Partition entropy and $\chi^2$ error.  
The improved MEMPHIS algorithm - Part I

Rafael Cubarsi & Santiago Alcobé

Dept. Matemàtica Aplicada IV
Universitat Politècnica de Catalunya
08034 Barcelona, Catalonia, Spain
E-mail:rcubarsi@ma4.upc.edu

July 2009

Abstract

The entropy of the population partition is studied as a function of the sampling parameter, so that within a particular interval of its graph, the plateau region, it is possible to get a stable estimation of the mixture parameters. The optimal estimation is associated with a local maximum of entropy. Alternatively, the $\chi^2$ error of the mixture approach may also be used to obtain an optimal segregation. The relationship between the fitting error and the population entropy has been analysed in detail. We have proved that, by using an appropriate sampling parameter, within a plateau region of the entropy graph, a local entropy maximum takes place simultaneously with a local minimum of the $\chi^2$ error. Therefore, the combined statistical method provides the best approximation mixture, as well as the less informative partition, to estimate the kinematic parameters of populations.

KEY WORDS: stars: kinematics – galaxies: kinematics and dynamics – galaxies: statistics – methods: statistical.

1 Introduction

The way in which a stellar sample is selected may obviously yield to very different parameter estimates to classify stellar populations. In general, a sample may contain stars belonging to the Galactic major components, like bulge, disk and halo, although each component can be decomposed in several subcomponents or stellar populations attending to some astrophysical properties and, in particular, to kinematics, which will focus our present attention. Usually the so-called population representative samples are drawn from large catalogues, like the Geneva-Copenhagen survey (GCS) of the Solar neighbourhood (Nordström et al. 2004), where the stars are chosen or rejected through an individual inspection in order to obtain a sample that suits, for example, some typical disk features. However, star selection attending to astrophysical criteria, like metallicity, may likely yield to some kinematically biased samples and population estimates, which would then match what theoretically we believe is a disk sample, and may then lead to the error described by the well known Conan Doyle’s character, ”Insensibly one begins to twist the facts to suit theories, instead to theories to suit facts”.

To avoid such a previous inspection of star catalogues Alcobé & Cubarsi (2005) (hereafter Paper I) had developed a purely statistical method named MEMPHIS algorithm (Maximum Entropy of Mixture Probability from Hierarchy Segregation), which was based in the principle of insufficient reason. The purpose was to obtain population representative subsamples, and to segregate them. In Paper I, for a local sample drawn from the HIPPARCOS catalogue (ESA 1998) containing the full set of space velocities, the neighbourhood stars were classified as belonging to the thin or thick disk, although within the thin disk it was also possible to distinguish between early-type, young-disk and old-disk stars. The existence of an intermediate population between the major disk components was rejected, since the appearance of a continuous transition between them was proved to be due to the overlapping wings of the major disk components. However, since halo stars are scarce in the solar neighbourhood, it was not possible to identify such a Galactic component. In the current work MEMPHIS algorithm is further analysed and improved in order to classify minor populations like the local halo. The application to local stellar samples is performed in Cubarsi & Alcobé (2009).

1.1 How the segregation works

MEMPHIS allows to identify the more representative mixture of populations among a number of feasible mixtures. The algorithm relies on the entropy of the partition induced by two statistical populations, which are assumed to be contained in the whole sample (Pop-I for the one with larger size, and Pop-II for the minor one). Instead of adapting the velocity distribution function to the whole sample, we choose the stellar subsample which best suits a two-normal trivariate mixture, which are the “a priori” less informative distributions. A brief outline of the procedure is given, which was explained in detail in Paper I, and it is reviewed in §2. Firstly, a sampling parameter has to be defined in order to introduce a hierarchy into the sample, so that a set of nested subsamples could be recursively drawn from the total sample, each one containing at least the same or an increasing number of stellar populations than previous subsamples. Some properties, which are associated with isolating integrals of the star motion (like the velocity component perpendicular to the Galactic plane, the rotational velocity, or the absolute value of the total velocity) can be used as sampling parameters to sort or to discriminate populations. Then, the bimodal pattern of two gaussian distributions is applied in order to identify differentiated kinematic behaviours within the subsample, by associating them with Pop-I and Pop-II populations. The segregation
method is designed to work under the circumstance where, by entering stars of a third population into the subsample, the populations are recursively redefined by joining both previously existing populations into Pop-I, and by reserving Pop-II for the newer population. This redistribution produces a discontinuity in the entropy of the mixture probability, which is seen as a drop in the entropy graph. The subsample corresponding to a sampling parameter providing the maximum entropy of the partition is taken as the more representative subsample to describe the actual population mixture, among all the set of nested subsamples containing the same number of stellar populations. In general, the criterion of maximum entropy of the population mixture can be applied in an independent way of the segregation method.

The algorithm was proposed as an alternative and more accurate procedure instead of choosing the partition minimising the $\chi^2$ fitting error of the mixture approach, as in Cubarsi & Alcóbé (2004), hereafter Paper II. The $\chi^2$ error basically gives the uncertainty of expressing the whole sample as a mixture of two gaussian distributions, and allows to describe mixtures of populations which are near of statistical equilibrium, like those of the main disk components, in contrast to other specific subpopulations, like early and young-disk stars, which lead to greater $\chi^2$ error.

The segregation method we used in Papers I and II was a variant of the moments method, which took into account the whole space of the heliocentric velocities, $v \in \mathbb{R}^3$, without any additional hypothesis about the symmetry of the distribution. Let us review the way in which the $\chi^2$ error was computed. For two normal populations, with mixture proportions $n'$ and $n''$, so that $n' + n'' = 1$, and with normalised probability density functions $f'$ and $f''$, the total velocity distribution is given by

$$f(v) = n' f'(v) + n'' f''(v)$$

The elements of the $p$-order tensor of moments are evaluated according to the expected value of the $p$-tensor powers of the velocity as $E[v_{i_1} \ldots v_{i_p}]$, for velocity component indices $i_k \in \{1, 2, 3\}$. In general, up to a given order, a set of $R$ relationships between total and partial moments is obtained, which can be written as a vector of $R$ equations in the following form

$$M_k = n'M'_k + n''M''_k$$

where $M_k$ are data computed from a working sample, with known sampling errors, and the right-hand side contains the moment estimates for both populations assumed to be contained in the sample. The partial moments are computed from a set of $\nu$ mixture parameters, which are the unknowns of the segregation problem. These parameters are the mixture proportions, and population means and covariance matrices. Nineteen unknowns in total. Thus, if $\varepsilon_k$ denotes the standard error of $M_k$, a calibration of the approximation can be done from the weighted mean of the squared errors,

$$\chi^2 = \sum_{k=1}^{R} \frac{1}{\varepsilon_k^2} |M_k - n'M'_k - n''M''_k|^2$$

The random variable $\chi^2$ is expected to have a chi squared distribution with $n = R - \nu$ degrees of freedom. A detailed analysis of the method can be found in Paper II, where, instead of working from the sample moments, the $k$-statistics of the sample were used, which are unbiased estimators of the distribution cumulants. Eq. 1 was considered up to order four, which is a set of 34 scalar equations. The standard errors were computed by using up to eighth-order $k$-statistics, being also proportional to a factor which, for large samples, depends on the inverse of the sample size.

With either of both criteria, maximum entropy or minimum $\chi^2$ error, the procedure for obtaining population representative subsamples as continuously increasing the sampling parameter, up to exhaust the whole sample, could be recursively applied to segregate more than two populations, under some minimal conditions.
1.2 Optimal samples

For any subsample the mixture proportions and population kinematic parameters are computed. They can be slightly different among subsamples, even describing the same mixture of populations. Therefore, we are led to choose an optimal subsample between them. If this subsample is selected according to the minimum $\chi^2$ error of the fit, we are then provided with the best mixture approach, though it has two disadvantages. As increasing the size of the sample the quadratic error dramatically decreases, which, in a natural way, blinds other feasible and perhaps better mixtures with minor subsample size. On the other hand, the quadratic error may have several local minimums where to choose. Alternatively, if the selection is done attending to the maximum partition entropy, the optimal subsample is then the most representative among those containing the same populations. That is, the less informative subsample which is consistent with the partition induced by the stellar populations. However, the entropy criterion does also present some troubles. For example, if the means of the partial distributions are too close, or their wings are too much overlapped, the inclusion of a new population in the nested subsamples does not produce a neat entropy drop. In addition, the entropy graph may also have several local maxima. Fortunately, by combining both preceding criteria it is possible to obtain an optimal segregation sample. Under some optimal conditions –where the new stars which are merged to the subsamples as increasing the sampling parameter do not introduce any significant variation of the population parameters, and therefore the new entering stars only contribute to build a more representative mixture– it is possible to obtain simultaneous local extremes of fitting error and entropy. Of course, both criteria work on the basis of an appropriate sampling parameter. Moreover, it is possible to use different sampling parameters depending on the populations to be segregated.

In the present work two new aspects of MEMPHIS algorithm are analysed. First, the use of several isolating integrals of the star motion as sampling parameters is studied. The integrals of motion determine univocally the orbit of each star in the phase space, so that we may expect that stars of the same stellar population have integrals which smoothly vary within a characteristic range of values, that may sensibly differ from stars belonging to other populations. However, as continuously scanning a stellar sample in terms of an integral of motion, discontinuities can be only observed if we are able to create a discrete partition of stellar populations. Second, the joint variations of entropy and $\chi^2$ error are studied. Once a new population is merged to the cumulative subsample, there is an interval of the sampling parameter where the entropy smoothly increases, and sometimes remains nearly constant up to the complete incorporation of the new population. We call it “plateau region” of the entropy graph, and it occurs when the core distribution of the new entering population has already been merged to the subsample. This nearly constant entropy allows a good estimation of the population parameters, since their computation is very stable. In such a plateau region it is shown that a local maximum of entropy occurs simultaneously with a local minimum of $\chi^2$. Moreover, while for large samples the minimum fitting error can be nearly undetectable, the entropy maximum is much more visible. Therefore, what was firstly a principle of insufficient reason, now is converted in a complementary criterion to choose a significative local minimum of the quadratic fitting error. Or, conversely, to refine the selection of the more representative subsample, within the plateau region we shall choose the sampling parameter giving the minimum fitting error.
2 Maximum partition entropy

According to Liouville’s theorem, the phase space density function can be explicitly expressed in terms of the isolating integrals of the equation of motion of a star for a given potential. The integrals of motion determine univocally the orbit of any star in the phase space, as well as the phase space density function, which is itself an integral of motion. But the phase density function, by its physical significance, must be a one-valued function of the six phase coordinates. Therefore, for a fixed time, the integrals define an uniparametric family of level curves, so that the phase space remains decomposed in a set of disjoint hypersurfaces corresponding to different integral values. Hence, for a given integral value $J = J_0$, a subsample containing stars with $J \leq J_0$ can be build up, and greater values of $J_0$ determine an increasing set of nested subsamples.

In general analytic integrals are always isolating, and up to three isolating integrals are found for all orbits under steady-state and axisymmetric potentials (e.g. Gilmore et al. 1989). Thus, by expressing the velocity components $(U, V, W)$ in a cartesian heliocentric coordinate system, with $U$ toward the Galactic centre, $V$ in the Galactic rotational direction, and $W$ perpendicular to Galactic plane, positive in the direction of the North Galactic pole, for a stationary potential $U$ the energy integral can be written as

$$I_1 = U^2 + V^2 + W^2 + 2U(x, y, z)$$

(4)

The integral for the axial component of the angular momentum, under an axisymmetric potential, is expressed in cylindrical coordinates $(r, \theta, z)$ as

$$I_2 = rV$$

(5)

Under a separable potential $U = U_1(r) + U_2(z)$, which is valid near the Galactic plane, it is also obtained a third integral, sometimes called Oort’s integral,

$$I_3 = W^2 + 2U_2(z)$$

(6)

Of course, any combination of above integrals is also conserved. Let us point out two very simple cases. For a fixed position, the quantity

$$I_4 = U^2 + W^2$$

(7)

is also an isolating integral. Similarly, the quadratic function

$$I_5 = U^2 + \alpha(V - V_0)^2 + \beta W^2$$

(8)

for any $V_0$, $\alpha$ and $\beta$ either constant or depending on the position, is also conserved. The later may be generalised, under appropriate hypotheses, to arbitrary quadratic functions in the peculiar velocities (Chandrasekhar 1942), which justifies the generalised use of ellipsoidal-type velocity distributions.

Therefore, any of the foregoing integrals induce a hierarchy in the phase space, which may be used to build a nested set of stellar subsamples, as well as to sort the stars within any particular stellar sample. The isolating property of the integrals of motion is indeed very close to the conditions required in Paper I for a sampling parameter, which will be reviewed in section §2.2.

Thus, let $P$ represent a sampling parameter associated with a property defining a stellar sample, in the sense that it determines a boundary for the stars composing the sample. It can be then obtained from either of the above isolating integrals of motion, for example, in a particular volume.
of the Galaxy, we may consider the maximum absolute value of the perpendicular velocity to the galactic plane, which would depend on the third integral $I_3$, or the greatest absolute value of the galactocentric star velocity, which is computed from the energy integral $I_1$. Similarly we may work with local quadratic integrals, referred to the local standard of rest of the population mixture, which would be obtained from $I_5$, as a combination of energy and angular momentum integrals, or even from the three basic isolating integrals, depending on the dynamical model. Specific requirements for such a parameter will be discussed later.

### 2.1 Entropy of the mixture

Then, by scanning the whole stellar sample with the sampling parameter $P$, a subsample $S(P)$ may be drawn, with a number of stars noted by $#S(P)$. We assume that a finite number $c$ of stellar populations is present in $S(P)$, according to a population partition $A = \{ A_1, A_2, ..., A_c \}$. For fixed time and position, the velocity density function $f(v)$ is therefore expressed as a mixture of $c$-partial density functions $f^{(i)}(v|X)$, which are associated with stellar populations depending on a set of mixture constants, namely $X$. In Cubarsi (1992) it was proved that the mixture proportions and the kinematic parameters of each population depend on some mixture constants, which are obtained in terms of the total cumulants of the sample. Thus, we can write

$$ f(v) = \sum_{i=1}^{c} p(A_i|X) f^{(i)}(v|X) \quad (9) $$

where $p(A_i|X)$ is the mixture proportion of the $i$-th population. Obviously $\sum_{i=1}^{c} p(A_i|X) = 1$. We are however interested in to stress the relationship between the mixture constants and the sampling parameter. Hence, for a subsample $S(P)$ we write $X = X(P)$ and $p(A_i|X(P)) = n^{(i)}(P)$. Thus, the density function becomes,

$$ f(v|P) = \sum_{i=1}^{c} n^{(i)}(P) f^{(i)}(v|X(P)) \quad (10) $$

Then, the entropy $H(A)$ of a partition $A = \{ A_1, A_2, ..., A_c \}$, each subset with probability $n^{(i)}(P)$, can be written explicitly depending on the parameter $P$ as follows

$$ H(A|P) = \sum_{i=1}^{c} -n^{(i)}(P) \ln(n^{(i)}(P)) \quad (11) $$

Above equation can be interpreted as the expected value of the uncertainty $I(A_i|P) = - \ln(n^{(i)}(P))$, so that the entropy variations measure the uncertainty variations in the mixture proportions (e.g. Papoulis 1989). The greater the entropy, the less the information of the population partition, and therefore, of their population parameters. In general, the most meaningful parameters are those describing largest populations, instead of the ones describing an excessively informative partition, for example with very few stars in a component (or even with one star alone). Hence, we are interested in to determine a sample $S(P)$ providing some mixture parameters associated with maximum values of $H$. This will be done by tuning the parameter $P$.

Indeed only a two-component mixture model is needed, since the entropy-based algorithm will detect entropy discontinuities every time that a new population is merged to the sample. In a two-component partition $\{ A_1, A_2 \}$, for a fixed sampling parameter $P$, by writing $n \equiv n^{(1)}$, $n^{(2)} = 1 - n$, we can express the entropy as

$$ H(n(P)) = -n(P) \ln(n(P)) - (1 - n(P)) \ln(1 - n(P)) \quad (12) $$
We assume hereafter that the first population $A_1$ is the more prominent one, $n \geq \frac{1}{2}$. Then, $H$ as a function explicitly depending on $n$ is a positive, decreasing and differentiable function which satisfies

$$0 < H(n) < \ln 2 \quad \text{and} \quad \frac{dH}{dn} < 0, \quad \text{for} \quad \frac{1}{2} < n < 1$$

(13)

In the following section we shall examine the behaviour of $H$ in the case of a three component partition.

### 2.2 Sampling parameter

The sampling parameter associated with an isolating integral induce an order in the velocity space so that, as increasing its value, a set of new stars are merged to the sample. However, the new stars may belong to a single population, or, in most cases, may be a mixture of several populations. Although it is not possible to find an ideal sampling parameter that distinguishes between single populations, under some minor requirements it is feasible to use a parameter that, depending on its value, gives priority to stars of some particular populations in front of others. For example, in a given volume of the Galaxy, the absolute value of the vertical velocity referred to the Galactic plane, that is the integral $I_3$ from Eq. 6, selects for low values thin and thick disk stars, and no stars from the halo. In contrast, for higher values, it allows the entry of halo stars alone. Similarly, the integral $I_4$ from Eq. 7, as we shall see in section §4, for low values selects thin disk stars, while greater values favours the income of thick disk and halo stars in an indiscriminate way. Thus, we wish to use sampling parameters that are, in some way, population informative. To this purpose they must fulfil some properties, which we tried to synthesise in Paper I in the four following conditions, that now we reproduce and improve, in order to give a clearer explanation of the entropy graph.

If a stellar sample is composed of a set of populations $\{A_1, A_2, ..., A_c\}$, then a hierarchical family of subsamples $S(P)$ should progressively incorporate the populations $A_1, A_2$ and so on –by preserving an order– as continuously increasing the parameter $P$. For a two-component partition $\{A_1, A_2\}$ the sampling parameter $P$ should satisfy the following properties:

(I) The number of stars of the subsample $S(P)$ increases with $P$, without losing any star. In other words, for two parameter values $P_1$ and $P_2$, we have

$$P_1 \leq P_2 \Rightarrow S(P_1) \subseteq S(P_2) \Rightarrow \#S(P_1) \leq \#S(P_2)$$

(14)

(II) $P$ should induce an ordered incorporation of stars to the subsample $S(P)$, so that, as $P$ increases, stars of population $A_1$ are firstly included in the sample and, afterwards, stars of population $A_2$ are progressively merged to it. Hence, the mixture proportion of the first population satisfies

$$P_1 \leq P_2 \Rightarrow n(P_1) \geq n(P_2)$$

(15)

Such a condition may be also fulfilled if stars of both populations are merged simultaneously to the subsample $S(P)$. Obviously, while the alone $A_1$-population composes the subsample $S(P)$, we have a constant population fraction $n(P) = 1$.

(III) If we assume that the transition between populations is enough smooth, then $n(P)$ is a continuous and differentiable function of $P$. Thus Eq. 15 is equivalent to $\frac{dn}{dP} \leq 0$, being
$n(P)$ a non-increasing function. Under previous assumptions, and taking into account Eq. 13, the entropy $H(n(P))$, or simply $H(P)$, is a non-decreasing function of $P$ in the interval $\frac{1}{2} < n < 1$:

$$\frac{dH}{dP} = \frac{dH}{dn} \frac{dn}{dP} \geq 0$$

(16)

Therefore, the entropy evolves, depending on the parameter $P$, according to the following property

$$P_1 \leq P_2 \Rightarrow H(P_1) \leq H(P_2)$$

(17)

Thus, for a stellar sample composed of two populations, if the second population is merged to a subsample containing only the first population, according to the hierarchy that has been induced by the ideal sampling parameter, the entropy of the mixture probability must be non-decreasing function of the sampling parameter.

However, for a sample with three populations $\{A_1, A_2, A_3\}$ the behaviour of the function $H(P)$ is as follows. Let us assume a continuous incorporation of stars to the subsample $S(P)$ by increasing $P$. Suppose also that all of the $A_1$-population has been completely merged to the subsample, and that the $A_2$-population is partially present in the subsample. Then, if a number of new stars is added to $S(P)$, one of the following situations may occur: Either the stars belong (or are very similar) to the population $A_2$, according to the situation described above, or obviously the new stars belong to a new $A_3$-population. In the later case, when the number of stars belonging to the new population is enough significative, and their kinematic parameters are enough different from $A_2$ and $A_1$-populations, the bimodal segregation model should mix up the two closest populations $A_1$ and $A_2$ into the first component Pop-I, and should reserve the second component Pop-II for the population $A_3$. Therefore, a fourth condition for an ideal sampling parameter is required:

(IV) A partition $\{A_1, A_2, A_3\}$ is reduced to the two-component partition $\{A_1 \cup A_2, A_3\}$, with a cumulative population $A_1 \cup A_2$. In this case the relationship of Eq.15 is no more fulfilled. Instead, the initial values $n = 1$ and $H(P) = 0$ are reset.

The failure of above conditions would indicate that the sampling parameter is not able to induce the desired hierarchy into the subsamples.

2.3 Entropy graph

The entropy may be plotted in terms of the sampling parameter, so that each time that a new population enters the increasing subsamples $S(P)$ a characteristic behaviour may be recognised in the graph. It can be easily explained with an example for a three-population partition, $\{A_1, A_2, A_3\}$. There exist a couple of values $P_A$ and $P_B$, which corresponds to sampling parameter values for which the second and third populations are respectively recognised. Then, according to the conditions of the preceding section, the following characteristic shapes are produced:

(a) If $P < P_A$ then $H(P) = 0$. The subsample $S(P)$ is only composed of $A_1$ stars.

(b) If $P \in R = (P_A, P_B)$ then $H(P) > 0$ and $\frac{dH}{dP} > 0$. Thus, during the mixture of two populations $A_1$ and $A_2$ there is an interval with continuously increasing entropy.

(c) If $P = P_B$, then $H(P) = 0$ and $n(P) = 1$. An entropy drop takes place when the cumulative population $A_1 \cup A_2$ is formed since the new $A_3$-population, which is enough differentiated from previous populations, is recognised.
Indeed, the cases (a) and (b) are only the boundaries, beginning and end, of an interval $R$ where the non-vanishing entropy continuously develops according to an active mixture. In practice, the case (c) shows only an entropy drop, and the null entropy is usually not reached, since the population $A_3$ is always detected after having been merged a sufficient number of new population stars to the sample $S(P)$.

If the above process occurs smoothly as increasing the size of the sample, within the interval $R$ defined in case (b), we can also find two specific behaviours which are associated with the following two regions:

$(R_1)$ An interval of the sampling parameter for which $H(P)$ is fast increasing, with a relatively high slope.

This occurs for values of the sampling parameter allowing the incorporation of stars up to the core of the $A_2$-population distribution. Therefore, in this region the mixing proportion sensibly varies. Similarly, the kinematic estimates of the $A_2$-population, Pop-II, will likely be unstable, since this population is just being detected, while those of Pop-I will mostly remain stable.

$(R_2)$ An interval of the sampling parameter where $H(P)$ slowly increases, or has a nearly vanishing slope. It may look like a plateau, where, towards the end, $H(P)$ reaches a local maximum. This occurs when the sampling parameter is scanning the farther wing of the $A_2$-population distribution. In this region the mixing proportions may slightly vary depending on the width of the wings. For example, for very flattened distributions we could expect sensible variations of the mixing proportion. In general, after the incorporation of the core distribution of the $A_2$-population, within the plateau region, we get some stable kinematic parameters for Pop-I as well as for Pop-II.

Such a scheme of entropy graph is shown in Fig. 1 by using a numerical simulation from univariate gaussian distributions. The shape of the graph is reproduced three times, according to a four population mixture, $\{A_1, A_2, A_3, A_4\}$, where three consecutive segregations, $\{A_1, A_2\}$, $\{A_1 \cup A_2, A_3\}$ and $\{A_1 \cup A_2 \cup A_3, A_4\}$ are carried out. Several Montecarlo simulations, even with non gaussian distributions, were performed in Paper I. We can also see in Fig. 2 such a behaviour for an actual stellar sample drawn from HIPPARCOS catalogue. A detailed view of the entropy graph and the population kinematic parameters are displayed in terms of the sampling parameter, in this case the absolute heliocentric velocity. The plots are similar to those obtained for the segregation of thin and thick disks in Paper I. For the plateau region of the entropy graph (left plot) the total velocity dispersion (trace of the covariance matrix) for both disk populations is displayed (right plot). The thin disk (dashed line) shows a nearly constant dispersion (in logarithmic scale), while the thick disk (continuous line) shows a slightly increasing rate, much more stable and smooth than before and after the plateau region of the entropy graph. In addition, along the plateau region, the entropy $H(P)$ shows several relative maxima, which are due to the non absolute continuity of the merging process.

The foregoing features can be easily extended to a greater number of populations, so that each entropy drop marks a new entering population. Notice that, for more than two populations, the informative population is the entering one, which produces the entropy drop, while the previous populations become indistinguishable in the cumulative component Pop-I.
Figure 1: (Left) Stellar density $N(P)$ in terms of the sampling parameter for a simulated univariate distribution, with superposition of four gaussian populations. (Right) Entropy graph of the mixture as scanning the whole sample. Both types of regions $R_1$ and $R_2$ are differentiated. The region $R_1$ appears at the beginning of being merged a new population, showing a fast increasing entropy, while the region $R_2$, the so-called plateau region, takes place after having been merged the core distribution of the new population.

Figure 2: (Left) Detailed entropy graph for a HIPPARCOS sample (Paper I). The plateau region allowing an optimal segregation of thin and thick disks takes place for a sampling parameter between $P = 160$ and $270$ km s$^{-1}$ (absolute heliocentric velocity). (Right) In a logarithmic plot, the figure displays the total dispersion $\sigma^2$ in terms of the sampling parameter $P$. For $160 < P < 270$ the plot shows a constant estimation of thin disk total dispersion (dashed line), while for the thick disk (continuous line), the total dispersion increases very slowly.
3 Minimum $\chi^2$ error of the mixture approach

As it has been explained in the introduction a $\chi^2$ test, according to Eq. 3, may be used to estimate the goodness of the approximation of the total moments through the mixture obtained for each set of values of the sampling parameter. The current segregation method works with velocity moments up to fourth order (3 first-order moments, 6 second moments, 10 third moments and 15 fourth moments, which make up to 34 relationships in total). The total moments are represented by a vector $M_i$ of $R = 34$ components, and are fitted from the kinematic parameters of populations ($2 \times 3$ means and $2 \times 6$ covariances) together with the mixture proportions, in order to fulfil the overdetermined set of relationships Eq. 2. Thus an amount of $v = 19$ mixture parameters have to be estimated. If $\varepsilon_i$ denotes the error of moment $M_i$, a calibration of the approximation can be done from the weighted mean of the squared errors, so that, by defining $\Delta_i = M_i - n'M'_i - n''M''_i$, we can write

$$\chi^2 = \sum_{i=1}^{34} \frac{1}{\varepsilon_i^2} |\Delta_i|^2$$  \hspace{1cm} (18)

The quadratic error $\varepsilon_i^2$ associated with a total sample moment $M_i$, which is usually called sampling variance, may be estimated as a sum of terms which are proportional to $\frac{1}{N}$ and $\frac{1}{N-1}$, and depends on higher distribution cumulants (e.g. Stuart & Ord 1987). For an enough large sample we may write $\varepsilon_i^2 \sim \frac{\sigma^2}{N}$, where the quantity $\sigma^2$ represents some intrinsic quantity of the distribution, which can be interpreted as the variance of a distribution mean value $\mu_i$, which sample estimate is $M_i$. Thus, for the given set of computed values $\Delta_i$, and in regard to the quantity $N$, the greater the sample size, the greater the $\chi^2$ fitting error. Similarly, we shall discuss in the following section how the quadratic error $\varepsilon_i^2$ may be estimated in terms of the sampling variances of the partial moments, and how the $\chi^2$ error depends on them.

Notice that in order to apply the $\chi^2$ test we should assume that the total error has a 34-gaussian probability distribution, so that the random variate expressed in Eq. 18 is expected to be a $\chi^2$ distribution with $q = R - v = 15$ degrees of freedom. Also notice that if the errors are given not by a gaussian but by a Poisson distribution, Eq. 18 obeys the $\chi^2$ distribution in the large limit $q$ anyway. It is known that, if $P(\chi^2; q)$ denotes its probability distribution, the relevant quantity to make decisions about goodness of the fit is the $\chi^2$ probability given by the integral

$$\pi(\chi^2; q) = \int_{\chi^2}^{\infty} P(\zeta; q) d\zeta$$  \hspace{1cm} (19)

which gives the probability that a function which describes a set of $q$ data points would give a value of $\chi^2$ as large, or larger, than the one we already have. Nevertheless, while comparing the fitting errors, more than the values $\chi^2$ and $\pi$, what is more significant and easy to visualise is the increase or decrease of the quantity $\chi^2$ itself.

3.1 Fitting error

The current purpose is to establish how the total quadratic error $\varepsilon_i^2$ should depend on the sampling variances of the mixture parameters, that is, on the quadratic errors associated either with the mixture proportions or the partial moment estimates. It will be first studied for two simple cases, depending on whether the parameter $P$ scanning the sample $S(P)$ is providing a sequence of subsamples which (a) maintain a nearly constant mixture proportion, but produces some slightly unstable kinematic estimates of populations, or (b) produces some constant kinematic parameters, although the mixture proportion may slightly vary.
Let us remember that at the beginning of being detected a new population, in the region of the entropy graph, we may expect a significant variation of both mixture proportion and population kinematic parameters. Thus, above cases may be particularly expected in the region $R_2$, the plateau region of the entropy graph.

(a) Let us assume a sample composed of a mixture of $N$ stars, with constant population sizes $N' = Nn'$ and $N'' = Nn''$, being $n' \geq n''$.

The sample partial moments $M_i'$ and $M_i''$ of populations may differ from their real distribution moments $\mu_i'$ and $\mu_i''$ accordingly to the approximate sampling variances $V[M_i'] = \frac{\sigma_i'^2}{N'}$ and $V[M_i''] = \frac{\sigma_i''^2}{N''}$ respectively, for enough large $N'$ and $N''$. Those quantities are the sampling variances of each partial moment, and the values $\sigma_i'$ and $\sigma_i''$ could be computed from higher-order distribution cumulants, if known. Hence, like in the total sample, the quantities $\sigma_i'$ and $\sigma_i''$ may be interpreted as population variances which are associated with the mean values $\mu_i'$ and $\mu_i''$ of the real populations, whose sample estimates are $M_i'$ and $M_i''$.

Total sampling variance $\varepsilon_i^2 = V[M_i]$ may be obtained by writing Eq. 2 in the form

$$NM_k = N'M_k' + N''M_k''$$

and by taking variances,

$$N^2V[M_i] = N'^2V[M_i'] + N''^2V[M_i''] + 2N'N''\text{cov}[M_i', M_i'']$$

Now, by using the proportions $n'$, $n''$, and by writing $\text{cov}[M_i', M_i''] = \frac{\rho_i\sigma_i\sigma_i''}{\sqrt{N\cdot N'}}$, in terms of the corresponding correlation factor $\rho_i$, we have

$$\varepsilon_i^2 = \frac{1}{N}[n'\sigma_i'^2 + n''\sigma_i''^2 + 2\sqrt{n'n''}\rho_i\sigma_i\sigma_i'']$$

Hence, such a calibration of the quadratic error obtained in Eq. 20 distributes the total error between partial’s, in terms of the population variances of the partial moments, by taking into account the likely dependence of them.

We must point out a particular situation which applies to our algorithm procedure. Let us assume that the variates $M_i'$ and $M_i''$ are independent, as it happens in the plateau region $R_2$ of the entropy graph, where $M_i'$ is nearly constant, and $M_i''$ may vary depending on Pop-II stars entering the sample $S(P)$. Then we have $\rho_i = 0$. Thus we can simply write $\varepsilon_i^2 = \frac{1}{N}(n'\sigma_i'^2 + n''\sigma_i''^2)$. It is however useful to obtain an approximate expression of this error in terms of the product $n'n''$, which takes values in the interval $[0, \frac{1}{2}]$ (by assuming $n'' \leq n'$). Obviously we can write $n' = \frac{1}{2} + \frac{1}{2}\sqrt{1 - 4n'n''}$ and $n'' = \frac{1}{2} - \frac{1}{2}\sqrt{1 - 4n'n''}$. The square root can be approximated by a first degree polynomial in terms of $n'n''$, so that $\sqrt{1 - 4n'n''} \sim 1 - \alpha n'n''$, with $2 \leq \alpha \leq 4$, by leading as well to the same qualitative result.

---

\[^1\]If needed, depending on the working interval for $n'$, we can use a more accurate approximation, such as $\sqrt{1 - 4n'n''} \sim 1 - n'n''$, with $2 \leq \alpha \leq 4$, by leading as well to the same qualitative result.
On the other hand, Eq. 2 can be thought in terms of the real distribution moments, instead of sample moments, with constant mixture proportions, where the exact total distribution moments $\mu_i$ are estimated through the partial ones as

$$\mu_i = n'\mu_i' + n''\mu_i''$$  \hfill (24)

Therefore the value $\mu_i$ must be considered constant, although there may be some uncertainty between the partial moments. In addition, the partial moments are assumed to be independent. Hence, the way of sharing the error between both populations may be obtained by differentiating Eq. 23, so that $n'd\mu'_i = -n''d\mu''_i$, and by taking variances,

$$n^2\sigma_i'^2 = n''^2\sigma_i''^2$$  \hfill (25)

Thus, the population variance $\sigma_i'^2$ of the moment $\mu_i'$, corresponding to the major population Pop-I, carries on with less error, while the the population variance $\sigma_i''^2$ of the moment $\mu_i''$, from the minor population Pop-II, is expected to have greater uncertainty $\sigma_i''^2$. Therefore, under above assumptions, we may assume that the following relationship is satisfied

$$\sigma_i''^2 - \sigma_i'^2 \geq 0$$  \hfill (26)

(b) Let us now assume some constant partial moment estimates, but a varying mixture proportion along the plateau region. The calibration error $\varepsilon_i^2$ may be then estimated in terms of the variance of mixture proportion as follows. By increasing the sampling parameter, we pick up a number of stars to build a sample $S(P)$. If the sample is large enough, for the $i$-th entering star, we can define a Bernoulli random variate $I_i$ with two possible outcomes, 1 or 0 depending on whether the star belongs either to Pop-I, the major population, or to Pop-II, the minor population. Obviously, the expected value of $I_i$ is $n'$ and the variance is $n'n''$. Then, the cumulative frequency of the stars belonging to Pop-I, $X_N = \sum_1^N I_i = N'$, follows a binomial distribution with mean $E[X_N] = Nn' = N'$, and variance $V[X_N] = Nn'n''$. Similarly, for the stars belonging to Pop-II, the random variate $Y_N = N - X_N = N''$ verifies $E[Y_N] = Nn'' = N''$, $V[Y_N] = Nn'n''$, and $\text{cov}[X_N, Y_N] = -Nn'n''$.

Notice that this is not exactly as stated, since the sample is finite. The correct distribution of the variate $X_N$ would be of hypergeometric type, but since the size $N$ is large we can assume the binomial hypothesis nearly in all the continuous region of the entropy graph. Nevertheless, when all the population Pop-II is merged to the subsample, also according to a hypergeometric distribution, the variances $V[X_N]$ and $V[Y_N]$ are null.

Thus, if the error associated with Eq. 20 is only due to the uncertainty of the mixture proportion, in a region of the entropy graph where the kinematic population parameters remain approximately constant, by applying the law of combination of errors we get

$$N^2V[M_i] = M_i'^2V[N'] + M_i''^2V[N''] + 2M_i'M_i''\text{cov}[N', N'']$$  \hfill (27)

which may be written as

$$N^2V[M_i] = M_i'^2V[X_N] + M_i''^2V[Y_N] + 2M_i'M_i''\text{cov}[X_N, Y_N]$$  \hfill (28)

Hence

$$N^2V[M_i] = M_i'^2Nn'n'' + M_i''^2Nn'n'' - 2M_i'M_i''Nn'n''$$  \hfill (29)
By dividing by \( N^2 \), and by rearranging terms, we now get the calibration error as

\[
\varepsilon_i^2 = \frac{1}{N} (M_i' - M_i'')^2 n'n'' \tag{30}
\]

In this case we also obtain a sampling variance which is proportional to the inverse of the sample size, but, in addition, such a calibration error depends on two factors, one of them involving the sample estimates of the mixture. It increases as both partial moments become more differentiated. The other factor is, one more time, the product \( n'n'' \). At the end of the plateau region \( \varepsilon_i^2 \) vanishes since \( n'' = 0 \).

A more general situation could be described from a combination of above cases, where mixing proportion and partial moments could vary somewhat.

### 3.2 \( \chi^2 \) graph

In the plateau region of the entropy graph we may assume:

(i) The mixture proportion and the kinematic parameters are independent variables, where both, population parameters and mixture proportion may separately vary depending on the randomly entering stars.

(ii) The kinematic parameters of the major population Pop-I are nearly constant, while the Pop-II kinematic parameters may slightly and smoothly vary, as it was shown in Fig. 2 for the HIPPARCOS sample. Therefore, both \( M_i' \) and \( M_i'' \) are also independent variates, with vanishing covariance.

Then, according to the previous section, the total squared calibration error can be written as the sum of equations Eq. 23 and Eq. 30,

\[
\varepsilon_i^2 = \frac{1}{N} \left( \sigma_i'^2 + \left[ (\sigma_i''^2 - \sigma_i'^2) + (M_i'' - M_i')^2 \right] n'n'' \right) \tag{31}
\]

In order to express the calibration error in terms of the entropy, the factor \( n'n'' \) can be written in the following form. If the entropy \( H(n') \) of Eq. 12 is approximated by a Lagrange polynomial \( L(n') \), with matching values at points \((0, H(0)), (\frac{1}{2}, H(\frac{1}{2})), \) and \((1, H(1)), \) we have \( H(n') \sim L(n') = 4 \ln 2 n'(n' - 1) \). Then

\[
n'n'' \sim \frac{1}{4 \ln 2} H(n') \tag{32}
\]

allowing Eq.31 to be approximated in terms of the entropy \( H(n') \). In this case, the calibration error for Eq. 20 becomes

\[
\varepsilon_i^2 = \frac{1}{N} \left( \sigma_i'^2 + \frac{1}{4 \ln 2} \left[ (\sigma_i''^2 - \sigma_i'^2) + (M_i'' - M_i')^2 \right] H(n') \right) \tag{33}
\]

By taking into account Eq. 26 and the positiveness of the terms involved in Eq. 33, we may write above equation in terms of the entropy associated with the sampling parameter \( P \), in the simpler form,

\[
\varepsilon_i^2 = \frac{1}{N} [A_i^2 + B_i^2 H(P)] \tag{34}
\]
Notice that the quantities $A_i^2 = \sigma_i'^2$ and $B_i^2 = \frac{1}{4\ln 2} \left[ \left( \sigma_i''^2 - \sigma_i'^2 \right) + \left( M_i'' - M_i' \right)^2 \right]$ are expected to be significantly stable within the plateau region of the entropy graph.

Therefore, under those assumptions, in the region $R_2$, by combining Eq. 18 and Eq. 34, the behaviour of $\chi^2$ as a function of either the sampling parameter $P$ or the entropy $H$ could be at last explained.

Thus, by taking derivative in Eq. 18, and bearing in mind Eq. 34, we have

$$\frac{d\chi^2(P)}{dP} = \frac{d\chi^2(P)}{dH} \frac{dH}{dP} = -N \sum_{i=1}^{34} \left( \frac{|\Delta_i| B_i}{A_i^2 + B_i^2 H(P)} \right)^2 \frac{dH}{dP}$$  \hspace{1cm} (35)

Then, since the summation terms are always positive, we are led to the relationship

$$\frac{d\chi^2(P)}{dP} \frac{dH}{dP} \leq 0$$  \hspace{1cm} (36)

Therefore, while varying the sampling parameter $P$ in a plateau region, an increasing entropy is always associated with a decreasing $\chi^2$ error, and conversely. Furthermore, within the region where MEMPHIS algorithm has been designed to obtain the optimal segregation, according to Eq. 35, local maxima of ones take place simultaneously with local minima of others.

Outside the plateau region, things can behave very different. At the beginning of the region $R_1$, with a suddenly income of a new population, above assumptions may be non valid.

Therefore, the optimal sampling parameter should provide: (1) one or several local maxima of entropy within the plateau region of the entropy graph, which indicate the less informative subsamples providing the more representative mixture parameters, and (2) one or several simultaneous minimums of the $\chi^2$ fitting error, which provide the optimal segregation. Thus, under the preceding circumstances, the “more representative” subsample is also the one providing the least error of the mixture approach.

The combination of both criteria will be used in Cubarsi & Alcobé (2009) to analyse two actual local samples, where the kinematic estimates of thin disk, thick disk, and halo will be obtained. Several sampling parameters will be compared, and a precise estimation of the mixture proportions will be carried out.
References

ESA, 1992, The Hipparcos Input Catalogue, ESA SP-1136
ESA, 1997, The HIPPARCOS Catalogue, ESA SP-1200