The Astrophysical Journal, 900:165 (21pp), 2020 September 10 https://doi.org/10.3847/1538-4357/abac61

Tracing the Assembly of the Milky Way's Disk through Abundance Clustering

Bridget L. Ratcliffe[1], Melissa K. Ness[2,3], Kathryn V. Johnston[2,3], and Bodhisattva Sen[1]
[1] Department of Statistics, Columbia University, 1255 Amsterdam Avenue, New York, NY 10027, USA
[2] Department of Astronomy, Columbia University, 550 West 120th Street, New York, NY 10027, USA
[3] Center for Computational Astrophysics, Flatiron Institute, 162 Fifth Avenue, New York, NY 10010, USA
Received 2020 February 12; revised 2020 August 3; accepted 2020 August 3; published 2020 September 14

Abstract

A major goal in the field of galaxy formation is to understand the formation of the Milky Way's disk. The first step toward doing this is to empirically describe its present state. We use the new high-dimensional data set of 19 abundances from 27,135 red clump Apache Point Observatory Galactic Evolution Experiment stars to examine the distribution of clusters defined using abundances. We explore different dimension reduction techniques and implement a nonparametric agglomerate hierarchical clustering method. We see that groups defined using abundances are spatially separated, as a function of age. Furthermore, the abundance groups represent different distributions in the [Fe/H]–age plane. Ordering our clusters by age reveals patterns suggestive of the sequence of chemical enrichment in the disk over time. Our results indicate that a promising avenue to trace the details of the disk's assembly is via a full interpretation of the empirical connections we report.

Unified Astronomy Thesaurus concepts: Milky Way Galaxy (1054); Red giant clump (1370); Stellar populations (1622); Clustering (1908); Stellar abundances (1577); Astrostatistics techniques (1886)

1. Introduction

The Milky Way's stellar mass is concentrated in the disk, the chemodynamical characteristics of which can constrain the Galaxy's formation history. The spatial and kinematic distribution of the Milky Way's disk has historically been separated into two populations, termed the "thin" and "thick" disk (Gilmore et al. 2002). These two populations have also been found to have different chemical properties, with the thin and thick disk structures having lower/higher [Fe/H] and higher/lower [α/Fe] respectively (e.g., Fuhrmann 1998; Bensby et al. 2014). However, the disk's distribution has also been argued to be better represented by a set of populations with a continuous sequence in morphological parameters, rather than described by a strict binary division (Bovy et al. 2012).

The advent of large spectroscopic surveys has provided the opportunity to examine the disk's chemodynamical structure in new ways. The Apache Point Observatory Galactic Evolution Experiment (APOGEE) survey, in particular Majewski et al. (2017), has afforded important insight into the disk, due to its deep infrared coverage (Zasowski et al. 2014), medium-resolution spectra ($R = 22{,}500$; Holtzman et al. 2015; García Pérez et al. 2016) and large number of stars (>300,000 in data release 14 (DR14); Blanton et al. 2017; Majewski et al. 2017; Abolfathi et al. 2018). Studies can now traverse a larger spatial extent (e.g., Nidever et al. 2014; Hayden et al. 2015) and simultaneously, across a diversity of chemical elements, beyond a bulk [Fe/H] and [α/Fe] (e.g., Weinberg et al. 2019).

APOGEE revealed bimodality in the two-dimensional [Fe/H]–[α/Fe] plane, with the stars clearly divided into a "high-α" and "low-α" sequence. These two sequences of high- and low-α stars have different ages and spatial and kinematic distributions. The relative fraction of high- and low-α stars varies depending on disk height and radius, with old high-α stars living primarily in $0.5 < |z| < 2$ kpc and $3 < R_{\rm GAL} < 11$ kpc and younger low-α stars living in $|z| < 1$ kpc for $3 < R_{\rm GAL} < 9$ kpc and evenly spread throughout $|z|$ for $R_{\rm GAL} > 9$ kpc (e.g., Nidever et al. 2014; Hayden et al. 2015; Silva Aguirre et al. 2018). Using stars that are sampled by APOGEE and Gaia, the high- and low-α sequences are seen to exhibit different orbital properties as well as age–velocity dispersion relations and distributions (e.g., Mackereth et al. 2019), including at fixed age (Blancato et al. 2019; Gandhi & Ness 2019).

It is tempting to associate these two sequences defined chemically directly with the thin and thick disks that have been defined in the stellar spatial and kinematic distribution, in apparent confirmation of the bimodal picture. However, it has become clear that the spatial-to-chemical mapping is discordant across large spatial extents, with the geometry of the two disks defined spatially differing from those where the separation is made in the chemical plane (Bland-Hawthorn et al. 2019).

The observed spatial and kinematic differences between the high- and low-α stars have led to the interpretation that there are two discrete modes of chemical evolution in the disk, with much debate regarding the origin of the sequences. Simulation work (e.g., Mackereth et al. 2018; Clarke et al. 2019) has been pursued to find the sources of these differences. From 133 Milky Way-like Evolution and Assembly of GaLaxies and their Environments (EAGLE) simulations, Mackereth et al. (2018) find that the Milky Way bimodality in chemistry is rare, evolving in only 5% of the simulation galaxies. With their simulations, they discuss how the bimodality is a consequence of an early gas accretion episode. On the other hand, Clarke et al. (2019) demonstrate with a Gasoline simulation that the bimodality is natural if the early gas-rich disk fragments, causing a mixture of clumpy and distributed star formation. The clumps enrich rapidly and migrate from the low- to high-α sequence due to the clumps' high star formation rate, while the distributed star formation produces the low-α sequence. Using a chemical evolution model, Lian et al. (2020) suggest that the inner disk formed from two main starbust episodes, with about 4 Gyr of secular, low-level star formation activity in between. The high-α and metal-poor low-α sequences are formed from the first and second starbust respectively, while the secular evolution phase is responsible for the metal-rich low-α stars.

Many different methods for dividing the stars in the [α/Fe]–[Fe/H] plane have been used to aid with interpretation and comparison to simulations (e.g., Masseron & Gilmore 2015; Blancato et al. 2019; Weinberg et al. 2019). In addition to density contours, Masseron & Gilmore (2015) used the [C/N] ratio to confirm their high- and low-α sequence populations. Weinberg et al. (2019) traced the groups in a two-dimensional density distribution and followed a shallow valley separating the two sequences in their division. Blancato et al. (2019) employed a more robust separation through a soft clustering using a Gaussian mixture model in the [α/Fe]–[Fe/H] plane, but found that their results were sensitive to the initialization parameters.

There are several limitations to these simple divisions in [α/Fe]–[Fe/H]. There is no standard way to separate the stars into the two sequences, and many approaches rely on an ad hoc division-by-eye. Moreover, any chemodynamical analysis of the disk that uses only [α/Fe]–[Fe/H] potentially omits important features from the larger chemical space. Perceiving two clusters in the two-dimensional plane does not necessarily mean that there are only two populations in the larger abundance space, and this approach may effectively marginalize over potentially physically meaningful sub-structures or signatures.

Large surveys that measure many dimensions of chemical-space, like APOGEE, offer the opportunity to investigate abundance distributions more generally. Working with high-dimensional data is nontrivial as the dimensions are difficult to visualize, there is a possibility of models over-fitting data, and the higher-dimensional space will be sparse without a large sample size—all often referred to as the curse of dimensionality. Previous work has attempted to reduce the high-dimensional data into something more manageable in an attempt to find clusters (e.g., Garcia-Dias et al. 2019). Ting et al. (2012) use principal component analysis (PCA) on 25 elements from High-Efficiency and high-Resolution Mercator Echelle Spectrograph (HERMES) in order to find the most informative lower-dimensional space for high- and low-metallicity stars. They establish a connection between the principal components and nucleosythesis mechanisms, and also allude to the fact that too many redundant dimensions reduce our ability to locate clusters. Through the employment of Expectation Maximized PCA on the full spectra, Price-Jones & Bovy (2018) find that only about 10 principal components accurately model the spectra. Recently, Casey et al. (2019) studied a data-driven model of nucleosynthesis with chemical tagging in a lower-dimensional latent space on the The GALactic Archaeology with HERMES data set. They found that $\sim$2500 stars divide into three clusters in the latent chemical abundance space; however, results were not consistent as more stars were added, indicating that a parametric Gaussian model is likely incorrect.

In this paper, we seek to leverage the >20 element abundances measured by the APOGEE survey to effectively cluster stars by their overall chemical similarity. We follow a similar approach to previous work, such as Ness et al. (2019), Bovy et al. (2016), and Price-Jones et al. (2020), by focusing on the abundances, and then analyzing the structure and ages of the clusters in abundance space. However, our contribution is unique because we are working with (i) the full [α/Fe]–[Fe/H] plane (unlike Ness et al. 2019), (ii) uninformed clustering rather than binning in two-dimensional abundances (unlike prior Bovy et al. 2016 work), and (iii) a larger set of 19 dimensions. We emphasize that for many analyses, it is convenient to make an ad hoc separation (e.g., Bland-Hawthorn et al. 2019) or statistical model of two populations in the [Fe/H]–[α/Fe] plane. This prior approach helps motivate our analysis, which instead explores how the data prefer to group in a larger abundance space using a clustering approach. Our results set the basis for asking the questions: (i) why do the data cluster as they do (including not into the ad hoc two populations often prescribed) and (ii) what can we learn from this?

Unlike other approaches, our approach is nonparametric and impartial to how many underlying populations comprise the data: hierarchical clustering (Ward 1963) in a 19-dimensional chemical abundance space. We choose this method rather than something like K-means (e.g., Hogg et al. 2016), as with hierarchical clustering there is no prior concern of choosing the correct number of clusters. Without being able to visualize all dimensions at once, choosing the wrong number of clusters could give rise to misleading results for a method requiring the number of clusters beforehand. Additionally, we investigate what abundances are driving the clustering results, and employ the standard data reduction techniques PCA (Pearson 1901; Hotelling 1933) and Isomap (Tenenbaum et al. 2000) to determine the most important features of the data and ensure clustering results are not over-fit. We also consider how many groups are justified as discrete populations.

This paper is organized as follows. In Section 2 we discuss the data sets used. Methods used for clustering, dimension reduction, and determining the number of stellar populations are outlined in Section 3. Section 4 briefly explores the abundance correlations, and investigates how the clusters defined from the large chemical abundance space appear in the [α/Fe]–[Fe/H] plane. In Section 5, the question of how many significantly different clusters is investigated. Finally, Sections 6 and 7 present the main conclusions and a discussion of our analysis.

2. Data

In this paper, we use APOGEE DR14 data (Majewski et al. 2017; Abolfathi et al. 2018) from Sloan Digital Sky Survey-IV (Blanton et al. 2017). We focus on the red clump (RC) stars found in the DR14 APOGEE RC catalog (Bovy et al. 2014), with their abundances processed by the APOGEE Stellar Parameter and Chemical Abundance Pipeline (ASPCAP; Pérez et al. 2016) and ages calculated by Sanders & Das (2018). Stellar parameters of $T_{\rm eff}$, $\log g$, and [Fe/H] as well as over 20 chemical element abundances are derived from the spectra with the ASPCAP pipeline.

2.1. APOGEE RC Abundances and Ages

In our analysis, we examine chemical abundance ratios for the RC stars given in the DR14 APOGEE RC catalog. RC stars span a narrow parameter space in effective temperature, $T_{\rm eff}$, and surface gravity, $\log g$. This avoids systematic effects induced in abundances by astrophysical diffusion (Liu et al. 2019) or imperfect stellar modeling (Weinberg et al. 2019), creating a high-fidelity sample of stars. As detailed in Bovy et al. (2014), the RC catalog was created through a combination of selection cuts in surface gravity, $T_{\rm eff}$, metallicity, and dereddened color $[J - K_s]_0$. Contamination from red giant branch stars is estimated to be less than 5%.

We also consider the stars' associated ages and errors calculated by Sanders & Das (2018). Through a Bayesian artificial neural network, posterior estimates for distances, masses, and ages are found from spectroscopic, photometric, and astrometric data. The priors assigned reflect the stars' likely population membership, with age priors differing depending on whether the star is likely located in the thin disk, thick disk, bulge, or stellar halo.

For our analysis, we select the 19 abundances from the available set of 25 in the APOGEE catalog that we deem to have the most reliable measurements. The set of abundances selected to work with consists of Ni, Co, Fe, Mn, Cr, V, Ti, TiII, Ca, Si, Mg, O, K, S, P, Al, N, CI, and C.

2.2. Quality Cuts on the Data

In order to remove outliers with anomalous abundance measurements, we consider stars with abundances $-1 \leqslant [X/Fe] \leqslant 1$ and $-1 \leqslant [Fe/H] \leqslant 1$. Similarly, since all but a handful of stars have ages greater than 0.1 Gyr, we remove the few very young stars. This provides a final sample of 27,135 RC stars with no missing parameters.

After removing outliers, the abundances still cover different ranges. For instance [Co/Fe] spans nearly the entire interval $[-1, 1]$ while [Ni/Fe] values only fall between $[-0.34, 0.39]$. This variance in the range leads to biases in dimension reduction techniques such as PCA, which we employ in this paper. In order to combat this, we standardize the data to give each abundance mean 0 and standard deviation 1. From here on we refer to these standardized abundances simply as abundances.

Our choice to work in the ([Fe/H], [X/Fe]) plane (for our clustering and dimension reduction approaches) is motivated by capturing the similarity of elements, with respect to the reference of (primarily) supernovae Ia (SNe Ia) enrichment. As discussed in Ting et al. (2012), the [X/Fe] space is less correlated than [X/H], so it can potentially better capture the subtle variations to differentiate populations that are most chemically similar.

3. Methods

In this section, we describe the methods used in our analysis. We first wish to examine how the data are organized into groups using a clustering algorithm. Further, how do the groups obtained with a clustering approach compare to the high- and low-α sequences as they are typically defined? The implementation of our selected clustering method, hierarchical clustering, is described in Section 4.1. In Section 4.3, we compare the clustering performed in 19 dimensions to groups determined using lower-dimensional representations of the data produced by PCA and Isomap. These best capture the most important features of the data and are a test of how robust our results are when done using the full dimensionality of the available data. We also address the question of how many stellar populations our data set can be decomposed into with clustering, and what this means. We examine how many groups justifiably exist in the abundance space using the Gap Statistic metric (Section 5) and Dunnett's modification of the Tukey–Kramer method (Section 5.1).

3.1. Clustering Methodologies

3.1.1. Agglomerative Hierarchical Clustering

Agglomerative hierarchical clustering is a clustering method that organizes stars into groups (Jain & Dubes 1988), working from individual data points (in our case each star with its set of abundances) that hierarchically merge into groups (in our case defined by their abundance similarity). The resulting clusters contain the stars that are most similar to one another, while the clusters themselves are most dissimilar from each other. The metric that splits the difference between what is similar and different is left up to the user. We discuss our choice of metric in Sections 4.1 and 5.1.

Many clustering methods, such as K-means (Hartigan & Wong 1979), require prior knowledge for how many clusters comprise the data. Hierarchical clustering, however, does not require a predefined number of clusters to be set. Rather, it hierarchically builds groups from the number of data points, N, to a single group, in a tree-like structure. That is, it first merges the most similar two stars and moves up the hierarchy until only one cluster remains after combining the least similar groups. The user can employ a metric to indicate the maximum number of different clusters that are justified given the data. The output is expressed in a so-called dendrogram, or tree diagram. This dendrogram illustrates the hierarchical arrangement of the clusters and allows us to visualize the similarity structure of the data. A dendrogram shows the transition from $N = 27{,}135$ groups to 1, where all 27,135 stars are plotted as the leaves of the tree and groups combine at certain "heights" to create nodes. In the metric we choose to implement for the clustering, the node heights represent the total within cluster sum of squares error for clusters $K = 1, \ldots, 27{,}135$,

$$\sum_{k=1}^{K} \sum_{i \in S_k} \sum_{j=1}^{19} (x_{ij} - \bar{x}_{kj})^2,$$

where S_k is the set of stars in cluster k, x_{ij} is the jth abundance of star i, and $\bar{x}_{kj}$ is the average of the jth abundance for the kth cluster. While a larger node height indicates the combining groups are more dissimilar, the difference between subsequent node heights can be thought of as a potential; the larger the potential, the more confidently we can say that the two adjoining groups are distinct. Once the user views the structure of the dendrogram, they are able to decide an appropriate number of clusters (Ward 1963). Thus hierarchical clustering suits not only our goal to cluster the data into two sequences, but also to determine if the RC sample forms more than two populations in the larger abundance space.

We choose to cluster the data via agglomerative hierarchical clustering, using Ward's minimum variance criterion (Ward 1963), which minimizes the total within-cluster variance at each step. This dissimilarity measure maximizes the distances between clusters while minimizing the distances between stars within a cluster (Ward 1963). This measure aligns with our goals for classifying stellar populations, where stars most similar in chemical space should be grouped together.

Specifically, we use the Ward2 algorithm described in Kaufman & Rousseeuw (2009) and Murtagh & Legendre (2011). Beginning with each star as its own cluster, at each step we combine the pair of clusters that leads to a minimum increase in total within-cluster variance until only one large cluster containing all the stars remains. The explicit steps are given as Algorithm 1 in the Appendix.

3.1.2. PCA

PCA (Pearson 1901; Hotelling 1933) is a linear dimension reduction technique. In our case, this transforms the data from a

19-dimensional abundance space to a much smaller set of variables that contain the most information about the data. The new smaller linearly uncorrelated set of variables, or principal components, are linear combinations of the 19 abundances. The first of these components is in the direction of the largest spread in the data. The second component, third component, and so forth, lie in the direction of the maximum variance subject to being uncorrelated to the previous components (Jolliffe 1990; Ringnér 2008).

As discussed in Section 2.2, we first normalize, or standardize, each variable to have mean 0 and standard deviation 1. Before standardization, each variable spans a different range and, if used directly, this would prevent each abundance from having equal importance in the analysis. Instead, a few variables would dominate and skew the results. Therefore, the 19×19 abundance covariance matrix that is calculated for PCA is computed using the standardized abundances as inputs. The covariance between any two abundances, x and y, is given as

$$\frac{1}{n-1}\sum_{i=1}^{n} x_i y_i.$$

The eigenvectors and eigenvalues of this covariance matrix are then calculated. It is established that the largest eigenvalues contain the most useful information regarding the spread of the data, and that the smallest eigenvalues primarily capture the noise. In fact, the eigenvalues capture the amount of variance explained by the associated eigenvector (Wold et al. 1987). Therefore, the first principal component is the eigenvector corresponding to the largest eigenvalue and so on (Jolliffe 1990). Additionally, the principal components are scaled to have unit norm in order to easily compare the amount each abundance contributes to a specific component.

3.1.3. Isomap

While PCA is beneficial as a linear dimension reduction methodology, a more generalized approach is to assume that our data lie on an embedded nonlinear manifold within the 19-dimensional space. This assumption requires a more flexible model that is able to capture additional (nonlinear) structure in the data, which may not be revealed with PCA. The nonlinear dimension reduction technique, Isomap (Tenenbaum et al. 2000), finds a low-dimensional embedding of the data that preserves the geodesic distances from 19 dimensions. Here, the geodesic distances represent the distances between stars in a neighborhood graph, as explained below. With the advantage of being less sensitive to noise, Isomap more reliably represents the data's global structure in the new lower dimensions (Tenenbaum et al. 2000; Silva & Tenenbaum 2003; Rosman et al. 2010). By comparing the results to PCA, we hope to understand the key features of the data.

The algorithm is introduced in Tenenbaum et al. (2000) and contains three primary steps. It is detailed in Algorithm 2 in the Appendix and described here. The first step establishes the five nearest neighbors for each data point. Star j is a neighbor of star i if it is one of the five closest stars to star i in the 19-dimensional abundance space using a Euclidean distance calculation. This creates a neighborhood graph G, where the edge length, $G_{i,j}$, is equal to the Euclidean distance between the two neighbors if star j is a neighbor of star i, and is set to 0 if not.

The second step creates a geodesic distance matrix, d_G. This approximates the shortest graph distance between star i and star j, using Dijkstra's algorithm (Dijkstra 1959), on the neighborhood graph G for all stars, i and j. Following the optimal path between nearest neighbors at a given stage, Dijkstra's algorithm finds the shortest path from star i to star j, $d_G(i, j)$. The explicit steps of the algorithm are given in Algorithm 3 in the Appendix.

In the last step, Isomap applies classical multidimensional scaling (MDS; Kruskal 1964) to the geodesic distance matrix, d_G, found in the previous step. MDS solves the problem of creating a coordinate matrix from distances provided (Kruskal 1964) and is given in detail in the Appendix as Algorithm 4. For our analysis, we choose to project our data into two dimensions.

3.2. Metrics to Quantify Cluster Significance

3.2.1. The Gap Statistic

Proposed by Tibshirani et al. (2001), the gap statistic is a method for estimating the number of clusters in a data set by formalizing the elbow method. The elbow method is explained in detail in Kodinariya & Makwana (2013). Briefly, in a plot of total within-cluster sum of squares (WSS) versus number of clusters, the elbow method requires one to infer where an "elbow" lies. The elbow method is usually subjective since in practice it can be unclear where exactly to cut off the number of clusters.

Assuming k clusters, in our analysis W_k is defined as the pooled WSS about the cluster means. The goal is to compare $\log(W_k)$ to what the expected value would be if the data were from a reference distribution, $E^*[\log(W_k)]$. The reference distribution is typically a uniform distribution taken over the range of observed values, with $E^*[\log(W_k)]$ being the mean $\log(W_k)$ for many bootstrapped samples from the reference distribution. The gap statistic is then defined as

$$\mathrm{Gap}(k) = E^*[\log(W_k)] - \log(W_k).$$

The determined number of clusters $\hat{k}$ is the smallest value k such that $\mathrm{Gap}(k) \geqslant \mathrm{Gap}(k+1) - s_{k+1}$ where s_k is the standard deviation of $E^*[\log(W_k)]$ for k clusters from Monte Carlo simulations.

3.2.2. Dunnett's Modification of the Tukey–Kramer Method for Determining Cluster Significance as a Function of Age

To find the maximum number of groups with statistically significant values that we do not use for clustering—in our case different mean ages in Section 5.1—we use Dunnett's modification of the Tukey–Kramer method to compare the mean ages of the clusters. Due to our large stellar sample, we choose a pairwise multiple comparison test over the Kolmogorov–Smirnov test, (Massey 1951) which compares distributions.

For each cluster of stars, k, we can calculate a mean age, μ_k. We wish to evaluate if our clusters are representative of populations with significantly different mean ages. To test the statistical significance of the mean age differences, we conduct a pairwise multiple comparison test, which compares the means of clusters 1 and 2, clusters 1 and 3, clusters 2 and 3, and so on. To quantify our results, we choose the confidence level to be the standard value 0.95 to create confidence intervals. A confidence level of $1 - \alpha$ means that if many samples were drawn using a given true parameter, we would expect the true

parameter to fall in $(1-\alpha)\times 100\%$ of the confidence intervals.

We specifically implement Dunnett's modification of the Tukey–Kramer method (Dunnett 1980), which allows for the clusters to have unequal sample sizes and different variances. In order to determine if two clusters i and j have significantly different mean ages, we look at the confidence interval

$$\mu_i - \mu_j \pm A_{ij,\alpha,k}\left(\frac{s_i^2}{n_i}+\frac{s_j^2}{n_j}\right)^{1/2},$$

where n_i, n_j are the number of stars in group i, j respectively, and s_i, s_j are the unbiased estimated standard deviations given by

$$s_i = \sqrt{\frac{1}{n_i-1}\sum_{x=1}^{n_i}(a_x-\bar{a}_i)^2}$$
$$s_j = \sqrt{\frac{1}{n_j-1}\sum_{x=1}^{n_j}(a_x-\bar{a}_j)^2}$$

for ages a with $\bar{a}_i$ and $\bar{a}_j$ being the mean ages for group i and j. Let q_α, k, ν be the critical value of a Studentized range distribution—with parameters k clusters and degrees of freedom ν—that gives the boundary of the acceptance region for the test with confidence level $1-\alpha$. Then, $A_{ij,\alpha,k}$ is a weighted average of the critical values given by

$$A_{ij,\alpha,k} = \frac{1}{\sqrt{2}}\frac{q_{\alpha,k,n_i-k}s_i^2/n_i + q_{\alpha,k,n_j-k}s_j^2/n_j}{s_i^2/n_i + s_j^2/n_j}.$$

In order to determine the largest number of clusters with distinct age distributions, we start with two clusters and increase the number of groups until the mean ages are no longer significantly different. We begin by evaluating if μ_1 is significantly different than μ_2 when the data are split into two clusters. We claim the two cluster means are significantly different if the confidence interval does not contain 0, and show no significant difference if 0 is included. If 0 is not contained in the interval, we then split the data into three clusters and compare μ_1 to μ_2, μ_1 to μ_3, and μ_2 to μ_3. If every one of these tests returns means that are significantly different then we proceed onto four clusters and compare μ_1 to μ_2, μ_1 to μ_3, μ_1 to μ_4 and so on. We continue to increase the number of clusters until one of the pairwise tests concludes that at least two means are not significantly different. Say this happens when there are K^*+1 clusters. Then the maximum number of clusters with distinct ages is K^*.

4. Results I: Clusters in Abundance Space Projections

Using Mg as our representative α-element, the bimodality of our sample in the [Mg/Fe]–[Fe/H] plane is clearly visible in a scatter plot of the data presented in Figure 1. Following Weinberg et al. (2019) and Blancato et al. (2019), we split the data by eye with two straight lines of differing slopes (orange line), joined at [Fe/H] ≈ 0. Hereafter, we label the 3211 stars above and the 23,924 stars below the orange line respectively as the high- and low-α sequence.

In this section we examine how two clusters defined in the full 19-dimensional chemical abundance space compare to the high- and low-α sequences determined by eye, after projection into the [Mg/Fe]–[Fe/H] plane. We then investigate the true dimensionality of the data, and cluster again in representative lower-dimensional hyperplanes to demonstrate that the clusters from 19 dimensions are genuine. We additionally explore the sequences in many different abundance planes to determine the key elements contributing and informing the clustering. We defer exploration of age and spatial relationships to Sections 5.1 and 5.2.

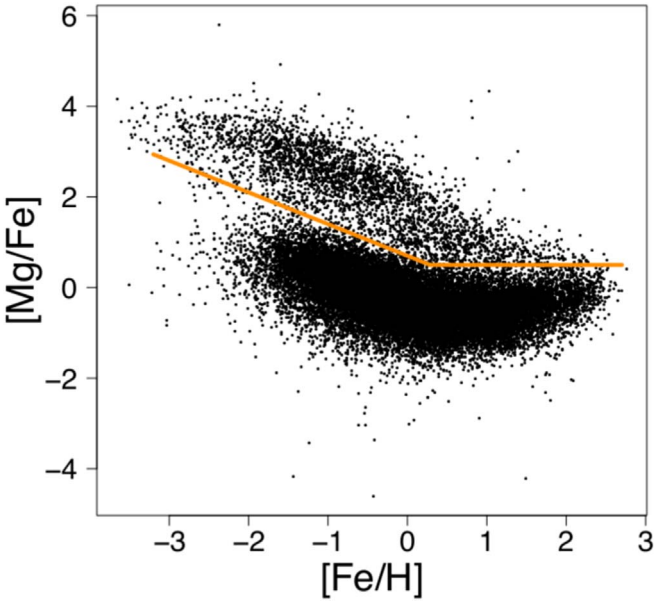

Figure 1. Our sample of 27,135 RC stars projected into the standardized [Mg/Fe]–[Fe/H] plane. The expected bimodality associated with the high- and low-α sequences is evident. Typically, the two sequences are divided by eye. One such division of the stars is given by the orange line, with the high-α sequence being above and the low-α sequence being below. The traditional bend at [Fe/H] ≈ 0 used by Weinberg et al. (2019) outlines the low-α stars to create the recognizable banana-like shape. We will refer back to this reference split of the sequences many times throughout this paper.

4.1. Clustering the Stars Using Their 19 Abundance Dimensions

We first cluster our 27,135 stars using their vector of 19 abundances with the hierarchical clustering method described in Section 3.1.1. We project the clustering hierarchy from individual stars to a single grouping in the dendrogram, shown in Figure 2. The y-axis of the dendrogram represents the potential difference between clustered groups. As discussed in Section 3.1.1, the larger this difference, the more significant the group classification. To first examine the most simple clustering case, we find a large jump in the total WSS from 457.2 to 281.9 in going from one to two clusters, which splits the data into two samples of 2292 and 24,843 stars, respectively.

The projection of these two clusters in the [Mg/Fe]–[Fe/H] plane is given as the leftmost plot in Figure 3. The two clusters look very similar to the previously defined high- and low-α sequences, with the black stars (cluster 1) corresponding to the high-α sequence and the green stars (cluster 2) corresponding to the low-α sequence, and part of what is typically included as the high-α sequence (highlighted as yellow stars in top left of Figure 8). While these two sequences are similar to the by-eye division, the sub-group of just over 900 stars with $0 \leqslant$ [Fe/H] $\leqslant 2$ and $0.5 \leqslant$ [Mg/Fe] $\leqslant 1.5$ is assigned to the same cluster as the bulk of the low-α stars.

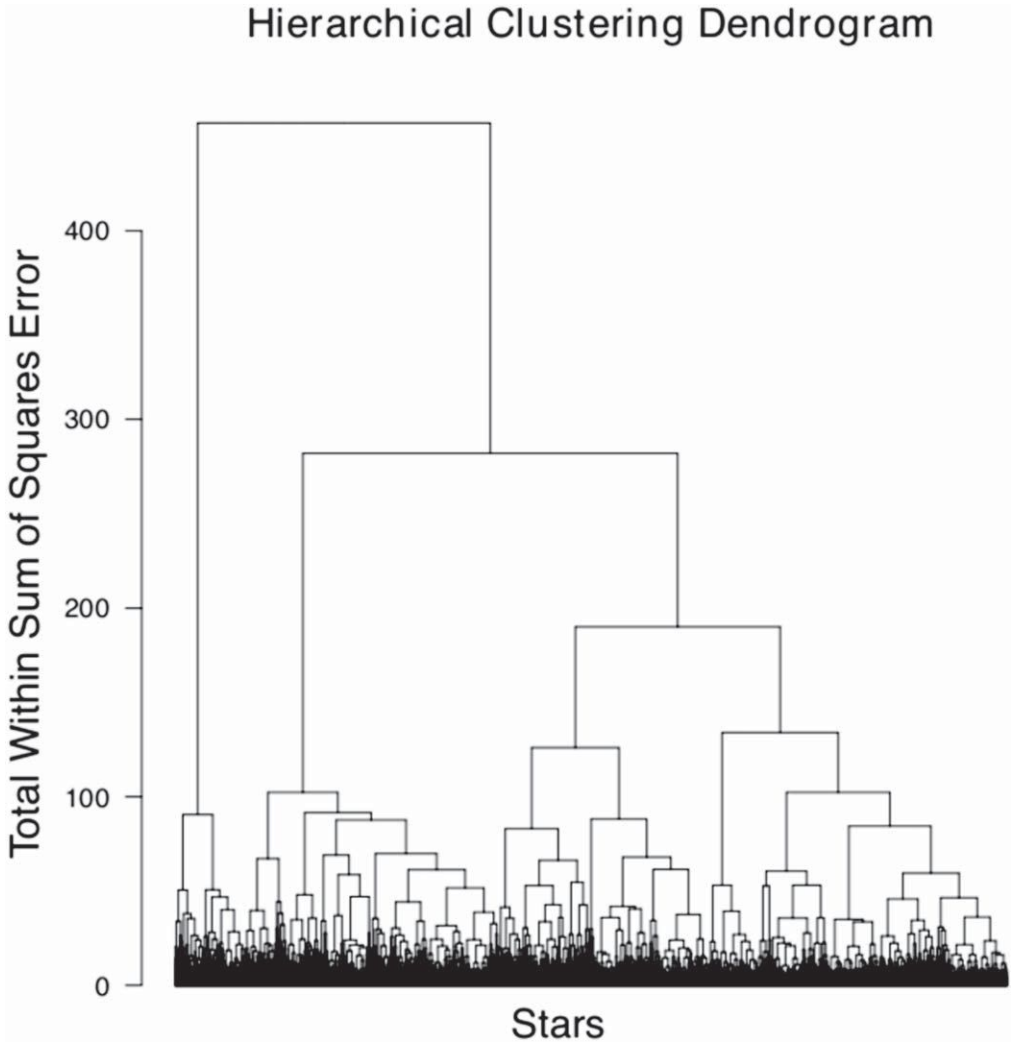

Figure 2. Dendrogram produced via agglomerative hierarchical clustering using Ward's minimum variance criterion on 27,135 RC stars. This dendrogram presents the clustering structure of the data, where the height of the *y*-axis represents the total within-cluster sum of squares (WSS) for the given number of clusters. The bottom (at height 0) is where each star is its own individual cluster and agglomerates to the top (at height ∼450) where where all the stars combine into one main cluster. In order to best determine the number of clusters our sample is comprised of, we compare the relative total WSS when successive groups combine. The total WSS for when there is one cluster is 457, and jumps down to 282 when the data split into two clusters. This jump is subjectively large, and justifies two main populations in our sample. However, the differences between total WSS for two through six clusters can all be considered "large," and thus make finding the number of meaningful populations in our data set nontrivial.

We validate our clustering results are robust to the underlying density distribution of the data by sampling a uniform number of stars across the [Fe/H]–[α/Fe] plane and repeating our analysis, which gives consistent results. Additionally, while we use the L2 norm throughout this paper, the main results are consistent for different norms.

4.2. Correlation Structure of the Data

In the previous subsection, we showed the two main clusters in 19 dimensions projected into the two-dimensional [Fe/H]–[Mg/Fe] plane. In order to determine the validity of the groupings found in 19 dimensions, we continue our analysis by investigating the structure of our data sample. Examining the correlation structure of each pair of abundances allows us to see if the data lie in a higher-dimensional ambient space than needed for doing work on the data such as clustering. Since our variables are normalized to have 0 mean and a standard deviation of 1, the correlation coefficient between abundance vector x and abundance vector y is

$$r_{x,y} = x \cdot y.$$

Figure 4 shows the correlation structure of the data that we calculate, in a matrix organized by elements divided into their families. A darker shade of either red or blue illustrates a stronger positive or negative linear relationship between the corresponding abundances, whereas a lighter shade indicates a very weak linear relationship between the two.

Our intention with this analysis is not to make strong statements about the strength of correlations within families. Rather, the purpose is to show connections between elements of the vector of chemical abundances. We will subsequently use the PCA reduction for a comparative analysis, to test how robust our results are when our analysis is performed using the full vector of elements (and dimensions) of the available data.

In order to determine if there are any intra-family relationships in the [X/Fe] space that suggest linear sub-structure, we categorize the abundances into four groups according to how

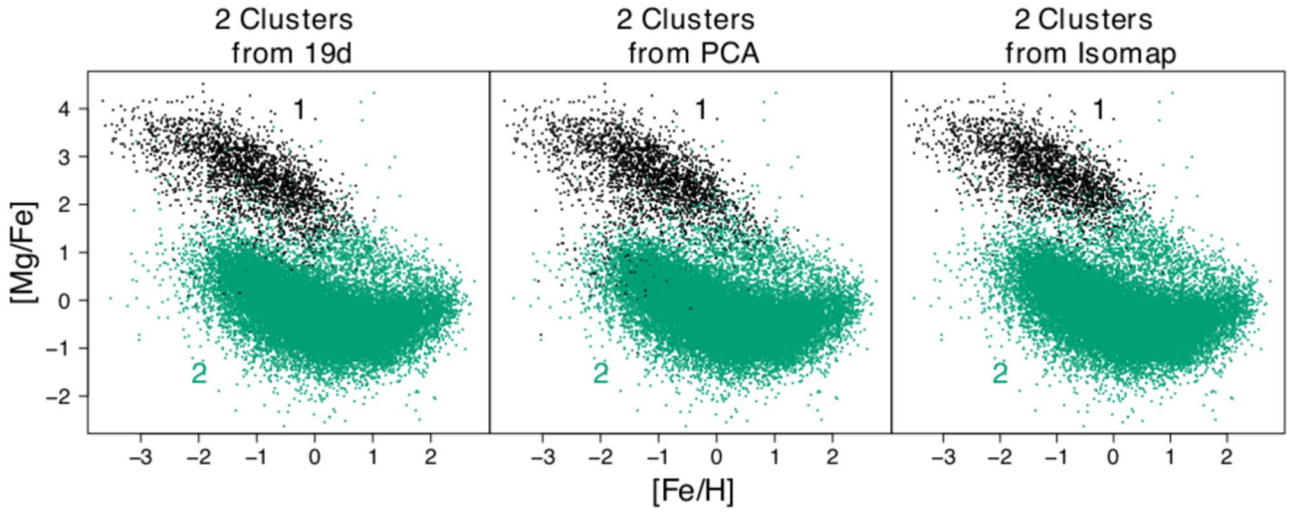

Figure 3. Two-dimensional projection in the standardized [Fe/H]–[Mg/Fe] plane of the two clusters that are found using agglomerative hierarchical clustering, with Ward's minimum variance criterion for the 27,135 RC stars, using (left) the 19-dimensional chemical abundance space, (middle) the first two dimensions of PCA decomposition, and (right) the first two dimensions of Isomap decomposition. We refer to these two clusters as 1 (black) and 2 (green). In all three plots there is a group of stars between $0 \leqslant$ [Fe/H] $\leqslant 2$ and $0 \leqslant$ [Mg/Fe] $\leqslant 1.5$ that typically are visibly classified as high-α stars, but more closely resemble the low-α sequence in 19 dimensions and the first two main dimensions of lower-dimensional embeddings. The stability of this grouping over the three different methods suggests that the results are a real yet unexpected feature of the data when using this algorithmic approach to clustering.

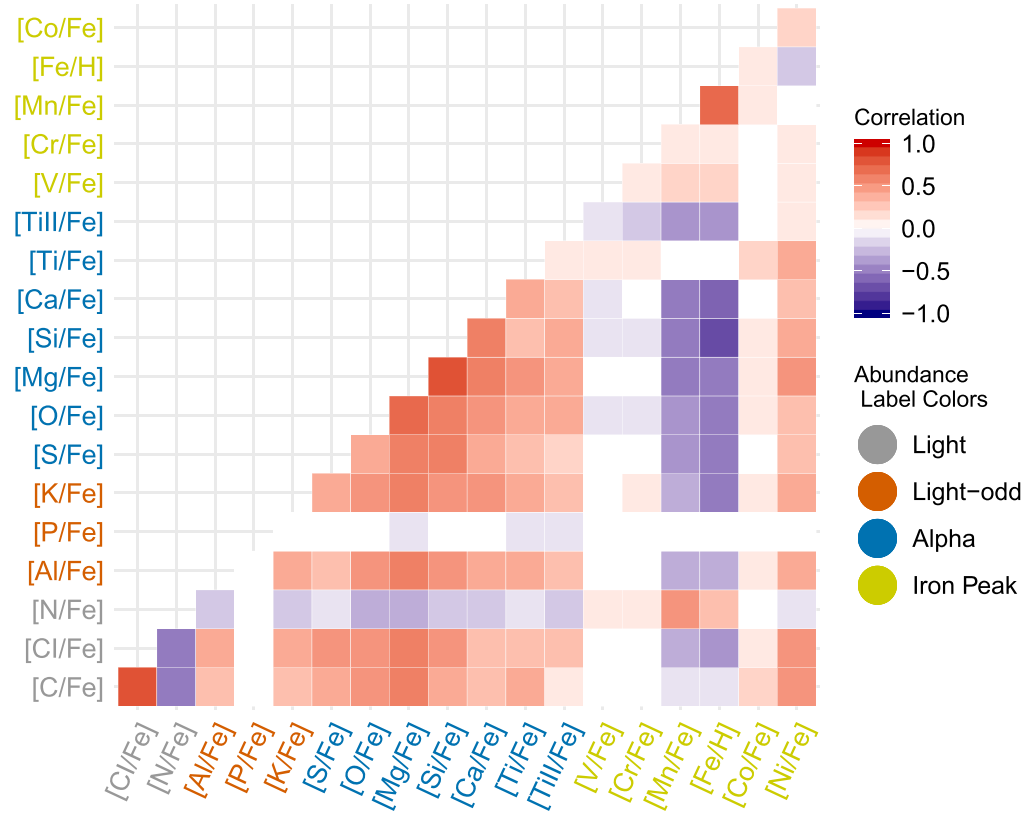

Figure 4. Correlation structure in the standardized abundances from the APOGEE red clump catalog for 27,135 stars. The correlations are quantified using the scale shown at the right. The abundance labels are colored according to their nucleosynthetic family and the elements are ordered within these families. Primarily, families either have positive correlations among their members (i.e., α- and iron-peak elements), or they have strong correlations that are both positive and negative (i.e., light elements). The strong intra-family relationships are indicative that the true dimensionality of the data is <19 dimensions of the individual abundances that are measured, and motivate our use of dimension reduction techniques.

the elements were produced: light elements with even atomic number (C, CI, N), light elements with odd atomic number (Al, P, K), α-elements (S, O, Mg, Si, Ca, Ti, Ti II), and iron-peak elements (V, Cr, Mn, Fe, Ni). Figure 4 reveals that there do appear to be connections within the different familial groups. The α-elements are fairly strongly positively correlated with one another, while the three light elements with even atomic number are very strongly linearly related. On the other hand,

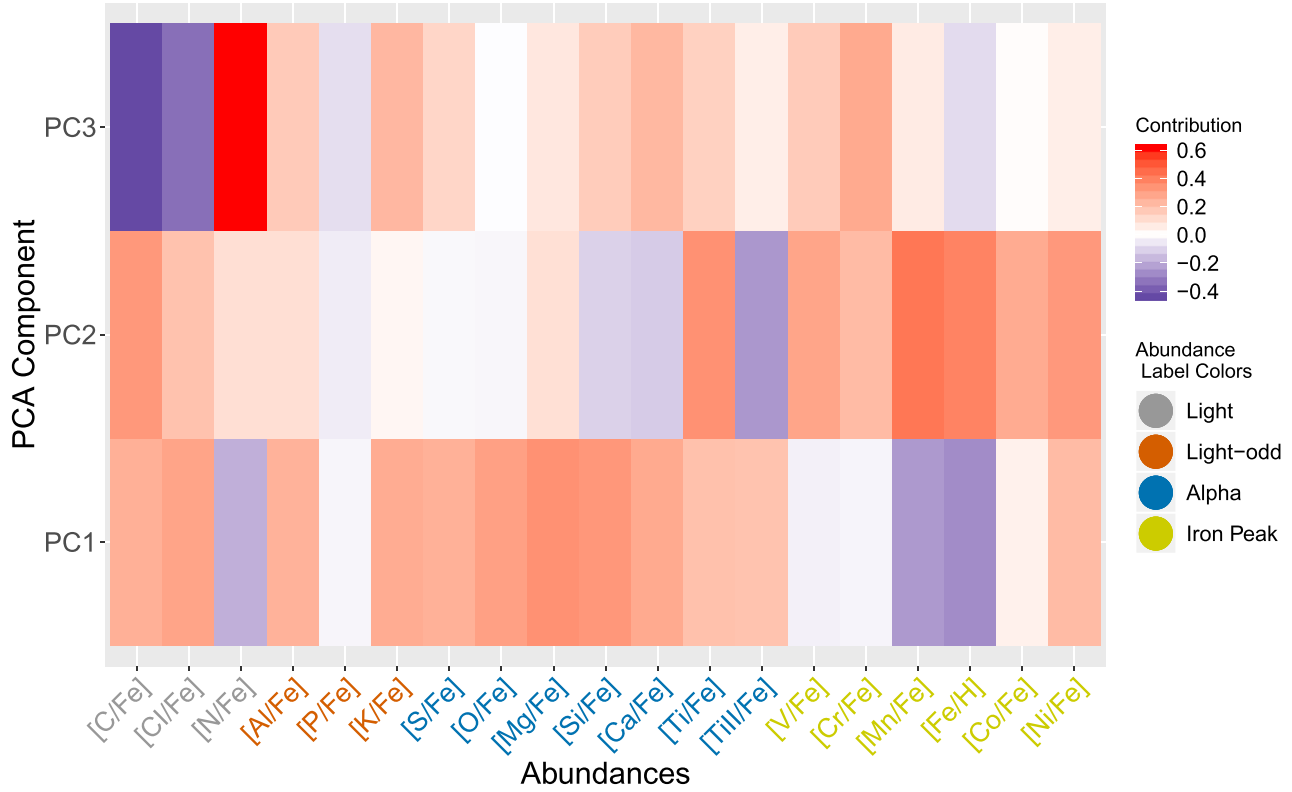

Figure 5. Depiction of how much each of the 19 chemical abundances contributes to the first three principal components. The first three components explain just over 50% of the variance of the 27,135 red clump stars in the 19 dimensions. Grouping and coloring the abundances by nucleosynthetic family allows us to quickly identify that the α-elements all positively contribute to the first component, the second component primarily captures the iron-peak elements, while the light elements with even atomic number contribute to the third.

the iron-peak elements are only weakly correlated with each other in this space. This is not particularly surprising since the abundances are normalized against Fe. Therefore these elements mostly represent the scatter due to measurement uncertainties (in the <0.05 dex precision regime). Looking at specific elements, [N/Fe], [Fe/H], and [Mn/Fe] are anti-correlated with nearly every other abundance. The poorly measured abundances in APOGEE ([P/Fe], [V/Fe], and [Cr/Fe]) have no strong correlation with any of the abundances, indicative of their large uncertainties in their measurements.

Due to some elements having quite strong linear correlations, we see from this simple investigation that the true dimensionality of the data is lower than the number of elements measured and considered in this work. Therefore, we proceed with implementing dimension reduction techniques in our clustering approaches.

4.3. Investigating Lower-dimensional Embeddings

4.3.1. Exploring the Influential Dimensions

Section 4.2 demonstrated that the abundances are highly correlated. Therefore, we now implement PCA on our 19-dimensional abundance data for our 27,135 stars to identify the primary features that dominate the clustering hierarchy. We find that the first three principal components explain 53% of the variance in the data. Therefore, focusing on these components allows us to examine how the abundances contribute to the key aspects of the data's distribution in higher dimensions. The relative contributions of each element to the first three principal components is shown in Figure 5. This figure demonstrates that the α-elements all positively contribute to the first principal component (PC1), the iron-peak elements all positively contribute to the second principal component (PC2), and the light elements with even atomic number are the main contributors to the third principal component (PC3). [P/Fe] does not strongly contribute to any of the first three components, rendering it insignificant in analyzing the data set's variability. It should also be noted that [Mg/Fe] provides the most absolute contribution to PC1, which is reassuring for [Mg/Fe] being a sound representative or proxy for an overall α-element in considering the [α/Fe]–[Fe/H] plane.

We now examine the three principal components from Figure 5 in two-dimensional projections. We show the first two components and the first and third components in the left and right of Figure 6 respectively to validate the significance of examining two groups. In both projections, the stars (colored according to their grouping defined in Section 4.1) are split into two primary populations, with each group corresponding to either the black (cluster 1) or green (cluster 2) cluster with little crossover. In addition to being spatially discrete in these planes, the two sequences show different behaviors in these projections. The black cluster (cluster 1) is evenly distributed along PC1 and PC2 and shows a strong positive correlation between PC1 and PC3, while the green cluster (cluster 2) is evenly distributed among all three components. This confirms the populations contain distinct differences.

4.3.2. Determining the Validity of the Clustering Model

One concern in working with high-dimensional data is the risk of over-fitting due to the curse of dimensionality. In order to determine if the group of $\sim$900 stars discussed in Section 4.1 are "misclassified" as a side effect of working in 19 dimensions, we cluster the stars using only the first two principal components (PC1 and PC2) rather than their 19 abundances as previously

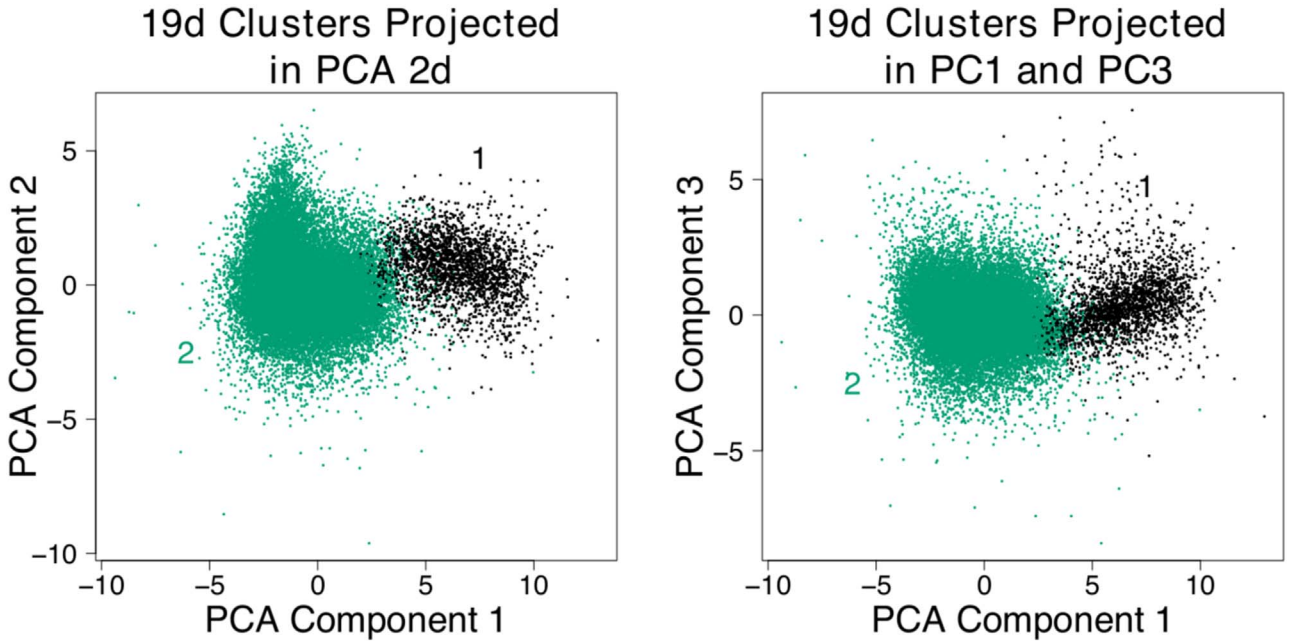

Figure 6. Red clump sample of 27,135 stars projected into (left) the first two components of principal component analysis (PCA), which explain 46% of the variance in the 19-dimensional data, and (right) the first and third PCA components. The stars are color coded according to their clusters found in 19 dimensions using agglomerative hierarchical clustering with Ward's minimum variance criterion. These clusters coincide with the two noticeable groups created in both planes. Since the first three components capture more than half of the data's variability from 19 dimensions, our choice of first examining two populations is validated.

done. We choose to only use the first two components as PC1 and PC2 together explain 46% of the variance of the data, thus capturing nearly half of the data's spread and ensuring we only examine the main features of the data. We again project the resulting two clusters back into [Mg/Fe]–[Fe/H] plane as shown in the middle panel of Figure 3. The results are almost identical to clustering in 19 dimensions: 89% of the small group of ∼900 stars in the green cluster (cluster 2) that join the black cluster (cluster 1) at high-α values are similarly identified as being part of the green cluster (cluster 2) using the first two principal components. This shows that this group of stars differentiates from the high-α sequence along the linear axes that describe the most spread.

In 19 dimensions we are unsure of the structure of the data, and a linear projection (e.g., PCA) may not be optimal. Therefore, we also run a nonlinear dimension reduction technique to compare our results with this methodology. This is the Isomap reduction discussed in Section 3.1.3. It is clear from Figure 7 that the first two components of Isomap look analogous to the first two components of PCA, demonstrating the merit of clustering into two primary groups that are consistent between different methodologies and assumed models.

Following the same approach as with PCA, we choose to cluster the stars in the first two dimensions of Isomap to examine if the two clusters are differently distributed and what happens to the group of metal-rich cluster 2 stars from the left of Figure 3, that are often associated with the high-α sequence (cluster 1). The two clusters determined with the data following Isomap dimension reduction are projected back into the [Mg/Fe]–[Fe/H] plane in the far right panel in Figure 3. Yet again, the two clusters show the same distribution as with PCA and without dimension reduction. The results are again almost identical to clustering in 19 dimensions: 97% of the small group of ∼900 stars in the green cluster (cluster 2) that join the black cluster (cluster 1) at high-α values are similarly identified as being part of the green cluster (cluster 2) using the first two first Isomap dimensions.

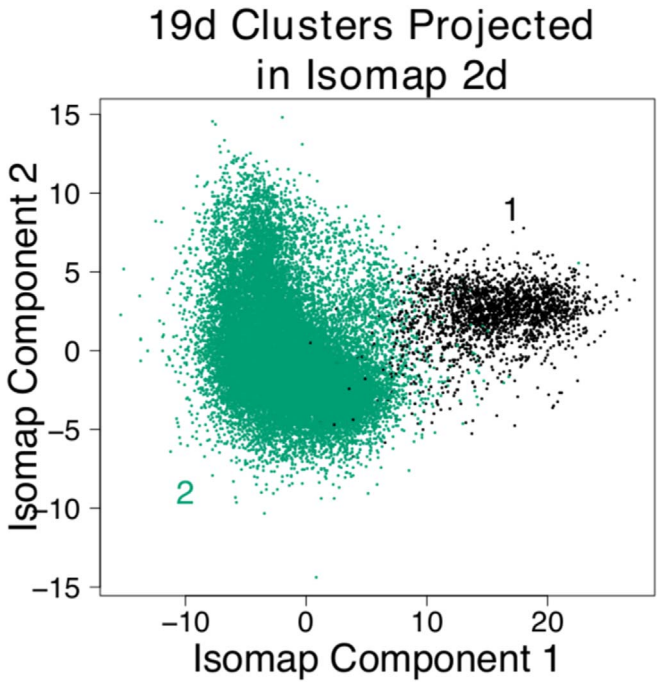

Figure 7. Red clump sample of 27,135 stars projected into the first two components of Isomap. The stars are colored according to their sequences from 19 dimensions, determined via agglomerative hierarchical clustering with Ward's minimum variance criterion. Similar to PCA, the first two components split primarily into two groups, further confirming that dividing the data into two groups is reasonable for this data set.

4.4. Projecting Our Two Clusters across Different Abundance Planes

We want to examine the projection of the two largest clusters from hierarchical clustering in different abundance planes. In particular, we are interested in the subset of

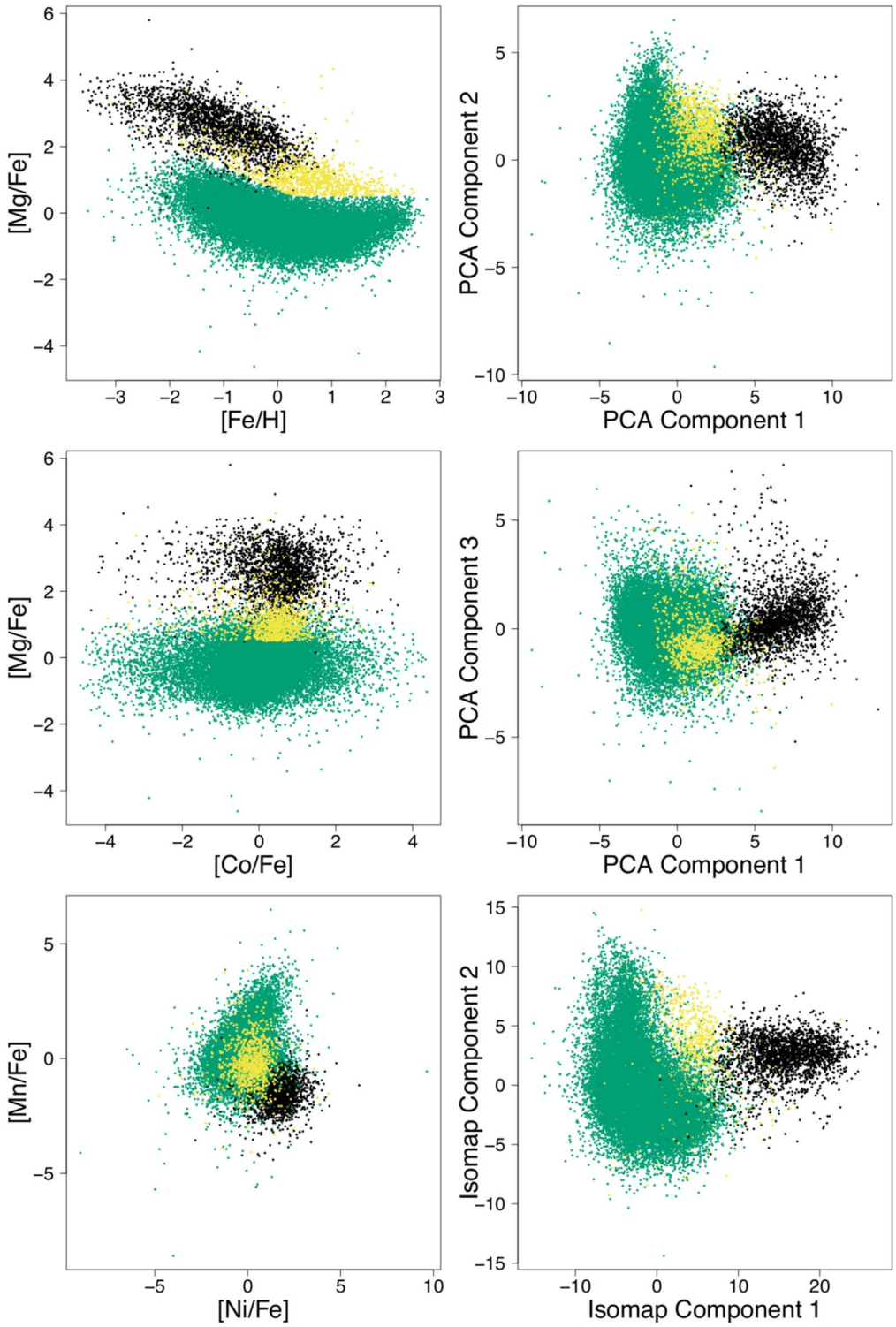

Figure 8. Red clump sample of 27,135 stars projected into (left) three abundance–abundance planes, (right) the first three components of PCA and first two dimensions of Isomap, in order to investigate where the ∼900 low-α stars (colored in yellow) live in different planes. The black group is the (higher-α) sequence of stars (cluster 1) from the left of Figure 3, while the green and yellow groups combine to create the green cluster (cluster 2) from the same figure. These plots were chosen as representatives from the full set of abundance–abundance planes. In some, there are visibly two groups of stars, where the yellow stars are part of the green group where the black group connects (middle left). In others, the yellow stars are dispersed throughout the green group (bottom left). The first two-dimensions of PCA and Isomap reveal that this group of stars is unique—closely resembling the green group but in a distinct, less dense region than the rest of the green stars. Similar to some abundance–abundance planes, PC1–PC3 shows the yellow group as an extension of the black group into the green group.

metal-rich stars that have been typically associated with the high-α sequence (see Figure 1) yet we find associated with our green cluster (cluster 2; Figure 3), which is comprised of low-α stars. We therefore differentiate this group of stars by eye from the rest of the green cluster (cluster 2) in a series of two-dimensional abundance planes to demonstrate that they sensibly align with the group the algorithm associates them with.

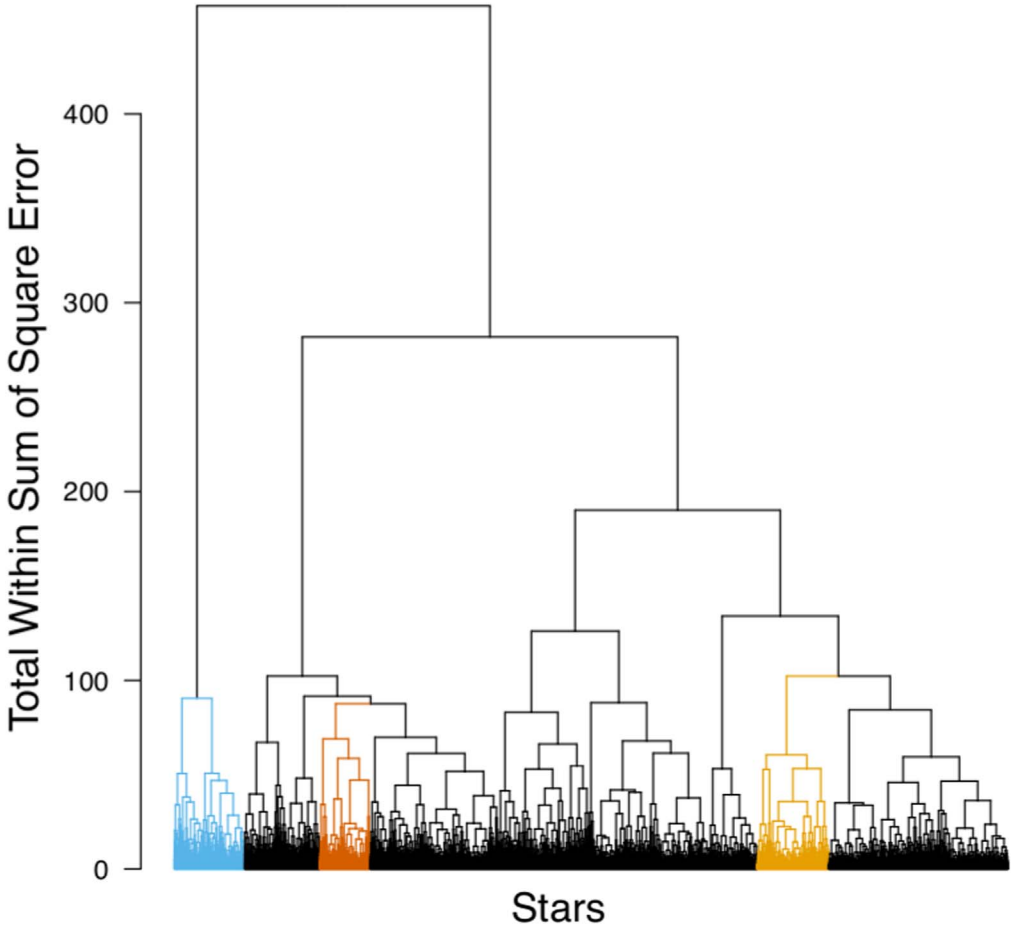

Figure 9. Nineteen-dimensional clustering dendrogram for 27,135 red clump stars created using agglomerative hierarchical clustering with Ward's minimum variance criterion. The two branches marked contain the majority of the ∼900 low-α stars that are typically classified as part of the high-α sequence in the [α/Fe]–[Fe/H] plane. The red-orange group contains ∼42% of this group of stars, followed by the orange group containing ∼20%. The other 38% are spread evenly throughout the other branches of the larger cluster. This shows that these stars contain large differences among each other, and do not show any resemblance in 19 dimensions to the more metal-poor high-α stars (cluster 1 in the left of Figure 3), which are highlighted here as light blue.

Specifically, we are interested in the yellow stars in Figure 8. After going through the $\binom{19}{2}$ abundance–abundance plots, we present in Figure 8 the unique plots that show the different scenarios in which the group of stars that are typically classified as high-α stars (black cluster) more closely resemble the low-α sequence (green cluster). Additional abundance–abundance plots are given as Figure A4 in the Appendix. In many of the plots, there are two blob-like structures, such as in the middle left plot. Here one group contains the black stars while the other mainly contains the yellow and green stars. Visually, the yellow stars are considered to be part of the green population, but they are located where the black group connects to the green one. In the bottom left plot we see that the two groups are less distinct, but here the yellow stars are condensed, located in the center of the green group. We also show at the top and bottom right of Figure 8 where the group of ∼900 stars live in the first two dimensions of PCA and Isomap. In both cases, but more strikingly in the Isomap plane, the stars are located in the less dense region of the green cluster near the connection to the black cluster. This suggests that, while these stars do more closely resemble the low-α sequence, they also contain qualities which make them unique to both clusters. From the PC1–PC3 plane (center right) the yellow stars appear to be an extension of the black cluster, but again located within the green group. These exploratory plots in Figure 8 (and the Appendix) show that the yellow subset of stars primarily coincide with the green low-α group, affirming their separation from the high-α sequence.

The natural question that follows on from this is whether this small subset of metal-rich stars is similar throughout the hierarchy of clusters, or contains distinct branches within itself. Figure 9 shows where the majority of the green clustered (cluster 2) metal-rich stars lie in the dendrogram. This subset of stars splits nearly in half when the entire sample of stars divides into three clusters, revealing that there are major differences within this unique group of ∼900 stars. Additionally, the majority of these stars lie within clusters with fairly small total WSS near 100, which is a very large jump to where the high-α sequence combines at a total WSS of 457. Between this and the fact that these stars are not all grouped together shows that they are well separated from the high-α sequence in the 19-dimensional chemical abundance space. This combined with the pairwise abundance plots indicates that a division of the abundance distribution into two clusters only is perhaps sub-optimal. That is, the data can be better described as breaking up into more than two groups.

5. Results II: Number and Properties of Significant Clusters

We have established that the data justifiably cluster into at least two groups using different approaches. Additionally, we

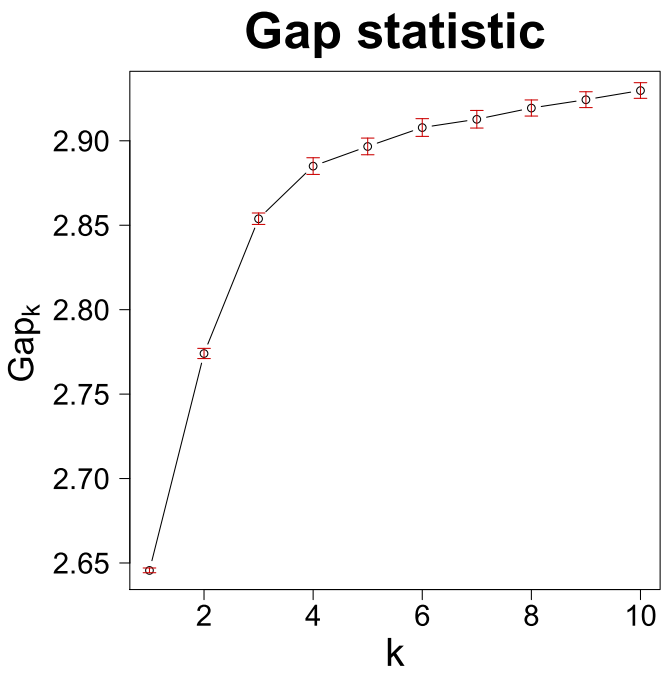

Figure 10. Gap statistic curve created using 10 bootstrapped samples of size 27,135 with 1σ error bars. The Gap statistic helps determine the number of clusters based on the 19-dimensional abundance space alone. The Gap statistic for six clusters only just meets the nominal criteria for deciding the highest number of clusters the data can be decomposed into, of being larger than the previous cluster's lower error bar. This is primarily due to the small error bars from our large sample size. This statistic, however, does suggest that there are at least six underlying populations in the data, but the Gap statistic for four through 10 clusters is very close.

have shown that these two groups are different from those typically visually assigned in the [Fe/H]–[α/Fe] plane. The next question we ask is: how many underlying clusters are our data potentially comprised of, and how do we determine this number? Given the dendrogram in Figure 2, two, three, four, and six clusters all seem reasonable for this data set given the large difference in the total WSS measurement on the y-axis. The second row of Figure 11 shows the split in the dendrogram for potential clusters for two through six groups. The top row shows the respective breakdown in the [Mg/Fe]–[Fe/H] plane.

As mentioned in Section 4.1, we compare the total WSS to infer the number of groups. Splitting the data into one, two, three, four, five and six clusters yields a total WSS of 457.2, 281.9, 190.2, 134.1, 126.1, and 102.3 respectively. To help determine the most appropriate cutoff, we examine the Gap statistic introduced in Section 3.2.1. The associated plot given in Figure 10 shows the Gap statistic, with 1σ standard deviation error bars for one through 10 clusters. The error bars are small primarily due to the large sample size. This is why we choose to run only 10 bootstrapped samples; increasing the number of bootstraps makes the error bars even more insignificant. The Gap statistic for four through 10 clusters is very close to reaching the significance threshold (a flattening of the statistic) but the turn-over is around six clusters. This suggests that the data can be meaningfully represented by up to six clusters. However, more generally, the roll of this statistic is indicative of a continuum of structure underlying the abundance distribution.

5.1. Ages As Tags of Cluster Populations

We now wish to test if the different clustered groups are also groups with different mean times of formation. Presumably, if the clustering of abundances links in any physically meaningful way to how the disk was formed over time, we might expect our statistically identified populations to have discrete ages. We find that our clusters do have different mean ages, with age distributions for the two to six clusters shown in the third row of Figure 11. Therefore, we now test the age separation as an independent metric via which we can justifiably decompose the data into discrete clusters.

The third row of Figure 11 presents the $\log_{10}$ age densities for each cluster as a function of the number of total groups. The black population of stars (cluster 1) remains unaffected when separating the data up to six clusters. It is comprised of primarily the oldest stars from the sample, and its age distribution is noticeably different from the other clusters. For the data to divide into three groups, the green group from two clusters (cluster 2; top row) splits into a left- and a right-skewed distribution which peak at the edges of the original plateau. Then the youngest stars break away from the right-skewed red cluster (cluster 3) to create four total clusters. Once the current blue group (cluster 4) separates, it becomes difficult to differentiate the middle-aged clusters. For instance the black (cluster 1), green (cluster 2), and light blue (cluster 6) densities are easy to distinguish for six populations, but the orange (cluster 5), blue (cluster 4), and red (cluster 3) densities are quite similar. Despite having similar densities, the pairwise multiple comparisons test outlined in Section 3.2.2 performed on the $\log_{10}$ ages from Sanders & Das (2018) suggest that our data separate into six populations at a confidence level of 0.95.

To investigate the sampling confidence of this result, we run 1000 Monte Carlo simulations by drawing new ages for each star, i, from a normal distribution with mean $\log_{10}$ age$_i$ and its associated error as the 1σ standard deviation. We report that 24.1%, 31.4%, and 44.5% of the simulations respectively produced four, five, and six clusters for a confidence level of 0.95. This implies that for up to six populations the stars separate into groups of different mean ages. This does not necessarily imply that there are six underlying clusters physically in the data, but rather age is a fundamental single variable by which the stars are linked together via their many abundances. Despite the age distributions being difficult to differentiate in Figure 11, our sample, clustered with abundances only, suggests that each of the six clusters has a statistically significant different mean age. This leads us to next examine the spatial distributions of the clusters.

5.2. Spatial Distributions of the Cluster Populations

We now examine the spatial distributions of our clustered groups by looking at the Galactocentric radial distance, $R_{\rm GAL}$, and the absolute distance from the mid-plane, $|z|$, (both in kiloparsecs and calculated assuming the Sun is 8 kpc from the Galactic center and 25 pc above the mid-plane). The bottom two rows of Figure 11 show that for two to six clusters, the clusters have different distributions in $R_{\rm GAL}$ and $|z|$. The structure in these distributions is driven by the APOGEE survey selection function (for example the relatively large number of stars in the specially targeted Kepler field near the Sun). It is the relative dissimilarity between the groups that is relevant and noteworthy here. This is indicative that stars that

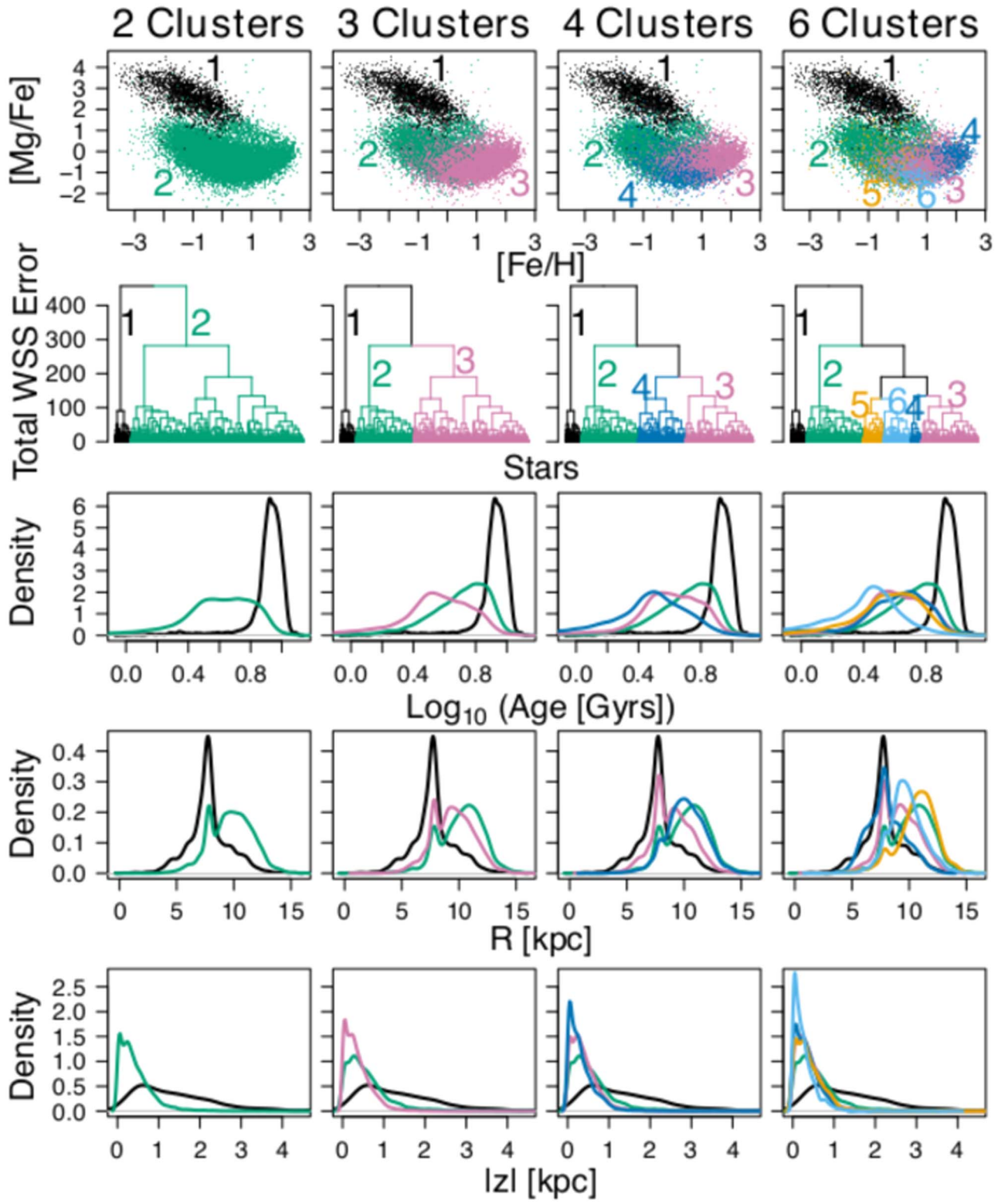

Figure 11. Colored top row: [Mg/Fe]–[Fe/H] planes, second row: dendrograms, third row: $\log_{10}$ age densities, fourth row: $R_{\rm GAL}$ densities, bottom row: $|z|$ densities for two, three, four, and six clusters. The colors order the clusters by mean age, with black (cluster 1) labeling the oldest group, then green (cluster 2), red (cluster 3), blue (cluster 4), orange (cluster 5), and finally light blue (cluster 6) defining the youngest group of stars.

are most chemically similar (according to our clustering) are also more spatially similar to one another and different between groups.

Starting with two clusters (left of Figure 11), we see the black cluster (cluster 1) is strongly peaked in $R_{\rm GAL}$ (fourth row) at $R_{\rm GAL} \approx 8$ kpc while the green cluster (cluster 2) is bimodal with a sharp peak at $R_{\rm GAL} \approx 8$ and broader peak from $R_{\rm GAL} \approx 9$–11 kpc. To create three clusters, the green cluster (cluster 2) divides into two bimodal distributions, with the new groups having their second peaks at $R_{\rm GAL} \approx 9$ (red cluster,

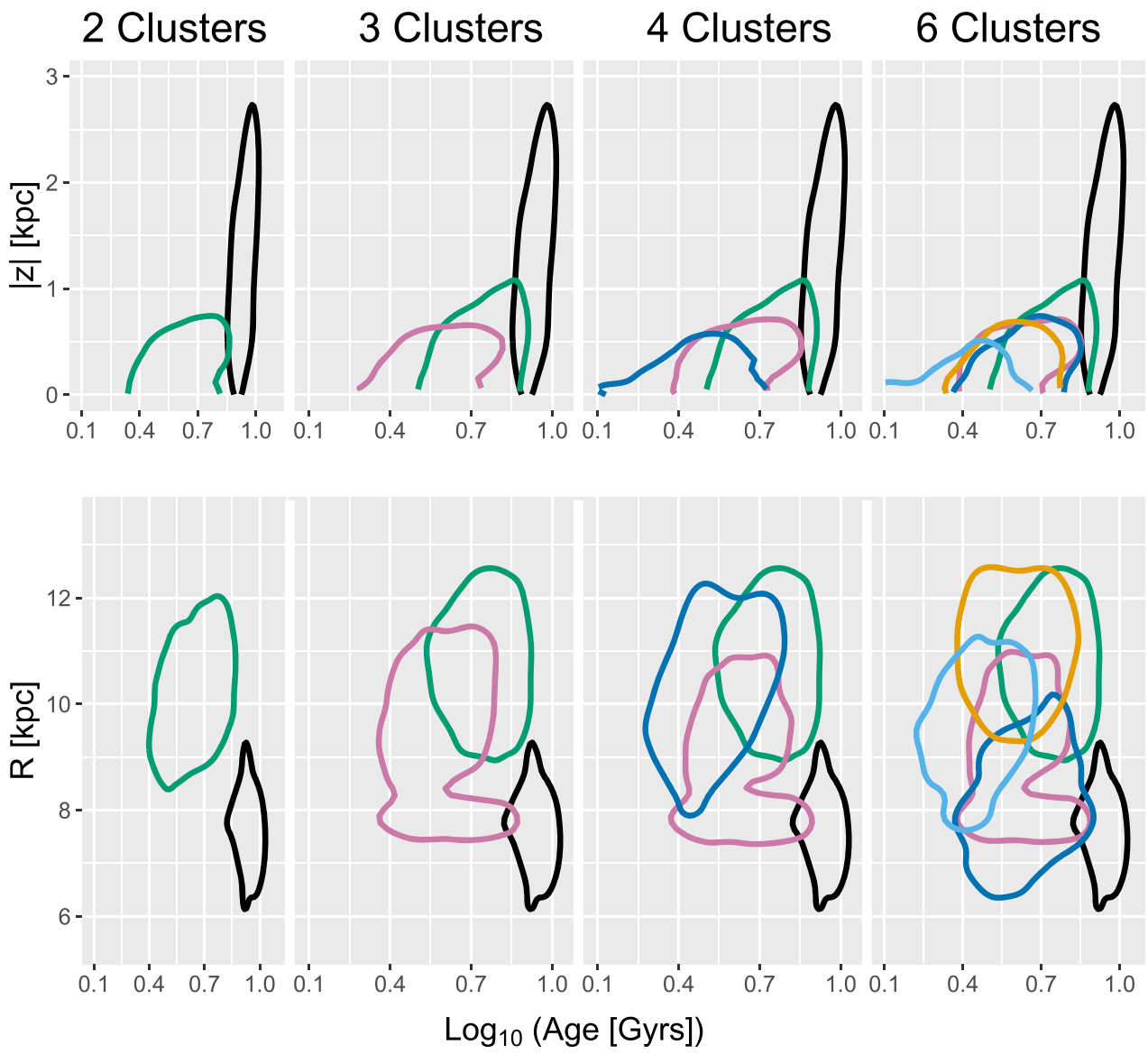

Figure 12. Clusters shown in Figure 11, projected into (R_{GAL}, $|z|$) spatial planes as a function of age. The clusters show different trends, with the centroids shifting in $|z|$–age and living in separate areas in R_{GAL}–age. The contour lines shown contain 50% of the stars in each cluster.

cluster 3) and 11 kpc (green cluster, cluster 2). The red group (cluster 3) then splits to create four clusters, with the new red group (cluster 3) taking on the majority of the $R_{GAL} \approx 8$ kpc stars compared to the blue group (cluster 4). Six groups become difficult to discern, with the majority of the clusters losing their bimodality and becoming somewhat unimodal.

The $|z|$ distributions (bottom row of Figure 11) allow us to observe the thickness of the distribution of the different clusters. The black cluster (cluster 1) takes on a broader range of higher $|z|$ values, while the green cluster (cluster 2) is more centralized at $|z| \approx 0$. As the green cluster (cluster 2) divides and more clusters are created, the $|z|$ densities stay peaked at $|z| \approx 0$, with each additional, younger, cluster becoming more localized near the center.

We further investigate the spatial–age distributions for two through six clusters. Figure 12 shows how our groups separate spatially (R_{GAL} and $|z|$) as a function of age. The distributions in $|z|$–age show a shift in the density centers, with the cluster centroids showing a positive correlation between $|z|$ and age. As the stellar sample is divided into more clusters, the youngest group of stars consistently gets more localized about the Galactic mid-plane. On the other hand, the primary centers of the clusters separate in no obvious pattern in the R_{GAL}–age plane; the oldest stars are localized to $R_{GAL} \approx 8$ kpc with the next oldest primarily centered at about $R_{GAL} \approx 11$ kpc, then the third oldest cluster is primarily centered at $R_{GAL} \approx 9$ kpc, the fourth oldest more strongly localized near $R_{GAL} \approx 8$ kpc, the fifth oldest spread around $R_{GAL} \approx 9$ kpc, and finally the youngest group of stars located mainly at $R_{GAL} \approx 9.5$ kpc.

6. Summary and Discussion

In the regime of high-dimensional data, we have the opportunity to go beyond a two-dimensional visual classification of stars into populations, and examine the information in

its entirety. In this paper, we use the APOGEE RC sample of stars, with stellar ages provided by Sanders & Das (2018) to investigate the distribution of the data in 19-dimensional abundance space as well as age and space. Our key findings can be summarized in three points.

1. There is coherent (correlated) structure in the abundance values themselves, as seen through their correlation matrix and PCA analysis (see Sections 4.2 and 4.3.1). The correlation structure between standardized abundances shows familial relationships in the light and α-elements, which motivates that the data can plausibly be projected into lower dimensions. The first three components of PCA, which describe just over 50% of the variance of the data, exhibit the same intra-family correlations.
2. In Section 4.1, we show that under the assumption that only two clusters describe the data, the projection of these into the [Mg/Fe]–[Fe/H] plane reveals that there is a subset of stars ($\sim$900 stars in this case) that visually appears to be a member of what is typically considered the high-α sequence, but in fact more closely resembles the low-α sequence using our algorithm. That is, clustering by a distance minimization algorithm in the 19-dimensional abundance space gives a different two-dimensional projection of abundance groups in the [Fe/H]–[α/Fe] plane than that given by a simple visual classification. Clustering in the first two components of both PCA and Isomap confirmed the robustness of this finding, as well as our overall clustering results to lower levels of the hierarchy (see Section 4.3.2).
3. In Section 5.1 we show our clusters, which are defined in abundance space alone, separate into populations that are also distinct in both age and spatial distribution. We can decompose our sample into at most six clusters, as this is the highest number of clusters that we can reasonably resolve into different populations in age and ($R_{\rm GAL}$, $|z|$) distributions given our age errors, which dominate our error budget (a median of $\approx$15%). However, the stars can also reasonably be described by a continuum of structure, with no definitive number of clusters comprising our sample (see Section 5).

Our first finding on correlations and PCA component structures in the chemical elements confirms the underlying simplicity of abundance space. In agreement with both Ting et al. (2018) and Price-Jones & Bovy (2018), we find that the dimensionality of the data set is much smaller than the full 19 dimensions from our sample. This simplicity can be interpreted as arising from the finite number of nucleosynthetic processes and distinct sites that yield enrichment with distinct abundance patterns (again, in agreement with Ting et al. 2018). We report that the first principal component corresponds to SNe II enrichment, the second principal component corresponds to enrichment from SNe Ia, and the third principal component captures contributions from dying low-mass stars.

The difference between the separation that assigns stars by-eye into the low- and high-α sequences and the distributions of our two clusters (found using information from 17 additional abundances) in the [α/Fe]–[Fe/H] plane calls into (further) question the extent to which disk decomposition in low dimensions can be physically interpreted. This difference gives insight into the discordant results when dividing into thin and thick disk populations using either space or [α/Fe]–[Fe/H]. Recently, in agreement with our findings, Ciucă et al. (2020) also argue that the subset of typically defined high-α stars is separate from the thick disk, and smoothly connects the high- and low-α sequences.

The clarity of the separation of spatial and age distributions of our abundance-defined clusters suggests that these higher dimensions add important constraints on Galactic history. For example, when ordered as a sequence in age, the distribution of our clusters in the [α/Fe]–[Fe/H] plane contradicts the simplest pictures of the monotonic chemical evolution in this plane. Rather, our intermediate-age and -metallicity clusters encompass both the low- and high-α sequences, likely reflective of a complex enrichment history by successive generations of both SNe II and SNe Ia. Indeed, in their hydrodynamical simulations of disk formation Clarke et al. (2019) find star-forming clumps that form on the low-α sequence, migrate to the high-α sequence, and eventually return to the low-α sequence.

Overall, our results provide confirmation of the remarkable simplicity underlying the abundance distributions of stars in the Galactic disk, as has already been reported from different perspectives in other studies. For example, Weinberg et al. (2019) show using APOGEE data that the abundance trends can be described by two populations across the Galaxy, and Ness et al. (2019) demonstrate that the age and [Fe/H] of stars on the low-α sequence can be used to predict [X/Fe] for other elements in the data to within 0.0 two-dimensionalex, on average. Our work is consistent with this picture, as apparent in Figure 13.

The age catalog we use has been determined using both spatial and metallicity priors (Sanders & Das 2018). In the Appendix (Section A.1), we compare the spatial–age and [Fe/H]–age distributions using two different age catalogs, where ages have been determined using the spectra directly to propagate asteroseismic ages (Pinsonneault et al. 2018) using data-driven modeling (Ness et al. 2019; Sit & Ness 2020). We obtain consistent results from all age catalogs; however, the clusters show slightly different distributions in these planes.

7. Conclusion and Future Prospects

We conclude that clustering in high-dimensional abundance space is informative about the sequence and locations of forming stellar populations. This is a particularly promising variant of chemical tagging (Freeman & Bland-Hawthorn 2002), even if we are unable to go as far as reconstructing the individual *physical* clusters in which stars are originally born (e.g., Ting et al. 2017; Ness 2018; Kamdar et al. 2019). We by no means claim that there is a discrete number of stellar populations comprising the disk of the Milky Way (i.e., six as shown in Figure 11). Rather, our clusters characterize the underlying continuous evolution in age and metallicity across the disk, analogous to analyses of mono-[Fe/H]–age populations (e.g., Bovy 2016). Our findings of a continuous evolution are in agreement with Bovy et al. (2012) in that there is no distinct bimodality of the thin and thick disk. However, we show this result by a completely different approach, making use of more abundance information.

This is a pilot study that outlines the potential of current and future data to capture the chemical enrichment sequence during the formation of the disk. With just under 30,000 stars used in this work, we see clustering organizes stars into age–abundance groups.

Figure 13. [Fe/H]–age contours containing 50% of the stars for each cluster, shown in Figure 12, showing the separation of the clusters in this plane.

However, this does not find individual birth clusters themselves. With much larger data sets, however, many stars from the same cluster may presumably fall within these groups, extending the effectiveness of high-dimensional clustering to aspirations of chemical tagging of individual birth clusters. Additionally, while we utilize the RC sample of stars, which have abundance measurements with reasonably small amplitudes, including the errors in the analysis will improve the assigned clustering. This is in both the weighting of the most informative elements and for recovering the true underlying populations that are most chemically similar, given the errors in the data. With the increase in both sample size and precision expected in future catalogs, stronger conclusions will be able to be drawn as to how the abundance clustering is tracing the birth properties of stars.

We seek to understand the formation of the Milky Way's disk. To do so, we must know the materials stars were born with ([Fe/H], [X/Fe]), when they were born ($t_{\rm birth}$), and where they were born ($R_{\rm birth}$). In observational data, we have access to precise ([Fe/H], [X/Fe]) and imprecise ages for a significant number of stars: data we work with in this paper. Only in a simulation however, do we have access to the full set of properties to trace formation and evolution of disks ([Fe/H], [X/Fe], $t_{\rm birth}$, and $R_{\rm birth}$). This enables us to use simulations to investigate and understand the dependencies and relationships between these properties in disk galaxies. As such, our work in preparation forms Paper II in this series. B. L. Ratcliffe et al. (2020, in preparation) uses hydrodynamical simulations of Milky Way analogs to interpret the origin of different abundance clusters in the disk. Due to the anticipated chemical homogeneity of stellar birth clusters (for clusters up to $10^5 M_{\odot}$; Bland-Hawthorn et al. 2010), we expect that in grouping chemically similar stars—as done in this work and Paper II—we are assigning stars within birth clusters to similar groups.

Several aspects of our approach could be improved. In order to determine the maximum number of clusters with significantly different age populations, we choose to compare the mean ages. The pairwise multiple comparison test design we propose is provisional—there are other ways to determine discrete aged clusters. Additionally, comparing means is not representative of distributions when the densities are non-normal or spread out. Ideally distributions would be compared, without the worry of small fluctuations misleading results.

In addition, we do not take into account the abundance measurement errors on the APOGEE data. Therefore all measurements are weighted equally in our clustering algorithm, whereas in reality some stars are measured more precisely than others and some elements are systematically measured more precisely than others. Future data sets (or more sophisticated analyses of existing data sets) are likely to give higher-precision measurements of abundances, ages, and locations, which will allow cleaner measurements of distributions in these spaces.

Overall, our results are indicative of the immense promise of future data sets, which could contain even more informative dimensions. One limitation of the APOGEE DR14 abundance catalog is that it is restricted to α, iron-peak, and light abundances, but is missing the channels from neutron capture processes, which are sensitive to different production mechanisms. The next data release of APOGEE, as well as future mission data (Kollmeier et al. 2017), will also contain valuable neutron capture element measurements (Hasselquist et al. 2016; Cunha et al. 2017) for millions of disk stars. This will enable a more complete abundance clustering analysis to inform our picture of the assembly of the disk. These additional elements could reveal additional details or structures that are not apparent in the 19 (highly correlated) abundances that we use, and explain (for example) the events which lead to the formation of the low- and high-α sequences.

We acknowledge helpful conversations with the Milky Way Stars group at Columbia University. K.V.J. is supported by NSF grant AST-1715582 and B.S. is supported by NSF grant DMS-1712822. M.K.N. is in part supported by a Sloan Foundation fellowship.

Appendix

A.1. Comparison Using Other Age Catalogues

We find that our clusters show separation in planes of [Fe/H]–age and ($R_{\rm GAL}$, $|z|$)–age. However, we use the age catalog from Sanders & Das (2018), which is generated using priors from a

model of the spatial distribution of populations in the Galaxy and metallicity priors (see also Das & Sanders 2018). We want to validate that we are not directly recovering these priors. Therefore, we examine two other age catalogs that have been generated differently to that of Sanders & Das (2018). We use two comparison catalogs: one generated using APOGEE DR12 spectra in Ness et al. (2016) and one generated using APOGEE DR14 spectra in Sit & Ness (2020). Both approaches employ the data-driven methodology of The Cannon directly on the stellar spectra (Ness et al. 2015). However, there are important differences. The earlier work employs a smaller training set using the then available APOKASC catalog (Pinsonneault et al. 2014), as well as DR12 spectra and propagates this age information to the remaining DR12 across the parameter range where the survey data are representative of the training data. The later work uses the far larger training set of stars in Pinsonneault et al. (2018) and the DR14 spectra, propagating age to a larger set of stars across a broader range in stellar parameters as well as simultaneously inferring individual abundances. The other major difference between the Ness et al. (2016) and Sit & Ness (2020) catalogs is that the latter uses a separate model to infer ages for low- and high-α stars, as motivated by Ness et al. (2019).

There are 17,152 stars in common with our analysis using the Sanders & Das (2018) ages from Ness et al. (2016), and 25,924 stars in common with Sit & Ness (2020). Given the different number of stars, we repeat our analysis in its entirety for these other age catalogs. Using the metric described in Section 5.1, we find we can go to fewer clusters (in both cases five clusters). This is a consequence of the smaller number of stars being analyzed and of the larger age errors (about 40% for both catalogs). We show the [Fe/H]–age and ($R_{\rm GAL}$, $|z|$)–age planes for these age catalogs in Figures A1 and A2 to compare with our results shown in the paper (Figures 12 and 13). The primary difference between these results and those in our main analysis is the less dramatic separation in the spatial planes with age. The discernible trend of younger stars living closer to the Galactic mid-plane found using the Sanders & Das (2018) ages is less apparent for both age catalogs, along with the separation in $R_{\rm GAL}$–age using the Ness et al. (2016) ages. However, we still see distinct cluster centers in $R_{\rm GAL}$–age using the Sit & Ness (2020) ages for up to four clusters. Additionally, both age catalogs show cluster separation in [Fe/H]–age, with trends reminiscent of the ones found in Figure 13. Overall, results using ages from the Ness et al. (2016) and Sit & Ness (2020) catalogs complement our main results found.

A.2. Isomap Physical Intuition

While Isomap uses the Euclidean distance to create the neighborhood graph G, it relies on the geodesic distances between points to create the new lower-dimensional embedding. Take, for example, the two-dimensional spiral roll in Figure A3, where the points marked with a red square are close in terms of their Euclidean distance. If we take into account the structure of the data, however, they are significantly separated from one another (along the geodesic path). The neighborhood graph constructed in Isomap yields the shortest path between points along the data structure itself, an approximation to the true geodesic path. Therefore, Isomap will project these two points further away than if we were to use a Euclidean distance metric, such as PCA.

A.3. Algorithms and Additional Figures

Here, we present the explicit steps for the different methodologies used in this work, as well as extra figures. The steps for Ward's minimum variance agglomerative hierarchical clustering, Isomap, Dijkstra's algorithm, and multidimensional scaling are given in Algorithms 1, 2, 3, and 4, respectively. Figures A1 and A2 show the ($R_{\rm GAL}$, $|z|$)–age and [Fe/H]–age planes for the clusters found in the main body of this work, with ages from Ness et al. (2016) and Sit & Ness (2020). Figure A3 aids in the physical interpretation of Isomap. Figure A4 is an extension of Figure 8.

Algorithm 1. Ward's Minimum Variance Agglomerative Hierarchical Clustering

Step 1: Start with each star as its own cluster.

Step 2: Create a dissimilarity matrix between cluster i and cluster j for all i and j. Let C_* be cluster $*$ and c_* be its cluster center. Then the dissimilarity matrix is:

$$\delta^2(C_i, C_j) = ||c_i - c_j||^2 = \sum_{k=1}^{d=19} (c_i(k) - c_j(k))^2$$

where

$$c_* = \frac{1}{|C_*|} \sum_{x \in C_*} x \in \mathbb{R}^{19}$$

Step 3: Combine the least dissimilar clusters C_i^* and C_j^* and update the dissimilarity matrix:

$$\delta^2(C_i^* \cup C_j^*, C_k) = \frac{|C_i^*| + |C_k|}{|C_i^*| + |C_j^*| + |C_k|}\delta^2(C_i^*, C_k) + \frac{|C_j^*| + |C_k|}{|C_i^*| + |C_j^*| + |C_k|}\delta^2(C_j^*, C_k) - \frac{|C_k|}{|C_i^*| + |C_j^*| + |C_k|}\delta^2(C_i^*, C_j^*)$$

Step 4: Repeat Step 3 until all clusters are combined and one large cluster containing all the stars remains.

Algorithm 2. Isomap

Step 1: Find the k nearest neighbors for each data point. Construct a neighborhood graph G, with the edge length $G_{i,j}$ equal to the Euclidean distance between the two neighbors if star j is a neighbor of star i.

Step 2: Create a geodesic distance matrix, d_G, that approximates the shortest path between all pairs of points using Dijkstra's algorithm (Algorithm 3).

Step 3: Apply classical multidimensional scaling to d_G (Algorithm 4).

Algorithm 3. Dijkstra's Alogrithm

This is pseudo-code to find $d_G(i, j)$, the geodesic distance from star i to star j.

Step 1: Start with all stars marked as unvisited nodes with tentative distance ∞ and star i as the current node.

Step 2: For each of the current node's unvisited neighbors, compare the neighbors current assigned distance to the tentative distance through the current node. Replace the assigned distance if tentative distance is smaller than assigned distance.

Step 3: Once all unvisited neighbors of the current node are considered, mark the current node as visited.

Step 4: Continue by selecting the unvisited node with the smallest tentative distance and repeat steps 2–4 until star j has been visited.

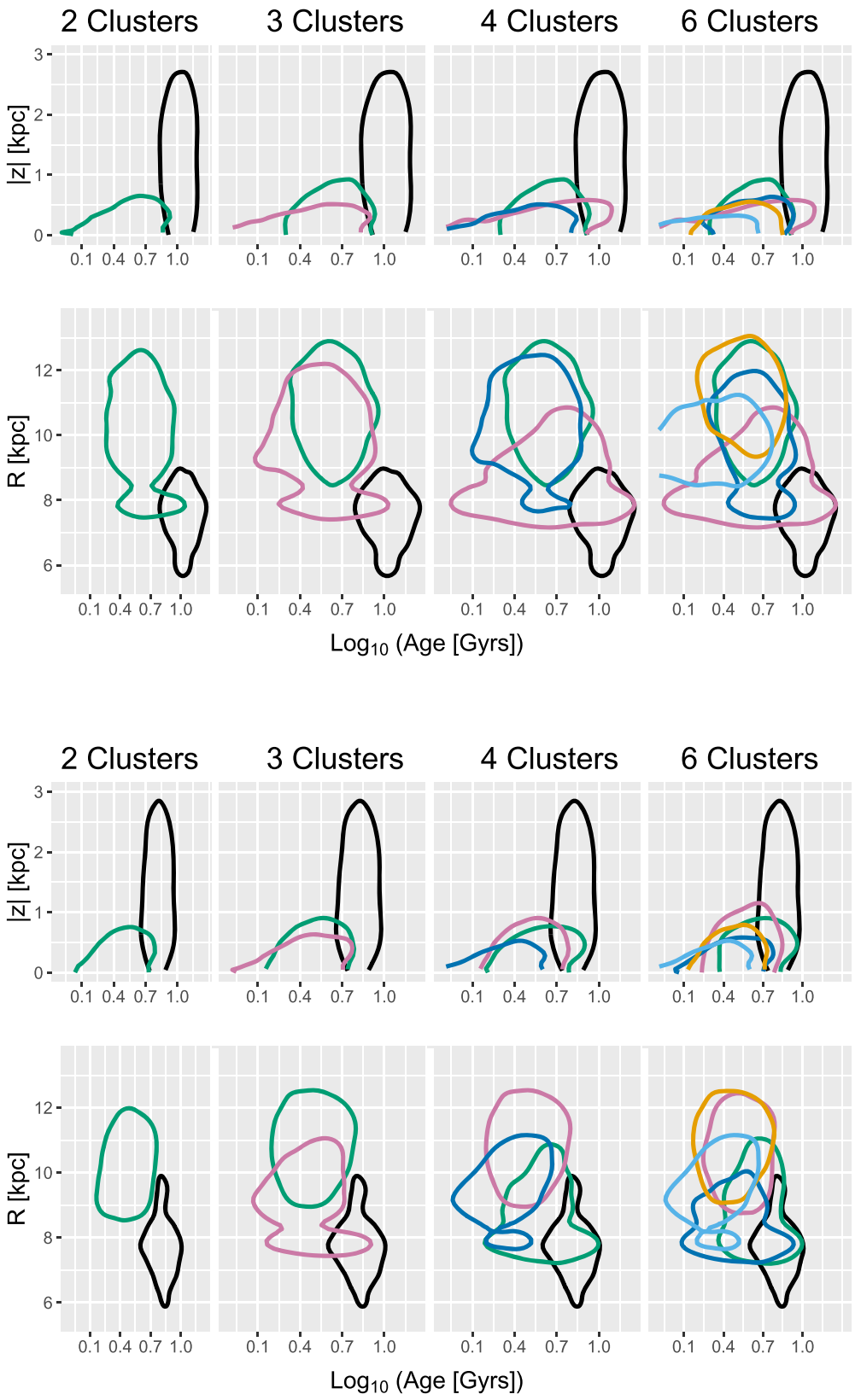

Figure A1. (R_{GAL}, $|z|$)–age planes of the clusters in Figure 12, only using the age catalog of Ness et al. (2016), at top, and of Sit & Ness (2020) at bottom. The clusters show more overlap in the (R_{GAL}, $|z|$)–[Fe/H] planes, but have different means and dispersions.

Figure A2. [Fe/H]–age planes of the clusters in Figure 13, only using the age catalog of Ness et al. (2016), at top, and of Sit & Ness (2020) at bottom. The age errors from these catalogs, at around 40%, are about double that of Sanders & Das (2018). However, they are derived using the stellar spectra themselves and have no prior probability assigned from spatial or abundance information. The different clusters show some overlap in the age–[Fe/H] plane, but have different means and dispersions.

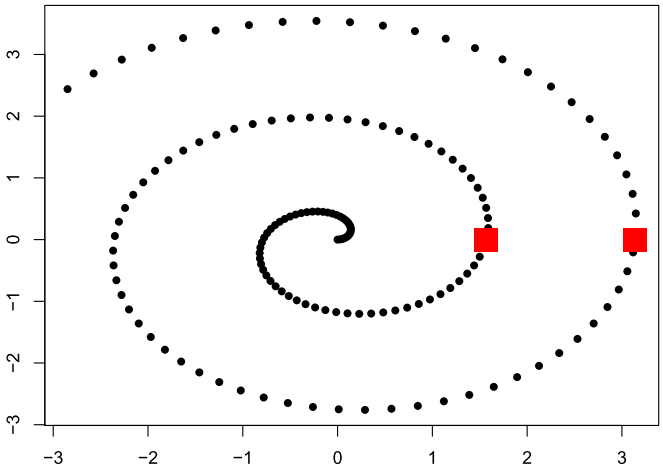

Figure A3. Simple toy model showing that the Euclidean distance metric may not accurately reflect the intrinsic similarity for data that lie on an embedded lower-dimensional nonlinear manifold. Here, the two points marked as red squares are close in this two-dimensional plane. However, taking into account the structure of the spiral, the two points are distant along the geodesic path.

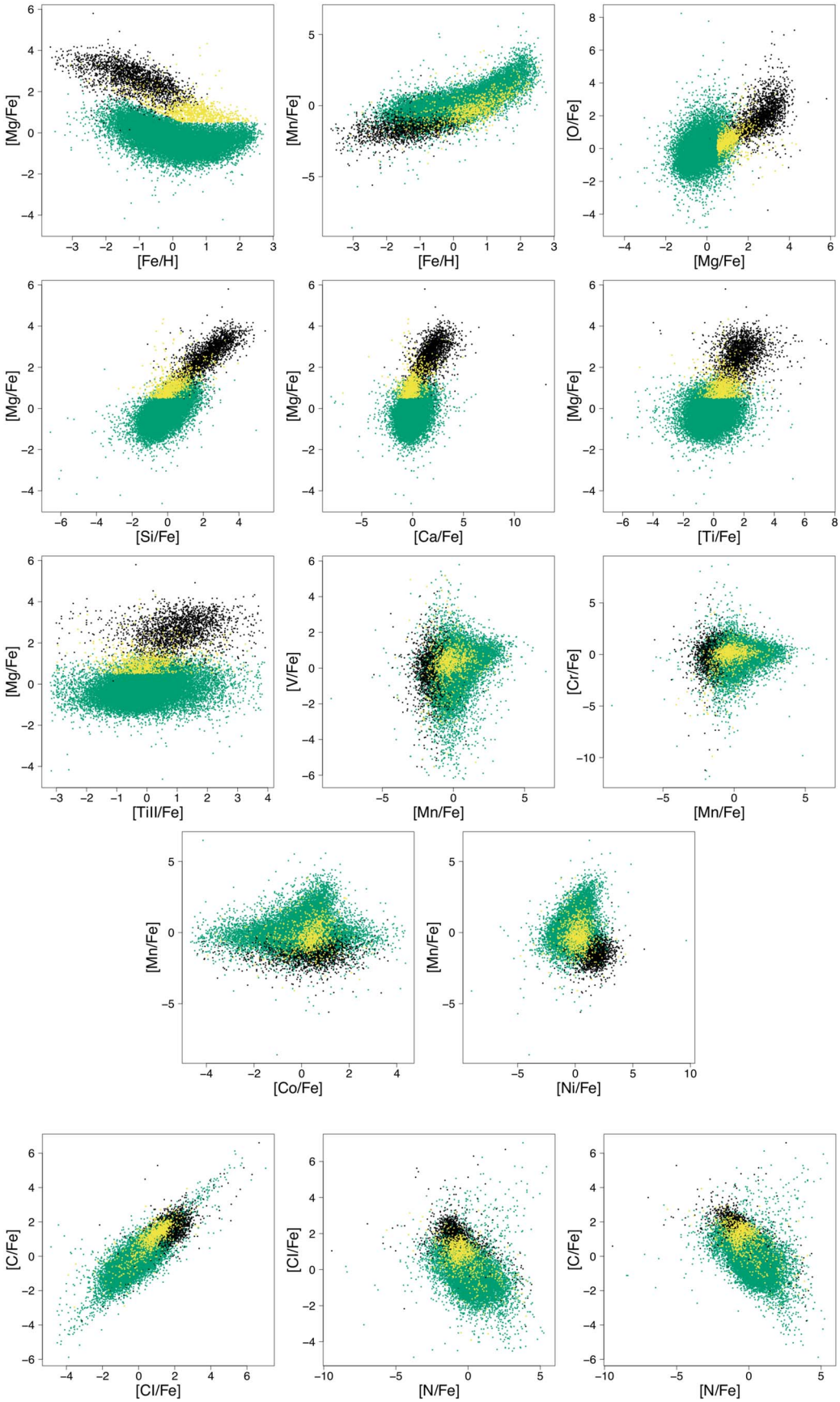

Figure A4. Additional abundance–abundance planes.

Algorithm 4. Multidimensional Scaling

MDS takes a distance matrix d_G and returns coordinates in a p -dimensional space, where p is the pre-specified desired number of dimensions which we choose to be 2.

Step 1: Create a squared distance matrix S where $S_{i,j} = d_G(i, j)^2$.

Step 2: Apply double centering through $B = -\frac{1}{2}JSJ$ where $J = 1 - \frac{1}{n}1\prime$ is the centering matrix.

Step 3: Calculate the eigenvalues and eigenvectors of B. Let $\lambda_1, \ldots, \lambda_p$ be the p largest eigenvalues and $e_1, \ldots, e_p$ be the associated eigenvectors.

Step 4: Define E_p as the $n \times p$ matrix with eigenvectors $e_1, \ldots, e_p$ as the columns and Λ_p as the diagonal matrix of the eigenvalues from step 3. The new lower-dimensional coordinates are $E_p\Lambda_p^{1/2}$.

ORCID iDs

Melissa K. Ness https://orcid.org/0000-0001-5082-6693

References

Abolfathi, B., Aguado, D., Aguilar, G., et al. 2018, ApJS, 235, 42
Bensby, T., Feltzing, S., & Oey, M. 2014, A&A, 562, A71
Blancato, K., Ness, M., Johnston, K. V., Rybizki, J., & Bedell, M. 2019, ApJ, 883, 34
Bland-Hawthorn, J., Krumholz, M. R., & Freeman, K. 2010, ApJ, 713, 166
Bland-Hawthorn, J., Sharma, S., Tepper-Garcia, T., et al. 2019, MNRAS, 486, 1167
Blanton, M. R., Bershady, M. A., Abolfathi, B., et al. 2017, AJ, 154, 28
Bovy, J. 2016, ApJ, 817, 49
Bovy, J., Nidever, D. L., Rix, H.-W., et al. 2014, ApJ, 790, 127
Bovy, J., Rix, H.-W., & Hogg, D. W. 2012, ApJ, 751, 131
Bovy, J., Rix, H.-W., Schlafly, E. F., et al. 2016, ApJ, 823, 30
Casey, A. R., Lattanzio, J. C., Aleti, A., et al. 2019, ApJ, 887, 73
Ciucă, I., Kawata, D., Miglio, A., Davies, G. R., & Grand, R. J. 2020, arXiv:2003.03316
Clarke, A. J., Debattista, V. P., Nidever, D. L., et al. 2019, MNRAS, 484, 3476
Cunha, K., Smith, V. V., Hasselquist, S., et al. 2017, ApJ, 844, 145
Das, P., & Sanders, J. L. 2018, MNRAS, 484, 294
Dijkstra, E. W. 1959, NuMat, 1, 269
Dunnett, C. W. 1980, J. Am. Stat. Assoc., 75, 796
Freeman, K., & Bland-Hawthorn, J. 2002, ARA&A, 40, 487
Fuhrmann, K. 1998, A&A, 338, 161
Gandhi, S. S., & Ness, M. K. 2019, ApJ, 880, 134
García Pérez, A. E., Allende Prieto, C., Holtzman, J. A., et al. 2016, AJ, 151, 144
Garcia-Dias, R., Allende Prieto, C., Sánchez Almeida, J., & Alonso Palicio, P. 2019, A&A, 629, A34
Gilmore, G., Wyse, R. F., & Norris, J. E. 2002, ApJL, 574, L39
Hartigan, J. A., & Wong, M. A. 1979, J. Royal Stat. Soc. Ser. C, 28, 100
Hasselquist, S., Shetrone, M., Cunha, K., et al. 2016, ApJ, 833, 81
Hayden, M. R., Bovy, J., Holtzman, J. A., et al. 2015, ApJ, 808, 132
Hogg, D. W., Casey, A. R., Ness, M., et al. 2016, ApJ, 833, 262
Holtzman, J. A., Shetrone, M., Johnson, J. A., et al. 2015, AJ, 150, 148
Hotelling, H. 1933, J. Educational Psych., 24, 417
Jain, A. K., & Dubes, R. C. 1988, Algorithms for Clustering Data (Englewood Cliffs, NJ: Prentice Hall)
Jolliffe, I. T. 1990, Wthr, 45, 375
Kamdar, H., Conroy, C., Ting, Y.-S., et al. 2019, ApJ, 884, 173
Kaufman, L., & Rousseeuw, P. J. 2009, Finding Groups in Data: An Introduction to Cluster Analysis (New York: Wiley)
Kodinariya, T. M., & Makwana, P. R. 2013, Applied Math. Info. Sci., 10, 1493
Kollmeier, J. A., Zasowski, G., Rix, H.-W., et al. 2017, arXiv:1711.03234
Kruskal, J. B. 1964, Psychometrika, 29, 1
Lian, J., Thomas, D., Maraston, C., et al. 2020, MNRAS, 497, 2371
Liu, F., Asplund, M., Yong, D., et al. 2019, A&A, 627, 117
Mackereth, J. T., Bovy, J., Leung, H. W., et al. 2019, MNRAS, 489, 176
Mackereth, J. T., Crain, R. A., Schiavon, R. P., et al. 2018, MNRAS, 477, 5072
Majewski, S. R., Schiavon, R. P., Frinchaboy, P. M., et al. 2017, AJ, 154, 94
Masseron, T., & Gilmore, G. 2015, MNRAS, 453, 1855
Massey, F. J., Jr 1951, J. Am. Stat. Assoc., 46, 68
Murtagh, F., & Legendre, P. 2011, arXiv:1111.6285
Ness, M. 2018, PASA, 35, e003
Ness, M., Hogg, D. W., Rix, H.-W., et al. 2016, ApJ, 823, 114
Ness, M., Hogg, D. W., Rix, H.-W., Ho, A. Y. Q., & Zasowski, G. 2015, ApJ, 808, 16
Ness, M. K., Johnston, K. V., Blancato, K., et al. 2019, ApJ, 883, 177
Nidever, D. L., Bovy, J., Bird, J. C., et al. 2014, ApJ, 796, 38
Pearson, K. 1901, London, Edinburgh, Dublin Phil. Mag. J. Sci., 2, 559
Pérez, A. E. G., Prieto, C. A., Holtzman, J. A., et al. 2016, AJ, 151, 144
Pinsonneault, M. H., Elsworth, Y., Epstein, C., et al. 2014, ApJS, 215, 19
Pinsonneault, M. H., Elsworth, Y. P., Tayar, J., et al. 2018, ApJS, 239, 32
Price-Jones, N., & Bovy, J. 2018, MNRAS, 475, 1410
Price-Jones, N., Price-Jones, B. J., Webb, J. J., et al. 2020, MNRAS, 496, 5101
Ringnér, M. 2008, NatBi, 26, 303
Rosman, G., Bronstein, M. M., Bronstein, A. M., & Kimmel, R. 2010, Int. J. Comput. Vision, 89, 56
Sanders, J. L., & Das, P. 2018, MNRAS, 481, 4093
Silva Aguirre, V., Bojsen-Hansen, M., Slumstrup, D., et al. 2018, MNRAS, 475, 5487
Silva, V. D., & Tenenbaum, J. B. 2003, Adv. Neural Info. Process. Systems, 15, 721
Sit, T., & Ness, M. K. 2020, ApJ, 900, 4
Tenenbaum, J. B., De Silva, V., & Langford, J. C. 2000, Sci, 290, 2319
Tibshirani, R., Walther, G., & Hastie, T. 2001, J. Royal Stat. Soc. Ser. B, 63, 411
Ting, Y.-S., Freeman, K. C., Kobayashi, C., De Silva, G. M., & Bland-Hawthorn, J. 2012, MNRAS, 421, 1231
Ting, Y.-S., Hawkins, K., & Rix, H.-W. 2018, ApJL, 858, L7
Ting, Y.-S., Rix, H.-W., Conroy, C., Ho, A. Y. Q., & Lin, J. 2017, ApJ, 849, 9
Ward, J. H., Jr 1963, J. Am. Stat. Assoc., 58, 236
Weinberg, D. H., Holtzman, J. A., Hasselquist, S., et al. 2019, ApJ, 874, 102
Wold, S., Esbensen, K., & Geladi, P. 1987, Chemometrics Intelligent Lab. Systems, 2, 37
Zasowski, G., Ménard, B., Bizyaev, D., et al. 2014, ApJ, 798, 35