Astronomy & Astrophysics

Globular clusters of the Local Group – statistical classification*

Tanuka Chattopadhyay[1] and Asis Kumar Chattopadhyay[2]

[1] Shibpur Dinobundhoo College, 412/1 G.T. Road (South), Howrah 711 102, India,
Visiting Associate, Inter-University Center for Astronomy and Astrophysics, Post Bag 4, Ganeshkhind, Pune 411007, India
e-mail: tanuka@iucaa.ernet.in

[2] Department of Statistics, Calcutta University, 35 Ballygunge Circular Road, Calcutta 700019, India
e-mail: akcstat@caluniv.ac.in

ABSTRACT

Aims. To find an objective classification of the globular clusters in our Galaxy, M 31, and LMC.

Methods. A new method of Cluster Analysis (CA)was carried out and the set of parameters for this method was selected through an objective process, Principal Component Analysis (PCA). Robustness of the classification was established using bootstrap samples.

Results. In every case they exhibit multi-population structure instead of bimodality as is found in many spirals and giant elliptical galaxies. The kinematics of MW and M 31 GCs are examined in support of these sub-populations in the cluster system. It is found that for MW and M 31 GCs a disc, inner halo, and outer halo populations of GCs are more likely to exist than only the disc and halo populations of GCs in MW as concluded by Zinn (1985, ApJ, 293, 424). This supports the existence of three populations more firmly explained by Zinn (1993, Globular Cluster-Galaxy Connection, 38) and Mackey & Gilmore (2004, MNRAS, 355, 504) whereas only two populations are found for LMC GCs. The new multivariate analysis increases the importance of the inclusion of many parameters while at the same time it eliminates less significant parameters and helps to enunciate a unique theory of galaxy formation.

Key words. Galaxy: kinematics and dynamics – galaxy: globular clusters: general – galaxies: Local Group – methode: statistical

1. Introduction

Globular clusters have long been considered as the unique tool for studying galaxy formation and evolution, but the first step is to study the formation process of the globular clusters (GCs) themselves. There are various theories regarding the formation of GCs. Among them (1) gaseous merger, (2) in situ GC formation, and (3) tidal stripping are important. According to merger theory, GCs in gE and cD galaxies were created by the gaseous merger of two progenitor spiral galaxies (Ashman & Zepf 1992; Zepf & Ashman 1993) so that there are two populations of GCs. One is the metal-poor GCs of the progenitor spirals and other is the metal-rich GCs formed in the collision of high velocity gas. One of the most noteworthy success of this model is the discovery of the protoglobular clusters in the currently merging galaxies (Whitmore et al. 1993; Schweizer et al. 1996). In the merger model the high S_N galaxies said to be created by the merger of several normal S_N galaxies. But in practice the GC systems of high S_N galaxies do not have more metal-rich GCs but instead have more metal-poor ones. This means that GC metallicity gradients are steeper in high S_N galaxies contrary to the prediction of Ashman & Zepf (1992). According to the most favourable theory (Forbes et al. 1997) GCs are considered to be formed in situ star formation episodes during a collapse process. In the first episode a chaotic merging of many small gaseous subunits (Searle & Zinn 1978; Katz 1992) occurs. During this phase a small fraction of the gas turns into stars and most of them reside in GCs, i.e. the ratio of stars in GCs to field stars is large (Forbes et al. 1997). So, the GCs formed in this phase are metal-poor. In the second phase, the stars enrich the medium and now field stars form in large numbers (due to more efficient cooling at high metallicity) and the GCs which form in this phase are metal-rich. In the third phase the remaining gas settles as a galactic disc at the centre of the galaxy. Spiral galaxies may be considered as an extreme case of this process. Here there is almost no pre-galaxy star formation. So the GCs which form in the second phase are metal-poor compared to the metal-rich GCs in the second phase for elliptical galaxies. Spirals then go on to form a prominent disc and associated GCs in the third phase of collapse. But the main difficulty of this process is the lack of a detailed mechanism for creating distinct phases of GCs formation from a single halo collapse. Also, this analysis is intended to find the bimodality of a single parameter which is colour.

In the present study we carry out a multivariate analysis, which is likely to be more appropriate in a multivariate set up. Early attempts to analyse the characteristics of the Galaxy and GCs by statistical methods were carried out by Brosche (1973), Peterson & King (1975), Brosche & Lentes (1984), Eigenson & Yatsyk (1989), Djorgovski (1991), Covino & Fracassini (1993). Some correlations of the slope of GCs present day mass function (PDMF) with other parameters were studied by Capaccioli et al. (1991) who found that two or at most three significant parameters determine the PDMF slope. This problem was also discussed by Djorgovski (1991). Several correlations among GCs metallicity and galaxy parameters were studied by van den Bergh (1975), Brodie & Huchra (1991), Forbes et al. (1996), Harris (1991), Djorgovski (1995). Djorgovski (1995) uses core radius, velocity dispersion, central surface brightness, and mass-to-light ratio to define a Fundamental Plane for the GCs of elliptical galaxies. These correlations provide rough distance indicators for GCs. In an earlier work Covino & Fracassini (1993) carried out a Principal Component Analysis (PCA) followed by

* Appendices A and B are only available in electronic form at http://www.aanda.org

Cluster Analysis (CA) for more than one parameter at a time to study the clustering nature of GCs in different galaxies in the Local Group. Their analysis suffers from several difficulties. First in their study the parameter set is not same for all galaxies. Also the number of parameters is very large (6 for M 31, 8 for LMC, and 10 for Milky Way) so that it is difficult to study the clustering nature in finer details and their interrelations. Finally, the number of clusters, which they selected in an ad hoc manner, is the eighth hierarchy stage. Also the sample size is small.

In the present problem we have first used PCA to search for the optimum set of parameters which gives the maximum variation for the globular clusters in Milky Way, M 31, and LMC. This method reduces the number of parameters to be selected for CA and hence serves the purpose of PCA. Then we took that optimum set of parameters and applied a new method of CA (Sugar & James 2003) which finds the optimum number of groups of GCs instead of choosing group in an ad hoc manner. Also in our case the sample size of Milky Way is almost double. This helps to sort out the optimum number of clusters and optimum set of parameters so that a more efficient theory of galaxy formation can be developed on the basis of this multivariate analysis. In this connection it is also important to mention that under the given set up a partitioning algorithm for cluster analysis is more appropriate than a hierarchical algorithm since the set up is not nested. The new method for finding the optimum number of clusters is based on partitioning algorithm.

In the present paper we closely follow the approach of Covino & Fracassini (1993) in order to classify GCs. The new aspects of our study are as follows:

1. We incorporate the recent catalogue of Milky Way GCs (Harris 1996) which contains data on 147 GCs which is almost double the sample size (74) used by Covino & Fracassini (1993). Also we used the recent catalogues of M 31 GCs (Barmby et al. 2002) and LMC GCs (Mackey & Gilmore 2003) which include the structural parameters together with photometric parameters unlike the sample used by Covino & Fracassini (1993). Also the core radii values used for LMC GCs were measured using HST which gives more accurate measurements than the previous values.
2. We use PCA to search for the optimum set of parameters giving maximum variation for all the GCs in Milky Way, M 31, and LMC instead of using the method for comparative study between the GCs among the galaxies.
3. We used that optimum set of parameters for CA instead of using different sets of parameters for different galaxies in an ad hoc manner and this process increases the consistency of the study.
4. We used a method of CA (Hartigan 1975) which is based on a partitioning algorithm and then applied a new technique to find the optimum number of groups, not a number selected in an ad hoc manner as the eighth hierarchy stage is selected in the paper by Covino & Fracassini (1993). We have taken different bootstrap samples generated from the original sample to test the robustness of the results of the analyses.

In the following sections we discuss the methods, sample sets and finally the results obtained from the analysis.

2. Method

In order to study the underlying nature of the data under consideration we have to start from the correlation matrix because Principal Component Analysis is based on this correlation or covariance matrix. Although a scatter plot is an essential first step in studying the the association between two variables, it is often useful to quantify the strength of the association by calculating a summary index. One commonly used measure is the correlation coefficient (Pearson's correlation coefficient) denoted by r or r_{xy}, which measures the strength of linear correlation between the values of two parameters x and y.

In Principal Component Analysis (Chattopadhyay & Chattopadhyay 2006) we are interested in discovering which parameters in a data set form coherent subgroups that are relatively independent of one another. The specific aim of the analysis is to reduce a large number of parameters to a smaller number while retaining maximum spread among experimental units. The analysis therefore helps us to determine the optimum set of parameters causing the overall variations in the nature of GCs. PCA has been discussed in detail in Appendix A.

Cluster analysis is the art of finding groups in data. Over the last forty years different algorithms and computer programs have been developed for CA. The choice of a clustering algorithm depends both on the type of data available and on the particular purpose. Generally clustering algorithms can be divided into two principal types viz. partitioning and hierarchical methods.

A partitioning method constructs K clusters i.e. it classifies the data into K groups which together satisfy the requirement of a partition such that each group must contain at least one object and each object must belong to exactly one group. So there are at most as many groups as there are objects ($K <= n$). Two different clusters cannot have any object in common and the K groups together add up to the full data set. Partitioning methods are applied if one wants to classify the objects into K clusters where K is fixed (which should be selected optimally). The aim is usually to uncover a structure that is already present in the data. The K-means method of (MacQueen 1967) is probably the most widely applied partitioning clustering technique.

Hierarchical algorithms do not construct single partition with K clusters but they deal with all values of K in the same run. The partition with $K = 1$ is a part of the output (all objects are together in the same cluster) and also the situation with $K = n$ (each object forms a separate cluster). In between all values of $K = 2, 3, ... n - 1$ are covered in a kind of gradual transition. The only difference between $K = r$ and $K = r + 1$ is that one of the r clusters splits in order to obtain $r + 1$ clusters or two of the $(r + 1)$ clusters combined to yield r clusters. Under this method either we start with $K = n$ and move hierarchically step-by-step, where at each step two clusters are merged, depending on similarity until only one is left i.e. $K = 1$ (agglomerative) or the reverse, i.e. start with $K = 1$ and move step-by-step, where at each step one cluster is divided into two- (depending on dissimilarity) until $K = n$ (divisive). Most of the previous works (Covino & Fracassini 1993) were done on the basis of hierarchical clustering. But we feel that for the problem under consideration the partitioning method is more applicable because (a) A partitioning method tries to select best clustering with K groups which is not the goal of hierarchical method. (b) A hierarchical method can never repair what was done in previous steps. (c) Partitioning methods are designed to group items rather than variables into a collection of K clusters. (d) Since a matrix of distances (similarities) does not have to be determined and the basic data do not have to be stored during the computer run partitioning methods can be applied to much larger data sets. For K-means algorithm (Hartigan 1975) the optimum value of K can be obtained in different ways.

By using this algorithm we first determined the structures of sub populations (clusters) for varying numbers of clusters taking $K = 2, 3, 4$ etc. For each such cluster formation we computed the values of a distance measure $d_K = (1/p)\min_x E[(x_K - c_K)'(x_K - c_K)]$ which is defined as the distance of the x_K vector (values of the parameters) from the centre c_K (which is estimated as the mean value) p is the order of the x_K vector. Then the algorithm for determining the optimum number of clusters is as follows (Sugar & James 2003). Let us denote by d'_K the estimate of d_K at the Kth point. Then d'_K is the minimum achievable distortion associated with fitting K centres to the data. A natural way of choosing the number of clusters is to plot d'_K versus K and look for the resulting distortion curve (Fig. 5). This curve is always monotonic decreasing. Initially one would expect much smaller drops for K greater than the true number of clusters because past this point adding more centres simply partitions within groups rather than between groups. According to Sugar & James (2003), for a large number of items the distortion curve when transformed to an appropriate negative power (p/2), will exhibit a sharp "jump" (if we plot K versus transformed d'_K). Then we calculated the jumps in the transformed distortion as $J_K = (d_K'^{-\mathbf{p}/2} - d_{K-1}'^{-\mathbf{p}/2})$.

The optimum number of clusters is the value of K associated with the largest jump. The largest jump can be determined by plotting J_K against K and the highest peak will correspond to the largest jump (Figs. 6, 7 and 17).

Table 1. Errors and extinctions in ages and colours of the GCs in Milky Way (MW) and LMC.

ID	MW GCs Errors Gyr	$E(B-V)$	ID	LMC GCs Errors log (age) (Gyr)	$E(B-V)$
104	1.3	Harris (1996)	1711	+.05−.05	
288	1.6		1754	+0.06−0.07	
362	2.0		1755		0.01
1261	1.3		1805	+.30−.10	
1851	1.0		1810		0.34
1904	1.5		1818	+.3−.1	0.25
2298	2.2		1831	+.3−.3	
2808	1.6		1835	+.07−.08	
3201	1.6		1850	+.2−.2	0.13
4147	1.6		1854		0.11
4590	1.0		1856	+.3−.3	0.07
5024	1.6		1866		0.01
5053	1.5		1885		0.23
5272	1.0		1898	+.3−.3	
5466	2.0		2004	+.2−.2	0.36
5897	2.1		2019	+.07−.09	
5904	1.3		2031	+.1−.1	.06
6101	1.3		2100	+.2−.2	
6121	2.0		2121	+.06−.07	
6171	2.3		2136	+.1−.1	
6205	2.6		2156	+.2−.2	-0.16
6218	1.3		2157	+.2−.2	
6254	2.0		2164	+.2−.2	
6341	1.7		2173	+.07−.09	
6352	1.3		2213	+.1−.12	
6397	1.9		2214	+.2−.2	
6535	2.4		2231	+.1−.13	

Table 1. continued.

ID	MW GCs Errors Gyr	$E(B-V)$
6584	1.4	Harris (1996)
6652	1.7	
6752	2.2	
6809	1.3	
6838	1.1	
7006	1.3	
7078	2.0	
7099	1.8	
7492	2.0	
Ter 7	0.8	
Ter 8	1.7	
Rup 106	0.9	
Pal5	1.6	
Pal12	1.7	
Ic4499	1.2	
Arp2	1.0	

3. Data set

Our analysis is based on three samples of GCs in Milky Way, M 31, and LMC which have appropriate photometric and structural parameter values.

Sample 1. This consists of 135 GCs taken from the catalogue of Harris (1996) which have nonzero values of all the parameters used for CA. The parameters used for PCA are distance from the galactic centre (R_{gc}), absolute visual magnitude (M_V), colour ($B-V$), concentration parameter (c), core radius (R_c), central surface brightness (μ_V), radial velocity (V_r), metallicity ([Fe/H]), and horizontal branch ratio (HBR). We are mainly concentrating on parameters which are intrinsic and independent in nature.

Sample 2. This consists of 35 GCs of the catalogue of Barmby et al. (2002). The parameters used are R_{gc}, M_V, $b(B-V)$, [Fe/H], c, R_c, V_r, and μ_V. We have converted the visual magnitudes from Barmby et al. (2000) to absolute visual magnitudes using a distance of 770 kpc (Sparke & Gallagher 2000) for M 31.

Sample 3. This consists of 23 GCs used in the paper by Mackey & Gilmore (2003). The parameters used are R_{gc}, M_V, $(B-V)$, c, R_c, V_r, [Fe/H], and μ_V. We have converted the visual magnitudes from Mackey & Gilmore (2003) to absolute visual magnitudes using a distance of 49 kpc (Sparke & Gallagher 2000) for LMC.

In Tables 3, 5, and 7 we have mentioned the elements of the correlation matrices corresponding to Samples 1, 2, and 3 respectively.

After finding the different groups by cluster analysis it is necessary to study the properties of the groups identified in terms of their ages along with other parameters. Again we have used $B-V$ as one of the parameters in PCA. As ages and $B-V$ values are subjected to measurement errors it is worthwhile to identify the nature of errors included in the data.

For Milky Way we considered 43 GCs and for LMC we considered 23 GCs. Their ages and corresponding errors can be obtained from Chaboyer et al. (1992) and Mackey & Gilmore (2003) respectively and are listed in Table 1. The means and standard deviations of these errors are listed in Table 2. The means and standard deviations (SD) of extinctions in colour $E(B-V)$ for Milky Way (MW) and LMC GCs are also listed in Table 2. It is interesting to note that most of the error and extinction distributions are Gaussian as indicated in Table 2 and Figs. 1−4 respectively. These show that the ages and colours of Milky Way and LMC GCs are consistent with respect to the measurement errors and extinctions. To study errors corresponding to ages for MW GCs we excluded some of the outliers and the fit is good which is evident from the Anderson Darling (AD) statistic. In Appendix B we have discussed Quantile Quantile Plot and Anderson Darling Statistic which have been used for fitting of Normal Distribution.

Table 2. Error analysis for the GCs of MW and LMC.

Name	Errors in Age		$E(B-V)$
	positive error		
LMC GCs	Mean	0.16	0.13
	SD	0.09	0.16
	AD	1.36	0.23
	Remark	Good fit	Very Good fit
	negative error		
	Mean	–0.15	
	SD	0.08	
	AD	1.27	
	Remark	Good fit	
MW GCs	Mean	1.59	1.95
	SD	0.44	0.62
	AD	0.45	0.11
	Remark	Good fit	Very good fit

Table 3. Correlation matrix for the parameters of Sample 1.

Parameter	R_{gc}	M_V	c	μ_V	$B-V$	[Fe/H]	R_c	HBR	V_r
R_{gc}	1								
M_V	0.22	1							
c	–0.32	–0.25	1						
μ_V	0.38	0.72	–0.52	1					
$B-V$	–0.29	–0.02	0.05	0.16	1				
[Fe/H]	–0.29	0.10	0.10	0.07	0.60	1			
R_c	0.04	0.11	–0.68	0.11	–0.16	–0.15	1		
HBR	–0.13	0.26	0.09	–0.26	–0.44	–0.77	0.11	1	
V_r	0.03	0.03	–0.05	0.09	–0.02	0.06	0.17	–0.03	1

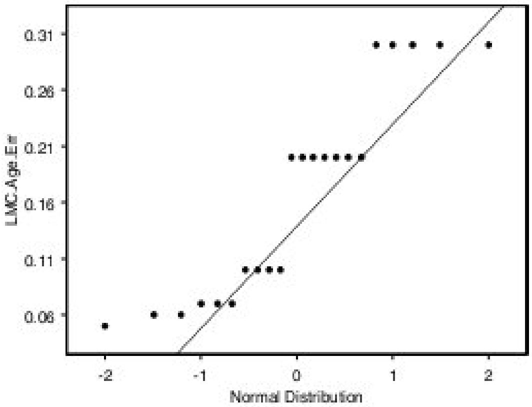

Fig. 1. QQ Normal plot for positive errors in ages of LMC GCs.

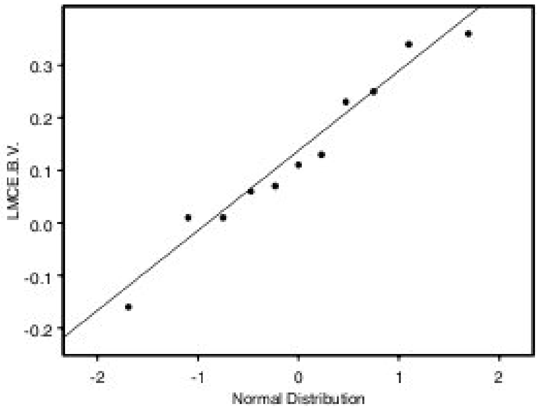

Fig. 2. QQ Normal plot for extinctions of LMC GCs.

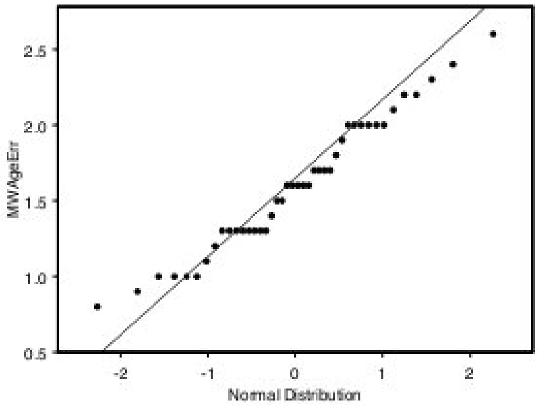

Fig. 3. QQ Normal plot for errors in ages of MW GCs.

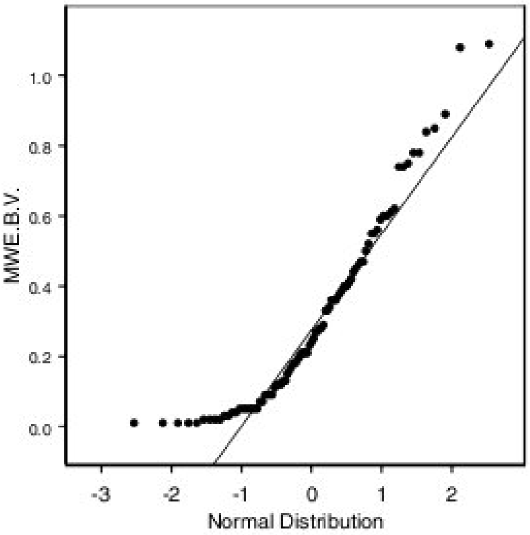

Fig. 4. QQ Normal plot for extinctions of MW GCs.

From the above findings it may be inferred that since the error distributions are Gaussian (symmetric), the errors are supposed to be averaged out in final analysis and results are not likely to be affected by them.

4. Results and discussions

4.1. Principal component analysis

We begin with a minimal number of parameters (selected by the trial and error method) and search for principal components giving the maximum percentage of total variation. In this respect we can say that we included many parameters like central surface brightness, colour and radial velocities, but they do not give maximum variation in PCA. Some of the parameter sets are given for comparison in Table 4 (e.g. S1(R_{gc}, M_V, R_c), S2(R_{gc}, M_V, c), S3($R_{gc}, B-V, V_r$), S4(R_{gc}, M_V, μ_V), S5(R_{gc}, M_V, c, R_c), S6($M_V, c,$ [Fe/H]), S7($c, R_c,$ [Fe/H]), S8($HBR,$ [Fe/H], c). Only the parameter set S7($c, R_c,$ [Fe/H]) has maximum variation (85.7 percent) as seen from last column of Table 4 with two principal components having eigen values greater than or nearly equal to 1. PCA analysis for Sample 2 and Sample 3 are listed in Tables 6 and 8 respectively. This also shows that the parameter sets S7 ([Fe/H], c, R_c) and S6 ([Fe/H], c, R_c) in Tables 6 and 8 respectively give maximum variation with two principal components with eigen values greater than or nearly equal to 1 (94.4 percent

Table 4. Results of PCA for Sample 1.

Set	Principal Component	Eigen value	Cumulative %
S1(R_{gc}, M_V, R_c)	1	1.6	53.1
	2	0.8	79.9
	3	0.6	100.0
S2(R_{gc}, M_V, c)	1	1.4	47.3
	2	1.1	82.9
	3	0.5	100.0
S3($R_{gc}, B-V, V_r$)	1	1.3	43.7
	2	0.9	75.3
	3	0.7	100.0
S4(R_{gc}, M_V, μ_V)	1	1.5	51.5
	2	0.9	81.9
	3	0.5	100.0
S5(R_{gc}, M_V, c, R_c)	1	2.0	50.7
	2	0.8	71.3
	3	0.7	88.9
	4	0.4	100.0
S6(M_V, c, [Fe/H])	1	1.6	53.0
	2	0.8	80.0
	3	0.6	100.0
S7(c, R_c, [Fe/H])	1	1.6	54.7
	2	0.9	85.7
	3	0.4	100.0
S8(HBR, [Fe/H], c)	1	1.5	49.7
	2	0.9	79.2
	3	0.6	100.0

Table 5. Correlation matrix for the parameters of Sample 2.

Parameter	R_{gc}	M_V	c	μ_V	$B-V$	[Fe/H]	R_c
R_{gc}	1						
M_V	–0.34	1					
c	–0.01	–0.058	1				
μ_V	–0.24	0.76	–0.33	1			
$B-V$	–0.25	0.16	0.39	–0.214	1		
[Fe/H]	–0.215	0.66	–0.26	0.48	0.68	1	
R_c	0.25	–0.01	–0.82	0.34	–0.46	0.09	1

Table 6. Results of PCA for Sample 2.

Set	Principal Component	Eigen value	Cumulative %
S1(R_{gc}, M_V, R_c)	1	1.4	47.7
	2	1.0	80.6
	3	0.6	100.00
S2(R_{gc}, M_V, c)	1	1.3	44.5
	2	1.0	78.2
	3	0.7	100.00
S3($R_{gc}, B-V, c$)	1	1.5	48.8
	2	1.0	81.8
	3	0.5	100.0
S4($B-V, c, R_c$)	1	2.1	71.0
	2	0.7	94.0
	3	0.2	100.0
S5(R_{gc}, M_V, c, R_c)	1	1.9	46.5
	2	1.3	79.9
	3	0.6	96.3
	4	0.2	100.0
S6(M_V, c, R_c)	1	1.8	60.5
	2	1.0	93.9
	3	0.2	100.0
S7([Fe/H], c, R_c)	1	1.9	62.7
	2	1.0	94.4
	3	0.2	100.0

Table 7. Correlation matrix for the parameters of Sample 3.

Parameter	R_{gc}	V_r	R_c	c	μ_V	[Fe/H]	M_V
R_{gc}	1						
V_r	0.32	1					
R_c	0.09	0.19	1				
c	–0.19	–0.03	–0.88	1			
μ_V	0.61	–0.13	0.13	–0.38	1		
[Fe/H]	0.46	0.44	0.16	–0.11	0.28	1	
M_V	0.52	–0.26	–0.03	–0.18	0.75	–0.12	1

and 96.1 percent in last columns). The group means are given in Tables 9 to 13.

It is found that the statistical dimensionality for the GCs in all the galaxies in the Local Group is two which involves the parameters core radius (R_c), concentration parameter (c) and metallicity. This is minimum but gives a total variation as high as 96 percent. This is maximum compared to all previous analyses (Kormendy & Djorgovski 1989; Djorgovski & de Carvalho 1990; Santiago & Djorgovski 1993; De Carvalho & Djorgovski 1992). This is the goal of PCA.

4.2. Cluster analysis

We found in the previous analysis that the maximum variation among the GCs in the Milky Way, M 31, and, LMC is due to the parameter set ([Fe/H], c, R_c). The metallicity values for GCs in M 31 are available for 35 clusters. So the sample size is slightly reduced for M 31 in our case compared to Covino & Fracassini (1993). The results for CA for Sample 1 are shown in Table 9 and Figs. 5 and 6 respectively. The jumps are at 4 and 6 respectively. To test the robustness of the classification we took several bootstrap samples generated from the original sample. In most of the situations it was found both from the jumps and from the ([Fe/H], R_c) plots that the proper number of groups (clusters) should be 3. We took this decision because of the following reasons.

1. From the plots of cluster vs jumps in most of the situations we observed that there was a maximum peak corresponding to 3 clusters (Fig. 7).
2. From the ([Fe/H], R_c) plots we also found if we choose the optimum number as 3 then the physical classification is quite clear whereas if we choose 4 or more clusters as optimum then the classification is rather messy. These features are presented in Figs. 13 and 8 respectively for the original Sample 1.
3. For our conclusion that the optimum number of clusters is 3 we have deviated slightly from the original algorithm (Sugar & James 2003) because it is a well known fact that cluster analysis is an exploratory data analytic technique and it depends heavily on proper physical explanation.

In support of the above discussion we have presented some of our findings related to bootstrap samples in Table 10 and in Figs. 8–12 respectively. In the tables we have mentioned the mean values for ([Fe/H], c, R_c) and the cluster points at which we have found peaks. Hence we have taken the optimum number of subgroups statistically and physically as three. Cluster 1, with high metallicities, low core radii, Cluster 2, with the lowest metallicity, low core radii, and Cluster 3 with still low metallicity, and high core radii. These are also reflected in the cluster means listed for these groups in the above tables. So we have carried out CA with $K = 3$ and the group means for all the

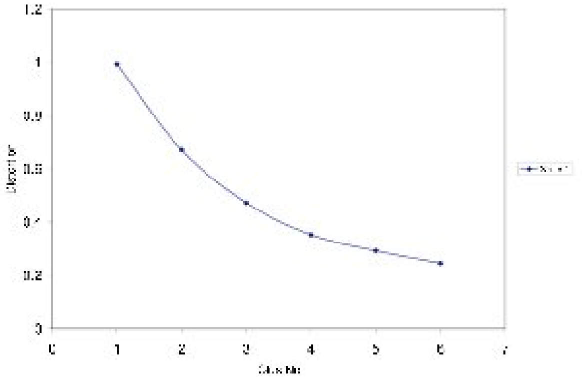

Fig. 5. *xy* diagram for the number of clusters (K) and corresponding distortions (d'_K) for Sample 1.

Table 8. Results of PCA for Sample 3.

Set	Principal Component	Eigen value	Cumulative %
S1(R_{gc}, M_V, R_c)	1	1.5	50.9
	2	1.0	84.6
	3	0.5	100.0
S2(R_{gc}, M_V, c)	1	1.6	54.4
	2	0.9	84.1
	3	0.5	100.0
S3(R_{gc}, M_V, c, R_c)	1	2.0	50.0
	2	1.4	85.7
	3	0.5	97.7
	4	0.1	100.0
S5($c, R_c, B-V$)	1	1.7	58.3
	2	1.0	91.9
	3	0.2	100.0
S6(c, R_c,[Fe/H])	1	1.9	64.1
	2	1.0	96.1
	3	0.1	100.0
S7(R_{gc}, M_V, μ_V)	1	2.3	75.5
	2	0.5	92.2
	3	0.2	100.0
S8(R_{gc}, [Fe/H], V_r)	1	1.8	60.5
	2	0.7	83.3
	3	0.5	100.0

Table 9. The group means for the parameters of the GCs of Sample 1.

No of clusters at the peaks	4, 6			
Variables	Cluster 1	Cluster 2	Cluster 3	Cluster 4
No. of members	37	41	45	12
⟨[Fe/H]⟩	–0.59	–1.54	–1.69	–1.29
⟨c⟩	1.66	2.11	1.21	0.92
⟨R_c⟩(pc)	2.66	0.59	3.86	17.53

parameters are shown in Table 11. The ([Fe/H], R_c),([Fe/H], c), and ([Fe/H], c, R_c) plots are shown in Figs. 13–15 respectively. The clustering is very prominent in Fig. 15 implying the fact that three dimensional parametric classification is more authentic than the two dimensional one as the classification is not clearly seen e.g. in ([Fe/H], c) plane.

Now the kinematics of MW GCs are studied to examine the physical consistency of the classification following Zinn (1985). The basic assumption is that the rotational velocity of each sub group in the classification is constant. For this it is necessary to know the distances of GCs from the galactic centre and the results depend slightly on the values adopted for $R_\odot$ and the velocity of the LSR about the galactic centre ($v_\odot$) which are taken as 8.2 kpc and 220 km s^{-1} for the present situation. The analysis follows the method of Frenk & White (1980). The values of rotational velocities (v_{rot}) and velocity dispersion (σ_{los}) are listed in Table 11 for each group. Also the group means for heights from the galactic plane ($|z|$), distances from the galactic centre (R_{gc}), metallicities, concentration parameters (c), core radii, and central surface brightness (μ_V) are listed for these sub groups. It is found that for [Fe/H] > –0.8 the cluster group has substantial rotation for R_{gc} < 4.4 kpc (Cluster 1) and for [Fe/H] < –0.8 and GCs have less rotation for R_{gc} > 4.4 kpc. This supports the analysis by Zinn (1985). The innermost group (Cluster 1) has substantial rotational velocity (~124 km s^{-1}), highly metal rich (~–0.64) having smaller core radii (~2.71 pc). They are concentrated near the galactic disc ($|z|$ ~ 1.73 kpc) and close to the Galactic centre (R_{gc} ~ 4.19 kpc). The velocity dispersion is comparatively smaller. But outside the inner region there are two groups instead of one as found by Zinn (1985). One group (Cluster 2) has very low metallicity, far from the Galactic disc, comparatively low rotational velocity (~5 km s^{-1}) with moderate core radii. The other group (Cluster 3) has low metallicity, high core radii, small rotation (~20 km s^{-1}), highest velocity dispersion (σ_{los} ~ 131 km s^{-1}), farthest from the Galactic centre (R_{gc} ~ 31.69 kpc) and concentrated farthest from the Galactic disc ($|z|$ ~ 18.81 kpc). So they may be associated with GCs of the outer halo. The ages of the MW GCs used are from Chaboyer (1992) and the ages vs metallicities for Clusters 1, 2, and 3 are

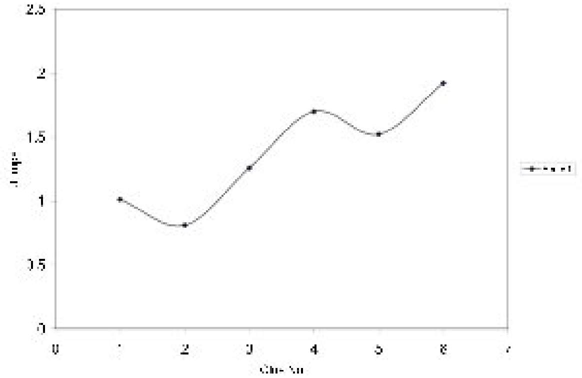

Fig. 6. *xy* diagram for the number of clusters (K) and jumps for Sample 1.

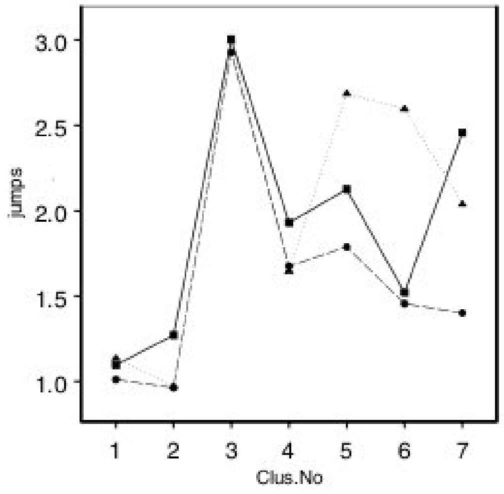

Fig. 7. *xy* diagram for the number of clusters (K) and jumps for Bootstrap Samples, dash for second, solid line for fourth and dotted line for third Bootstrap Sample respectively.

Table 10. The group means for the parameters of the GCs of Bootstrap samples.

Bootstrap samples	No of clusters at peaks	Variables	Cluster 1	Cluster 2	Cluster 3
		No. of members	36	68	31
1	3, 5	$\langle$[Fe/H]$\rangle$	–0.59	–1.46	–1.80
		$\langle c \rangle$	1.42	2.00	0.93
		$\langle R_c \rangle$(pc)	2.84	0.92	12.66
		No. of members	41	66	28
2	3	$\langle$[Fe/H]$\rangle$	–0.73	–1.61	–1.67
		$\langle c \rangle$	1.50	1.86	0.82
		$\langle R_c \rangle$(pc)	3.06	1.26	13.36
		No. of members	43	68	24
3	3,5	$\langle$[Fe/H]$\rangle$	–0.57	–1.58	–1.60
		$\langle c \rangle$	1.70	1.8	0.87
		$\langle R_c \rangle$(pc)	2.84	1.32	10.45
		No. of members	56	65	14
4	3,7	$\langle$[Fe/H]$\rangle$	–0.89	–1.65	–1.01
		$\langle c \rangle$	1.97	1.44	1.24
		$\langle R_c \rangle$(pc)	1.17	2.81	17.49

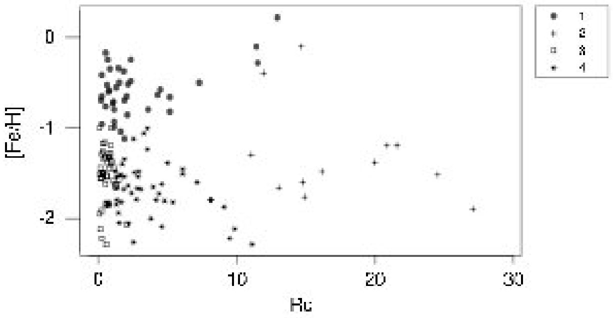

Fig. 8. [Fe/H] vs. R_c (pc) diagram for Sample 1. The suffixes indicate the cluster number.

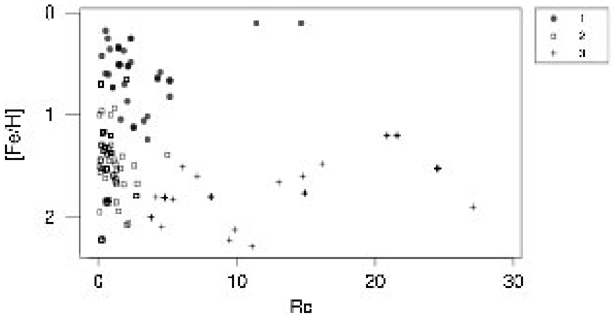

Fig. 9. [Fe/H] vs. R_c (pc) diagram for first Bootstrap Sample 1. The suffixes indicate the cluster number.

shown in Fig. 16. The mean ages for these groups are also listed in Table 11. The GCs of the outer halo are younger compared to those of inner halo and their age- metallicity scatter plot shows no correlation. The GCs in the inner halo (Cluster 2) are the oldest population ($\sim 10^{14.83}$ yr) and the age- metallicity diagram shows a correlation with considerable scatter. The ages of all the GCs in each cluster are not known. The GCs whose ages are available from Chaboyer (1992) are used. So these diagrams suffer from complicity and firm conclusion. Also the rotational velocities calculated for these groups (Clusters 2 and 3) show that GCs in the inner and outer halo have substantially smaller rotation and higher velocity dispersion. All these facts are consistent with the work carried out by Zinn (1993) and subsequently by many authors (van den Bergh 1993; Lynden-Bell & Lynden-Bell 1995; Silk & Wyse 1993). The present analysis differs from the former work in the sense that the classification is done in an objective and more scientific way on the basis of multiple parameters at a time instead of taking color or metallicity or horizontal branch ratio parameter one at a time and making revisions every time e.g. in the works of Zinn (1985) and Zinn (1993) where there are two and three sub populations respectively. So the present analysis selects the optimum, unique group and helps to enunciate unique theory for galaxy formation. Also the choice of the optimum set of parameters for CA was done in an objective way through PCA. So from the present classification it can be concluded that initially there was a halo in which metal poor GCs formed (inner halo). Then the medium got enriched from the evolving stars and a disc of GCs formed which are comparatively metal rich. The GCs of the outer halo might have been accreted from the neighbouring galaxies through tidal accretion. This is concluded from the kinematic properties, absence of age-metallicity correlation and metallicity gradient etc in Cluster 3. This is also suggested by Mackey & Gilmore (2004) on the basis of HB morphology of the Galactic halo GCs and those in the neighbouring dwarf galaxies.

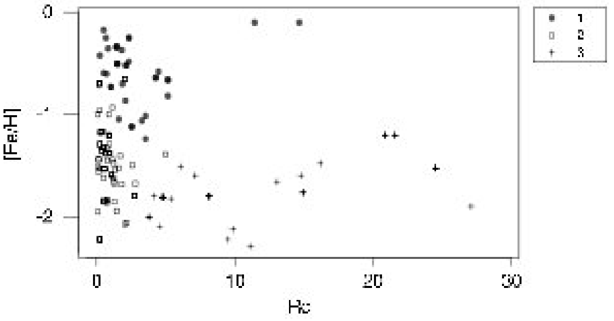

Fig. 10. [Fe/H] vs. R_c (pc) diagram for second Bootstrap Sample 1. The suffixes indicate the cluster number.

The CA for M 31 GCs are shown in Table 12 and Fig. 17 respectively. The optimum number in this case is also three like those in MW. Cluster 1 has high metallicity, very low core radii and close to the galactic centre. Cluster 2 has minimum metallicity, comparatively lower core radii, and is very far from the centre. Cluster 3 has comparatively low metallicity, maximum core

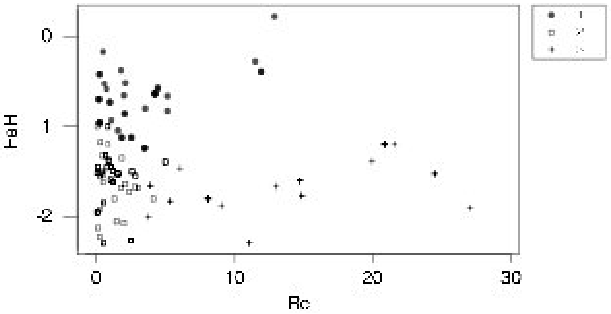

Fig. 11. [Fe/H] vs. R_c (pc) diagram for third Bootstrap Sample 1. The suffixes indicate the cluster number.

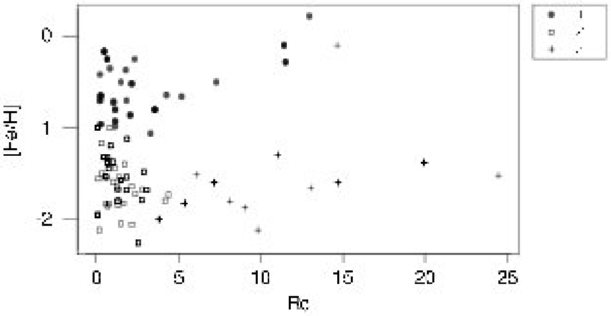

Fig. 12. [Fe/H] vs. R_c (pc) diagram for fourth Bootstrap Sample 1. The suffixes indicate the cluster number.

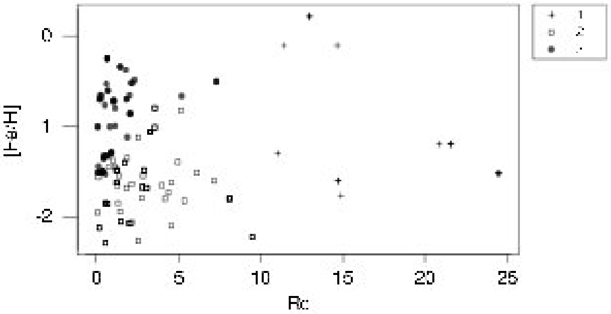

Fig. 13. [Fe/H] vs. R_c (pc) diagram for final cluster analysis result for Sample 1. The suffixes indicate the cluster number.

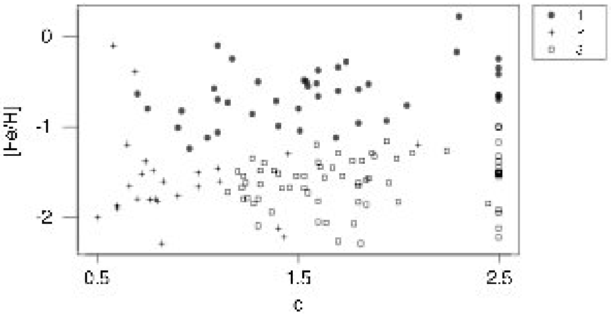

Fig. 14. [Fe/H] vs. c diagram for final cluster analysis result for Sample 1. The suffixes indicate the cluster number.

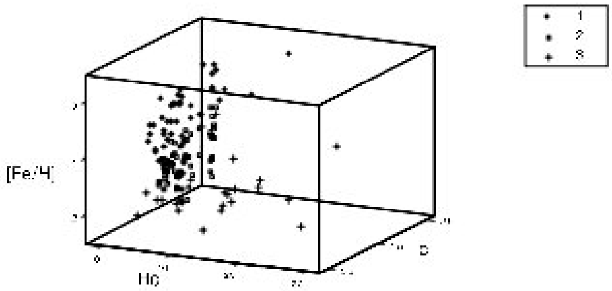

Fig. 15. [Fe/H], R_c (pc), c diagram for final cluster analysis result for Sample 1. The suffixes indicate the cluster number

Table 11. The group means for the parameters of the GCs of Sample 1 in the final analysis.

Variables	Cluster 1	Cluster 2	Cluster 3
No. of members	41	69	25
$\langle$[Fe/H]$\rangle$	–0.64	–1.61	–1.58
$\langle c\rangle$	1.60	1.82	0.91
$\langle R_c\rangle$(pc)	2.71	10.16	12.12
$\langle\lvert z\rvert\rangle$(kpc)	1.73	5.49	18.81
$\langle R_{gc}\rangle$(kpc)	4.196	10.16	31.69
$\langle\mu_V\rangle$)	18.792	17.28	22.32
$\langle v_{rot}\rangle$ (km s^{-1})	124	5	20
σ_{los}	88	129	131
$\langle$log (Age)$\rangle$(yr)	11.81	14.83	14.66

Table 12. The group means for the parameters of the GCs of M 31.

Variables	Cluster 1	Cluster 2	Cluster 3
No. of members	9	19	7
$\langle$[Fe/H]$\rangle$	–0.43	–1.51	–1.41
$\langle B-V\rangle$	1.06	0.76	0.73
$\langle c\rangle$	1.67	1.55	1.08
$\langle R_c\rangle$(pc)	0.545	0.92	1.93
$\langle R_{gc}\rangle$(kpc)	5.28	8.78	8.17
$\langle\mu_V\rangle$	16.28	16.01	16.56
$\langle v_{rot}\rangle$ (km s^{-1})	52.87	31.16	51.90

radii. For calculating the rotational velocities of these groups the function

$$v = v_{sys} + v_{rot}\sin(\phi)$$

(Perett et al. 2002) is fitted to the radial velocities of the GCs of M 31 where ϕ is the position angle taken from Barmby et al. (2000), v_{sys} is the mean velocity of the M 31 cluster system (de Vaucouleurs et al. 1991) taken as -300 ± 4 km s^{-1}. They are shown in Table 12. It is found that Cluster 1 and Cluster 3 have comparable rotational velocities while Cluster 2 has somewhat lower rotational velocity. Also Cluster 1 is the most metal rich component of the system and closest to the galactic centre. So it can be associated with the disc part like MW Cluster 1. On the other hand Cluster 2 is the most metal poor component and farthest from the galactic centre. So it can be associated with the outer halo like MW Cluster 3 and Cluster 3 of M 31 can be associated with the inner halo like MW Cluster 2. Also from the ([Fe/H], R_c) diagram (Fig. 18) it is clear that the groups are very similar to those of MW GCs (Fig. 13). So it can be concluded that formation history of MW and M 31 are more or less similar.

The CA for LMC GCs shows that there are aparantly three groups (Table 13).The third group containing only 3 GCs has more or less similar characteristics as first group (Cluster 1). The GCs in these two groups are almost coeval ($\sim 10^8$ yr). Also the mean metallicities (~ -0.37 and ~ -0.34) of these groups and mean distances (~ 3 kpc) from the galactic centre are similar. Only the core radii differ. Almost similar features have been found for several other bootstrap samples. As a result we may consider them as a single group. The number of GCs in the second group (Cluster 2) is also very small (4) but this may be due to lack of data points. Since metallicity values are available only for

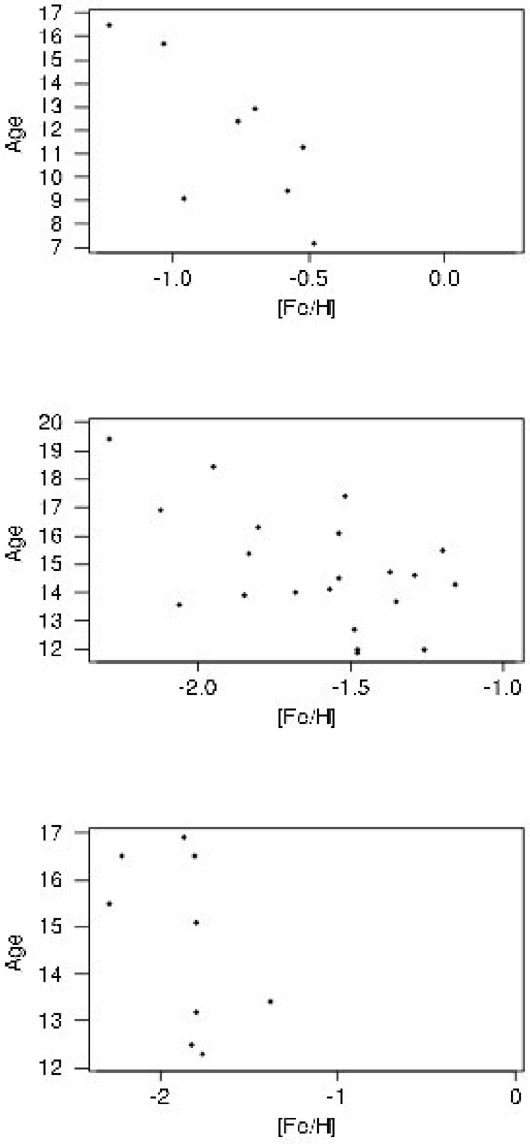

Fig. 16. Scatter diagram of age in logarithmic scale (yr) vs. metallicity with available values of ages in clusters 1, 2 and 3 respectively as a result of cluster analysis for Sample 1.

Table 13. The group means for the parameters of the GCs of LMC.

Variables	Cluster 1	Cluster 2	Cluster 3
No. of members	16	4	3
$\langle$[Fe/H]$\rangle$	–0.37	–1.6	–0.34
$\langle B-V\rangle$	0.33	0.55	0.18
$\langle\mu_V\rangle$	18.33	17.40	18.49
$\langle c\rangle$	1.39	1.26	0.68
$\langle R_c\rangle$(pc)	2.62	3.72	10.87
$\langle R_{gc}$ (kpc)	3.18	1.5	2.85
$\langle$log (Age)$\rangle$ (yr)	8.09	10.20	7.77

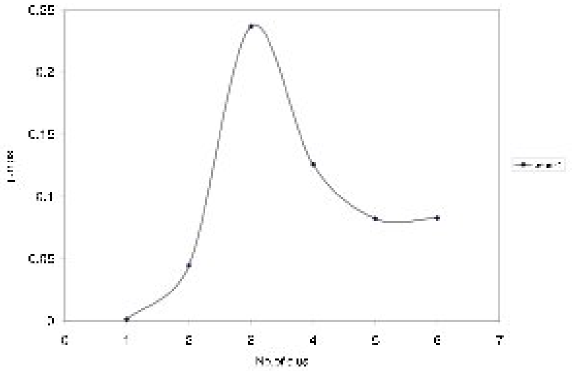

Fig. 17. *xy* diagram for the number of clusters (*K*) and jumps for Sample 2.

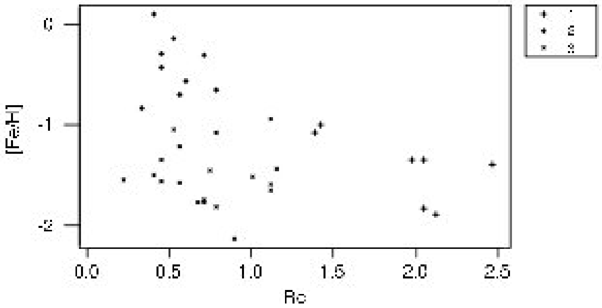

Fig. 18. [Fe/H], R_c (pc) diagram in the cluster analysis for Sample 2. The suffixes indicate the cluster number.

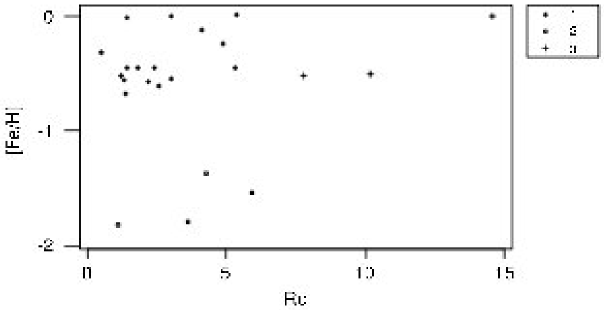

Fig. 19. [Fe/H], R_c(pc) diagram in the cluster analysis for Sample 3. The suffixes indicate the cluster number.

23 GCs of LMC (Mackey & Gilmore 2003) the sample size has been reduced from that one (39) used by Covino & Fracassini (1993). Now it is a well known fact that metallicity has good correlation with colour ($B - V$). So if the CA is carried out with (c, R_c and $B - V$) with the Covino & Fracassini (1993) sample as ($B - V$) is available for 39 GCs in that sample, then the classification (Fig. 20) shows a good concentration in Cluster 2. On the basis of the above discussion it may be inferred that the actual number of clusters in LMC is likely to be two instead of three as in the cases of MW and M 31. But this feature is not directly reflected through CA due to small sample size. The group means are shown in Table 13. The metallicity vs core radii diagram is also shown in Fig. 19. The ages of the LMC GCs used are from Mackey & Gilmore (2003). The mean ages of the groups are also listed in the table. It is seen that outer GCs (Cluster 1) are younger and more metal rich than those of inner GCs (Cluster 2) which exhibits reverse nature as compared with that of MW and M 31 GCs. Also the outer GCs of LMC are much younger (of the order of Myr) than those of MW GCs. These indicate that the formation history of LMC may be different from that of MW or M 31. This will be clear if a spectroscopic study of the GCs can be carried out and the sample size is increased.

5. Conclusions

We have applied multivariate analysis for the reclassification of the globular clusters of our Galaxy, M 31, and LMC. First

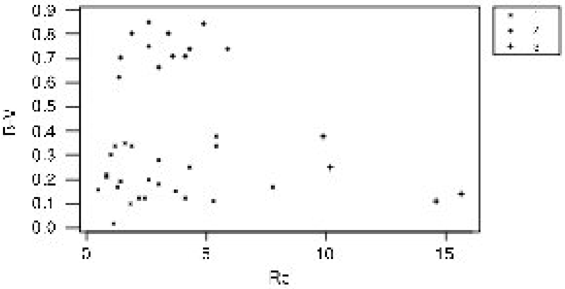

Fig. 20. $B-V$, R_c (pc) diagram in the cluster analysis for sample of LMC GCs of Covino & Fracassini (1993). The suffixes indicate the cluster number.

a Principal Component Analysis is performed to search for the optimum set of parameters giving the maximum over all variation among the GCs in these galaxies. It is found that metallicity, concentration parameter and core radius are the parameters responsible for maximum variation in the GCs of Milky Way and M 31. The statistical dimensionality is two in every situation which is less than those of elliptical galaxies (Santiago & Djorgovski 1992). This procedure is completely different from the method of studying two point correlation or studying the properties of GCs with respect to a single parameter like colour or HB morphology as done by previous authors (Zinn 1985, 1993; Mackey & Gilmore 2004). As the present set up is multivariate, it is quite likely to carry out the analysis by multivariate methods. Also there are various parameters responsible for the variation among GCs. It is better to select the optimum set giving maximum variation. This reduces the difficulty in handling large number of parameters simultaneously and drawing any physical conclusions while at the same time restores the significant parameters responsible for maximum variation. This is the goal of PCA. In the present situation the optimum set does not include HB morphology parameter which is HBR but includes chemical composition and morphological parameters instead (Table 4, S7, and S8). So it is more scientific to perform classification on the basis of these significant parameters selected objectively instead of taking any parameter in a subjective way. Then Cluster Analysis is carried out with respect to this optimum set. Here two and three clusters are found in case of MW, M 31, and LMC GCs. The robustness of the classification is tested by taking a few bootstrap samples generated from the original one. The classification differs from that by Zinn (1985) and is in agreement with Zinn (1993) and Mackey & Gilmore (2004). For MW and M 31 three clusters, disc, inner halo, and outer halo GCs are found. The kinematic properties, age metallicity diagram, metallicity gradients studied for these groups also support the true nature of the classification. The analogous behaviour of the GCs with MW GCs in different parametric planes shows that they have a similar nature as those of MW GCs. Perrett et al. (2002) have calculated the kinematic properties of some 200 GCs in M 31 and have found two groups, disc and halo GCs with metallicities peaked at −0.5 and −1.41 respectively. In the present analysis, three groups have been found with mean metallicities −0.43, −1.41, and −1.51 respectively. The existence of the third group with minimum metallicity is analogous in properties with inner halo GCs of MW. For LMC GCs two groups have been found instead of three though the conclusion is not very firm due to the small size of the sample. The evolution history is likely to differ from those of MW and M 31. As there is controversy in the formation of GCs in disc (Schommer et al. 1992) or in pressure supported halo (van den Bergh 2004) so more clear picture will come out from the spectroscopic study of GCs with a larger sample size.

Acknowledgements. One of the the authors (Tanuka Chattopadhyay) wishes to thank University Grants Commission (UGC), India for awarding her a major research project for the work.The authors are grateful to Prof. Rajaram Nityananda for his useful suggestions. The authors are also thankful to Prof. Duncan Forbes for his careful reading and important suggestions. The authors are grateful to the referee for his careful reading and valuable suggestions to improve the quality of the work.

References

Ashman, K. M., & Zepf, S. E. 1992, ApJ, 384, 50
Barmby, P., Holland, S., & Huchra, J. P. 2002, AJ, 123, 1937
Brosche, P. 1973, A&A, 23, 259
Brosche, P., & Lentes, F. T. 1984, A&A, 139, 474
Brodie, J. P., & Huchra, J. 1991, ApJ, 379, 157
Capaccioli, M., Ortolani, S., & Piotto, G. 1991, A&A, 244, 298
Chaboyer, B., Sarajedini, A., & Demarque, P. 1992, ApJ, 394, 515
Chattopadhyay, T., & Chattopadhyay, A. 2006, AJ, 131, 2452
Covino, S., & Pasinetti Fracassini, L. E. 1993, A&A, 270, 83
De Carvalho, R. R., & Djorgovski, S. 1992, ApJ, 389, L49
deVaucouleurs, G., deVaucouleurs, A., Corwin, H. G. Jr., et al. 1991, Third Reference Catalogue of Bright Galaxies, version 3.9 (New York: Springer)
Djorgovski, S. G. 1991, ASPC, 13, 112
Djorgovski, S. G. 1995, ApJ, 438, L29
Djorgovski, S. G., & de Carvalho, R. 1990, in Windows on Galaxies, ed. G. Fabiano, et al. (Dordrecht: Kluwer), 9
Eigenson, A. M., & Yatsyk, O. S. 1989, SvA, 33, 280
Forbes, D. A., Brodie, J. P., & Huchra, J. 1996, AJ, 112, 2448
Forbes, D. A., Brodie, J. P., & Grillmair, C. J. 1997, AJ, 113, 1652
Frenk, C. S., & White, S. D. 1980, MNRAS, 193, 295
Harris, W. E. 1991, A&A, 29, 543
Harris, W. E. 1996, AJ, 112, 1487
Hartigan, J. A. 1975, Clustering Algorithms (New York: Wiley)
Kormendy, J., & Djorgovski, S. 1989, A&A, 27, 235
Katz, A. 1992, ApJ, 391, 502
Lynden-Bell, D., & Lynden-Bell, R. M. 1995, MNRAS, 275, 429
Mackey, A. D., & Gilmore, G. F. 2003, MNRAS, 338, 85
Mackey, A. D., & Gilmore, G. F. 2004, MNRAS, 355, 504
MacQueen, J. 1967, Fifth Berkeley Symp. Math. Statist. Prob., 1, 281
Perett, K. M., Bridges, T. J., Hanes, D. A., Brodie, J. P., & Carter, D. 2002, AJ, 123, 2490
Peterson, C. J., & King, I. R. 1975, AJ, 80, 427
Santiago, B. X., & Djorgovski, S. 1993, MNRAS, 261, 753
Schommer, R. A., Olszewski, E. W., Suntzeff, N. B., & Harris, H. C. 1992, AJ, 103, 447
Schweizer, F., Miller, B. W., Whitmore, B. C., & Fall, S. M. 1996, AJ, 112, 1839
Searle, L., & Zinn, R. 1978, ApJ, 225, 357
Silk, J., & Wyse, R. F. G. 1993, Phys. Rep., 231, 293
Sparke, L. S., & Gallaher, J. S. 2000, in Galaxies in the Universe (Cambridge University Press), 134
Sugar, A. S., & James, G. M. 2003, JASA, 98, 750
van den Bergh, S. 1975, A&A, 13, 217
van den Bergh, S. 1993, AJ, 105, 971
van den Bergh, S. 2004, AJ, 127, 897
Whitmore, B. C., Schweizer, F., Leitherer, C., Borne, K., & Robert, C. 1993, AJ, 106, 1354
Zepf, S. E., & Ashman, K. M. 1993, MNRAS, 264, 611
Zinn, R. 1985, ApJ, 293, 424
Zinn, R. 1993, in the Globular Cluster-Galaxy Connection, ed. G. H. Smith, & J. P. Brodie (San Fransisco: ASP), 38

Appendix A: Appendix on principal component analysis

The purpose of the principal component analysis is to reduce the complexity of multivariate data by transforming the data into the principal components space and choosing the first n principal components that explain most of the variation in the original variables.

Let the components x_i of a random vector $x = (x_1, x_2, \ldots, x_p)'$ are measured in the same or comparable units. Also assume that the variances of the variables do not vary too much. Let Σ denote the covariance matrix of the random vector x. Σ is assumed to be at least positive semi definite rank $r(\leq p)$. All the eigen values of Σ are real and non negative. Let $\lambda_1 \geq \lambda_2 \ldots \geq \lambda_p \geq 0$ be the ordered eigen value of Σ. Then there exists an orthogonal matrix $G^{p\times p} = (g_1^{p\times 1}\ g_2^{p\times 1} \ldots g_p^{p\times 1})$, $GG' = I_p$ such that $G'\Sigma\, G = D_\lambda$, $D_\lambda = \text{Diag}(\lambda_1, \lambda_2, \ldots, \lambda_p)$

Consider the transformation

$$y = G'x \tag{A.1}$$

Then

$$\text{Cov}\,(y) = D_\lambda. \tag{A.2}$$

The vectors $g_1, g_2, \ldots, g_p$ of the matrix G are called eigen vectors corresponding to the eigen values $\lambda_1, \ldots, \lambda_p$. It can be shown that

$$\begin{aligned}\lambda_1 &= g_1'\Sigma g_1 = \text{DM}\max\nolimits_{a:a'a=1} a'\Sigma a \\ \lambda_2 &= g_2'\Sigma g_2 = \text{DM}\max\nolimits_{b:b'b=1,b'g_1=0} b'\Sigma b \\ &\vdots\end{aligned}$$

and so on.

From the above result it follows that for any set of orthogonal vectors $h_i, h_i'h_i = 1$ and for any $k \leq p$

$$\lambda_1 + \lambda_2 + \ldots + \lambda_k = \text{DM}\Sigma_{i=1}^k g_i'\Sigma g_i \geq \text{DM}\Sigma_{i=1}^k h_i'\Sigma h_i$$

where by DM we denote a diagonal matrix.

For this reason, the component

$$y_1 = g_1'x \tag{A.3}$$

is called the first principal component.

$$y_2 = g_2'x \tag{A.4}$$

is called the second principal component and so on.

Here $\text{Var}(y_i) = \lambda_i$, Because $\lambda_1 \geq \lambda_2 \ldots \lambda_p$, y_1 has the largest variance λ_1, y_2 has the second largest variance λ_2 and so on. Since

$$\lambda_1 + \lambda_2 + \ldots + \lambda_p = tr\Sigma = \text{DM}\Sigma_{i=1}^p \sigma_{ii} \tag{A.5}$$

the sum of the variances of p principal components is the same as the sum of the variances of the original variables $x_i \ldots x_p$. Thus the components with smaller variances could be ignored without significantly affecting the total variance and thereby reducing the number of variables from p to say $k \leq p$.

Many criteria have been suggested by different authors for deciding how many principal components to retain. Some of these criteria are as follows :

1. Include just enough components to explain some arbitrary amount (say 90%) of the variance.
2. Exclude those principal components with eigen values below the average. For principal components calculated from the correlation matrix, this criteria excludes components with eigen values less than 1. This criterion has been used in the present paper.
3. Use of the screeplot technique.

Appendix B: Appendix on q-q plot and anderson darling test

Quantile-Quantile Plot

The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution. A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. By a quantile, we mean the fraction (or percent) of points below the given value. That is, the 0.3 (or 30%) quantile is the point at which 30% of the data fall below and 70% fall above that value.

A 45-degree reference line is also plotted. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line. The greater the departure from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions.

In order to fit a theoretical distribution to a data set we take the data set as the first sample and observations from the theoretical distribution under consideration (in our case Normal) as the second sample.

Anderson-Darling test

The Anderson-Darling test (Stephens 1974) is used to test if a sample of data came from a population with a specific distribution. It is a modification of the Kolmogorov-Smirnov (K-S) test and gives more weight to the tails than does the K-S test. The K-S test is distribution free in the sense that the critical values do not depend on the specific distribution being tested. The Anderson-Darling test makes use of the specific distribution in calculating critical values. This has the advantage of allowing a more sensitive test and the disadvantage that critical values must be calculated for each distribution.

The Anderson-Darling test is an alternative to the chi-square and Kolmogorov-Smirnov goodness-of-fit tests.

Definition: The Anderson-Darling test is defined as:
H0: The data follow a specified distribution.
Ha: The data do not follow the specified distribution.
Test Statistic: The Anderson-Darling test statistic is defined as

$$A^2 = -N - S$$

where

$$S = \Sigma_{i=1}^N ((2i-1)/N)[\ln(F(Y_i) + \ln(1 - F(Y_{N+1-i}))]$$

F is the cumulative distribution function of the specified distribution. Note that the Y_i are the ordered data.

Significance Level: Critical Region: The critical values for the Anderson-Darling test are dependent on the specific distribution that is being tested. Tabulated values and formulas have been published (Stephens 1974, 1976, 1977, 1979) for a few specific distributions (normal, lognormal, exponential, Weibull, logistic, extreme value type 1). The test is a one-sided test and the hypothesis that the distribution is of a specific form is rejected if the test statistic, A, is greater than the critical value.