Analysing the red-clump star population of the Milky Way with abundance-space extreme deconvolution

Author: Aleix Cuevas Bullich
Facultat de Física, Universitat de Barcelona, Diagonal 645, 08028 Barcelona, Spain.

Advisor: Friedrich Anders
(Dated: January 26, 2022)

Abstract: In this work we use the Extreme Deconvolution algorithm to extract the underlying distribution function from an heterogeneous and noisy star sample of 10,941 Red Clumps from the GALAH+ DR3 survey. This algorithm allows us to describe the density probability function as a sum of Gaussians that we can interpret as distinct Galactic components. We are able to clearly distinguish the thin and thick disk components, and the hαmr group.

I. INTRODUCTION

Galactic Archaeology is the subfield of Galactic astronomy that aims at understanding the formation and evolution of galaxies, in particular the Milky Way [1]. Explaining how the Milky Way formed is one of the principal goals of astrophysical research of recent years, and thanks to the advent of the Gaia mission [2] it is beginning to become within our reach.

In order to infer the chemo-dynamical history of the Milky Way, it is necessary to collect and analyze millions of stellar spectra across all Galactic components, from the Galactic Centre to the halo. Thanks to large-scale spectroscopic surveys, like RAVE [3], APOGEE [4] or GALAH [5], the amount of spectroscopic data taken has increased by several orders of magnitude in the last 15 years. The analysed dimensions have also increased, from kinematics and metallicities to detailed 6D phase-space + precise chemical abundances for several elements + ages. This large amount of data also means that new machine-learning techniques (supervised or unsupervised learning) are necessary to make sense of the complex datasets and to produce new scientific results.

In this work we explore a new technique in the context of multi-dimensional abundance-space analysis: extreme-deconvolution Gaussian Mixture Modelling [6]. In short, this technique expands the classical two-dimensional abundance analysis of the Tinsley-Wallerstein Diagram ([α/Fe] vs. [Fe/H]; [7]) to taking into account all available elemental abundances at the same time.

Our aim is to check if the method works well with abundance data (using the high-quality dataset of GALAH DR3 red-clump star catalogue) and if we are able to reproduce meaningful results or find new chemically defined subgroups of stars.

II. DATA

We use the chemical abundances published in the the third data release of the Galactic Archaeology with HERMES survey (GALAH DR3 [15]). The data release contains 678,423 spectra from 588,571 mostly nearby stars taken with the HERMES spectrograph, built into the Anglo-Australian Telescope. The DR3 catalogue provides stellar atmospheric parameters and individual elemental abundances for up to 30 different elements.

In order to be able to work with very homogeneous abundances and little systematic trends (e.g. with effective temperature), we select a sample of red-clump stars (metal-rich red giant stars in the core He-burning phase; [8]). The selection is achieved by applying some basic cuts to the GALAH DR3 catalogue with the astronomical software TOPCAT [16]:

$$Red\ clump\ cuts = \begin{cases} 4500K < T_{eff} < 5100K \\ 2.3 < log(g) < 2.55 \\ is\_redclump\_bstep > 0.5 \end{cases} \quad (1)$$

After making some quality checks for reliably measured abundances, we further reduce the sample size by requiring $flag\_sp = 0$ & $flag\_fe\_h = 0$ & $flag\_X\_fe$ for each of the considered abundances. We end up with a sample of 10,941 red-clump stars with good-quality chemical abundances of 24 elements: Fe, O, Na, Mg, Al, Si, K, Ca, Sc, Ti, V, Cr, Mn, Co, Ni, Cu, Zn, Y, Zr, Ba, La, Ce, Nd, Eu.

III. EXTREME DECONVOLUTION

There is a wide variety of methods to describe the probability distribution function (PDF) of a data set. The mathematical problem of finding an adequate PDF is called density estimation. Another (related) problem of general interest is to find and characterise fluctuations or overdensities within the data. This is called clustering.

In many cases both problems can simultaneously be solved by Gaussian Mixture Modelling (GMM), which models the data as a sum of Gaussians that can sometimes be interpreted as subgroups. This is a parametric method; unlike methods such as Kernel Density Estimation (KDE) or Nearest Neighbors Density Estimation, which do not specify a functional model to fit the data on [9]. It is the most common mixture modelling technique and often used to adjust probability densities

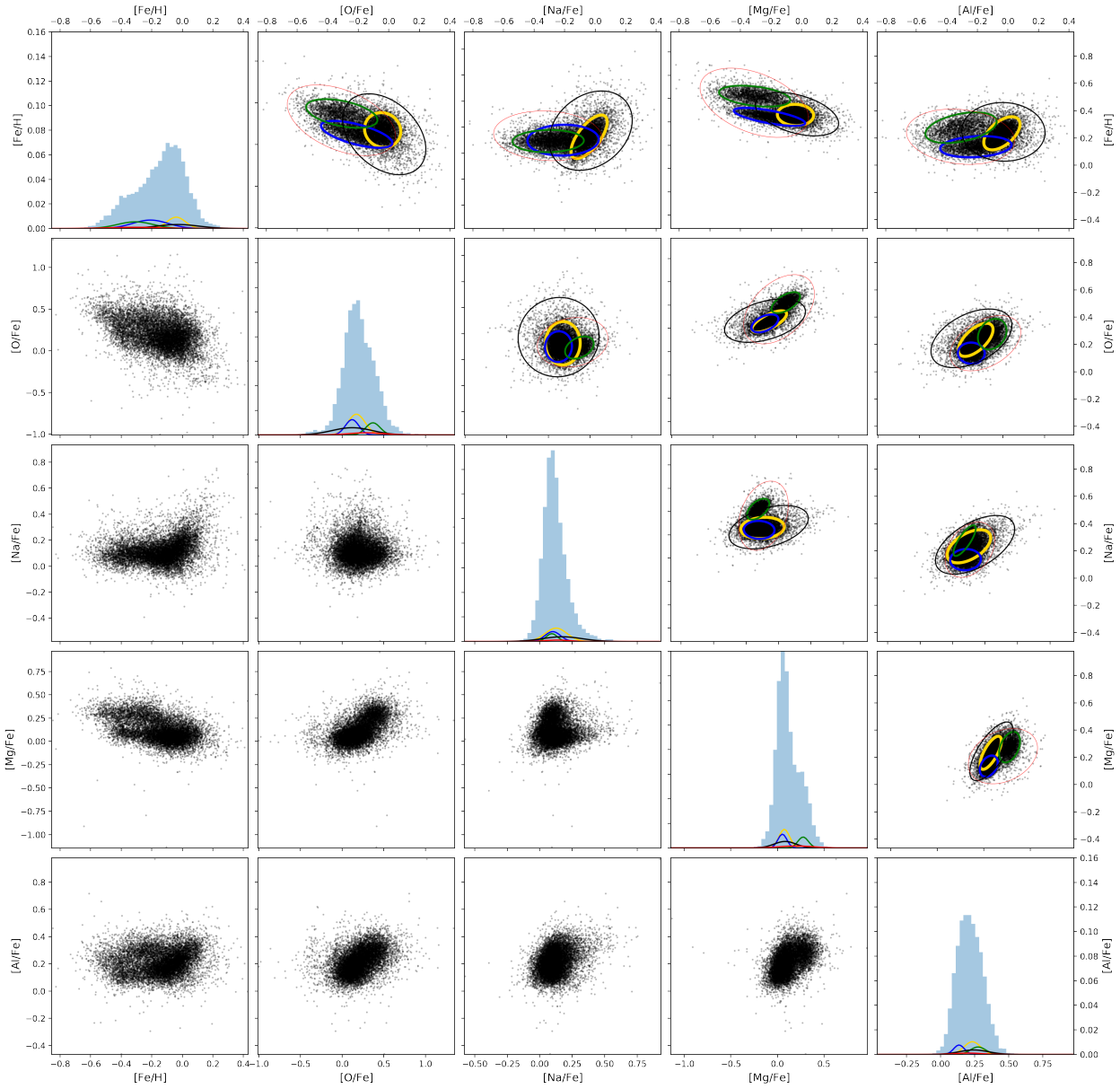

FIG. 1: Result of the application of XDGMM on the RC sample of GALAH DR3, for the 24-dimensional abundance space. For visibility we only show the 2D correlations of 5 of the considered elemental abundance ratios: [Fe/H], [O/Fe], [Na/Fe], [Mg/Fe], [Al/Fe]. The panels below the diagonal shows the GALAH abundances as reported in the catalogue (i.e. before applying XDGMM). The panels above the diagonal show the same abundance ratio diagrams, but now after XDGMM has been applied (so that the axis labels for the plots above the diagonal have to be inverted). In each of the upper-diagonal panels we overplot ellipses whose colours and line widths correspond to the found Gaussian clusters and their weights, respectively. The diagonal panels shows the one-dimensional histograms of the respective abundance ratios (before applying XDGMM) as well as the respective distribution of the five clusters.

to subpopulations from a more general population.

In a one-dimensional (1D) GMM, the probability density is modelled as [9]:

$$\rho(x) = N \sum_{j=1}^{M} \alpha_j \mathcal{N}(\mu_j, \sigma_j), \tag{2}$$

where we sum over M Gaussians, each one with a weight α_j, average μ_j and a covariance matrix σ_j.

An advantage of mixture models is that they do not require knowing which subpopulation each point belongs to. They are unsupervised methods, so no prior information is given on the data. Unsupervised clustering techniques use all the available data to find the optimum

number of classes. The only prior information needed is the number of clusters (but even this can be optimised, e.g. by minimising the Bayesian Information Criterion; BIC [9]).

When the data set is subject to significant (and known) measurement errors, we can use a generalised version of GMM called 'Extreme Deconvolution' (XDGMM; [6]). The algorithm takes into account the known data uncertainties to deliver the deconvolved means and covariances of each Gaussian. It basically assumes that the observed value is given by a real value operated with a projection matrix (computed by GMM), following this relation [9]:

$$x_i = R_i v_i + \epsilon_i, \tag{3}$$

where x_i and v_i are the observed and the true values respectively, R_i is the projection matrix and ϵ_i the known error of each measure. In this work we use the astroML [17] implementation of [9], in particular the function `astroML.density_estimation.XDGMM`.

IV. RESULTS

A. Analysing abundance space

Figure 1 demonstrates how XDGMM works on chemical abundance space data. It shows 2D projections for five of the 24 abundances spanned by the GALAH DR3 red-clump sample. For example, the [Mg/Fe] vs [Fe/H] plot (panel positions 4,1 and 1,4) is the typical [α/Fe] vs [Fe/H] diagram in which we can, almost by eye even before applying XDGMM, distinguish the two main local disk components: thin and thick disk [10]. The panels for [O/Fe] vs. [Fe/H] (O being another α element) also show signs of a chemical bimodality.

The XDGMM analysis with 5 components (shown in the panels above the diagonal in Fig. 1) clearly allows us to detect and characterise these two disk components (blue and green ellipses representing thin and thick disk respectively) and to enhance the visibility of the gap in abundance space that was partly blurred by observational uncertainties. We also clearly detect another component that has emerged as a separate group in the last 10 years: the high-α metal-rich (hαmr) stars discovered by [11] are clearly picked up as a separate group by XDGMM (yellow ellipses in Fig. 1).

The remaining two components (red and black) found by XDGMM in Fig. 1 correspond to stars with larger abundance uncertainties and thus can be considered as noise (see also Table I).

B. Analysing kinematics space

After applying XDGMM, we may calculate the probability of each point to belong to each component [9]:

$$p(j \mid x_i) = \frac{\alpha_j \mathcal{N}(\mu_j, \sigma_j)}{\sum_{j=1}^{M} \alpha_j \mathcal{N}(\mu_j, \sigma_j)}, \tag{4}$$

and paint each star according to their highest membership probability $p(j|x_i)$. We can now project the different abundance groups in kinematics space.

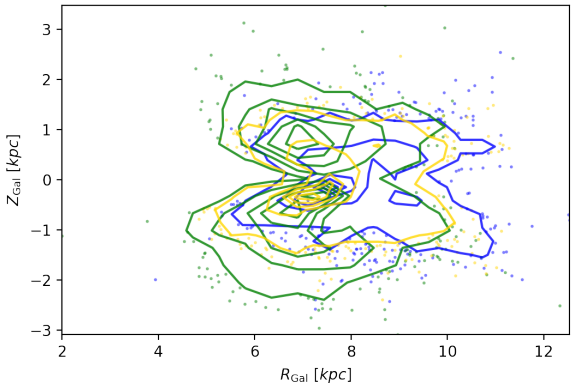

FIG. 2: Galactic distribution (in Galactocentric cylindrical coordinates) of the three main subgroups found by XDGMM with 5 components (using the same colours as in Fig. 1). For each group we show four iso-density contours; outliers are shown as individual points.

To visualize the Galactic distribution of the sample, in Fig. 2 we plot Galactocentric coordinates $R_{\rm Gal}$ vs. $Z_{\rm Gal}$ for the three main groups found by XDGMM. While a quantitative analysis is impossible due to the important selection biases, we comment on some qualitative results.

The green (thick disc) and yellow (hαmr) components concentrate closer to the inner disk, while the blue (local thin disc) extends to larger Galactocentric radii; some stars reach distances of $R_{\rm Gal} > 10$ kpc. The green group is more extended in vertical direction too, while the other two components are flatter. Fig. 2 thus clearly justifies the names we have attached to the blue and green components based on their chemical abundances in Sect. A, and confirms that the scale length of the thick disc is shorter than that of the thin disc [12].

The yellow group is concentrated near the Sun and does not extend as far into the outer disc as the blue group. It is likely that a large part of the hαmr group is made up of migrated stars that were born in the inner disk and migrated to the solar environment.

Fig. 3 shows the kinematics of the three main chemical groups in the GALAH RC sample in the classic Toomre diagram. We see that most stars are co-rotating with the Galactic disc; only 8 of them have retrograde orbits ($v < v_\phi^{GAL} = 0$) and thus belong to the retrograde halo. Stars with a total velocity $v_{LSR} > 180\ km \cdot s^{-1}$ are most likely halo stars [13], which means that the green component contains a number of halo stars; although we have seen most of it is part of the thick disk (section IV A).

In contrast, yellow and blue components are clear candidates for kinematically cold disk populations. Not only have they an average velocity close to the Local Standard of Rest but also the perpendicular component $\sqrt{{v_R}^2 + {v_Z}^2}$ is quite low.

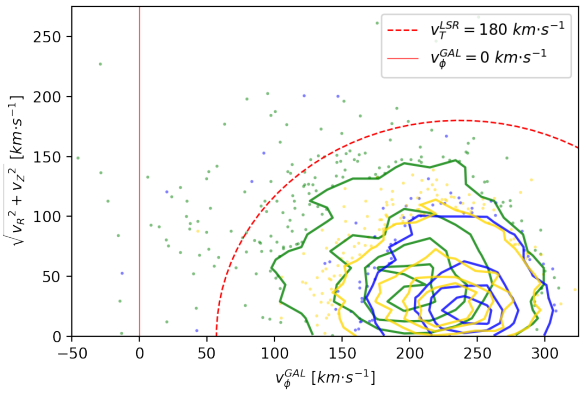

FIG. 3: Toomre diagram of the three main subgroups (using the same colours as in Fig. 1, in the same style as Fig. 2). The x axis shows the azimuthal velocity v_ϕ, while the y axis shows the absolute velocity component perpendicular to v_ϕ. The dashed red line indicates a total velocity of $v_T^{LSR} = 180\ km \cdot s^{-1}$ with respect to local standard of rest, while the continuous red line indicates zero tangential velocity.

The Sun has an angular velocity of $v_\phi^{GAL} \approx 248\ km \cdot s^{-1}$, which combined with its peculiar motion with respect to the local standard of rest ($+11$ km$\cdot s^{-1}$) means that the average orbital velocity of the thin disk is around 237 $km \cdot s^{-1}$ [14]. We find mean velocities of 235.1 $km \cdot s^{-1}$, 221.6 $km \cdot s^{-1}$ and 194.6 $km \cdot s^{-1}$ for the blue, yellow and green groups respectively, which reaffirms that most of the stars from the first two groups kinematically belong to the thin disk.

As in the previous Fig. 2, the Toomre diagram in Fig. 3 does not contain the red and black populations (noisy abundance measurements) that would appear as very scattered points.

Finally, Fig. 4 shows the orbit distribution ($Z_{\max}$ vs. *ecc* in doubly logarithmic scale) of the three abundance groups (as reported in the GALAH DR3 catalogue). We see that the two quantities are correlated for all subgroups, albeit with significant scatter. We also see, in accordance with Figs. 2 and 3, that the green (thick-disc) population moves on more eccentric orbits that also reach higher altitudes above the Galactic plane. The yellow population falls somewhat in-between the blue and the green population in this diagram.

Kinematically, stars with $ecc > 0.55$ ($log(ecc) > -0.26$) are on very radial orbits and cannot be considered disk stars; so the top right of the diagram are candidates for (kinematically defined) halo stars. We find that the range of $0.5 < ecc < 0.7$ ($-0.25 < log(ecc) < -0.15$)

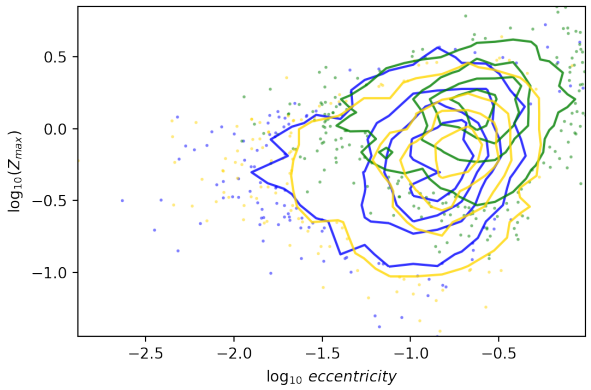

FIG. 4: Orbit distribution. The colors of the components are the same we used in Fig. 1. We show the relation beetween the maximum height with respect to the Galactic plane and the eccentricity (we made the logarithm of the two variables for easier visualization).

is sparsely populated, possibly indicating a physical gap between thick disk and halo stars.

C. Higher-order description of abundance space

Gaussian Mixture Models with few components are commonly used for clustering. Increasing the number of Gaussians is more useful for approximating the probability density function, at the expense of an increase in the number of parameters. We have run XDGMM for a higher number (up to 30) of Gaussian components to describe the chemical-abundance space of the local disk more accurately.

The results in the 24-dimensional GALAH DR3 abundance space are, unfortunately, too complex to show them here in the same style as in Fig. 1, so we exemplify them in tabular form in Table I. The first table of I summarises and quantifies our findings from Sect. A: the thin disk component (blue group) has higher mean metallicity than the thick disk (green group). The enhancement of α elements in the thick disk is evident in the abundance of [O/Fe] of the green component. The table also lists the minimum abundance elements in each group. For example, the low [Y/Fe] for yellow group suggests that this group is probably old (no time for s-process enhancement).

The bottom table of Table I shows the results for an XDGMM analysis with 10 Gaussian components. With higher number of components, the results become harder to interpret, but possibly also more interesting, as it should become possible to detect groups of outliers with peculiar chemical composition.

Component colour	% in sample	Mean [Fe/H]	[Fe/H] dispersion	Max.[X/Fe]	Min.[X/Fe]
Black	11.6%	-0.024	0.15	[V/Fe]	[Ce/Fe]
Red	3.6%	-0.309	0.19	[Ba/Fe]	[Fe/H]
Yellow	37.1%	-0.037	0.07	[V/Fe]	[Y/Fe]
Green	21.1%	-0.308	0.13	[O/Fe]	[Fe/H]
Blue	26.6%	-0.213	0.13	[Ba/Fe]	[Fe/H]

Component colour	% in sample	Mean [Fe/H]	[Fe/H] dispersion	Max.[X/Fe]	Min.[X/Fe]
Black	12.7%	-0.159	0.10	[V/Fe]	[Fe/H]
Red	19.9%	-0.073	0.07	[Ba/Fe]	[Ce/Fe]
Gold	16.2%	0.013	0.06	[V/Fe]	[Y/Fe]
Green	10.0%	-0.412	0.09	[O/Fe]	[Fe/H]
Blue	2.8%	0.033	0.12	[Ba/Fe]	[Ce/Fe]
Orange	7.4%	-0.021	0.13	[Ba/Fe]	[Fe/H]
Cyan	11.4%	-0.305	0.10	[Ba/Fe]	[Fe/H]
Lime	1.9%	-0.332	0.17	[Ba/Fe]	[Fe/H]
Magenta	15.9%	-0.187	0.09	[O/Fe]	[Fe/H]
Gold	1.7%	-0.242	0.21	[V/Fe]	[Fe/H]

TABLE I: Results of the 5-component XDGMM (top table) and 10-component XDGMM (bottom): Each row refers to an abundance group. For each group, we show percentage, metallicity, metallicity spread, maximum and minimum abundance.

V. CONCLUSIONS

Extreme Deconvolution (XDGMM) is effective and precise in finding and denoising overdensities in stellar chemical abundance space. The algorithm allows us to distinguish disk components for a low but arbitrary number of Gaussians. However, it does not scale favourably with the size of the data set and the number of Gaussians. The optimal number of components could be calculated, e.g. by minimizing the BIC [9].

XDGMM is a promising way to describe the multi-dimensional abundance-space PDF. Modeling the data for a larger number of components should work well for analysing the composition of the local disk environment, perhaps even quantitatively.

Acknowledgments

I would like to thank my advisor Dr. Friedrich Anders for the guidance and constant help during the work. I would also like to thank my family for their unconditional support.

[1] K. Freeman and J. Bland-Hawthorn, ARA&A **40**, 487 (2002), astro-ph/0208106.
[2] Gaia Collaboration, A. G. A. Brown, A. Vallenari, T. Prusti, J. H. J. de Bruijne, F. Mignard, R. Drimmel, C. Babusiaux, C. A. L. Bailer-Jones, U. Bastian, et al., A&A **595**, A2 (2016), 1609.04172.
[3] M. Steinmetz, G. Matijevič, H. Enke, T. Zwitter, G. Guiglion, P. J. McMillan, G. Kordopatis, M. Valentini, C. Chiappini, L. Casagrande, et al., AJ **160**, 82 (2020), 2002.04377.
[4] S. R. Majewski, R. P. Schiavon, P. M. Frinchaboy, C. Allende Prieto, R. Barkhouser, D. Bizyaev, B. Blank, S. Brunner, A. Burton, R. Carrera, et al., AJ **154**, 94 (2017), 1509.05420.
[5] S. Buder, S. Sharma, J. Kos, A. M. Amarsi, T. Nordlander, K. Lind, S. L. Martell, M. Asplund, J. Bland-Hawthorn, A. R. Casey, et al., MNRAS **506**, 150 (2021), 2011.02505.
[6] J. Bovy, D. W. Hogg, and S. T. Roweis, Annals of Applied Statistics **5**, 1657 (2011), 0905.2979.
[7] B. M. Tinsley, ApJ **229**, 1046 (1979).
[8] L. Girardi, ARA&A **54**, 95 (2016).
[9] Ž. Ivezić, A. J. Connelly, J. T. VanderPlas, and A. Gray, *Statistics, Data Mining, and Machine Learning in Astronomy* (2014).
[10] B. Edvardsson, J. Andersen, B. Gustafsson, D. L. Lambert, P. E. Nissen, and J. Tomkin, A&A **500**, 391 (1993).
[11] V. Z. Adibekyan, N. C. Santos, S. G. Sousa, and G. Israelian, A&A **535**, L11 (2011), 1111.4936.
[12] T. Bensby, A. Alves-Brito, M. S. Oey, D. Yong, and J. Meléndez, ApJ **735**, L46 (2011), 1106.1914.
[13] K. A. Venn, M. Irwin, M. D. Shetrone, C. A. Tout, V. Hill, and E. Tolstoy, AJ **128**, 1177 (2004), astro-ph/0406120.
[14] J. Bland-Hawthorn and O. Gerhard, ARA&A **54**, 529 (2016), 1602.07702.
[15] `https://www.galah-survey.org/dr3/the_catalogues/`
[16] `http://www.star.bristol.ac.uk/~mbt/topcat/`
[17] `http://www.astroml.org/`