so dim that are not detected. In a more general case, we have no access to the emission of the individual stars, but just to the emission of the total ensemble, without further information of the individual components. It is the common case in extragalactic studies.

2 Modeling stellar ensembles

The modeling of stellar ensembles aims to provide information about the physical parameters of an stelar ensemble (star formation history, mass of the system, chemical evolution history) from just the integrated light obtained from the ensemble. Mathematically it means to recover a the primitive of a definite integral. Although the problem looks to be highly degenerate it can be solved (or at least we can suggest a suitable range of solutions) thanks to the restrictions imposed by stellar evolution theory, as it is the case of anlysis of color-magnitude diagrams. Let me explain it in some detail.

The emission of the ensemble is usually dominated by just a few high luminous stars, and most of low luminous stars in the system (i.e. the ones that defines the total mass in the system) are undetected. In the other hand, the most luminous stars are Post-MS stars, which relative densities in each evolutionary phase is just proportional to the lifetime of hte phase. These lifetimes depends on the initial mass and the age of the stars in the Post-MS. In addition, there is a proportionality between the density of Post-MS stars (which dominates the integrated light) and the MS stars (with a low contribution to the integrated light) given by the IMF. Finally, the different relative contributions are strongly dependent on the observed wavelength range. So, combining the information from different wavelengths we can make inferences about the Post-MS population and infer from that the physical properties of

the ensemble including the total amount of mass into stars, star formation histories etc.

This situation is related with the properties of the, so called, *wild distributions* [9], or distributions where the highest possible value, although with a low probability, is able to dominate the mean value of the distribution. The *wild distribution* responsible of the success on to obtain information from the integrated light is the stellar Luminosity Distribution Function, sLDF, it means, the probability of find an star with a given luminosity. Let us illustrate it with a simple example (we refer [5] for more details):

Let us assume a system where all stars are in the MS and that the stars follows a mass-luminosity relation $\ell \propto m^{\beta}$. Assuming a power-law IMF, $\phi(m) \propto m^{-\alpha}$, we can define the sLDF $\varphi_{\rm L}(\ell)$ as:

$$\varphi_{\rm L}(\ell) = \phi(m) \times \left(\frac{d\ell(m)}{dm}\right)^{-1} = A\ell^{-\frac{\alpha}{\beta}} \cdot \frac{1}{\beta}\ell^{-\frac{\beta-1}{\beta}} = \frac{A}{\beta}\ell^{\frac{1-\alpha-\beta}{\beta}}. \qquad (1)$$

being A a normalization constant so $\varphi_{\rm L}(\ell)$ is normalized to one. The mean value of the sLDF is then:

$$\mu_1' = \frac{A}{\beta}\int_{\ell_{\rm min}}^{\ell_{\rm max}} \ell \cdot \ell^{\frac{1-\alpha-\beta}{\beta}} d\ell = \frac{A}{1+\beta-\alpha} \cdot \left(\ell_{\rm max}^{\frac{1+\beta-\alpha}{\beta}} - \ell_{\rm min}^{\frac{1+\beta-\alpha}{\beta}}\right). \qquad (2)$$

If $1+\beta-\alpha > 0$, the mean luminosity is driven by $\ell_{\rm max}$. In a typical situation with $\beta \approx 3$, the most luminous stars will dominate the luminosity if $\alpha < 4$: this is the case of Salpeter's IMF [8].

Trivially, if $\varphi_{\rm L}(\ell)$ is normalized to the number of stars in the ensemble, $\mathscr{N}_{\rm tot}$, the value obtained, that is $\langle \mathscr{L}_{\rm tot} \rangle = \mathscr{N}_{\rm tot} \times \mu_1'$, corresponds to the *mean* total luminosity of the ensemble (I will back to this point latter). When Post-MS stars enter in the game the situation is ever more extreme since their luminosity are even larger than the one they had in the MS. So, the sLDF turns into a power-law distribution due to MS stars plus a high luminosity tail with variable structure (according the age of Post-MS stars). Given that the *mean* amount of gas transformed into stars, $\langle \mathscr{M}_{\rm tot} \rangle$ is also proportional to $\mathscr{N}_{\rm tot}$ (also provided by the IMF) we can obtain age dependent mass to luminosity ratios, $\langle \mathscr{L}_{\rm tot} \rangle / \langle \mathscr{M}_{\rm tot} \rangle$, which allow to obtain a value of $\mathscr{M}_{\rm tot}$ from the observed $\mathscr{L}_{\rm tot}$ once the age is obtained.

The main technique, called evolutionary population synthesis, was mainly developed by B. Tinsley [11] in the 70's. Currently there are several codes which provide the mean values of the sLDF (although normalized to different values and defined as *integrated* emission instead a mean value), like *Starburst99* by Leitherer et al. [7] or Bruzual & Charlot models [2].

However, since we work with a *wild* distribution function (the sLDF) the mean value of such distribution is not the full history: The mean value is not a good proxy to make inferences (contrary to the gaussian case). Our many question know is how the functional form of the sLDF change when we consider ensembles of stars. So we must combine each of the possible star in the ensemble properly [5].

As a general rule, the probability distribution function, PDF, resulting from the sum of several variables is obtained as the convolution of the PDFs of the individual variables. For example, let $\varphi_{\rm x}(x)$ be the PDF of a variable x and $\varphi_{\rm y}(y)$ the PDF of a variable y independent of x. The probability density of a variable $u = x+y$ is given by the product of the probabilities of $\varphi_{\rm x}(x)$ and $\varphi_{\rm y}(y)$ summed over all the combinations of x and y such that $u = x+y$, which is the definition of convolution:

$$\varphi_{\rm u}(u) = \int_{-\infty}^{\infty} \varphi_{\rm x}(z)\,\varphi_{\rm y}(u-z)\,dz = \varphi_{\rm x}(x) \otimes \varphi_{\rm y}(y). \tag{3}$$

In our case, we are assuming that all the stars have luminosities distributed following the same distribution function, $\varphi_{\rm L}(\ell)$, and that the stars are independent on each other. Therefore, the population Luminosity Distribution Function, pLDF, of an ensemble of $N_{\rm tot}$ stars is obtained by convolving $\varphi_{\rm L}(\ell)$ with itself $N_{\rm tot}$ times:

$$\varphi_{\rm L_{tot}}(\mathscr{L}) = \overbrace{\varphi_{\rm L}(\ell) \otimes \varphi_{\rm L}(\ell) \otimes \ldots \otimes \varphi_{\rm L}(\ell)}^{\mathscr{N}_{\rm tot}}. \tag{4}$$

Hence, if the sLDF is known, the pLDF of an ensemble of $\mathscr{N}_{\rm tot}$ stars can be computed by means of a convolution process. Self-convolutions have some additional interesting properties, in particular that the cumulants of the pLDF are just $\mathscr{N}_{\rm tot}$ times the cumulants of the sLDF. So, trivially,

$$\begin{aligned} \mu_1'(\mathscr{L}) &= N_{\rm tot}\,\mu_1(\ell), & \kappa_2(\mathscr{L}) &= \sigma^2(\mathscr{L}) = N_{\rm tot}\,\kappa_2(\ell), \\ \gamma_1(\mathscr{L}) &= \frac{1}{\sqrt{N_{\rm tot}}}\,\gamma_1(\ell), & \gamma_2(\mathscr{L}) &= \tfrac{1}{N_{\rm tot}}\,\gamma_2(\ell), \end{aligned} \tag{5}$$

where κ_2 is the variance and γ_1 and γ_2 are the skewness and the kurtosis of the corresponding distribution. Note that, in agreement with the central limit theorem, $\gamma_1(\mathscr{L}) \to 0$ and $\gamma_2(\mathscr{L}) \to 0$ for large enough $N_{\rm tot}$ values, i.e. the distribution tends to a Gaussian with a relative dispersion which also tends to zero.

Just for reference values, a gaussian approximation of the pLDF is reached for stellar ensembles with total mass $\mathscr{M}_{\rm tot} > 10^5 {\rm M}_\odot$ for visible bands and $\mathscr{M}_{\rm tot} > 10^7 {\rm M}_\odot$ for infrared bands [3, 5].

Previous relations are useful to unveil the scale properties of LDFs, and obtain situations where the pLDF can be properly approximate by a gaussian, so its mean value can be use as a proxy for data analysis (I will back to this point latter). However, it is not sufficient for current astronomical research: the increasing spacial resolution and sensibility of current facilities implies a reduction in the number of stars per resolution element (pixel, IFU, etc); the observation of faint sources provide access to systems with an intrinsically low number of stars; the drastic reduction of the observational error provides that observational data shows the *physical variance* (due to the pLDF variance among others) of stelar ensembles.

The convolution method, although theoretically plausible, contains technical difficulties: the wild nature of the sLDF, including gaps and bumps in the high luminosity tail due to fast stellar evolutionary phases, needs a high resolution in the binning of the sLDF for convolution. Alas, the large dynamic range in luminosities, from 10^{-2} $L_\odot$ to 10^6 $L_\odot$, makes the numerical computation unfeasible. So Monte Carlo simulations are more useful to describe the resulting pLDF outside the gaussian regimen.

3 Data mining on stellar ensembles simulations

The needing of sampling the pLDF for different situations is not just the only reason to use Monte Carlo simulations. In the previous Sections I have just shown the case of a single pLDF, which would corresponds to a given band or wavelength bin. But a real analysis of observational data makes use of several bands or wavelength points. Given that different regions of the possibles sLDF (with a fixed set of ensemble physical conditions) are dominated by the same set of stars in particular evolutionary stages, strong (non necessarily linear) correlations between the sLDFs are expected. In addition, the distribution nature of extensive quantities produce non trivial rupture of the intensive character of assumed intensive quantities like color and spectral indices commonly used in data analysis (see Fig. 2 below and [6] for details). A formal solution is a multidimensional convolution process with a number of dimensions similar to the wavelength resolution in our observations, but it is technically unfeasible as in the case of simple pLDFs, and, currently, the problem remains unsolved.

Additional advantages of Monte Carlo simulations is to study situations where the pLDF shows a bimodal behavior. These situations are expected for stelar ensembles with a number of stars such the mean luminosity of the pLDF is near maximum luminosity $\ell_{\rm max}$ of the sLDF [4]. Bimodal distributions also appear when the simulations make use of power-law distributions of $\mathscr{N}_{\rm tot}$. Unfortunately, there is no way, but just Monte Carlo simulations, to identify and explore the situations when its happens (see [4, 5] for more details).

The situation can be strongly improved by the use of Data Mining techniques over simulations. As an example, I show in Fig. 2 a serendipity result of the analysis of the Monte Carlo simulations of young stellar ensembles (see the caption and [12]

for details). Although the result, once found, can be explained by the wild nature of the corresponding pLDF for a low number of stars ($\mathscr{N}_{\rm tot} < 10^4$), it was not expected *a priori* when simulations where performed.

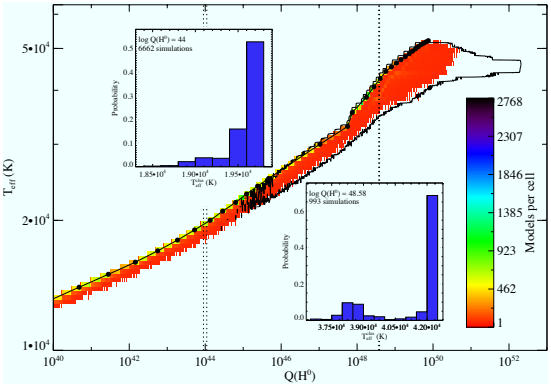

Fig. 2 Extensive, $Q(\mathrm{H}^0)$, vs. Intensive, $T_{\rm eff}$, quantities in the case of stellar population Monte Carlo simulations. Note that the intensive quantity (formally independent of the size of the sistem) is not longer *intensive* in the case of low populated clusters, but correlates strongly with the extensive quantity. Also note bimodal features (right-bottom box) in the region just before the intensive quantity becomes really intensive. Figure from [12].

4 The inverse problem: induced sampling

I must remind that interesting the results of stelar ensembles Monte Carlo simulations would be, the final goal of the simulations is to provide analysis tools to infer physical quantities from observational data. Stellar ensembles models, whatever are used in the form of pLDF mean values, cumulants oor the whole distribution, have an intrinsic undefined parameter: the number of stars in the ensemble, $\mathscr{N}_{\rm tot}$, which is in fact one of the physical parameters aimed to obtain from the models (remember the discussion about the $\langle \mathscr{L}_{\rm tot} \rangle$ / $\langle \mathscr{M}_{\rm tot} \rangle$ relationships).

Even in the case of gaussian pLDF, not just the mean value of the distribution must be correctly fitted, but also their associated variance (which intimately depends on $\mathscr{N}_{\rm tot}$). The only way to do that is to use traditional methods to guest a value of the physical parameters in the model comparison, and iterate the method up using the variance of the pLDF as a metric of fitting. Obviously, the method is not valid for no gaussian distributions the the meaning of the mean and variance can not be translated neither to representative values nor confidence intervals. So, new methods for analysis are needed in that cases.

Finally, for the case of observation with spatial resolution, we can advantage of the intrinsic distribution of $\mathscr{N}_{\rm tot}$ in the observational set: different resampling of the observational set (varying artificially the size of the resolution element) must produce self consistent results in terms of physical parameters, since related ultimately with the sLDF of the system, with scale in mean value and variance with $\mathscr{N}_{\rm tot}$ of the considered resolution element. This methodology of induced sampling provides

additional test about our inference of $\mathscr{N}_{\rm tot}$ in the system. This method is similar to bootstrapping the data, but including the physical model (the $\mathscr{N}_{\rm tot}$ dependent pLDF) in the analysis of stelar ensembles.

However, the methodology I proposed here, has not been yet developed properly in no analysis method (as far as I know).

Acknowledgements I acknowledge Valentina Luridiana for the developing of the probabilistic theory of population synthesis all along several years. I also acknowledge the third author of the [5] paper (only available in the astro-ph version of the paper) for a practical example of wild distribution in real life. I acknowledge Luisma Sarro Baró the opportunity to attend this meeting among many others things. This work has been supported by the MICINN (Spain) through the grants AYA2007-64712 and AYA2010-15081.

References

[1] Bayo, A., Rodrigo, C., Barrado Y Navascués, D., Solano, E., Gutiérrez, R., Morales-Calderón, M., & Allard, F. 2008, A&A, 492, 277
[2] Bruzual, G., & Charlot, S. 2003, M.N.R.A.S., 344, 1000
[3] Buzzoni, A. 1989, Ap.J.S.S. , 71, 871
[4] Cerviño, M., & Luridiana, V. 2004, A&A, 413, 145
[5] Cerviño, M., & Luridiana, V. 2006, A&A, 451, 475
[6] Cerviño, M., & Valls-Gabaud, D. 2003, M.N.R.A.S., 338, 481
[7] Leitherer, C. et al. 1999, Ap.J.S.S., 123, 3
[8] Salpeter, E. E. 1955, Ap.J., 121, 161
[9] Sornette, D. 2004, *Critical phenomena in natural sciences : chaos, fractals, selforganization and disorder : concepts and tools*, 2nd ed. by Didier Sornette. Springer series in synergetics. Heidelberg: Springer
[11] Tinsley, B. 1980, *Fun. Cosm. Physics* 5, 287
[12] Villaverde, M., Cerviño, M., & Luridiana, V. 2010, A&A, 522, A49