Mario Pasquato
Marie Skłodowska-Curie Fellow
Padua Observatory
(Italy)

Collaborators:
Federico Abbate
Abbas Askar
Ammar Askar
Alessandro Ballone
Chul Chung
Michela Mapelli
Mario Spera
Piero Trevisan

Clustering Clusters

Unsupervised machine learning on globular cluster structural parameters

arxiv:1901.05354

Bologna 22/01/2019

The problem

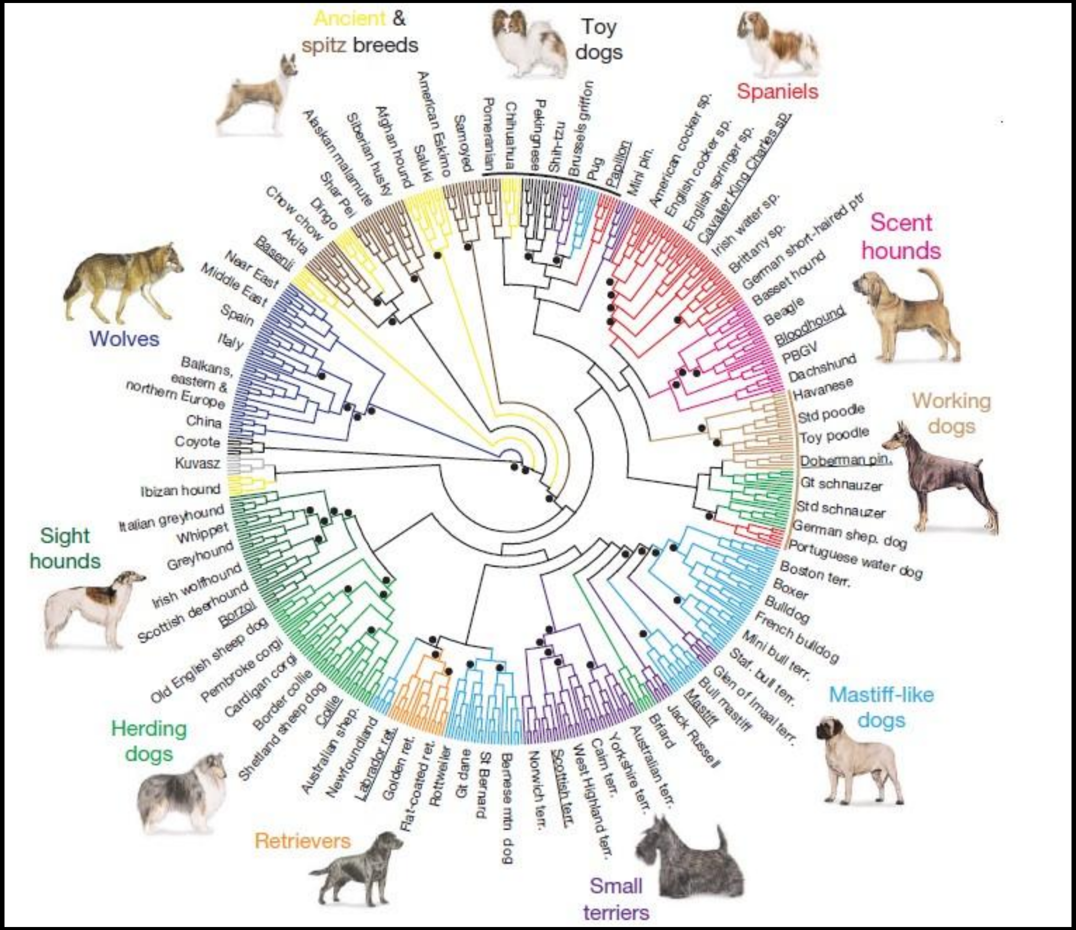
- How many groups of Globular Clusters (GC) are there?
  - 2 (disk/halo; Zinn 1985)
  - 3 (disk/inner-/outer-halo Lee et al. 1988, Zinn 1993)
  - or more?
- How strongly is each GC associated to its group?
- How good is the clustering structure in general?

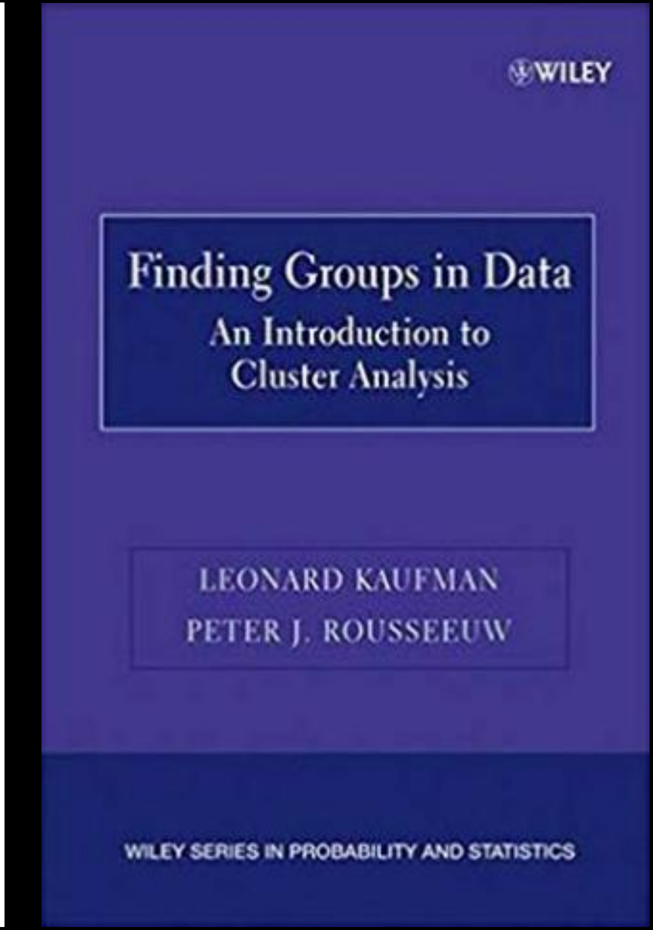
The solution

- How many groups of Globular Clusters (GC) are there?
  - 2 (disk/halo; Zinn 1985)
  - 3 (disk/inner-/outer-halo Lee et al. 1988, Zinn 1993)
  - or more? Objectively measure the number of groups by
- How strongly is each GC associated to its group? quantifying the strength of the association for each GC and
- How good is the clustering structure in general? finding the number of groups that maximizes the average of these `association strengths’

Unsupervised classification

A standard problem in unsupervised machine learning: finding groups in data (clustering)





WILEY

Finding Groups in Data

An Introduction to Cluster Analysis

LEONARD KAUFMAN

PETER J. ROUSSEEUW

WILEY SERIES IN PROBABILITY AND STATISTICS

Clustering: two ways

- Partitioning methods

  given the requested number of groups, find an ~~optimal~~ good way to assign items: K-means, PAM...

- Hierarchical methods
  - divisive: split the dataset into two groups, then each group into subgroups, all the way to items: DIANA
  - agglomerative: merge items into groups, groups into larger groups, all the way to the whole dataset: AGNES

  typically greedy, no global optimum

  return a dendrogram (like a philogenetic tree)

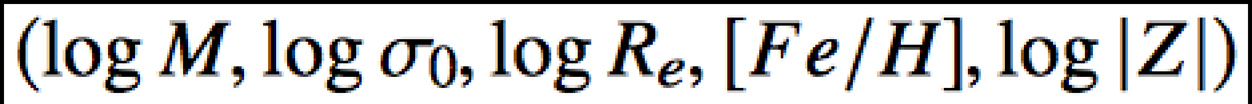
Partitioning around medoids

Kaufman L., Rousseeuw P., 1987

- Partitioning method
- Given data represented by points in $\mathbf{R}^n$ and the desired number of clusters k, finds k data points - centroids for each group (medoids)
- Assigns each point to the group that minimizes the mean distance from medoids
- Recalculates the medoids for each group to further lower the mean distance
- Iterates until convergence

Parameter space

- GCs represented by five numbers



$$(\log M, \log \sigma_0, \log R_e, [Fe/H], \log |Z|)$$

- Mass, central velocity dispersion, half-mass radius, metallicity, height on the Galactic plane

- All *ratio quantities*
- All logs (units of meas. have no effect)

PAM + GCs

I applied PAM to Baumgardt & Hilker 2018, catalog of GC structural parameters

End product

- 110 GCs classified into disk/inner halo/outer halo or disk/halo: ideal for statistical studies
- three (or two) medoids – representative GCs: ideal for case studies
- Silhouette widths: quantifier of the strenght of association of each GC with parent group

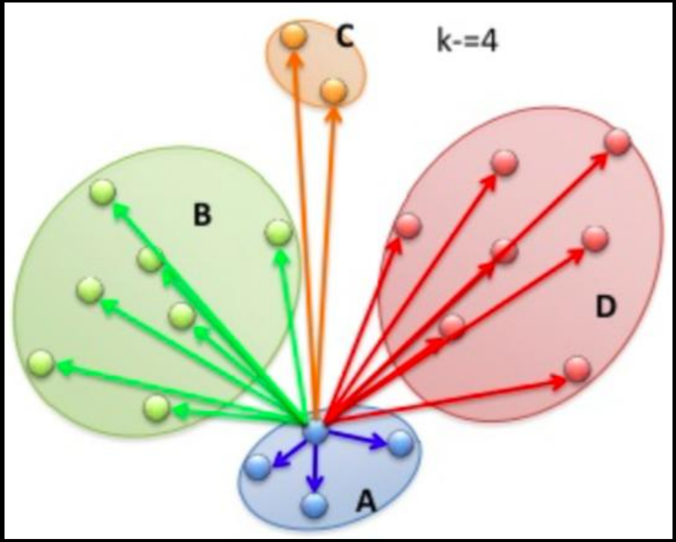
Silhouette width

measures how well an object fits in its assigned cluster (rather than in a neighbor)

mean distance from elements of the nearest other group - mean distance from elements of the same group

---

maximum of these two distances



$-1 < S_i < 1$

S_i ~ 1: good fit, well within group
S_i ~ 0: bad fit, on the fringe between groups
S_i ~ -1: very bad fit, in wrong group

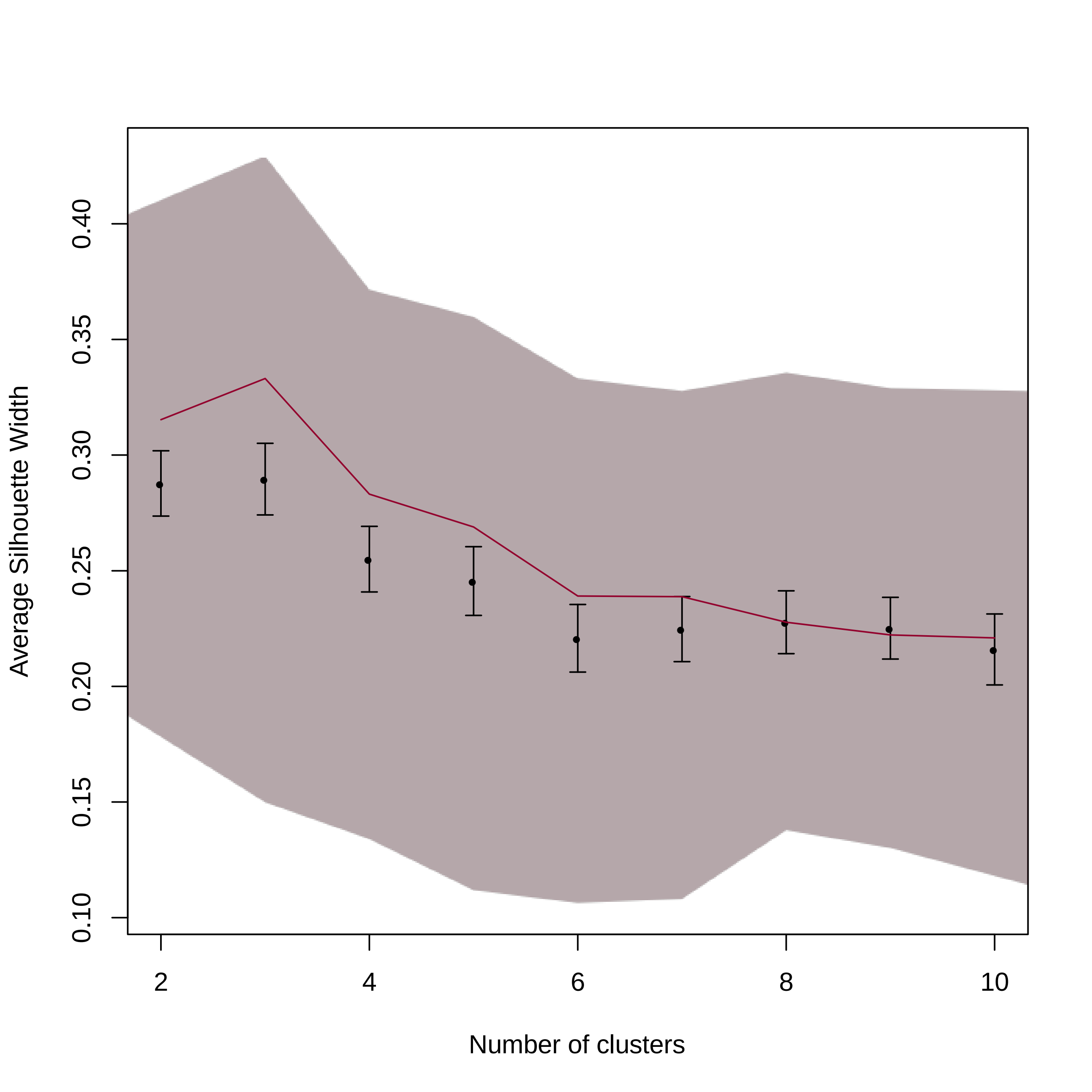
How many groups?

Average silhouette width as a function of the number of groups

Remember: higher silhouette width = better fit of each GC in its parent group

Natural number of GC groups: three (maximizes average silhouette width)

Two groups also good, more groups (4, 5...) are excluded

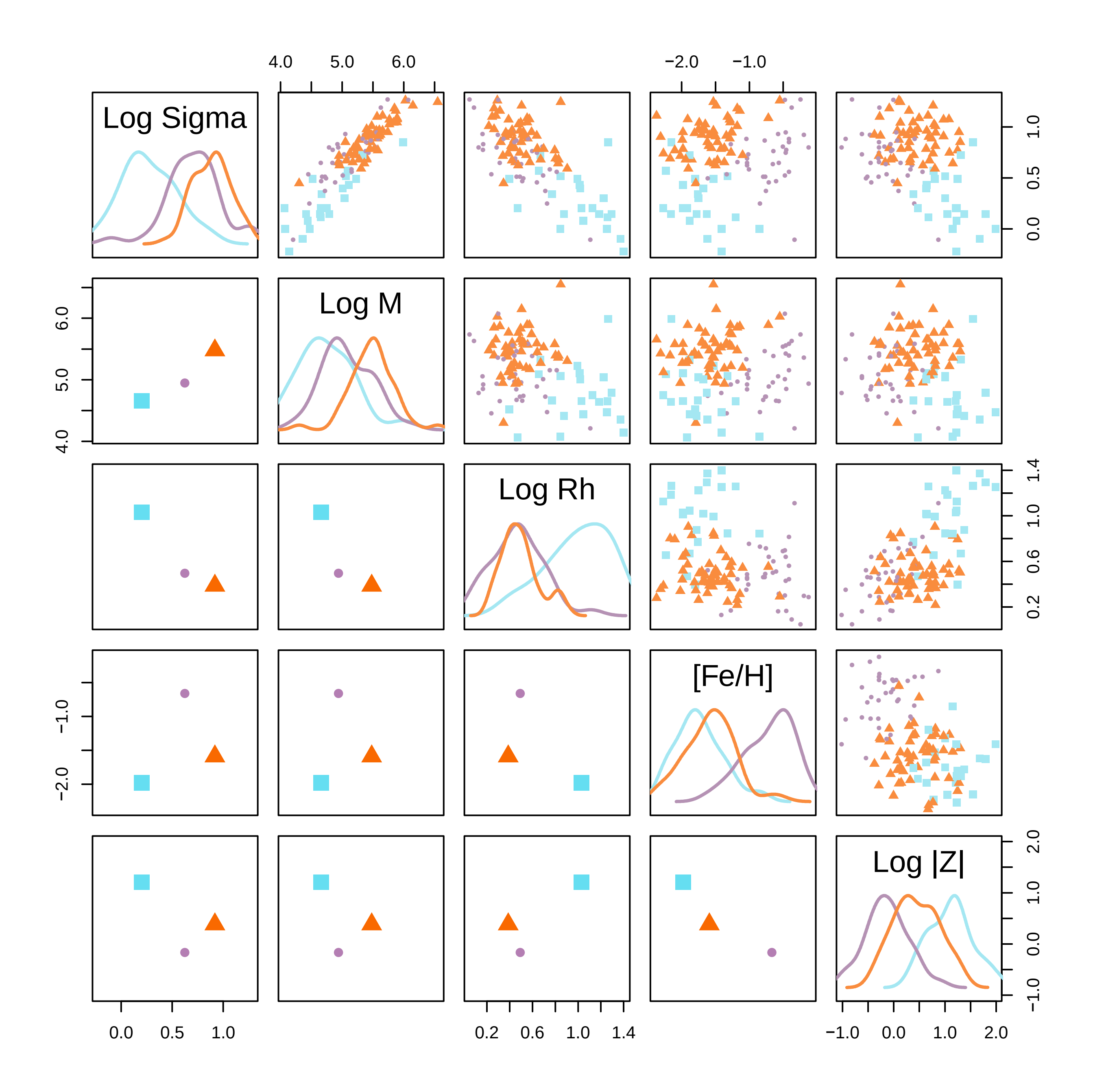


pair plot view
3 groups

Purple: low height on the Galactic plane, high [Fe/H]

Orange: intermediate height on the Galactic plane, low metallicity, high mass

Cyan: furthest from the Galactic plane, low metallicity, low mass but big radius



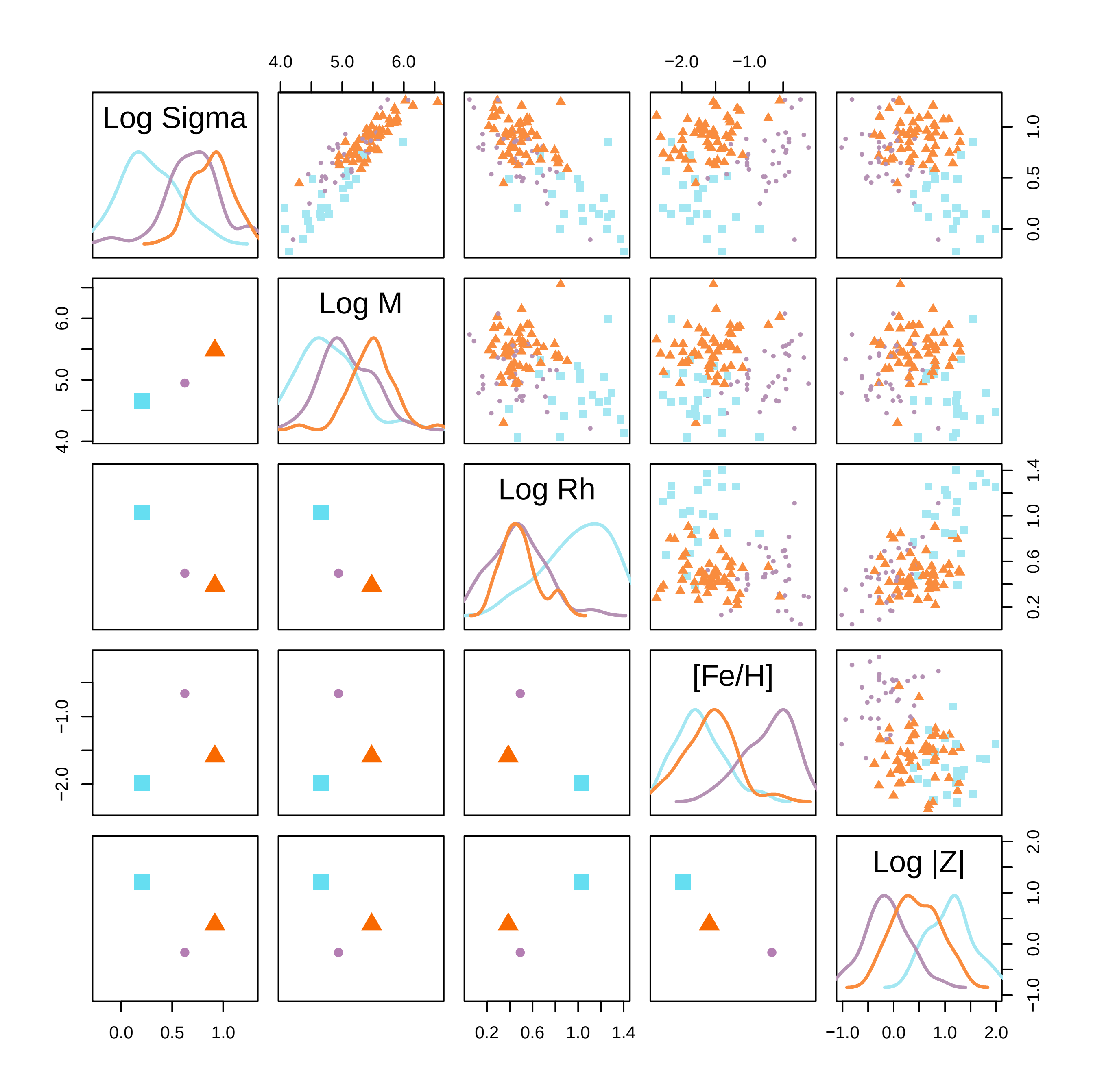
pair plot view
3 groups

Purple: disk

Orange: inner halo

Cyan: outer halo

we recovered
the Zinn trichotomy

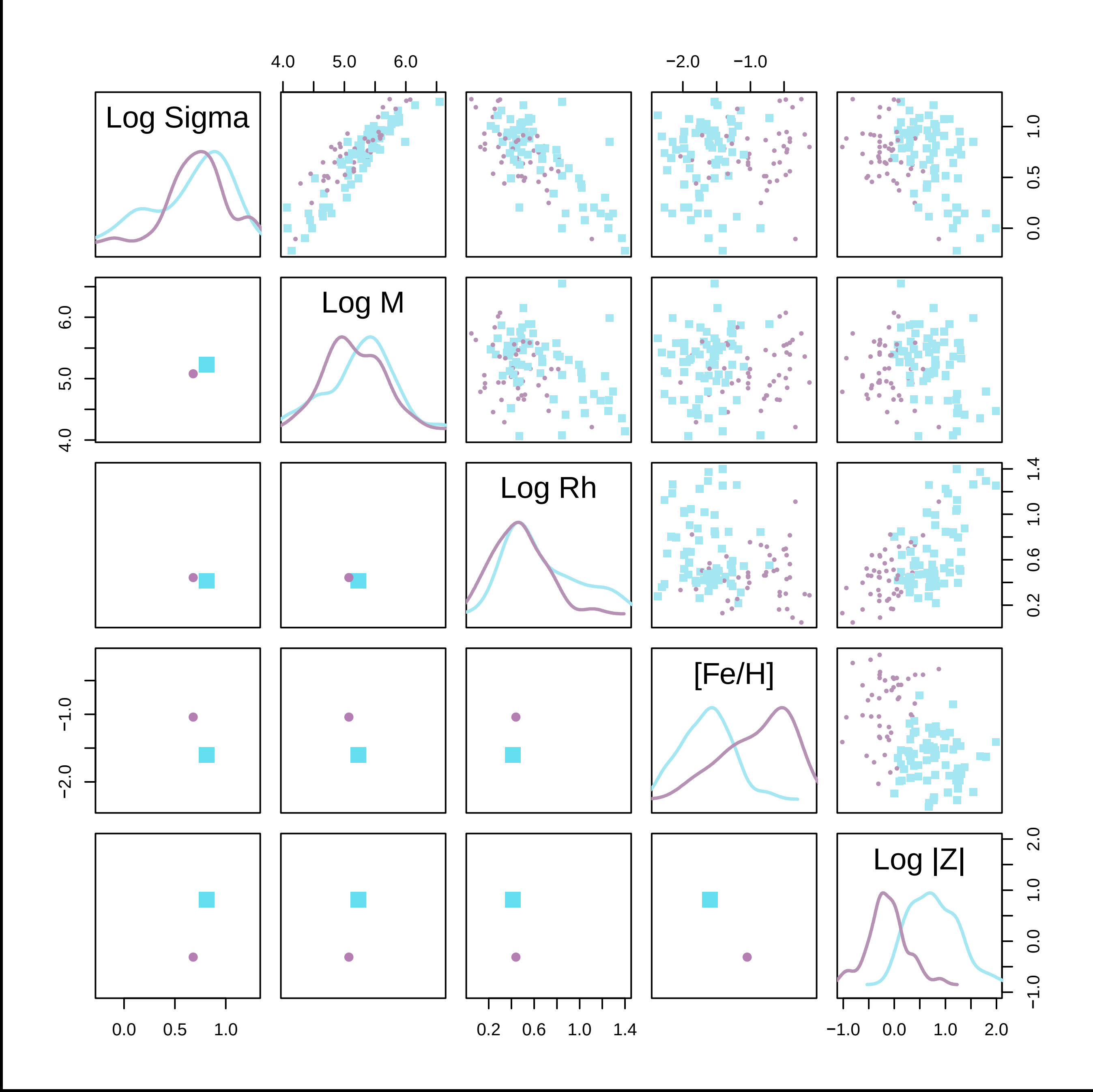


pair plot view
2 groups

Purple: disk

Cyan: halo

Inner halo is split, disk stays disk, outer halo stays halo

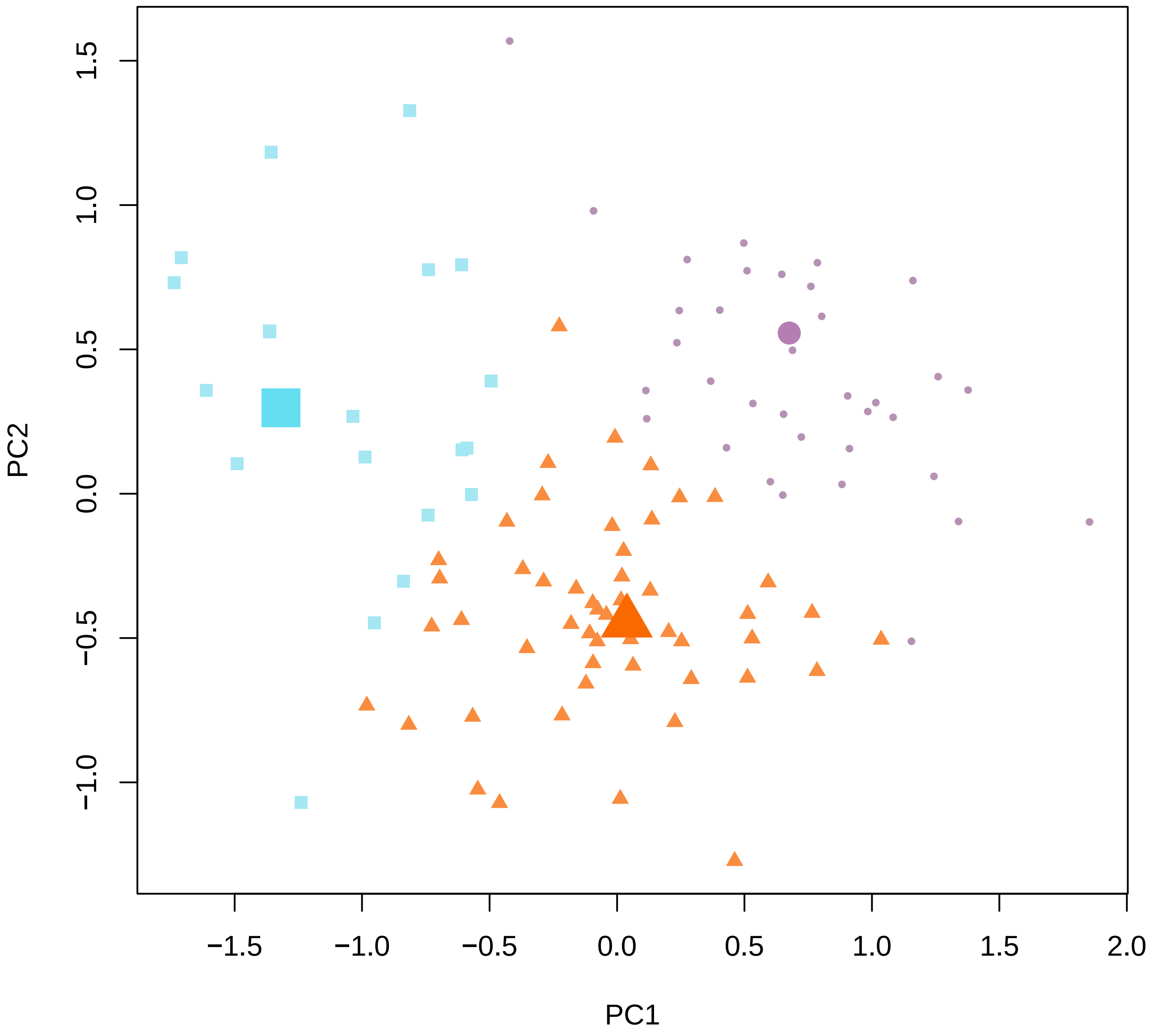


princip. comp. view
3 groups

Purple: disk

Orange: inner halo

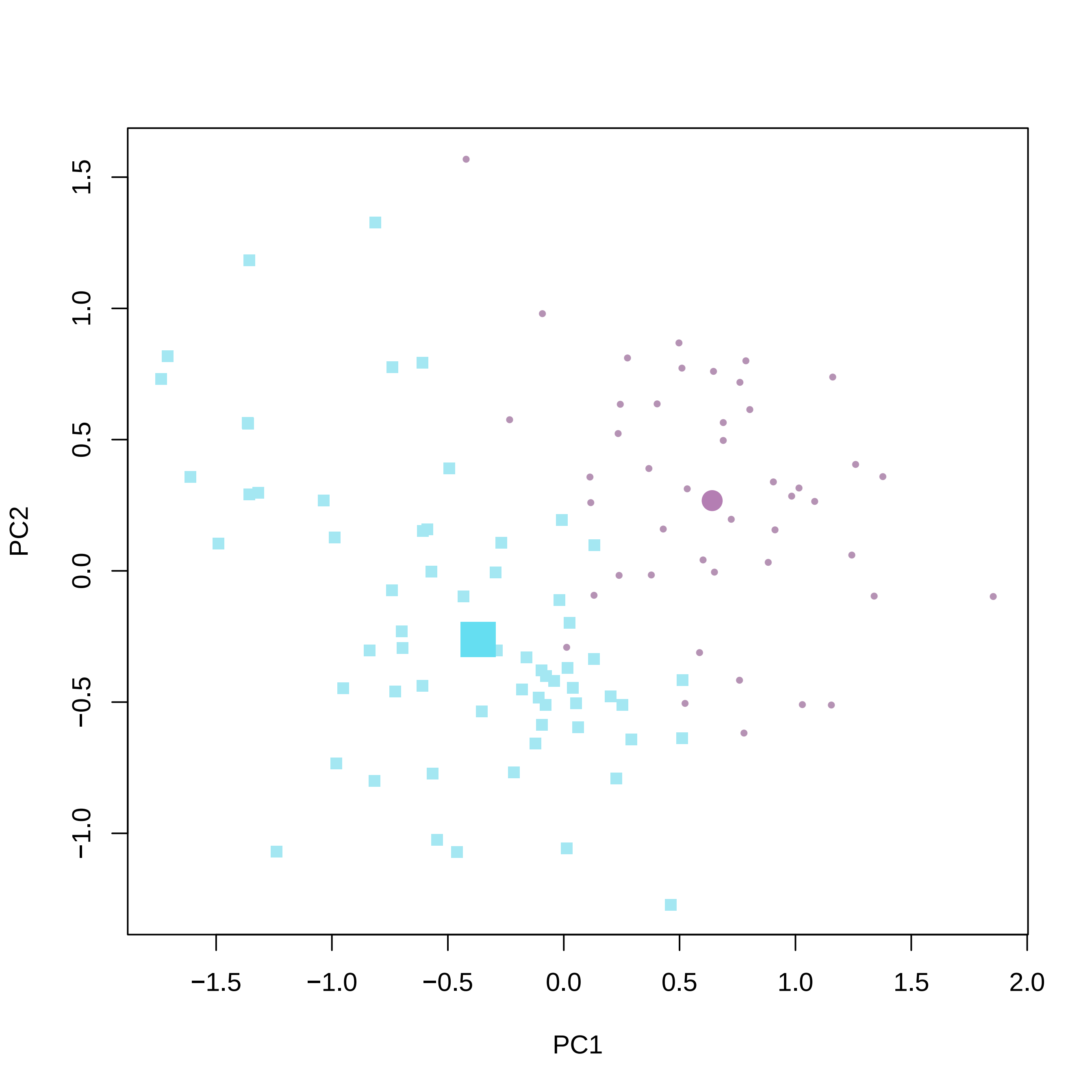
Cyan: outer halo



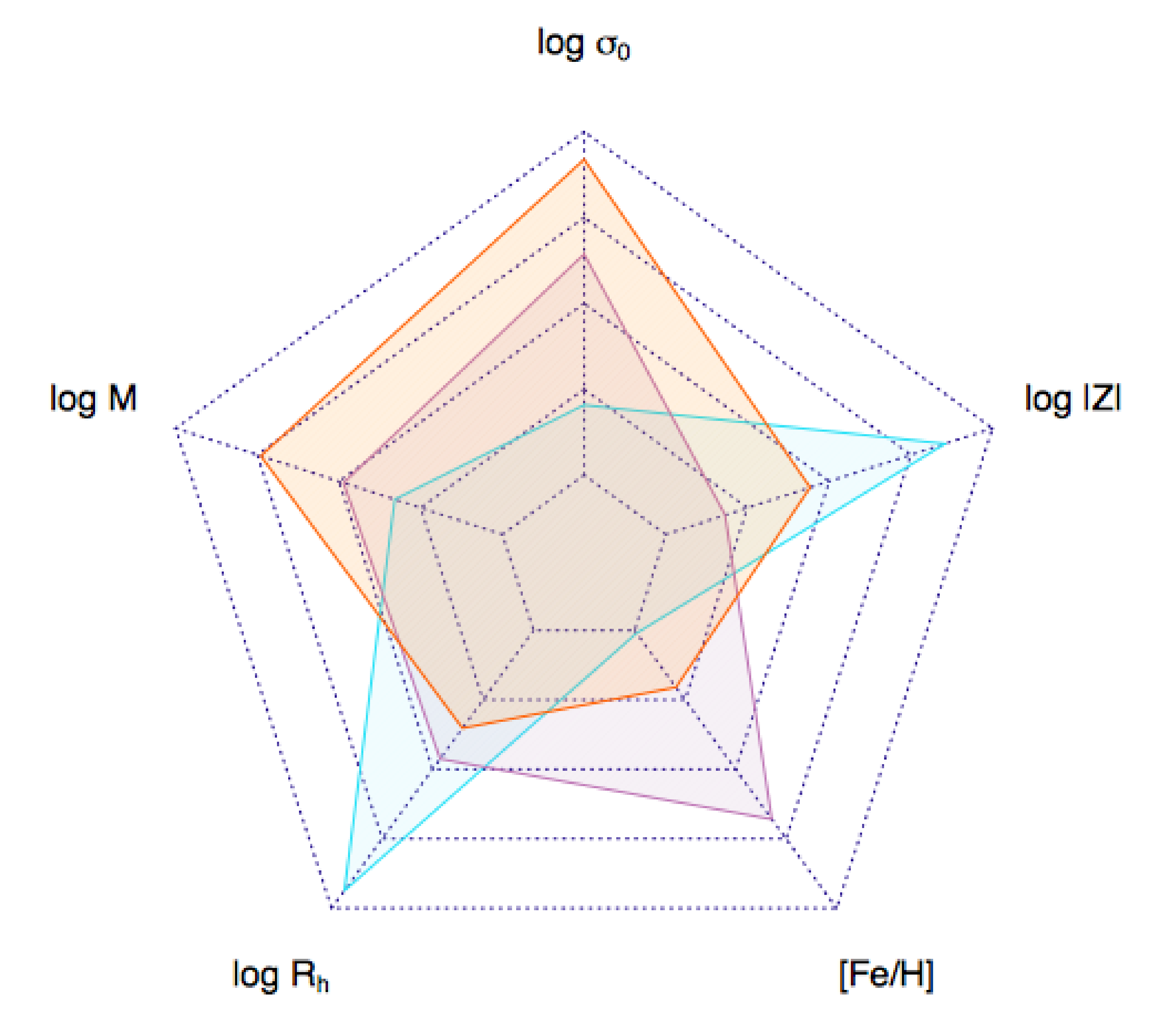
princip. comp. view
2 groups

Purple: disk

Cyan: halo



Who are the medoids?



Disk: NGC 6352

Inner halo: NGC 5986

Outer halo: NGC 5466

Notable correlations – relaxation time

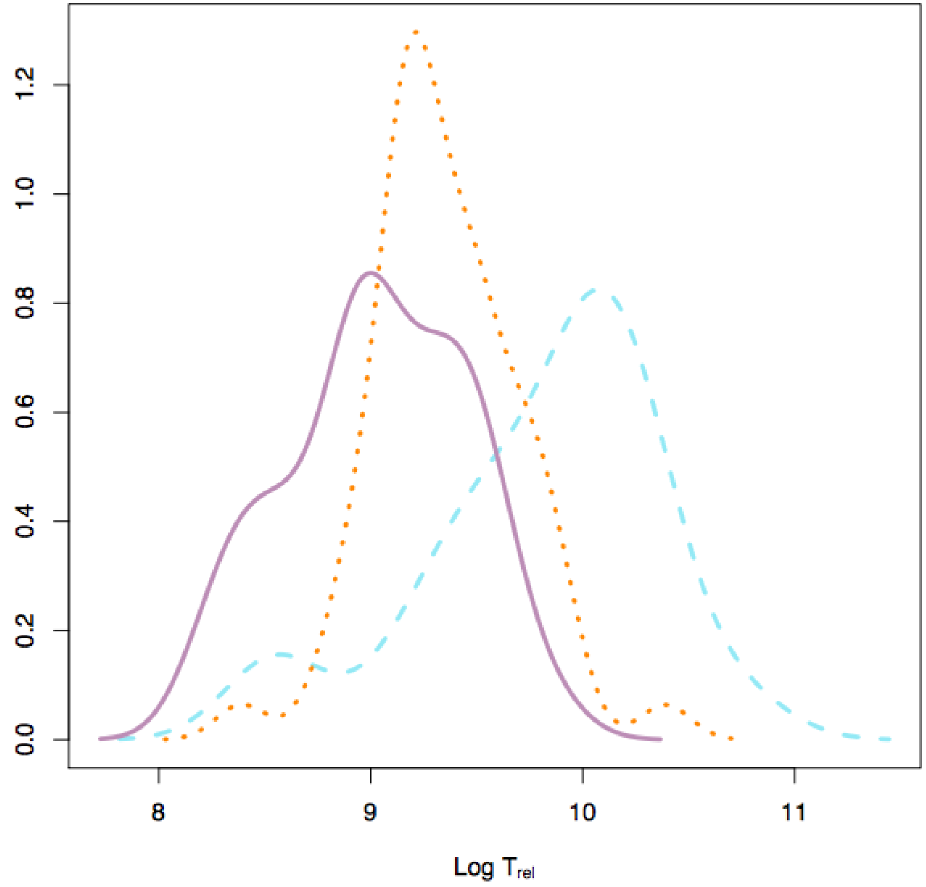


Figure 13. Log half-mass relaxation time distributions estimated with kernel density estimation for the three groups: disk (solid purple line), inner halo (dotted orange line), and outer halo (dashed light blue line).

Outer halo clusters are unrelaxed (dynamically young)

Disk clusters are more relaxed (dynamically older)

Inner halo are in-between

Notable correlations – MF slope

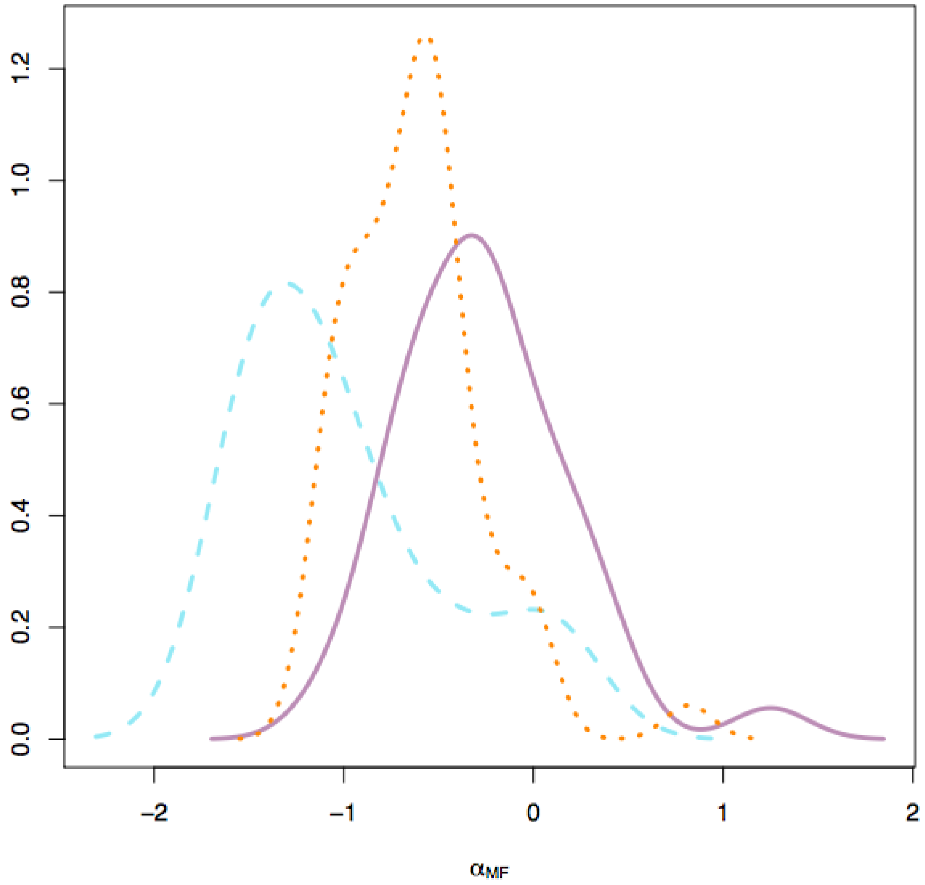


Figure 12. Mass function slope distributions estimated with kernel density estimation for the three groups: disk (solid purple line), inner halo (dotted orange line), and outer halo (dashed light blue line).

Outer halo clusters are less depleted of low mass stars (dynamically young)

Disk clusters are more depleted of low mass stars (dynamically older)

Inner halo are in-between

Unfortunately could not compare with Ferraro et al. 2012 relaxation cathegories based on BSS distribution, because almost all of those GCs are inner halo

Notable correlations – Age

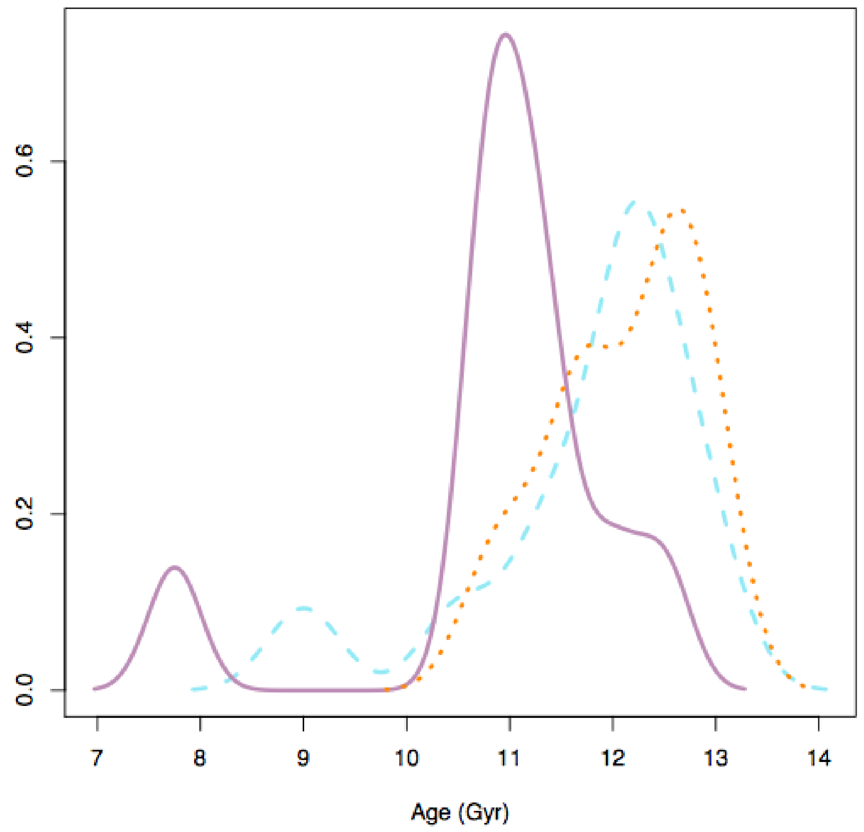


Figure 14. Age distributions estimated with kernel density estimation for the three groups: disk (solid purple line), inner halo (dotted orange line), and outer halo (dashed light blue line). The bumps at about 8 and 9 Gyr are due to Terzan 7 and Palomar 12, which are outliers in the distributions of their respective groups (disk and outer halo).

Outer halo and inner halo clusters are old

Disk clusters are young

Ages from Recio-Blanco 2018

Silhouette width VS age

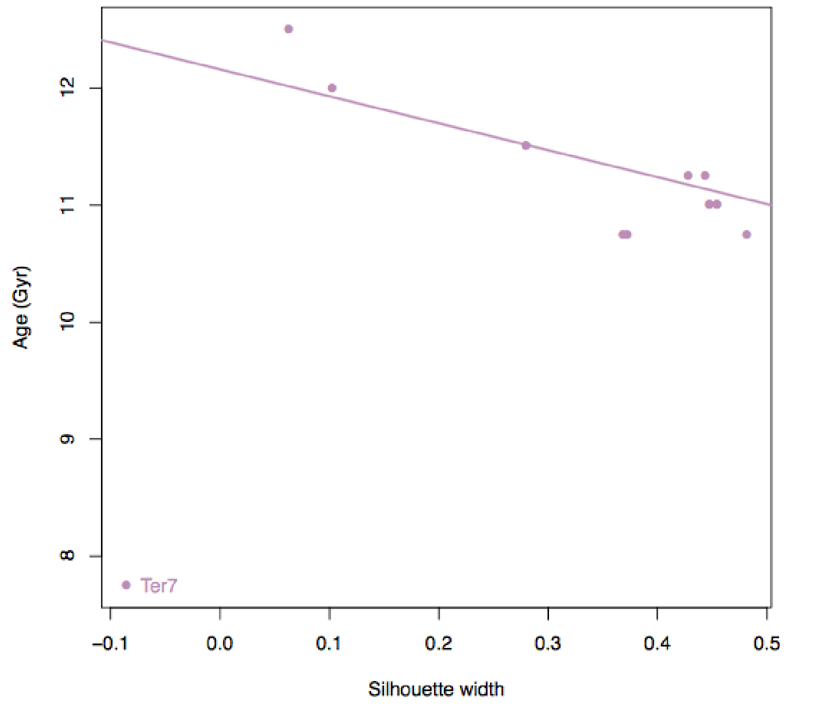


Figure 15. Correlation between silhouette width and age for disk GCs. GCs that are more strongly associated with the disk group (have a higher silhouette width) are younger, despite the fact that no information on age was used to determine the clustering and the relevant silhouette widths. The line is a robust linear fit, and one outlier (Terzan 7) is labeled.

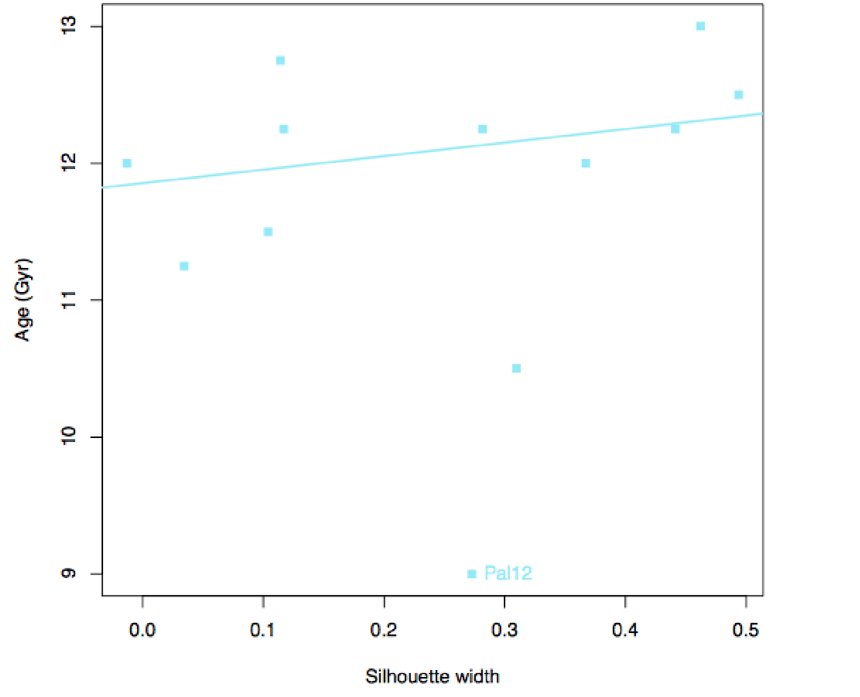


Figure 16. Correlation between silhouette width and age for outer halo GCs. GCs that are more strongly associated with the outer halo group (have a higher silhouette width) are older, despite the fact that no information on age was used to determine the clustering and the relevant silhouette widths. The line is a robust linear fit, and one outlier (Palomar 12) is labeled.

More firmly classified as disk = younger, more firmly classified as outer halo = older even though age was not used to determine the groups

End product

- 110 GCs classified into disk/inner halo/outer halo or disk/halo: ideal for statistical studies
- three (or two) medoids – representative GCs: ideal for case studies
- Silhouette widths: quantifier of the strenght of association of each GC with parent group: correlations with other parameters can be sought

https://arxiv.org/abs/1901.05354

Acknowledgment

“This project has received funding from the European Union’s Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 664931”

Project	Machine learning	Algorithm details	Data	Collab.	Paper
Finding IMBHs in GCs	Supervised on surface density profile features	SVM, NN (dense), RF, knn	MOCCA simulations	M. Mapelli A. Askar M. Giersz	Pasquato et al. 2019, MNRAS submitted
	Supervised on pulsar a, j, s	SVM	Direct N-body simulations	M. Spera F. Abbate	in prep.
Finding BH subsystems in GCs	Supervised on structural parameters	Dec. tree, GBT, SVM, knn Naive Bayes	MOCCA and N-body simulations GC catalogues	Ab. Askar Am. Askar	Askar et al. 2019, MNRAS submitted arXiv: 1811.06473
Clustering Clusters	Unsupervised	PAM	GC catalogues	C. Chung	Pasquato & Chung 2019, MNRAS submitted arxiv:1901.05354
Turbulence index in molecular clouds	Supervised on images	NN (convolutional)	RAMSES hydro simulations	P. Trevisan A. Ballone M. Mapelli	in prep.

IMBH = Intermediate Mass Black Hole (10^2 - 10^5 M_{sun}) **GC** = Globular Cluster (spherical star cluster, 10^{10} yr old, 10^6 stars) **MOCCA** = MOnte Carlo Cluster simulAtor (Fokker-Plank code; Giersz et al. 2013, moccacode.net) **Direct N-body** = Code that solves the equation of motion for all particles in a self-gravitarting system **RAMSES** = Adaptive Mesh Refinement code (Teyssier 2002, bitbucket.org/rteyssie/ramses) **SVM** = Support Vector Machines (Cortes & Vapnik 1995) **NN** = Neural Network **knn** = k-nearest neighbor (lazy learning) **RF** = Random Forest, **GBT** = Gradient Boosted Trees (Decision tree ensemble methods) **PAM** = Partitioning Around Medoids (clustering algorithm, Kaufman & Rousseeuw 1990) **Molecular cloud** = Cold (10 K) gas cloud, 10-10^2 parsec in size, will become a star cluster **Surface density profile** = Mass in visible stars per unit area as a function of distance from GC center (projected on the sky)