Serb. Astron. J. № 173 (2006), 57 - 64 **UDC 524.45M67–16–17**
DOI:10.2298/SAJ0673057U Original scientific paper

THE EXPECTATION MAXIMIZATION ALGORITHM AS A POWERFUL TOOL TO SOLVE THE STELLAR MEMBERSHIP IN OPEN CLUSTERS. AN APPLICATION TO M67

A. Uribe, R. Barrera and E. Brieva

Observatorio Astronómico, Facultad de Ciencias, Universidad Nacional de Colombia, Apartado Aéreo 2584, Bogotá, Colombia

(Received: August 29, 2006; Accepted: September 21, 2006)

SUMMARY: The EM algorithm is a powerful tool to solve the membership problem in open clusters when a mixture density model overlaping two heteroscedastic bivariate normal components is built to fit the cloud of relative proper motions of the stars in a region of the sky where a cluster is supposed to be. A membership study of 1866 stars located in the region of the very old open cluster M67 is carried out via the Expectation Maximization algorithm using the McLachlan, Peel, Basford and Adams EMMIX software.

Key words. Methods: analytical – Methods: data analysis – Methods: statistical – open clusters and associations: individual: M67

1. INTRODUCTION

M67 is a very old open cluster located at $\alpha_{2000} = 8^h50^m26,1^s$ and $\delta_{2000} = +11^o48^{'}46^{''}$. This famous galactic object has been extensively studied through theory and observation. Regarding the proper motions and membership determination, one can quote the papers by Sanders (1977), Girard (1989) and Zhao (1993). M67 is located at an estimated distance of 870 pc (Girard et al. 1989). Most studies agree to an age between 4 and 5 Gyr, making it a representative object for population of solar age stars (Demarque et al. 1992, Van den Berg et al. 2004). M67 is an appropriate target of observation for the study of solar type stars since it has the same chemical composition of the Sun (Barry and Cromwell 1974, Giampapa 2000). A mixture density model that overlaps two heteroscedastic bivariate normal components has been built to fit the relative proper motions measured in mas per year of 1866 stars located in the region of M67, as a model for the cluster and field stars. Proper motions greater in absolute value than 1.5 were pruned (Fig. 1). The data were obtained by Sanders (Sanders 1971, 1977). The parameters of the mixture have been estimated via the expectation maximization algorithm following Dempster (1977). Membership probabilities are obtained applying a Bayesian rule and using the McLachlan, Peel, Basford and Adams EMMIX software (McLachlan et al. 1999).

2. THE MODEL AND THE LIKELIHOOD FUNCTION

A Gaussian Mixture Model has been often used to classify into two homogeneous subpopulations π_i, a random sample of n observations $x_1, \ldots, x_n$, coming from a heterogeneous population (MacLachlan et al. 2000). Let $x_j^T = (x_{j1}, x_{j2})$ be the transposed vector formed by measuring two continuous random variables X_{j1}, X_{j2} in the individual

j, $j = 1, \ldots, n$. If the probability density function (pdf) of the ith normal subpopulation, $i = 1, 2$, is $f_i(x_j, \mu_i, \Sigma_i)$, with vector of means μ_i and variance-covariance matrix Σ_i, then the multivariate variable $X_j^T = (X_{j1}, X_{j2})$, $j = 1 \ldots, n$ has a pdf given by the mixture $f(x_j; \Theta)$ of two normal components:

$$f(x_j; \Theta) = \sum_{i=1}^{2} \alpha_i f_i(x_j; \mu_i, \Sigma_i), \tag{1}$$

where the components of the vector Θ are the unknown proportions $\alpha = (\alpha_1, \alpha_2)$, constrained by $0 < \alpha_i < 1$ and $\sum_{i=1}^{2} \alpha_i = 1$, the unknown components of the vector of means μ_1, μ_2 and the unknown parameters that come from the two heteroscedastic variances Σ_i. In a first step we consider the n $\times$ 2 matrix X with its j line, $j = 1, \ldots, n$, given by X_{j1}, X_{j2}. This is an incomplete matrix in the sense that it is clearly unkown to which subpopulation i, $i = 1, 2$, should be assigned the jth individual with a vector of observations x_j^T, realizations of the random vector X_j^T. In order to solve this incomplete data problem, a complete n x (2+2) matrix, let us say $[X, Z]$, is formed, where Z is defined by latent Bernoulli variables Z_{ji}, with Z_{ji} being one, if the jth individual belongs to the ith subpopulation π_i, and zero in other case. To each individual j is then associated a vector Z_j, $Z_j^T = (Z_{j1}, Z_{j2})$, with one component equals to unity an the other being zero. It follows that Z_j has a multinomial distribution: $Z_j \sim$ Multinomial$(1; \alpha_1, \alpha_2)$: $f(Z_j; \alpha_1, \alpha_2) = \prod_{i=1}^{2} \alpha_i^{Z_{ji}}, j = 1, \ldots, n$, and the conditional density function $f(X_j \mid Z_j)$ is given by: $f(X_j \mid Z_j) = \prod_{i=1}^{2} \{f_i(X_j; \mu_i, \Sigma_i)\}^{Z_{ji}}$. Then, the joint density function $f(X_j, Z_j; \Theta)$ can be written:

$$f(X_j; Z_j; \Theta) = \prod_{i=1}^{2} \left\{ \alpha_i (2\pi)^{-1} \mid \Sigma_i \mid^{-1/2} \right.$$
$$\left. \exp\left[-\frac{1}{2} (X_j - \mu_i)^T \Sigma^{-1} (X_j - \mu_i) \right] \right\}^{Z_{ji}}. \tag{2}$$

Denoting by $L_c(\Theta; X, Z)$ the likelihood function for the complete variable (X, Z), it clearly follows that, $L_c(\Theta; X, Z) = \prod_{j=1}^{n} f(X_j, Z_j; \Theta)$, and the loglikelihood function l will be $l(\Theta; X, Z) = \sum_{j=1}^{n} \log f(X_j, Z_j; \Theta)$. In an equivalent way, $l(\Theta; X, Z) = \sum_{j=1}^{n} \sum_{i=1}^{2} Z_{ji} \{\log \alpha_i + \log f_i(X_j; \mu_i, \Sigma_i)\}$. Estimates $\hat{\mu}_i$, $\hat{\Sigma}_i$ and $\hat{\alpha}_i$ can be found following a maximum likelihood approach for incomplete data problems using the Expectation and Maximization steps of the EM Algorithm proposed by Dempster, Laird and Rubin (Dempster et al. 1977, Mclachlan et al. 2000).

The Expectation Maximization Algorithm

The EM algorithm can be applied to find estimates $\hat{\alpha}_i$, $\hat{\mu}_i$, $\hat{\Sigma}_i$ for the parameters of the mixture density function(4). The algorithm approaches to the problem by solving in an indirect way the incomplete data loglikelihood equations:

$$\frac{\partial \log f(X_j; \Theta)}{\partial \Theta} = \vec{0}. \tag{3}$$

This is done proceeding iteratively in terms of the complete data loglikelihood function $\log L_c(\alpha_i, \mu_i, \Sigma_i; X, Z)$; as it is unobservable, it is replaced by the conditional expectation:

$$E_{Z|X} \left\{ \log L_c \left(\alpha_i, \mu_i, \Sigma_i; X, Z \right) \right\}. \tag{4}$$

The algorithm is iterative and at each iteration it alternates the two operations of Expectation E and Maximization M. More specifically, let $\Theta^{(0)}$ be some initial value for Θ. Then on the first iteration the E-step requires the calculation of the Θ function $Q(\Theta, \Theta^0)$:

$$Q(\Theta; \Theta^0) = E_{Z|X} \left\{ \log L_c \left(\Theta; X, Z \right) \right\}. \tag{5}$$

The M-step requires the maximization of $Q(\Theta, \Theta^0)$ with respect to Θ over the parameter space Ω; that is, we choose $\Theta^{(1)}$ so that

$$Q(\Theta^{(1)}; \Theta^0) \geq Q(\Theta; \Theta^0), \tag{6}$$

for all $\Theta \in \Omega$.

The E and M steps are then carried out again, but this time with $\Theta^{(0)}$ replaced by the current fit $\Theta^{(1)}$. In the $(k+1)$th iteration, the E and the M steps are defined as follows:

(i) E-step. Calculate $Q(\Theta; \Theta^{(k)})$ where

$$Q(\Theta; \Theta^k) = E_{Z|X} \left\{ \log L_c \left(\Theta; X, Z \right) \right\}. \tag{7}$$

(ii) M-step. Choose $\Theta^{(k+1)}$ to be any value of Θ that maximizes $Q(\Theta, \Theta^{(k)})$, that is,

$$Q(\Theta^{(k+1)}; \Theta^k) \geq Q(\Theta; \Theta^k) \tag{8}$$

for all Θ in the parameter space Ω.

The E and M steps are iterated repeatedly until the difference $L(\Theta^{(k+1)}) - L(\Theta^{(k)})$ becomes arbitrarilly small. Dempster, Laird and Rubin (Dempster et al. 1977) show that

$$L(\Theta^{(k+1)}) \geq L(\Theta^{(k)}), \tag{9}$$

for $k = 1, 2, 3, \ldots$. Hence, convergence must be obtained for a sequence of likelihood values that are to be bounded from above. In this way the sequence $\Theta^{(0)}, \ldots, \Theta^{(n)} \ldots$ leads to a likelihood estimate $\hat{\Theta}$.

Uniqueness and some other discussions about this $\hat{\Theta}$ can be found in McLachlan et al. (1997). Applying this two steps algorithm to the mixture of two normal components requires then:
E-**Step:** Define $Q(\Theta, \Theta^{(0)})$ by:

$$Q = \left\{ \sum_{j=1}^{n} \sum_{i=1}^{2} Z_{ji} \left[\log \alpha_i + \log f_i(X_j; \mu_i, \Sigma_i) \right] \right\}. \tag{10}$$

Then:

$$Q(\Theta, \Theta^{(0)}) = E_{Z|X,\Theta^{(0)}} \tag{11}$$

The expectation operator E has the subscript $Z \mid X, \Theta^{(0)}$ to convey explicitly that this expectation is being made using the density function for Z with given X, and $\Theta^{(0)}$ as an initial value for Θ. It follows that, in the $(k+1)$th iteration, the E-step requires the calculation of $Q(\Theta, \Theta^{(k)})$, where $\Theta^{(k)}$ is the value of Θ after the kth EM iteration. As the looklikelihood L_c is linear with respect by the unobserved data Z_{ji}, the E-step in the $(k+1)$th iteration simply requires the calculation of the current conditional expectation of Z_{ji} given the observed data X. Now,

$E_{Z|X,\Theta^{(k)}}$ = Probability$\{Z_{ji} = 1 \mid X\}$, and then it clearly follows:

$$E_{Z|X,\Theta^{(k)}} \left[Z_{ji} = 1 \mid X \right] = \frac{\hat{\alpha}_i^{(k)} f_i(X_j, \hat{\Theta}^{(k)})}{\sum_{i=1}^{2} \alpha_i^{(k)} f_i(X_j \hat{\mu}_i^{(k)}, \hat{\Sigma}_i^{(k)})}, \tag{12}$$

or:

$$E_{Z|X,\Theta^{(k)}} \left[Z_{ji} = 1 \mid X \right] = \hat{\tau}_{ji}(X_j, \Theta^{(k)}). \tag{13}$$

In this way, the function $Q(\Theta, \hat{\Theta}^{(k)})$ built in the kth iteration of the M-step is given by:

$$\begin{aligned} Q(\Theta, \hat{\Theta}^{(k)}) &= \sum_{j=1}^{n} \sum_{i=1}^{2} \hat{\tau}_{ji}(X_j; \hat{\Theta}^{(k)}) \left\{ \log \alpha_i + \right. \\ &+ \left. \log f_i(X_j; \mu_i, \Sigma_i) \right\}. \end{aligned} \tag{14}$$

M-**Step:** Obtain $\hat{\Theta}^{(k+1)}$ as the value of Θ that maximizes the Θ function $Q(\Theta, \hat{\Theta}^{(k)})$, $k = 1, \ldots,$

$$\hat{\Theta}^{(k+1)} = \max_{\Theta \in \Omega} Q(\Theta, \hat{\Theta}^{(k)}), \tag{15}$$

or:

$$\begin{aligned} \hat{\Theta}^{(k+1)} &= \max_{\Theta \in \Omega} \left\{ \sum_{j=1}^{n} \sum_{i=1}^{2} \hat{\tau}_{ji}(X_j, \hat{\Theta}^{(k)}) \left[\log \alpha_i + \right. \right. \\ &+ \left. \left. \log f_i(X_j; \mu_i, \Sigma_i) \right] \right\} \end{aligned} \tag{16}$$

Clearly, to obtain $\hat{\alpha}_i^{(k+1)}$ it suffices to maximize $Q_1(\alpha_i; \hat{\Theta}^{(k)})$ given by:

$$Q_1(\alpha_i \hat{\Theta}^{(k)}) = \sum_{j=1}^{n} \hat{\tau}_{ji}(X_j, \hat{\Theta}^{(k)}) \log \alpha_i - \lambda \left(\sum_{i=1}^{2} \alpha_i - 1 \right), \tag{17}$$

$i = 1, \ldots, g$. The λ term in the above equation takes into account the constraint $\sum_{i=1}^{2} \alpha_i = 1$, using a Lagrange multiplier λ. In order to obtain $\hat{\mu}_i^{(k+1)}$, it suffices to maximize the function $Q_2(\mu_i, \hat{\Theta}^{(k)})$,

$$\begin{aligned} Q_2(\mu_i, \Theta^{(k)}) &= \sum_{j=1}^{n} \hat{\tau}_{ji}(X_j, \hat{\Theta}^{(k)}) \left[X_j^T \Sigma_i^{-1} \mu_i - \right. \\ &- \left. \frac{1}{2} \mu_i^T \Sigma_i^{-1} \mu_i \right] \end{aligned} \tag{18}$$

In a similar way, $\hat{\Sigma}_i^{(k+1)}$ is obtained maximizing the function $Q_3(\Sigma_i; \hat{\Theta}^{(k)})$ given by:

$$Q_3 = \sum_{j=1}^{n} \hat{\tau}_{ji} \left\{ -\frac{1}{2} \log \mid \Sigma_i \mid - \frac{1}{2} \left(X_j - \mu_i \right)^T \Sigma_i^{-1} \left(X_j - \mu_i \right) \right\} \tag{19}$$

with $i = 1, 2$. Thus, $\hat{\alpha}_i^{(k+1)}$, $\hat{\mu}_i^{(k+1)}$ and $\hat{\Sigma}^{(k+1)}$ are eventually found by the following equations:

$$\hat{\alpha}_i^{(k+1)} = \frac{1}{n} \sum_{j=1}^{n} \hat{\tau}_{ji}(X_j, \hat{\Theta}^{(k)}), \tag{20}$$

$$\hat{\mu}_i^{(k+1)} = \frac{1}{n \hat{\alpha}_i^{(k)}} \sum_{j=1}^{n} \hat{\tau}_{ji}(X_j, \hat{\Theta}^{(k)}) X_j, \tag{21}$$

$$\hat{\Sigma}_i^{(k+1)} = \frac{1}{n \hat{\alpha}_i^{(k)}} \sum_{j=1}^{n} \hat{\tau}_{ji} \left(X_j - \hat{\mu}_i^{(k+1)} \right) \left(X_j - \hat{\mu}_i^{(k+1)} \right)^T. \tag{22}$$

The EM estimates $\hat{\Theta}^{(k+1)} = (\hat{\alpha}_i^{(k+1)}, \hat{\mu}_i^{(k+1)}, \hat{\Sigma}_i^{(k+1)})$ are numerically found using the EMMIX software. It is somehow amazing that the *EM* analytic procedure leads finally to the so called Wolfe equations obtained by Wolfe, but following a different approach (Wolfe 1970, Hand 1981). Those equations have also been used to solve the membership problem in open clusters (Cabrera et al. 1985). Both Wolfe and the *EM* procedures find solutions to the same system of nonlinear likelihood equations.

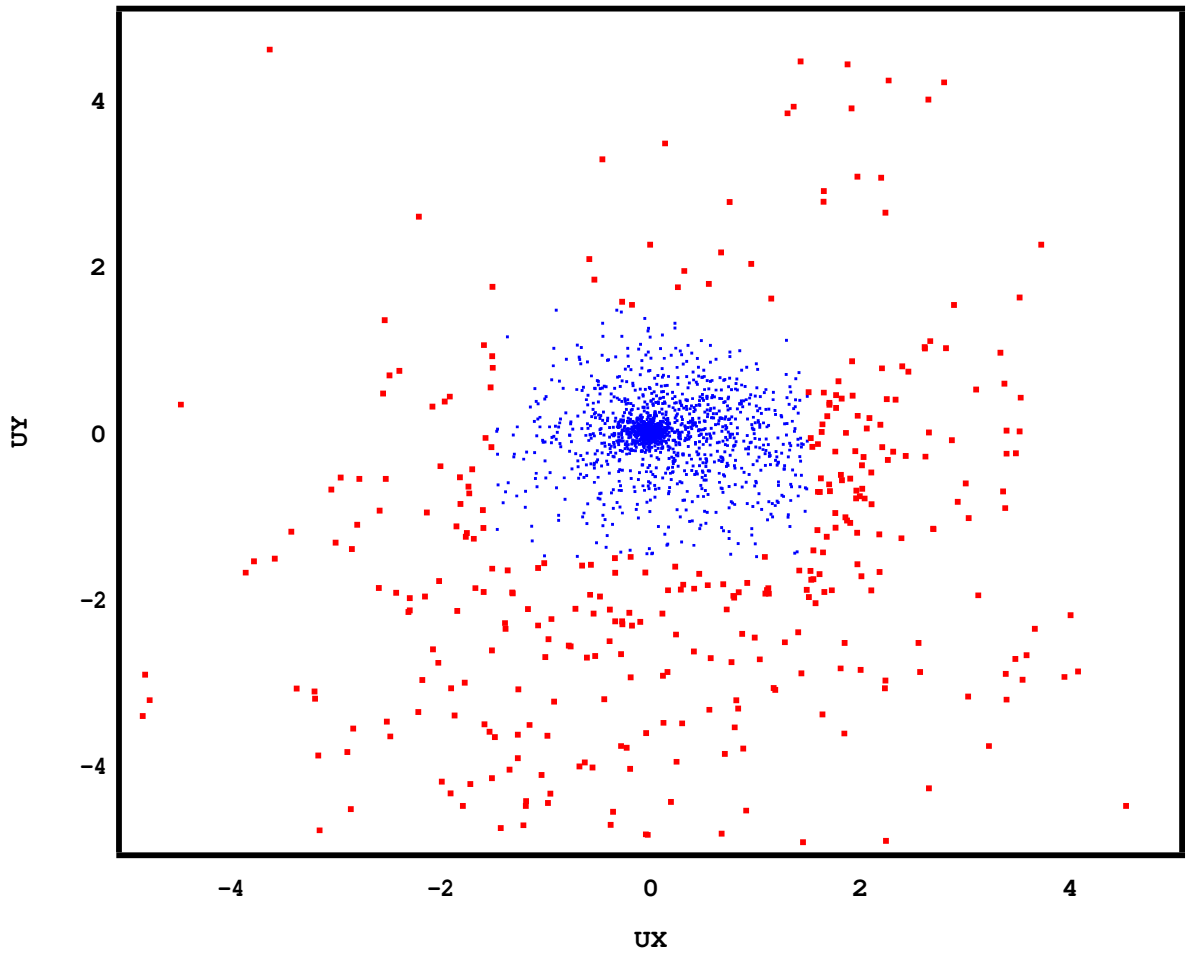

Fig. 1. *The 356 large squares are M67 Sanders stars with absolute proper motions UX or UYgrater than 1.5 arcseconds per century and have been pruned from our membership study following Zhao et al. astrometric criterion. The 1510 small squares are stars that are the subject of this membership study working from proper motions via the EM algorithm.*

3. MEMBERSHIP RESULTS IN M67

The above outlined theory was applied to solve the stellar membership problem in the region of the open cluster M67 using the McLachlan EMMIX software (McLachlan et al. 1997, 2000). The program EMMIX built for a mixture of normal components, uses an appropiate starting point obtained by different multivariate cluster techniques, mainly K-means. The vector with estimates of the parameters of the model is found following the EM algorithm and is given by (0.1959, -0.0441, 0.6346, 0.5320, -0.0427, -0.0132, -0.0013, 0.0995, 0.0829, 0.0300, 0.327), where the centroids, standard deviations and correlation coefficient estimates for the field and for the cluster are respectively given by the first five and by the sixth to the tenth components; the last value is the proportion of cluster stars. From the total number of 1510 stars considered in the M67 region only 534 stars have membership probabilities grater than 0.50; they are divided into probable and most probable members; this last group is formed with 322 stars with membership probabilities grater than 0.90; 310 of them with V and B-V complete published photometric data were plotted in the Hertzsprung-Russell diagram given in Fig. 4. The photometry was taken from Montgomery (1989) and Sanders (1989); for a few stars Girard et al. (1989) data were used. Taking into account spectral types and metallicity, the following twelve G2V solar type stars were found: 724, 777, 779, 945, 991, 1012, 1218, 1452, 1462, 1477, 1484, 1616, where the Sanders identification number was used. EMMIX uses the BIC and the AIC criteria of clustering and gives 0.933 and 0.944 as estimates of correct allocation rates for each component, the field and the cluster, and the estimate of total correct allocation rate equals 0.937. The Member-

ship probabilities are listed in Table 1;[1] ID1 and ID2 are respectively a sequential number and the Sanders identification number; μ_x and μ_y are the proper motions data; P1 and P2 are membership probabilities given respectively by Sanders and in our study. The membership results are in good agreement with previously published papers by Sanders (1977), Girard (1989) and Zhao (1993), that are the main M67 membership studies (Fig. 2 and Fig. 3).

4. CONCLUSIONS

The EM algorithm is a powerful tool to solve the membership problem in open clusters when a mixture of two bivariate normal components models the cloud of proper motions of a region of the sky where a cluster is supposed to be. It leads directly to a point of the parameter space that is a local maximum of the likelihood function. Then, membership probabilities are found using the Bayes theorem. The program EMMIX built for a mixture of normal components, uses an appropiate starting point obtained by different multivariate cluster techniques, mainly K-means, and it uses the BIC and AIC criteria to estimate allocation rates for each component, the field and the cluster.

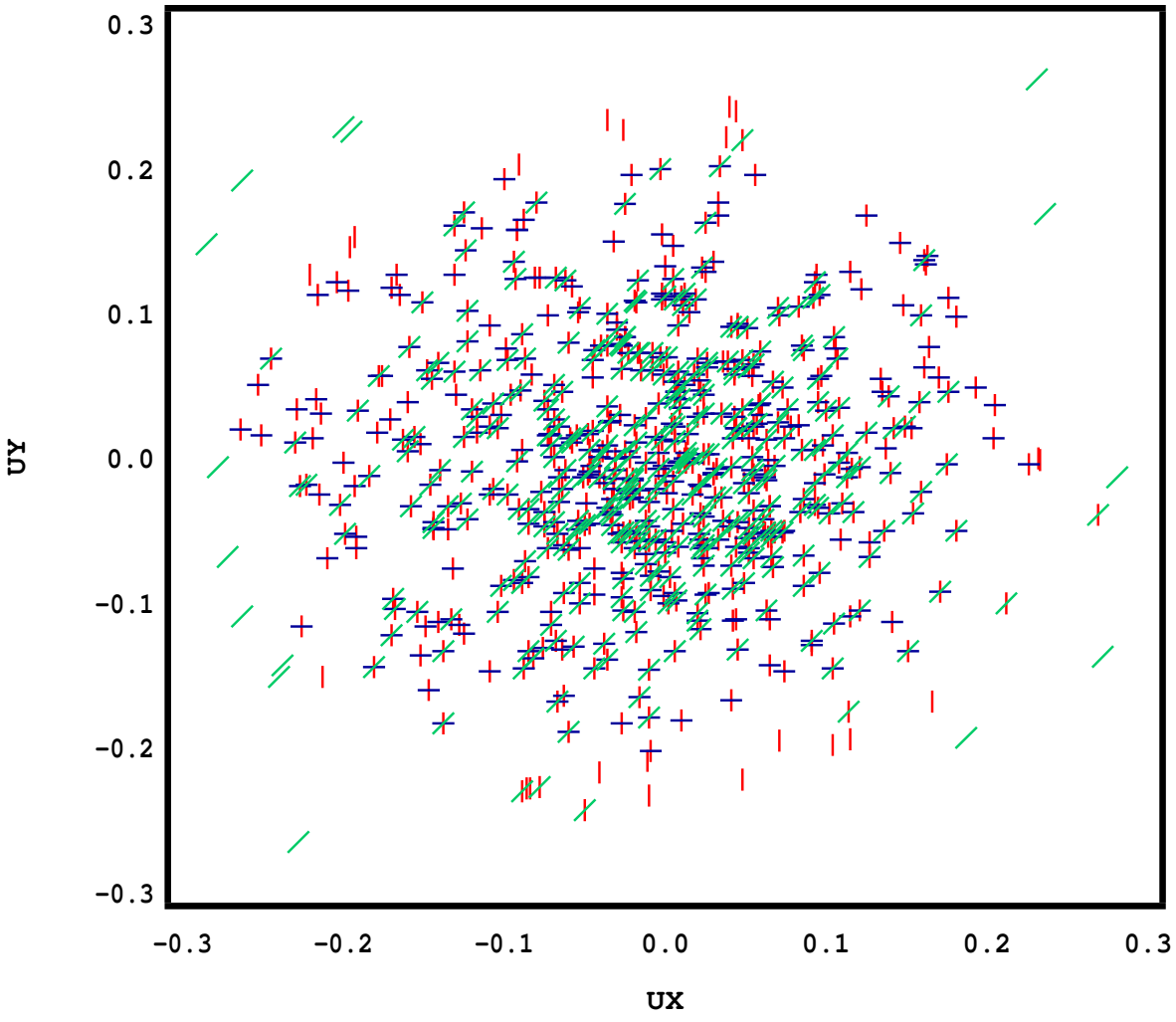

Fig. 2. *M67 564 Sanders probable members with membership probabilities grater than 0.50. (vertical lines) and 534 M67 probable members according to our study (horizontal lines); the inclined lines are 430 Girard members at the same probability level.*

[1] http://saj.matf.bg.ac.yu/173/pdf/Table1.pdf

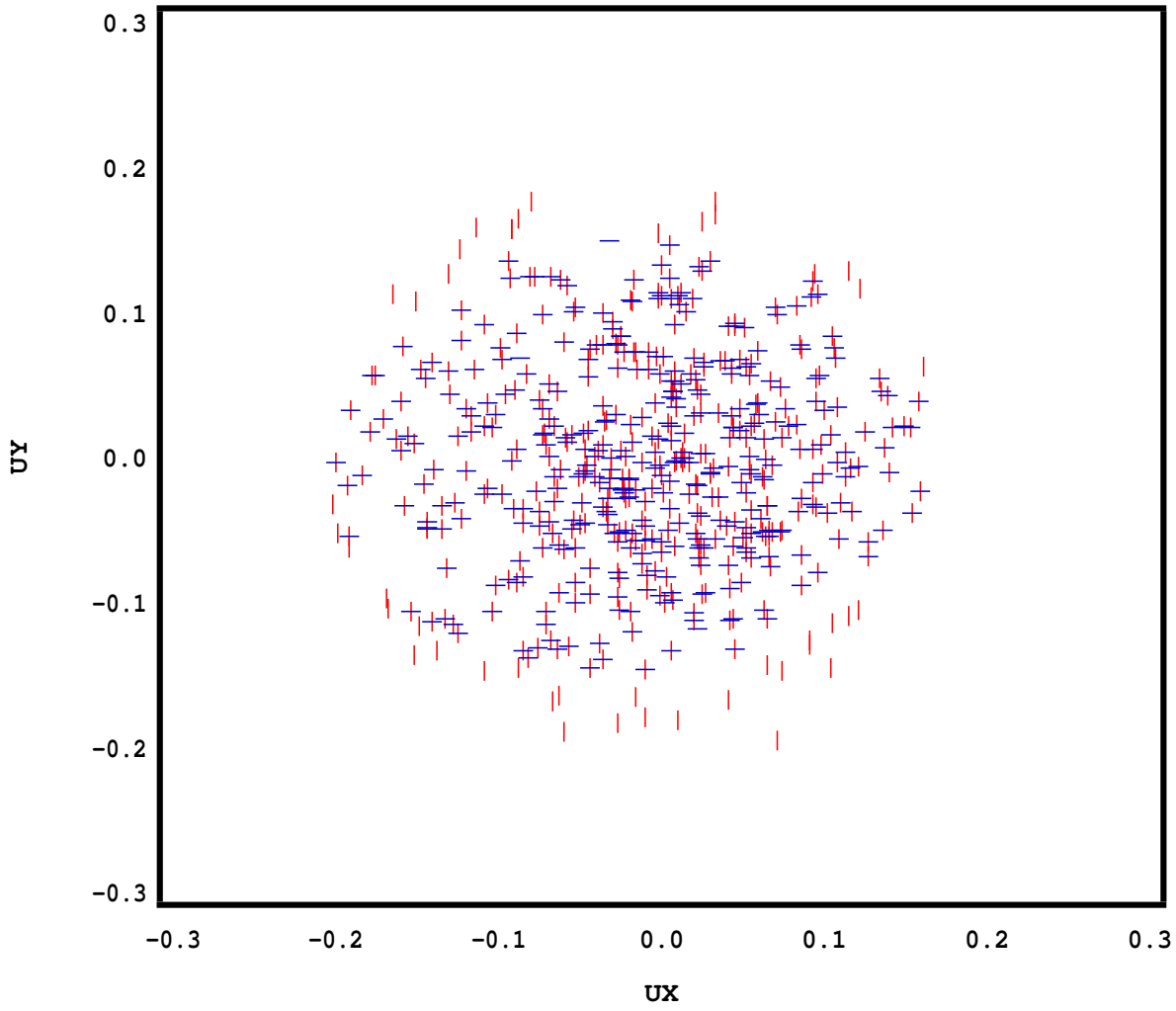

Fig. 3. *M67 480 Sanders probable members with membership probabilities grater than 0.80. (vertical lines) and 438 M67 probable members at the same probability level according to our study (horizontal lines).*

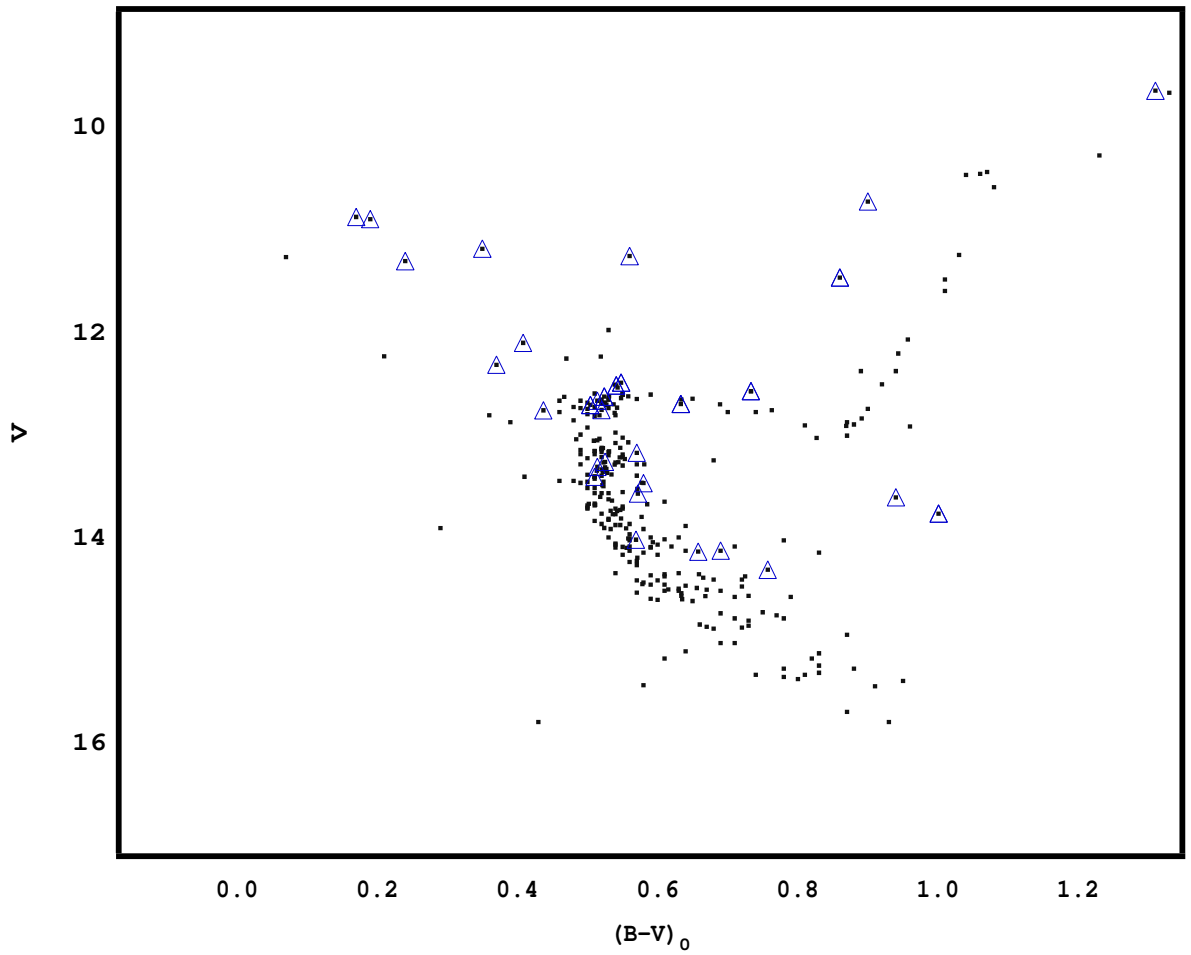

Fig. 4. *M67 Color Magnitude Diagram built with the 310 most probable members. The triangles are 39 M67 binary stars.*

Acknowledgements – The authors thank an anonymous refereé for comments and suggestions. We want to acknowledge the Division of Reasearch of Universidad Nacional de Colombia (DIB), for continuous support given to the Research Project Membership in the region of M67, and to COLCIENCIAS, for giving recognition to the Galactic Astronomy Group of the Observatorio Astronómico Nacional as a research group.

REFERENCES

Barry, D.C., Cromwell, R.H.: 1974, *Astrophys. J.*, **187**, 107.
Cabrera, C., Alfaro, E.: 1985, *Astron. Astrophys.*, **150**, 298.
Demarque, P., Guenther, D.B., Green, E.M.: 1992, *Astrophys. J.*, **103**, 151.
Dempster, A., Laird, N. and Rubin, D.J.: 1977. *Journal of the Royal Statistical Society, Series B*, **39**, 1.
Girard, T.M, Grundy, W.M., Lopez, C.E., Van Altena, W.F.: 1989, *Astron. J.*, **98**, 227.
Giampapa, M.S., Radick, R.R., Hall, J.C., Baliunas, S.L.: 2000, *Bull. American Astron. Soc.*, **32**, 832.
Hand, D.J.: 1981, Discrimination and Classification, Chichester, Wiley.
McLachlan, G., Krishnan, T.: 1997. The EM Algorithm and Extensions, John Wiley.
McLachlan, G.J., Peel, D., Basford, K.E., Adams. P.: 1999, Fitting of mixtures of normal and t components. *Journal of Statistical Software*, **4**, n. 2.
McLachlan, G., Peel, D.: 2000, Finite Mixture Models, Wiley.
Montgomery, K.A., Marschall, L.A., Janes, K.A.: 1993, *Astron. J.*, 1106,181.
Sanders, W.L.: 1971, *Astron. Astrophys*, **14**, 226.
Sanders, W.: 1977, *Astron. Astrophys. Suppl. Series*, **27**, 89.
Sanders, W.: 1989,*Rev. Mex. Astron. Astrophys.*, **17**, 31.
Van den Berg, M., Verbunt, F., Mathieu, R.D.: 2000, Proceedings from ASP Conference, **198**, 503.
Wolfe, J.H.: 1970, *Multivariate Behavioral Research*, **5**, 329.
Zhao, J.L., Tian, K.P., Pan, R.S., He, Y.P., Shi, H.M.: 1993, *Astron. Astrophys. Suppl. Series*, **100**, 24.

АЛГОРИТАМ ЗА МАКСИМИЗАЦИЈУ ОЧЕКИВАЊА КАО МОЋНА АЛАТКА ЗА ОДРЕЂИВАЊЕ ПРИПАДНОСТИ ЗВЕЗДА РАЗВЕЈАНОМ ЈАТУ, ПРИМЕНА НА M67

A. Uribe, R. Barrera and E. Brieva

Observatorio Astronómico, Facultad de Ciencias, Universidad Nacional de Colombia, Apartado Aéreo 2584, Bogotá, Colombia

UDC 524.45M67–16–17
Оригинални научни рад

ЕМ алгоритам је моћна алатка за решавање проблема припадности у развејаним јатима када се образује модел густине мешавине који преклапа две хетероседастичке биваријантне нормалне компоненте да би се добило добро слагање за облак релативних сопствених кретања звезда у области неба где се претпоставља да ће јато да буде смештено. Испитује се припадност за 1866 звезда које се налазе у пољу врло старог развејаног јата M67 коришћењем алгоритма за максимизацију очекивања, а применом софтвера EMMIX чији су творци МекЛафлин, Пил, Басфорд и Адамс.