UNIVERSITY OF CAMBRIDGE

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 28, 2024

1 Introduction

Our Milky Way Galaxy is composed of the main galactic disk, galactic bulge, bar and surrounded by the galactic halo [19]. Large population of stars contained within the galactic halo are believed to be accreted into galaxy as a result of past galactic merger events. Analysis of the galactic halo has been of particular interest thanks to the capacity of the stellar halo to preserve orbital information of the stars deposited in the halo in contrast to the stars in the galactic disk which have their dynamical properties impacted by the effects radial mixing [22]. Additionally, with the availability of the new astrometric data from the Gaia space observatory [7] - in the phase space - combined with the chemical information provided by Apache Point Observatory Galactic Evolution Experiment (APOGEE) [1] and Galactic Archaeology with HERMES (GALAH) [5] it is possible to analyse the halo with combined chemical and dynamical information.

Gaia space observatory is an European Space Agency mission designed to collect high precision astrometric data of the stellar objects across the Milky Way galaxy. APOGEE survey provides high-resolution infrared spectroscopy data of more than 100,000 red giant stars across galactic bulge, disk and halo producing detailed chemical abundances [1]. Equivalently, GALAH survey produced dataset containing chemical abundance of 588,571 nearby stars [5]. Combining chemical analysis from both of these surveys with kinematics measured by Gaia mission, we can obtain chemo-dynamical information of individual stars across the Galaxy, providing valuable insight into the evolution history of the Milky Way galaxy.

Analysis in chemo-dynamical space with classical machine learning technique of fitting a mixture of Gaussian distributions into the data (un-supervised learning) has been the aim of the recent publication [17]. Findings described in the publication describe four independent components, three of which were previously known - *Aurora*, *Splash* and Gaia-Sausage/Enceladus (GS/E) - and one new component *Eos*. The aim of this data analysis project is to reproduce the findings described in the publication [17].

The extended stellar halo is composed of the remnants of past accreted systems, which are detected in Solar system neighourhood [16]. As for the massive galaxy accretion, only one such merger event has been identified from the vast debris structure left behind, now known as Gaia Sausage/Enceladus (GS/E) but predicted before Gaia [6]. It is hypothesised that the progenitor galaxy was massive enough for its orbit to radialise as a result of complex interaction of the dynamical deceleration, host recoiling and friction resulting in the deposition of the stars in the Milky Way in its last apo-centre at around 30 kpc. As a result, the region of Galactic halo within this break radius is dominated by GS/E stars as was confirmed by *Gaia* data [17], unique chemical fingerprint [8] and large group of globular clusters [16]. Number of estimates using different methods - local estimate [3] and global estimate [10] - show GS/E merger contributing almost as much as half of the total halo mass. Timelines for the galactic accretion place GS/E merger at around

8-10 Gyr in agreement with models based on White Dwarf cooling as well as main-sequence-turn-off stars [17]. GS/E merger is however not the only significant merger event Milky Way galaxy experienced. Remnants of more recent accretion events may be present at the outskirts of the galaxy similar to two recent and massive accretion events: that of the Sagittarius dwarf galaxy [9] and the of the Magellanic Clouds [13]. Additionally debris from several smaller systems eg. Sequoia [16] has been found.

Motivation of the analysis presented in the recent study [17] is to provide unbiased decomposition of the stellar halo in the Solar neighbourhood. This can provide insight into evolution of the Milky Way galaxy and confirm current assumptions stemming from the Lambda-CDM (ΛCDM) cosmological model suggesting multiple accretion and merger events contribute to the formation of the galaxies. Goal of the study is to identify and characterise all significant and independent halo components without any prior using Gaussian Mixture model. Population of the stars included in the analysis include only stars on the orbits with high eccentricity to exclude majority of thin disc stars which otherwise could dominate our analysis [17]. However not all stars with high eccentricity have been accreted. Some of the in-situ stars with the [Fe/H] > -1 dex have been born inside the Milky Way and scattered into galactic halo through the interaction with the GS/E progenitor.

Elemental abundances were one of the main features, in addition to dynamical information, used in the study for the model fitting purposes. Metallicity (elemental abundance) of the star is defined as an abundance of the elements heavier than hydrogen and helium in the star. These elements were created by the process of nucleosynthesis in the star through the process of nuclear fusion. Elemental abundance provides information about the stellar origin and the age. Elements are assigned into groups by the "tracers" (type of stars producing these elements throughout the lifetime) and the timescales it takes for the elements to be created. Table 1 shows group of abundances with example elements and tracer stars.

Group	Tracer	Mechanism of entry into interstellar region	Timescale	Example Elements
α (alpha)	High mass stars $M > 8M_{\odot}$	Core collapse supernovae	0 - 100 Myr	O, Na, Mg, Al, Si, Ca, Ti
Iron Peak	Low mass stars $M < 8M_{\odot}$	Type Ia supernovae	100 Myr - 1 Gyr	Ti, V, Cr, Mn, Fe, Co, Ni, Cu, Zn
Iron	Low and high mass stars	CC SNe, Type Ia SNe	0 - 100 Myr 100 Myr - 1 Gyr	Fe
s-Process	Low mass stars $1M_{\odot} < M < 3M_{\odot}$	Winds during asymptotic giant branch phase	300 Myr - 5 Gyr	Sr, Y, Zr, Ba, La, Ce, Nd
r-process	High mass stars $8M_{\odot} < M < 22M_{\odot}$	CC SNe, Neuron star merger	0 - 100 Myr 50 Myr - 14 Gyr	Nd, Eu, Th, U

Table 1: Elemental abundances [23]

Abundances are measured by relative fraction and not by explicit mass. [X/H] denotes logarithmic abundance ratio of element X relative to hydrogen and is defined as:

$$\left[\frac{X}{H}\right] = \log_{10}\left(\frac{N_X}{N_H}\right)_* - \log_{10}\left(\frac{N_X}{N_H}\right)_{\odot} \tag{1}$$

where $*$ symbol indicates target star and $\odot$ symbol our Sun. They are placed on "log 12" scale which is calibrated such that $\log_{10}\left(N_{H\odot}\right) = 12$.

Careful choice of the chemical abundances used in the study is motivated by representation of the elements groups following different tracers. Different tracer groups offer the perspective of the chemical evolution of the galaxy directly translating to the timescales of the evolution. For example clusters with the stars which are rich in r-process elements have longer star formation history as production of these elements takes the longest out of all tracer groups as shown in table 1. This understanding of the past helps us form the view of the evolutionary processes as well as current state of the galaxy.

Methodology applied in this study incorporates machine learning method categorised as "unsupervised learning" where the true distribution of the data is not known (labels for the samples are not known). Unsupervised learning is a set of statistical tools which attempt to discover the

unknown structure in the data by identifying unknown subgroups. These tools are often used as a mean of *exploratory data analysis*. Clustering methods work by partitioning data into clusters (subgroups) so that the data points in the same cluster are similar to each other while data points in different groups are as different as possible [11]. Clustering methods are especially suitable for application to the astronomical observations where objects naturally form clusters as a result of physical interactions between them. This study uses Gaussian Mixture Model clustering method which assumes that the objects are distributed following Gaussian distribution where overall structure of the samples is composed of the mixture of Gaussian distributions. It is important to note however, that physical processes involved in the galactic evolution do not result in perfect Gaussian distribution of the stars based on their chemo-dynamical properties therefore applying Gaussian distribution is only an approximation and some of the galactic halo components might not follow this distribution.

This report is structured as follows. Section 2 (Data) describes pre-processing and filtering applied to the baseline dataset in order to obtain clean data used for GMM fitting. Section 3 (Methodology) talks about Gaussian Mixture Model fitting, model evaluation criteria and assignment of the stars to the halo components. Section 4 (Halo Constituents) presents and discusses obtained results. Lastly section 5 (Dimensionality reduction) covers extension to the original work presenting analysis of results obtained from dimensionality reduction methods.

2 Data

Main and Value Added Catalogue (VAC) of APOGEE Data Release (DR) 17 [1] which provides stellar parameters, radial velocity and abundances for up to 20 chemical species for 372,458 unique targets (main red star sample) was used for the study. It was cross-matched with Gaia EDR3 [5] which was used for calculating orbital dynamics of the stars using galaxy modelling library AGAMA [24].

APOGEE samples were filtered by applying various quality cuts. Stars with `STAR_BAD` flag in `ASPCAPFLAG`, and `PROGRAMNAME = magclouds` were rejected from the filtered dataset. Only the main red star samples with `EXTRATARG = 0` were used. For a red giant star samples `logg < 3.0` was applied. Data was further filtered with `x_fe_flag = 0`, `x_fe_err < 0.1 dex` for almost all elements with the exception of [Ce/Fe] for which `ce_fe_err < 0.15 dex` was used instead of `0.1`. To obtain the halo or halo-like Galactic components such as the GS/E and the Splash an `eccentricity > 0.85` cut was adopted to obtain stars on nearly radial orbits. To minimise contamination from the bulge populations apocenter $>$ 5kpc was used. Lastly the distance error $<$ 1.5 kpc and orbital energy $E < 0$ km^{2}s^{-2} were applied [17]. Total of $\approx$ 1700 APOGEE stars passed the quality cuts.

For the purpose of the unsupervised clustering, 5 dimensions of elemental abundances [Fe/H], [α/Fe], [Al/Fe], [Ce/Fe], [Mg/Mn], and the orbital energy (E) information as 6th dimension was fed into Gaussian Mixture Model (GMM). 6-dimensional space was motivated by the desire to include dynamical properties (E) as well as the "most reliable" (or "reliable" if necessary) elemental production channels for the categories of odd-Z (Al), α (α, Mg), iron-peak (Mn), and s-process (Ce) elements, following APOGEE's recommendation for giant stars [17].

APOGEE elemental abundances in the base catalogue are provided in the standard format, relative to hydrogen ([X/H]) and iron ([X/Fe]) as they are good reference elements. Number of other useful combinations of elemental abundances knows to be effective for distinguishing the stellar populations with different origin were derived. Namely [Mg/Mn] and $[\alpha/Fe]$ had to be derived from the measurements provided in the base catalogue. Since abundance quantities are in logarithmic scale as shown in equation 1, in order to derive [Mg/Mn] quantity, simple formula $[Mg/Mn] = [Mg/Fe] - [Mn/Fe]$ was applied to obtain this derived quantity. Each quantity is however associated with measurement uncertainty which needed to be propagated to the derived quantity. For the error propagation, simple error propagation formula of $\sqrt{[Mg/Fe]_{err}{}^2 + [Mn/Fe]_{err}{}^2}$ was applied. Equivalently derived quantity [α/Fe] was derived from abundances present in APOGEE catalogue $[\alpha/Fe] = [\alpha/M] + [M/H] - [Fe/H]$ with error propagation as $\sqrt{[\alpha/Fe]_{err}{}^2 + [\alpha/M]_{err}{}^2 + [Fe/H]_{err}{}^2}$.

Main and VAC data of GALAH DR3 [5] release 678,423 spectra of 588,571 mostly nearby stars was used as a cross-check. The catalogues provide stellar parameters, radial velocity, and abundances for up to 30 chemical species. Recommendations from the GALAH Collaboration for the choice of columns and quality cuts were applied to the data samples. Samples were first filtered with `snr_c3_iraf > 30` and `flag_sp = 0`. Similar cuts from APOGEE were applied with

the exception of using `e_x_fe < 0.2` dex for elements `x` in the dataset [17].

For the GALAH-Gaia GMM fit, 12-dimensional chemo-dynamical space containing ≈ 1100 stars passing above quality cuts were included in the dataset. Abundances used were [Fe/H], [α/Fe], [Na/Fe], [Al/Fe], [Mn/Fe], [Y/Fe], [Ba/Fe], [Eu/Fe], [Mg/Cu], [Mg/Mn], [Ba/Eu], in addition to energy E. This includes similar tracers as the APOGEE sample, plus the r-process element (Eu) available in GALAH [17]. Similarly to APOGEE dataset, number of quantities used for the model fitting were derived from the abundances provided by the GALAH catalogue. Abundance [Mg/Cu] was derived from [Mg/Fe] and [Cu/Fe], [Mg/Mn] from [Mg/Fe] and [Mn/Fe], [Ba/Eu] from [Ba/Fe] and [Eu/Fe]. Same error propagation method as applied to derived quantities in case of APOGEE was used for errors on measurements for GALAH case.

3 Clustering methodology

A Gaussian mixture model is a probabilistic model that assumes all the data points are generated from a mixture of a finite number of Gaussian distributions with unknown parameters. One can think of mixture models as generalizing k-means clustering to incorporate information about the covariance structure of the data as well as the centers of the latent Gaussians [21]. The optimal (maximum likelihood) model parameters can be found using Expectation-maximization algorithm (EM), an iterative method for finding maximum likelihood estimate of parameters in statistical models [15]. Analysis discussed in this report used `EXTREME DECONVOLUTION` (XD) [4] algorithm. One of the main advantages of XD is the ability to incorporate uncertainties associated with the observed data. In case of astronomical data uncertainties on the measurements are heteroscedastic but well defined for each observation resulting in low signal-to-noise data. In such scenario uncertainties have to be incorporated into the model instead of subtracting them from the data. XD achieves this by computing likelihood for each data point individually such that model likelihood is convolved with unique uncertainty distribution of the data point. Final likelihood of the data set is then obtained by simply multiplying obtained individual likelihoods.

XD algorithm requires initial values for the parameters of the multivariate Gaussian distribution as a starting point for the iteration - means, covariance matrix and amplitudes. Initial values for the means of each of the features were initialised to one of the data points in the dataset - this ensures that the mean of the distribution in corresponding dimension is within the range of the values of the features. Amplitudes were randomly initialised using Dirichlet distribution which has crucial property that all of the values drawn from the distribution sum to 1 ensuring that the sum of initial amplitudes is also 1. Covariance matrix was simply initialised to a diagonal matrix of 1s. Sensitivity of XD to the initialization was observed during development manifesting itself as instability of the solution when measured with different metrics.

Main metric used to evaluate model fit was Bayesian Information Criterion (BIC) [20]. BIC score was used to measure performance of the model while preventing over-fitting. When fitting a mixture of distributions, the best fit would simply be the one with single distribution for every data point in the dataset. Such fit is unrealistic therefore good metrics specify regularisation term penalising the complexity of the model. The ability of the BIC score to prevent model over-fitting is achieved through penalty term as defined in equation 2 [25] where k corresponds to the number of the parameters.

$$\mathrm{BIC}(\mathcal{M}) = -2\mathcal{L}(\hat{\Theta}) + k\ ln(n) \tag{2}$$

For a mixture of Gaussian distributions, term k is directly proportional to the number of components (N_{comp}) fit into the data as well as dimensionality (p) of the feature space with k being defined as:

$$N_{covar} = N_{comp}\ \frac{p\ (p-1)}{2} \tag{3}$$

$$N_{mean} = p\ N_{comp} \tag{4}$$

$$k = N_{covar} + N_{mean} + N_{comp} - 1 \tag{5}$$

Alternative metric commonly used is Akaike Information Criterion (AIC) [2] with similar properties as BIC including penalty term for number of parameters. However AIC score is defined as:

$$AIC(\mathcal{M}) = -2\mathcal{L}(\hat{\Theta}) + 2\ k \tag{6}$$

Penalty term 2 k is more tolerant of the higher number of the free parameters which, in case of the GMM fit for the two datasets used in this study, wasn't sufficient to penalise the model for overfitting. When using AIC score to evaluate the goodness of fit, metric had always decreasing trend as more and more components were fitted into the data without any signs of convergence. Consequently this metric had very low informational value in assessing number of components present in the dataset and was not used for the datasets in this study.

Likelihood surface in the higher dimension can have multiple local maxima therefore finding the global maximum requires exploration of the large part of the surface. To achieve that, at each iteration, randomly initialised parameters were passed to the model fitting algorithm (XD) ensuring EM algorithm starts at a different starting point on the likelihood surface. This approach maximises coverage of the feature space resulting in likelihood surface being more broadly explored and reducing risk of algorithm converging to local maximum in each run. Consequently, fitting procedure ran 100 iterations for each number of components, computing BIC score for each fit. Listing below shows pseudo-code outlining fitting and results collection loop:

Listing 1: Pseudocode outlining GMM fitting loop

```
run_xd(features, uncertainties):
    err_covar = construct_covariance(uncertainties)
    fitted_params = list()
    bics = list() #Bayesian Information Criterion
    for num_of_components in range 1..10: # try fitting up to 10 components
        for fit in range 1..100: # 100 fits for each component
            amplitudes, means, covars = generate_initial_guesses(features,
                num_of_components)
            likelihood = extreme_deconvolution(features, err_covar, amplitudes, means,
                covars) # amp, covar and mean modified in place by XD
            bic = bayesian_infromation_criterion(likelihood, num_of_components,
                number_of_features, number_of_samples)
            bics.append(bic, num_of_components)
            fitted_params.append(amplitudes, means, covars, num_of_components)

    return bics, fitted_params
```

For each number of components, minimum value of the BIC score was chosen out of the 100 fits. Best fit was minimum of minimums across all fitted number of components. Figure 1 shows BIC score for the APOGEE dataset.

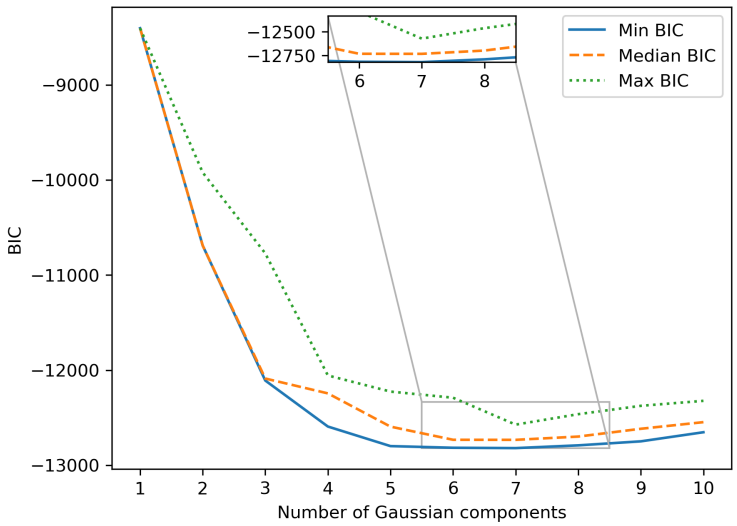


Figure 1: BIC score for the APOGEE dataset

Results indicates that the best fit is between 5 to 7 components. Scores for these components are comparable, however 7 components scored the lowest score in the majority of the runs. Median and max BIC scores for each number of the components confirm the same downward trend towards 6-7 components. All fits between 5-7 components capture major galactic halo components (GS/E, *Aurora*, *Splash*) and newly discovered *Eos* component. The difference between 5-7 component

fits was sensitivity to the background (single low weight background component vs 2 background components) and the algorithm finding sub-structures within main components (dividing GS/E into two similar sub-clusters).

Similarly to APOGEE-Gaia dataset, GMM fitting was applied to the GALAH dataset as a cross-check. BIC scores for the GALAH dataset are presented in the figure 2. It is immediately obvious that the GMM fit for the GALAH dataset produced significantly different results - only capturing 3 major components of the galactic halo. There is a very strong trend illustrated by all of the minimum, median and maximum BIC scores with the line plot initially decreasing until a sharp bend at 3 components and rapidly increasing score for subsequent components. Despite higher dimensionality of the dataset (12 dimensions) measurements provided by the GALAH instrument have higher uncertainties associated with the measurements resulting in lower resolution measurement compared to APOGEE measurements. Taking into consideration lower resolution of the data, likelihood of the GMM fit improves significantly less after 3 components fit. Since BIC score penalises more complex models with higher number of parameters it is disfavouring more granular model with larger number of components (parameters) as is the case of the GALAH-Gaia dataset despite having higher likelihood. As is discussed in the later chapter, fitting the model with 5 components for GALAH-Gaia dataset using prior knowledge shows agreement with APOGEE-Gaia when comparing 2-dimensional projections of the features present in both datasets. It is important to note here that solely relying on the standard evaluation metric is often not sufficient and the results should be evaluated in the context of the data. Assessing results obtained from unsupervised clustering methods is one of the challenges posed by these techniques as there are no universally accepted mechanisms for performing any cross-validation [11].

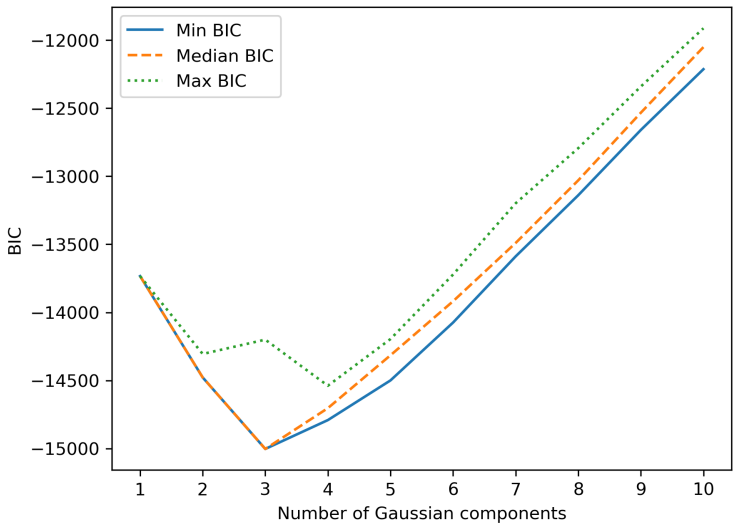


Figure 2: BIC score for the GALAH dataset

Membership of the each individual star in the cluster component was established with the help of `GaussianMixture` implementation in the *scikit-learn* library [18]. Since XD algorithm only provides likelihood value together with updated values for amplitude, mean and covariance for each component, assignment of the stars into the clusters was achieved by evaluating probability density function (Gaussian distribution) of the stars with respect to each individual component. As a result of the fit, each star has certain probability of being member of each of the clusters. Stars would then be assigned to the cluster with the highest probability. `GaussianMixture` implementation of *scikit-learn* encapsulates this calculation. Values for means, amplitudes and covariances as computed by XD were retrofitted into `GaussianMixture` class which was then subsequently used to establish cluster membership. The main reason for not fitting model to the data with the *scikit-learn* implementation in the first place was that its implementation does not take uncertainties of the measurements into consideration. As a result, combination of XD to obtain parameters of the underlying Gaussian distributions and *scikit-learn* to assign stars to clusters was used to establish components of the galactic halo.

4 Constituents of the galactic halo

In the agreement with the publication [17] GMM modelling technique was able to identify chemical signatures of the three known components of the Galaxy - GS/E, *Splash* and *Aurora* - which were previously identified by selecting samples with the application of hand made cuts to the stellar halo dataset.

Table 2 presents parameters describing weight, the mean and standard deviation for each GMM component in the APOGEE-Gaia dataset. The value for the combined GS/E which was divided into two substructures by the GMM fit is marked as GS/E_{sum}. Equally the sum of the two components associated with the background ($back_1$ and $back_2$) is marked $back_{sum}$ for a reference.

Component	Weight (%)	Count	Energy ($10^5 km^2 s^{-2}$)	[Fe/H]	$[\alpha/Fe]$	[Ce/Fe]	[Al/Fe]	[Mg/Mn]
GS/E_1	27.59	470	-1.51 ± 0.14	-1.05 ± 0.20	0.16 ± 0.05	-0.22 ± 0.06	-0.14 ± 0.07	0.43 ± 0.09
GS/E_2	23.09	393	-1.51 ± 0.19	-1.38 ± 0.19	0.20 ± 0.03	-0.25 ± 0.06	-0.22 ± 0.15	0.57 ± 0.14
Splash	25.19	429	-1.76 ± 0.12	-0.61 ± 0.16	0.29 ± 0.02	0.25 ± 0.06	-0.19 ± 0.12	0.49 ± 0.07
Aurora	5.67	96	-1.76 ± 0.15	-1.19 ± 0.15	0.32 ± 0.05	0.13 ± 0.19	-0.05 ± 0.19	0.64 ± 0.09
Eos	9.63	164	-1.66 ± 0.12	-0.65 ± 0.21	0.14 ± 0.05	0.03 ± 0.13	-0.15 ± 0.07	0.31 ± 0.08
$back_1$	5.90	100	-1.60 ± 0.18	-1.17 ± 0.33	0.17 ± 0.09	-0.17 ± 0.22	0.33 ± 0.63	0.44 ± 0.18
$back_2$	2.92	49	-1.60 ± 0.25	-0.92 ± 0.66	0.14 ± 0.06	0.20 ± 0.33	-0.01 ± 0.25	0.29 ± 0.22
GS/E_{sum}	50.68	863	-1.51 ± 0.16	-1.21 ± 0.20	0.18 ± 0.04	-0.23 ± 0.06	-0.18 ± 0.11	0.50 ± 0.11
$back_{sum}$	8.82	149	-1.60 ± 0.21	-1.05 ± 0.49	0.15 ± 0.08	0.01 ± 0.28	0.19 ± 0.44	0.37 ± 0.20

Table 2: Summary of GMM fit result for APOGEE-Gaia dataset

Component	Weight (%)	Count	Energy ($10^5 km^2 s^{-2}$)	[Fe/H]	$[\alpha/Fe]$	[Na/Fe]	[Al/Fe]
GS/E	41.58	453	-1.58 ± 0.18	-0.97 ± 0.18	0.12 ± 0.07	-0.25 ± 0.10	-0.15 ± 0.10
Splash/back	45.54	496	-1.81 ± 0.12	-0.64 ± 0.17	0.23 ± 0.08	0.06 ± 0.10	0.24 ± 0.14
Rest/back	12.88	140	-1.67 ± 0.20	-0.95 ± 0.28	0.19 ± 0.13	-0.04 ± 0.25	0.13 ± 0.28

[Mn/Fe]	[Y/Fe]	[Ba/Fe]	[Eu/Fe]	[Mg/Cu]	[Mg/Mn]	[Ba/Eu]
-0.36 ± 0.09	0.09 ± 0.10	0.37 ± 0.18	0.45 ± 0.11	0.61 ± 0.07	0.48 ± 0.12	-0.08 ± 0.21
-0.19 ± 0.08	0.10 ± 0.20	0.23 ± 0.24	0.28 ± 0.08	0.30 ± 0.14	0.45 ± 0.14	-0.05 ± 0.27
-0.32 ± 0.18	0.43 ± 0.43	0.77 ± 0.50	0.42 ± 0.22	0.54 ± 0.18	0.52 ± 0.18	0.35 ± 0.46

Table 3: Summary of GMM fit result with 3 components for Galah dataset

Component	Weight (%)	Count	Energy ($10^5 km^2 s^{-2}$)	[Fe/H]	[α/Fe]	[Na/Fe]	[Al/Fe]
GS/E	33.10	360	-1.53 ± 0.15	-0.96 ± 0.17	0.11 ± 0.06	-0.27 ± 0.10	-0.18 ± 0.09
Splash	31.15	339	-1.83 ± 0.10	-0.61 ± 0.13	0.27 ± 0.04	0.11 ± 0.07	0.31 ± 0.09
Aurora	12.36	134	-1.80 ± 0.15	-1.09 ± 0.11	0.23 ± 0.07	-0.12 ± 0.10	0.09 ± 0.16
Eos	14.91	162	-1.70 ± 0.14	-0.67 ± 0.18	0.11 ± 0.05	-0.07 ± 0.10	0.02 ± 0.14
Back	8.48	92	-1.67 ± 0.20	-0.89 ± 0.32	0.19 ± 0.14	0.03 ± 0.25	0.14 ± 0.30

[Mn/Fe]	[Y/Fe]	[Ba/Fe]	[Eu/Fe]	[Mg/Cu]	[Mg/Mn]	[Ba/Eu]
-0.36 ± 0.09	0.07 ± 0.08	0.39 ± 0.19	0.47 ± 0.10	0.61 ± 0.06	0.47 ± 0.11	-0.09 ± 0.22
-0.18 ± 0.06	0.07 ± 0.18	0.17 ± 0.23	0.28 ± 0.07	0.30 ± 0.11	0.49 ± 0.08	-0.11 ± 0.26
-0.38 ± 0.08	0.32 ± 0.22	0.46 ± 0.25	0.32 ± 0.11	0.60 ± 0.10	0.63 ± 0.07	0.15 ± 0.28
-0.21 ± 0.09	0.10 ± 0.16	0.35 ± 0.20	0.32 ± 0.13	0.31 ± 0.19	0.32 ± 0.12	0.03 ± 0.22
-0.29 ± 0.20	0.46 ± 0.50	0.84 ± 0.57	0.47 ± 0.23	0.51 ± 0.21	0.47 ± 0.20	0.37 ± 0.51

Table 4: Summary of GMM fit result with 5 distinct components for Galah dataset

Figure 3 shows number of 2-dimensional projections of the original 6-dimensional chemo-dynamical space. Best fit model contains 7 Gaussian components shown with 2σ ellipses. GMM fitting using XD algorithm was also applied to the GALAH-*Gaia* dataset. GALAH-*Gaia* contained additional chemical dimensions - for the odd-Z and iron-peak elements Na and Cu were included in addition to abundances present in the APOGEE-*Gaia* sample. For the s-process, Y and Ba are used instead of Ce and for the r-process Eu is included. APOGEE-*Gaia* sample does not contain r-process elements which makes this additional dimension very valuable [17]. Table 3 presents parameters of the GALAH-*Gaia* dataset with 3 components with 2-dimensional projections from the 12-dimensional space shown in figures 4 and 5. Similarly for the comparison with the original publication [17] table 4 shows 5 component summary and figures 6, 7 show 2-dimensional abundance projections.

The GMM for the APOGEE-*Gaia* dataset clearly identified 4 strong components, three out of which were previously known constituents (*GS/E*, *Splash*, *Aurora*) with the 4th component being new population identified as *Eos* [17]. Algorithm was sensitive enough to find substructure within GS/E component which was divided into two substructures. This can be attributed to the fact that the dataset clearly captures significant evolutionary stages of GS/E such as the SN Ia onset, which alters the main channel of chemical enrichment in the environment. As a result, evolutionary trend of GS/E is altered - especially [Fe/H] - which can not be covered with a single Gaussian distribution. Similarly for the background, two different structures which are treated as a single background were identified.

When comparing APOGEE-*Gaia* fit to GALAH-Gaia two comparisons need to be made with respect to the different fits - three component fit favoured by BIC score and five component fit which is based on some prior knowledge. Three components fit clearly identified GS/E component which agrees with APOGEE-*Gaia* in all dimensions when error is taken into consideration. Most of the fitted GS/E dimensions agree with APOGEE dataset, taking into consideration the error in the region of 10^{-2}. The exception is the [Fe/H] dimension where the difference between APOGEE and GALAH fit is slightly higher. Second component identified in this fit mostly captures *Splash* and part of the background. This is the most obvious from [Al/Fe] abundance with APOGEE-*Gaia* fit attributing mean of -0.19 ± 0.12 compared to 0.24 ± 0.14 from GALAH indicating additional contribution from the different component(s). Final component (marked as *Rest/back*) is the residuals and the background components similar to the case in APOGEE-*Gaia* sample.

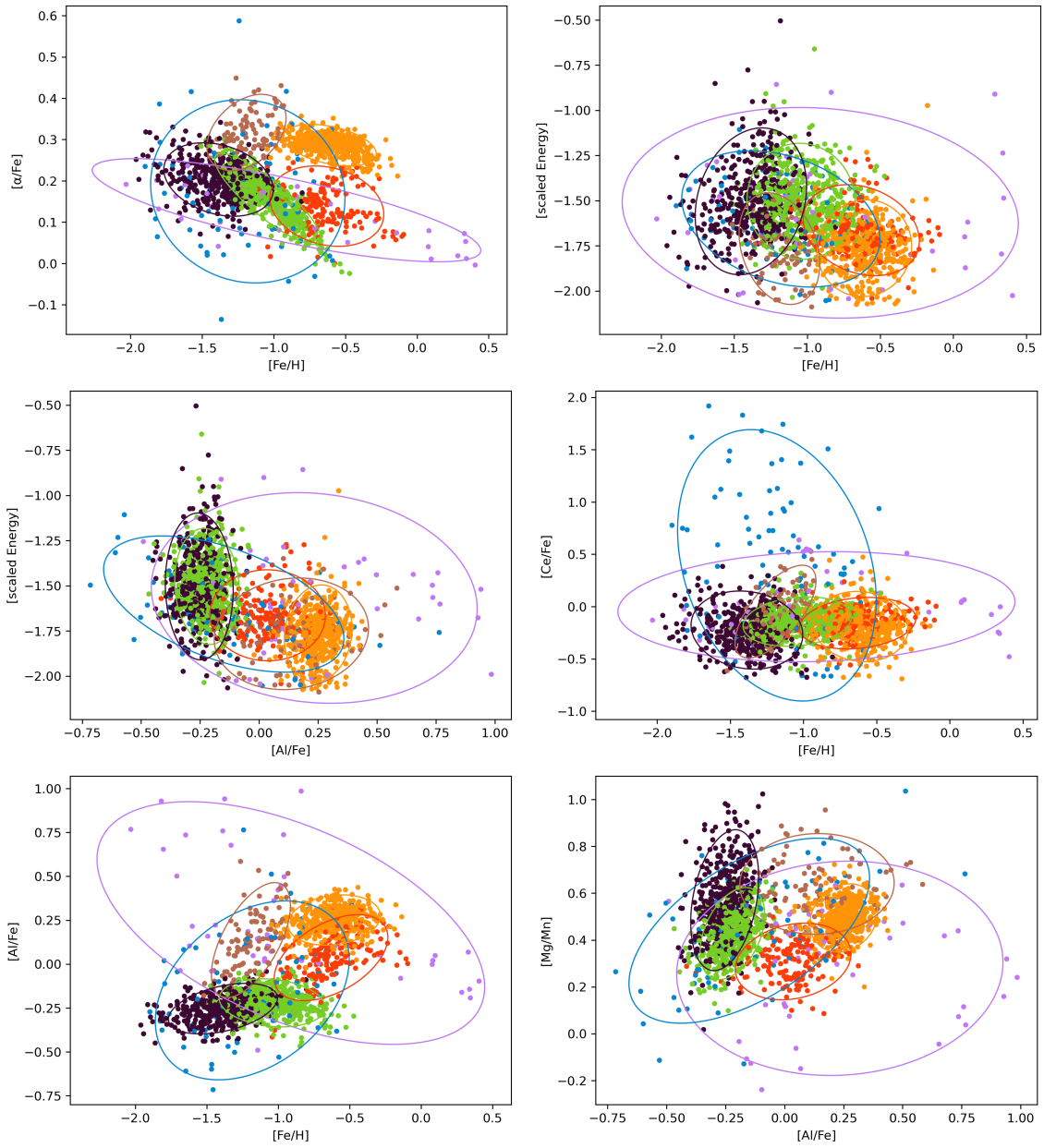


Figure 3: Distribution of Galaxy's eccentric populations in APOGEE-Gaia sample in chemo-dynamical spaces. Dark purple colour marks GS/E_1, green colour GS/E_2, orange *Splash*, brown *Aurora*, red *Eos*, pink and blue $back_1$ and $back_2$

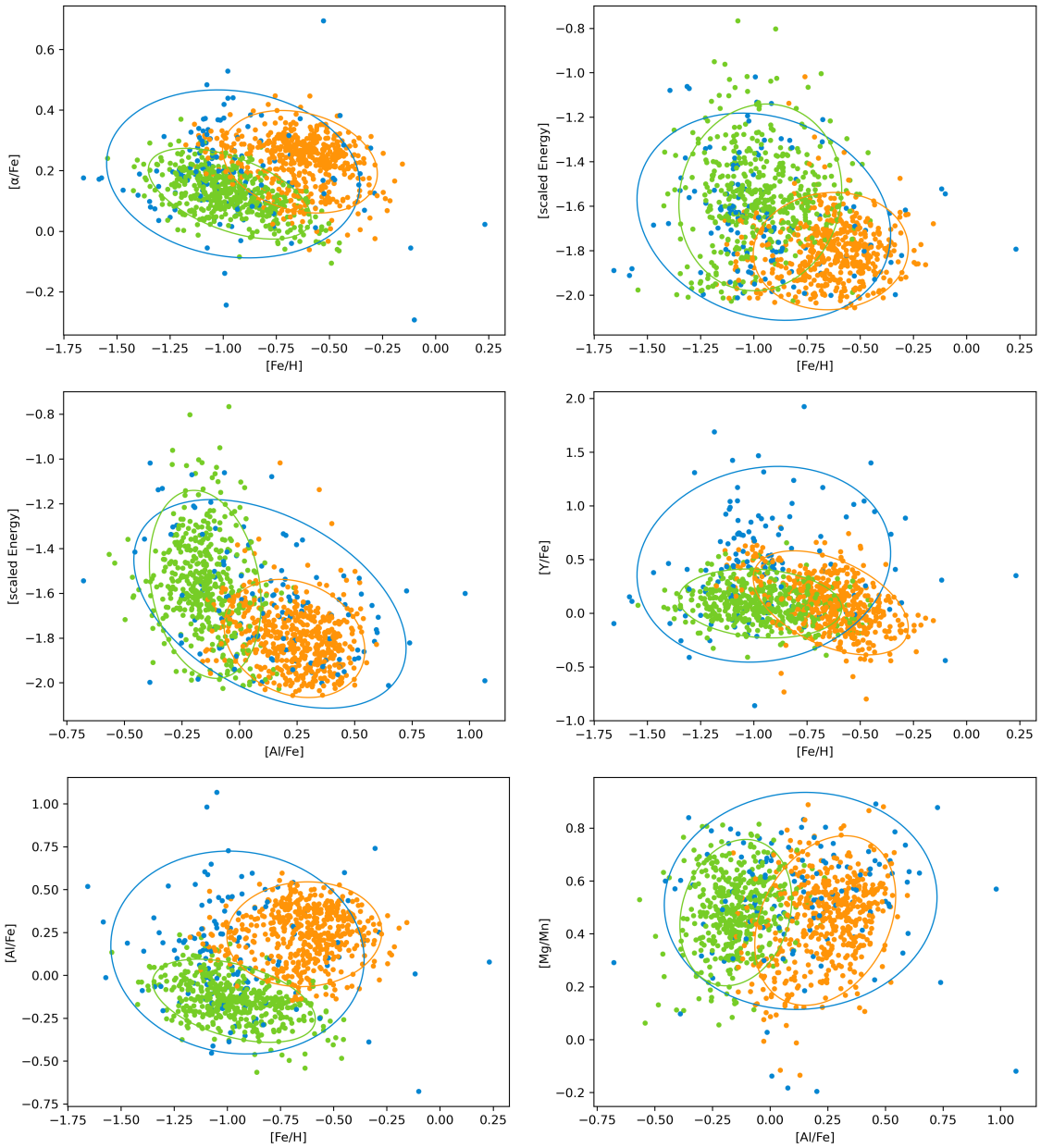


Figure 4: Distribution of Galaxy's eccentric populations in GALAH sample in chemo-dynamical space for three halo components. Green *GS/E*, orange *Splash* and blue colour mixture of the components

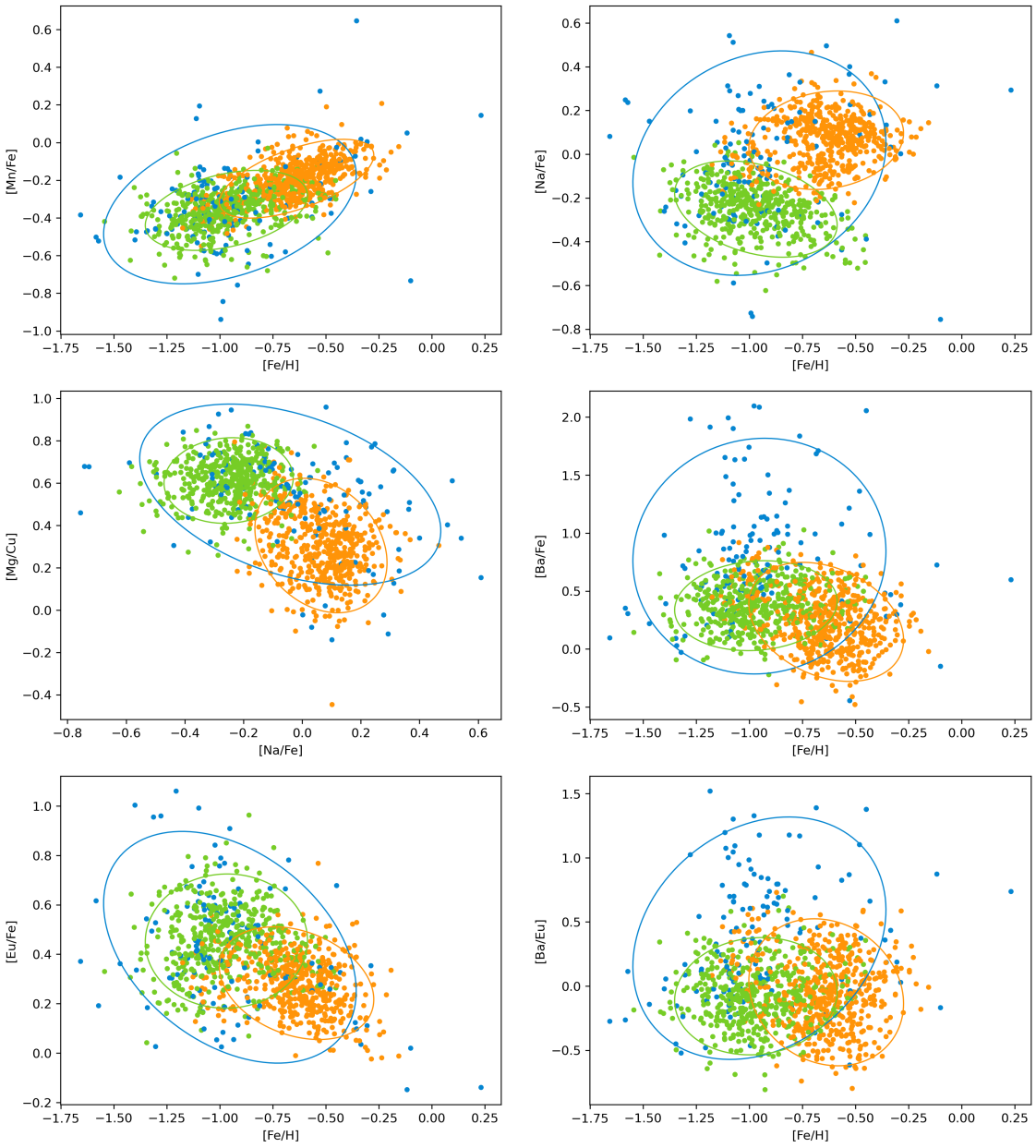


Figure 5: Distribution of Galaxy's eccentric populations in GALAH sample in chemo-dynamical space for three halo components. Similar to figure 4 but for different 2-dimensional projections. Green *GS/E*, orange *Splash* and blue colour mixture of the components

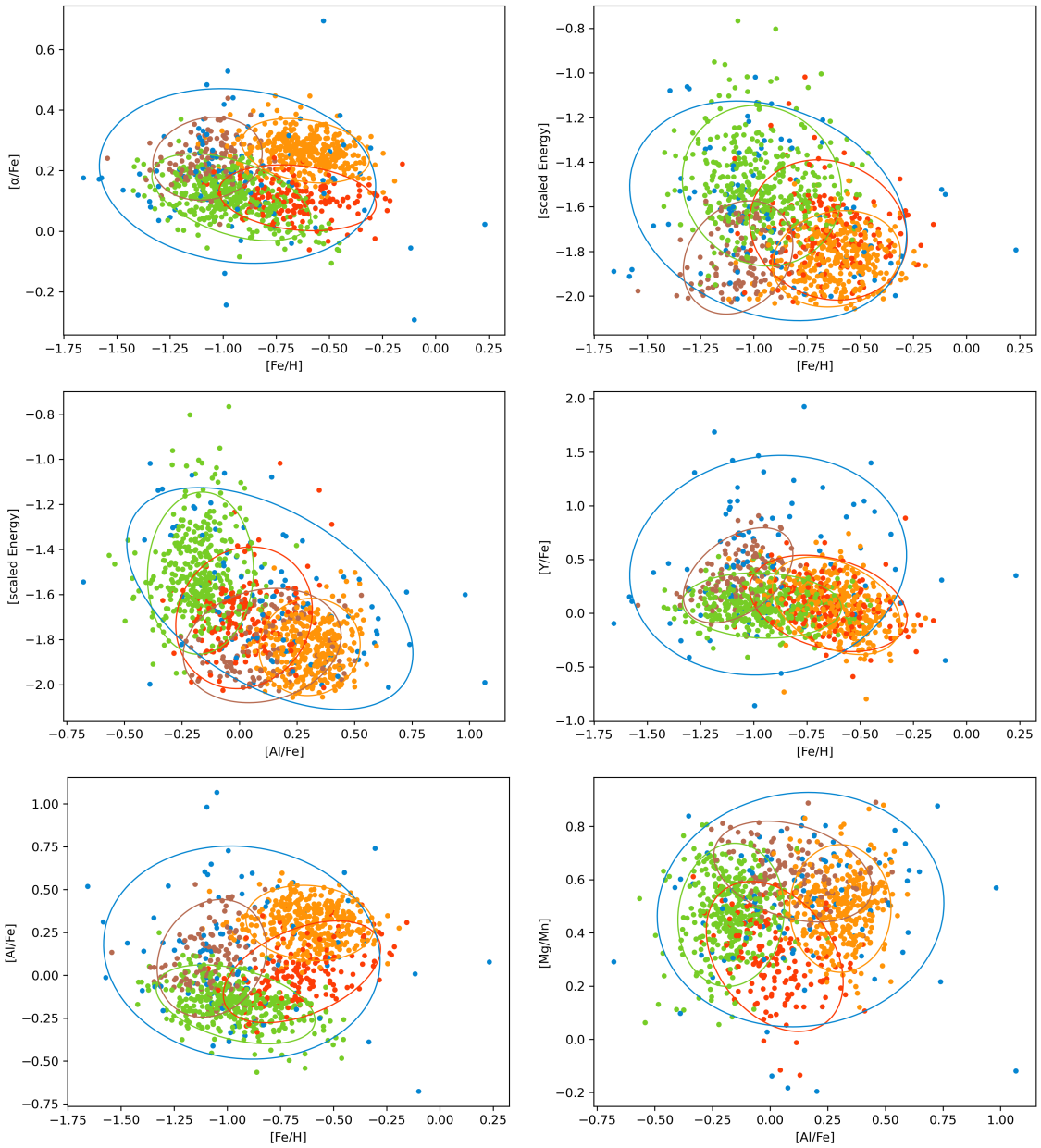


Figure 6: Distribution of Galaxy's eccentric populations in GALAH sample in chemo-dynamical spaces for 5 halo components. Green colour marks *GS/E*, orange colour *Splash*, brown colour *Aurora*, red colour *Eos*, blue colour background

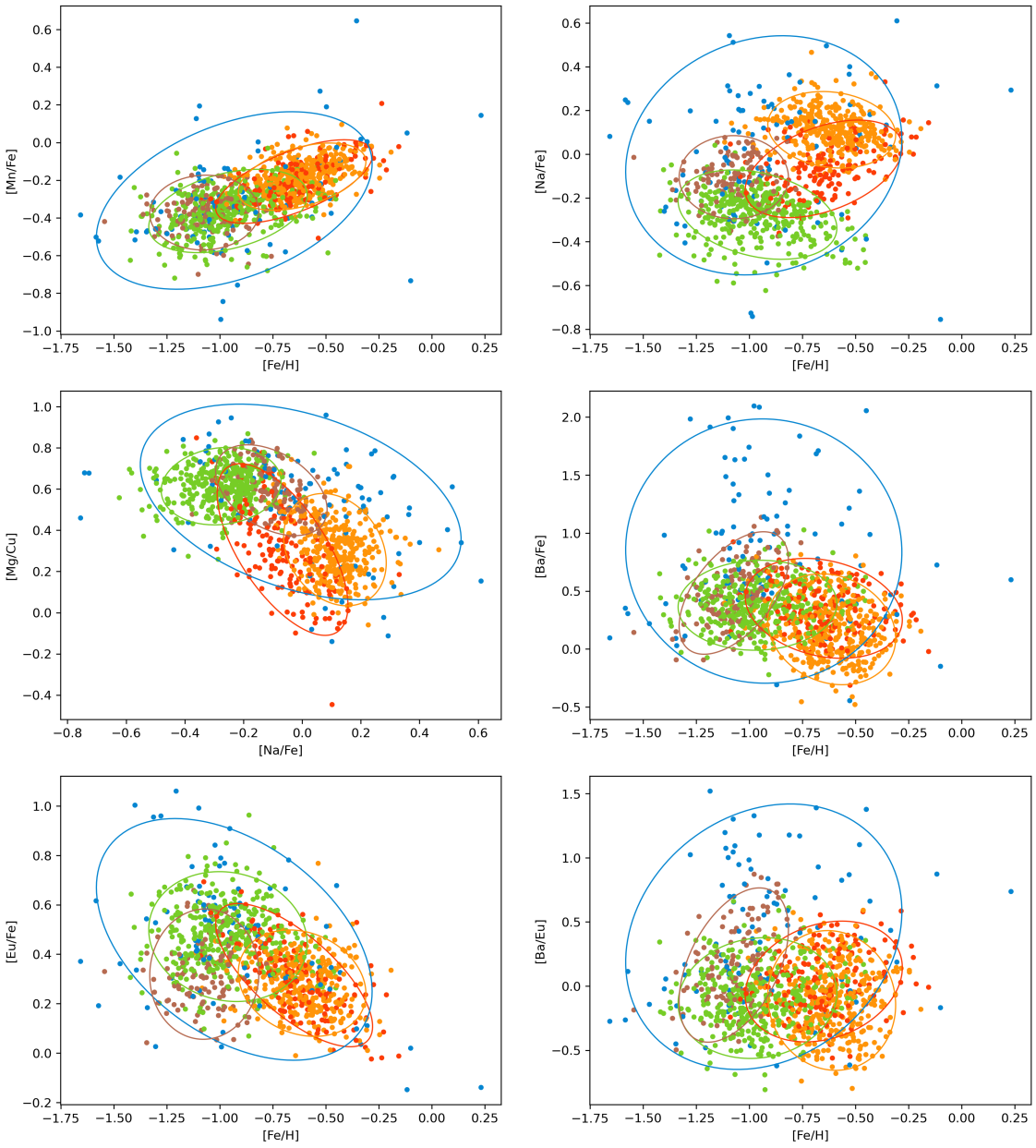


Figure 7: Distribution of Galaxy's eccentric populations in GALAH sample in chemo-dynamical space for 5 halo components. Similar to figure 4 but for different 2-dimensional projections. Green colour marks *GS/E*, orange colour *Splash*, brown colour *Aurora*, red colour *Eos*, blue colour background

Five components fit disfavoured by BIC score agrees with the findings described in the baseline publication [17] as well as APOGEE-*Gaia* fit. There are consistencies between same groups of elements such as α elements [α/Fe], iron-peak [Mn/Fe] or odd-Z [Al/Fe], [Na/Fe] describing similar trends in the stellar population for both datasets. Despite both fits identifying 4 strong components there are disparities between relative weightings - APOGEE-*Gaia* attributes half of the population of the stars to the GS/E while GALAH-*Gaia* only one third. This disparity can be explained by difference in survey's coverage on populations mainly GALAH-*Gaia* containing smaller fraction of stars below [Fe/H] < -1.3 dex resulting in GMM over-weighting the *Splash*, the *Aurora* and the *Eos* while underweighting GS/E component compared to APOGEE-*Gaia*. Another notable difference is how GMM is able to break GS/E into two separate but closely related components based on its internal chemical structure[17].

Splash component is very distinct from the other components having high metallicity such as [α/Fe], [Al/Fe] and [Fe/H] prominent in the APOGEE-Gaia 2D projection shown in figure 3 while having low orbital energy. High metallicity of this component is also captured in the GALAH-Gaia dataset figure 6. Both of these results agree with the idea that Splash component has *in-situ* origin but was displaced from the galactic disk by merger event [17].

GS/E is composed of the stars with relatively high orbital energy when compared to the rest of the stars in the dataset indicating that this component originated *ex-situ* and was accreted into Milky Way halo. It is poorer in the [Fe/H], [Al/Fe] elements as shown in APOGEE-Gaia dataset figure 3 suggesting inefficient star formation. Additionally, APOGEE-Gaia dataset covers wide range of metallicities resulting in the chemical feature space being non-Gaussian. It manifests itself in GMM fit identifying 2 different components of the GS/E that can not be covered by single Gaussian component. GALAH-Gaia dataset reveals high [Eu/Fe] abundance (figure 7) which is produced from r-process channels (CC SNe and neutron star mergers) indicating longer star formation of the cluster which allows stars to get more enriched in Eu [17].

Aurora component is the smallest of the main components with low orbital energy confirming earlier studies suggesting *in-situ* origin. It is very metal rich with high metallicity of [Fe/H] ≈ -1 with the highest abundance ratio of α-elements as well as [Al/Fe], [Ce/Fe], [Y/Fe], [Ba/Fe], [Ba/Eu] compared to the other halo components. High abundance of s-process elements have been suggested to originate from low- and intermediate-mass Asymptotic Giant Branch (AGB) stars. Low ratio of [Eu/Fe] produced by r-process from core collapse supernovae indicate at history of rapid star formation in this population. [17].

GMM decomposition of the local halo identified new component which is referred to as *Eos*. It has similar metalicity as *Splash* and contributes $\approx 10\%$ in the APOGEE sample and $\approx 15\%$ in GALAH. Compared to Splash, Eos has lower α-elements abundance but still higher than GS/E laying somewhere in between. Similarities in the levels of [Ce/Fe], [Na/Fe], [Eu/Fe], [Ni/Fe], [Eu/Fe] and [Eu/Mg] indicate that Eos might have origins outside of the galaxy being accreted into halo. High levels of [Al/Fe] and correlation between [Al/Fe] and [Fe/H] however indicate that Eos origin is *in-situ* [17].

Lastly both dataset fits identified "background" population of stars exhibiting outlier properties such as spanning across large abundance space (represented as high value of 2σ). This low-weight background population encompasses residuals from Gaussian model fitting.

5 UMAP and t-SNE projection

One of the approaches used in the analysis of the higher dimensional data is to employ dimensionality reduction technique which embeds high-dimensional data in low-dimensional space for the purposes of the visualisation. Important distinction to note is that, as opposed to clustering, dimensionality reduction techniques discard information in the process of projecting features into lower dimensions and it is primarily used to visualise structure of the data in low-dimensional space. Clustering on the other hand maintains all of the information present in the data while defining boundaries between groups of clusters.

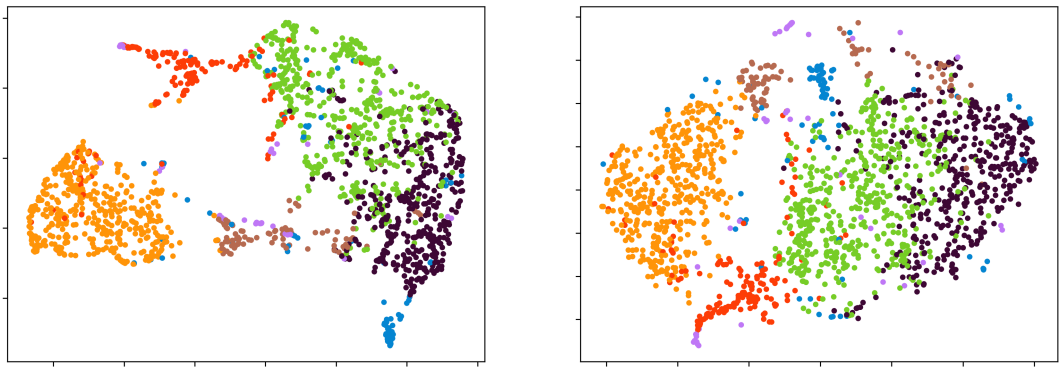
Two dimensionality reduction methods, t-SNE and UMAP were applied to the samples for APOGEE-Gaia and GALAH-Gaia. t-distributed stochastic neighbor embedding (t-SNE) method is non-linear dimensionality reduction techinique. It works by computing similarity of the data points in high-dimensional space using Gaussian distribution and similarity of the points in the mapped low-dimensional projection using Cauchy distribution. Kullback-Leibler (KL) divergence is used as a metric to minimise difference between the two distributions. Hyperparameter used in the mapping is number of the nearest neighbours to consider referred to as *perplexity* [12].

Uniform Manifold Approximation and Projection (UMAP) method is alternative dimensionality reduction method based manifold folding in Riemannian geometry and algebraic topology. It is computationally more efficient compared to t-SNE while also preserving more of the global structure of the data. Computationally UMAP is graph based learning algorithm operating on weighted graph [14].

Both UMAP and t-SNE methods are non-linear dimensionality reduction techniques. Non-linear aspect refers to the fact that the distances between data points in the high-dimensional space are not preserved in the low-dimensional projection. Distances between groups of points have no physical interpretation in the embedded low-dimensional space and only logical grouping of points should be used when interpreting the results.

In this study both APOGEE-Gaia and GALAH-Gaia datasets were transformed with t-SNE and UMAP algorithms embedding 6-dimensional and 12-dimensional chemo-dynamical feature space to 2-dimensional for visualisation purposes. Cluster membership calculated from GMM fitting was then applied to the data points represented with colours of the components matching projections presented earlier in figure 3 for APOGEE-Gaia and in figures 6 and 7 for GALAH-Gaia dataset.

Figure 8 presents UMAP and t-SNE embedding of the 6-dimensional APOGEE-Gaia dataset with 7 coloured components as identified by GMM fitting.



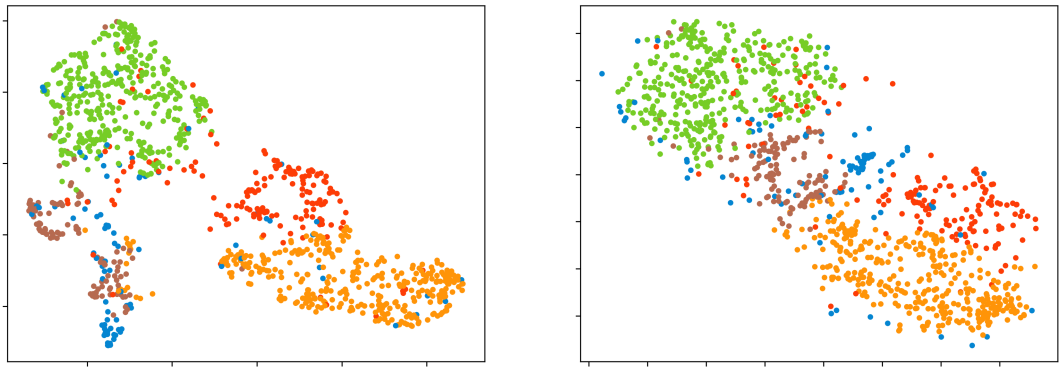
(a) UMAP embedding

(b) t-SNE embedding (perplexity = 100)

Figure 8: UMAP and t-SNE embeddings for APOGEE-Gaia dataset. Dark purple colour marks GS/E_1, green colour GS/E_2, orange *Splash*, brown *Aurora*, red *Eos*, pink and blue $back_1$ and $back_2$

In both embeddings GS/E components (GS/E_1 purple and GS/E_2 green) are grouped together with some datapoints mixed between the two. Equally there are islands for orange *Splash*, brown *Aurora* and red *Eos*. Purple points for background are scattered around the plots as is expected from the model residuals. One interesting thing subject to further investigation is strong cluster of blue points present in both UMAP and t-SNE embeddings which were categorised as background based on the GMM fit. These embeddings however indicate that a part of this blue cluster of points might be representing as of yet unidentified structure.

Figure 9 shows UMAP and t-SNE embeddings of the 12-dimensional GALAH-Gaia dataset with 5 coloured components identified by GMM fit. There are islands of the points representing main components - purple GS/E, green *Splash* and red *Eos*. Noticeable difference relates to the orange *Aurora* component which is not as strongly grouped in the UMAP embedding. It is even more prevalent in the t-SNE embedding where *Aurora* is separated into two isolated islands. Similarly to APOGEE-Gaia case background component appears to be grouped into island of its own in the UMAP embedding with weaker grouping in the t-SNE. Both of this results could be subject of further investigation.



(a) UMAP embedding

(b) t-SNE embedding (perplexity = 100)

Figure 9: UMAP and t-SNE embeddings for GALAH-Gaia dataset. Green colour marks *GS/E*, orange colour *Splash*, brown colour *Aurora*, red colour *Eos*, blue colour background

6 Conclusion

This study focused on the identification of the components within galactic halo by means of unsupervised machine learning algorithm. Gaussian Mixture model clustering algorithm was applied to the APOGEE-Gaia and GALAH-Gaia datasets in chemical and dynamical space to produce unbiased decomposition of the galactic halo structure. Goal of the study was to identify chemo-dynamical structure of the Milky Way's halo without any further prior knowledge to support evolutionary theory of a large galaxy based on ΛCDM model. Extreme Deconvolution fitting library was used to fit the GMM model as it offers benefit of incorporating measurement uncertainties into the model in contrast to other popular implementations such as *scikit-learn*. Bayesian

Information Criterion metric was used to evaluate the goodness-of-fit of the model. Despite discrepancy in BIC score between APOGEE-*Gaia* and GALAH-*Gaia* datasets, resulting from lower resolution of metalicities in GALAH dataset, both fits agree on main galactic halo constituents. Following number of iterations across different number of components, GMM fit identified 4 main halo 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*. Results reproduced in this study are in agreement with the results reported in the original work [17]. Additionally, this study looked into application of dimensionality reduction methods as a support to model fits. UMAP and t-SNE identified same main components as GMM fit while additionally suggesting further structure in the background and Aurora components.

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] H. Akaike. "A new look at the statistical model identification". In: *IEEE Transactions on Automatic Control* 19.6 (1974), pp. 716–723. DOI: 10.1109/TAC.1974.1100705.

[3] V Belokurov et al. "Co-formation of the disc and the stellar halo". In: *Monthly Notices of the Royal Astronomical Society* 478.1 (June 2018), pp. 611–619. ISSN: 0035-8711. DOI: 10.1093/mnras/sty982. eprint: https://academic.oup.com/mnras/article-pdf/478/1/611/25005885/sty982.pdf. URL: https://doi.org/10.1093/mnras/sty982.

[4] 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.

[5] 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.

[6] N. Wyn Evans. "The early merger that made the galaxy's stellar halo". In: *Proceedings of the International Astronomical Union* 14.S353 (2019), pp. 113–120. DOI: 10.1017/S1743921319009700.

[7] 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.

[8] Amina Helmi et al. "The merger that led to the formation of the Milky Way's inner stellar halo and thick disk". In: *Nature* 563.7729 (Nov. 2018), pp. 85–88. ISSN: 1476-4687. DOI: 10.1038/s41586-018-0625-x. URL: https://doi.org/10.1038/s41586-018-0625-x.

[9] R. A. Ibata, G. Gilmore, and M. J. Irwin. "A dwarf satellite galaxy in Sagittarius". In: *Nature* 370.6486 (July 1994), pp. 194–196. ISSN: 1476-4687. DOI: 10.1038/370194a0. URL: https://doi.org/10.1038/370194a0.

[10] Giuliano Iorio and Vasily Belokurov. "The shape of the Galactic halo with Gaia DR2 RR Lyrae. Anatomy of an ancient major merger". In: *Monthly Notices of the Royal Astronomical Society* 482.3 (Oct. 2018), pp. 3868–3879. ISSN: 0035-8711. DOI: 10.1093/mnras/sty2806. eprint: https://academic.oup.com/mnras/article-pdf/482/3/3868/26768322/sty2806.pdf. URL: https://doi.org/10.1093/mnras/sty2806.

[11] Gareth James et al. *An Introduction to Statistical Learning: with Applications in Python*. Springer, 2023. URL: https://link.springer.com/book/10.1007/978-3-031-38747-0.

[12] Laurens van der Maaten and Geoffrey Hinton. "Visualizing Data using t-SNE". In: *Journal of Machine Learning Research* 9.86 (2008), pp. 2579–2605. URL: http://jmlr.org/papers/v9/vandermaaten08a.html.

[13] D.S. Mathewson, M.N. Cleary, and J.D. Murray. "The Magellanic Stream." In: *apj* 190 (June 1974), pp. 291–296. DOI: 10.1086/152875.

[14] Leland McInnes, John Healy, and James Melville. *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*. 2020. arXiv: 1802.03426. URL: https://arxiv.org/abs/1802.03426.

[15] Xiao-Li Meng and David Van Dyk. "The EM Algorithm—an Old Folk-song Sung to a Fast New Tune". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 59.3 (Jan. 2002), pp. 511–567. ISSN: 0035-9246. DOI: 10.1111/1467-9868.00082. eprint: https://academic.oup.com/jrsssb/article-pdf/59/3/511/49588939/jrsssb_59_3_511.pdf. URL: https://doi.org/10.1111/1467-9868.00082.

[16] G C Myeong et al. "Discovery of new retrograde substructures: the shards of ω Centauri?" In: *Monthly Notices of the Royal Astronomical Society* 478.4 (June 2018), pp. 5449–5459. ISSN: 0035-8711. DOI: 10.1093/mnras/sty1403. eprint: https://academic.oup.com/mnras/article-pdf/478/4/5449/25105825/sty1403.pdf. URL: https://doi.org/10.1093/mnras/sty1403.

[17] 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.

[18] F. Pedregosa et al. “Scikit-learn: Machine Learning in Python”. In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[19] S. E. Sale et al. “The structure of the outer Galactic disc as revealed by IPHAS early A stars”. In: *Monthly Notices of the Royal Astronomical Society* 402.2 (Feb. 2010), pp. 713–723. ISSN: 1365-2966. DOI: 10.1111/j.1365-2966.2009.15746.x. URL: http://dx.doi.org/10.1111/j.1365-2966.2009.15746.x.

[20] Gideon Schwarz. “Estimating the Dimension of a Model”. In: *The Annals of Statistics* 6.2 (1978), pp. 461–464. DOI: 10.1214/aos/1176344136. URL: https://doi.org/10.1214/aos/1176344136.

[21] *scikit-learn: Gaussian mixture models*. URL: https://scikit-learn.org/stable/modules/mixture.html. (accessed: 09.06.2024).

[22] J. A. Sellwood and J. J. Binney. “Radial mixing in galactic discs”. In: *Monthly Notices of the Royal Astronomical Society* 336.3 (Nov. 2002), pp. 785–796. ISSN: 0035-8711. DOI: 10.1046/j.1365-8711.2002.05806.x. eprint: https://academic.oup.com/mnras/article-pdf/336/3/785/2953908/336-3-785.pdf. URL: https://doi.org/10.1046/j.1365-8711.2002.05806.x.

[23] Monty Stephanie. *Lecture Four in Galaxtic Archeology (Lent Term)*. Feb. 2024.

[24] Eugene Vasiliev. “AGAMA: action-based galaxy modelling architecture”. In: *Monthly Notices of the Royal Astronomical Society* 482.2 (Oct. 2018), pp. 1525–1544. ISSN: 0035-8711. DOI: 10.1093/mnras/sty2672. eprint: https://academic.oup.com/mnras/article-pdf/482/2/1525/26288818/sty2672.pdf. URL: https://doi.org/10.1093/mnras/sty2672.

[25] Ernst Wit, Edwin van den Heuvel, and Jan-Willem Romeijn. “‘All models are wrong...’: an introduction to model uncertainty”. In: *Statistica Neerlandica* 66.3 (2012), pp. 217–236. DOI: https://doi.org/10.1111/j.1467-9574.2012.00530.x. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-9574.2012.00530.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9574.2012.00530.x.