

Chemical tagging and age estimation with the GALAH DR4 survey

Author: Alejandra Laguarda González, alaguago65@alumnes.ub.edu
Facultat de Física, Universitat de Barcelona, Diagonal 645, 08028 Barcelona, Spain.

Advisors: Friedrich Anders, Chloé Padois; {fanders,chloe.padois}@fqa.ub.edu

Abstract: Stellar chemical abundances encode valuable information about the formation and evolution of the Milky Way. In this work, we explore two complementary approaches to extract this information from the GALAH DR4 survey. First, we explore the use of a supervised machine-learning algorithm to estimate stellar ages for red giant stars from their chemical abundances and atmospheric parameters, using asteroseismic ages as training data. While the model is able to recover a global age trend, the predicted ages show an unexpectedly poor precision. Second, we analyse the multidimensional chemical abundance space of red clump stars using an unsupervised clustering method, identifying chemically coherent groups. These groups display distinct chemical patterns that can be associated with different components of the Galactic disc.

Keywords: Chemical abundances, Machine learning, Stellar ages, Chemical tagging, Stellar populations, Galactic disc.

SDGs: 4. Quality education.

I. INTRODUCTION

The chemical compositions of stars encode fundamental information about the formation and evolution of the Milky Way. Large spectroscopic surveys such as GALAH provide high-quality measurements of stellar atmospheric parameters and elemental abundances for hundreds of thousands of stars, enabling the study of Galactic structure and stellar populations from a chemical perspective. This approach, commonly referred to as Galactic Archaeology, aims to reconstruct the evolutionary history of the Galaxy.

Stellar age is one of the most informative parameters for understanding Galactic evolution. Knowledge of stellar ages allows the study of how different Galactic components formed and evolved over time, as well as the investigation of age-dependent trends within the stellar disc. However, obtaining accurate stellar ages remains challenging. Even the most reliable methods, e.g. asteroseismology, are dependent on models of stellar evolution and unfeasible for enormous samples of stars. Therefore, alternative methods are needed in order to obtain stellar ages for large samples. A previous study ([1]) based on GALAH data has confirmed that the elemental abundances of a star correlate directly with its age, thus it is possible to estimate stellar ages using chemical abundances.

However, stellar ages do not capture the full information encoded in the multidimensional chemical abundance space. Stars with similar ages can display a wide range of chemical patterns due to differences in their formation environment, enrichment histories and radial migration [2]. For this reason, analysing the chemical abundance space offers complementary insight, allowing the identification of chemically defined stellar subpopulations that share similar formation and enrichment histories.

In this work, we explore two different approaches to extract information from stellar chemical abundances in

the GALAH DR4 survey. First, we estimate spectroscopic ages for a sample of red giant stars using a supervised method called Extreme Gradient Boosting (Sect. II). Second, we apply unsupervised clustering methods to a sample of red clump stars to identify chemical subgroups (Sect. III). Together, these approaches provide complementary perspectives on the temporal and chemical structure of the Galactic disc and illustrate the information encoded in stellar chemical abundances.

II. PREDICTING THE AGE OF RED GIANT STARS USING MACHINE LEARNING

A. Data

We use data from the fourth data release of the Galactic Archaeology with HERMES survey (GALAH DR4; [3, 4]), which provides stellar atmospheric parameters and individual elemental abundances derived from high-resolution spectroscopy for 917,588 stars in the Milky Way. For the chemical abundances, we rely on the improved calibration of Kos et al. [5]. The ages provided in the GALAH catalogue were obtained using isochrone fitting. However, an independent catalogue (Warfield catalogue; [6]) provides asteroseismic ages for red giants co-observed by the K2 mission [7] and the APOGEE survey [8], which are expected to be more accurate. To use these age estimates as a reference, we need the list of common stars between the Warfield catalogue and GALAH DR4.

We selected the red giant stars from the GALAH DR4 catalogue by applying cuts in effective temperature ($4000 \text{ K} < T_{\text{eff}} < 5200 \text{ K}$) and surface gravity ($1.0 < \log(g) < 3.5$). Then, we crossmatched this red giants sample with the Warfield catalogue, yielding 2,484 stars in common.

To construct a reliable training set, we applied quality cuts to ensure well-measured stellar parameters, chemi-

cal abundances, and age estimates. We discarded abundances with poor-quality flags, and only retained stars with good-quality measurements for the selected abundances and with relative age uncertainties below 4 Gyr. As a consequence, the final training sample is strongly reduced to 721 stars. Although this reduction is severe, it is necessary to minimise the noise in the training variables. The resulting clean sample is composed of red giant stars and their effective temperature, surface gravity, 21 chemical abundances and asteroseismic ages and is used to train the machine learning model described in the following section.

B. XGBoost Algorithm

To predict stellar ages we use Extreme Gradient Boosting (XGBoost; [9]), a supervised machine-learning algorithm based on ensembles of decision trees. XGBoost searches for complex and non-linear relations between a set of input variables and a target variable by combining the predictions of many simple decision trees. Each tree applies a sequence of conditions on the input features, and the trees are trained sequentially so that each new tree improves the predictions by correcting the errors of the previous ones. In addition, the algorithm includes regularisation techniques that limit the complexity of the trees, helping to reduce overfitting and improve the generalisation of the model.

In this work, stellar age estimation is formulated as a regression problem, where the goal is to predict a continuous output variable (stellar age) from a set of input features. The input variables used to train the model are T_{eff} , $\log(g)$, and a set of 21 well measured chemical abundances from GALAH DR4. The target variable is the stellar age provided by the Warfield catalogue.

The clean dataset of 721 stars described in the previous section is randomly divided into a training sample (80%) and a test sample (20%). We used the training sample to train the model and learn the relationship between the input variables and stellar age, while the test sample is used to evaluate the performance of the model on unseen data. Once trained, the model can be applied to other red giant stars to estimate their spectroscopic ages.

C. Results

After training the XGBoost model, we evaluate its performance on the test sample. Fig. 1 shows the comparison between the predicted spectroscopic ages and the seismic ages, with the dashed diagonal line indicating the ideal one-to-one relation.

From Fig. 1 it is clear that the model does not produce completely random predictions. A global trend is recovered, such that stars with larger reference ages are generally assigned larger predicted ages. The small bias

of 0.06 Gyr indicates that the model does not systematically overestimate or underestimate stellar ages, showing that it is well calibrated on average. However, the scatter around the one-to-one relation is large, with a typical precision of 2.38 Gyr. This means that, although the model captures a meaningful relation between the parameters and stellar age, the predicted ages differ substantially from the catalogue values and are, therefore, not very precise. This behaviour is typical of supervised machine-learning regression models and reflects a regression-towards-the-mean effect [10].

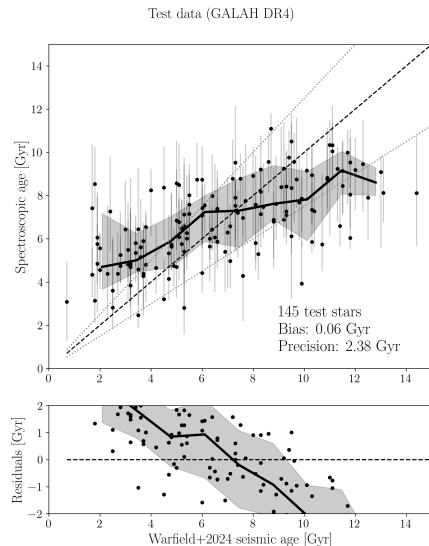


FIG. 1: Predicted spectroscopic ages from the XGBoost model versus seismic ages from the Warfield catalogue for the test sample (145 stars). The dashed line shows the one-to-one relation. The solid line and shaded region indicate the binned median and dispersion. The lower panel displays the residuals as a function of seismic age. The reported bias (0.06 Gyr) and precision (2.38 Gyr) summarise the global performance of the model.

The limited precision of the predictions suggests that the available training data are insufficient to constrain a tighter relation between spectroscopic parameters and age. In particular, the relatively small size of the clean training sample restricts the ability of the model to learn more accurate predictions.

To further investigate this issue, we examined the quality of the age labels used for training. In red giant stars, the surface $[C/N]$ ratio, which is the abundance ratio between carbon and nitrogen, is expected to correlate with age [11]. Fig. 2 shows the relation between the $[C/N]$ ratio and the asteroseismic age for the training sample.

The diagram shows a very large dispersion and no clear global trend, indicating that the age labels themselves are highly scattered. Since the model can only learn the relations present in the training data, a noisy and weakly defined age-parameter relation inevitably limits the pre-

cision that can be achieved.

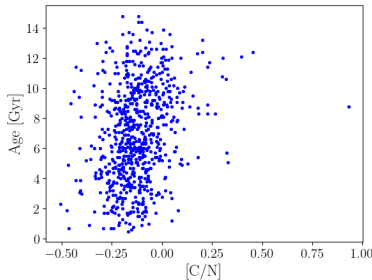


FIG. 2: Asteroseismic age (from the Warfield catalogue) versus $[C/N]$ ratio for the stars in the training sample.

Overall, these results show that estimating spectroscopic ages using machine learning techniques is feasible, but strongly dependent on the quality of the training set. In the mid-term future, more precise and accurate stellar ages provided by PLATO asteroseismology [12] are expected to significantly alleviate this limitation by enabling more reliable training samples.

III. CHEMICAL TAGGING WITH XDGMM

A. Data

In the second part of this work, we continue to use data from GALAH DR4. However, instead of predicting spectroscopic ages, we now aim to interpret the full chemical-abundance distribution. To this end, in order to minimise stellar-parameter biases, we select a sample of red-clump stars by applying cuts in effective temperature and surface gravity, $4500 \text{ K} < T_{\text{eff}} < 5100 \text{ K}$ and $2.3 < \log(g) < 2.55$, yielding a sample of 75,848 stars.

Some abundances have high measurement uncertainties for the majority of the sample. To discard them, we examined the quality flags of all elements and retained only the elements which have reliable measurements for more than 70% of our sample, resulting in a set of 27 chemical abundances. Then, we applied a final quality cut, selecting stars with $\text{flag} = 0$ (the best quality) for all 27 abundances, leading to a clean sample of 37,306 red clump stars. This dataset is used to define chemical subpopulations in the following sections.

B. UMAP

Uniform Manifold Approximation and Projection (UMAP; [13]) is a dimensionality reduction technique that allows high-dimensional data to be represented in a low-dimensional space. In our sample, each star is characterised by 27 chemical abundances, defining a high-dimensional chemical space that cannot be directly visu-

alised. UMAP projects this information into two dimensions while preserving local similarities, such that stars with similar chemical compositions are placed close together in the resulting map. In this work, UMAP is used as a visualisation tool to determine whether it is meaningful to search for chemically defined stellar subpopulations.

The resulting two-dimensional UMAP projection is shown in Fig. 3, coloured by $[Fe/H]$ and $[Al/Fe]$, the iron over hydrogen and aluminum over iron ratios, respectively. These abundances display a clear gradient across the UMAP space, indicating that chemical information is driving the structure seen in the UMAP rather than random noise.

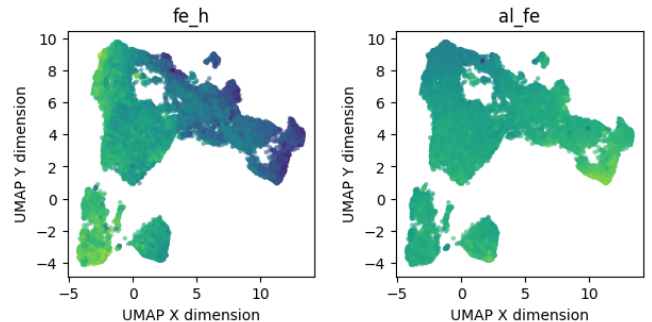


FIG. 3: UMAP projection coloured by $[Fe/H]$ (left panel) and $[Al/Fe]$ (right panel). The presence of a smooth gradient demonstrates that iron and aluminum contribute to the structure of the chemical abundance space.

The presence of a clear structure in the UMAP motivates the application of a clustering method to quantitatively identify distinct stellar populations. This is addressed in the following section.

C. XDGMM

In this part of the work, we use an Extreme Deconvolution Gaussian Mixture Model (XDGMM; [14]), an unsupervised clustering method. This technique improves the standard Gaussian Mixture Model (GMM) by taking into account the individual measurement uncertainties of the data.

The stellar sample is represented in a multidimensional chemical abundance space, where each star corresponds to a point defined by its 27 chemical abundances. The underlying assumption of Gaussian mixture modelling is that the distribution of stars in this space can be described as a combination of several groups, each represented by a Gaussian distribution. In this framework, each Gaussian component corresponds to a chemically distinct stellar subpopulation.

The model determines the chemical centre of each group given by the mean abundances, the shape of each group described by its covariance matrix, and the weight

of each component, which reflects how many stars belong to it. In addition, for each star, the model provides the probability of belonging to each Gaussian component.

The XDGMM model uses an Expectation-Maximization (EM) algorithm, which is an iterative technique that alternates between assigning stars to Gaussian components and updating the parameters of each component until a stable solution is reached.

The input to the model consists of the abundance matrix and the corresponding covariance matrix, both constructed from the red clump sample described previously. An important aspect of this method is the choice of the number of Gaussian components. An optimal number of clusters is expected to exist, balancing the ability of the model to describe the data and its complexity. We determined this optimal number of Gaussian components calculating the Bayesian Information Criterion (BIC) for different numbers of components, the optimal one corresponding to the minimal BIC value. In our case, the BIC analysis indicates an optimal solution with 20 Gaussian components.

Using this configuration, we apply the XDGMM model to the data. The resulting chemical subpopulations are analysed in the following section.

D. Results

We run the XDGMM algorithm on our clean sample of red clump stars, using 20 Gaussian components. The obtained groups are shown in the lower panel of Fig. 4, for the $[\text{Ti}/\text{Fe}]$ - $[\text{Fe}/\text{H}]$ space as an example. The results for all the chemical abundances are shown in Appendix A (Fig. 6).

The upper panel of Fig. 4 shows the UMAP projection coloured by the XDGMM clusters. The map displays extended regions dominated by single colours, indicating that stars that are close in chemical space are assigned to the same cluster. If either of the clustering methods were not capturing meaningful chemical structure, stars belonging to the same cluster would appear dispersed across the map. Instead, neighbouring points in the UMAP mostly share the same colour, showing that XDGMM separates the sample into chemically coherent groups and is able to capture real structure in the abundance space, as we can also see from the lower panel of Fig. 4.

To interpret the origin of the clusters, we examine their mean chemical abundance patterns. By comparing these patterns, we find that several clusters exhibit very similar chemical behaviour. Based on these similarities, we grouped the clusters into five characteristic chemical regimes.

Fig. 5 shows these patterns, where clusters with similar chemical behaviour are displayed together in the same panel.

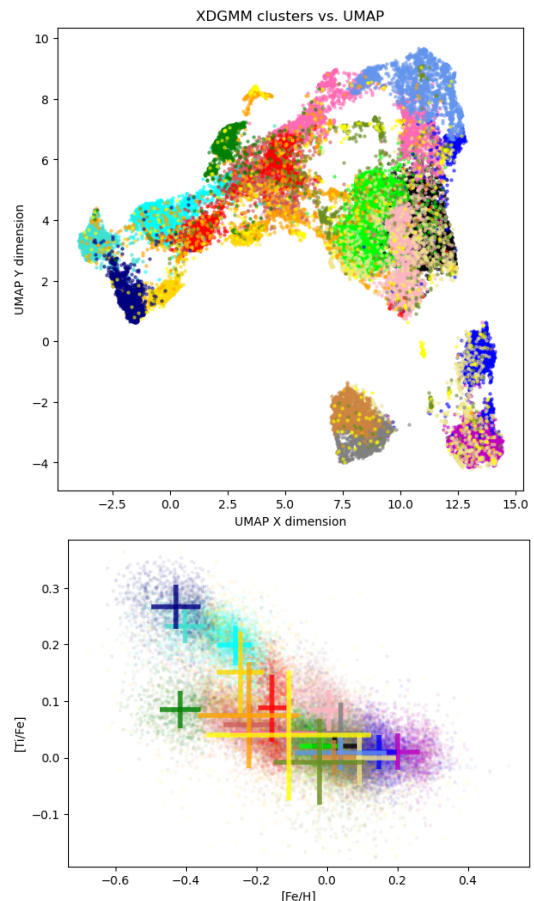


FIG. 4: Upper panel: UMAP projection of the red clump sample, coloured by the clusters identified with XDGMM. Each point represents a star, and colours indicate the most probable Gaussian component. Lower panel: Stellar distribution in the $[\text{Ti}/\text{Fe}]$ versus $[\text{Fe}/\text{H}]$ plane, coloured by the Gaussian components. Crosses indicate the mean chemical abundances and covariances of each component.

The first panel shows clusters characterised by high $[\alpha/\text{Fe}]$ ratios and low to intermediate $[\text{Fe}/\text{H}]$ ratios, reflecting enrichment dominated by type II supernovae. This chemical pattern is consistent with thick disc-like populations.

The second panel contains clusters with low $[\alpha/\text{Fe}]$ ratios and subsolar metallicities. These abundance patterns are consistent with metal-poor thin disc populations, commonly associated with the outer regions of the Galactic disc, where star formation and chemical enrichment proceed more slowly, resulting in lower metallicities.

The third panel shows clusters with low $[\alpha/\text{Fe}]$ and solar-like metallicities ($[\text{Fe}/\text{H}] \approx 0$). These clusters represent the thin disc in the solar neighbourhood. This chemical pattern reflects the prolonged formation of the thin disc, which allows type Ia supernovae to contribute significantly to the iron enrichment of the interstellar medium.

The fourth panel groups clusters with low $[\alpha/\text{Fe}]$ and super-solar metallicity. These clusters are chemically

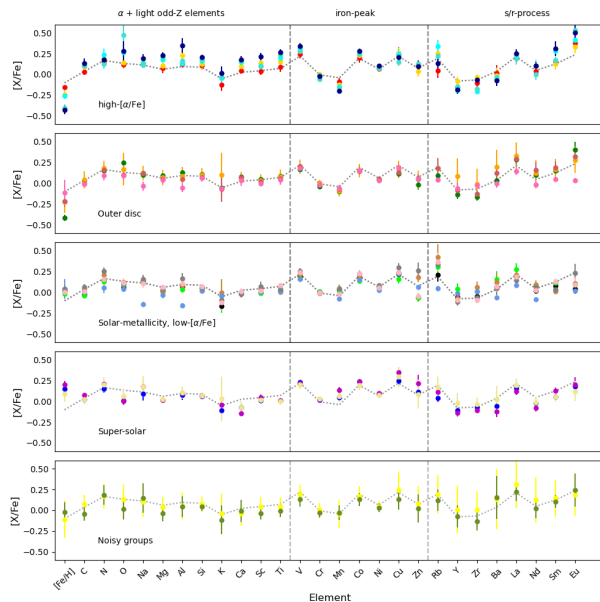


FIG. 5: Mean chemical abundance patterns of the identified clusters, separated into five panels according to their chemical similarity. Each color corresponds to the average abundance pattern of a cluster. The grey dashed line shows the global mean abundance pattern of the full sample. The elements are divided into three groups based on their origins (e.g. α -elements, iron-peak elements, and neutron-capture elements).

similar to thin disc populations but extend to higher metallicities, which are typically associated with the inner regions of the Galactic disc, where star formation has been more efficient and the interstellar medium more strongly enriched.

Finally, the fifth panel contains two clusters with large dispersions and poorly defined abundance patterns, which include relatively few stars. The colours corresponding to these clusters are very dispersed in the coloured UMAP, which confirms that they are noisy groups.

The identification of these chemical regimes and their

association with different Galactic disc components is based on well established relations between chemical abundance patterns and disc structure reported in previous studies [e.g. 15].

IV. CONCLUSIONS

In this work, we explored the information contained in stellar chemical abundances using two complementary approaches. First, we applied a supervised machine-learning method to estimate spectroscopic ages of red giant stars. While the model is able to recover a global age trend without significant bias, the limited precision of the predictions highlights the strong dependence of chemical age estimates on the quality and size of the training sample.

Second, we used an unsupervised clustering method to analyse the chemical abundance space of red clump stars. The results show that XDGMM is effective in identifying chemically coherent groups, which then can be interpreted and associated with different Galactic disc components.

Overall, this work illustrates the potential of combining supervised and unsupervised learning techniques to extract different information from stellar chemistry. Future asteroseismic surveys like PLATO [12], providing larger samples and more precise data, will be crucial to improve chemical age estimations and further reconstruct the formation and evolution of the Galactic disc.

Acknowledgments

I would like to sincerely thank my advisors Friedrich Anders and Chloé Padois for their guidance and help throughout this project. I am also very grateful to my family and friends for their constant support and for believing in me and my work.

-
- [1] M. R. Hayden et al., *Monthly Notices of the Royal Astronomical Society* **517**, 5325 (2022).
 - [2] K. Freeman and J. Bland-Hawthorn, *ARA&A* **40**, 487 (2002).
 - [3] S. Buder et al., *Publications of the Astronomical Society of Australia* **42**, e051 (2025).
 - [4] <https://www.galah-survey.org/dr4/overview/>.
 - [5] J. Kos et al., *Astronomy and Astrophysics* **703**, A104 (2025).
 - [6] J. T. Warfield et al., *The Astronomical Journal* **167**, 208 (2024).
 - [7] S. B. Howell et al., *Publications of the Astronomical Society of the Pacific* **126**, 398 (2014).
 - [8] Abdurro'uf, D. S. Aguado, C. Allende Prieto, et al., *The Astrophysical Journal Supplement Series* **259**, 35 (2022).
 - [9] T. Chen and C. Guestrin, *arXiv e-prints* (2016), arXiv:1603.02754.
 - [10] Y.-S. Ting, *The Open Journal of Astrophysics* **8**, E95 (2025).
 - [11] M. Martig, M. Fouesneau, H.-W. Rix, et al., *Monthly Notices of the Royal Astronomical Society* **456**, 3655 (2016).
 - [12] H. Rauer et al., *Experimental Astronomy* **59**, 26 (2025).
 - [13] <https://umap-learn.readthedocs.io/en/latest/>.
 - [14] J. Bovy, D. W. Hogg, and S. T. Roweis, *The Annals of Applied Statistics* **5**, 1657 (2011).
 - [15] M. R. Hayden et al., *The Astrophysical Journal* **808**, 132 (2015).
 - [16] <https://scikit-learn.org/stable/>.

Etiquetatge químic i estimació d'edats amb el sondeig GALAH DR4

Author: Alejandra Laguarda González, alaguago65@alumnes.ub.edu
Facultat de Física, Universitat de Barcelona, Diagonal 645, 08028 Barcelona, Spain.

Advisors: Friedrich Anders, Chloé Padois; {fanders,chloe.padois}@fqa.ub.edu

Resum: Les abundàncies químiques estel·lars codifiquen informació valuosa sobre la formació i l'evolució de la Via Làctia. En aquest treball, explorem dos mètodes diferents per extreure aquesta informació a partir del sondeig GALAH DR4. En primer lloc, explorem l'ús d'un algoritme d'aprenentatge automàtic supervisat per estimar les edats estel·lars de gegants vermelles a partir de les seves abundàncies químiques i paràmetres atmosfèrics, utilitzant edats asterosísmiques com a referència. Tot i que el model és capaç de recuperar una tendència global amb l'edat, les edats estimades presenten una precisió limitada. En segon lloc, analitzem l'espai multidimensional d'abundàncies químiques d'estrelles red clump mitjançant un mètode de clustering no supervisat, identificant grups químicament coherents. Aquests grups presenten patrons químics diferenciats que es poden associar amb diferents components del disc galàctic.

Paraules clau: Abundàncies químiques, Aprenentatge automàtic, Edats estel·lars, Etiquetatge químic, Poblacions estel·lars, Disc galàctic.

ODSs: Aquest TFG està relacionat amb l'Objectiu de Desenvolupament Sostenible (SDG) 4. Educació de qualitat.

Objectius de Desenvolupament Sostenible (ODSs o SDGs)

1. Fi de la es desigualtats		10. Reducció de les desigualtats	
2. Fam zero		11. Ciutats i comunitats sostenibles	
3. Salut i benestar		12. Consum i producció responsables	
4. Educació de qualitat	X	13. Acció climàtica	
5. Igualtat de gènere		14. Vida submarina	
6. Aigua neta i sanejament		15. Vida terrestre	
7. Energia neta i sostenible		16. Pau, justícia i institucions sòlides	
8. Treball digne i creixement econòmic		17. Aliança pels objectius	
9. Indústria, innovació, infraestructures			

Appendix A: Full chemical abundance space and XDGMM Gaussian components

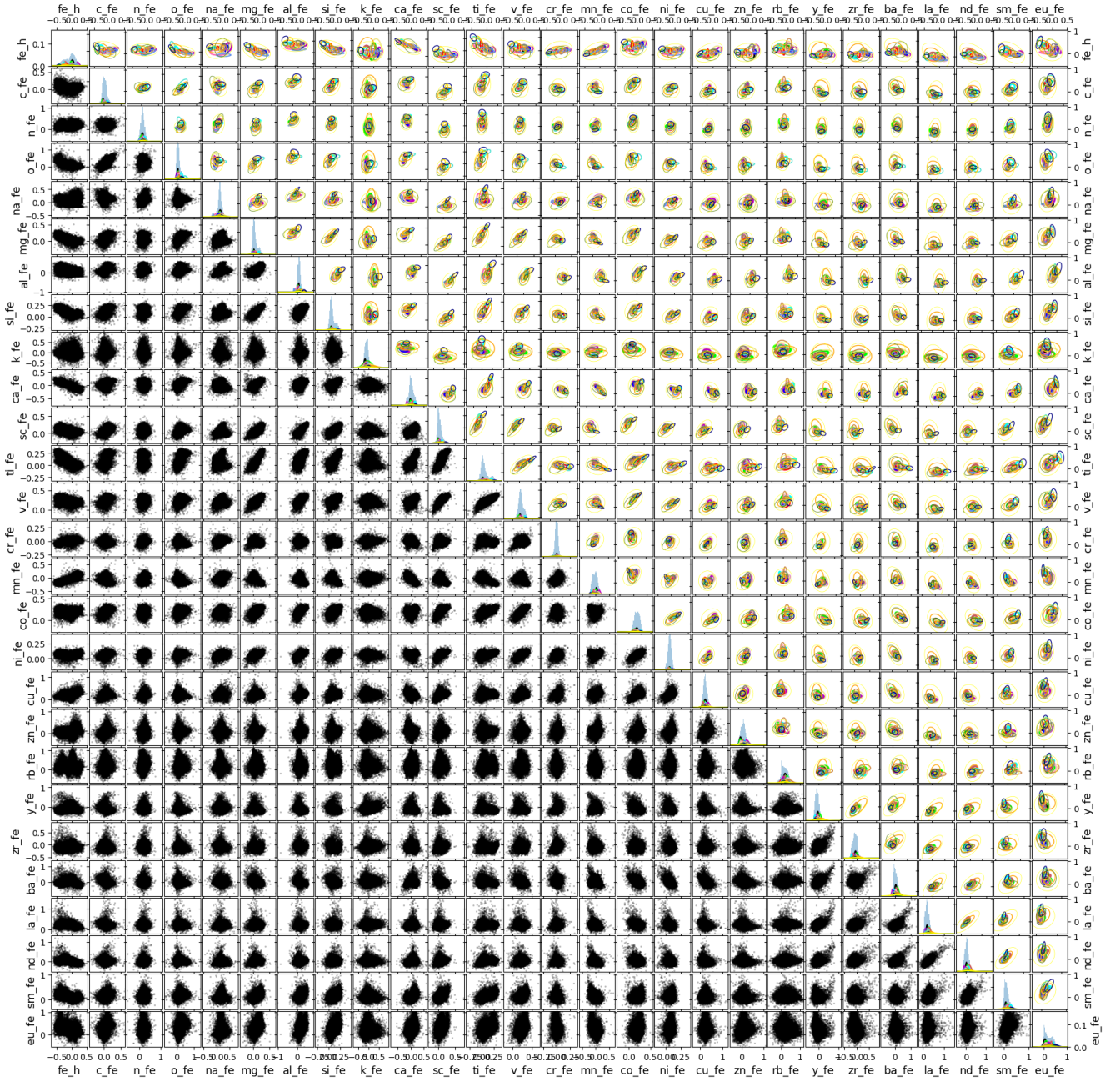


FIG. 6: Result of the application of the XDGMM algorithm to the red clump sample. The figure shows the two-dimensional correlations of the 27 chemical abundances. Panels below the diagonal display the original GALAH abundances prior to applying XDGMM, while the panels above the diagonal show the same projections after XDGMM has been applied. The Gaussian components identified by XDGMM are shown as ellipses in the upper panels. Diagonal panels show the one-dimensional abundance distributions before applying XDGMM, together with the corresponding distributions of the identified clusters.