Likelihood-free Inference of Chemical Homogeneity in Open Clusters

Aarya Patil*† Jo Bovy*† Gwendolyn Eadie*‡

Abstract

Star clusters are excellent astrophysical laboratories to study the history of star formation and chemical enrichment in our Galaxy. These are groupings of stars born out of the same gas cloud, and are theoretically expected to have similar chemical compositions. Empirically validating this chemical homogeneity is important yet difficult because the measurement of accurate and precise chemistry of stars using stellar spectroscopic data is statistically challenging. We perform high-fidelity Likelihood-free Inference of chemistry of stars using state-of-the-art Neural Density Estimation to observationally determine the level of chemical homogeneity in open clusters. We make our model computationally efficient by incorporating active learning and dimensionality reduction of stellar spectroscopic data through Functional Principal Component Analysis. Our constraints on chemical homogeneity will not only help understand the detailed evolution of star-forming clouds but also allow us to trace the chemical and dynamical history of our Galaxy through *chemical tagging*.

Key Words: methods: data analysis — techniques: spectroscopic — Galaxy: abundances — Galaxy: disk — Galaxy: formation

1. Introduction

A star cluster is a set of stars that are gravitationally bound to each other and are believed to form in the same gas cloud at the same time (Shu et al., 1987; Lada and Lada, 2003). Simulations show that there is turbulent mixing in these clouds which makes the star-forming gas chemically homogeneous (Feng and Krumholz, 2014). Due to this, we expect stars in a cluster to have similar chemistry, and there is now a growing need to observationally test this theory in order to understand the exact nature of star-forming clouds. The detailed evolution and mixing of the gas in these clouds, especially during the initial stages of star-formation, is not yet properly understood because the young star clusters are difficult to observe (McKee and Tan, 2002; Feng and Krumholz, 2014). Over a timescale of a few million years, high mass stars die and produce heavy elements that enrich the gas, and this complicates our simple picture of star-formation. Thus, we need to measure the birth chemistry of long-lived stars in a cluster and constrain their chemical homogeneity to understand the evolution of star-forming clouds.

The chemistry of long-lived stars measured through high-resolution spectroscopic observations can be traced back to the birth chemistry if we understand the processes of stellar evolution that affect a star's chemistry over its lifetime. Despite several efforts to develop a comprehensive theory of stellar evolution and orders-of-magnitude improvements in the observed spectral data, stellar spectroscopy presents many challenges due to high-dimensional parameter spaces and the instrumental factors that affect astronomical observations. Physically-motivated models for the complex astrophysical processes in stars are highly non-linear and numerically driven, and these can present challenges in traditional statistical frameworks (Feigelson and Babu, 2012). Additionally, there is the problem of

*David A. Dunlap Department of Astronomy and Astrophysics, University of Toronto, 50 St. George Street, Toronto, ON, M5S 3H4, Canada

†Dunlap Institute for Astronomy and Astrophysics, University of Toronto, 50 St. George Street, Toronto, ON M5S 3H4, Canada

‡Department of Statistical Sciences, University of Toronto, 100 St George St, Toronto, ON M5S 3G3, Canada

dealing with noisy, heterogeneous data streams that are incomplete. What this means is that estimating the birth chemistry of stars in a cluster and providing strong limits on chemical homogeneity is difficult.

Several studies on the chemical homogeneity of open clusters, which are clusters of stars found in the disk, suggest homogeneity (e.g. Reddy et al., 2012; Bovy, 2016). However, some authors have found contradictory results. Studies have found inhomogeneities in open cluster M67 member stars and attributed these inhomogeneities to stellar evolution (e.g. Liu et al., 2019).

Bovy (2016) use forward modeling and Approximate Bayesian Inference (ABC), a traditional likelihood-free inference method, to infer the chemical homogeneity in open clusters given stellar spectroscopic data. Despite being successfully used in several applications in astronomy and many other branches of science, ABC methods are inefficient when simulations are expensive since they require a large number of simulations to be run. Density-Estimation Likelihood-Free Inference (DELFI; Bonassi et al., 2011; Fan et al., 2013; Papamakarios and Murray, 2016; Lueckmann et al., 2018; Alsing et al., 2018; Papamakarios et al., 2018; Lueckmann et al., 2018; Alsing et al., 2019) is a new Bayesian inference method in which density estimation is employed to perform statistical inference, and is an improvement over ABC approaches by orders-of-magnitude. We develop a technique based on this novel Likelihood-free Inference method and infer accurate and precise chemistry of stars in open cluster M67 given spectroscopic data.

Chemical homogeneity in open clusters strongly motivates us to tag individual star formation events in the Milky Way by determining the chemistry of large samples of stars and finding chemically similar groups. Open clusters get dispersed over a lifetime of $\sim$100 million years through random interactions in the Galaxy (Lada and Lada, 2003). Thus, stars that were born together might currently be at very different locations in the Galaxy but could be traced back to their birth cluster using just their chemical signatures. This *chemical tagging* has the ability to map the most detailed chemical and dynamical evolution of our Galaxy, and our studies to constrain chemical homogeneity in open clusters will help validate this promising approach.

2. Data

We use the Apache Point Observatory Galactic Evolution Experiment (APOGEE - Majewski et al., 2017) spectroscopic data for our work. APOGEE is a high-resolution (R $\sim$22,500) spectroscopic survey in the infrared (H-band 1.51 to 1.70 μm) that observes multiple targets simultaneously using a 300-fiber spectrograph (Wilson et al., 2010). We use the APOGEE data released as part of the public Data Release 14 of Sloan Digital Sky Survey-IV (SDSS-IV) which includes spectroscopic data for over 250,000 stars. We test the chemical homogeneity of open cluster M67 using open cluster data in the APOGEE/Open Cluster Chemical Abundances and Mapping (OCCAM) Data Release 14 sample (Donor et al., 2018). We use open cluster M67 because it is a well-studied open cluster, with age and chemistry similar to those of the Sun. The data is high-dimensional with a wavelength grid of $\sim$8000 and has several masked or missing data values. Also, the data is very noisy and the APOGEE noise characterisation is known to have issues. Our work is tailored to deal with these types of issues.

3. Methods & Analysis

Simulators are common in science and engineering to model mechanistic processes, and can be used to generate data using parameters of physical theories. However, the challenge

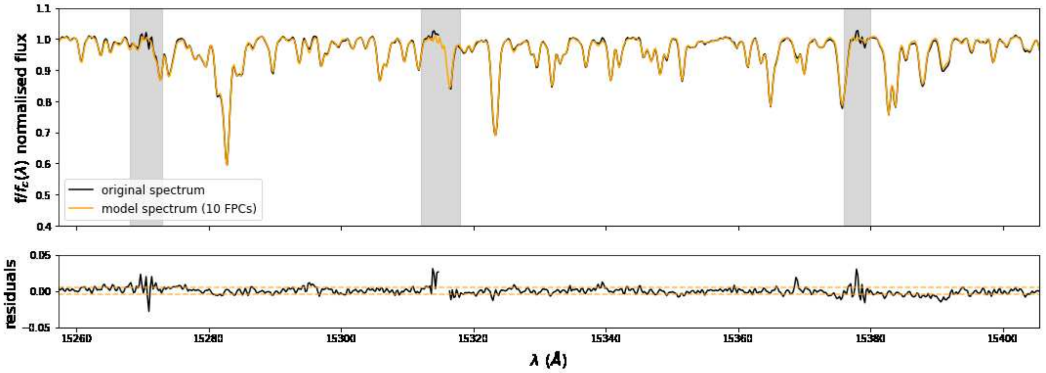

Figure 1: Modelling a spectrum using the first 10 functional principal components (FPCs). The top panel shows the original spectrum and an overplotted reconstructed model. The bottom panel shows the residuals between the original spectrum and model with horizontal lines marking the APOGEE base uncertainty (0.005). The grey regions highlight what FPCA does best: remove instrumental noise in the continuum and generate missing data.

in inference of parameters of a simulator given observed data is that $p(x \mid \theta)$, the likelihood, is often intractable. This is especially true in astronomy because problems involve high-dimensional parameter spaces and data. Here, we infer the spread in chemistry of star clusters using spectroscopic models without making any assumptions on the likelihood; we perform Likelihood-Free Inference.

We build an ensemble of Neural Density Estimators (NDEs) using state-of-the-art Masked Autoregressive Flow (Papamakarios et al., 2017) and Mixture Density Networks (Bishop, 1994) to make the model architecture flexible. We incorporate dimensionality reduction for building computationally efficient models and active learning to explore relevant regions of parameter space on the fly into DELFI (Papamakarios et al., 2018; Lueckmann et al., 2018). Dimensionality reduction is especially important for this problem since the spectral data is high-dimensional. Thus, our model ensures fast, high-fidelity likelihood-free inference.

Price-Jones and Bovy (2018) successfully reduce the dimensionality of stellar chemical space using Expectation-Maximization Principal Component Analysis (EMPCA) to 10 Principal Components (PCs) (Roweis and Saul, 2000; Bailey, 2016). Their PCs show large scale structures which cannot be explained by theory because EMPCA cannot distinguish between data and noise variability if the noise characterisation is imperfect. APOGEE spectra do not have a proper noise characterisation and it is hard to disentangle all the different kinds of noise that can affect spectral information. To tackle this, we use Functional Principal Component Analysis (FPCA), the functional version of PCA, to reduce the dimensionality of spectral data to 10 PCs that have narrow theoretical features (Ingrassia and Costanzo, 2005). Spectroscopic measurements are discrete samples of continuous functions of wavelength $f(\lambda)$. Each spectrum $f(\lambda)$ is a single object or "point" in a large "spectroscopic functional space". This motivates the use of FPCA that transforms the data into a functional form using basis functions as follows:

$$f_n(\lambda) \approx \sum_{k=1}^{K} \alpha_{n,k}\phi_k(\lambda) \tag{1}$$

where $f_n(\lambda)$ is the n^{th} observed spectrum and it is regressed onto $\phi_k(\lambda)$, the K basis functions.

This analysis is further improved by using simulated spectra as basis functions instead

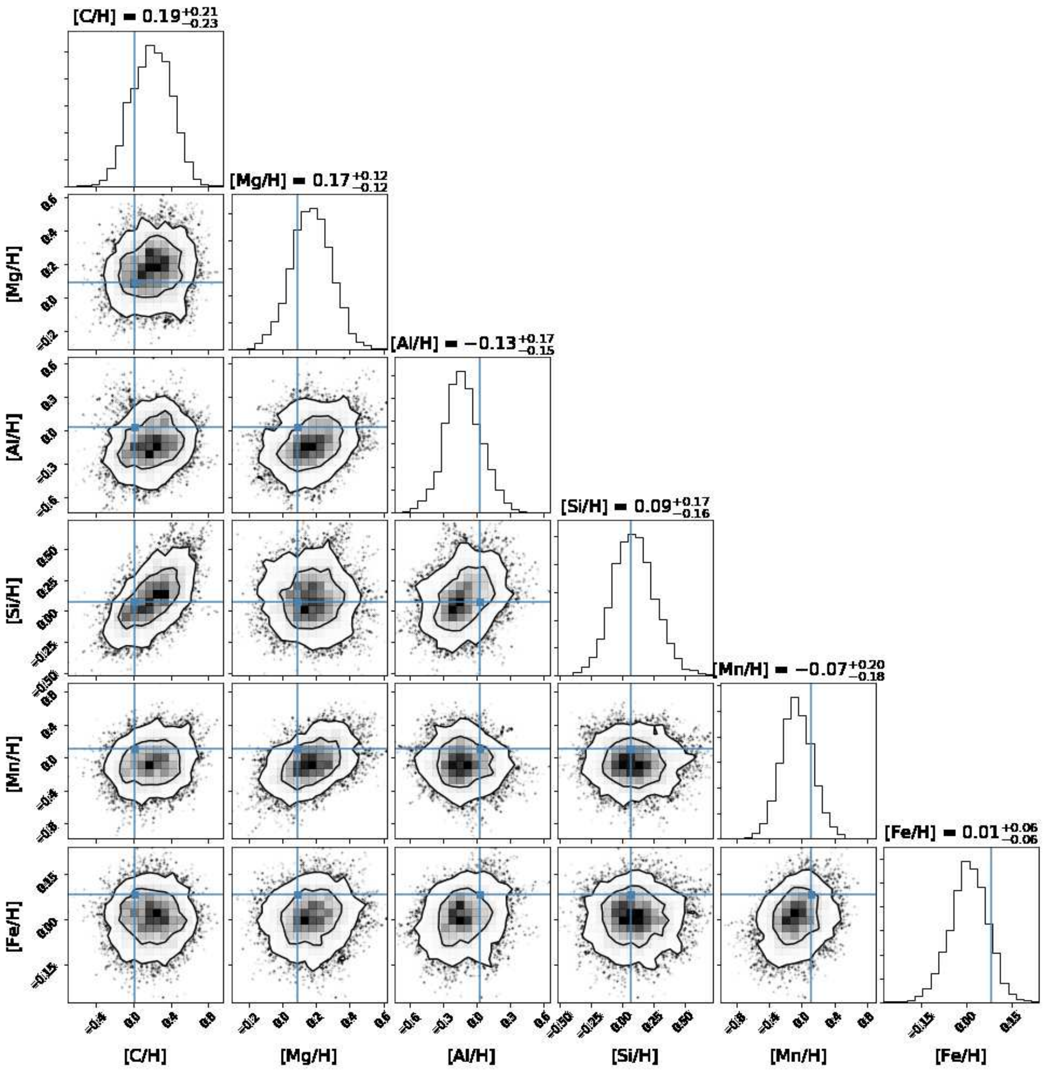

Figure 2: Posterior probability distribution of a M67 red giant APOGEE spectrum using DELFI. The contours represent 68 and 95% credible intervals. Blue lines represent the chemical abundances estimated in the APOGEE pipeline that uses a χ^2 minimization technique and internal calibration relations that assume homogeneity in clusters.

of traditional orthogonal basis functions like Legendre polynomials. By using only 50 simulations with varying chemical parameters, the ~8000 dimensional spectroscopic space can be reduced to PCs capturing theoretical stellar features.

4. Results & Conclusion

Using FPCA, we have successfully reduced the dimensionality of stellar spectroscopic space with a basis of only 50 simulations. Figure 1 shows an example fit of an APOGEE M67 red giant spectrum using 10 functional principal components. These components have been computed using a sample of ~50,000 red giant stars in APOGEE with $-0.15 \leq [Fe/H] \leq 0.15$[1]. Using FPCA and DELFI, we have accurately and efficiently inferred the abundance of 15 different elements in open cluster M67 member stars given their APOGEE

[1] [X/H] refers to the logarithm of the abundance of element X in a star compared to its Hydrogen content with respect to that of the Sun. The unit used is the dex which stands for *decimal exponent*.

spectra. Figure 2 shows an example posterior distribution of 5 different chemical abundances, [C/H], [Mg/H], [Al/H], [Si/H], [Mn/H], [Fe/H] given a M67 red giant member APOGEE spectrum. Our work is in progress, and we are currently constraining the level of chemical homogeneity in open cluster M67. We will extend this to other open clusters, and potentially to globular clusters as well. Once we determine the level of chemical homogeneity in star clusters, we will put theoretical constraints on models of star-forming clouds. Eventually, we want to apply this method to infer the chemistry of the entire APOGEE spectroscopic survey that includes millions of stars. This will allow us to perform chemical tagging and provide the finest details on the evolution of our Galaxy.

5. Acknowledgements

The authors would like to thank Prof. Sebastian Jaimungal (University of Toronto, Department of Statistical Sciences, Toronto, Ontario, Canada).

References

Alsing, J., Charnock, T., Feeney, S., and Wand elt, B. (2019). Fast likelihood-free cosmology with neural density estimators and active learning. , 488(3):4440–4458.

Alsing, J., Wandelt, B., and Feeney, S. (2018). Massive optimal data compression and density estimation for scalable, likelihood-free inference in cosmology. *Monthly Notices of the Royal Astronomical Society*, 477(3):2874–2885.

Bailey, S. (2016). Weighted EMPCA: Weighted Expectation Maximization Principal Component Analysis.

Bishop, C. M. (1994). Mixture density networks. Technical report.

Bonassi, F., You, L., and West, M. (2011). Bayesian Learning from Marginal Data in Bionetwork Models. Statistical Applications in Genetics and Molecular Biology. , (10).

Bovy, J. (2016). THE CHEMICAL HOMOGENEITY OF OPEN CLUSTERS. *The Astrophysical Journal*, 817(1):49.

Donor, J., Frinchaboy, P. M., Cunha, K., Thompson, B., O'Connell, J., Zasowski, G., Jackson, K. M., McGrath, B. M., Almeida, A., Bizyaev, D., Carrera, R., García-Hernández, D. A., Nitschelm, C., Pan, K., and Zamora, O. (2018). The open cluster chemical abundances and mapping survey. II. precision cluster abundances for APOGEE using SDSS DR14. *The Astronomical Journal*, 156(4):142.

Fan, Y., Nott, D. J., and Sisson, S. A. (2013). Approximate bayesian computation via regression density estimation. *Stat*, 2(1):34–48.

Feigelson, E. D. and Babu, G. J. (2012). *Modern Statistical Methods for Astronomy: With R Applications*. Cambridge University Press.

Feng, Y. and Krumholz, M. (2014). Early turbulent mixing as the origin of chemical homogeneity in open star clusters. *Nature*, 513.

Ingrassia, S. and Costanzo, G. D. (2005). Functional principal component analysis of financial time series. In Bock, H.-H., Gaul, W., Vichi, M., Arabie, P., Baier, D., Critchley, F., Decker, R., Diday, E., Greenacre, M., Lauro, C., Meulman, J., Monari, P., Nishisato, S., Ohsumi, N., Opitz, O., Ritter, G., Schader, M., Weihs, C., Vichi, M., Monari, P., Mignani, S., and Montanari, A., editors, *New Developments in Classification and Data Analysis*, pages 351–358, Berlin, Heidelberg. Springer Berlin Heidelberg.

Lada, C. J. and Lada, E. A. (2003). Embedded Clusters in Molecular Clouds. , 41:57–115.

Liu, F., Asplund, M., Yong, D., Feltzing, S., Dotter, A., Meléndez, J., and Ramírez, I. (2019). Chemical (in)homogeneity and atomic diffusion in the open cluster M 67. , 627:A117.

Lueckmann, J.-M., Bassetto, G., Karaletsos, T., and Macke, J. H. (2018). Likelihood-free inference with emulator networks. *arXiv e-prints*, page arXiv:1805.09294.

Majewski, S. R., Schiavon, R. P., Frinchaboy, P. M., Allende Prieto, C., Barkhouser, R., Bizyaev, D., Blank, B., Brunner, S., Burton, A., Carrera, R., Chojnowski, S. D., Cunha, K., Epstein, C., Fitzgerald, G., García Pérez, A. E., Hearty, F. R., Henderson, C., Holtzman, J. A., Johnson, J. A., Lam, C. R., Lawler, J. E., Maseman, P., Mészáros, S., Nelson, M., Nguyen, D. C., Nidever, D. L., Pinsonneault, M., Shetrone, M., Smee, S., Smith, V. V., Stolberg, T., Skrutskie, M. F., Walker, E., Wilson, J. C., Zasowski, G., Anders, F., Basu, S., Beland, S., Blanton, M. R., Bovy, J., Brownstein, J. R., Carlberg, J., Chaplin, W., Chiappini, C., Eisenstein, D. J., Elsworth, Y., Feuillet, D., Fleming, S. W., Galbraith-Frew, J., García, R. A., García-Hernández, D. A., Gillespie, B. A., Girardi, L., Gunn, J. E., Hasselquist, S., Hayden, M. R., Hekker, S., Ivans, I., Kinemuchi, K., Klaene, M., Mahadevan, S., Mathur, S., Mosser, B., Muna, D., Munn, J. A., Nichol, R. C., O'Connell, R. W., Parejko, J. K., Robin, A. C., Rocha-Pinto, H., Schultheis, M., Serenelli, A. M., Shane, N., Silva Aguirre, V., Sobeck, J. S., Thompson, B., Troup, N. W., Weinberg, D. H., and Zamora, O. (2017). The Apache Point Observatory Galactic Evolution Experiment (APOGEE). , 154(3):94.

McKee, C. F. and Tan, J. C. (2002). Massive star formation in 100,000 years from turbulent and pressurized molecular clouds. , 416(6876):59–61.

Papamakarios, G. and Murray, I. (2016). Fast ϵ - free inference of simulation models with bayesian conditional density estimation. In Lee, D. D., Sugiyama, M., Luxburg, U. V., Guyon, I., and Garnett, R., editors, *Advances in Neural Information Processing Systems 29*, pages 1028–1036. Curran Associates, Inc.

Papamakarios, G., Pavlakou, T., and Murray, I. (2017). Masked Autoregressive Flow for Density Estimation. *arXiv e-prints*, page arXiv:1705.07057.

Papamakarios, G., Sterratt, D. C., and Murray, I. (2018). Sequential Neural Likelihood: Fast Likelihood-free Inference with Autoregressive Flows. *arXiv e-prints*, page arXiv:1805.07226.

Price-Jones, N. and Bovy, J. (2018). The dimensionality of stellar chemical space using spectra from the Apache Point Observatory Galactic Evolution Experiment. , 475(1):1410–1425.

Reddy, A. B. S., Giridhar, S., and Lambert, D. L. (2012). Comprehensive abundance analysis of red giants in the open clusters NGC 752, 1817, 2360 and 2506. , 419:1350–1361.

Roweis, S. T. and Saul, L. K. (2000). Nonlinear Dimensionality Reduction by Locally Linear Embedding. *Science*, 290(5500):2323–2326.

Shu, F. H., Adams, F. C., and Lizano, S. (1987). Star formation in molecular clouds - Observation and theory. , 25:23–81.

Wilson, J. C., Hearty, F., Skrutskie, M. F., Majewski, S., Schiavon, R., Eisenstein, D., Gunn, J., Blank, B., Henderson, C., Smee, S., Barkhouser, R., Harding, A., Fitzgerald, G., Stolberg, T., Arns, J., Nelson, M., Brunner, S., Burton, A., Walker, E., Lam, C., Maseman, P., Barr, J., Leger, F., Carey, L., MacDonald, N., Horne, T., Young, E., Rieke, G., Rieke, M., O'Brien, T., Hope, S., Krakula, J., Crane, J., Zhao, B., Carr, M., Harrison, C., Stoll, R., Vernieri, M. A., Holtzman, J., Shetrone, M., Allende-Prieto, C., Johnson, J., Frinchaboy, P., Zasowski, G., Bizyaev, D., Gillespie, B., and Weinberg, D. (2010). The Apache Point Observatory Galactic Evolution Experiment (APOGEE) high-resolution near-infrared multi-object fiber spectrograph. In *Ground-based and Airborne Instrumentation for Astronomy III*, volume 7735 of *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, page 77351C.