Chapter 1
Introduction

Statistical tools and techniques play a major role in many areas of science like biology, economics, physics etc. In particular, a systematic development in statistical methodology for biological problems started long back and constituted a new area of bio-statistics. Compared to biology, the rate of progress in the development of statistical techniques for physical science is very slow. In areas like Astronomy and Astrophysics the application of sophisticated statistical analysis is a comparatively recent phenomenon. Astronomy & Astrophysics is the science to study the different features of planets, stars, galaxies and the Universe as a whole. Astrophysicists try to model observed astronomical properties by using laws governing physical processes. The problem is to make inference for the underlying properties on the basis of observations related to a few external characteristics. During the past two decades the inter-disciplinary field of astrostatistics has newly emerged in order to study important astrophysical issues through appropriate statistical analysis.

Astronomy in recent past has developed a lot with the launch of several missions like GALEX (Galaxy Evolution Explorer), Kepler Space Telescope, Hubble Space Telescope (HST) etc. through which terabytes of data are available for preservation. This increasing proportionality of huge data demands data access efficiency. The implication of the above statement is important in a sense that most of the astrophysical phenomena are being observed in terms of light intensity as a function of wavelength or frequency which are snapshots of experiments which cannot be repeated as such. So one can easily
understand why every single observation needs to be preserved.

A group of stars within the galaxy that resemble each other in spatial distribution, chemical composition or age are called stellar population. Stellar populations are not discrete in their properties, but rather have a continuum of characteristics that reflect the changes in star formation with time. Walter Baade introduced in 1944 the concept of two stellar populations. There are two stellar population types. The stars in the globular clusters with those similar to them and stars in the solar neighborhood including the stars in Galactic clusters. This classification was mainly based on physical properties and space distribution of objects, but the difference in their kinematical behavior is also important. There are systematic variations in stellar properties also. Before understanding these differences it is necessary to know something about the spherical clusters of stars called globular clusters (GCs). Globular clusters are compact, dense clusters of very old stars, typically containing $10^4$ to $10^6$ members in a spherical region of space less than about 50 parsec in diameter. Advances in observational technology and astrophysical theory have made it possible to understand that stellar populations can now be defined in terms of the age, metal content and location of the stars.

Individual globular clusters and globular cluster systems are of interest in their own right, but they also provide unique insights into a wide range of astrophysical processes and systems. As individual objects, they constitute isolated laboratories in which many aspects of stellar evolution and dynamics can be studied. The study of globular clusters has been and still is essential for furthering our knowledge of such astrophysical phenomena as stellar and galactic evolution, variable and X-ray emission stars, chemical abundances (primordial nucleosynthesis), etc. When considered as a system, the globular clusters in a galaxy provide a fossil record of the dynamical and chemical conditions when the galaxy was in the process of formation. Thus observations of globular cluster systems can be used to constrain models of the formation and evolution of galax-
ies. Globular clusters are ideal laboratories for testing theories of stellar evolution, the chemical evolution of the Universe and the dynamics of N-body systems.

Furthermore, globular clusters are some of the oldest stellar systems known and thus estimates of their ages can be used to constrain the age of the Universe as a whole. The distribution of their ages and correlations between cluster ages and metal abundances make these systems valuable probe into the processes of galaxy formation.

Study of horizontal branches (HB) of globular clusters in color-magnitude diagram is very important in understanding the evolution of various stellar aggregates in them. Horizontal branch stars are characterized by helium burning core and hydrogen burning shell which have evolved off the red giant branch (RGB) having almost constant effective luminosity (hence the name 'horizontal') but widely varying color range.

There are various theories regarding the formation of GCs, e.g., (1) gaseous merger, (2) in situ GC formation and (3) tidal stripping. Merger theory says that GCs in giant elliptical (gE) and cD galaxies were created by the gaseous merger of two progenitor spiral galaxies (Ashman & Zepf 1992 [ME92]; Zepf & Ashman 1993 [EM93]). The protoglobal clusters in the currently merging galaxies were discovered by Whitmore et al. (1993) [FC93]; Schweizer et al. (1996) [FC96]. According to the most favorable theory (Forbes et al. 1997 [APJ97]) GCs are considered to be formed in situ star formation episodes during the collapse process of a large gaseous ensemble. There are three phases in the entire process. In the first phase a chaotic merging of many small gaseous subunits (Searle & Zinn 1978 [LR78]; Katz 1992 [N.92]) 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 [APJ97]). The metallicity of the GCs formed in this phase is low, i.e., they are metal-poor. In the second phase, the stars enrich the medium and now a large number of field stars are formed (due to more efficient cooling at high metallicity) and the metallicity of the GCs...
formed in this phase is high, i.e., they are metal-rich. In the third phase the remaining gas settles as a galactic disk at the center of the galaxy. Formation of spiral galaxies is the extreme case of this process. For elliptical galaxies, the GCs which form in the second phase are metal-rich than those of the spiral galaxies formed in the second phase. The final theory is the theory of acquiring globular clusters by a larger galaxy due to its tidal force from satellite dwarf galaxies. This is known as tidal stripping.

Globular clusters (GCs) are touchstones of astrophysics. Their study addresses many important issues ranging from stellar evolution to the formation of galaxies and cosmology.

Classical formation of galaxies can be divided into five major categories: (1) the monolithic collapse model, (2) the major merger model, (3) the multiphase dissipational collapse model, (4) the dissipationless merger model and (5) accretion and in situ hierarchical merging.

According to the monolithic collapse model, an elliptical galaxy is formed through the collapse of an isolated massive gas cloud at high redshift (Larson 1975 [B.75]; Carlberg 1984 [G.84]; Arimoto & Yoshii 1987 [NY87]). In the major merger model, elliptical galaxies are formed by the merger of two or more disk galaxies (Toomre 1977 [A.77]; Ashman & Zepf 1992 [ME92]; Zepf et al. 2000 [EAJ+00]).

The multiphase dissipational collapse has been proposed by Forbes et al. (1997) [APJ97]. According to this model, the GCs form in distinct star formation episodes through dissipational collapse.

It should be mentioned that kinematic studies of GCs in several giant elliptical galaxies show results which differ from galaxy to galaxy (Woodley et al. 2007 [AFA+07]).

In order to investigate the above mentioned astrophysical situations through statistical data analysis, one has to take into account the following points. The first requirement
is representative data set which contains relevant information related to parameters (variables) responsible for the actual variation. In many situations such parameters may not be directly available and may have to be derived through different physical relations. Secondly, it is necessary to know about the nature of data type of the different variables (e.g. continuous, categorical, binary etc.) from the physical properties as the final analysis will depend upon the nature of data. Thirdly, there should not be a large percentage of mixing values. One may take care of mixing values through different techniques but the number of mixing observations should be small. As the observations are collected through different types of telescopes and the distances of the objects are in astronomical units, the chance of mixing observations is very high. Fourthly, since it is difficult to collect information accurately on many variables, it is very likely that there will be measurement errors and one should take care of this problem. Finally, for the proper analysis of the data, it is necessary to select the appropriate method of statistical analysis. Since most of the standard statistical techniques are based on different model assumptions (like Gaussianity, independence, homogeneity etc.), it is necessary to test these assumptions before the actual analysis as the nature of astronomical data are generally unknown and even the “cause and effect” relationship is not clear under different situations.

All the above requirements create necessity for the refinement of existing techniques or development of new techniques for appropriate statistical analysis.

For many real life situations the number of variables under consideration as well as the number of observations are very large. In order to analyze such multivariate data, it is necessary to reduce the dimension properly. A smaller dimension is necessary for further analysis like classification or clustering. In statistics, Principal Component Analysis (PCA) is the most popular among the dimension reduction techniques. Although basically PCA is an exploratory technique, for making inference it is necessary to make
normality assumption regarding the underlying multivariate distribution. In practice, decisions regarding the quality of the Principal Component approximation should be made on the basis of eigenvalue-eigenvector pairs. In order to study the sampling distribution of their estimates the multivariate normality assumptions become necessary as otherwise it is too difficult.

More recently, independent component analysis (ICA) has emerged as a strong competitor to PCA and factor analysis. Whereas ICA finds a set of source data that are mutually independent, PCA finds a set of data that are mutually uncorrelated. ICA was primarily developed for non-Gaussian data in order to find independent components responsible for a larger part of the variation.

Under non-Gaussian multivariate situation the test for equality of location or scale parameters is not quite common and one may have to take help of nonparametric tests as the standard statistical techniques are optimal under the assumptions of independence, homoscedasticity and normality.

Regression analysis is a widely used method in astronomical research. It is used for two purposes: (i) to develop a quantitative relationship among astronomically observed properties of a set of objects and (ii) to predict the values of a particular property in terms of other properties of that set, e.g., relations between X-ray temperatures and velocity dispersions for galaxy clusters, the color-luminosity relations for field galaxies, period luminosity relation for variable stars, Tully-Fisher relation (maximum rotation velocity vs. luminosity relation) for galaxies and other Fundamental Plane (FP) relations when considering more than two variables.

The extension from regression line to regression plane is an important aspect of studying Fundamental Plane of galaxies. Now, there are various characteristics of galaxies which are correlated, e.g., a galaxy with a higher luminosity has a larger central velocity dis-
persion ($\sigma$) (Faber & Jackson 1976 [ME76]) or a galaxy with a larger size (viz. effective radius $r_e$) have fainter effective surface brightness ($<\mu_e>$) (Kormendy 1977 [J.77]). The usefulness of these correlations is, when a characteristic that can be determined without prior knowledge of galaxy's distance, e.g., central velocity dispersion and it is correlated with a characteristic, such as luminosity or in turn effective radius, that can be determined only for galaxies with known distances then, with this correlation, one can determine the distances to distant galaxies which is a difficult task in astronomy. The Fundamental Plane relation is applicable for giant early type galaxies and extends to faint and low-mass galaxies (Nieto et al. 1990 [JLERP90]). The symmetric plane plays the role of the so called FP used generally in Astrophysics. The above problems will become more complicated when the "cause and effect" relationship is not known and measurement errors are present in the data.

1.1 Review of earlier work

While studying problems related to globular cluster and galaxy formation theory and related issues, it is found that most of the previous works suffer from the defects like consideration of one or two variables at a time, proper choice of the variables, proper validation of the model, presence of outliers, use of appropriate statistical tools and techniques etc. For example, it has been seen that there are kinematic differences between the sub populations of GCs in different galaxies. These differences can be used as an observational constraint on the galaxy formation model. In many previous works the GCs are classified as metal rich and metal poor on the basis of the value of a single metallicity parameter [Fe/H] > or < -1 which is subjective in nature and also inappropriate in a multivariate setup. Concentrating on a single parameter means that one ignores the joint effect of several parameters.
The globular cluster system of the Milky Way consists of over 150 known members. It is likely that not all the clusters have been detected, primarily because of obscuration by the Galactic bulge. The best estimate for the total population is around 180 objects. The system is centrally concentrated, with roughly half of the globular clusters residing within about 5 kpc of the Galactic center. The majority of globular clusters form a roughly spherical, metal-poor, halo distribution. A smaller number of globular clusters are relatively metal-rich and have the spatial and kinematic characteristics of a thick disk population. One of the most striking aspects of the properties of Milky Way globular clusters is the lack of correlations between them. Some of the most detailed statistical analyses of correlations between globular cluster properties have been carried out by Djorgovski and collaborators (e.g. Djorgovski 1993 [S.93]; Djorgovski & Meylan 1994 [SG94]; Djorgovski 1995 [G.95]). Metallicity does not correlate with any other globular cluster property, with the exception of position within the Galaxy and the stellar mass function slope. The two most notable correlations, other than those depending on position within the Galaxy, involve globular cluster velocity dispersion which increases with both luminosity and central surface brightness. The possibility that the halo globular cluster system is comprised of two distinct populations is a significant development in this area of research. Zinn (1993) [J.93] suggests that the younger halo globular clusters have an accretion origin, as envisaged by Searle and Zinn (1978) [LR78]. This is supported by the correlations found by van den Bergh (1993a) [vdBS3a]. However, this still leaves considerable scope for the question of how these clusters actually formed. The discovery of the Sagittarius dwarf and its globular cluster system suggests that at least some halo globular clusters in the Milky Way originally formed around dwarf galaxies. It therefore seems likely that the halo globular cluster system, and the younger halo system in particular, may consist of clusters with more than one origin.

As discussed earlier, for studying galaxy formation and evolution, globular clusters
are considered as the unique tool as their ages are comparable to the ages of their host galaxies (viz. $\sim$ Gyr). So it is essential to study the formation process of the globular clusters (GCs). Recent developments in the study of Milky Way globular clusters (GCs) have generated considerable excitement with the suggestion that the halo system of globular clusters is composed of two subsystems (Zinn 1993 [J.93]; van den Bergh 1993 [vdBS93], 1994 [vdBS94]; Mackey & van den Bergh 2005 [DvdBS05]; Chattopadhyay & Chattopadhyay 2007 [TK07], hereafter CC07): (1) the inner halo, consisting of comparatively older GCs with a metallicity gradient, in a flattened spatial distribution around the galactic disk and (2) the outer halo, consisting of comparatively younger GCs having a negligible metallicity gradient, and high retrograde orbits. Several others have attempted to identify candidates for accreted GCs. But all these previous studies have taken one parameter at a time for comparison, thus neglecting the joint effects of the remaining ones.

There are kinematic differences among the sub-populations of GCs in different galaxies. These differences can be used as an observational constraint on the galaxy formation model. With the above objective in mind, we have carried out a multivariate analysis of extragalactic GCs (Chattopadhyay et al. 2009 [KTE'09]) In this context, NGC 5128 is of interest because it is the nearest giant elliptical galaxy whose large sample of GCs is amenable to spectroscopic observations (Beasley et al. 2008 [ATE08]) and whose structural parameters have been derived by fitting models to surface brightness profiles based on Hubble Space Telescope/Advance Camera for Surveys (HST/ACS) imaging (McLaughlin et al. 2008 [EPE08]). Also the radial velocities are available for a large sub-sample of GCs (Woodley et al. 2007 [AEA+07]).

Work in the past few decades has shown that the stars in globular clusters are among the oldest stars in the Galaxy, with ages of the order of $10^9$ years. Previous studies have revealed that the variation in the HB morphology is due to metallicity of the cluster. But
metallicity variation is not sufficient to explain in many cases the observed differences between the HB of Galactic (our Galaxy) globular cluster (GGC), e.g., clusters M2 and M3 have similar metallicities [Fe/H] ~ -1.6 but their HB morphologies are different (Stetson et al. 1996 [BA\textsc{m}96]). M2 has a blue HB exhibiting a long tail and M3 on the other hand has comparable number of stars on each side of the RR Lyrae gap. These inhomogeneities in the chemical composition of RGB stars have been explained to have primordial origin. In the previous literatures it appears that scientists have used different HB parameters to explain the morphology.

Regression analysis is a widely used method in astronomical research. In case of two variables X and Y, one is treated as independent and the other dependent and ordinary least squares method gives a single linear regression of the dependent variable Y against the independent variable X, denoted by $\text{OLS}(Y|X)$. OLS$(Y|X)$ is one which minimizes the sum of squares of the Y residuals and predicts Y in terms of X. For astronomical purpose several problems are faced with the above choice. If the choice of the independent variable is not clear then there is alternative option of OLS$(X|Y)$ and the distinction between these two approaches is often not clear (Bandiera & Hunt 1989 [RL89]) though a third robust process is discussed by Branham (1982) [Jr.82] and Lutz (1983) [E.83] which is not least squares procedure at all. So one single relationship treating X and Y symmetrically is required. In the above discussion measurement errors have not been considered. Measurement errors are the errors which arise in the measurement process of the instrument, e.g., signal to noise ratio, repeated measurements of some property etc. Heteroscedastic measurement errors are mentioned in Isobe et al. (1990) [TDGJ90] as well as in Feigelson & Babu (1992) [DJ92]. In the paper Isobe et al. (1990) [TDGJ90], the authors deal with data having no measurement errors. Feigelson & Babu (1992) [DJ92] performed regression between two variables including measurement errors. Nonlinear regression using ORDPACK (Boggs et al. 1990 [TRHB90]; Press et
al. 1986 [HPAT86]) have been performed also. Akritas & Bershady (1996) [GA96] have developed regression regarding two variables including known measurement errors, (a) allowing measurement errors of both the variables, (b) allowing measurement errors of the variables to be dependent on each other, (c) measurement errors depending on measurement and (d) finding other symmetric lines, e.g., bisector and orthogonal regression etc. Then these techniques are applied to various astrophysical regression situations like color-luminosity relation for field galaxies, Tully-Fisher relation and Tolman test as mentioned above.

1.2 Dimensions of the present work

The modest contribution to "Astrostatistics" made in this dissertation can be briefly stated as follows:

1. Study the origin of Milky Way Globular Clusters through horizontal branch morphology (Babu et al. 2009 [JTKS09]).

2. Study the formation of Galactic halo using Globular Clusters as archeological probe by a nonparametric approach (Mondal et al. 2008 [SKT08]).

3. To study the individual galaxy formation through Globular Clusters on the basis of modified Principal Component Analysis (PCA) and model based Clustering (Chattopadhyay et al. 2009 [KTE+09]).

4. To use the Independent Component Analysis (ICA) for a further improvement (Chattopadhyay et al. 2013 [KST13]).

5. To set up regression with unknown nature of interdependence among the variables with measurement errors with an application to fundamental plane of galaxies
1.3 Preview of the current work

In the present work multivariate statistical analyses have been carried out to study problems related to search for the origin of Milky Way Globular Clusters, formation of galactic halo, NGC 5128 galaxy formation through globular clusters and prediction of galaxy parameters under unknown nature of dependence among the parameters.

In chapter 2, we have started our study using principal component analysis (PCA) to identify the proper HB morphology parameter. Once the proper HB morphology parameter is selected, our next quest is to find significant independent, intrinsic parameters of the globular clusters in our Galaxy which are mostly responsible for variation of the HB morphology. For this, first a classification of the Galactic globular clusters is performed with respect to an independent set of intrinsic parameters assuming there are three coherent groups. Then HB morphology parameter is predicted in terms of these parameters in each homogeneous group. Finally, as HB stars are the post evolution stage of the RGB stars, a cluster analysis is carried out for the RGB stars of Galactic inner halo to examine their formation origin. So various steps followed are as follows:

1. We have selected the proper HB morphology parameter in an objective way using PCA (Chattopadhyay & Chattopadhyay 2006 [TK06]).

2. We have taken three subgroups of GGC following Chattopadhyay & Chattopadhyay (2007) [TK07] instead of one as taken by Recio-Blanco et al. (2006) [AAG+06].

3. We have applied stepwise multiple regression technique instead of ordinary regression used by Recio-Blanco et al. (2006) [AAG+06] for these Galactic globular clusters.
4. We have applied a Cluster Analysis (CA) (Chattopadhyay & Chattopadhyay 2007 [TK07]) technique to the different samples of RGB stars of GGC to verify the actual number of sub-populations and compared them with the previous results. A change point analysis has also been carried out to verify the nature of different subgroups of GGC obtained by CA.

In chapter 3, we have compared the origin of GCs of Milky Way belonging to disk, inner halo and outer halo separately with those in the neighboring dwarf galaxies under a nonparametric multivariate setup since the sizes of all the data sets are either moderate or very small and the tests for multivariate normality were rejected for all of them.

In chapter 4, we have carried out a multivariate analysis of extragalactic GCs. In this context, NGC 5128 is of interest because it is the nearest giant elliptical galaxy whose large sample of GCs is amenable to spectroscopic observations (Beasley et al. 2008 [ATE+08]) and whose structural parameters have been derived by fitting models to surface brightness profiles based on Hubble Space Telescope/Advance Camera for Surveys (HST/ACS) imaging (McLaughlin et al. 2008 [EPE+08]). Also the radial velocities are available for a large sub-sample of GCs (Woodley et al. 2007 [AEA+07]).

In the present study, we have first used a modified technique of principal component analysis (PCA; Salibian Barrera et al. 2006 [MSG06]) to search for the optimum set of parameters which gives the maximum variation for the GCs in NGC 5128. This method helps to extract the significant parameters from the large set of photometric, structural, and kinematic parameters. Then we have classified the GCs on the basis of these significant parameters using a model based method of cluster analysis (CA; Qiu & Tamhane 2007 [DC07]) which finds the structure of the optimum groups of GCs instead of choosing groups in an ad hoc manner on the basis of a single parameter. This multivariate analysis helps us to enunciate a more efficient theory of GC formation.
In this context, it should be mentioned that kinematic studies of GCs in several giant elliptical galaxies show results which differ from galaxy to galaxy (Woodley et al. 2007 [AEA+07]). In a recent study (Hwang et al. 2008 [SGS08]), the rotations of metal-poor and metal rich GCs have been studied in six giant elliptical galaxies; two systems of GCs show strong rotation, while the other ones show moderate or weak rotation. We have studied NGC 5128 as the representative of the latter group, while it remains important to study representatives of the former group in a multivariate setup, once the adequate data are available.

In chapter 5, we used Independent Component Analysis (ICA), which is a strong competitor of PCA to improve the analysis of the previous chapter. Although the above mentioned modified PCA is quite robust in the presence of outliers, but it performs better when the sample size is quite large in comparison with the dimension of the data set. Further, test for normality shows that the observations are from non-Gaussian distribution.

In the present study we have done K-means clustering on the basis of Independent Components (ICs) as well as Principal Components (PCs) to identify the proper method applicable to the present data set on the basis of within cluster sum of squares.

In chapter 6, attempts have been made to divide a sample of 430 early type galaxies into several homogeneous groups and then to find multiple regression planes (e.g. FP) for prediction purpose by considering all possible combinations of dependent and independent variables as the "cause and effect" relationship is not known. Finally a symmetric regression plane has been developed for each group in which any one variable among the three can be considered as a dependent variable.

Throughout this work the data used are compiled from different archives (like Vizier, Aladin etc.), published works and sky surveys. Some of the parameters are also derived
from physical relations. R codes have been written for the different algorithms used for data analysis and are listed in the appendix.