Unsupervised machine learning techniques for chemical analysis in spectroscopic stellar surveys

Author: Jaume Dolcet Monés.
Facultat de Física, Universitat de Barcelona, Diagonal 645, 08028 Barcelona, Spain.

Advisor: Friedrich Anders

(Dated: February 4, 2021)

Abstract: In this work, we use the dimensionality reduction technique UMAP (Uniform Manifold Approximation and Projection) and a clustering algorithm (HDSCAN) on a large sample of stellar abundance ratios from a high-quality sample of the APOGEE DR16 survey (16000 red clump stars). We are able to reliably differentiate groups of stars corresponding with the chemical thick disk and thin disc, as well as a group corresponding to *high α metal rich* stars, and groups with anomalous abundances of certain elements, some of which are due to low precision on the abundances of P, Co, and Na determined by the pipeline. .

I. INTRODUCTION

For centuries humanity has wondered about the nature of the stars, but it wasn't until spectroscopy was first implemented in a telescope by Joseph Fraunhofer [1] that we became able to observe the spectrum of the celestial bodies and began to analyse their chemical composition.

Now stellar surveys can provide data of the chemical abundances of massive amounts of stars, and by analysing their composition, we can learn about the conditions under which the stars in our galaxy were formed, which can inform us on the past of the Milky Way.

Our aim is to characterize the population by identifying groups of stars with similar abundances and thus, similar origin. This is no trivial task since we will have to find clusters in an N-dimensional distribution of data points.

The usual approach has been to limit the analysis to a 2-D distribution of abundances, usually $[\alpha/Fe] vs [Fe/H]$ (α corresponds to elements formed through the alpha process), and it has been shown that doing so allows distinguishing two separated groups corresponding to the Thick (low metallicity, high α) and thin disk (high metallicity, low α) [13][14]. However, with big sample sizes, it becomes difficult to systematically distinguish between the two. Moreover, this method doesn't allow us to use the abundances of multiple elements in our characterization. That's why we use *dimensionality reduction.*

Dimensionality reduction techniques project data from a high dimensional space into a low dimensional one, while maintaining some structure of the original data, in order to facilitate its analysis.

They have recently started being applied to stellar abundances [20][19]. In [5], Tsne was applied to a solar neighbourhood sample of stellar abundances (HARPS-GTO) and several distinct populations were identified.

In this work, we attempt to continue this effort by using an arguably more efficient dimensionality reduction method (UMAP) and by applying it to a bigger sample.

The programming language python is the main tool used in this work and we implement the dimensionality reduction technique UMAP and the clustering algorithm HDBSCAN through their respective python libraries. Other tools used in the characterization of groups of stars and data visualization are the software TOPCAT and the tool *Science Archive Webapp*, from the SDSS website [18].

II. DATA

We will use data from the APOGEE DR16 survey [4], which contains high resolution, high signal-to-noise ratio, infrared spectra from over 100,000 red giant stars across the Milky Way, and provides the abundance ratios of several elements, computed by the *APOGEE Stellar Parameters and Abundances Pipeline* ASPCAP [16]. The details of the data reduction and calibration applied for DR16 data are described in [17].

The sample used in this study is comprised only of stars with a low value of χ^2 on their ASPCAP fit, and that have not been flagged by ASPCAP; so we are only considering *high-quality spectra.* The ASPCAP data provides abundances of 26 species, but in this study, we will use 20 of them (C, Cl, N, O, Na, Mg, Al, Si, P, S, K, Ca, Ti, Ti-II, V, Cr, Mn, Fe, Co, Ni, Cu).

III. DIMENSIONALITY REDUCTION

Uniform Manifold Approximation and Projection (UMAP) is a manifold learning method that allows us to create a projection of a multidimensional space where the distances between points are correlated with the distances in the multidimensional space; in our case, the abundance space where each element constitutes a dimension.

It produces results in a similar way to the widely used manifold learning algorithm t-SNE [9] which has been used with the abundance space of stellar populations [10][11][5]. UMAP, however, presents some advantages:

1. It uses less computation time than t-SNE, especially in big sample sizes.

2. It preserves *global data structure*, while in t-SNE only local structure is preserved. In other words, in the UMAP projection distances between clusters are correlated with distances in the original space.

The details on the mathematics of UMAP are described in [8].

A. Choosing the UMAP hyperparameters

As a manifold learning method, UMAP has multiple hyperparameters that affect its output:

n_components determines the dimensions of the map where the data will be visualized. We set it to 2.

metric determines how the distance is obtained in the n-dimensional space. We use *euclidean*.

n_neighbours balances the importance given to local or global structure in the data.

min_dist determines the minimum distance points will be allowed to be from each other in the reduced dimension map.

For n_neighbours and min_dist we use a range of values in order to determine if the structure observed in the 2-D plot is dependent on the hyperparameters.

As can be seen in Fig.(1), the choice of hyperparameters doesn't change the overall structure of the plot, but it can determine how clearly some groups are clustered together. For example, we see that clusters are more clearly defined when lower values of *min_dist* are used, as the data points are allowed to be more closely clustered than for higher values of *min_dist*.

In order to determine the hyperparameters, we use the distances on the 2-D plot between groups of stars belonging to known globular clusters.

Stars belonging to the same globular cluster share a similar origin and thus, have a similar chemical composition. So by minimizing the distances between stars from the same cluster we determine which hyperparameters result in a 2-D plot that best retains the *chemical closeness* between groups of stars.

We use 3 known clusters with stars that appear in our sample: NGC 188, with 4 members; NGC 6791, 6 members; NGC 6819, 11 members.

For each of them, we compute the standard deviation of the x and y positions. We then compute the sum of the x and y deviations for each cluster and compare it with the results obtained from each pair of hyperparameters.

The set of hyperparameters that minimizes the x and y deviations of stars in these clusters is:

$$n\_neighbours = 100 \quad , \quad min\_dist = 0.1.$$

In this study, we use the projection resulting from this choice of parameters.

IV. CLUSTERING

We do a first analysis of the UMAP plot by defining the groups manually with the software TOPCAT, simply by selecting groups of stars that look separate from the rest of the data points. This allows us to visualize the selected groups in 2-D abundance plots and to check if they are consistent with the structure of the UMAP plots obtained from different hyperparameters.

This way of defining the groups is, however, arguably arbitrary. So, in order to identify groups of stars from the UMAP 2-D projection in a systematic way, we use a clustering algorithm: Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDSCAN) [7].

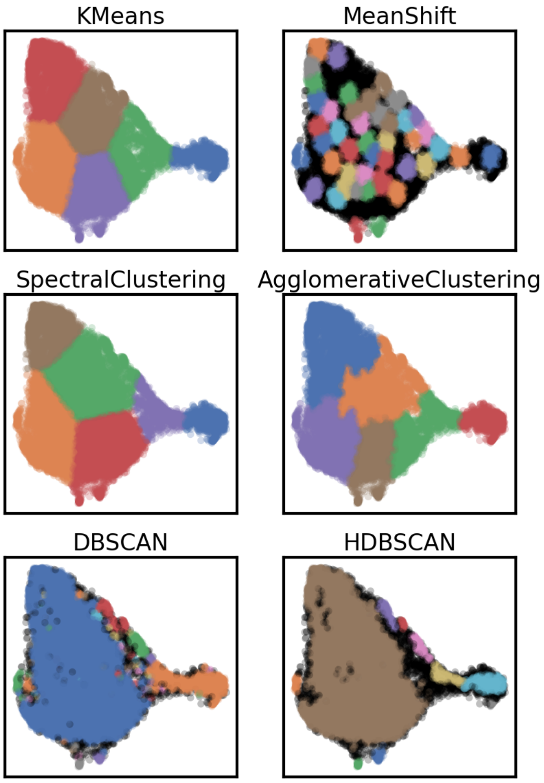

FIG. 1: Comparison of the clusters identified by 6 different clustering algorithms, using the parameters that seem to provide the best results for our data

In Fig. (1) we can see a comparison of the clusters identified by multiple clustering algorithms for the UMAP projection of our data. DBSCAN and HDBSCAN seem to best identify the structure of the data, and provide similar results, but DBSCAN defines many small clusters that don't appear to be significant. HDBSCAN labels most of these clusters as *noise*, and it seems to provide the best results.

HDBSCAN uses data point density to determine what constitutes a *cluster*, and it assumes that some amount of points are *noise* and do not belong to any particular cluster.

Similarly to UMAP, there are multiple parameters that will affect the cluster selection:

min_cluster_size determines the minimum size a group must have to be considered a cluster.

min_samples determines how conservative the clustering will be, so a higher value will lead to more points being labeled as noise.

With our data, using very low values for min_cluster_size and min_samples we get a large amount of very small clusters throughout the plot, and for values larger than approximately 10 we get roughly the same main clusters independently of the concrete values of the parameters. For high enough values of min_samples some of the smaller clusters are considered noise and others merge. Similarly, for high values of min_cluster_size some clusters also get merged together. However, for values of both parameters between 10 and 30, the clusters obtained are generally the same and independent of the particular parameter combination.

For our study, we set:

$min\_samples = 15 \quad , \quad min\_cluster\_size = 30.$

There are more parameters that effect the result of HDSCAN that have not been discussed here, but we set them to their default value in the python library as recommended by its documentation [12].

The resulting clusters extracted from HDBSCAN, as well as multiple 2-D chemical abundance plots, are shown in Fig.(2)

V. GROUP CHARACTERIZATION

In this section we will discuss the properties of the groups defined by the clustering algorithm (HDBSACN). 12 different groups of stars have been identified, as well as a 13th group which contains the stars labelled as noise by HDBSCAN (the gray group in Fig.(2)).

A first step towards characterizing the groups is to check their positions in the $[\alpha/Fe] vs [Fe/H]$ plot. An example of such plot is the top left panel of Fig.(2) ($[Mg/Fe] vs [Fe/H]$).

At first glance, we see that there is a clear distinction between what could be considered *high* $[\alpha/Fe]$ and *low* α/Fe groups.

Interestingly, this distinction can also be observed in the UMAP projection: all groups that appear to have high α element abundances are clustered at the right of the UMAP projection, and the ones that show low α abundances are on the left, in proximity of the large group labelled *thin disk I*. This is an example of global structure being preserved in the UMAP projection space.

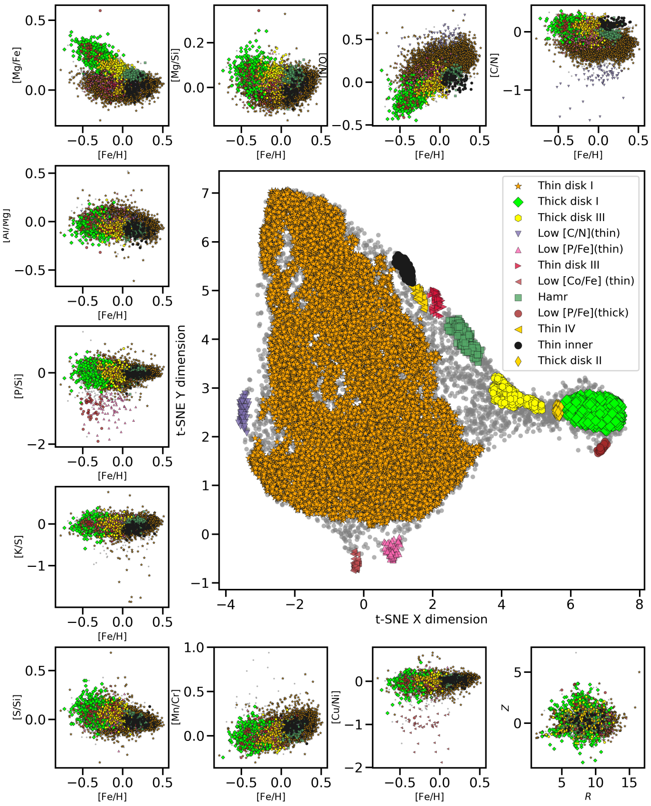

FIG. 2: Clusters obtained with HDBSCAN shown in the UMAP projection plot as well as in multiple 2-D abundance plots. The labels of the groups are discussed in the section *Group Characterization.*

A. Thin and thick disk identification

The discontinuity between the high and low $[\alpha/Fe]$ populations has been observed numerous times [13][14], and is interpreted as a chemical distinction between two galactic components:

1. The *thick disk*: with a much older population, usually with lower metallicity and enriched in α elements (O, Mg, Si, S, Ca, Ti) that result from type II supernovae;

2. The *thin disk*: with a younger population, higher metallicity, and lower α element abundances.

This components can also be defined through kinetic properties, but the chemical composition is less dynamic, so it retains information from the formation of the stars in a more direct way and provides a more reliable way of defining these populations.

Based on these distinctions, we label the two main populations identified in the previous section as *thin disk* stars (the low $[\alpha/Fe]$ population) and *thick disk* stars (high $[\alpha/Fe]$). However, for high metallicities, the distinction between these two components becomes less clear, and the thick disk stars aren't usually considered to exceed $[Fe/H] > 0$. So the *green square group* from Fig.(2) may be considered a distinct population described as *high α metal rich*.

Such a population was described for the HARPS sample in [15].

B. Description of individual groups

The plots of Fig.(3), allow us to more precisely discuss the chemical abundances of the individual groups identified by UMAP:

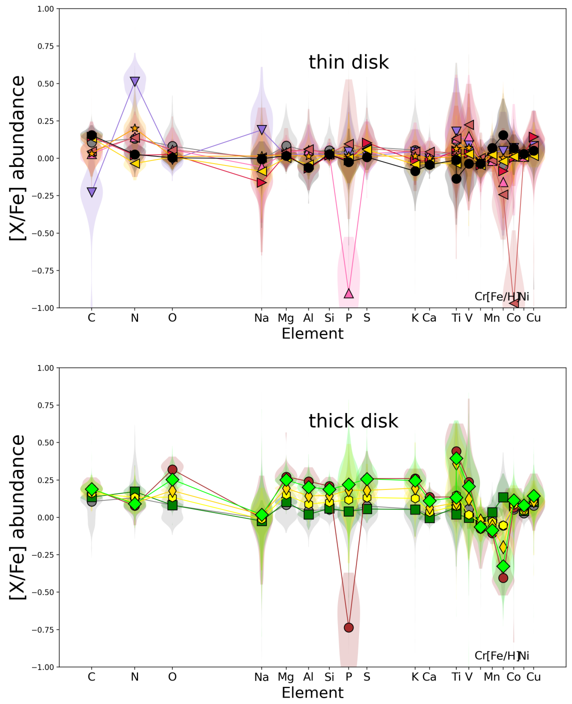

FIG. 3: Relative chemical abundances of the groups defined by HDBSCAN, divided into *thin disk* stars (top panel) and *thick disk* stars. The symbols are the same we used in Fig.(2), and they represent the average abundance of all the stars in the group. The whole distribution of abundances of each group is also plotted. Notice that the x axis represents Z, so the abundances of Ti and Ti II share their position in the x axis.

Examining the *bottom plot*, we can see that the groups show the same hierarchical order in $[\alpha/Fe]$ abundance across all α elements. If we analyze the [Fe/H] abundance ratio the reverse order is observed. This result is consistent with the observations in $[\alpha/Fe] vs [Fe/H]$ space, where the thick disk shows an inverse correlation between α element abundance and metallicity.

We can conclude that the groups labelled *Thick disk I* through *III* are subgroups of what we consider the chemical *thick disk* that differ from each other in metallicity.

The group *Thick disk I* has the highest α element abundances ($[\alpha/Fe] \simeq 0.25$), and *Thick disk III*, the lowest ($[\alpha/Fe] \simeq 0.1$).

The group labelled *hαmr* (green square) is the highest metallicity subgroup among the high $[\alpha/Fe]$ population ($[Fe/H] \simeq 0.15$). So, as we already discussed, we consider it distinct from the thick disk and label it *high α metal rich*.

The remaining group of the high α population (Low P (thick)) has some interesting characteristics: Across all elements, its abundance ratios are very similar to the *Thick disk I* group. However, its average Phosphorous abundance is $[P/Fe] \simeq -0.75$, which is much lower than any other group. Some of the stars of this group reach $[P/Fe] = -1.75$.

Examining now the top plot, we can see that there is almost no variance in $[\alpha/Fe]$ among the low α population. However, there are some groups that, similarly to the *Low P(thick)* group, have anomalous abundances of certain elements.

First of all, we consider the most numerous group in the sample: which we have labelled *Thin disk I.* It has a high variance in $[Fe/H]$ but a considerably low one in α elements, with an average value around ($[\alpha/Fe] \simeq 0$).

Most groups in the low α population share this same average abundance ratio. Interestingly, the *noise* group, which is shown in grey in both panels of Fig.(3), presents average α element abundances higher than all groups considered as *thin disk* and lower than those considered *thick disk*. This can easily be explained because there were probably stars from both populations that were considered noise by HDSCAN.

One example of a group that has similar abundances to *Thin disk I*, but differs in some elements is the group labelled *Low [C/N](thin)*. It has a low average abundance of C: $[C/Fe] \simeq -0.25$, a high abundance of N: $[C/Fe] \simeq 0.5$, and a high abundance of Na ($[Na/Fe] \simeq 0.2$).

Another such group is *Low P (thin)*, which has a low abundance of P, similar to *Low P (thick)* in average value and variance, but in the rest of elements shares the same abundances of *Thin disk I.*

The group labelled *Low Co (thin)* has an extremely low average abundance of Co ($[Co/Fe] \simeq -1$) and a slightly low $[Fe/H]$ abundance.

The group labelled *Thin inner* has a high metallicity similar to *hαmr.* Taking into consideration the metallicity gradient in galactic radius, we interpret this group as belonging to the *inner disk.*

The last two groups (*Thin disk III and IV*) seem hard to characterize because their abundances don't differ noticeably from *Thin disk I.* However, some differences can be observed: *Thin disk III*, for instance, has the lowest average Na concentration of any group, and *Thin disk IV* has the lowest average concentration of N.

VI. CONCLUSIONS

Throughout this work, we have shown that UMAP, as a dimensionality reduction method, has proven successful in creating a 2D projection of the 20 element abundance data of a sample of over 16,000 stars, that retains the structure of the original data and thus allows for a systematic chemical classification and characterization of the groups of stars according to their chemical compositions with independence of the hyperparameters chosen.

We have also shown that the clustering algorithm HDBSCAN can serve as a reliable and reproducible way to determine the boundaries of the groups of stars in the UMAP projection space.

This process has produced satisfactory results, as it has allowed us to chemically define the thick disk and thin disk populations in a more reliable way than using the 2-D $[\alpha/Fe]vs[Fe/H]$ abundance plot, as well as to identify subgroups inside of these main populations, that have different abundances of α elements and metallicity (groups *Thick disk I* through *III*, *hαmr* or *Thin inner*), or that have abundances of certain elements that deviate from the main populations (groups *Low [P/Fe](thin), Low [P/Fe](thick), Low [Co/Fe](thin), Low [C/N](thin))*.
Our results, obtained with data from a sample of over 16,000 Red Clump stars from the APOGEE survey are compatible with those of [5], in which data from the solar neighborhood HARPS sample was analyzed using the dimensionality reduction algorithm T-sne, and the same major groups were identified: *Thin disk and thick disk*, as well as *hαmr*. We can, thus, conclude that the split between the two disk components, as well as the separation of a hαmr group are not local phenomenon nor selection effects, and can be recovered with different dimensionality reduction algorithms.

However, the same groups of stars with anomalous abundances of particular elements were not observed in [5], so it's possible that they are exclusive to our sample or even an error of the pipeline. As is discussed in detail in [17], the abundance of P, Co and Na are among the less precisely determined ones in the sample, as they are measured using few weak absorption lines, so it is recommended that they be avoided. Citing from [17]: ”Phosphorous is measured from a few very weak lines, and is the least precisely determined element abundance in DR16” ”The cobalt abundances are derived from a single line, and therefore it is not unexpected that they show significant scatter”.

Taking this into consideration, we can conclude that the extreme abundances observed in the groups labelled *Low [P/Fe](thin), Low [P/Fe](thick)*, and *Low [Co/Fe](thin)*, are most likely non physical, and they are caused by pipeline errors. The abundances observed in *Low [C/N](thin)*, on the other hand may be physical, as the [C/N] ratio is considered to be correlated to stellar parameters such as mass and age.

In conclusion, as stellar surveys provide increasingly bigger amounts of data, machine learning seems to be the most suitable approach to analysing it, and dimensionality reduction techniques, in particular, are the most useful tool to help us analyse structure in stellar chemical abundance space. Even if some of the abundances we used were not precisely measured, our method proved useful in detecting outliers with suspicious abundances.

In the future, it would be an interesting effort to complement the chemical data with other stellar parameters of the selected groups, in hopes of gaining more insight in the causes of the anomalous abundances.

Acknowledgments

I would like to sincerely thank my advisor, Dr. Friedrich Anders, for providing his constant assistance, guidance and knowledge throughout this work, and for providing me an opportunity to develop my TFG in this most interesting field. I would also want to express my gratitude to my parents, who have always supported me in every possible way.

[1] Fraunhofer, Joseph, 1817, Annalen der Physik, 56: 264–313

[2] Delgado Mena, E., et al., 2017, A&A, 606, A94

[3] Fuhrmann, K. 1998, A&A, 338, 161

[4] Majewski, S. R., et al., 2017, AJ, 154, 94

[5] Anders, F., et al., 2018, A&A, 619, A125.

[6] Anders, F., et al., 2014, A&A, 564, A115

[7] Campello R.J.G.B., et al., PAKDD 2013, Lecture Notes in Computer Science, vol 7819.

[8] McInnes, L., et al., 2018, arXiv:1802.03426

[9] van der Maaten, L., et al., 2008, J. Mach. Learn. Res., 9, 85

[10] Matijevic, G., et al. 2017, ˇ A&A, 603, A19

[11] Anders, F., et al. 2018, IAU arXiv:1708.09319v2

[12] L. McInnes, J., et al., 2017 Journal of Open Source Software, The Open Journal, 2, 11

[13] Edvardsson, B., et al., 1993, A&A, 275, 101

[14] Fuhrmann, K. 1998, A&A, 338, 161

[15] Adibekyan, V. Z., et al., 2011, A&A, 535, L11

[16] García Pérez, Ana E., et al., 2016, The Astronomical Journal, 151, 6, 144

[17] Jönsson, H., et al., 2020, The Astronomical Journal, 151, 6, 144

[18] https://dr16.sdss.org/infrared/spectrum/search

[19] Boesso, R., et al., 2018, MNRAS, 474, 4010

[20] Jofré, P., et al., 2017, MNRAS, 467, 1140