Comparing Parametric and Nonparametric Statistical Methods for Studying the Velocity Distributions of Population I Stars

F. ARENOU
URA D0335 CNRS, Observatoire de Paris-Meudon
F-92195 Meudon, France

M. L. BOUGEARD
URA 1125 CNRS: I.E.R.S., Observatoire de Paris
61 Av. de l'Observatoire, F-75014 Paris, France

Abstract: *This Paper investigates population I star samples from the viewpoint of the velocity distributions, each being viewed as a mixture. To obtain information on the class centers, clustering techniques are first applied. Secondly, we use a parametric maximum likelihood formulation that we solve by Redner-Walker's E.M. algorithm. The obtained results are compared and discussed from an astrophysical viewpoint.*

Key words: - velocity distribution - mixtures - clusters - E.M. algorithm.

1. INTRODUCTION

The stellar distribution in Our Galaxy is a function of several variables and it is quite known that Stellar Kinematics play a major role in studies of the galactic structure. Adopting the usual description of the observed space velocity of a star with respect to the Sun in terms of a three dimensional vector (U, V, W), the observed distribution in each component has been long assumed as nearly gaussian. Gómez *et al.* (1990) defined particular homogenous Population I samples that are nearly complete up to a limiting magnitude of stars located in the solar neighborhood at a distance smaller than 200 parsecs from the Sun, covering ages up to about 6 10^8 years. They proved that the multivariate velocity densities can be viewed as a *mixture* of populations related to different star formation bursts. Here, our aim is to compare different statistical methods of separation of the components by focusing our attention on their A2 V-sample (size $n = 97$).

In Section 2 nonparametric clustering techniques are applied. Then, assuming a mixture of multivariate Gaussians in Section 3, we use a maximum likelihood

I.I. Mueller and B. Kołaczek (eds.),
Developments in Astrometry and Their Impact on Astrophysics and Geodynamics, 265–269.

approach that we solve numerically by the recent and powerful E.M. algorithm. The obtained results are compared and discussed.

2. SEPARATING COMPONENTS: NONPARAMETRIC METHODS

2.1 From an univariate approach to a multivariate one

Although U,V are highly correlated, it becomes apparent by using Kolmogorov-Smirnov univariate tests of normality, that, among U,V,W, only U seems nearly gaussian. Moreover the U distribution is assymetric and on a (U,V) graph, one can guess a mixture of at least two groups. So, an analysis will be performed on the global (U,V,W) distribution.

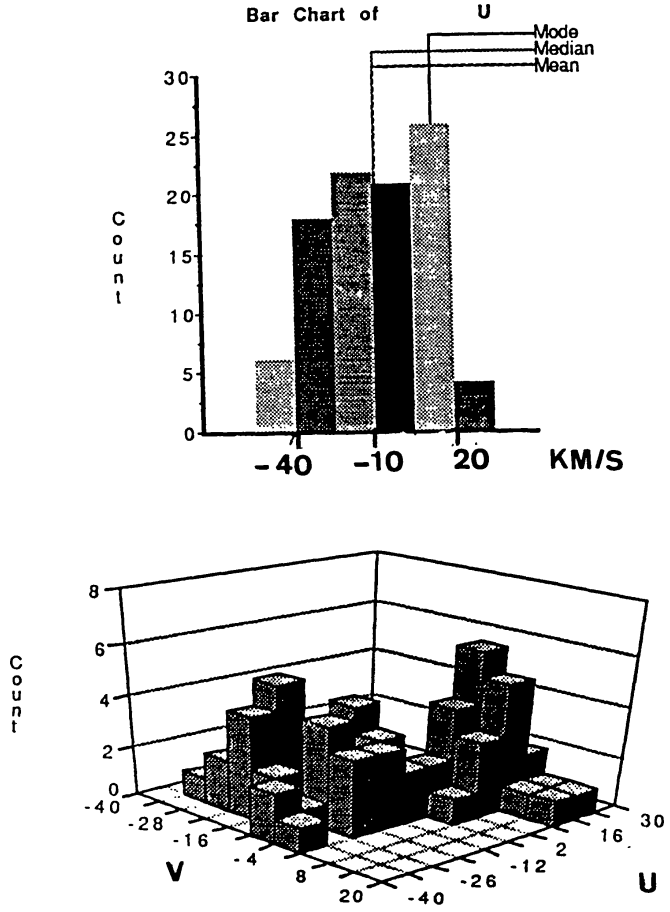

Fig.1. On the top, the U-distribution, on the bottom the bidimensional (U,V) one

2.2 Method

Several hierarchical clustering approaches are available. Here, we apply the *"Nearest Neighbour chain"* NN-algorithm that classifies observations into groups using a nonparametric method (Hand, 1981).

2.3 Results

We obtain a natural partition into two components (A_c, B_c) that is shown on Fig.2 drawn in the U-V galactic plane. The centroid of each class is given by:

A_c: $\bar{U} = -22.5$ $\bar{V} = -14.1$ $\bar{W} = -6.6$ ($n = 57$ stars among 97)
B_c: $\bar{U} = +10.3$ $\bar{V} = -\ 0.4$ $\bar{W} = -7.6$ ($n = 40$ stars among 97)

We have noticed that, in the present case, a non hierarchical classification based on *aggregation around moving centers* leads to similar results.

3. PARAMETRIC ESTIMATION

3.1 Statistical formulation

In order to apply a maximum likelihood method, we assume in the following that the velocity distribution $f(x)$, $x = (U, V, W)$ of the A2 V-sample can be realistically represented as a *convex combination of Gaussians* Φ_i, $i = 1 - k$ with mean m_i, variance Σ_i, respective weights p_i:

$f(x, \theta) = p_1 \Phi_1(x \mid m_1, \Sigma_1) + \ldots + p_k \Phi_k(x \mid m_k, \Sigma_k)$

Methods for estimating the unknown parameters $\theta = (p_i, m_i, \Sigma_i, i = 1 - k)$ have been and are still widely discussed in the statistical litterature. As it is well known, some difficulties arise in the case of overlapping in the components and when the sample size is "statistically" small (the present case).

3.2 The Redner-Walker (1984) E.M. algorithm

In order to maximize the log-likelihood function $L(\theta)$ of the sample, we use the Redner-Walker E.M. algorithm that is known to converge and to preserve the parameter constraints compared to the Newton method. Recall that the updating rule is based on the evaluation at step *E (expectation step)* of an ad-hoc estimate p_{ij} of the posterior probability that an observation x_j belongs to the ith group, given the current estimate of θ. Then, step *M (maximization)* updates θ with reference to this p_{ij} (see formulation in: Bougeard and Arenou, 1990).

3.3 Results and interpretation

Several independent runs have to be performed since the resulting likelihood surface is littered with singularities that potentially occur on the boundary of the parameter

space. For the A2 V sample under consideration here, a 2-mixture solution (A, B) with respective weights 0.64 and 0.36 has been found as significant. We obtain as estimates of the mean values and of the diagonal of the variance matrix:

A: $(m_U, m_V, m_W) = (-20.8, -14.3, -6.8)$ $\quad (\sigma_U, \sigma_V, \sigma_W)^2 = (13.3, 10.2, 8.7)^2$

B: $(m_U, m_V, m_W) = (+11.4, +1.7, -7.4)$ $\quad (\sigma_U, \sigma_V, \sigma_W)^2 = (7.1, 5.8, 7.1)^2$

Moreover the p_{ij} permit a classification of the stars in each group that is shown on Fig.2.

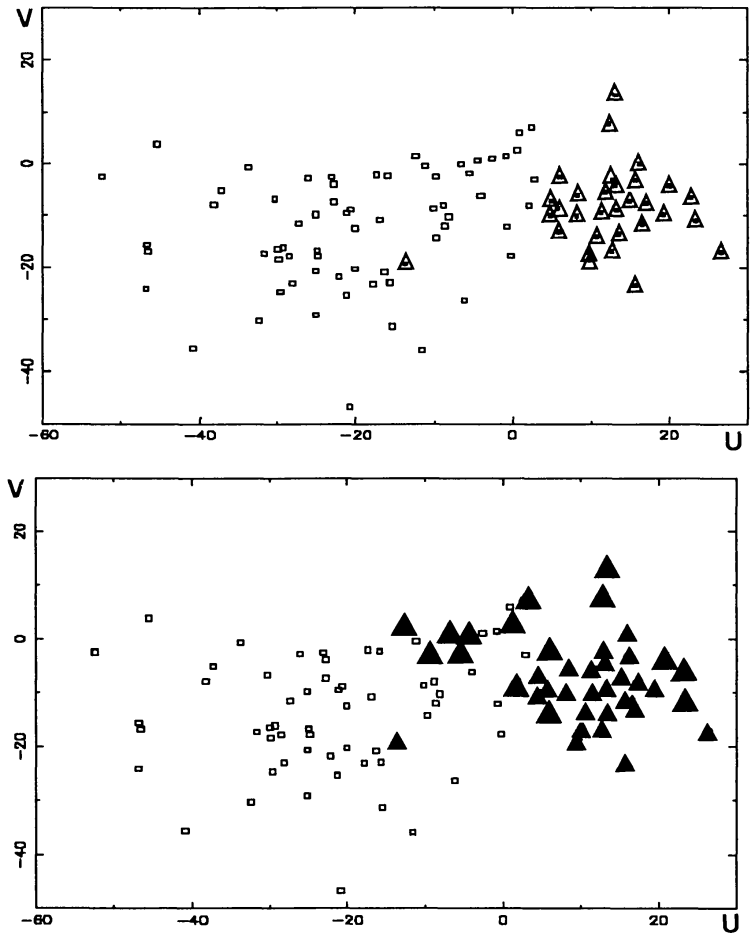

Fig.2. Visualisation of the 2-mixture in the galactic plane: on the top, the nonparametric clusters, on the bottom the E.M. groups

It can be noticed that the above parametric results are in agreement with those of the clustering method. This supports the hypothesis that this sample is in fact a mixture of two populations.

Gómez *et al.* (1990) interpret each group as formed in a particular star burst. Moreover, they assume that each burst also produced open clusters. At this time, the individual ages of the stars in the sample under study are not available; the age is clearly the discriminant variable which could support this interpretation in terms of star formation bursts. Nevertheless, we can notice that the above statistical estimations of the A group mean values agree with the velocity components of the IC 2391 young cluster and the B group with U Ma cluster that is older.

4. CONCLUSION

In this Paper we have been concerned with the problem of separating components of a mixture that is a problem of importance in Astronomy (study of velocity distributions in terms of star formation bursts,...). Different methods were examplified in application to an A2 V-sample. Further statistical investigations are to be found in a forthcoming paper. Nevertheless, larger difficulties were encountered for the analysis of samples with smaller size. Recall that for a 2-component mixture in three dimensions, the estimation of 1 (proportion) + 3 (mean-values) + 6 (covariance-terms) = 10 parameters are needed.

The results will be improved as soon as a better knowledge of the star velocities will be obtained, allowing the use of larger samples.

5. REFERENCES

Bougeard M.L., Arenou F.,1990: in *Errors, Biais, Uncertainties in Astronomy*, Cambridge Press, 277-280

Bougeard M.L., Arenou F., Gómez A.E., 1989, *I.S.I. 49th Session*

Gómez A.E., Delhaye J., Grenier S., Jaschek C., Arenou F., Jaschek M., 1990, *Astron. Astrophys.* **236**, 95

Hand D.J., 1981, *Discrimination and classification*, N. York, J. Wiley

Redner, Walker, 1984, *S.I.A.M.*, 26 n°2, 195-239