UNIVERSITY OF CAMBRIDGE

Executive Summary

Chemo-dynamical analysis of Milky Way's stellar populations with un-supervised multi-dimensional clustering

Michal Dorko (md2018)

Supervisor: GyuChul Myeong (gm564)

The Cavendish Laboratory: Department of Physics
University of Cambridge

June 27, 2024

Summary

Current understanding of the evolution history of Milky Way galaxy is based on Lambda-CDM (ΛCDM) cosmological model, which suggests multiple accretion and merger events contribute to the formation of the galaxies. Identifying structures within Milky Way galactic halo can help us explain evolutionary history of the galaxy and confirm validity of the current cosmological models. Stars in the Milky Way galactic halo have properties which make them suitable for analysis in chemical and dynamical space. Stars deposited in the galactic halo were suitable for this study as they preserve their orbital dynamics, in contrast to the stars in galactic disk which are affected by radial mixing. This study aims at reproducing results in the original work, which applied un-supervised machine learning method to the dataset containing subset of the stars in the galactic halo observable from our Solar neighbourhood [5].

Gaussian Mixture Model (GMM) clustering method, which falls under un-supervised machine learning methods, was applied to the pre-processed dataset without specifying any prior probability distribution. Dataset in the study was created by combining astrometric data from Gaia space observatory [4] with the chemical information provided by Apache Point Observatory Galactic Evolution Experiment (APOGEE) [1] and Galactic Archaeology with HERMES (GALAH) [3] surveys. Data was further filtered by applying set of quality cuts to obtain set of stars with suitable dynamical properties and chemical abundances. Processed APOGEE-Gaia dataset contained $\approx$ 1700 samples represented as 6-dimensional space with 5 chemical and 1 dynamical dimension. Similarly GALAH-Gaia dataset contained $\approx$ 1100 samples in 12-dimensional space. Since all astronomical measurements from the instruments mentioned above contain uncertainties, GMM fitting library has to be able to incorporate them into model. Extreme Deconvolution (XD) [2] fitting library was used for this purpose as it is designed specially for noisy and incomplete data, taking uncertainties into account when fitting the model. Bayesian Information Criterion metric was used to select the model best fitting the data.

BIC scores initially disagreed on the number of components present in the galactic halo. Due to lower resolution in the GALAH dataset (larger uncertainties on the measurements) regularisation term disfavoured all fits with more than 3 components. In contrast, model fit for APOGEE-Gaia dataset identified 7 components. However, after reconciliation of the difference and selecting number of components based on the existing knowledge, fits for both of the datasets, APOGEE-Gaia and GALAH-Gaia, identified 4 main components. Three of these components were previously known - GS/E, *Splash* and *Aurora* - in agreement with existing studies. The forth newly discovered component was named *Eos*. Chemical and dynamical properties match those reported in the original work [5] as well as previous studies of the galactic halo.

In addition to GMM fit, dimensionality reduction methods t-SNE and UMAP were applied to both datasets. Dimensionality reduction techniques embed high-dimensional data into low-dimensional space discarding information in the process. In case of datasets in this study both were embedded into 2-dimensional space. Both t-SNE and UMAP embeddings contained islands of samples corresponding to the main galactic halo components, in agreement with GMM fit. Agreement between GMM, t-SNE and UMAP further supports the theory of the origin of the galactic halo components. Additionally, a small island emerged in the 2D embedding among the cluster of points initially classified as residual "background" by GMM fit which is subject to further investigation.

References

[1] Abdurro'uf et al. "The Seventeenth Data Release of the Sloan Digital Sky Surveys: Complete Release of MaNGA, MaStar, and APOGEE-2 Data". In: *The Astrophysical Journal Supplement Series* 259.2 (Mar. 2022), p. 35. ISSN: 1538-4365. DOI: 10.3847/1538-4365/ac4414. URL: http://dx.doi.org/10.3847/1538-4365/ac4414.

[2] Jo Bovy, David W. Hogg, and Sam T. Roweis. "Extreme deconvolution: Inferring complete distribution functions from noisy, heterogeneous and incomplete observations". In: *The Annals of Applied Statistics* 5.2B (June 2011). ISSN: 1932-6157. DOI: 10.1214/10-aoas439. URL: http://dx.doi.org/10.1214/10-AOAS439.

[3] Sven Buder et al. "The GALAH+ survey: Third data release". In: *Monthly Notices of the Royal Astronomical Society* 506.1 (May 2021), pp. 150–201. ISSN: 1365-2966. DOI: 10.1093/mnras/stab1242. URL: http://dx.doi.org/10.1093/mnras/stab1242.

[4] Gaia Collaboration et al. "Gaia Early Data Release 3 - Summary of the contents and survey properties". In: *A&A* 649 (2021), A1. DOI: 10.1051/0004-6361/202039657. URL: https://doi.org/10.1051/0004-6361/202039657.

[5] G. C. Myeong et al. "Milky Way's Eccentric Constituents with Gaia, APOGEE, and GALAH". In: *The Astrophysical Journal* 938.1 (Oct. 2022), p. 21. ISSN: 1538-4357. DOI: 10.3847/1538-4357/ac8d68. URL: http://dx.doi.org/10.3847/1538-4357/ac8d68.