FOSM-2022/23 Statistical Analysis Project
Overview of the globular star clusters dynamics in the Milky Way: analysis of the properties that affect the central velocity dispersion

Mattia Bennati
7122582
mattia.bennati@edu.unifi.it

Abstract

The objective of this report is to study and evaluate the intrinsic characteristics of globular clusters in order to identify which properties affect the central velocity dispersion and how they influence each other. To investigate the correlations between the data, both analytical and graphical approaches have been used. In this case, the $GlobClus\_prop$ *dataset has been examined.*
The related source code will be released publicly.

Future Distribution Permission

The author of this report gives permission for this document to be distributed to Unifi-affiliated students taking future courses.

1. Introduction

Globular clusters are spherical collections of stars that orbit a galactic core as a satellite. They are very tightly bound by gravity, which gives them their spherical shapes and relatively high stellar densities toward their centers. Stars in globular clusters typically have a lower proportion of heavy elements, because they formed before the universe had been enriched by successive generations of stars. In the Milky Way, globular clusters are distributed in the halo of the galaxy and revolve around the galactic center in highly elliptical orbits. Globular clusters have interesting dynamical characteristics due to their dense stellar environments, such as the possibility of stellar collisions and the presence of blue stragglers, which are stars that appear younger than the rest.

2. The GlobClus_prop Dataset

2.1. Structure

The given dataset contains 20 fundamental properties of 147 clusters in the Milky Way. The data contains missing values with NaN placeholders and must be pre-processed before usage. Properties include Galactic location, integrated stellar luminosity, metallicity, ellipticity, central surface brightness, color, and seven measures of dynamical state. See appendix A.

2.2. Pre-processing phase

A first analysis showed that some records contained missing data that has been replaced with ”NaN” placeholders, and the cluster names feature is useless for the study. The first step actually consisted of cleaning the dataset by removing the incomplete records, and excluding the ”Name” column.
Another possible step could be to convert galactic longitudes and latitudes into distances, but the current study concerns variables that already take distances into account implicitly (e.g.: distance-independent brightness measurements or already scaled for distances).
Clusters with specific values of R.sol and R.GC are compared within a specific consistent context. So the conversion has been deemed as not useful in this case.
However, this step would have been necessary if the task were to determine the spatial distribution of clusters in three dimensions, or if calcu-

lations requiring explicit distance measurements were performed. (E.g.: kinetic or potential energy within the galaxy, or simulations of cluster motions).

3. Model selection

3.1. Identification of the significant variables

In order to evaluate the impact of each variable on the central velocity dispersion, several approaches have been used starting from p-value and penalization criteria proceeding in all directions: forward, backward and mixed.

- **Forward procedure:** It's an iterative process that starts from the complete linear model (containing all the features) and consists of updating it by removing the less significant feature at each step, until all the remaining ones are significant. If a penalization criteria is used, the removed feature is the one that corresponds to the higher penalization value.
- **Backward procedure:** It's an iterative process that starts from the null linear model instead (containing only the intercept), and updates it by adding the most significant feature at each step, until there is no more significant feature to add.
- **Mixed procedure:** Consists of using the forward procedure at first, but for each step there is an additional evaluation. It consists in checking if the addition of a new significant feature to the model makes another one not significant anymore.
  The feature that is no more significant is removed, as it would happen with the backward approach, and the process goes forward until there are no more features to add or remove.

There are two main penalization methods, AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion) and the penalization terms are: $penAIC(d) = 2|d|$ and $penBIC(d) = |d|nlog(n)$ respectively.
The criteria suggest choosing the model $F_d(\theta; X)$ with a minimum $AIC_d = penAIC(d) - l(\hat{\theta}d; X)$ or $BIC_d = penBIC(d) - l(\hat{\theta}d; X)$

3.2. Results

The different approaches identified the following five unique models:

a) 6 occurrences:
$S0 \sim Mv + log.t + V.esc + E.B.V + CSBt$

b) 1 occurrence:
$S0 \sim Mv + V.esc + Conc + log.rho + log.t$

c) 1 occurrence:
$S0 \sim Mv + V.esc + log.rho + log.t$

d) 1 occurrence:
$S0 \sim V.esc + Conc + CSBt + E.B.V + log.rho + Mv + log.t$

e) 1 occurrence:
$S0 \sim V.esc + Conc + CSBt + E.B.V + log.rho$

The model $'$**a**$'$ has been selected here since: it was the one with the highest features significance, the p-values were very low, it had the lowest standard error, and it has been identified many times with different approaches.
This is the summary of the identified model:

```
Call:
lm(formula = S0 ~ Mv + log.t + V.esc + E.B.V + CSBt)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.44914 -0.07025  0.01326  0.06725  0.75342 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -2.404720   0.200281  -12.01   <2e-16 ***
Mv           0.536888   0.040365   13.30   <2e-16 ***
log.t        1.346157   0.067872   19.83   <2e-16 ***
V.esc        0.233623   0.002616   89.32   <2e-16 ***
E.B.V        1.541982   0.094606   16.30   <2e-16 ***
CSBt        -0.506816   0.032777  -15.46   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’
↪  0.1 ‘ ’ 1

Residual standard error: 0.1636 on 107 degrees of
↪  freedom
Multiple R-squared:  0.9978,	Adjusted R-squared:
↪  0.9977 
F-statistic:  9661 on 5 and 107 DF,  p-value: < 2.2e-16
```

4. Model Analysis

4.1. Initial evaluation

According to the selected model, the features that affect the central velocity dispersion the most are: the absolute magnitude (Mv), the central relaxation time (log.t), the central escape velocity (V.esc), the color excess (E.B.V) and the central superficial brightness (CSBt).
Effects of each feature on the velocity dispersion:

- **Mv:** an increase in magnitude of one unit leads to an increase of approximately 0.536888 Km/s
- **log.t:** one year of relaxation time consists of an increase of 1.346157 Km/s
- **V.esc:** an increase in escape velocity of 1 Km/s results in an increase of 0.233623 Km/s
- **E.B.V:** an increase of one unit in the magnitude of the colour excess results in an increase of 1.541982 Km/s
- **CSBt:** the one-unit increase in surface brightness results in a decrease of 0.506816 Km/s

By looking at the model summary we can see that:

- The **standard error estimate** ”0.1636” out with 107 degrees of freedom is relatively low, and this indicates good estimation accuracy.
- The **determination coefficient** is ”0.9977”, and by being this high, indicates that the model explains the variance of the central velocity dispersion, almost completely.
- The **test function F** is extremely significant [9661 out of 5 and 107 GDL], and the residuals are well distributed with the median close to zero.

The model seems to accurately fit the data.

4.2. Residuals inspection

By plotting the model characteristics, it is clear that the residuals follow an approximately normal distribution.
It seems anyway that the dataset contains some external points that might interfere with the model (possible outliers)

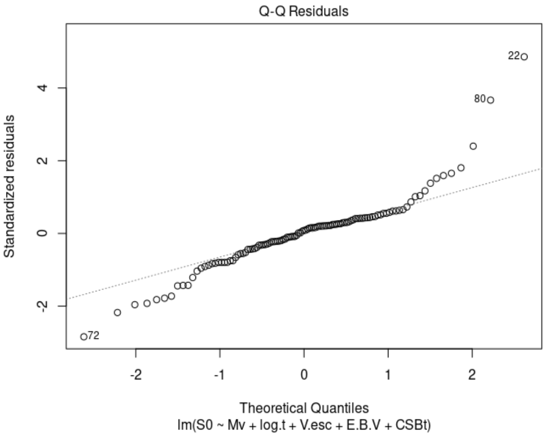

Figure 1. Theorical quantities and standardized residuals

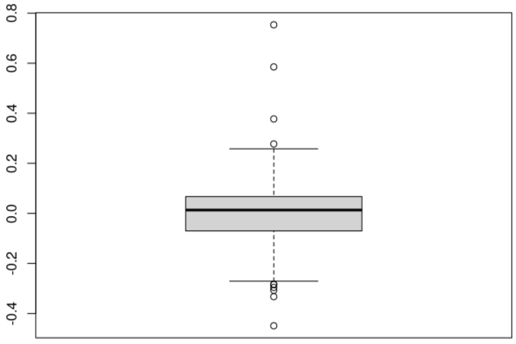

Figure 2. Residuals boxplot

A further inspection of the residuals with a boxplot confirmed the possibility of having some outliers.
However, by creating a linear model without the outliers and comparing it with the original one, it is clear that they don’t have a big impact, since their graphs look the same.
The model seems very robust and has a great tolerance in relation to abnormal values.

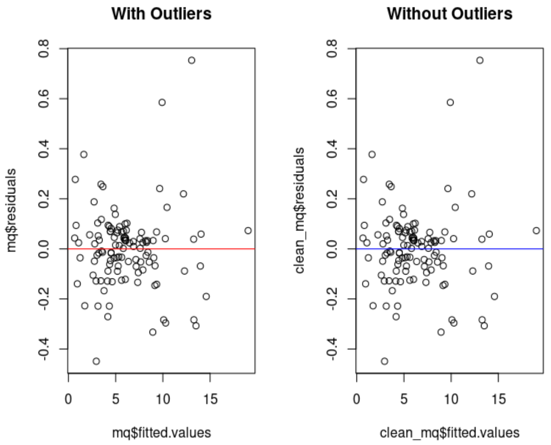

Figure 3. Comparison between the original model and the one without outliers

4.3. Evaluating the impact of single features

To evaluate the effect of each feature over the central velocity dispersion, generic linear models with single independent variables have been created and evaluated.
This is a summary of the detected measurements:

- **Mv:** confirmed a high significance.
  An increment to the level of magnitude leads to a decrease in velocity dispersion of 2.2190 Km/s.

- **log.t:** the significance is still high, but slightly reduced compared to the model that considers the identified features.
  For each relaxation year, there is a reduction of velocity dispersion of 1.0902 Km/s.

- **V.esc:** confirmed a high significance.
  The increase in the escape velocity of 1 Km/s leads to an increase of the central velocity dispersion of 0.23609 Km/s.

- **E.B.V:** slightly reduced significance compared to the initial model, but still high.
  An increase of one unit in color excess magnitude corresponds to an increment of 1.80018 Km/s in velocity dispersion.

- **CSBt:** highly significant.
  An increment of the superficial brightness of 1 unit leads to a reduction in velocity dispersion of 1.1231 Km/s

4.4. Features correlation analysis

The analysis initially consisted of plotting the features' graphs to analytically determine the possible dependency relationships, and the following potential correlations have been found:

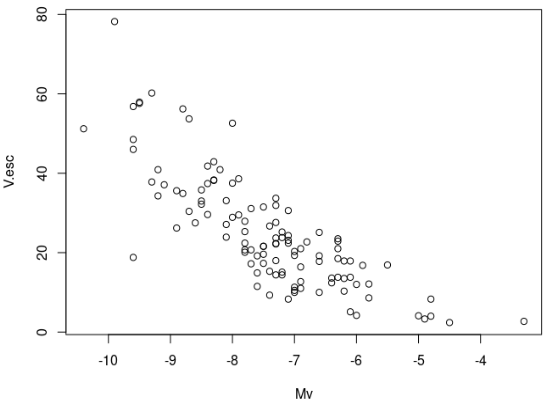

Figure 4. Potential correlation between absolute magnitude and escape velocity

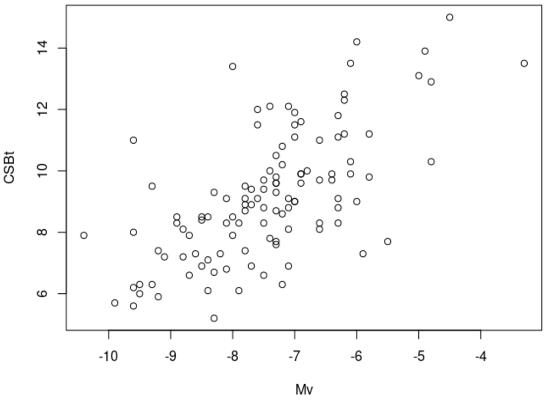

Figure 5. Potential correlation between absolute magnitude and superficial brightness

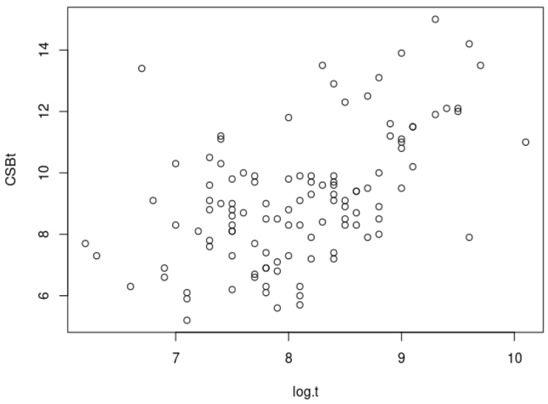

Figure 6. Potential correlation between central relaxation time and superficial brightness

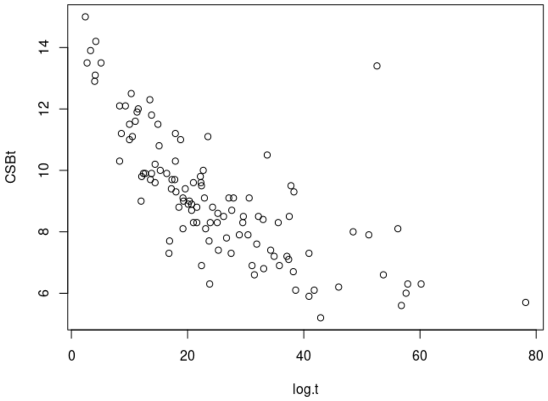

Figure 7. Potential correlation between escape velocity and superficial brightness

The second step consisted in the analysis of more complex models, containing the single interactions between the possibly correlated features, the computation of the correlation indexes, and confidence intervals.

The significant correlations that concern only the superficial brightness, are with: escape velocity, relaxation time and absolute magnitude.

The indexes are as follows:

```
# Testing Mv _|_ CSBt #
Statistic (DEV):   63.453 df: 1 p-value: 0.0000 method:
↪  CHISQ
Cor. Idx: 0.6554871
```

```
# Testing log.t _|_ CSBt #
Statistic (DEV):   34.811 df: 1 p-value: 0.0000 method:
↪  CHISQ
Cor Idx: 0.5149092
# Testing V.esc _|_ CSBt #
Statistic (DEV):   85.399 df: 1 p-value: 0.0000 method:
↪  CHISQ
Cor. Idx: -0.7282425
```

The definition of a new model containing all the interactions led to a lower AIC value and showed that the correlation with the escape velocity is not significant to estimate the central velocity dispersion.

The model is now more balanced in terms of adaptivity and complexity. So the previous one has been discarded, and this is the final summary:

```
Call:
glm(formula = S0 ˜ Mv * CSBt + log.t * CSBt + V.esc +
↪  E.B.V)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.115921   0.920163  -3.386 0.000998 ***
Mv           0.852553   0.100381   8.493 1.43e-13 ***
CSBt        -0.291760   0.103315  -2.824 0.005676 **
log.t        1.556646   0.097119  16.028  < 2e-16 ***
V.esc        0.248964   0.004556  54.645  < 2e-16 ***
E.B.V        1.027439   0.136557   7.524 1.90e-11 ***
Mv:CSBt     -0.031114   0.008673  -3.587 0.000509 ***
CSBt:log.t  -0.038621   0.010098  -3.825 0.000223 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’
↪  0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be
↪  0.02194682)

    Null deviance: 1296.3894  on 112  degrees of
    ↪  freedom
Residual deviance:    2.3044  on 105  degrees of
↪  freedom
AIC: -101.18

Number of Fisher Scoring iterations: 2
```

By plotting the features and the central velocity dispersion, a very high correlation has been found with the escape velocity. The two quantities go almost hand in hand (linear growth):

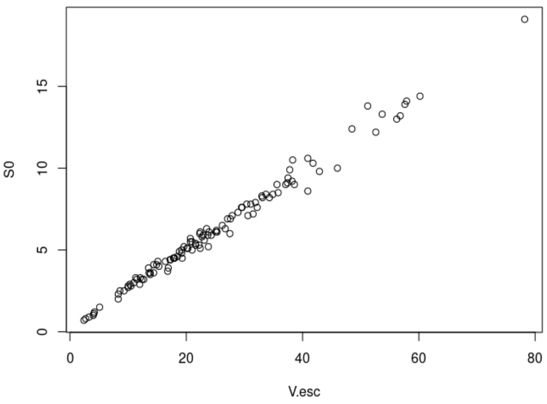

Figure 8. Correlation between escape velocity and central velocity dspersion

4.5. Multivariate Analysis (graph-based)

This phase consisted of evaluating the identified model accuracy and correlations while estimating the central velocity dispersion's trend, based onto graphical techniques. More specifically, undirected graphs with penalization criteria and bayesian networks have been used.
The first step concerns the creation of a graphical gaussian model starting from continuous data that considers only the features present in the final model.
The model has been used to fit the data, and a stepwise BIC "forward" logic has been applied to it.
The resulting graph appears as follows:

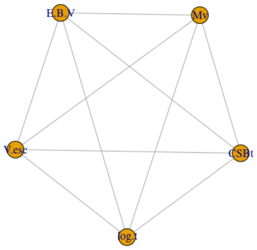

Figure 9. Initial graph of dependencies based onto the final model's features

The graph shows that all the features are dependent on each other, but a more in-depth evaluation must be carried out to investigate further and determine which correlations are actually significant.
This is the evaluation of the correlation matrix that has been used to determine the strength of the dependencies:

```
                Mv       log.t      V.esc        E.B.V       CSBt
Mv      1.00000000 -0.08430665 -0.8085964  0.05965678  0.6554871
log.t  -0.08430665  1.00000000 -0.3096412 -0.39958765  0.5149092
V.esc  -0.80859641 -0.30964120  1.0000000  0.20965656 -0.7282425
E.B.V   0.05965678 -0.39958765  0.2096566  1.00000000  0.2083057
CSBt    0.65548711  0.51490924 -0.7282425  0.20830574  1.0000000
```

By evaluating the correlation matrix and excluding the weak dependencies, this is the final representation of the model:

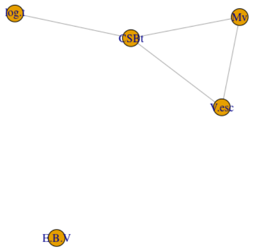

Figure 10. Graph of the strong correlations between the features

This clearly shows that the superficial brightness is strictly dependent on the relaxation time, escape velocity and absolute magnitude, as previously stated.
The creation of a bayesian network that considers the same features, shows identical relationships between them, but also the direction of each dependency.
In particular, it is possible to see that the superficial brightness trend depends on the relaxation time, escape velocity and absolute magnitude, not the contrary:

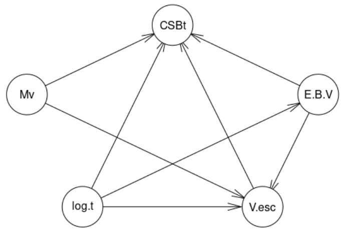

Figure 11. Bayesian network of the chosen model

5. Conclusions

The conducted study helped into understanding the dynamics of globular clusters concerning the central velocity dispersion in an agnostic manner. It highlighted the correlations between the involved features, it helped to understand some of their characteristics and trends while evaluating their overall and single significance for the model. The model that has been identified seems really accurate, and it correctly fits the data contained in the dataset.

Using many approaches to identify the best models reduced the possibility of overfitting the data. And the implementation of graphical models confirmed the strength of the analytical process applied.

Looking at the results obtained, it is clear that the central velocity dispersion depends primarily on the central relaxation time, the escape velocity, the superficial brightness, the color excess and the absolute magnitude.

It is important to denote that some of these properties are linked to each other, so the velocity dispersion is more likely dependent on their interactions. All the features in the final model are highly significant, and the standard error estimate is very low.

This is the resulting generic linear model:

$$S0 \sim Mv{*}CSBt{+}log.t{*}CSBt{+}V.esc{+}E.B.V$$

A. Features description

- **$Name$** : This is the common name of the cluster
- **$Gal.long$** : Galactic longitude (degrees). With respect to the center of the galaxy
- **$Gal.lat$** : Galactic latitude (degrees). With respect to the center of the galaxy
- **$R.sol$** : Distance from the sun (kiloparsecs, kpc) [1pc = 3.26 light years]
- **$R.GC$** : Distance from the center of the galaxy (kpc)
- **$Metal$** : Logarithmic metallicity of the cluster. With respect to the sun's
- **Mv** : Absolute visual magnitude. Luminosity measure of a celestial object seen from a fixed standard distance [Approximate indication of the mass]
- **$r.core$** : Radius from the core/nucleus (parsecs, pc). The Distance from the cluster's center within which the superficial density of stars reaches half of its average value
- **$r.tidal$** : Tidal radius (pc). The Spatial region around the cluster outside which the tidal forces generated by a more massive external body (e.g.: another galaxy), overcome the internal gravity.
- **$Conc$** : Nucleus concentration parameter (a-dimensional). It Indicates how dense or concentrated the cluster's core is with respect to its peripheral regions
- **$log.t$** : Logarithmic central relaxation time's scale (years). Describes the time required by a solar system to reach its dynamic balance given the gravitational forces between its stars
- **$log.rho$** : Cluster's central density logarithm (solar masses times cubic parsecs). Indicates the density of stars located in the central cluster's region
- **$S0$** : Central velocity dispersion (km/s). Average movement speed of stars around the cluster's core. Varies within a range for each star (dispersion)
- **$V.esc$** : Central escape velocity (km/s). Minimum speed required by a star to overcome the cluster's/galaxy's gravity starting from its center without any further propulsion. Strictly dependent on the cluster's/galaxy's mass.
- **VHB** : Horizontal branch level (Magnitude). Specifies the evolution phase of stars found in the Hertzsprung-Russell diagram (shows the relations between stars luminosity and their colors or superficial temperatures)
- **$E.B.V$** : Color excess (Magnitude). Describes the amount of starlight absorption and dispersion caused by the interstellar dust
- **$B.V$** : Color index (Magnitude). Color measure of an object based onto the different magnitudes seen in two different wavelength bands, typically visible light and infrared
- **$Ellipt$** : Cluster's ellipticity. Refers to the flattening level of a cluster with respect to a perfect sphere
- **$V.t$** : Integrated V magnitude (Magnitude). Describes the total luminosity of a cluster, as it appears when seen from an observation point on the earth, in the visible light bandwidth V (visible)
- **$CSBt$** : Central superficial brightness (Magnitude times squared arc-seconds). Measure of the emitted light for unit of area, usually calculated for the central portion of the cluster.