A&A 510, A78 (2010)
DOI: 10.1051/0004-6361/200912886
© ESO 2010

Astronomy & Astrophysics

Cluster radius and sampling radius in the determination of cluster membership probabilities

N. Sánchez, B. Vicente, and E. J. Alfaro

Instituto de Astrofísica de Andalucía, CSIC, Apdo. 3004, 18080 Granada, Spain
e-mail: nestor@iaa.es

Received 14 July 2009 / Accepted 24 November 2009

ABSTRACT

We analyze the dependence of the membership probabilities obtained from kinematical variables on the radius of the field of view around open clusters (the sampling radius, R_s). From simulated data, we show that optimal discrimination between cluster members and non-members is achieved when the sampling radius is very close to the cluster radius. At higher R_s values, more field stars tend to be erroneously assigned as cluster members. From real data of two open clusters (NGC 2323 and NGC 2311), we infer that the number of identified cluster members always increases with increasing R_s. However, there is a threshold value $R_{s,opt}$ above which the identified cluster members are severely contaminated by field stars and the effectiveness of membership determination is relatively small. This optimal sampling radius is ≃14 arcmin for NGC 2323 and ≃13 arcmin for NGC 2311. We discuss the reasons for this behavior and the relationship between cluster radius and optimal sampling radius. We suggest that, independently of the method used to estimate membership probabilities, several tests using different sampling radius should be performed to evaluate possible biases.

Key words. methods: data analysis – open clusters and associations: general – open clusters and associations: individual: NGC 2311 – open clusters and associations: individual: NGC 2323

1. Introduction

Large astrometric catalogues derived from surveys covering very wide areas of the sky are allowing the systematic searching of new star systems (see, for example, López-Corredoira et al. 1998; Hoogerwerf & Aguilar 1999; Kazakevich & Orlov 2002; Myullyari et al. 2003; Caballero & Dinis 2008; Zhao et al. 2009, and references therein). The searching process is based on the detection of clearly defined structures in subsets of phase space. Both spatial density peaks and proper motion peaks are indicative of star clusters; peaks detectable in only the proper motion distributions suggest the existence of moving groups, whereas more spread-out and less dense velocity-position correlated structures could be associated with stellar streams. Once these structures have been detected, the next step is to identify possible members of the star system. In the particular case of open clusters, the most often used procedure for selecting possible cluster members is the algorithm designed by Sanders (1971). This algorithm is based on a former model proposed by Vasilevskis et al. (1958) for the proper motion distribution. The model assumes that cluster members and field stars are distributed according to circular and elliptical bivariate normal distributions, respectively. The Sanders' algorithm, or some variation or refinement of it, has been and is still being widely used to estimate cluster memberships either as the only method or as part of a more complete treatment that includes, for example, spatial and/or photometric criteria. Some representative references are Wu et al. (2002), Jilinski et al. (2003), Balaguer-Núñez et al. (2004), Dias et al. (2006), Kraus & Hillenbrand (2007), and Wiramihardja et al. (2009).

With the advent of large catalogues and databases available via internet and future surveys such as the forthcoming *Gaia* mission of ESA, the interest in developing and applying fully automated techniques is of increasing interest among the astronomical community. However, special care must be taken to avoid obtaining biased results. In this work, we show that the results obtained when using the Sanders' algorithm depend significantly on the choice of the size of the field of view surrounding the cluster. So, once a possible open cluster is detected, it is natural to ask which area of the sky should be sampled to obtain the most reliable membership determinations. It is equally important to ask about the robustness of used methodology, i.e., how the solution changes when the sampled area is varied? Here we explore these subjects by using both simulated and real data. In Sect. 2, we briefly present the method used to determine memberships and describe the simulations that we performed to analyze the expected behavior. The results of applying the Sanders' algorithm on the simulated data are discussed in Sect. 3. After this, in Sect. 4 we use real astrometric data of two open clusters (NGC 2323 and NGC 2311) to evaluate the performance of the algorithm. We discuss strategies to estimate the optimal sampling radius, i.e., the maximum radius beyond which the identified cluster members are expected to be severely contaminated by field stars. The main results of the present work are summarized in Sect. 5.

2. Description of the method

2.1. Membership determination

The key point of the membership discrimination method is the assumption that the distribution of observed proper motions (μ_x, μ_y) can be described by means of two bivariate normal distributions, one circular for the cluster and one elliptical for the field

(Vasilevskis et al. 1958). We define Φ_c and Φ_f to be the cluster and field probability density functions, respectively. Then,

$$\Phi_c(\mu_x,\mu_y) = \frac{1}{2\pi\sigma_c^2}\exp\left\{-\frac{1}{2}\left[\left(\frac{\mu_x-\mu_{x,c}}{\sigma_c}\right)^2+\left(\frac{\mu_y-\mu_{y,c}}{\sigma_c}\right)^2\right]\right\} \quad (1)$$

and

$$\Phi_f(\mu_x,\mu_y)=\frac{1}{2\pi\sigma_{x,f}\sigma_{y,f}\sqrt{1-\rho^2}}\exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{\mu_x-\mu_{x,f}}{\sigma_{x,f}}\right)^2\right.\right. \quad (2)$$
$$\left.\left.+\left(\frac{\mu_y-\mu_{y,f}}{\sigma_{y,f}}\right)^2-2\rho\left(\frac{\mu_x-\mu_{x,f}}{\sigma_{x,f}}\right)\left(\frac{\mu_y-\mu_{y,f}}{\sigma_{y,f}}\right)\right]\right\},$$

where $(\mu_{x,c},\mu_{y,c})$ is the cluster distribution centroid with standard deviation σ_c, $(\mu_{x,f},\mu_{y,f})$ is the field centroid with standard deviations $\sigma_{x,f}$ and $\sigma_{y,f}$, and ρ is the correlation coefficient of field stars. The probability density function for the whole sample is simply

$$\Phi(\mu_x,\mu_y) = n_c\Phi_c(\mu_x,\mu_y) + n_f\Phi_f(\mu_x,\mu_y), \quad (3)$$

n_c and n_f being the normalized numbers of cluster and field stars, respectively. To obtain the unknown parameters (centroids, standard deviations, numbers of members and non-members), an iterative procedure is used by applying the maximum likelihood principle (Sanders 1971). Here we use the algorithm proposed by Cabrera-Caño & Alfaro (1985), which first detects and removes outliers that can produce unrealistic solutions, and then uses a more robust and efficient iterative procedure for the model parameter estimation. Once these parameters are known, the membership probability of the ith stars can be calculated directly as

$$p(i) = \frac{n_c\Phi_c(i)}{\Phi(i)}\,. \quad (4)$$

2.2. Simulations

We consider a cluster with a given radius R_c. We define "cluster radius" as the radius of the smallest circle that can completely enclose its stars. In true situations, R_c is an unknown quantity that has to be estimated a posteriori, but here its value is known and remains constant throughout each simulation. The total number of stars belonging to the cluster is denoted by $N_{c,max}$ and the number of field stars lying exactly within the same sky area of the cluster is $N_{f,cri}$. The independent variable is the radius of the field encircling the cluster. This radius might represent the radius of the field in which the observations are made or the field around the cluster extracted from an astrometric catalogue. We call this variable the sampling radius R_s, which can be larger or smaller than the cluster radius R_c.

The numbers of cluster stars and field stars to be simulated are represented by $N_{c,sim}$ and $N_{f,sim}$, respectively. Obviously, the number of clusters stars and field stars *within the field of view* depend on the size of this field, that is, both $N_{c,sim}$ and $N_{f,sim}$ are functions of R_s. If the field stars are distributed nearly uniformly in space, then $N_{f,sim}$ should increase as the sampling radius increases as

$$N_{f,sim}(R_s) = N_{f,cri}\,(R_s/R_c)^2\,. \quad (5)$$

The rate at which $N_{c,sim}$ increases with R_s depends instead on the radial profile of the surface density of cluster stars ($\Sigma_{c,sim}$). For simplicity, we assume that the surface density at r is given by (Caballero 2008)

$$\Sigma_{c,sim}(r) = \frac{\delta N_{c,max}}{2\pi R_c^\delta}r^{\delta-2}, \quad (6)$$

where the index $\delta \le 2$. For the extreme case $\delta = 2$, we have $\Sigma_{c,sim} = N_{c,max}/(\pi R_c^2) =$ const.. Integrating Eq. (6), we obtain the number of cluster stars within a given sampling radius (for $R_s \le R_c$),

$$N_{c,sim}(R_s) = N_{c,max}\,(R_s/R_c)^\delta\,. \quad (7)$$

Negative δ values make no sense, so this approach is limited to the range $0 < \delta \le 2$. The role of the parameter δ is to control how rapidly $N_{c,sim}$ increases as R_s increases. Thus, we do not need to know the exact functional form as long as we are able to simulate either completely flat ($\delta = 2$) or extremely peaked ($\delta \simeq 0$) density profiles.

To perform the simulations, we distribute $N_{f,sim}$ field stars and $N_{c,sim}$ cluster stars according to bivariate Gaussian distributions in the proper motion space (μ_x,μ_y). The routine "gasdev" from the Numerical Recipes package (Press et al. 1992) is used to generate normally distributed random numbers. The fields are centered on $(0,0)$ with standard deviations of $\sigma_{x,f} = \sigma_{y,f} = \sigma_f$. The tests performed using elliptical (rather than circular) distributions for the field stars yielded essentially the same results and trends. The clusters are centered on $(\mu_{x,c},\mu_{y,c})$ and have standard deviations $\sigma_{x,c} = \sigma_{y,c} = \sigma_c$. Thus, for a given sampling radius R_s and according to Eq. (5), we randomly generate $N_{f,sim}$ field stars that follow a bivariate normal distribution in the proper motion space. For the cluster, we generate $N_{c,sim}$ stars according to Eq. (7) when $R_s \le R_c$, and we generate $N_{c,sim} = N_{c,max} =$ const. stars when $R_s \ge R_c$. The three free parameters, excluding those describing the Gaussians, are the total number of stars in the cluster ($N_{c,max}$), the number of field stars within the cluster area ($N_{f,cri}$), and the cluster star density profile (δ). For each set of parameters, we performed 100 simulations and calculated both the average values of the studied quantities and their corresponding standard deviations.

3. Results from simulations

For each simulation, we calculated cluster membership probabilities using the method described in Sect. 2.1. We performed several simulations by varying the input parameters (the number of stars in both the cluster and the field, the centroid distance in the proper motion space, and standard deviations) within reasonable ranges. Apart from minor differences, such as the error bars being larger when cluster and field distributions are more similar, all the results and trends remained essentially identical to those described in this section. We begin by showing how the algorithm works. In Fig. 1, we can see an example of a simulation of a cluster of 200 stars, which has been adequately sampled with $R_s = 1.1R_c$ The right panel clearly shows the occasional but inevitable "failures" of the method. First, cluster stars in the tails of their own distribution may not be recognized as members. Second, field stars located by chance below the cluster distribution may be selected as probable members.

What would happen if we select a larger field? To address this point, we calculated membership probabilities as a function of the sampling radius. Here we consider as cluster members stars with membership probabilities ≥0.5 in a Bayesian sense. We performed several tests of different selection criteria. As expected, the number of assigned members depends on the selection criterion used, although the main results and trends presented here remain unchanged. Figure 2 shows the number of stars classified as members (which we denote by N_c) or non-members (N_f) by the algorithm as a function of the sampling radius. In these particular simulations, the number of assigned

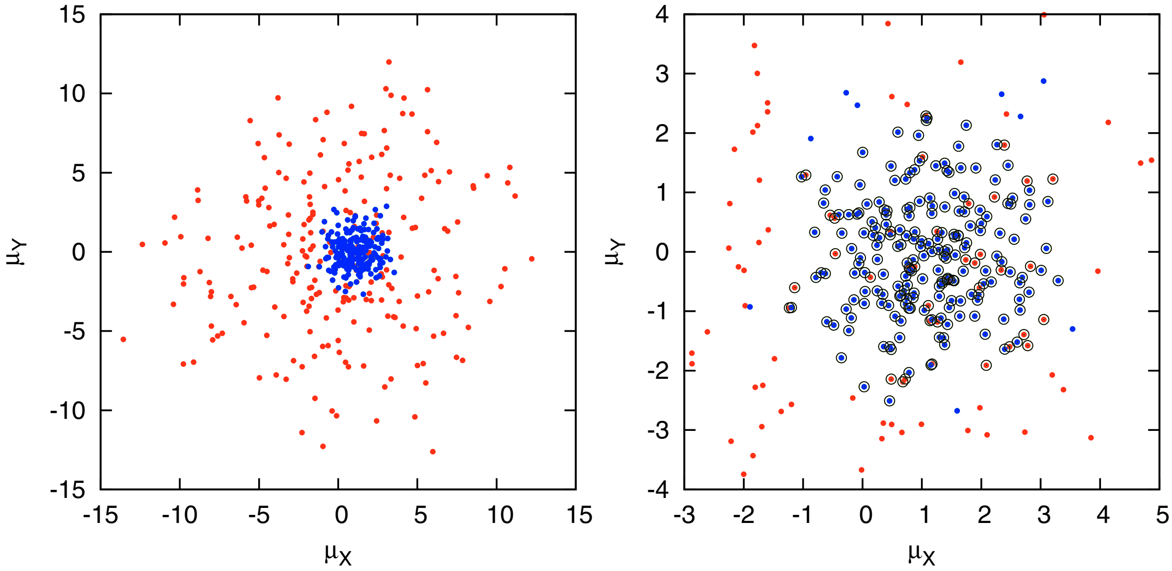

Fig. 1. Proper motion for the stars of a random simulation with $N_{\mathrm{c,max}} = N_{\mathrm{f,cri}} = 200$, $\delta = 2$, and $R_{\mathrm{s}}/R_{\mathrm{c}} = 1.1$ (see text for details of the meaning of each of these quantities). *Left panel* shows the distribution for all the 442 simulated stars. Red circles are the field stars centered on (0, 0) with $\sigma_{\mathrm{f}} = 5$ and blue circles are the 200 cluster stars centered on (1, 0) with $\sigma_{\mathrm{c}} = 1$. *Right panel* is a magnification of the central region in which we have marked with circles the stars whose resulting cluster membership probabilities are higher than 0.5 according to the algorithm used.

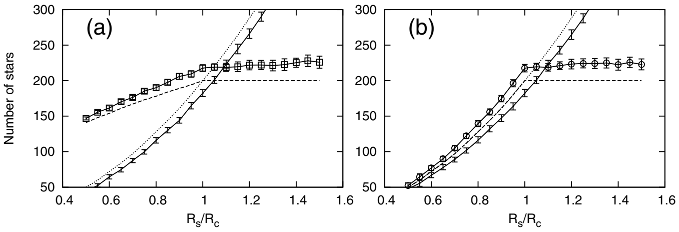

Fig. 2. Calculated number of field and cluster stars as a function of the sampling radius in units of the cluster radius, $R_{\mathrm{s}}/R_{\mathrm{c}}$, for simulations with the same set of parameters as Fig. 1. **a)** Simulation with peaked density profile ($\delta = 0.5$), assigned members are indicated by squares connected by lines. **b)** Simulation with flat density profile ($\delta = 2$), members are indicated by circles connected by lines. Assigned field stars are indicated by vertical bars connected by lines, the length of the bars indicating one standard deviation. The real numbers of simulated stars are shown by dashed lines (cluster) and dotted lines (field).

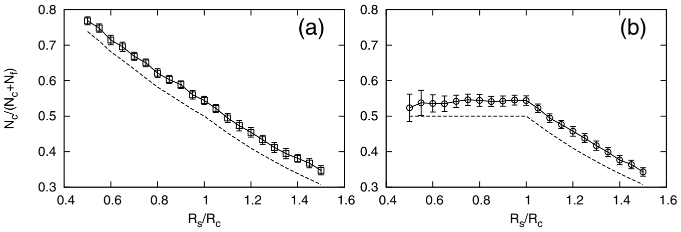

Fig. 3. Calculated fraction of cluster stars as a function of the sampling radius for the same simulations as in Fig. 2. The real (simulated) values are shown by dashed lines.

members N_{c} is always higher than the true number of cluster stars. Most of the cluster stars are well identified but, as mentioned before, field stars falling below the cluster distribution are also considered as members. For the same reason, the number of field stars is always lower than expected. For $R_{\mathrm{s}} < R_{\mathrm{c}}$ (subsampled cluster), N_{c} increases with R_{s} because obviously the number of cluster stars in the sample increases as R_{s} increases. The rate at which this occurs depends on the cluster density profile, which for simulations with $\delta = 2$ in Fig. 2 is exactly the same as for the field (homogeneous distribution). For $R_{\mathrm{s}} \geq R_{\mathrm{c}}$, we observe a change in the behavior of N_{c}. In this case, we do not include new cluster stars in the sample as R_{s} increases, and N_{c} increases slightly because of the new field stars that are erroneously classified as possible members. On the other hand, the number of field stars always increases at a rate roughly proportional to R_{s}^2. It is easy to see that, in general, the fraction of cluster stars (shown in Fig. 3) should be a decreasing function of R_{s} for any cluster with $\delta < 2$. Only in the extreme case of homogeneous clusters, the fraction of cluster stars remains constant with R_{s} for $R_{\mathrm{s}} < R_{\mathrm{c}}$.

Figures 2 and 3 show the number of stars classified as members, although we do not know whether this classification is indeed reliable. To quantify the correctness of the result, we define the matching fraction of the cluster M_{c} to be the net proportion of cluster stars that are well classified. If N_{ok} is the total number of cluster stars correctly classified as members minus the number of cluster stars incorrectly classified as non-members, then $M_{\mathrm{c}} = N_{\mathrm{ok}}/N_{\mathrm{c,max}}$. The value of M_{c} can be a negative number if the number of misclassifications is higher than the number of correct classifications and M_{c} is exactly 1 only when the algorithm classifies correctly all the stars of the cluster. In Fig. 4, we see that the highest M_{c} value occurs precisely when the sampling radius equals the cluster radius. At smaller sampling radii, the matching fraction of the cluster obviously decreases because the cluster is being subsampled. Interestingly, the matching fraction is also smaller at $R_{\mathrm{s}} > R_{\mathrm{c}}$, but the reason in this case is that more field stars are being erroneously assigned to the cluster as R_{s} increases. The most robust classification is achieved when the sampling radius is very close to the cluster radius, although, as expected, even in this case the matching fraction does not reach its maximum value $M_{\mathrm{c}} = 1$. However, the matching fraction is

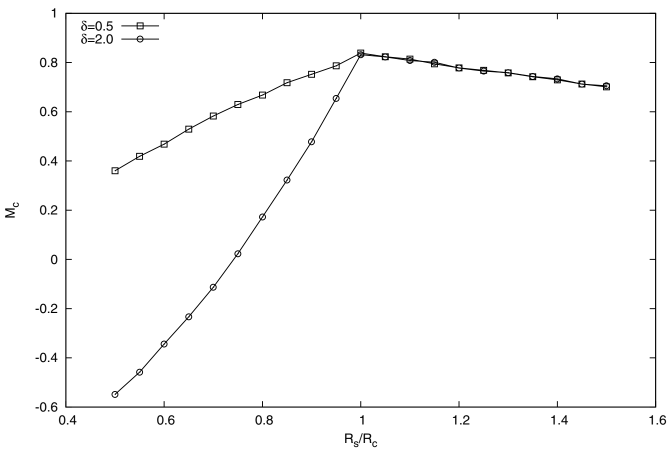

Fig. 4. Matching fraction of the cluster (see text) as a function of the sampling radius for the same simulations as in Fig. 2. The error bars are of the order of the symbol sizes but are not shown for clarity.

relatively high ($M_\mathrm{c} = 0.83$) at $R_\mathrm{s} = R_\mathrm{c}$ and decreases slowly to 0.71 at $R_\mathrm{s} = 1.5\ R_\mathrm{c}$. Moreover, the behaviors of N_c and N_f with R_s are very similar to those expected (Figs. 2 and 3). This is because both cluster and field stars were assumed to have perfect normal distributions and, therefore, both populations can be well detected by the algorithm since it assumes the same kind of underlying distribution. When using real data, the situation becomes more complex, as discussed in the next section.

4. Results using real data

We use the CdC-SF Catalogue (Vicente et al. 2010), an astrometric catalogue with a mean precision in proper motion of 2.0 mas/yr (1.2 mas/yr for reliable measurements, typically for stars with $V < 14$). Given the position of a known open cluster, we extract circular fields of varying radius centered on it and then we calculate membership probabilities by using the same algorithm as in Sect. 3. Here we analyze two open clusters that are included in the area covered by this catalogue: NGC 2323 (M 50) and NGC 2311. To minimize the influence of possible outliers on our results, we restrict the sample to $|\mu| \leq 20$ mas/yr. The number of probable members N_c, i.e., stars with membership probabilities higher than 0.5, is shown in Fig. 5 as a function of the sampling radius. In general, N_c always increases with increasing R_s and there are no relatively flat regions analogous to those observed in Fig. 2 for $R_\mathrm{s} > R_\mathrm{c}$. Without previous knowledge of the approximate value of the cluster radius, how can we determine the most reliable result? This is not a trivial question given the large uncertainties involved in the estimation or definition of the cluster radius (see discussion in Sect. 4.2). For example, the radius of the total extent of NGC 2323 estimated by different authors varies widely: 10 arcmin (Claria et al. 1998), 16.7 arcmin (Nilakshi et al. 2002), 15 arcmin (Kalirai et al. 2003), 22.2 arcmin (Kharchenko et al. 2005), 17 arcmin (Sharma et al. 2006, using their own optical data), or 22 arcmin (Sharma et al. 2006, using 2MASS data). Our calculations yield $N_\mathrm{c} = 198$ probable members in a field of radius $R_\mathrm{s} = 17$ arcmin, but this number increases to $N_\mathrm{c} = 336$ for $R_\mathrm{s} = 22$ arcmin. This means that there could be more than 100 undetected members, if we use $R_\mathrm{s} = 17$ arcmin and the cluster radius is actually $R_\mathrm{c} = 22$ arcmin. Conversely, there are more than 100 spurious members if we use $R_\mathrm{s} = 22$ arcmin and $R_\mathrm{c} = 17$ arcmin. The fraction of cluster members is shown in Fig. 6. The trend in which $N_\mathrm{c}/(N_\mathrm{c} + N_\mathrm{f})$ decreases with R_s is qualitatively consistent with the expected behavior (Fig. 3). However, there is a R_s value from which the fraction of members increases as R_s increases and, as mentioned in the previous section, this behavior is only possible if N_c increases faster than N_f does (i.e., at a rate higher than $\sim R_\mathrm{s}^2$). The only way that this could happen is if the algorithm introduces many spurious members as R_s increases. In other words, there is a critical R_s value above which a significant number of spurious members are erroneously included as part of the cluster (see also Piatti et al. 2009). Here we call this critical value the optimal sampling radius, $R_\mathrm{s,opt}$, and obviously do not recommend using a sampling radius larger than this value. From Fig. 6, we obtain $R_\mathrm{s,opt} \simeq 14$ arcmin for NGC 2323 and $R_\mathrm{s,opt} \simeq 13$ arcmin for NGC 2311, but we emphasize that these values are valid for the data that we use and, in principle, they cannot be extrapolated to other data sets.

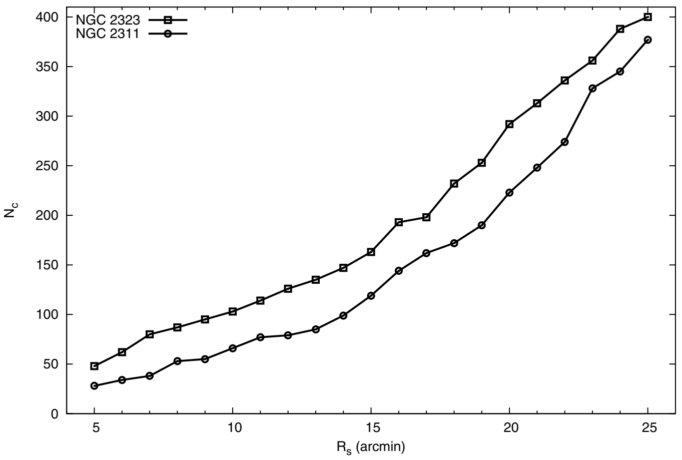

Fig. 5. Number of cluster stars N_c as a function of the sampling radius R_s in arcmin for the open clusters NGC 2323 (squares connected by lines) and NGC 2311 (circles connected by lines).

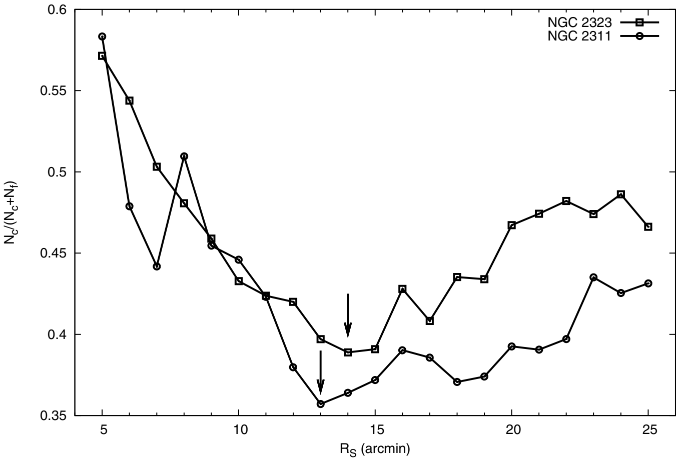

Fig. 6. Fraction of cluster stars as a function of the sampling radius for NGC 2323 (squares connected by lines) and NGC 2311 (circles connected by lines). Vertical arrows indicate the optimal sampling radii (see text).

The main reason behind the behavior observed in Fig. 6 is the disagreement between the assumed and the "true" underlying distributions of proper motion of field stars. A circular normal bivariate function is a good representation of the cluster probability density function (PDF), the standard deviation being

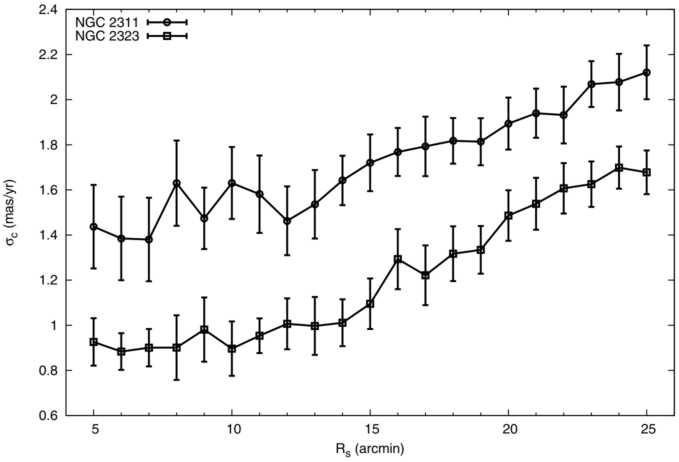

Fig. 7. Estimated standard deviations as a function of the sampling radius for the clusters NGC 2323 (squares connected by lines) and NGC 2311 (circles connected by lines). The bars indicate the uncertainties obtained from bootstrapping.

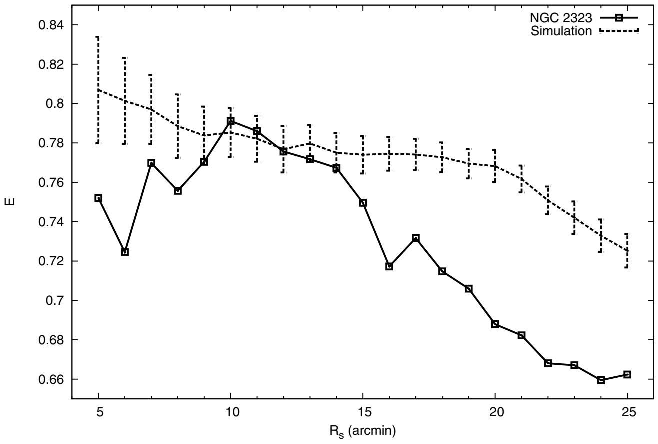

Fig. 8. Effectiveness of membership determination (see Eq. (8)) as a function of the sampling radius for the open cluster NGC 2323 (open squares connected by solid lines) and for simulations using parameter values corresponding to those obtained for NGC 2323 (dashed lines).

the result of observational errors that prevent the intrinsic velocity dispersion of the cluster from being completely resolved. However, it is known that an elliptical normal bivariate function is not always the most reliable model for the field PDF (see discussions on this subject in Cabrera-Caño & Alfaro 1990; Uribe & Brieva 1994; Balaguer-Núñez et al. 2004; Sánchez & Alfaro 2009; Griv et al. 2009). The combination of several factors, such as galactic differential rotation or peculiar motions, may affect the field star distribution, which usually tends to exhibit non-Gaussian tails. Non-parametric models, which make no a priori assumptions about the cluster or field star distributions, were introduced and used to overcome this problem (cf. Cabrera-Caño & Alfaro 1990; Chen et al. 1997). We note that both the classical parametric and non-parametric methods agree reasonably well with each other only for nearly Gaussian field distributions (see Fig. 5 in Sánchez & Alfaro 2009). When the number of field stars increases and the algorithm tries to fit a Gaussian function to the PDF, the fit tends to produce a wider and flatter function. As a consequence, the membership probabilities (defined as the ratio of the cluster to the total proper motion distribution function) increases and the number of assigned members therefore also increases. This effect is magnified when the cluster distribution becomes "contaminated" by many field stars, because the standard deviation of the cluster then tends to increase with the consequent increase in the number of spurious members. The standard deviations estimated for the two clusters being studied are shown in Fig. 7. The error bars were estimated using bootstrap techniques: the calculation is repeated for a series of 100 random resamplings of the data and the standard deviation of the obtained set of values is taken as the associated uncertainty. The standard deviations remain nearly constant ($\sigma_c \simeq 1.4$–1.6 for NGC 2311 and $\sigma_c \simeq 0.9$–1.0 for NGC 2323) in the region in which $R_s \lesssim R_{s,opt}$ (see also Fig. 6). This is the expected behavior because, in principle, σ_c should not depend on the sample size. However, above the optimal sampling radius, we can see a gradual increase in σ_c because of the effect mentioned previously.

4.1. Effectiveness of membership determination

It is not possible in practice to quantify the degree of correlation between identified and true cluster members, such as the matching fraction in Fig. 4. Instead, we can use the concept of effectiveness of membership determination, which is defined to be (Tian et al. 1998; Wu et al. 2002)

$$E = 1 - \frac{N \sum_{i=1}^{N} \{p(i)\,[1 - p(i)]\}}{\sum_{i=1}^{N} p(i) \sum_{i=1}^{N} [1 - p(i)]} \,, \tag{8}$$

where $p(i)$ is the membership probability of the ith star and N is the sample size. This index measures the effectiveness of the membership determination by measuring the separation between field and cluster populations in the probability histogram. The higher the index E, the more effective the membership determination. The maximum E value is obtained when there are two perfectly separated populations of N_c stars with membership probabilities $p(i) = 1$ and N_f stars with $p(i) = 0$. Figure 8 shows E for the open cluster NGC 2323 as a function of the sampling radius. For the sake of comparison, we also show the result for simulations using the same parameters obtained for NGC 2323. Our most reliable estimation for this cluster ($R_s = R_{s,opt} = 14$ arcmin) yielded the following values of proper motions (in mas/yr): $\mu_{x,c} = 1.09$, $\mu_{y,c} = 1.13$, $\sigma_{x,c} = \sigma_{y,c} = 1.01$, $\mu_{x,f} = +0.77$, $\mu_{y,f} = -2.54$, $\sigma_{x,f} = 6.41$, and $\sigma_{y,f} = 5.84$. According to the result shown in Fig. 9 (next section), we assume $R_c = 20$ arcmin and $\delta = 1.7$ for the cluster. In addition, we choose $N_{c,max} = 250$ and $N_{f,cri} = 500$ to obtain the measured values $N_c = 147$ and $N_f = 231$ at $R_s = 14$ arcmin. The superimposed dashed lines in Fig. 8 are the average values (and their standard deviations) for these simulations. The simulated E value remains fairly constant (within the uncertainties) as R_s increases up to the value $R_s \simeq R_c = 20$ arcmin, beyond which it decreases at a relatively rapid rate. For NGC 2323, we see that E begins to decrease more rapidly as R_s increases just beyond $R_s \simeq R_{s,opt} = 14$ arcmin. The clearest separations between cluster and field stars and the closest agreement with the simulations are achieved in the range $10 \lesssim R_s \lesssim 14$ arcmin.

4.2. Cluster radius and optimal sampling radius

When using only kinematical criteria, we propose that the sample size can substantially alter the results obtained (the memberships and the remaining properties derived from there). Thus, the strategy of choosing a field large enough to be sure of covering more than the entire cluster must be performed carefully, especially in dense star fields. According to our simulations (Sect. 3), the most robust membership estimation is achieved when $R_{\rm s} \simeq R_{\rm c}$. This would seem an obvious result, given that for $R_{\rm s} < R_{\rm c}$ the cluster is subsampled, whereas for $R_{\rm s} > R_{\rm c}$ the probability of contamination by field stars is increased. This illustrates that it is important to know the cluster radius reliably before estimating memberships. It is difficult to determine precisely the radius of a cluster because the definition of radius is ambiguous itself, since star clusters have no clearly defined natural boundaries. In this work, we have used the usual definition of $R_{\rm c}$, which is the radius of the circle containing all the cluster members. Most "geometric" definitions tend to overestimate the true size, especially for irregularly shaped clusters (Schmeja & Klessen 2006). But this is not the main problem, which is instead that the independent estimations of cluster radii available in the literature usually differ significantly. Angular sizes listed in catalogues as Webda[1] were compiled from older references (e.g., Lynga 1987) in which most of the apparent diameters were estimated from visual inspection. According to Webda, for NGC 2323 $R_{\rm c}$ = 7 arcmin, whereas Sharma et al. (2006) estimate $R_{\rm c} \sim 20$ arcmin. As mentioned above, it is usual practice to choose a field larger than the apparent area covered by the cluster (taken from the literature) to estimate membership probabilities. However, at least when applying the Sanders' method, assigned members will be spread throughout the selected area because of contamination by field stars. It is probably not coincidental that this is true, for example, for probable members in the Dias catalogue (Dias et al. 2002). How reliable are memberships derived from proper motions? It depends on the "true" $R_{\rm c}$ values. Thus, again, a reliable assessment of membership should use some robust estimation of the radius.

A commonly used procedure for determining (or defining) the cluster radius is based on the analysis of the projected radial density profile. Usually, some particular analytical function (for example, a King-like model) is fitted to the density profile and the cluster radius is extracted from this fit. A systematic determination of cluster sizes based on objective and uniform estimations of radial density profiles was performed by Kharchenko et al. (2004). One limitation of this method is the sensitivity of the fit to small variations in the distribution of stars, especially for poorly populated open clusters. The most reliable fits are obtained by using only cluster members, but we are then affected again by the problem of membership determination. As an example, we consider Fig. 9, which compares the density profiles obtained for the open cluster NGC 2323 for two different sampling radii $R_{\rm s} = R_{\rm s,opt}$ = 14 arcmin and $R_{\rm s}$ = 25 arcmin. According to our results (Sect. 4), our most reliable estimation is achieved when $R_{\rm s} = R_{\rm s,opt}$. In this case, the least squares fit to a power law function infers a cluster radius in the range ~20–25 arcmin. However, if we consider a sample of size $R_{\rm s}$ = 25 arcmin, the contamination by field stars tends to cause the overestimation of the star density and both the index of the power law and the estimated cluster radius change significantly (see Fig. 9). But the main drawback of this method is that simple analytical fits are not always a good representation of the star distribution in open clusters (Sánchez & Alfaro 2009). The radius defined by fitting a density profile may be useful in analyzing and comparing the properties of several clusters systematically, but great care must be taken when using these model-dependent definitions to estimate the "true" cluster radius. The point at which the fitted star density equals the background (or drops to zero) does not even necessarily agree with the outer boundary of an open cluster. In principle, new-born stars in a young cluster should be spread throughout the region that is collapsing to form the cluster. At a certain distance from the high density peak in the molecular cloud, the required conditions are no longer fulfilled and the star formation efficiency may decrease abruptly. So, a radial star density distribution that decays smoothly to $R_{\rm c}$ may not always be suitable, especially for compact and/or very young star clusters. Moreover, if the clusters exhibit some degree of substructure, this type of procedure yields totally unrealistic results (Sánchez & Alfaro 2009). Young embedded clusters often show hierarchical structure (Elmegreen 2009), so that these methods cannot in principle be applied to embedded clusters but only to centrally concentrated open clusters.

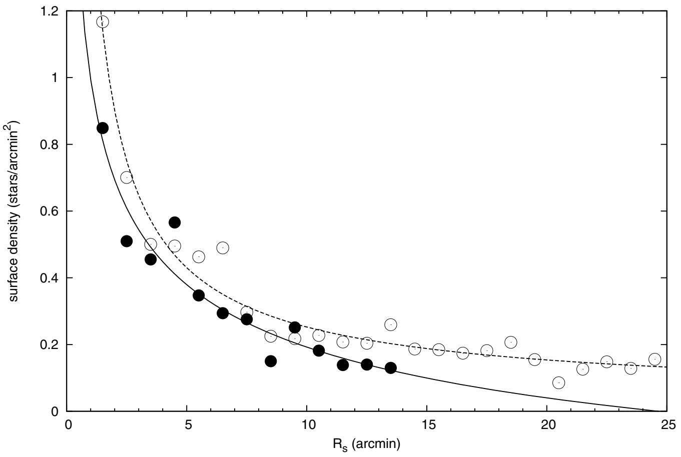

Fig. 9. Radial density profiles for the cluster NGC 2323 calculated for the cases $R_{\rm s}$ = 14 arcmin (solid circles) and $R_{\rm s}$ = 25 arcmin (open circles). Lines show the best fits to functions of the form $\sim r^{\delta-2}$ (see Eq. (6)). The solid line represents the case $R_{\rm s}$ = 14 arcmin for which $\delta \simeq 1.7$, and the dashed line corresponds to $R_{\rm s}$ = 25 arcmin for which $\delta \simeq 1.2$.

Obviously, any reliable estimation of the cluster radius ultimately depends on the membership determination. Field star contamination may affect the determination of $R_{\rm c}$, and what we have demonstrated in this work is that this contamination can become a significant problem if not taken into consideration. Furthermore, even though cluster and field populations were well separated, the estimated radius would depend on the limit magnitude if, for instance, there was mass segregation. This kind of problems is particularly relevant to the development of automated techniques in which it is necessary to establish objective criteria when determining the size of the sample to be processed. What we propose here is to apply any suggested method to several sample sizes $R_{\rm s}$ and analyze the behavior obtained. It is difficult to establish simple rules for evaluating this behavior because the results will depend directly on both the membership determination algorithm and the input data. However, for the method considered in this work, which is based on two Gaussian populations, the basic procedure can be outlined as follows:

1. An upper limit to $R_{\rm s}$ can be estimated by fitting the spatial star density to, for example, a King profile. The estimated tidal radius (or, to be conservative, twice its value) may be considered an upper limit to the optimal sampling radius and would define the range of $R_{\rm s}$ values to be scanned.
2. For each $R_{\rm s}$ value, cluster memberships and all the relevant quantities (numbers of cluster stars and field stars, centroids with their standard deviations, effectiveness of membership determination) have to be estimated.

[1] http://www.univie.ac.at/webda

3. The next step is to plot the number of cluster members N_c as a function of the sampling radius R_s. If the membership determination works reasonably well, meaning that it presents little contamination by field stars, then we would observe a behavior similar to that seen in Fig. 2: N_c increases as R_s increases until some point (when $R_s = R_c$) and then N_c remains approximately constant for higher R_s values (or increases at a much slower rate). In this way, we can estimate the cluster size directly from the data and the membership criteria without making any additional assumptions. The optimal sampling radius at which we achieve the most reliable membership estimation is precisely $R_{s,opt} = R_c$ (Fig. 4)
4. If the parametric model does not adequately describe the real data and/or if the internal noise does not have simple properties, then the behavior of the estimated parameters with respect to R_s would differ from that expected. If this were the case, we should plot the fraction of members $N_c/(N_c + N_f)$ versus R_s, where we would identify the optimal sampling radius $R_{s,opt}$ with the minimum in this plot (Fig. 6). In the absence of more accurate information, this value should correspond to the radius at which the membership classification is the most reliable (for this method in a given astrometric catalogue).
5. Our experience indicates that the properties derived from the Sanders' method tend to exhibit noise and it is not always easy to identify precisely the position of specific features (such as the minimum in the $N_c/(N_c + N_f)$ versus R_s plot). Some complementary strategies may be useful in identifying or confirming the optimal sampling radius. First, one can consider the variation in the proper motion standard deviation with radius. The dispersion in the cluster proper motions should exhibit a change of slope at radius close to the optimal sampling radius (Fig. 7). Secondly, the maximal effectiveness of membership determination should be reached around $R_{s,opt}$ (Fig. 8).

The strategy proposed in this work, i.e., to estimate and analyze cluster memberships as a function of R_s, should in principle allow us to identify the optimal sampling radius. However, we emphasize that it may not always be possible (or at least not always unambiguous) to determine $R_{s,opt}$ in the way described above. For instance, for very peaked cluster density profiles the change in N_c at $R_s = R_c$ may not be significant enough to be easily detected (e.g., Fig. 2a). In spite of this, it seems appropriate and useful to perform these tests before further analysis.

5. Conclusions

We have evaluated the performance of the commonly used Sanders' method (Vasilevskis et al. 1958; Sanders 1971; Cabrera-Caño & Alfaro 1985) for determining star cluster memberships. In general, the results depend on the radius of the field containing the sampled cluster (the sampling radius, R_s). The main reason for this dependence is the difference between the assumed Gaussian and the true underlying proper motion distributions. The contamination of cluster members by field stars increases as the sampling radius increases. The rate at which this effect occurs depends on the intrinsic characteristics of the data set. There is a threshold value of R_s above which the identified cluster members are highly contaminated by field stars and the effectiveness of membership determination is relatively small. Thus, care must be taken when applying the Sanders' method (by itself or as part of a more extensive procedure) especially when we do not have reliable information about the true cluster radius and/or when the sampling radius is larger than the cluster radius. If this type of effect is not taken into consideration in automated data analysis then significant biases may arise in the derived cluster parameters. The optimal sampling radius can be estimated by plotting the number of cluster members and/or the fraction of members as a function of the sampling radius. Moreover, this type of analysis can also be used as an objective procedure that can be applied systematically to determine cluster radii.

Acknowledgements. We thank the referee for his/her comments which improved this paper. We acknowledge financial support from MICINN of Spain through grant AYA2007-64052 and from Consejería de Educación y Ciencia (Junta de Andalucía) through TIC-101 and TIC-4075. N.S. is supported by a post-doctoral JAE-Doc (CSIC) contract. E.J.A. acknowledges financial support from the Spanish MICINN under the Consolider-Ingenio 2010 Program grant CSD2006-00070: "First Science with the GTC".

References

Balaguer-Núñez, L., Jordi, C., Galadí-Enríquez, D., et al. 2004, A&A, 426, 819
Caballero, J. A. 2008, MNRAS, 383, 375
Caballero, J. A., & Dinis, L. 2008, Astron. Nachr., 329, 801
Cabrera-Caño, J., & Alfaro, E. J. 1985, A&A, 150, 298
Cabrera-Caño, J., & Alfaro, E. J. 1990, A&A, 235, 94
Chen, B., Asiain, R., Figueras, F., et al. 1997, A&A, 318, 29
Claria, J. J., Piatti, A. E., & Lapasset, E. 1998, A&AS, 128, 131
Dias, W. S., Alessi, B. S., Moitinho, A., et al. 2002, A&A, 389, 871
Dias, W. S., Assafin, M., Flório, V., Alessi, B. S., & Líbero, V. 2006, A&A, 446, 949
Elmegreen, B. G. 2009, Ap&SS, 153
Griv, E., Gedalin, M., & Eichler, D. 2009, AJ, 137, 3520
Hoogerwerf, R., & Aguilar, L. A. 1999, MNRAS, 306, 394
Jilinski, E. G., Frolov, V. N., Ananjevskaja, J. K., Straume, J., & Drake, N. A. 2003, A&A, 401, 531
Kalirai, J. S., Fahlman, G. G., Richer, H. B., et al. 2003, AJ, 126, 1402
Kazakevich, E. É., & Orlov, V. V. 2002, Astrophysics, 45, 302
Kharchenko, N. V., Piskunov, A. E., Röser, S., Schilbach, E., & Scholz, R.-D. 2004, AN, 325, 740
Kharchenko, N. V., Piskunov, A. E., Röser, S., Schilbach, E., & Scholz, R.-D. 2005, A&A, 438, 1163
Kraus, A. L., & Hillenbrand, L. A. 2007, AJ, 134, 2340
López-Corredoira, M., Garzón, F., Hammersley, P. L., et al. 1998, MNRAS, 301, 289
Lynga, G. 1987, Computer Based Catalogue of Open Cluster Data, 5th edn. (Strasbourg: CDS)
Myullyari, A. A., Flynn, C., & Orlov, V. V. 2003, Astron. Rep., 47, 169
Nilakshi, Sagar, R., Pandey, A. K., & Mohan, V. 2002, A&A, 383, 153
Piatti, A. E., Clariá, J. J., Parisi, M. C., et al. 2009, New Astron., 14, 97
Press, W. H., Teukolsky, S. A., Vetterling, W. T., et al. 1992, Numerical recipes in FORTRAN. The art of scientific computing (Cambridge: University Press)
Sánchez, N., & Alfaro, E. J. 2009, ApJ, 696, 2086
Sanders, W. L. 1971, A&A, 14, 226
Schmeja, S., & Klessen, R. S. 2006, A&A, 449, 151
Sharma, S., Pandey, A. K., Ogura, K., et al. 2006, AJ, 132, 1669
Tian, K.-P., Zhao, J.-L., Shao, Z.-Y., et al. 1998, A&AS, 131, 89
Uribe, A., & Brieva, E. 1994, Ap&SS, 214, 171
Vasilevskis, S., Klemola, A., & Preston, G. 1958, AJ, 63, 387
Vicente, B., Abad, C., Garzón, F., et al. 2010, A&A, 509, A62
Wiramihardja, S. D., Arifyanto, M. I., & Sugianto, Y. 2009, Ap&SS, 319, 125
Wu, Z. Y., Tian, K. P., Balaguer-Núñez, L., et al. 2002, A&A, 381, 464
Zhao, J., Zhao, G., & Chen, Y. 2009, ApJ, 692, L113