Dissertation in Astronomy

# Statistical methods in the era of large astronomical surveys

By

Arvind Christopher Nagarajah Hughes

Thesis Supervisors:
PD. Dr. Coryn Bailer-Jones
Prof. Dr. Daniel Zucker

Thesis dissertation committee:
PD. Dr. Coryn Bailer-Jones
PD. Dr. Andreas Koch-Hansen
Prof. Dr. Luca Amendola
Prof. Dr. Joachim Wambsganß

A thesis submitted to,
the Combined Faculties of the Natural Sciences and Mathematics
of Heidelberg University, Germany
for the degree of Doctor of Natural Sciences
and
The School of Mathematical and Physical Sciences, Macquarie
University, Australia,
for the degree of Doctor of Philosophy
June 2023

Typeset in $\LaTeX\,2_\varepsilon$.

This thesis is being submitted to Macquarie University and Heidelberg University in accordance with the Cotutelle agreement dated 12.06.2023.
To the best of my knowledge and belief, the thesis contains no material previously published or written by another person except where due reference is made in the thesis itself

Arvind Christopher Nagarajah Hughes

# Acknowledgements

This body of work would not have been achieved without the guidance and support of many people.

First and foremost, I extend my sincere appreciation to my supervisors, Daniel Zucker in Australia and Coryn Bailer-Jones in Germany. Their unwavering support, guidance, and patience throughout my PhD journey, particularly during the challenging times of the pandemic, played a crucial role in completing this endeavour. I am immensely grateful for their continuous motivation and mentorship. Above all, I am deeply thankful for the trust, independence, and flexibility they granted me, allowing me to grow and thrive as a researcher. Thank you both.

Thank you to the Gaia Group at MPIA for their invaluable support throughout the last year. Morgan Fouesneau, thank you for your coding assistance and Rene Andrae, thank you for answering my questions about Gaia and surveys. But a special thanks goes to Sara Jamal, who provided both conceptual insights and coding guidance, and assisted greatly with the editing of a Chapter in this work.

To my family, my parents and my four younger brothers, thank you for always being the loudest and most vocal supporters in all my endeavours (especially in the sports stands) and encouraging this pivot to a different career path.

The pandemic added a unique aspect to my PhD journey, the lost opportunity to build camaraderie and shared experiences with fellow researchers facing similar challenges. However, this changed when I relocated to Heidelberg for the latter part of my studies. I cherish the moments spent with Siddhant Deshmukh, Jacob Isbell, Paul Joseph, Jonas Syed, and Oliver Volkel, indulging in beer drinking, playing pool at Metropol, trash talking, and watching sports. Additionally clubbing with the raving mad techno lovers Alex Dimoff, Callie Clontz, Guillaume Guiglion and Steve Hannon. I would also like to express my gratitude to my office mates, Matt Gent and Zhang-Liang Xie, for the engaging conversations and endless supply of office snacks. To all of you, I extend my heartfelt thanks for the incredible memories we created together and the lasting friendships we have formed. Lastly to my friends in Australia, while there are too many to name individually, I want to give special thanks to Faraz Syed and family, as well as Dylan Lamprecht, for always being there, no matter the distance.

Finally my partner, Liv, what an experience this has been from the cramped confines of our small apartment at the beginning of the pandemic, to moving to the other side of the world together. Thank you for embarking on this adventure with me. Your openness, thoughtfulness, care and unrelenting belief has made the experience easy and thoroughly enjoyable.

I am sincerely grateful to each and every one of you.

# List of Publications

**Journal (peer reviewed)**

**2022**

1. **Hughes A. C. N.**, Spitler L. R., Zucker D. B., Nordlander T., Simpson J., da Costa G. S., Ting Y.-S., et al.,*”The GALAH Survey: A New Sample of Extremely Metal-poor Stars Using a Machine-learning Classification Algorithm”*, 2022, ApJ, 930, 47
2. **Hughes A. C. N.**, Bailer-Jones C. A. L., Jamal S., *”Quasar and galaxy classification using Gaia EDR3 and CatWise2020”*, 2022, A&A, 668, A99
3. Da Costa G. S., Bessell M. S., Nordlander T., **Hughes A. C. N.**, Buder S., Mackey A. D., Spitler L. R., et al.,*"Spectroscopic follow-up of statistically selected extremely metal-poor star candidates from GALAH DR3"* 2022, arXiv, arXiv:2210.05161

# Abstract

Statistical methods play a crucial role in modern astronomical research. The development and understanding of these methods will be of fundamental importance to future work on large astronomical surveys. In this thesis I showcase three different statistical approaches to survey data. I first apply a semi-supervised dimensionality reduction technique to cluster similar high resolution spectra from the GALAH survey to identify 54 candidate extremely metal-poor stars. The approach shows promising potential for implementation in future large-scale stellar spectroscopic surveys. Next, I employ a method to classify sources in the Gaia survey as stars, galaxies or quasars, making use of additional infrared photometry from CatWISE2020 and discussing the importance of applying adjusted priors to probabilistic classification. Lastly, I utilise a method to estimate the rotational parameters of star clusters in Gaia, with an application to open clusters. This is done by considering the rotation of a cluster as a 3D solid body, and finding the best fitting parameters by sampling constructed likelihood functions. The methods developed in this thesis underscore the significant contributions statistical methodologies make to astronomy, and illustrate how the development and application of statistical methods will be essential for extracting meaningful insights from future large scale astronomical surveys.

# Zusammenfassung

Statistische Methoden spielen in der modernen astronomischen Forschung eine entscheidende Rolle. Die Entwicklung und das Verständnis dieser Methoden sind von grundlegender Bedeutung für die zukünftige Arbeit an großen astronomischen Durchmusterungen. In dieser Arbeit stelle ich drei verschiedene statistische Ansätze für Durchmusterungsdaten vor. Zunächst wende ich ein halbüberwachtes Verfahren zur Dimensionalitätsreduktion an, um ähnliche hochauflösende Spektren aus der GALAH-Durchmusterung zu bündeln und 54 Kandidaten für extrem metallarme Sterne zu identifizieren. Der Ansatz zeigt ein vielversprechendes Potenzial für den Einsatz in zukünftigen groß angelegten stellaren spektroskopischen Durchmusterungen. Als Nächstes wende ich eine Methode zur Klassifizierung von Quellen in der Gaia-Durchmusterung zu Sternen, Galaxien oder Quasare an, wobei ich zusätzliche Infrarot-Photometrie von CatWISE2020 verwende und die Bedeutung der Anwendung angepasster Prioritäten für die probabilistische Klassifizierung erörtere. Schließlich verwende ich eine Methode zur Schätzung der Rotationsparameter von Sternhaufen in Gaia, mit einer Anwendung auf offene Sternhaufen. Dabei wird die Rotation eines Sternhaufens als 3D-Volumenkörper betrachtet, und die am besten passenden Parameter werden mit Hilfe von Stichproben konstruierter Likelihood-Funktionen ermittelt. Die in dieser Arbeit entwickelten Methoden unterstreichen den bedeutenden Beitrag, den statistische Methoden für die Astronomie leisten, und verdeutlichen, wie wichtig die Entwicklung und Anwendung statistischer Methoden für die Gewinnung aussagekräftiger Erkenntnisse aus künftigen groß angelegten astronomischen Durchmusterungen sein wird.

# Contents

# List of Figures

# List of Tables

# 1
# Prolegomenon

## 1.1 Outline

The thesis explores promising areas in astrostatistics for more efficient extraction of astronomical results from, and applications in, future astronomical surveys.

Chapter 2 begins with an overview of the history and future prospects of astrostatistics and large astronomical surveys. It also delves into the different surveys and the astronomical objects that are under consideration.

In Chapter 3, a semi-supervised approach is presented for identifying various stellar types in large spectroscopic surveys, particularly when only a limited number of stellar types are known. The method utilises t-stochastic neighbor embedding (t-SNE), a dimensionality reduction technique that facilitates visualising object similarity in a 2D space. By overlaying unknown objects near known objects, the method improves the identification process, with close proximity suggesting (in this case) spectral similarity. This technique is applied to the Galactic Archaeology with HERMES (GALAH) spectroscopic survey to identify rare, extremely metal-poor stars.

Shifting the focus from spectra, Chapter 4 introduces a statistical method to classify extragalactic sources into three classes – stars, galaxies, and quasars – based on their positions and photometry in the Gaia survey. This method builds upon the Gaussian Mixture and Gaia-only model discussed in Bailer-Jones et al. (2019), by incorporating two gradient boosted models, and infrared photometry from the CatWISE 2020 survey. Additionally a latitude and magnitude-dependent prior is applied to enhance the representativeness of the results.

Chapter 5 investigates the evidence of solid-body rotation in star clusters using Gaia data. A method inspired by the work of Sollima et al. (2019) is employed to identify and quantify rotation focusing on open clusters. The validity of the method is tested using simulated clusters and known globular clusters, before being applied to open clusters.

The final chapter concludes the thesis, providing a brief discussion on future prospects and an outlook in the field of astrostatisics and large surveys.

# 2 Background

## 2.1 Astronomy and Statistics

### 2.1.1 Brief history

Astronomy and statistics have an interwoven history. The astronomer is restricted to observing external characteristics of objects populating the universe, and inferring from these data their properties and underlying physics (Feigelson, 2009). Even in ancient times, civilisations conducted significant quantitative measurements of celestial phenomena. One of the earliest statistical methods, developed by the Greek astronomer Hipparchus, involved the calculation of a form of arithmetic mean and variance. By observing the duration of the day and the interval between solstices, Hipparchus estimated the day's variability by employing half the range of his measurements (Plackett, 1958).

The challenges encountered in astronomy prompted early researchers to devise new statistical methodologies. Initially, statistics in the 18th century focused on data collection and compilation but gradually evolved to concentrate on the development of mathematical techniques for data analysis and interpretation. The late 18th and 19th centuries witnessed the ascendance of statistical methods in astronomy, thanks to mathematical visionaries like Johann Carl Friedrich Gauss and Pierre-Simon Laplace. Both Gauss and Laplace developed the method of least squares within a probabilistic framework, demonstrating its superiority in determining orbital parameters from astronomical observations (Feigelson, 2009). Consequently, the method of least squares swiftly emerged as the principal tool connecting astronomical observations with celestial mechanics. Gauss also pioneered methods for handling observational measurement errors and introduced his renowned Gaussian, or "Normal", distribution, which was commonly known as the "astronomical error function" throughout much of the 19th century (Gauss and Stewart, 1995).

For early 20th century observational astronomers, the method of least squares remained the primary statistical tool. One notable example of a linear relationship derived from observations was Hubble's Law, which describes the expansion of the universe by establishing a linear equation between the recessional velocity of external galaxies and their proper distance (Hubble, 1929). Despite this, two significant statistical advancements, namely Bayes' Theorem by Thomas Bayes (Bayes and Price, 1763) and Maximum Likelihood Estimation by Ronald Fisher (Fisher, 1922), were sporadically applied in astronomy throughout the early 20th century. By the 1980s and 1990's however, Maximum Likelihood Estimation, in particular, had already left a profound impact on a variety of fields, including image restoration (Lucy, 1974) and the calculation of the galaxy luminosity function in extragalactic astronomy (Efstathiou et al., 1988).

Early in the 21st century, however, astronomy found itself experiencing an unprecedented flood of data (Ball and Brunner, 2010). Due to the rapid advancement of astronomical instruments, particularly the CCD, and significant progress in computer technology, the field of astronomy witnessed remarkable developments in the quantity of available data, and in greatly enhanced data processing and storage capabilities. Data-driven disciplines are increasingly faced with the problem of how to store, organise, use, and interpret the enormous amounts of data being generated by new research infrastructure (Szalay and Gray, 2001). This push into "big data" has seen astrostatistical methodology grow rapidly and emerge as an active area of research.

### 2.1.2 Astrostatistics

Astrostatistics, an interdisciplinary field that merges statistical techniques and data science with the realm of astronomy, holds immense significance in the extraction of meaningful information from vast datasets gathered through large-scale astronomical surveys. Its objectives encompass identifying data patterns, discerning trends, and characterising novel astronomical objects. Moreover, astrostatistics serves as a vital component in the planning and design of astronomical surveys, aiding in optimising observations and data collection for maximal scientific returns.

One of the foremost challenges encountered in astrostatistics involves grappling with substantial levels of noise and uncertainty within the data. This encompasses addressing sources of error like measurement noise, as well as managing missing or incomplete data. To overcome these hurdles and extract reliable information from the data, we can employ a variety of techniques, including Bayesian inference and machine learning.

Astrostatististics is applicable to a wide array of astronomical problems and will play a pivotal role in forthcoming large-scale astronomy surveys such as the imminent Legacy Survey of Space and Time (Ivezic et al., 2019) and ESA's Euclid mission (Euclid Collaboration et al., 2022). These surveys will amass enormous amounts of data, necessitating the implementation of astrostatistical methods for data analysis and comprehension. In addition to traditional astronomical surveys, astrostatistics also assumes a vital role in time-domain astronomy, that is, the collection of data over temporal intervals. This field encompasses the study of variable stars, supernovae, and other transient phenomena. Astrostatistical techniques are employed to model the light curves of these objects, facilitating their classification into distinct types (e.g., Richards et al., 2011; Sanders et al., 2015; Lochner et al., 2016).

Cosmology is another realm where astrostatistics finds extensive application. Within the study of the universe's large-scale structure and evolution, astrostatistics enables the analysis of vast-scale galaxy surveys and the exploration of matter distribution throughout the cosmos. This includes investigations into the cosmic microwave background radiation, the distribution of galaxy clusters and quasars. By scrutinising such observed and simulated data, we can glean insights into the properties of dark matter and dark energy, while also deepening our

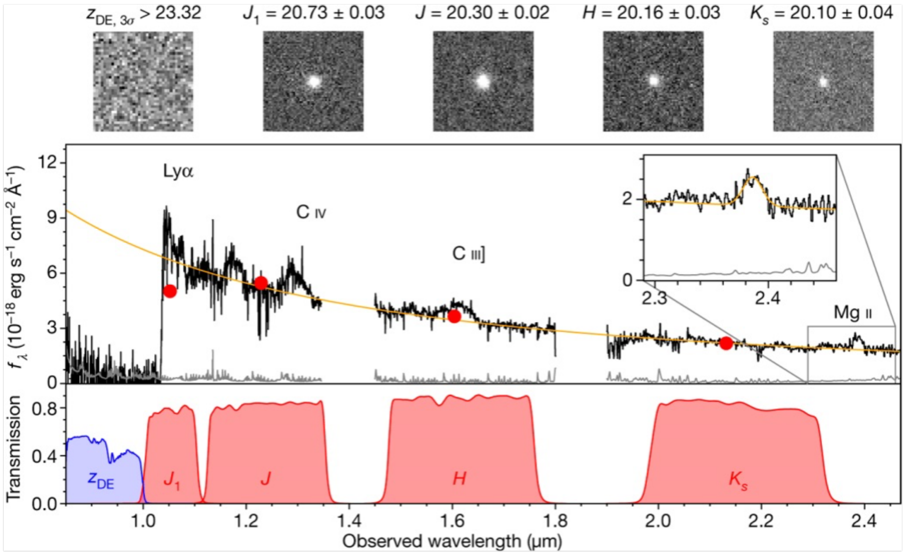


Figure 2.1: Highest redshift quasar found by Bañados et al. (2018), at $z \sim 7.5$. Image credit: Bañados et al. (2018)

understanding of the universe's origin and evolution (e.g., Brehmer et al., 2019; Ntampaka et al., 2020; Villaescusa-Navarro et al., 2021). Techniques for data mining large surveys for quasars, including AllWISE, DECALS, and the Sloan Digital Sky Survey (SDSS) have led to significant progress in the hunt for high-redshift quasars. Focusing on this search, several groups have used artificial neural networks (Claeskens et al., 2006; Carballo et al., 2008), SVM and learning vector quantisation (Zhang and Zhao, 2003), and kernel density estimation (Richards et al., 2009b). Many of these works combine multiwavelength data, particularly X-ray, optical and radio. An example of a high-redshift quasar found through these methods can be seen in Fig.2.1, taken from the work of Bañados et al. (2018).

Astrostatistics also proves invaluable in the realm of exoplanetary studies, focusing on planets orbiting stars beyond our solar system. Typically, data from exoplanet surveys are sparse and tainted by noise; see Fig. 2.2 for an example of a lightcurve from a seminal paper by Southworth (2011). Astrostatistical methods prove invaluable in modelling and analysing this type of data to identify exoplanets. These methods involve the application of algorithms, which can effectively classify exoplanets based on their specific properties by extracting information, potentially not seen with traditional methods from large-scale surveys (e.g., Márquez-Neila et al., 2018; Shallue and Vanderburg, 2018; Kunimoto and Matthews, 2020; Giacalone et al., 2021).

Astrostatistics stands as a rapidly expanding field critical to the exploration of astronomy and cosmology. By uniting statistical techniques, data science, and astronomical studies, astrostatistics empowers us to extract valuable insights from extensive datasets and unlock novel discoveries about the universe. As astronomical surveys advance, accumulating ever-increasing volumes of data, the role of astrostatistics will continue to grow in importance. There are five key reasons for our interest in this work: accurate inference, characterising uncertainty, dealing with complex data, dealing with incomplete and/or biased data and, lastly, the most fruitful use of the methodology, data driven discoveries:

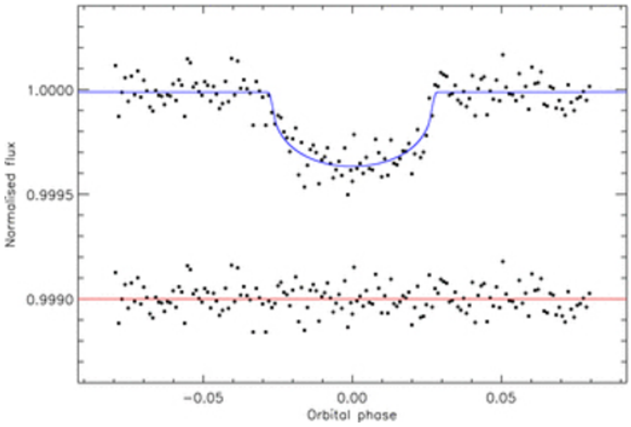


Figure 2.2: An example of a transiting exoplanet. Image credit: Southworth (2011)

- Accurate inference: Astronomy is an observational science and our understanding of the universe relies heavily on interpreting data collected from telescopes and other instruments. The astrostatistical method allows us to make robust inferences and draw meaningful conclusions from these observations. By ensuring the accuracy and reliability of our statistical analyses, we can confidently interpret the data and make sound scientific claims about the nature of celestial objects and phenomena.

- Characterising uncertainty: Uncertainty is an inherent aspect of any scientific measurement or observation. Astrostatistical methodology provides us with tools to quantify and characterise this uncertainty in a rigorous manner. By understanding the uncertainties associated with our measurements, we can establish the confidence levels of our results and avoid drawing overly confident or misleading conclusions. This is particularly crucial when making predictions or making claims about rare or extreme objects or events.

- Dealing with complex data: Astronomical datasets are often complex, featuring various sources of noise, systematic effects, and correlations. Astrostatistical techniques help us model and account for these complexities, allowing us to separate genuine astrophysical signals from the noise and other confounding factors. This enables us to extract valuable information from the data and uncover subtle patterns fundamental to our understanding of the universe.

- Handling incomplete and biased data: We frequently encounter incomplete and biased data due to various observational limitations and selection effects. Astrostatistical methods provide us with powerful tools to handle such data and mitigate the impact of these limitations. By carefully accounting for selection biases and developing appropriate statistical corrections, we can obtain more accurate and representative measurements of astronomical properties.

- Data driven discovery: With the advent of large-scale astronomical surveys and projects, we are witnessing an unprecedented growth in the volume and complexity of data. Astrostatistical techniques play a crucial role in analysing and extracting meaningful information from these massive datasets. By developing and applying sophisticated statistical methods, we can identify subtle trends, patterns, and correlations that might have otherwise gone unnoticed.

Astrostatistics will remain indispensable for data analysis, survey design, and optimisation, fueling continued progress in the field, and this thesis aims to showcase its application to these areas.

### 2.1.3 Statistical approaches

Having outlined some of the diverse application of statistics within astronomy, we turn our attention to the framework of astrostatistics and expand upon concepts described in Section 2.1.2. Two key components in astronomy are inference and prediction. Inference creates a mathematical model of the data generation process to formalise understand or test a hypothesis about how the system behaves. Prediction aims at forecasting unobserved outcomes or future movements, such as orbital dynamics, or the redshift of an object.

In statistics and machine learning (ML), there exists a distinction between the goals of prediction and inference. While both approaches can be employed for prediction and inference, statistical methods have traditionally focused on inference by constructing and fitting project-specific probability models. These models enable the computation of a quantitative measure of confidence in the presence of a discovered relationship, indicating that it represents a genuine effect rather than mere noise. Moreover, with sufficient data, assumptions such as equal variance can be explicitly verified, and the model can be refined if necessary.

On the other hand, ML primarily emphasises prediction by utilising versatile learning algorithms to identify patterns within complex and extensive datasets. ML methods prove particularly valuable when dealing with "wide data", where the number of input variables exceeds the number of subjects, as opposed to "long data", where the number of subjects surpasses the input variables. ML techniques require minimal assumptions about the data-generating system, allowing them to be effective even in the absence of a carefully controlled experimental design and in the presence of intricate nonlinear interactions. However, despite delivering compelling prediction outcomes, the lack of an explicit model can make it challenging to directly connect ML solutions with existing astronomical knowledge.

As the number of input variables and potential associations among them increases, the complexity of the model needed to capture these relationships also grows. This presents a challenge for statistical inference, as the precision of inferences tends to diminish. In this context, the boundary between statistical approaches and ML approaches becomes less distinct and more ambiguous. The intricate interplay between variables and the intricate nature of the relationships make it harder to rely solely on traditional statistical methods, prompting the need for ML techniques that can effectively handle such complexity.

I will briefly define two fundamental approaches in ML: supervised and unsupervised learning. Figure 2.3, from scikit-learn, elegantly illustrates the methods one can consider given the data at hand. I will focus mainly on classification, a core aspect of this thesis.

Supervised learning involves training a model using a labelled dataset, where each data point is associated with a known target variable or class label. The objective is to establish a mapping between the input variables and the desired output variable, enabling the model to make predictions on new, unseen data. In the context of astrostatistical classification tasks, supervised learning algorithms can be employed to assign objects into predefined classes or categories based on their input features. For instance, in the identification of various stellar types, a supervised learning algorithm can be trained on labelled spectroscopic data, with each spectrum annotated with a specific stellar type. Subsequently, the trained model can be utilised to classify new spectra into the appropriate stellar type based on their features.

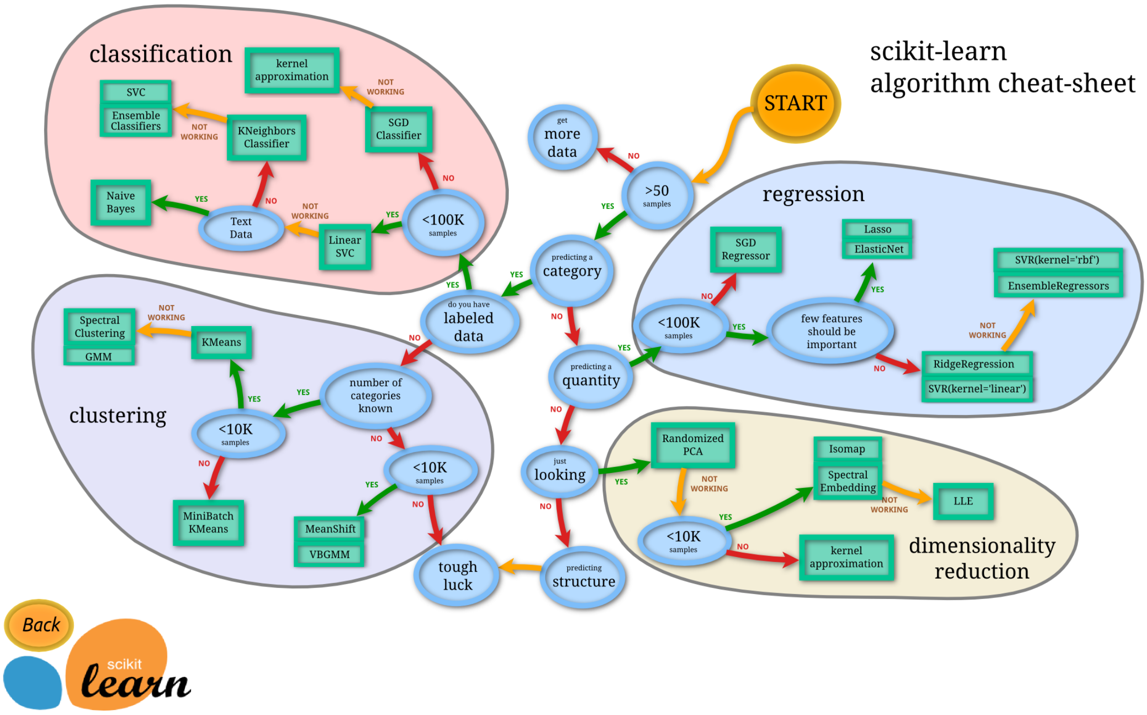


FIGURE 2.3: The statistical approach flow chart. Image credit: The team at Scikit-learn (Pedregosa et al., 2011).

Classification, an integral task in machine learning, holds significant importance in astrostatistics. It involves assigning data points or objects to predefined categories or classes based on their features. Classification algorithms learn patterns and relationships from labelled training data, enabling them to make predictions on unseen instances. Classification techniques prove valuable in identifying different types of astronomical objects, including stars, galaxies, or quasars, based on their observed properties. By training a classification model on labelled data, it becomes possible to classify new observations into the appropriate astronomical class, thereby facilitating automated object identification and categorisation.

On the other hand, unsupervised learning deals with datasets that lack explicit labels or target variables. The main objective is to uncover inherent patterns, structures, or relationships within the data without any predefined categories or classes. Unsupervised learning algorithms strive to identify natural clusters or groups within the data, discover underlying dimensions, or detect anomalies. In the realm of astrostatistics, unsupervised learning techniques find application in various tasks, such as clustering similar objects based on their properties or detecting rare astronomical phenomena. For example, unsupervised learning algorithms can be applied to categorise galaxies into distinct groups based on their observational characteristics, without prior knowledge of their classes or types.

## 2.2 Large astronomical surveys

Astronomical surveys are systematic observations of the night sky on a large scale, aiming to discover and map the positions and properties of the objects detected. These surveys have evolved significantly since the early 20th century, when photographic plates and more advanced telescopes were introduced to capture detailed sky images.

One of the most notable surveys in the history of astronomy is the Harvard Photometry, also known as the Harvard College Observatory Photographic Plate Collection. Initiated in the late 19th century, this survey utilised photographic plates to record images of the night sky. These images were then used to measure the brightness and positions of stars and other celestial objects, resulting in the creation of the first comprehensive star catalogue (Turner, 2003).

The advent of digital imaging technology in the late 20th century revolutionised the field of astronomy, enabling more precise and efficient surveys. Among these advancements, the Sloan Digital Sky Survey (SDSS) (York et al., 2000) stands out as a significant milestone. The SDSS, which began in 1998 and has continued with evolving instrumentation and objectives to the present day, was initially a large-scale astronomical survey conducted using a dedicated telescope at the Apache Point Observatory in New Mexico, USA. Equipped with a state-of-the-art camera, the telescope captured detailed images of the night sky. The primary objective of the SDSS was to map the three-dimensional structure of the universe by observing and measuring the properties of galaxies, quasars, and stars such as the redshifts for extragalactic objects (Abazajian et al., 2003) . Covering about one-quarter of the entire sky, the survey produced a massive dataset containing over 930,000 galaxies, 120,000 quasars, and more than a billion stars. The SDSS also made significant contributions to our understanding of dark matter, unveiling its distribution in the universe by studying the motion and spatial arrangement of galaxies.

The success of the SDSS inspired the initiation of other large-scale astronomical surveys. Two notable ongoing surveys are the Gaia mission and the GALAH survey. Large astronomical surveys such as GALAH and Gaia (described in more detail below) offer several advantages over smaller, targeted surveys. Their wide sky coverage increases the likelihood of discovering rare or unique objects. The larger sample size facilitates statistical analysis and the identification of patterns and trends in the data. Moreover, these surveys allow for the observation and study of a diverse range of objects and phenomena, leading to a more comprehensive understanding of the universe. In the future, large astronomical surveys will continue to play a crucial role in advancing our understanding of the universe. Surveys like Euclid will collect even more extensive data, enabling new discoveries and providing deeper insights into the properties and evolution of the universe.

### 2.2.1 Galactic Archaeology with HERMES (GALAH) Survey

GALAH – the GALactic Archaeology with HERMES (High Efficiency and Resolution Multi-Element Spectrograph) survey – is a large-scale astronomical survey aimed at studying the structure, evolution, and history of the Milky Way galaxy. Launched in 2013, the GALAH Survey is a collaboration between institutions in Australia, Europe, and Asia, and involves the use of the aforementioned HERMES instrument, which is mounted on the 3.9-meter Anglo-Australian Telescope (AAT) in New South Wales, Australia. HERMES is a high-resolution spectrograph capable of measuring the spectra of up to 400 stars simultaneously, providing detailed information about the chemical composition, age, and kinematics of these stars (Simpson et al., 2016). The primary objective of GALAH is to obtain a comprehensive understanding of the Milky Way's formation and evolution (De Silva et al., 2015). To achieve this, GALAH aims to collect approximately 1,000,000 high-resolution stellar spectra (with a resolution of approximately 28,000) for elemental abundance analysis. GALAH strives for a precision of 0.05 dex in elemental abundance, necessitating a signal-to-noise ratio of $\sim 100$. Consequently, the observed magnitude range is limited to $12 < V < 14$ towards the Galactic plane. Based on these constraints, it is anticipated that the final GALAH survey sample will consist of roughly 77% thin disk stars, 22% thick disk stars, 0.8% bulge stars, and 0.2% halo stars (Martell et al., 2017). The chemical compositions of these stars will yield insights into the formation and evolution of the Milky Way, including the roles played by mergers, accretion, and star formation in shaping the Galaxy's structure and chemical makeup.

This thesis uses detailed high resolution GALAH spectra to identify extremely metal-poor stars within the survey, which is discussed in Chapter 3.

### 2.2.2 Gaia Survey

The Gaia survey is an astrometry mission of the European Space Agency (ESA) that was launched in 2013. The primary goal of the mission is to create a precise three-dimensional map of the Milky Way, by measuring the position, distance, and motion of over a billion stars in our Galaxy (Gaia Collaboration et al., 2016). One of the most significant achievements of the Gaia survey is the creation of the largest and most precise 3D map of the Milky Wayto date. This map shows the position and motion of stars in our galaxy in unprecedented detail, allowing study of the structure, dynamics, and evolution of the Galaxy. Gaia's measurements have enabled the study of the properties and behavior of stars in our galaxy, including their formation, evolution, and interactions. Gaia has also been used to study the distribution and composition of dark matter in the Milky Way and to search for exoplanets (i.e., planets outside our solar system) by detecting their gravitational influence on nearby stars. In addition, Gaia has enabled incredibly accurate 3D mapping of dust within our Galaxy. Overall, the Gaia survey is a groundbreaking mission that has revolutionised our understanding of the Milky Way and the universe.

This thesis uses astrometry and broad band photometry in the G, $G_{\mathrm{BP}}$, and $G_{\mathrm{RP}}$ bands for about 1.8 billion sources in Gaia Early Data Release 3 (eDR3) (Riello et al., 2021) to classify sources into stars, galaxies and quasars discussed in Chapter 4. Chapter 5 uses the precise positions, distances, and proper motions in Gaia Data Release 3 (DR3) (Gaia Collaboration et al., 2022), to estimate the internal rotational properties of clusters, particularly focusing on open clusters.

### 2.2.3 CatWISE2020

In 2013, NASA's Wide-field Infrared Survey Explorer mission (WISE) was repurposed as NEOWISE, with a primary focus on searching for near-Earth objects. The mission's AllWISE catalog, released in the same year, was the outcome of combining multiple exposures from the first year of WISE surveying. Since then, NEOWISE has been releasing individual exposures annually. Meisner et al. (2019) utilised the unWISE processing to create an image atlas by combining the data from the 2010 and 2011 exposures used in AllWISE with the 2013-2016 NEOWISE data. This was further utilised by Schlafly et al. (2019) to generate an unWISE catalog, which identifies the sources found in these combined exposures using a point-source photometry code called “crowdsource”, specifically designed for crowded fields. CatWISE is a program that aims to catalogue sources by combining data from the WISE and NEOWISE all-sky surveys at wavelengths of 3.4 and $4.6\mu m$ (W1 and W2). The CatWISE Preliminary Catalogue comprises 900,849,014 sources, measured using data collected from 2010 to 2016. This dataset represents a significant increase in both the number of exposures (four times as many) and the time baseline (over ten times as long) compared to the AllWISE Catalogue. CatWISE employs the software from AllWISE to measure sources in coadded images created from six-month subsets of this data, with each subset covering the entire sky during a particular epoch. The CatWISE2020 Catalogue (Marocco et al., 2021), however, extends the time baseline by two years compared to the CatWISE Preliminary Catalogue. It also utilises the "Crowdsource" code as the detection software. These two enhancements result in approximately twice as many sources being included in the CatWISE 2020 Catalogue compared to the CatWISE Preliminary Catalogue, with a total of 1,890,715,640 sources spanning the entire sky. The photometry in the W1 and W2 bands is used in addition to the Gaia photometry in Chapter 4, as the combined use of infrared data with optical photometry should, in principle, enhance the performance of an extragalactic classifier.

## 2.3 Astronomical objects

A primary focus of this thesis is on classification and the various methods we may be able to use to identify different astronomical objects. The following section gives a description of the types of objects considered: metal-poor stars, extragalactic objects and star clusters.

### 2.3.1 Metal-poor stars

The metallicity of a star is typically expressed as its [Fe/H] ratio, which logarithmically compares the amount of iron in the star (relative to hydrogen) to the ratio of those elements found in the Sun; a star with $[\mathrm{Fe/H}] = 0$ would have the same ratio of iron to hydrogen as the Sun, while a star with $[\mathrm{Fe/H}\ ] = -1$ would have 10% of the Sun's iron abundance. Stars with $[\mathrm{Fe/H}] < -3$ – that is, stars with less than 0.1% of the Sun's iron abundance – are classified as extremely metal-poor (EMP) stars.

EMP stars are of particular interest as they are among the oldest and most chemically primitive objects in the universe. These stars formed during the early universe when there were very few heavy elements, and the universe was primarily composed of hydrogen and helium (e.g., Beers and Christlieb, 2005). Metal-poor stars provide information about the process of stellar nucleosynthesis, which is the process by which elements heavier than hydrogen and helium are formed in stars. Metal-poor stars have a low metallicity because the gas out of which they formed had not yet been significantly enriched by the nucleosynthetic products of previous generations of stars, which disperse heavy elements into the interstellar medium through stages of stellar evolution and supernovae. As such, EMP stars offer valuable insights into the conditions that existed during the universe's infancy, as well as the processes that led to the formation of the first stars (Frebel and Norris, 2015).

One of the most significant challenges in studying EMP stars is their rarity. Since these stars are chemically primitive and formed early in the universe's history, they tend to only be found in the oldest parts of the Galaxy. The majority of EMP stars are located in the halo of the Milky Way (Tumlinson, 2010), a region surrounding the Galaxy's central bulge that contains some of the Galaxy's oldest stars. Finding and studying these stars requires specialised techniques and instruments, which limits our ability to study them in detail.

Despite these challenges, significant strides have been made in understanding EMP stars. One notable discovery in this field was the identification of HE 0107-5240, an EMP star with [Fe/H] $< -5$ [dex]. This star was discovered in 2001 by Christlieb et al. (2002) using the Hamburg/ESO survey, a large-scale survey of the southern sky that aimed to identify metal-poor stars. The discovery of HE 0107-5240 was significant as it was the first star found with [Fe/H] $< -5$, indicating its extremely low metal abundance. Subsequently more metal-poor stars have been found over the years, and the lowest [Fe/H] star to date is SMSS J031300.36-670839.3, which Keller et al. (2014) estimated to have a metallicity of [Fe/H] $< -7$ [dex], later revised to be [Fe/H] $< -6.53$ by Nordlander et al. (2019).

As noted above, metal-poor stars are typically found in the halo of the Milky Way (e.g., Cordoni et al., 2021; Sestito et al., 2019; An et al., 2013). The halo is the roughly spherical region that surrounds the disk of the galaxy and contains the oldest stars. Metal-poor stars in the halo are thought to have formed early in the history of the Milky Way, when the Galaxy was still forming and there were smaller amounts of heavy elements available. These stars are also found in globular clusters, which are tightly packed groups of stars that orbit around the galaxy. Globular clusters are some of the oldest structures in the Milky Way, and hence the metal-poor stars in these clusters can provide important information about the early universe. Studies of metal-poor stars have led to important discoveries in astrophysics. For example, the metal-poor star HD 140283, also known as the Methuselah star, has been found to be one of the oldest stars in the Milky Way. Its age was at one point estimated to be around 14.5 billion years (Bond et al., 2013), which is greater than the consensus age of the universe itself; but whatever its true age, the star still has significant amounts of heavy elements, indicating that there must have been at least one (and likely more) preceding generation of stars. This discovery strongly suggests that there are even older stars that have not yet been found.

In addition to being important for understanding the early universe, metal-poor stars are also useful for studying the properties of stars themselves. Metal-poor stars have a simpler chemical composition than stars with a higher metallicity, which can make them easier to model and understand, which in turn can help us to better understand the processes that govern the evolution of stars (e.g., Placco et al., 2019; Sakari et al., 2018; Spite et al., 2018; Lee et al., 2013).

Metal-poor stars are typically identified through spectroscopic analysis, utilising methods like equivalent width measurements to compare absorption lines of elements such as calcium (Ca) and iron (Fe) with metal-poor spectral templates. This allows for the derivation of stellar parameters including effective temperature, surface gravity, and overall metallicity. However, despite EMP stars showing few spectroscopic absorption features from metals, trying to identify a relatively line-free spectrum in real spectra that contain noise can in practice be problematic. Photometric surveys, such as the SkyMapper survey (Da Costa et al., 2019) in the southern hemisphere, have also proven effective in finding metal-poor stars by examining color indices like B-V or U-B, as metal-poor stars tend to have bluer colors compared to their metal-rich counterparts. Subsequently, these candidates undergo spectroscopic follow-up. Given the rarity of metal-poor stars and the high number of large spectroscopic and photometric surveys, advancements in computing power and the development of efficient statistical techniques have facilitated more effective methods for exploring these surveys. Chapter 3 describes a potential statistical methodology that could be employed for this purpose, as well as for identifying any other stellar type of interest.

### 2.3.2 Extragalactic objects

Extragalactic objects are celestial bodies that are located beyond the boundaries of our Milky Way. These objects offer unique opportunities to explore the Universe on a larger scale, and to test our understanding of astrophysical processes under extreme conditions. Among the various broad categories of extragalactic objects, quasars and galaxies are perhaps the most intriguing and scientifically well-studied.

Quasars, short for "quasi-stellar objects," are located at cosmological distances and powered by matter accretion onto supermassive black holes. They serve as probes for black hole physics and the early Universe (e.g, Croom et al., 2009). Quasar spectra reveal broad emission lines, indicating gas motion and providing information about the gas distribution. Studying quasars has revealed important properties and constraints on the growth of supermassive black holes and host galaxies. Absorption lines in quasar spectra, generally due to foreground material, shed light on the intergalactic medium and early galaxy formation.

Galaxies, on the other hand, are vast collections of stars, gas, dust, and dark matter that trace the Universe's structure and evolution, dark matter and dark energy, and star and planet formation. The two categories of objects are not completely independent; quasars' high luminosity, resulting from accretion onto their black holes, generates strong radiation pressure and outflows that impact surrounding gas and stars, potentially influencing star formation and galaxy evolution.

As quasars emit across the entire electromagnetic spectrum they can be identified in optical surveys such as the SDSS by searching for point sources with unusual colors or strong emission lines, in X-rays surveys using telescopes like Chandra due to high energy emissions from jets and accretion disks, in radio wavelengths using instruments such as the Australia Telescope Compact Array, and in infrared surveys such as the Wide-field Infrared Survey Explorer (WISE). This results in many different approaches for finding quasars, the most straightforward being simple colour selection (Schneider et al., 2007), ranging to Bayesian probabilistic methods (e.g., Bailer-Jones et al., 2019; Richards et al., 2009a) and then machine learning methods using artificial neural networks (Yèche et al., 2010).

Galaxies exhibit a diverse range of spectral features that provide crucial information about their physical properties and evolution. By analysing galaxy spectra, we can uncover details about star formation, the presence of young or old stars, gas and dust distribution, interstellar medium composition, and the occurrence of active galactic nuclei and other exotic phenomena. These spectra offer insights into the star formation history, chemical enrichment, and dynamical evolution of galaxies. Moreover, galaxies serve as valuable probes for studying dark matter, which constitutes a significant portion of the Universe's matter. By observing the motions of stars and gas within galaxies, we can infer the distribution and quantity of dark matter, leading to the discovery of dark matter halos that extend beyond the visible components and shape the large-scale structure of the Universe.

Morphology, or shape, is another crucial aspect of galaxies. They are broadly classified into three main types: elliptical, spiral, and irregular. Elliptical galaxies possess a round or oval shape and consist predominantly of old stars. Spiral galaxies, on the other hand, exhibit a flattened disk-like structure with spiral arms containing young stars and gas. Irregular galaxies lack a well-defined structure and do not fall within the standard classes defined in the Hubble sequence, but typically show signs of star formation.

The properties of galaxies can be strongly influenced by their environment, with factors like size, morphology, and star formation rate varying based on the density of surrounding gas and other galaxies. The study of galaxy distribution and clustering has yielded valuable information about the large-scale structure of the Universe and the characteristics of dark matter.

Technological advancements and improved observational techniques have revolutionised the study of extragalactic objects, with observatories spanning the electromagnetic spectrum, from radio waves to gamma rays, driving the exploration of the universe beyond the Milky Way. Large-scale surveys like the SDSS and the Dark Energy Survey (Abbott et al., 2018) have facilitated the study of galaxy and quasar distribution and clustering on unprecedented scales, providing critical insights into the universe's large-scale structure and its evolution since the Big Bang.

Future studies of extragalactic objects will continue to push the boundaries of our understanding of the universe. Advances in technology, coupled with new facilities like the James Webb Space Telescope and the Square Kilometer Array, will enable even more detailed and sensitive observations. These developments will undoubtedly yield new insights into the structure and evolution of the cosmos.

In Chapter 4, I present a statistical methodology for identifying quasars and galaxies using astrometry and photometry within the Gaia survey.

### 2.3.3 Star clusters

Star clusters – associations of stars that formed together – are celestial objects that offer crucial insights into the processes driving the formation and evolution of the stars of which they are composed. Star clusters are generally classified as one of two types: open clusters and globular clusters.

Open clusters are relatively small groups of up to a few thousand stars held together by gravitational attraction. They are typically found in the disk of the Milky Way and are characterised by their relative youth, with ages ranging from a few million to a few billion years. Open clusters are key probes of the structure and history of the Galactic disk (Cantat-Gaudin et al., 2019). They serve as excellent laboratories for studying stellar evolution, and are excellent tracers used to follow the stellar metallicity gradient of the Milky Way (Yong et al., 2005). Of particular relevance to this thesis, the study of the kinematics of OCs and reconstructions of their individual orbits (Cantat-Gaudin et al., 2016) help us understand radial migration (Anders et al., 2017) and the internal processes of dynamical heating (Quillen et al., 2018).

Globular clusters, in contrast, are much larger and more tightly bound. They consist of hundreds of thousands to millions of stars, forming roughly spherical collections with diameters ranging from 10 to 200 light-years. Globular clusters are primarily situated in the halo of the Milky Way, far from regions of active star formation (Gratton et al., 2012). They are significantly older, with ages spanning 10 to 13 billion years. These clusters provide insights into the structure and evolution of the Milky Way and offer glimpses into the early universe. Their properties, such as ages, chemical compositions, and kinematics, help us understand galactic formation processes.

The formation and evolution of open and globular clusters differ due to their distinct characteristics. Open clusters are believed to form from the collapse of giant molecular clouds, triggering the formation of multiple stars. Over time, tidal forces from the galaxy cause these clusters to disperse. Globular clusters, on the other hand, have older stars and are more densely packed, and their formation scenarios are less well understood. Their gravitational interactions can lead to ejections and stellar collisions, the latter giving rise to peculiar objects like blue stragglers.

In addition to their individual significance, star clusters offer valuable information about the formation and evolution of galaxies. The synchronised formation of stars within a cluster allows us to study the interstellar medium and gain insights into chemical enrichment histories and star formation mechanisms. Furthermore, the distribution of different types of clusters within galaxies can help investigate different Galactic components, such as the halo and the disk.

The parameterisation of clusters is an ongoing field of research, from studies of globular clusters looking at their chemical compositions (e.g., Smith et al., 2000) to an overall census of open clusters that can be found in Gaia (Cantat-Gaudin et al., 2019). However, the internal rotation properties, especially for open clusters, are not well understood.

The identification of internal rotation within a cluster, whether it be a globular or open cluster can be done in a few ways :

- Proper Motion and Radial Velocity measurements: By measuring the proper motions of stars in a cluster, we can derive their tangential velocities. This information, combined with the radial velocities, can provide a complete three-dimensional picture of stellar motion (e.g., van Leeuwen et al., 2000; Vasiliev, 2019).

- Spatial Distribution Analysis: The spatial distribution of stars within a cluster can provide insights into its rotation. Analysing the positions of stars relative to the cluster center can reveal any systematic patterns or asymmetries that may indicate rotation. Methods such as the center-of-mass techniques are often employed to estimate the cluster center and assess its rotation (Lützgendorf et al., 2012).

- Modelling and Simulations: Complex dynamical models and simulations can be employed to study the internal rotation of star clusters. These models consider the gravitational interactions between stars, including the effects of rotation and other relevant factors. (e.g., Bekki, 2010; Combes et al., 1999) By comparing simulated models with observational data, we can infer the rotation properties of star clusters.

In Chapter 5, I use the Gaia proper motions, radial velocities and, tests on simulated data to estimate the rotational parameters of star clusters by considering each cluster as a 3D solid body. I then apply the method to a list of open clusters with unknown rotational parameters.

# 3 The GALAH Survey: A New Sample of Extremely Metal-Poor Stars Using A Machine Learning Classification Algorithm

This Chapter is based upon work done in Hughes et al. (2022b), and is the first insight into applying a particular statistical technique on a large stellar spectroscopic survey to extract interesting science in an efficient way. The focus on this work was classification methods, and the ability to identify the proverbial "needle in a haystack". I was the lead author and responsible for redoing the GALAH data reduction of the spectra, and the subsequent analysis. The co-authors (L. R. Spitler, D. B. Zucker, T. Nordlander, J. Simpson, G. S. Da Costa, YS. Ting, C. Li) provided either helpful feedback on the data modelling and manuscript structure or additional data to make this analysis rigorous. The remaining authors are the GALAH builders, without whom this work would not have been possible. Follow-up work on the identified candidates was done in Da Costa et al. (2023), but that work is not discussed in this Chapter as I only provided the candidates.

## Brief Summary

Extremely Metal-Poor (EMP) stars provide a valuable probe of early chemical enrichment in the Milky Way. Here we leverage a large sample of $\sim$ 600,000 high-resolution stellar spectra from the GALAH survey plus a machine learning algorithm to find 54 candidates with estimated [Fe/H] $\leqslant$ -3.0, 6 of which have [Fe/H] $\leqslant$ -3.5. Our sample includes $\sim$ 20% main sequence EMP candidates, unusually high for EMP star surveys. We find the magnitude-limited metallicity distribution function of our sample is consistent with previous work that used more complex selection criteria. The method we present has significant potential for application to the next generation of massive stellar spectroscopic surveys, which will expand the available spectroscopic data well into the millions of stars.

## 3.1 Introduction

Extremely metal-poor stars (EMP, [Fe/H] $\leqslant -3.0$) are interesting stellar objects, as they provide a window into the history of the early Universe. The EMP stars that exist today formed in environments with much less chemical enrichment than is typically found in the interstellar medium today. As a result, they record the chemical yields produced by the first generations of stars after the Big Bang, and thereby provide crucial clues to the properties of early supernovae and their progenitors. Hence EMP stars are essentially a log of some of the earliest events in the Galaxy's chemical evolution.

The significance of metal-poor stars has been be reviewed extensively (e.g., Beers and Christlieb, 2005; Frebel and Norris, 2015). However, to date very few EMP stars have been discovered, especially considering the fact that entire observational surveys have been dedicated to that aim. Querying the high-resolution SAGA database (Suda et al., 2008) we see only $\sim 1000$ stars with [Fe/H] $< -3$, $\sim 200$ with [Fe/H] $< -3.5$ and $\sim 50$ with [Fe/H] $< -4$ have been found. As noted above, EMP stars offer a unique window into the chemical enrichment of the Milky Way as it was forming, yet their relative rarity constrains our ability to probe those early times. Hence expanding the known sample is of critical importance for creating a comprehensive picture of the processes dominating the life cycle of stars and the interstellar medium in the early Universe.

With the development of highly multiplexed astronomical spectrographs, many current stellar surveys are producing spectroscopic datasets that are too large for traditional analysis (e.g., RAVE Steinmetz et al. 2020; APOGEE Ahumada et al. 2020; GALAH Buder et al. 2020). The next generation of surveys – including WEAVE (Dalton et al., 2014), 4MOST (de Jong et al., 2019) and SDSS V's MWM (Kollmeier et al., 2017) – will expand the available spectroscopic data well into the millions of stars. Thanks to recent improvements in computing and statistical methods however, we are now able to develop more refined tools and processes to sift through these huge datasets in order to reliably identify rare and interesting science targets, such as EMP stars. This paper seeks to identify EMP stars in the GALAH spectroscopic survey, and develops a novel machine-learning approach that could be used to identify other scarce objects in large astronomical datasets.

The machine learning method adopted in this paper is t-SNE (Maaten and Hinton, 2008), a dimensionality reduction technique. This method has been successfully applied to astronomical data for identification of sub-structure within a parameter space and the classification of stellar objects; in particular, t-SNE has been applied in the stellar and chemical abundance space, to identify membership in stellar-clusters and streams (e.g., Anders et al., 2018; Kos et al., 2018). Matijevič et al. (2017) and Jofré et al. (2017) applied t-SNE to RAVE survey spectra to identify very metal-poor stars and stellar twins, respectively, and Traven et al. (2020) used t-SNE on GALAH spectra to find FGK-type binary stars. The application of t-SNE in these cases followed a top-down approach, i.e., running t-SNE on the dataset in question and then exploring the resulting space, which can be inefficient in identifying objects of particular interest. In this paper however, we show an alternative approach, following the process described in Hughes (2017) and similar to Hawkins et al. (2021), in which we flag objects we are interested in prior to running t-SNE, and then see where they appear on the projected t-SNE space. This works because any unclassified star falling near the flagged stars can be considered a potential candidate because it will have similar spectral features.

In this paper, data from the GALactic Archaeology with HERMES spectroscopic survey (GALAH) are analysed to show how machine learning methods, applied to spectra, can be used to identify extremely metal-poor stars. The paper is organised as follows: the GALAH

| s_object_ID | Object Name | $T_{eff}^{L}$ | log $g^{L}$ | $[Fe/H]^{L}$ | $T_{eff}^{Est}$ | log $g^{Est}$ | $[Fe/H]^{Est}$ | $T_{eff}^{DR3}$ | log $g^{DR3}$ | $[Fe/H]^{DR3}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| 140209005201151* | HD 122563 Kirby et al. (2010) | 4367 | 0.60 | -3.15 | 5000 | 0.50 | -2.80 | 4616 | 1.46 | -2.51 |
| 140307003101095 | 2MASS J13274506-4732201 Simpson et al. (2012) | 4661 | 1.50 | -2.70 | 5000 | 0.50 | -2.10 | 4616 | 1.36 | -1.87 |
| 140412001201388 | HE 1207-3108 Yong et al. (2012) | 5294 | 2.85 | -2.70 | 5300 | 0.50 | -3.10 | 5404 | 2.97 | -2.56 |
| 140810004701232 | UCAC4 157-208544 Placco et al. (2019) | 4651 | 1.24 | -2.52 | 5000 | 0.50 | -2.10 | 4539 | 1.46 | -1.87 |
| 150409002601337 | TYC 4934-700-1 Sakari et al. (2018) | 4614 | 1.03 | -2.52 | 5100 | 0.50 | -2.60 | 4687 | 1.45 | -2.25 |
| 150718004401358* | BPS CS 22892-0052 McWilliam et al. (1995) | 4760 | 1.30 | -3.10 | 5200 | 3.75 | -3.50 | 5657 | 2.33 | -2.19 |
| 151008003501121* | HE 0124-0119 Li et al. (2015) | 4330 | 0.10 | -3.57 | 4000 | 5.00 | -4.25 | 4367 | 1.63 | -3.38 |
| 160401003901201 | DENIS J133748.8-082617 Sakari et al. (2018) | 4265 | 0.25 | -2.62 | 4800 | 0.50 | -2.40 | 4289 | 0.73 | -2.44 |
| 160403004201044 | 2MASS J13273676-1710384 Placco et al. (2019) | 5223 | 1.67 | -2.55 | 5200 | 0.75 | -2.60 | 5127 | 2.12 | -2.17 |
| 160424004701042 | UCAC4 053-017641 Placco et al. (2019) | 4832 | 1.61 | -3.41 | 5000 | 0.50 | -2.90 | 4795 | 2.05 | -2.51 |
| 160519002601142 | UCAC4 226-057537 Placco et al. (2019) | 4619 | 1.07 | -2.54 | 4900 | 0.50 | -2.30 | 4526 | 1.40 | -2.05 |
| 160813003601164* | 2MASS J21260896-0316587 Hollek et al. (2011) | 4725 | 1.15 | -3.22 | 5100 | 0.50 | -3.10 | 5056 | 2.15 | -2.71 |
| 161009003801062 | UCAC4 464-129364 Mardini et al. (2019) | 4945 | 1.53 | -2.52 | 5000 | 0.50 | -2.70 | 4743 | 2.10 | -2.36 |
| 161104002301201 | 2MASS J22045836+0401321 Spite et al. (2018) | 4700 | 1.20 | -2.90 | 5000 | 0.50 | -2.80 | 4632 | 1.82 | -2.57 |
| 161118004701028 | SMSS J051008.62-372019.8 Jacobson et al. (2015) | 5170 | 2.40 | -3.20 | 5300 | 0.50 | -3.20 | 5342 | 3.31 | -2.68 |
| 170601003101219 | 2MASS J14175995-2415463 Schlaufman and Casey (2014) | 4914 | 1.45 | -2.40 | 5000 | 0.50 | -2.60 | 4724 | 1.47 | -2.21 |
| 170615004401258* | 2MASS J18082002-5104378 Meléndez, Jorge et al. (2016) | 5440 | 3.00 | -4.07 | 5500 | 0.50 | -4.25 | 5741 | 3.48 | ··· |
| 170805005101110 | HE 0048-6408 Placco et al. (2014a) | 4378 | 0.15 | -3.75 | 4800 | 0.50 | -3.80 | 4221 | 1.18 | -3.83 |
| 170904000601186 | 2MASS J21303218-4616247 Masseron et al. (2010) | 4100 | -0.30 | -3.39 | 5000 | 0.50 | -3.10 | 3987 | 0.89 | -3.76 |
| 170906004601038 | HE 0105-6141 Barklem et al. (2005) | 5218 | 2.83 | -2.55 | 5300 | 0.75 | -2.60 | 5190 | 2.87 | -2.36 |
| 170906004601108 | BPS CS 22953-0003 Spite et al. (2018) | 5100 | 2.30 | -2.80 | 5100 | 0.50 | -3.10 | 5044 | 2.36 | -2.73 |
| 171001003401116 | HE 0433-1008 Beers et al. (2017) | 4708 | 1.31 | -2.62 | 4900 | 0.50 | -2.70 | 4423 | 1.54 | -2.77 |
| 171205002101255* | SMSS J031300.36-670839.3 [Keller]Nordlander et al. (2017b) | 5150 | 2.20 | $< -6.53$ | 5000 | ··· | ··· | ··· | ··· | ··· |

Table 3.1: Extremely metal-poor stars from the literature found in GALAH DR3, designated with the GALAH identifier s_object_ID. The different superscripts in the stellar parameters reflect the source of the parameters: **L** indicates values from the literature, **Est** shows the results of our parameter estimation method and **DR3** represents the output of the GALAH DR3 pipeline. The stars marked with an asterisk are the "known" EMP stars used in the application of the methodology outlined in this paper; the remainder were subsequently identified as a verification of the method.

survey and the data are described in Section 3.2, Section 3.3 outlines the methodology used to find EMP stars, the results and candidates from applying the methodology to the data are shown in Section 3.4, in Section 3.5 we discuss those results and we summarise our conclusions in Section 3.6.

## 3.2 Data

The following section outlines the datasets used in this paper. Section 3.2.1 and Section 3.2.2 briefly introduces the GALAH Survey and discusses how the stars used in this analysis were selected. Section 3.2.3 describes how we label known EMP stars and other classified stars within GALAH. Lastly, Section 3.2.4 details how the synthetic templates that will be used in deriving stellar parameters are constructed.

### 3.2.1 GALAH

The Galactic Archaeology with HERMES (GALAH) survey is a high resolution spectroscopic survey of the Milky Way which uses the High Efficiency and Resolution Multi-Element Spectrograph (HERMES) on the 3.9m Anglo-Australian Telescope. By its finish GALAH will obtain $\sim 1,000,000$ high resolution spectra ($R \sim 28,000$) of stars at Galactic latitudes of

$|b| > 10\circ$ and declinations $-80\circ < \text{Dec} < +10\circ$, across the four discrete spectral arms of HERMES: 4713-4903Å (blue channel), 5648-5873Å (green channel), 6478-6737Å (red channel), and 7585-7887Å (IR channel). The spectrograph can typically achieve a signal to noise ratio (SNR) of $\sim$ 100 per resolution element at magnitude V $\sim$ 14 in the red arm during a 1-hour exposure (De Silva et al., 2015). By measuring radial velocities, stellar parameters and abundances for as many as 30 elements, the goal of the GALAH survey is to produce a comprehensive view of the formation and chemodynamical evolution of the Milky Way.

This paper uses GALAH data release 3 (DR3; Buder et al., 2021), in which all observed stellar spectra were extracted as one dimensional spectra, continuum normalised and radial velocity-corrected to the barycentric reference frame. This data release includes spectra of $\sim 600,000$ unique stars. The GALAH data reduction pipeline is described in Kos et al. (2017); for the data analysis pipeline, DR3 stellar parameters and abundances were estimated via the spectrum synthesis code Spectroscopy Made Easy (SME; Valenti and Piskunov, 1996; Piskunov and Valenti, 2017) using theoretical 1D hydrostatic models taken from the Marcs grid and 1D non-LTE grids as described in Amarsi et al. (2020) for 11 elements (Li, C, O, Na, Mg, Al, Si, K, Ca, Mn, Fe and Ba).

### 3.2.2 Sample Selection

We selected a subset of the $\sim 600,000$ DR3 stellar spectra tailored to the needs of our analysis. We considered spectra taken between November 2013 and February 2018 and limited ourselves to one spectrum per star, thus avoiding problems encountered with stacked spectra in DR3 (Sec 6.2 in Buder et al., 2021); in addition, we only used spectra that passed the reduction pipeline quality control (i.e., `red_flag==0`). We did not include poor-quality spectra with low signal-to-noise in the green channel (S/N < 35 per resolution element) and stars with GAIA $G_{BP} - G_{RP} < 0.6$ (Gaia Collaboration et al., 2018), as these typically represent hot stars that may appear extremely metal-poor but are not.

### 3.2.3 Literature stellar labels

To be able to classify stellar spectra using semi-supervised machine learning, which is the combination of labelled and unlabelled data, we first have to assign a label to a proportion of the spectra. In our case the label is the stellar classification of the spectra.

The sample of stars with a known stellar classification was compiled by cross matching the stellar classification labels defined by Traven et al. (2017) to the GALAH survey data using `s_object_ID`, a unique star identifier in DR3. The 5 stellar classes chosen based on SIMBAD are:

1. Binary stars
2. Cool metal-poor giants
3. H$\alpha$/H$\beta$ emission
4. Hot stars
5. Stars with molecular absorption bands

These stellar classifications were determined manually by Traven et al. (2017) after having run t-SNE in combination with Density-Based Spatial Clustering of Applications with Noise

(SCAN or DBSCAN) (Ester et al., 1996) on GALAH DR1 (Martell et al., 2017). We therefore do not treat these labels as definite, as they only represent potential identifications. In addition, we defined two other categories of labelled objects: Extremely Metal-Poor stars (`ExtMetalPoor`), and one specific EMP star, the Keller star (`Keller`), which are described immediately below.

A sample of 8 EMP stars was manually identified by cross-matching stars observed in GALAH with stars in SIMBAD determined to have [Fe/H] $\lesssim -3$; this latter sample yielded a table with 538 unique entries. A 10 arcsecond positional cross-match of this table against the GALAH data resulted in a list of 7 possible EMP stars based upon literature measurements compiled by SIMBAD. However, two of the stars failed to pass the GALAH quality checks because of poor spectrum normalisation, and an additional star did not satisfy the metallicity requirement of [Fe/H] $\sim -3$ as a detailed examination of literature measurements showed that it most likely has [Fe/H] $\approx -2$. These spectra are given the label `ExtMetalPoor`.

The star SMSS J031300.36-670839.3 (Keller et al., 2014) was not a GALAH target, but was observed with HERMES on 5 December 2017 and then processed by the GALAH pipeline. The spectrum is almost featureless aside from hydrogen absorption, which is not surprising given its initial upper limit estimate of [Fe/H] $\approx -7.1$ (Keller et al., 2014), subsequently updated to [Fe/H] $< -6.53$ by Nordlander et al. (2017a). This star is given the classification of `Keller`.

The sample of known EMP stars described above – the five `ExtMetalPoor` stars, and the `Keller` star – are presented in Table 3.1, in which they are highlighted by a *.

The stars that do not have a stellar classification label after cross-matching by `s_object_ID` are given the classification `unlabelled`. These observations are combined with the labelled dataset to define our full GALAH dataset.

To summarise the GALAH dataset used in this paper, after applying our sample selection criteria for signal-to-noise and $G_{BP} - G_{RP}$ colour, we have a dataset with 9058 labelled and 590514 unlabelled stars. Thus the total number of stars analysed in this work is 599855.

### 3.2.4 Synthetic Templates

To determine the stellar parameters (see Section 3.3.2) of any potential EMP candidate we fit the observed spectra with synthetic spectral templates with known stellar parameters. Following Nordlander et al. (2019), the 6045 synthetic spectra templates were produced with Plez (2012) in 1D LTE using standard MARCS model atmospheres (Gustafsson et al., 2008). We used $v_{mic}$ = 1 km s$^{-1}$ and plane-parallel model atmospheres for models with $\log g >$ 3.5, and $v_{mic}$ = 2 km s$^{-1}$ and spherical geometry for models with $\log g <= 3.5$. We fixed [$\alpha$/H]= 0.4 and initially considered $T_{\rm eff}$ = 4000K$-$8000K in steps of 500K, $\log g$ = 0.0 $-$ 5.0 in steps of 0.5 and [Fe/H] = $-7.0 - 0.0$ in steps of 0.5 dex, and a limited range of carbon enhancements, [C/H]= $0.0, 0.5, 1.0$. All synthetic spectra were broadened by $v_{broaden}$ = 10 km s$^{-1}$ to represent the instrumental resolution, from an initial resolution of 1 km s$^{-1}$.

A finer grid was subsequently generated with $T_{\rm eff}$ = 4000K$-$7500K in steps of 100K, $\log g$ = 0.0 $-$ 5.0 in steps of 0.25 dex, and varying step sizes in metallicity for different ranges of [Fe/H]:
$-7.00$ to $-5.50$ in steps of 0.5 dex,
$-5.00$ to $-4.25$ in steps of 0.25 dex,
$-4.00$ to $-2.10$ in steps of 0.1 dex, and
$-2.00$ to $-1.00$ in steps of 0.25 dex,
to give further detail on the range of [Fe/H] values that we can estimate. When applying the

finer grid we set the carbon abundance [C/Fe] = 0, as informed by our simulation analysis in Appendix A.1, and also because the wavelength ranges in GALAH cannot be used to meaningfully constrain the carbon abundance in EMP stars.

## 3.3 Method

Here we discuss a methodology that can be used to identify EMP stars within GALAH data, but also potentially adapted to find any stellar type within a given spectroscopic dataset. The methodology is a hybrid of machine learning and a more traditional model-fitting approach. Section 3.3.1 outlines how to identify similar stars using a branch of machine learning known as dimensionality reduction, with a focus on targeting EMP stars. Once candidate EMP stars have been found, Section 3.3.2 describes the estimation of their stellar parameters.

### 3.3.1 Identifying Similar Stars

Identification of EMP stars in spectroscopic surveys typically involves the fitting of select regions of an observed spectrum with a synthetic spectrum, employing a metric to define their similarity (usually $\chi^2$)

EMP stars have, however, a relatively featureless spectrum, where distinguishing the difference between a real spectral feature and the inherent noise is challenging. Similarly, the *synthetic* spectrum of a metal-poor star is close to featureless, but lacks the noise that is present in an observed spectrum. Hence when applying a $\chi^2$-fitting method which entails comparing an observed spectrum with a synthetic spectrum, we expect stars that aren't extremely metal-poor to be identified as such (and vice-versa) due to model systematics. Ideally we would like a method that can 1) be independent of synthetic templates, 2) self-identify important features of a spectrum and 3) categorise similar spectra.

To be able to create a method as described above is challenging for metal-poor stars. If, however, we could visualise the similarity of objects within a dataset, and group them visually, then we could reduce the search space for finding objects of interest, prior to running a $\chi^2$-fitting method. Furthermore, if we had a sample of the objects we were trying to identify, we could flag these before running a visualisation technique and see which groups they are clustered in, and hypothesise that the surrounding group must contain similar observations. By approaching finding similar objects this way, we remove the necessity for synthetic template comparison at the identification phase of the process, resulting in a purer candidate sample.

Dimensionality reduction techniques, a branch of machine learning, are a standard way of extracting important features from large datasets. Dimensionality reduction is the process of representing a high-dimensional data set $X = x_1, x_2, \ldots, x_N$, by a set $Y$ of vectors $y_i$ in two or three dimensions, and then placing similar observations in close proximity in the new parameter space, $y_i$, while keeping dissimilar observations at larger distances. The resulting reduced parameter space, $Y$, may then be visualised to determine similar and dissimilar input data. The most common dimensionality reduction techniques used in astronomy are principal component analysis (PCA) and multidimensional scaling (MDS). Principal component analysis has been used by Yip et al. (2004) to classify quasars using SDSS spectral data; Connolly and Szalay (1999) demonstrated that PCA can be used to build galaxy SED's from data that may be noisy or incomplete; and Re Fiorentin et al. (2007) showed that PCA can be used to estimate stellar atmospheric parameters with SDSS/SEGUE spectra and Ting et al.

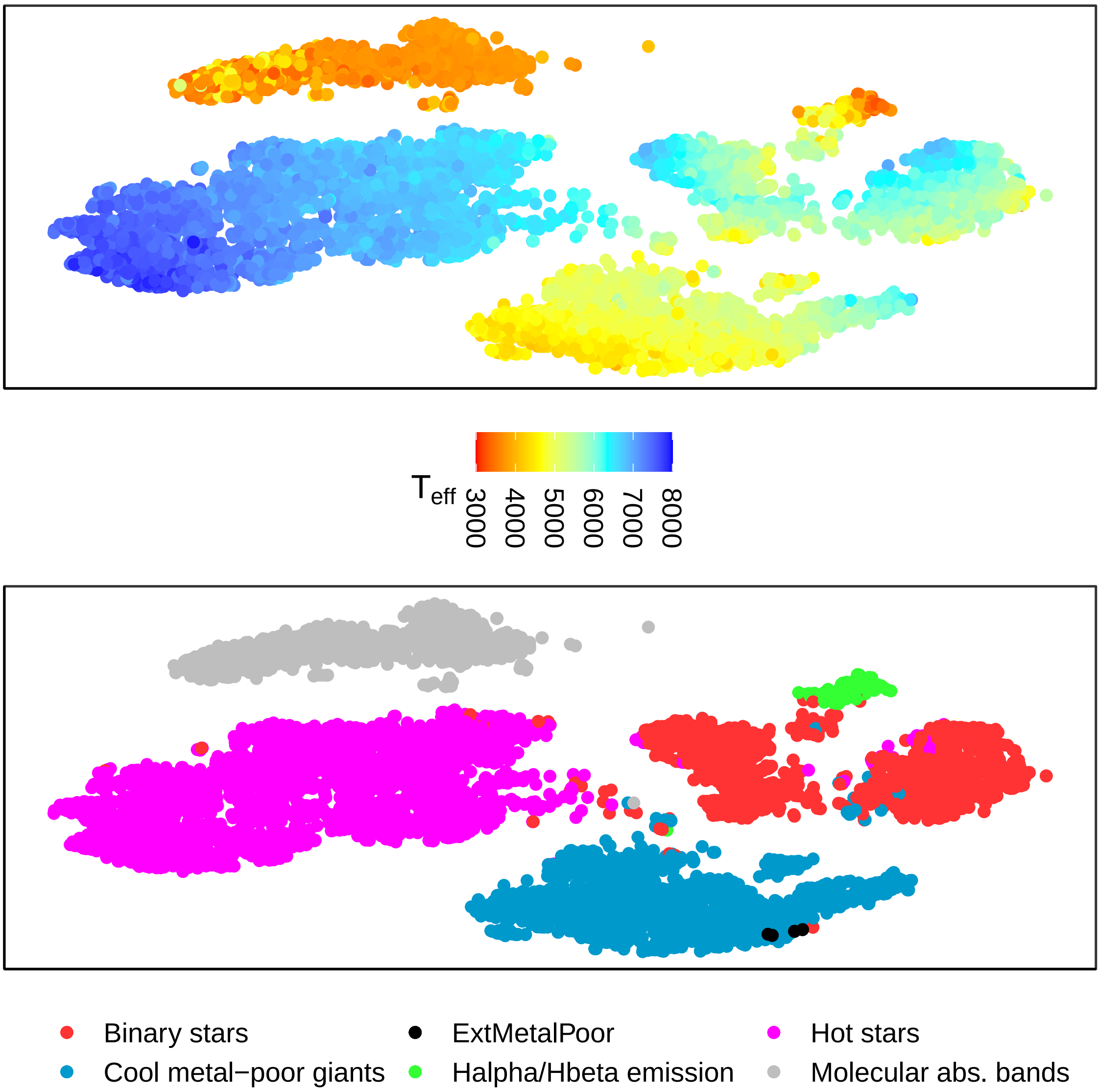


Figure 3.1: Illustrative t-SNE maps constructed only using the labelled portion our dataset. Top panel is coloured by effective temperature from GALAH DR3. Bottom panel the t-SNE map is coloured by stellar classification labels. Each point represents a star which has had its spectral information collapsed into two points in a 2-dimension parameter space produced by t-SNE; the axes are in arbitrary units reflecting only the dynamic range of the data in this space. Comparing the two panels shows that t-SNE is sensitive to effective temperature and the stellar classification. Note that neither spectroscopic temperature or classifications were included in the *input* to t-SNE.

(2012) used PCA to explore the stellar chemical abundance space. The frequent use of PCA underscores the importance of dimensionality reduction in the area of classification.

A significant weakness in PCA, however, is that it is intrinsically linear. PCA does not consider the structure of the manifold; there may be data points that form a nonlinear manifold, which PCA will not be able to deconstruct. In addition, while dimensionality reduction techniques have been used in astronomy before, they generally were applied to smaller datasets. The effective parameter space for the GALAH data is 4 channels × 4096 pixels × 65,536 flux levels × 600,000 stars $\simeq 6 \times 10^{14}$ values; with a dataset of this magnitude, traditional techniques face a computational challenge.

### t-SNE

Like other dimensionality-reduction techniques, t-SNE (Maaten and Hinton, 2008) can be used to visualise how similar points are within a dataset. t-SNE assesses the similarity of features in the higher-dimensional space by using the Euclidean distance metric (alternative metrics may be applied). A similarity matrix of probabilities, representing the higher-dimensional space, is calculated by converting these Euclidean distances using a standard Normal distribution. The feature space is then reduced to 2 or 3 dimensions, and the process above is repeated for this lower dimensional space; however, the t-distribution is used instead of a standard Normal distribution to construct the similarity matrix. To finally determine the lower dimensional representation of distances within our dataset, the Kulback-Leibler (KL) divergence between the two joint probability distributions is minimised using gradient descent. We chose the Barnes-Hut gradient descent version of t-SNE, implemented in the R package Rtsne[1] by Krijthe (2015), as it substantially speeds up t-SNE and allows t-SNE to be applied to much larger datasets that would be computationally intractable with the original t-SNE algorithm. t-SNE  unlike techniques such as PCA, is able to produce more visually compelling clusters because t-SNE 's non-linearity enables it to maintain the trade-off between local and global similarities between points. This makes t-SNE well suited to the purpose of finding and visualising the distribution of similar spectra in a large dataset. Refer to Traven et al. (2017) for a more detailed description of the t-SNE algorithm.

To illustrate the effectiveness of t-SNE at classifying stars using only their spectra, we consider our defined labelled dataset of 9058 classified stars. For this application, only the spectral data for the labelled stars was passed in to t-SNE. The labels and additional stellar parameter information were not used. t-SNE 's input is a set of 9058 high-dimensional objects $x_i....x_N$, where each object is described by 12288 wavelength values (for this analysis we ignore the infrared channel), representing a single star. Top panel of Figure 3.1 shows the t-SNE map coloured by effective temperature ([$\mathrm{T_{eff}}$]). Bottom panel of Figure 3.1 is the same map coloured by the classification labels in Traven et al. (2017). In both panels, the cluster of hot stars is clearly distinguishable by both temperature and label, highlighting that by applying t-SNE to only spectra, we can visualise sensitivity in both the stellar parameter and stellar classification space.

### Determining which Wavelength Regions to Fit

In searching for EMP stars it is important to understand the significant spectral features that are key indicators of extremely low metallicity. When dealing with low and medium-resolution spectra, traditionally the infrared calcium triplet or ultraviolet calcium H and K

[1] `https://github.com/jkrijthe/Rtsne`

lines have been used as standard indicators of low metallicity. These lines are, however, outside of the GALAH wavelength range. Therefore the first step in applying our method to the entire GALAH dataset is to determine which metal lines within the GALAH wavelength range would be most useful for identifying EMP stars.

The optimal wavelength ranges were selected by determining the lower limit for [Fe/H] using the spectral templates. This was achieved by running a series of $\chi^2$ fitting simulations using different elemental line combinations. The simulation which resulted in the highest level of certainty for the lowest [Fe/H] was selected. The optimal restricted wavelength ranges used are the regions around H$\alpha$, H$\beta$, 4867-4872Å and 4887-4892Å ; the latter two ranges contain the strongest Fe I lines in the blue channel (4875.88Å, 4890.76Å and 4891.49Å). Additionally we found that the OI triplet (7771.94Å- 7775.39Å) in the IR channel was a useful discriminant for removing hot stars that were contaminating the sample, and thus this range was also included. In Figure 3.2 we show that, using spectra in the GALAH wavelength range, we can say with reasonable confidence that a star has a metallicity as low as [Fe/H] $\sim -3.5$ (or potentially below that value); see Appendix A.1 for further details.

By applying a method like t-SNE the idea is to reduce any bias that may arise in choosing which wavelength ranges to consider, as the technique may be able to better identify significant wavelength ranges not considered. To use the entire wavelength range, however, is a) computationally unfeasible and b) given the relatively featureless nature of EMP star spectra, can introduce noise that may skew the final results. We were thus unable to avoid having to select which wavelength ranges to input into t-SNE.

## 3.3.2 Estimating Stellar Parameters for EMP stars

To confirm the identification of any EMP candidate found using the t-SNE methodology outlined in Section 3.3.1, we require an estimation of their basic stellar parameters, $T_{\rm eff}$, $\log g$ and [Fe/H]. The GALAH DR3 pipeline provides measurements of these stellar parameters for many of our candidates; however the DR3 pipeline is not tailored towards metal-poor stars with weak metal-lines. We developed a simple iterative procedure to estimate each stellar parameter for a candidate EMP star, which is described below (see Section 3.4.2 for the application).

### Effective Temperature, Surface Gravity and Metallicity

To estimate $T_{\rm eff}$ , $\log g$ and [Fe/H], we apply a $\chi^2$-minimisation routine between the observed spectra and the synthetic templates defined in Section 3.2.4 using only the H$\alpha$ and H$\beta$ regions and a select few metallicity lines around 4870Å and 4890Å. We fit the lines simultaneously to account for the degeneracies that arise between the stellar parameters, and select the template corresponding to the minimum $\chi^2$.

The upper half of Figure 3.3 shows the fitting of the H$\beta$ region to 3 synthetic templates for a given star, with the optimal fit of $T_{\rm eff}$ equal to 5250, shown in red, and the best fit of $\log g$ as 2.5. The bottom panels of Figure 3.3 show the two line regions considered in the fitting of metallicity, which suggest this observed star may have a metallicity in the range $-3.5 <$ [Fe/H] $< -3.0$.

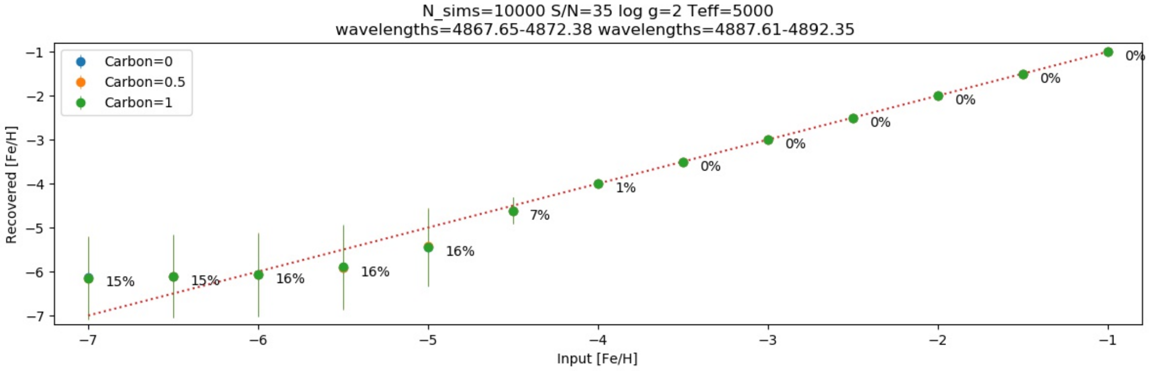


Figure 3.2: The output of a simulation fitting synthetic spectra with templates, showing that, given these stellar and observational parameters ($\log g = 2$, $T_{\rm eff} = 5000$K, S/N $= 35$) we can reliably estimate metallicities down to [Fe/H] $\lesssim -3.5$ with GALAH data. Percentages indicate fractional uncertainty from scatter in the recovered metallicities rounded to the nearest percentage. Due to the relative insensitivity to carbon in the GALAH wavelength ranges, the blue and orange points are masked by the green points. See Appendix A.1 for further details.

## 3.4 Results

The following section describes the results of applying the outlined methodology to the full GALAH dataset inclusive of unclassified stars, as defined in Section 3.2.2. Section 3.4.1 applies our hybrid t-SNE methodology, Section 3.4.2 and Section 3.4.3 estimates the stellar parameters and applies some further analysis on the candidate sample.

### 3.4.1 Applying the t-SNE methodology

Applying the methodology to find EMP stars described in Section 3.3.1, we consider the entire GALAH sample, subsetted by the optimal wavelength regions as determined in 3.3.1 and with the EMP stars and the Keller star flagged. We will use the additional classification labels defined in Section 3.2.3, to flag other structures in the t-SNE plane.

The t-SNE method was calculated with the perplexity set to 40, the number of iterations set to 2000 and the other hyperparameters (see Wattenberg et al. (2016)) left to their default values. The processing was run on an Ubuntu server, with 344GB of RAM and an Intel Xeon CPU E5-2695 v3 @ 2.30 GHZ with 30 threads.

The resulting map is shown in Figure 3.4. A separate "island" containing all 5 known EMP stars and the Keller star is located in the top left of the map. We infer that the unlabelled stars surrounding the known Keller and EMP stars form a potential metal-poor cluster on the map. This cluster is then extracted and passed into our stellar parameter fitting routine described in Section 3.3.2, reducing the search space to fit stellar parameters of potential EMP stars from 600,000 to approximately 2500. A zoomed in image of this cluster is shown in Figure 3.5.

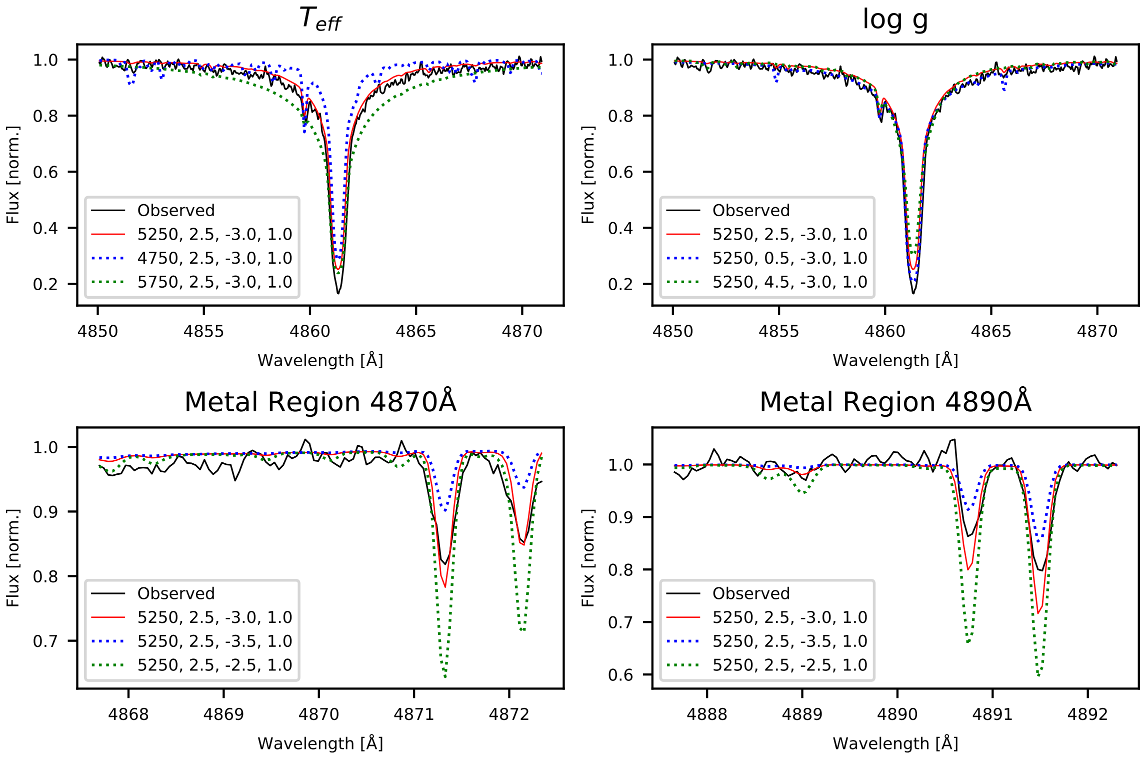


Figure 3.3: Fitting the observed H$\beta$ region for a given star with the wider synthetic template grid as a test of the fitting of the stellar parameters, $T_{\text{eff}}$, $\log g$ and [Fe/H]. It is clear from the upper panels that this star is best fit by a $T_{\text{eff}}$ of 5250 and a $\log g$ of 2.5. The bottom panels represent the line regions considered in fitting [Fe/H], and show that this star likely has a metallicity [Fe/H] between -3.0 and -3.5. The parameters of the synthetic templates as given in the panels are $T_{\text{eff}}$, $\log g$, [Fe/H] and [C/Fe] (the assumed [C/Fe] may be ignored for these fits).

### 3.4.2 Stellar Parameter Estimation for the EMP stars Cluster

Taking the hypothesised metal-poor only island, we estimate the stellar parameters for each star in the island using our simple stellar parameter fitting routine described in Section 3.3.2.

The estimated $T_{\text{eff}}$ and $\log g$ values for our metal-poor island are shown in the upper two panels of Figure 3.6. Here we see a similar distribution of $T_{\text{eff}}$ and $\log g$, with cooler giant-type stars on the left, going to hotter, higher $\log g$ stars to the right of the island. [Fe/H] is shown in the bottom panel of Figure 3.6 and displays a gradient of metallicity, higher to lower, from the bottom to the top edge of the island. The previously defined extremely metal-poor coast (as seen in Figure 3.5) is evident.

Before identifying and analysing EMP candidates in our cluster, we note that our stellar parameter fitting routine is relatively simple and is only used as a guide. The method was necessary, as we see a significant scatter with respect to DR3 pipeline-derived metallicities, as well a systematic tendency toward higher measured metallicities in DR3. This may be attributed to the GALAH analysis pipeline being optimised for thin and thick disk stars, with typical metallicities $[\text{Fe/H}] \geqslant -2$. Moreover, a comparison of our metallicity estimates for the stars with both the (admittedly heterogeneous) literature metallicities and GALAH DR3

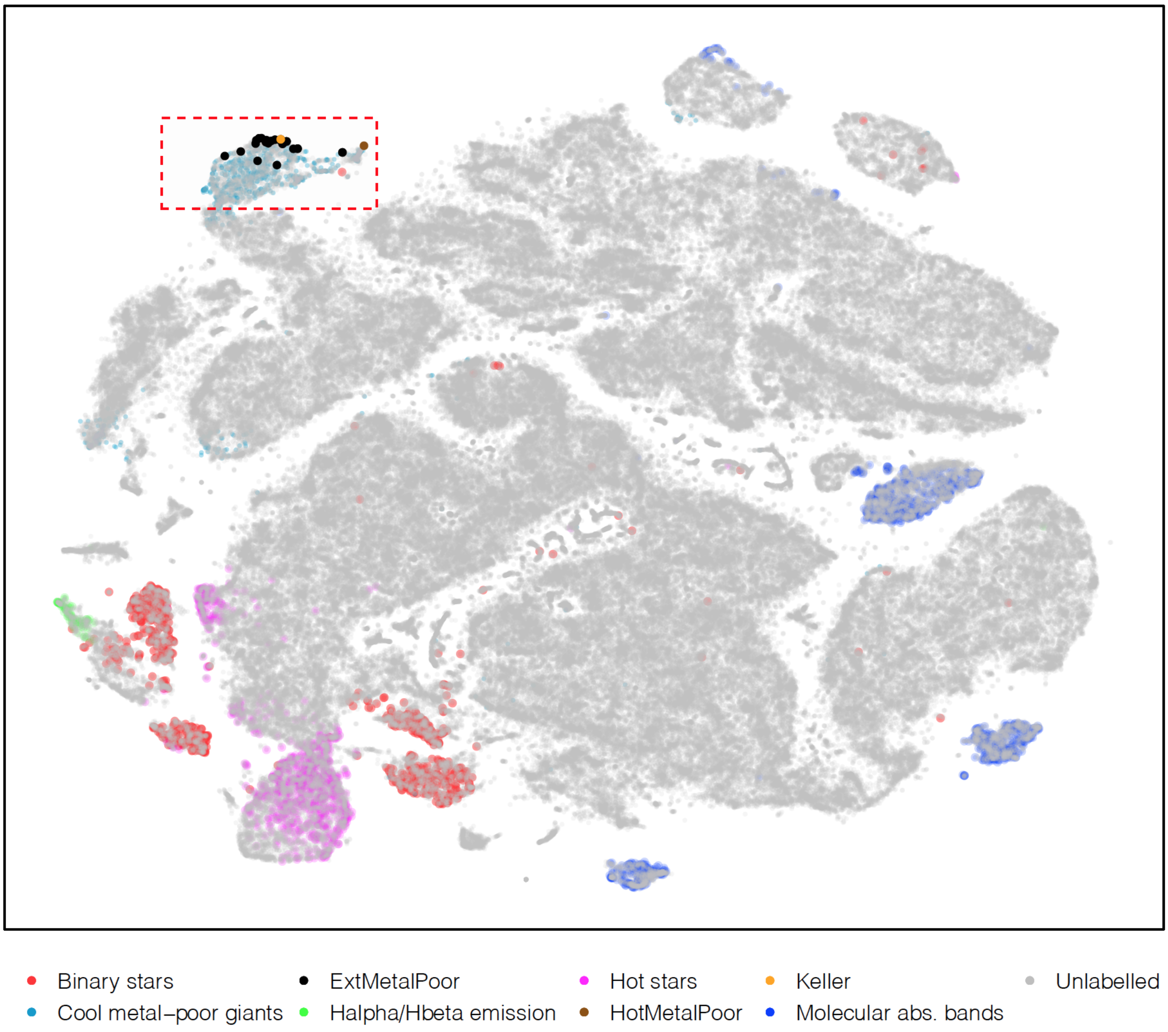


Figure 3.4: t-SNE map with the unknown (unlabelled) stars plotted in grey and the known extremely metal-poor stars – corresponding to the stars shown in Table 3.1 – overlaid in black, brown and orange. A region containing all 5 (*) stars and the additional known metal-poor stars is located to the top left of the map. The dashed box represents the "island" selected for further analysis, with the extremely metal-poor stars focused on the upper 'coast' of the island.

metallicities, shown in Figure 3.7, suggests that our method is yielding reasonable estimates for [Fe/H]. Similar comparisons for our estimates of $T_{\rm eff}$ and $\log g$ (Figures 3.8 and 3.9, respectively) also show acceptable agreement with values from both the literature and GALAH DR3.

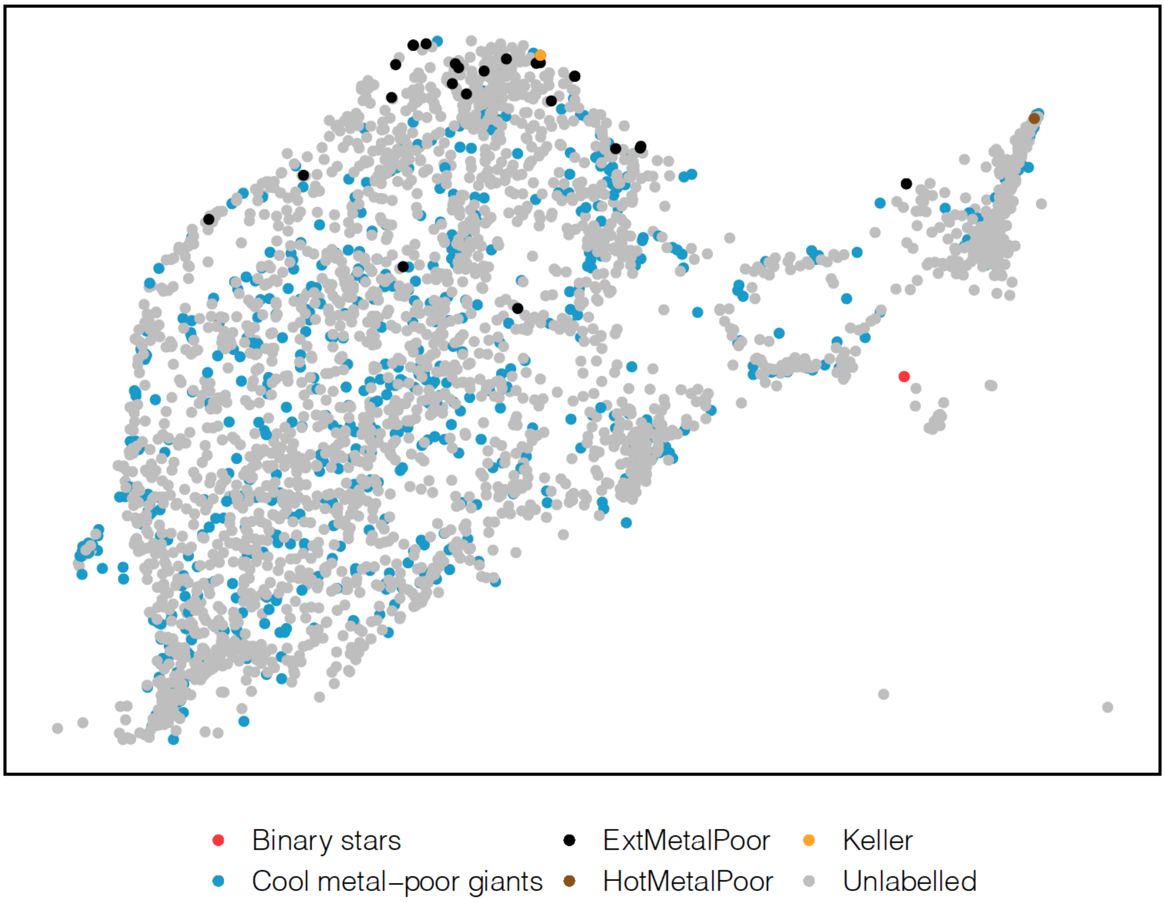


Figure 3.5: A zoomed in view of the island highlighted by the red-dashed box in Figure 3.4 with 2487 potential metal-poor stars. The known EMP stars lie on the upper extremely metal-poor “coast" of the island.

### 3.4.3 Candidates

Having used the template fitting process described in Section 3.3.2 to estimate the stellar parameters of our cluster in Section 3.4.2, we find 380 stars that have [Fe/H] $\leqslant -2.5$ and 135 stars with [Fe/H] $\leqslant -2.7$. For the rest of the discussion, however, we only consider stars that have [Fe/H] $\leqslant -3$, to satisfy the “extremely" metal-poor star designation. This results in 54 EMP candidates, 6 of which have an estimated [Fe/H] $\leqslant -3.5$. We note that 9 SIMBAD-sourced EMP stars from the literature in Table 3.1 are all contained in this t-SNE sample, and 7 (2 of which overlap with the literature) were identified as potential EMP stars by the GALAH DR3 pipeline, resulting in a net total of 40 potentially previously unidentified candidate EMP stars.

The spectra of the 54 candidate EMP stars are relatively featureless (with the exception of H$\alpha$ and H$\beta$) across the HERMES wavelength ranges; Table 3.2 shows the derived parameters for a few of our candidates. The spectrum of a representative EMP candidate from our selection is shown in Figure 10. This star has an [Fe/H] $< -4.5$, as determined by our stellar parameter routine.

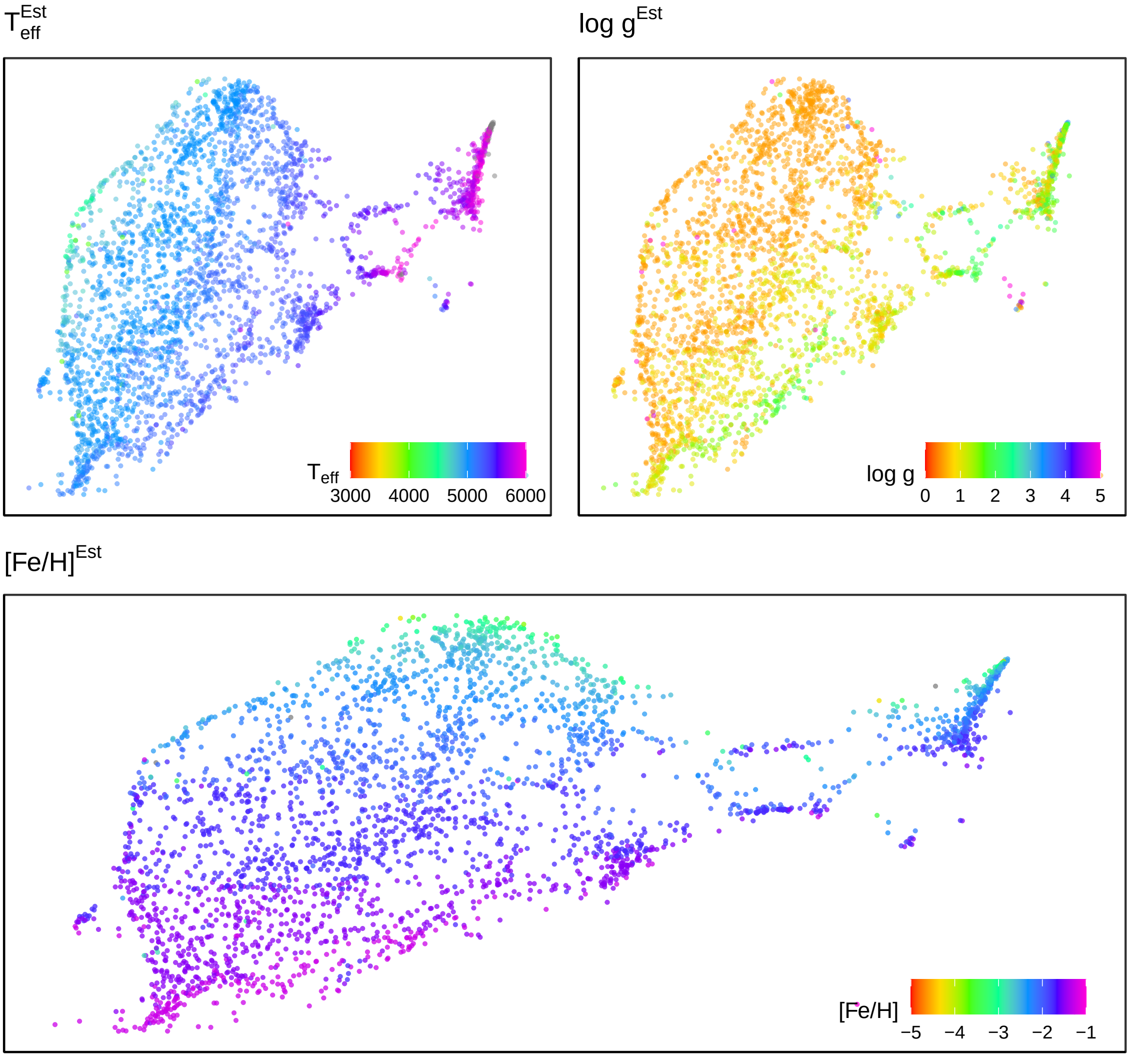


Figure 3.6: Three panels showing the selected island coloured by each stellar parameter. The top-left panel shows our estimated $T_{\rm eff}$, having a distribution of temperature from cold to hot going left to right. Similarly the top-right panel shows estimated $\log g$ having a similar left to right distribution. The lower panel shows estimated [Fe/H] having a gradient of high to low metallicity from bottom to the top, with the top edge in agreement with the previously seen “extremely metal-poor coast".

As a first attempt for confirmation that the candidates are likely EMP stars, we display their photometric properties in a parameter space that has successfully been used to select EMP stars. Figure 3.11 shows our sample cross-referenced with the SkyMapper photometric catalogue (Onken et al., 2019). Here $m_i$ represents a metallicity index, defined as $(v-g)_0 - 1.5(g{-}i)_0$, and $(g{-}i)_0$, as a proxy for $T_{\rm eff}$. We show the EMP selection region in the figure from Da Costa et al. (2019), and find that most of our candidates are red giants and our sample fits within this region. This suggests, at least in terms of the broad metallicity-sensitive features targeted by the SkyMapper photometry, that our sample contain *bona fide* EMP stars. We

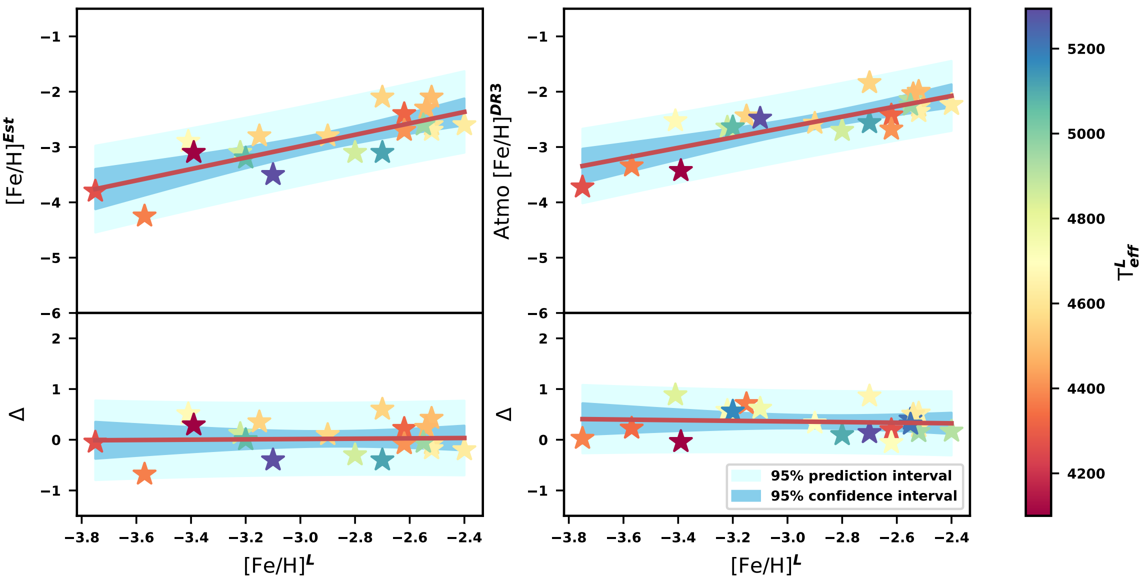


Figure 3.7: Two plots comparing our estimated [Fe/H] (left) and GALAH DR3 [Fe/H] (right) values with the literature. The top panels show the respective values while the bottom panels represent the difference between the method/s and the literature. The red line shows a linear best-fit to the data, with the prediction and confidence intervals as indicated by the shaded regions. Overall both methods have a 95% confidence band of approximately ±0.5 but the GALAH DR3 measured [Fe/H] values are higher on average than literature metallicities in this low-metallicity range.

note that Da Costa et al. (2019) find that 7% of the stars within the SkyMapper selection region ultimately prove to be EMP stars based on follow-up spectroscopy.

What about the candidates that fall *outside* of that selection box? We plot our candidates and the known literature stars on a color-magnitude diagram, using `pho_g_mean_mag` and the color $G_{BP} - G_{RP}$ from GAIA DR2 (Gaia Collaboration et al., 2018), and distances from GAIA (Bailer-Jones et al., 2018) in Figure 3.12. The majority of our candidates fall on the red giant branch along with some literature EMP stars, suggesting they are mostly red giants. Some of our candidates, however, are located near the main sequence turn-off. A significant portion of our EMP candidates indeed show higher surface gravities, suggesting they are actually main sequence turn off stars.

## 3.5 Discussion

At a high-level, given that we already had a large sample of high resolution spectra our candidate selection was relatively straightforward compared to previous EMP work: we have a magnitude-limited sample of stars and simply identified the population using iron and

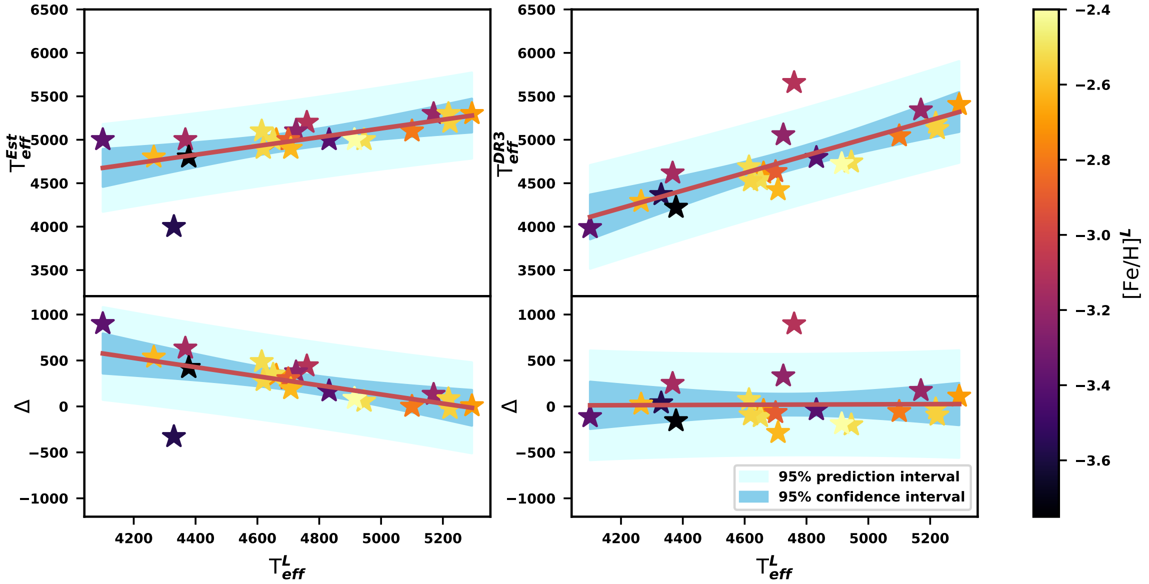


Figure 3.8: Two plots comparing our estimated $T_{\text{eff}}$ (left) and GALAH DR3 $T_{\text{eff}}$ (right) with the literature. The top panels show the respective values while the bottom panels represent the difference between the method/s and the literature. The red line shows a linear best-fit to the data, with the prediction and confidence intervals as indicated by the shaded regions. There is a trend in the errors of our estimation method, in that we have higher $T_{\text{eff}}$ at the lower end but overall have a similar error band to that of GALAH DR3.

hydrogen absorption lines[2]. We note this is only possible because, even with the relatively limited wavelength coverage of GALAH spectra, that there is still sufficient sensitivity to spectral features indicative of EMP-like metallicities (see Appendix A.1).

One question which arises is how our sample compares to previous work on the metallicity distribution function (MDF) for EMP stars. The topic has been explored in a number of recent studies (e.g., Da Costa et al., 2019; Youakim et al., 2020; Yong et al., 2021), with Yong et al. (2021) finding a slope for the MDF of $\Delta(\log \text{N})/\Delta[\text{Fe/H}] = 1.51$ dex per dex for $-4.0 < [\text{Fe/H}] < -3.0$, with an apparent steep drop-off below -4.0 (below -4.0 it would appear virtually all stars are C-enhanced, with the [Fe/H] values likely varying stochastically depending on Population III supernova yields).The left panel of Figure 3.13 shows the MDF for our candidate sample and the right panel shows a log-scaled histogram with the gradient of 1.51 from Yong et al. (2021) overlaid (red-dashed line). Here we can see that the “unbiased” nature of the current sample, which provides another way of investigating the form of the MDF, yields reasonably consistent results. However, given the MDF presented in Yong et al. (2021) and the current sample size ( 50 stars with $[\text{Fe/H}] < -3.0$), the probability of any of the current EMP candidates having $[\text{Fe/H}] < -4.0$ is not very high, as most will be closer to -3.0. Hence a significantly larger EMP sample is required for probing the low-metallicity end of the stellar MDF; in this paper we have demonstrated that applying our approach to larger

[2] As noted previously, we also used an oxygen feature, but only as a discriminant to remove hot stars that were contaminating the sample.

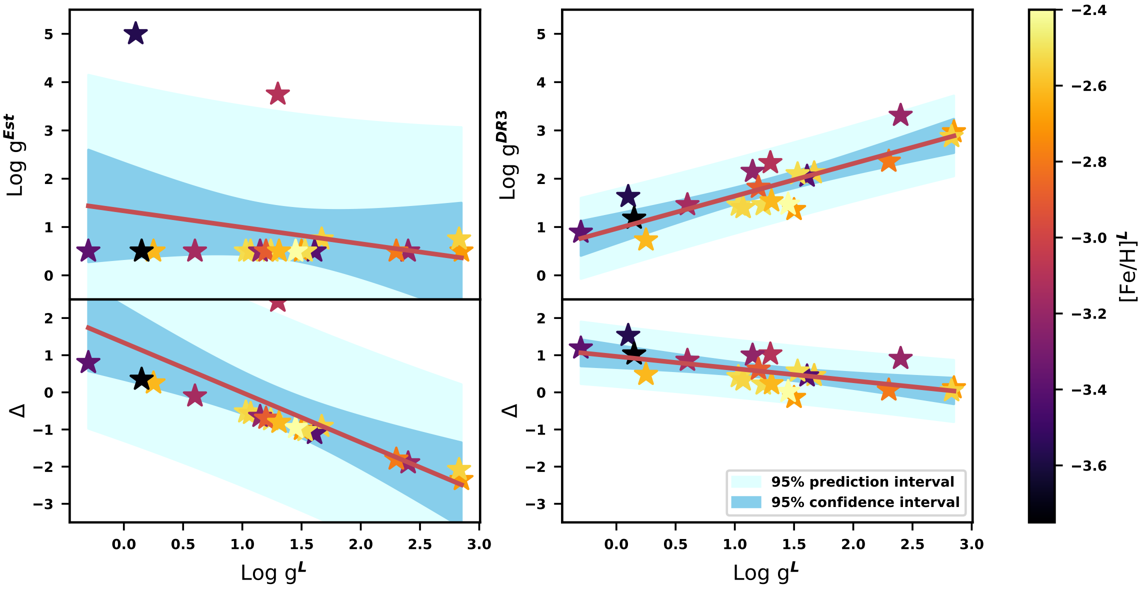


Figure 3.9: Two plots comparing our estimated log $g$ (left) and GALAH DR3 log $g$ (right) values with the literature. The top panels show the respective values while the bottom panels represent the difference between the method/s and the literature. The red line shows a linear best-fit to the data, with the prediction and confidence intervals as indicated by the shaded regions. Here you can clearly see that the GALAH DR3 estimates of log $g$ are a much better match, which is to be expected, given the relative simplicity of our method.

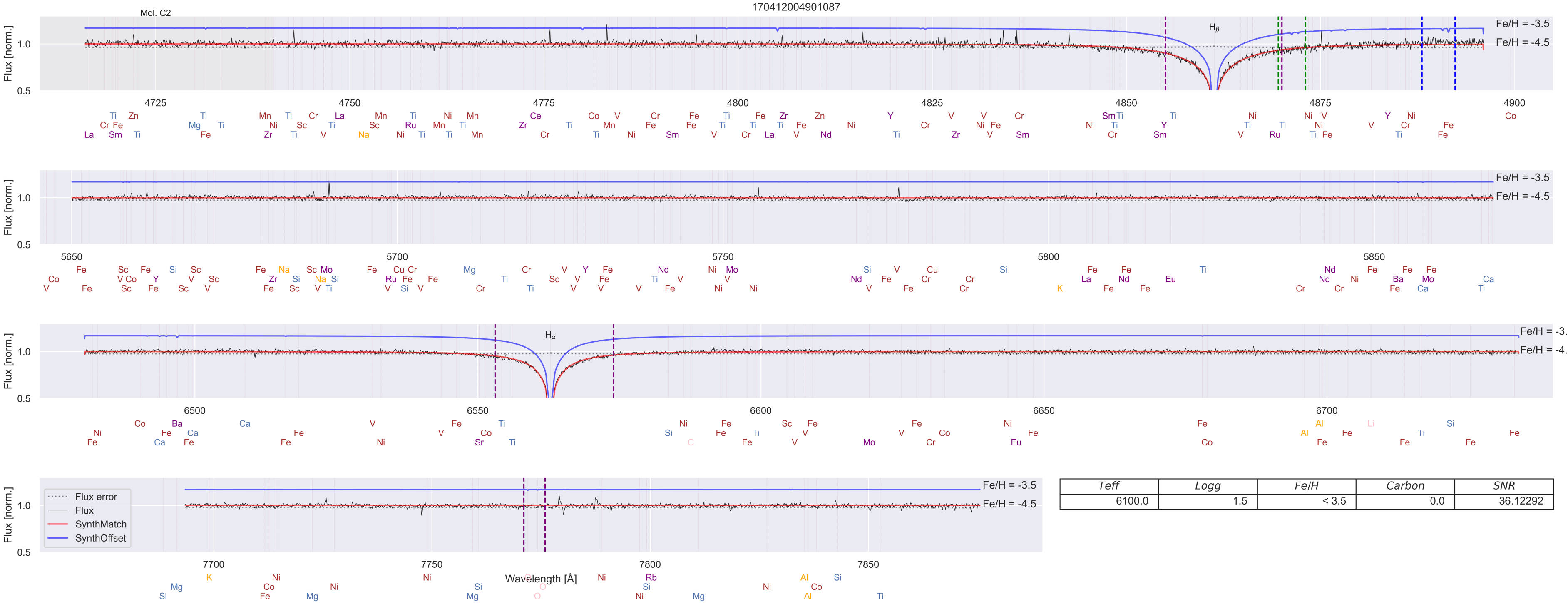
samples reaching fainter magnitudes is a key way to generate such an EMP sample.

Although the sample requires further spectroscopic observations to confirm our stellar parameter estimates, its relatively unbiased nature means there are a number of promising properties of the sample that suggest the method developed here has some advantages over other techniques for finding and understanding the EMP population.

Firstly, as shown in Figure 3.11 and Figure 3.12, we appear to have identified some main-sequence or main-sequence turnoff candidates. This is interesting because the sample of EMP stars from the literature observed serendipitously by GALAH (see e.g.Table 3.1) consists of essentially all giant stars, reflecting the fact that previous work (e.g., Starkenburg et al., 2017) prioritised probing larger volumes in order to obtain large samples of relatively rare EMP stars. For this reason most surveys specifically targeted stars with giant-like properties, whose high luminosities allow them to be studied at greater distances.

If even one of our main sequence EMP candidates turns out to be a *bona fide* main sequence or main sequence turn-off EMP star, this is an exciting opportunity to explore a less-studied population of EMP stars. The abundance patterns of main sequence stars are comparatively easy to understand because they have not yet been affected by evolution in the post main sequence phase. The ages of these stars are also more accessible through comparison to isochrones, which is important for placing these EMP stars into the context of the formation and assembly of the Milky Way.

Another advantage of our method, which differs from other EMP selection methods – e.g.,



| *Teff* | *Logg* | *Fe/H* | *Carbon* | *SNR* |
|---|---|---|---|---|
| 6100.0 | 1.5 | < 3.5 | 0.0 | 36.12292 |

Figure 3.10: A candidate from the t-SNE EMP island identified in Figure 3.5 plotted over the 4 GALAH wavelength ranges, and overlaid with the best synthetic spectrum match in red, as well as a vertically-shifted comparison spectrum offset by +1.0 in [Fe/H] shown in blue. The best-fit stellar parameters are listed in the bottom-right inset table and the vertical dashed lines represent the different wavelength regions used, as defined in Section 3.3.1.

| s_object_ID | RA | DEC | $T_{\mathrm{eff}}^{\mathrm{Est}}$ | log $g^{\mathrm{Est}}$ | [Fe/H]$^{\mathrm{Est}}$ |
|---|---|---|---|---|---|
| 131123002501215 | 63.5677656 | -60.151311 | 5000 | 0.50 | -3.00 |
| 131217002301168 | 64.8334861 | -58.678350 | 5200 | 0.50 | -3.10 |
| 140312003501132 | 203.154833 | -38.009181 | 4900 | 0.50 | -3.30 |
| 140711001301222 | 242.630802 | -25.337563 | 5000 | 0.50 | -3.20 |
| 140808004701080 | 28.0619680 | -72.320519 | 5600 | 0.75 | -3.40 |
| 140809004901060 | 40.9968414 | -70.248597 | 4000 | 5.00 | -3.00 |
| ... | ... | ... | ... | ... | ... |

Table 3.2: A subset of EMP candidates, with the full candidate list available electronically.

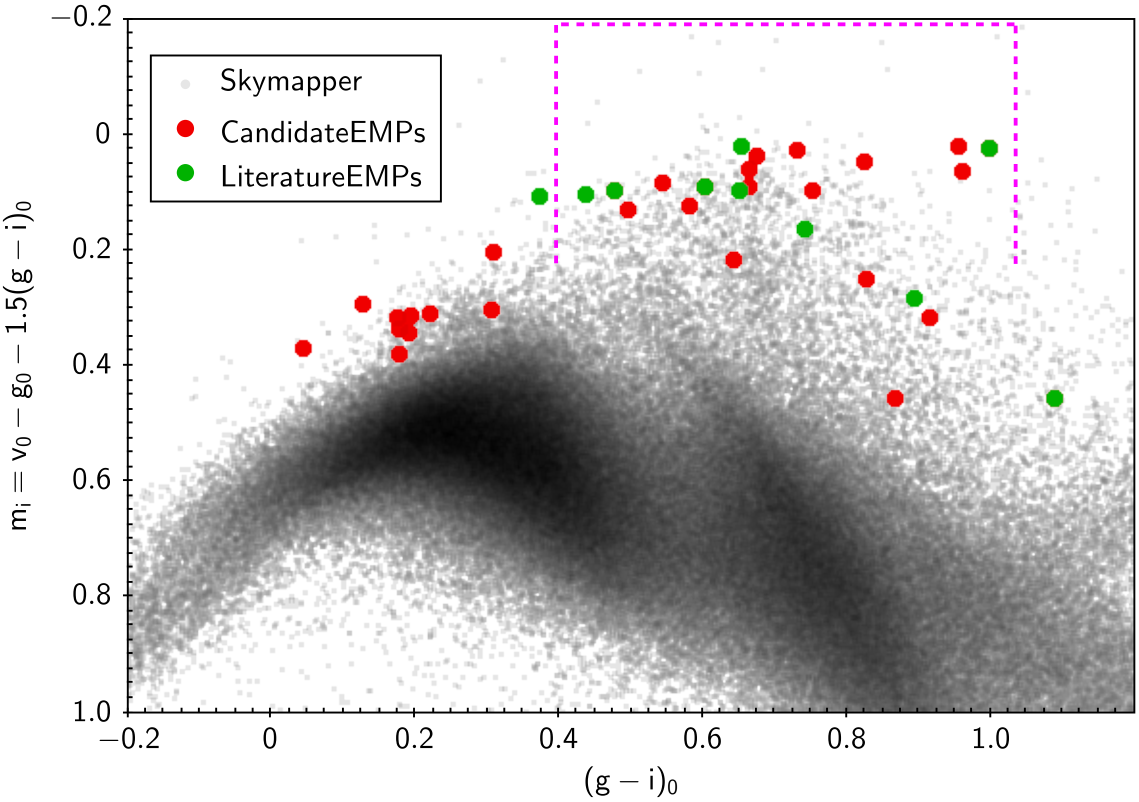


Figure 3.11: A Skymapper metallicity-sensitive diagram, showing most of our candidates are likely red giants falling within the SkyMapper selection window (dashed magenta lines, from Da Costa et al., 2019). The compact grouping to the left represents candidate EMP main sequence turn-off stars.

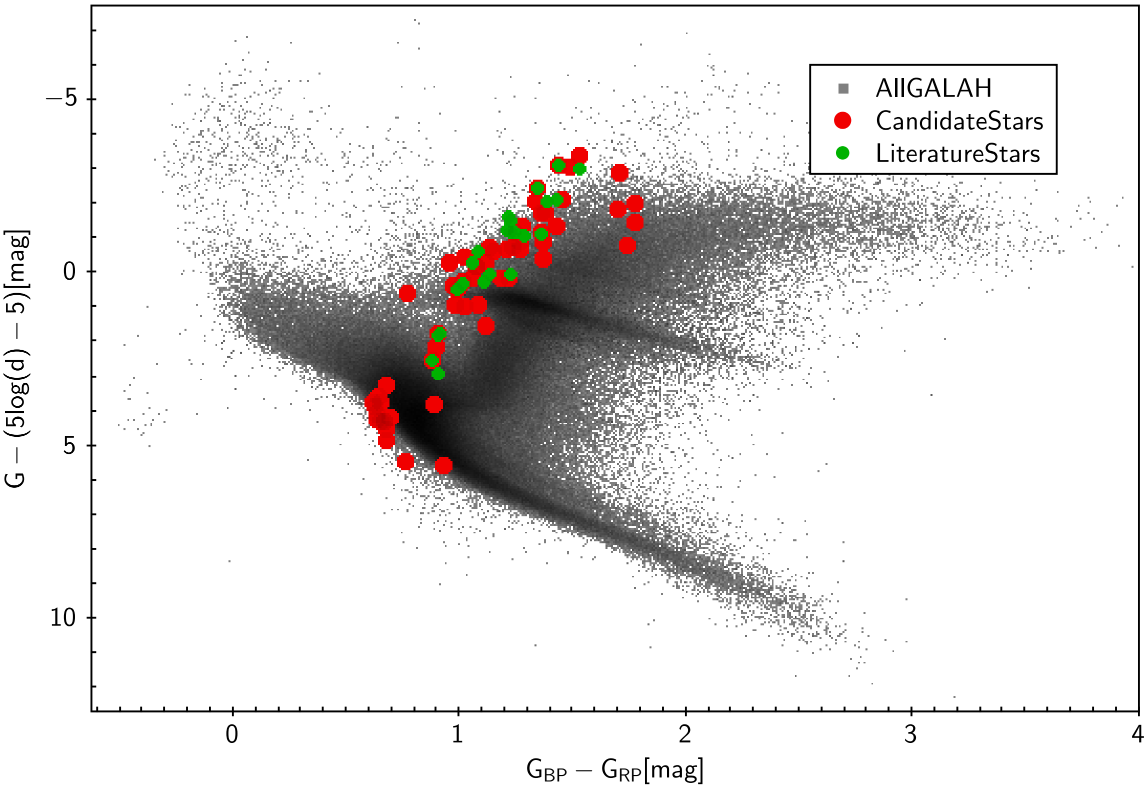


Figure 3.12: The color-magnitude diagram using magnitudes and distances from GAIA DR2 for the candidate EMP stars (yellow circles).The majority of the EMP candidates are red-giants, while 20% appear to be consistent with main-sequence turn off stars. Green circles are known EMP stars as defined in Table 3.1, including the most iron-poor star known, SMSS J031300.36–670839.3 (Keller et al., 2014).

some combinations of photometric filters (Da Costa et al., 2019) – is that carbon features did not affect our candidate selection. This means we have a relatively unbiased sample with respect to carbon abundance. Carbon-enhanced metal-poor stars (which have [C/Fe] $> 0.7$), become increasingly more frequent as [Fe/H] decreases (e.g., Placco et al. 2014b), and for [Fe/H]$\leqslant -4.0$, carbon-enhanced metal-poor stars dominate the known sample. Hence this candidate sample presents an opportunity to explore the relative fraction of carbon-enhanced metal-poor stars as a function of [Fe/H] free of carbon-influenced selection bias. In fact, GALAH does not cover the wavelength ranges required to estimate carbon at extremely low metallicities, making follow-up observations of this magnitude limited candidate sample essential for studying its carbon abundances.

The orbital information for our EMP candidates is captured in the vertical action and azimuthal action plot shown in Figure 3.14, similar to Figure 1 of Sestito et al. (2020) and Figure 5 of Cordoni et al. (2021). While our admittedly smaller set of candidates does not extend as high in vertical action as the Sestito et al. (2020) sample, we do see a significant near-“planar” component, biased toward prograde motion, in agreement with the results of both those authors and the Skymapper-based study of Cordoni et al. (2021). Hence, while these kinematic data are not proof of the EMP nature of our candidates, they are consistent

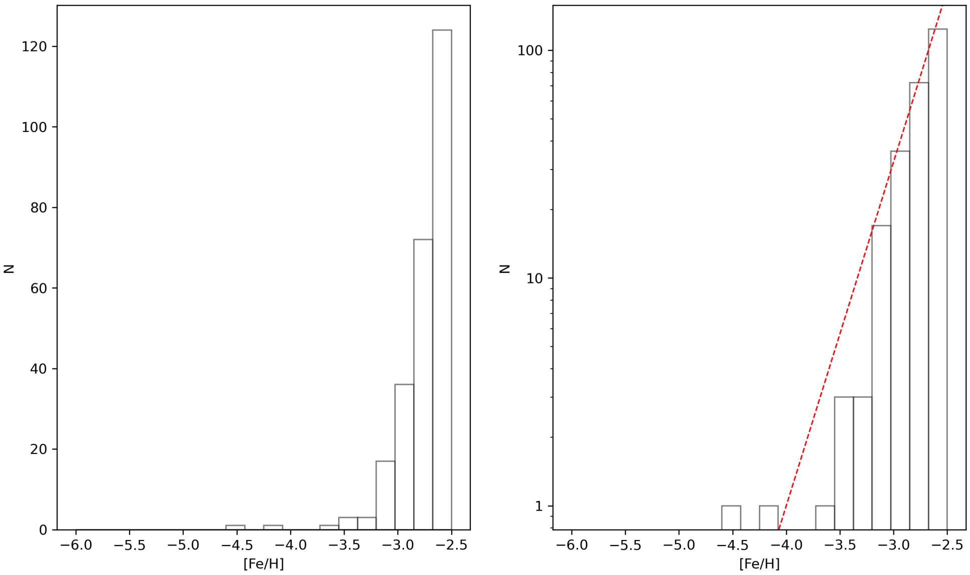


Figure 3.13: Metallicity distribution function for our candidate sample (left) and the log-scaled distribution function with the slope of 1.51 as determined in Yong et al. (2021) overlaid (red-dashed line). For this histogram we only show candidates from the main GALAH survey, which is a magnitude-limited sample. We specifically exclude stars in GALAH DR3 targeted by other surveys (K2-HERMES Wittenmyer et al. (2018), TESS-HERMES Sharma et al. (2018) and GALAH-faint) because they incorporated fainter stars. The MDF follows a similar trend to that as seen in Yong et al. (2021), except for steeper fall-off at [Fe/H] $< -3.3$.

with the observed properties of confirmed EMP stars.

Finally, we note that the method presented here evolved from Hughes (2017), which employed t-SNE on GALAH spectra to classify them and identify several interesting classes of objects, including metal-poor stars. In that work the fit used a relatively simple set of absorption features in the spectra. In the present work, we found a significant improvement in the quality of the candidates by assessing different line combinations in order to improve our metallicity sensitivity (see Appendix A.1). We furthermore included the GALAH infrared fourth channel because it contains an oxygen feature – not previously considered – which served as a discriminant to reject spurious hot stars.

### 3.5.1 Advantages of a machine learning-based approach over more traditional $\chi^2$ fitting methods

As shown in Figure 3.5, our method uses t-SNE to isolate candidate EMP stars in a region with spectra similar to known EMP stars from the literature. By fitting synthetic spectra to the candidate EMP stars, we refine the selection of EMP candidates in the t-SNE space of EMP candidates to the top portion of the data shown in Figure 3.6. The clustering of known and candidate EMP stars in essentially a localised region in the entire t-SNE parameter space

| Method | EMP stars (%) | Extraneous sources (%) | Total EMP Candidates |
|---|---|---|---|
| $\chi^2$ only | 19 | 81 | 126 |
| t-SNE and $\chi^2$ | 90 | 10 | 60 |

Table 3.3: Accuracy percentages between our method (i.e. t-SNE classification, then a $\chi^2$ fit to models) and a traditional $\chi^2$-fitting technique for finding EMP stars. Percentage of EMP stars is the fraction of the total count that passed a visual inspection. Extraneous sources included both candidates with bad fits and those with strong absorption features, indicating that they are not likely to be EMP stars. The accuracy percentage of candidates that were found to be good EMP stellar candidates is higher using our method.

illustrates the potential power of our method. Nevertheless, a valid question is whether there are any improvements on our machine learning-based method in terms of finding EMP stars over a simple $\chi^2$ fit to a wide range of synthetic spectral templates.

We tested the $\chi^2$ stellar parameter routine on the full GALAH dataset, to potentially identify EMP stars that did not fall within the t-SNE EMP island identified in Figure 3.5. The results of this run are compared to the t-SNE run in Table 3.3. The purely $\chi^2$ method returned more potential candidates, but upon visual inspection, 81% of those candidates were poorly fit, and some had strong absorption features, indicating that they are not good EMP candidates. The increased fraction of bad fits is likely because of model systematics – the minimum $\chi^2$ might not be representative of actual EMPs. Applying t-SNE before running a $\chi^2$ fitting routine minimises this effect.

Moreover, we found that the $\chi^2$ method does not contain all the t-SNE EMP candidate sample: only 10 of our total sample of 60 (54 candidates and 6 spurious stars) are found. Finally we also note that not all of the literature stars from Table 3.1 were recovered in the $\chi^2$ sample: only 3 of 23 are found.

## 3.6 Conclusions

We have demonstrated a methodology for finding EMP stars within a spectroscopic dataset – in this case spectra of $\sim 600,000$ stars from the GALAH high-resolution survey – that is both computationally efficient and accurate, and may potentially be adapted to find other specific types of stars. Furthermore we have shown that, using the GALAH wavelength ranges, we can derive metallicities down to [Fe/H] $\sim -3.5$.

The candidate list we have identified is distinct from the results of many past surveys targeted specifically at EMP stars (e.g., Da Costa et al., 2019; Starkenburg et al., 2017). Given the nature of the GALAH dataset – essentially a magnitude-limited sample of stellar spectra – our candidate list does not preferentially select giant stars (although, given their greater luminosity, giant stars probe a larger volume). This means we are sensitive to main-sequence and main-sequence turnoff stars, which are an interesting EMP population because, not having undergone dredge-ups, they are more likely to retain their original abundance patterns, and in the case of main-sequence turnoff stars, they can potentially yield useful stellar ages. Moreover the lack of strong carbon features in the GALAH wavelength windows means we are not biased against carbon-enhanced metal-poor stars – a significant fraction of

EMP stars (Yong et al., 2013; Lee et al., 2013) – unlike some photometric-based EMP star surveys (cf. Da Costa et al., 2019).

With regard to our methodology, we found hybrid approach, i.e., pre-selection using t-SNE focused on specific wavelength regions, followed by parameter estimation via $\chi^2$-fitting, to be the most efficient way to identify candidate EMP stars. Simpler "brute-force" methods, for example applying t-SNE to the entire spectral range, or skipping machine-learning-based pre-selection and going straight to $\chi^2$-fitting to template spectra, proved to be both much more computationally intensive and much more likely to include extraneous spectra in their output. Although our method was tailored to GALAH spectra, we expect that similar techniques should be applicable to datasets from other ongoing and future large spectroscopic surveys, including WEAVE (Dalton et al., 2014) and 4MOST (de Jong et al., 2019).

While we have demonstrated that our metallicity estimates – along with those from the GALAH DR3 pipeline – are fairly reliable with regard to identifying EMP star candidates, follow-up observations, ideally covering additional regions of the optical spectrum more sensitive to low-metallicity measurements, are required to confirm these estimates, as well as to determine the abundances of carbon and other specific elements of interest (see, e.g., Beers and Christlieb, 2005; Frebel and Norris, 2015). To this end we are engaged in a program of follow-up spectroscopy, with results seen in the work of Da Costa et al. (2023).

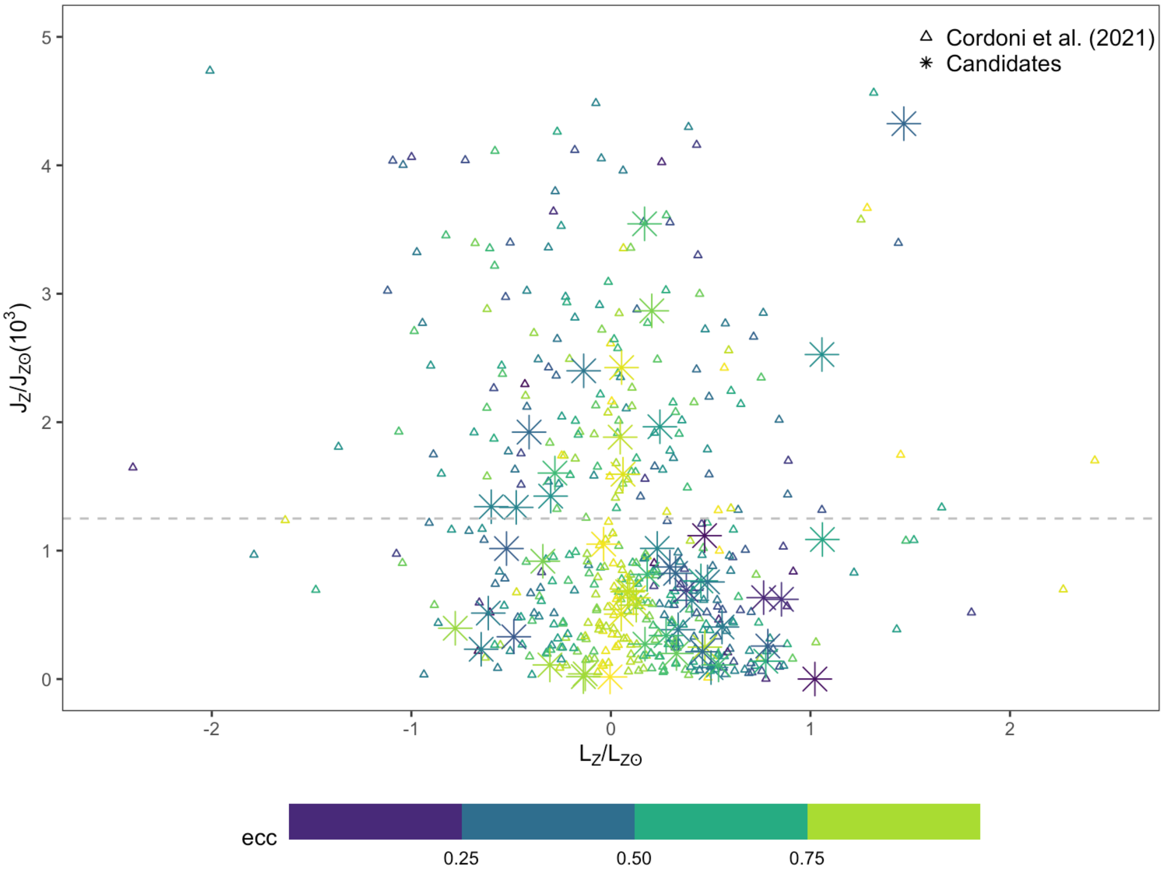


Figure 3.14: Vertical versus azimuthal action components color-coded by eccentricity for our EMP candidates with metallicities [Fe/H] $< -3$ (star symbol), as well as literature values from Cordoni et al. (2021, triangle symbol). The action quantities are scaled by the solar values (i.e., $L_{z\odot} = 2009.92$ km s$^{-1}$ kpc, $J_{z\odot} = 0.35$ km s$^{-1}$ kpc). In this parameter space, we adopt the same horizontal dashed line at $^{J_z}/_{J_{z\odot}} = 1.25 \times 10^3$ as in Cordoni et al. (2021) to distinguish between planar and non-planar orbits. The distribution of our candidates appears to be consistent with the observed orbital properties of confirmed EMP stars shown in Figure 1 of Sestito et al. (2020) and Figure 5 of Cordoni et al. (2021).

# 4

# Quasar and galaxy classification using Gaia EDR3 and CatWise2020

Having discussed the application of a clustering method to spectra, we now turn our attention to the classification of extragalactic sources using astrometry and photometry. The base dataset for this is the Gaia survey with additional photometry in the infrared region from CatWISE2020. The statistical approach here is an assessment of different tree based algorithms in comparison to Gaussian Mixture models and a discussion about the use of priors. This chapter is based upon work done in Hughes et al. (2022a) which builds upon the work previously done by (Bailer-Jones et al., 2019). Both co-authors (C. Bailer-Jones and S. Jamal) contributed to the development of the methods, the analysis and interpretation.

## Brief Summary

We assess the combined use of Gaia photometry and astrometry with infrared data from CatWISE2020 in improving the identification of extragalactic sources compared to the classification obtained using Gaia data. Here we perform a comprehensive study in which we assess different input feature configurations and prior functions to identify extragalactic sources in Gaia, with the aim of presenting a classification methodology that integrates prior knowledge stemming from realistic class distributions in the Universe. In this work, we compare different classifiers, namely Gaussian mixture models (GMMs) and the boosted decision trees, XGBoost and CatBoost, in a supervised approach, and classify sources into three classes, namely star, quasar, and galaxy, with the target quasar and galaxy class labels obtained from the Sloan Digital Sky Survey Data release 16 (SDSS16) and the star label from Gaia EDR3. In our approach, we adjust the posterior probabilities to reflect the intrinsic distribution of extragalactic sources in the Universe via a prior function. In particular, we introduce two priors, a global prior reflecting the overall rarity of quasars and galaxies, and a mixed prior that incorporates in addition the distribution of the extragalactic sources as a function of Galactic latitude and magnitude. The best classification performances, in terms of completeness and purity of the extragalactic classes, namely the galaxy and quasar classes, are achieved using the mixed prior for sources at high latitudes and in the magnitude range G = 18.5 to 19.5. We apply the identified best-performing classifier to three application datasets from Gaia Data Release 3 (GDR3), and find that the global prior is more conservative in what it considers to be a quasar or a galaxy compared to the mixed prior. In particular, when applied to the quasar and galaxy candidate tables from GDR3, the classifier using a global prior achieves purities of 55% for quasars and 93% for galaxies, and purities of 59% and 91%, respectively, using the mixed prior. When compared to the performances obtained on the GDR3 pure quasar and galaxy candidate samples, we reach a higher level of purity, 97% for quasars and 99.9% for galaxies using the global prior, and purities of 96% and 99%, respectively, using the mixed prior. When refining the GDR3 candidate tables via a cross-match with SDSS DR16 confirmed quasars and galaxies, the classifier reaches purities of 99.8% for quasars and 99.9% for galaxies using a global prior, and 99.9% and 99.9% using the mixed prior. We conclude this work by discussing the importance of applying adjusted priors that portray realistic class distributions in the Universe and the effect of introduction infrared data as ancillary inputs in the identification of extragalactic sources.

## 4.1 Introduction

Classification of galactic and extragalactic sources is fundamental for statistical analyses of large populations, as well as for probing the properties of individual objects. For instance, quasars (quasi-stellar objects) refer to highly luminous active galactic nuclei (AGNs), which are used as probes to investigate fundamental questions in Cosmology such as galaxy evolution (e.g., Harrison et al., 2018), the composition of the interstellar medium (e.g., Li et al., 2022), and supermassive black-hole formation and evolution (e.g., Croom et al., 2009).

All-sky surveys, such as the Sloan-Digital Sky Survey (SDSS) (York et al., 2000) and the Wide-field Infrared Survey Explorer (WISE) (Wright et al., 2010), have created detailed maps of the Universe at optical and infrared wavelengths. Infrared data is highly informative for the classification of stars, quasars, and galaxies. As demonstrated in the work by Kurcz et al. (2016), the authors exploited the infrared colours from WISE and reported a 90%–95% classification accuracy across all object types despite the limitations observed for galaxy sources with a high dust component. The combined use of infrared data with optical photometry should, in principle, enhance the classification accuracy and reduce the number of false positives across all object types.

However, a large fraction of work on classification fails to consider the intrinsic distribution of sources of different classes, and only reports results —in particular the accuracy (i.e., the fraction of correct predictions per target class)— using a test set that is typically not representative of the observable Universe. Moreover, such test sets often under-represent the stellar contaminants that would, in practice, lower the purity of extragalactic classification. To account for the actual distribution, we introduce a prior (discussed in detail in Sect. 4.3.3) which, in a Bayesian framework, is used to adjust the estimated model posterior probabilities in order to reflect the class distribution of sources we would expect to exist. Furthermore, after a model has been applied, we apply an adjustment factor to the distribution of sources, such that the performance metrics are computed as if the model had been applied to the dataset with a realistic expected distribution. Applying both the prior and the adjustment factor result in classification performances that are more representative of what we can achieve —although inevitably lower— than the performances obtained when the prior and adjustment factor are not applied. Despite the lower results of some models, applying the prior correction is a necessary step because it will reveal the real classification performances, especially for large-scale surveys for which the observed sources are unknown.

The Gaia mission is an optical mapping survey designed to focus on stars in our Galaxy (Gaia Collaboration et al., 2016). During Gaia's scan of the entire sky, the satellite observes all point-source-like objects down to a magnitude limit of $G \simeq 21$, including extragalactic objects (Gaia Collaboration et al., 2022). A reliable methodology to identify extragalactic sources would benefit the construction of comprehensive catalogues useful for addressing fundamental questions in astronomy.

To design this method, we followed a similar approach to that of Bailer-Jones et al. (2019), who obtained a classification of extragalactic sources using Gaussian mixture models (GMMs; Fraley and Raftery (2002)) applied to Gaia Data Release 2 (DR2) photometry and astrometry. In this study, we consider photometry and astrometry from Gaia Early Data Release 3 (GeDR3) and the addition of infrared photometry from CatWISE2020 (Marocco et al., 2021), as well as the application of gradient-boosting decision trees, namely XGBoost (Chen and Guestrin, 2016) and CatBoost (Dorogush et al., 2017) to construct a three-class classifier (quasar, galaxy, star). The objective of our work is to assess the effects of additional information from infrared photometry and the omission of parallax and proper motions on the classification of extragalactic sources. We also aim to evaluate different classification algorithms and the appropriate use of different priors to ensure that the reported performance results are reflective of reality.

## 4.2 Data

Our input data comprise astrometry and photometry from the GeDR3 catalogue and infrared photometry from the CatWISE2020 catalogue. The training and test datasets for the quasars and galaxies are based on the sixteenth data release of SDSS (SDSS-DR16, Ahumada 2020) while the star sample is built from the Gaia GeDR3 catalogue. We are aware that SDSS is not complete and does not cover the same magnitudes as Gaia; however, we accept these limitations when building our class samples.

The Gaia GeDR3 catalogue (Gaia Collaboration et al., 2021) was published on 3 December 2020 for observations acquired between 25 July 2014 and 28 May 2017, spanning a period of 34 months. GeDR3 consists of astrometry, and broad band photometry in the G, $\mathrm{G_{BP}}$, and $\mathrm{G_{RP}}$ bands for about 1.8 billion sources. In this work, we set a limit in magnitude up to $G > 14.5$ mag. This work commenced prior to Gaia Data Release 3 (GDR3) and therefore made use of the public data in GeDR3. However, as the photometry and astrometry remain unchanged between GeDR3 and GDR3, our findings are applicable to DR3.

The CatWISE2020 catalogue consists of about 1.8 billion sources observed across the entire sky selected from the WISE and NEOWISE survey data in the W1 and W2 (3.4 µm and 4.6 µm) bands (Marocco et al., 2021). In our study, we chose CatWISE2020 instead of All/unWISE as the CatWISE2020 catalogue extends to fainter magnitudes and the associated data processing pipeline uses the full-depth unWISE co-addition of AllWISE and NEOWISE 2019 Data Release for aperture photometry (Marocco et al., 2021), which results in a significant improvement over the AllWISE data. A five-arcsecond positional cross-match of CatWISE2020 with GeDR3 identifies about 1.5 billion sources.

### 4.2.1 Classes

The goal of our classification is to identify objects in the target star, quasar, and galaxy classes. The definitions of the target classes are similar to those used by Bailer-Jones et al. (2019), but are augmented with the aforementioned CatWISE2020 cross-match. However, in our application we do not use —and therefore do not require the availability of— parallax and proper motions. This approach results in a much larger set of galaxies, because most galaxies observed by Gaia lack published parallax and proper motions due to a poor fit of the astrometric model on account of their physical extent. As these may indicate a different type of galaxy, this effectively changes our class definition. We ensure there are no common sources between the three class datasets.

#### Quasars

The SDSS-DR16 quasar catalogue (Lyke et al., 2020) contains 750 414 quasars confirmed by optical spectroscopy. Its authors estimate the contamination to be around 0.5%. We select objects with a `zWarning` flag equal to zero, indicating a higher reliability in the classification or the redshift estimation. We cross-match the selected sample to GeDR3 by sky position with a one-arcsecond search radius using the CDS X-match tool, finding 489 581 matches in total. This constructed sample is then compared with the cross-matched sample from GeDR3 and the CatWISE2020 catalogue, resulting in 484 749 objects with GeDR3 features and CatWISE2020 magnitude measurements in the W1 and W2 bands.

#### Galaxies

The sample of galaxies in our train and test datasets is constructed from SDSS-DR16 Ahumada (2020); Blanton (2017). We select 777 409 objects from the `SDSS SpecObjAll` table on the SDSS Skyserver identified as `GALAXY` with `zWarning` equal to zero, and are identified as neither `AGN` nor `AGN BROADLINE` in the subclass field. The selected sample is similarly cross-matched with GeDR3, finding all objects. In our selection, we relax the requirement of the parallax and proper motions, as such information may be unavailable for several sources in Gaia, particularly amongst galaxy sources. Applying the defined criterion, we retain about 90% of the galaxy sources. Furthermore, supplementing the CatWISE2020 colours to our constructed sample results in a total of 766 310 objects. Following the work by Bailer-Jones et al. (2019), we apply a colour cut to the galaxy sources using the same colour-edge locus as shown in Fig. 4.1. Objects below this locus represent stellar contaminants within the galaxy sample. Potential sources of contamination include errors in the SDSS classification or the Gaia BP/RP spectra affected by blends of nearby objects (De Angeli et al., 2022). The colour cut removes 1061 contaminants from our galaxy sample.

#### Stars

The spectroscopic selection for stars from SDSS data is complex, ill-defined, and likely affected by a biased distribution of stellar types. We therefore do not use SDSS to define the star class. We exploit the fact that the majority of observed sources in Gaia are expected to be stars. We therefore construct our star sample via a random subset of 3 million sources from the Gaia catalogue in which known galaxies and quasars are filtered out. We augment the sample with the CatWISE2020 cross-match, resulting in 1.8 million sources identified as stars. In

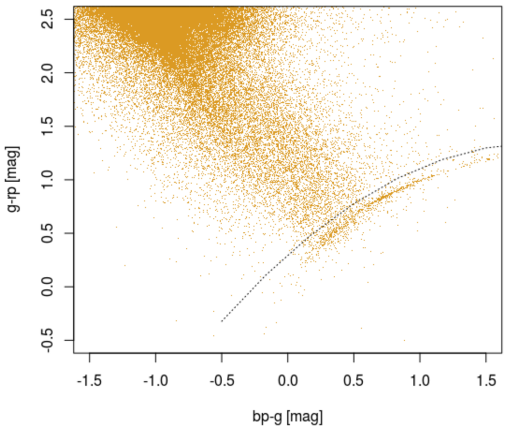


Figure 4.1: Colour–colour diagram of the galaxy class. Sources below the dashed line are contaminants that are removed from the galaxy sample.

our constructed star sample, we could expect a non-zero level of contamination from non-stellar sources. This contamination level is unknown, but our prior defined in Sect. 4.3.3 is our expectation. Ideally, our classifier trained on the cleaned sample would be robust to contamination.

### 4.2.2 Training, validation, and test sets

The full dataset is the combination of the quasar, galaxy, and star samples. The data are split into two equal parts at random. The first part is then split again into ten equal parts, with nine being used for training, and one part for validation, to monitor the performance during the training. For brevity, we often refer to these two together as the 'training data'. The second part is the test set which is kept back to assess the fixed models.

During the training phase, the training dataset is used to train the statistical model while the validation set is used to assess the performance of the trained model. After convergence, the trained model is stored and the test set, that is, a subset of the data unseen during training, is used to evaluate the performances of the classifier.

For the classifier trained on the balanced dataset, we select a random subset of 200 000 sources of each class for the training (90 000 for training and 10 000 for validation) and test datasets. By constructing a balanced classifier, we are able to directly compare the intrinsic performance of the models trained on different feature configurations and classification methods and identify the best performing method. The class imbalance is addressed in the discussion of the priors in Sect. 4.3.3.

Having selected the best-performing model using the balanced training and test dataset discussed in Sect. 4.4.1, we re-define our training and test datasets to use as many of the available sources in the quasar and galaxy class as possible by sampling a random subset of 900 000 stars, 200 000 quasars, and 370 000 galaxies from the full dataset. We train the feature configuration and classification method identified using the balanced dataset on this imbalanced training and test set in Sect. 4.4.2. This resulting classifier is applied to the application sets in Sect. 4.5.

### 4.2.3 Application sets

We use three datasets derived from GDR3 (Vallenari et al., 2022) to demonstrate the application of our best-performing classifier.

- A subset of 50 million randomly selected GDR3 sources that have CatWISE2020 photometry.
- Quasar candidate table from GDR3: The quasar tables described in Gaia Collaboration et al. (2022) represent datasets where there is an estimation of the number of quasars within GDR3. The quasar candidate table contains 6 649 162 sources with a purity of 0.52, and is refined further in the pure subsample (1 942 825 sources) with a purity of 0.96.
- Galaxy candidate table from GDR3: Similarly defined in Gaia Collaboration et al. (2022), the full table reports 4 842 342 candidates with a purity of 0.69, and the pure sample (2 891 132 sources) reaches a purity of 0.94.

There are 144 109 sources in common between the quasar and galaxy candidates (Gaia Collaboration et al., 2022). For ease of interpretation of our results **on** these tables, we therefore choose to remove these sources from the subsequent analysis.

### 4.2.4 Feature selection

In feature selection, an important condition is the completeness of each feature, as missing data often cause many statistical methods to fail. As noted in section 4.2.1, a large fraction of galaxies do not have published parallaxes and proper motions in GeDR3. We therefore disregard both as input features in order to retain as many sources as possible. As inputs to the classifier, we test various combinations of eight features: six of the eight features are defined in Bailer-Jones et al. (2019), which we refer to as 'Gaia_f', and the other two features are W1-W2 and the G-W1 colour constructed from the CatWISE2020 catalogue. The six features from Gaia_f are apparent magnitude (G), sine of the Galactic latitude ($sinb$), g-rp (G-RP), bp-g (G-RP), relative variability in the G band (relvarg), and the astrometric unit weight error (UWE). We report the distribution of each feature in Fig. 4.2 and their descriptions below:

- Figure 4.2 shows the distribution of the broadband G magnitude in Gaia and the colours BP-G, G-RP, W1-W2, and G-W1. Quasars have characteristic optical-infrared colours. In the colour–colour and colour–magnitude space, quasars can be discerned from other stellar objects as well as from galaxies; see Fig. 4.3. Additionally in Fig. 4.3, we can see the clear distinction from the galaxy class. Due to the clear separation between the distinct classes seen in Figs. 4.2 and 4.3, we consider the colour information as one of the main discriminating features of the target classes.
- Galactic longitude and latitude ($l, b$) can also be useful discriminants. Compared to stars, for which the distribution is concentrated towards the Galactic disk and the bulge, extragalactic objects are expected to be uniformly distributed across the entire sky (Copernican principle). However, such distribution is not observed due to the strong interstellar extinction in the disk of the Galaxy concealing extragalactic sources at low

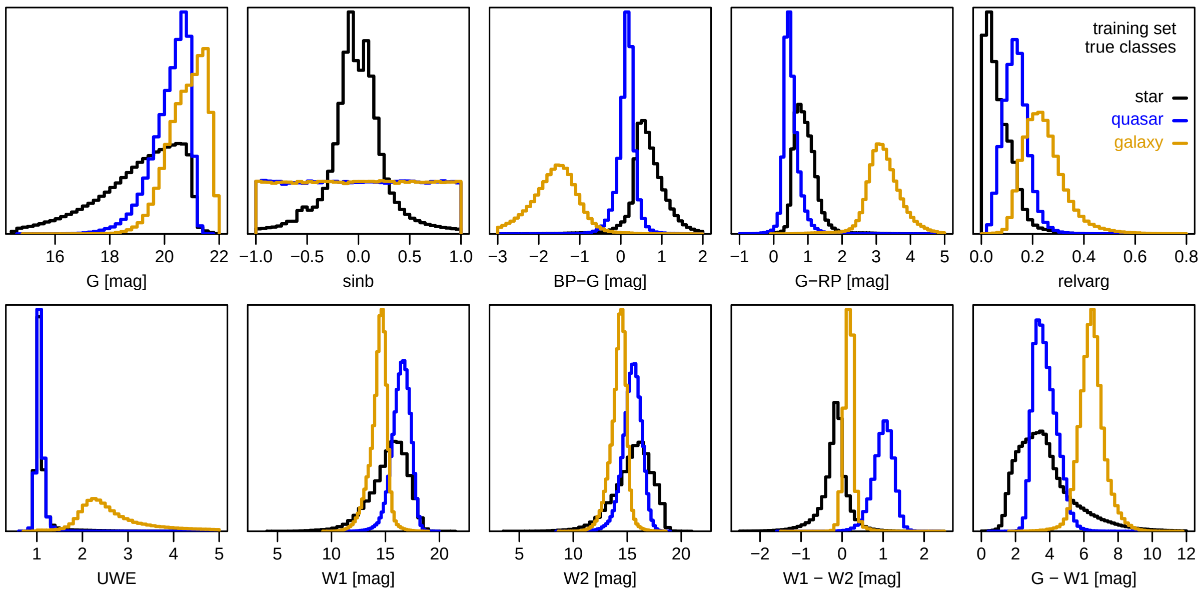


Figure 4.2: Distribution of the features from the training dataset, coloured according to their true classes. Black: stars. Blue: quasars. Orange: galaxies. Each distribution is separately normalised and the *sinb* has been randomised for quasars and galaxies (constant probability per unit sky area).

latitudes. Due to the SDSS sky coverage, the extragalactic objects in our training and test datasets follow a non-uniform distribution. We corrected our sample from this selection effect by randomising the latitude of these objects in our training and test datasets with values drawn from a uniform distribution in *sinb*. This approach may not be a perfect solution because, as mentioned, we do not expect to see a large fraction of extragalactic objects at low latitudes. While this may help us find otherwise-difficult-to-detect extragalactic objects at low latitudes, it may also lead to a higher number of false positives. We accept this limitation. Galactic longitude is a problematic parameter because it wraps at $l = 0\circ = 360\circ$ and is not used as a model feature. However, we do use $l$ when computing our priors to account for the footprint of SDSS in comparison with Gaia (see Sect. 4.3.3).

- The relative variability in the G-band, which we call 'relvarg' following the work by Bailer-Jones et al. (2019), is defined as the ratio of the standard deviation of the epoch photometry to its mean. Relvarg can be computed from the fields in GeDR3 as phot_g_n_obs/phot_g_mean_flux_over_error. Figure 4.2 shows a higher variability in quasars compared to stars. Galaxies also show large levels of variability, although in Gaia this effect is a spurious artefact due to galaxies being extended in their surface-brightness profile. At each epoch scan, Gaia will determine a slightly different photocentre possibly related to a different photometry. However, we can exploit this behaviour to help distinguish galaxies.

- The astrometric unit weight error, UWE, is defined as the square root of the $\chi^2$ multiplied by the number of degrees of freedom of the astrometric solution. A larger UWE value correlates to a weak fit to the astrometric solution and generally an enhanced value for some galaxies. We do not use the re-normalised UWE (RUWE), which also removes dependencies on colour and magnitude, because RUWE is not defined when the parallax and proper motions are missing.

## 4.3 Classification

In our work, the goal of the classification task is to find an optimal mapping between a class label (i.e. star, quasar, or galaxy) and a set of features characteristic of a given object. Several methods proposed in the literature have exploited supervised classification to determine the best mapping between input features and discrete classes. In our work, we seek a probabilistic classification, whereby a trained classifier generates a probability that an object belongs to a class, offering more flexibility in the determination of the final class prediction. Moreover, exploiting a Bayesian framework allows, on one hand, to define the posterior probability for an object to belong to a specific class, and on the other hand to incorporate the use of a highly informative prior function on the target classes, such as the expected distribution of objects across the Universe.

The following section introduces the terminology and the classification metrics used in our evaluation, the probabilistic models used to identify extragalactic sources, and the prior functions we exploit to address the issue of class imbalance.

### 4.3.1 Terminology and metrics

In this section, we define key terms and the metrics used to assess a classifier performance. Classes may be defined as true and predicted. The true class refers to what has been defined in the training and testing datasets as the object's assumed class as defined by SDSS (for galaxies and quasars) or Gaia (for stars), and is therefore bound to have some inherent misclassification errors which will add noise to our classifier. The predicted class refers to the class that has been assigned using the probabilities outputted from our estimated classifier, which may be taking the maximum posterior probability or by considering a probability threshold. We define our predicted class as being the maximum posterior probability for a given source. To compare the predicted and true classes, we construct a confusion matrix, where entry row $i$ and column $j$ of the matrix refer to the number of objects with the true class $i$ classified into predicted class $j$. The confusion matrix is of dimension $K \times K$, with $K^2$ numbers when classified using the maximum posterior probability.

During training, we seek to minimise a loss function and monitor the performances of the model across all iterations using an evaluation metric. In multiclass classification, the standard loss function is the cross-entropy, defined in Eq. 4.1, for which an ideal model would be able to correctly predict all objects (i.e. a cross entropy loss value equal to 0), in contrast with the opposite case of a larger value when the predictions diverge from the true class:

$$\text{Cross entropy loss} = \frac{-1}{N} \sum_{n=1}^{N} \sum_{j=1}^{K} y_{nj} \log(p_{nj}), \tag{4.1}$$

where $N$ refers to the sample size, $K$ the number of classes, $y_{nj}$ the outcome equal to 1 for the true class and 0 otherwise, and $p_{nj}$ the probability that object $n$ belongs to class $j$.

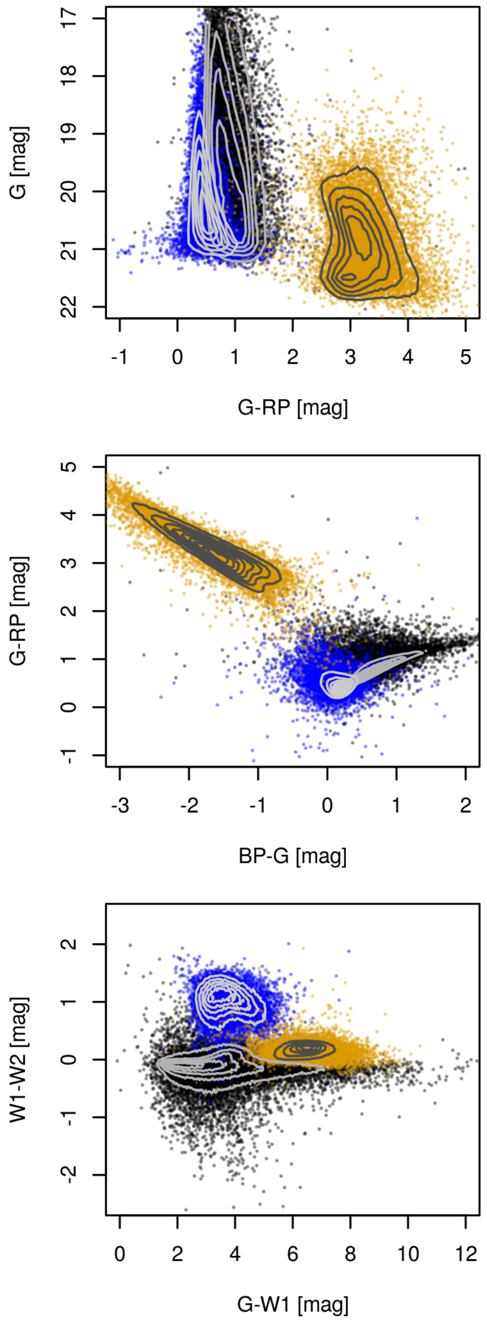


Figure 4.3: Colour–magnitude diagram (*top*) and two colour–colour (*middle and bottom*) diagrams highlighting the distribution of each class, with contours designating the density on a linear-scale for a random sample of 10 000 observations. The colour black corresponds to stars, blue to quasars, and gold to galaxies. Distinct aggregates can be identified for each class, although a significant interclass overlap still occurs.

Classification performances are evaluated on a dataset unseen during the training phase, that is, the test dataset. Performances are evaluated through metrics such as purity, completeness, and the F1-score. The purity, also known as precision (cf. Eq. 4.2), refers to the number of true positives (TPs) over the full count of objects in the target class. Purity can also be considered as a measure of contamination (1 - purity), representing the false positive (FP) rate. The higher the purity, the lower the contamination.

$$\text{Purity} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \tag{4.2}$$

The completeness, also known as the recall or sensitivity (cf. Eq. 4.3), refers to the number of TPs over the number of objects in the target class, that is, the total sum of correct predictions and true non-detections (false negatives (FNs)). A perfect model has purity and completeness both equal to 1.

$$\text{Completeness} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{4.3}$$

The F1-score is computed as the harmonic mean of a model's completeness and purity:

$$\text{F1} = 2 \times \frac{\text{Purity} \times \text{Completeness}}{\text{Purity} + \text{Completeness}}. \tag{4.4}$$

We define the objective function during training as the cross-entropy (Eq. 4.1), but we use the F1-score as the evaluation metric applied to the validation dataset for the statistical methods described in Sect. 4.3.2. A perfect model has an F1-score of 1.

We report the completeness and purities in the discussion of each of our classifiers, as these metrics are of the greatest interest when considering the rare classes, quasars and galaxies, and because these objects are harder to classify in comparison to the large number of stars observed in Gaia.

### 4.3.2 Statistical methods

Gaussian mixture models and gradient-boosting methods have been shown to be effective in numerous classification tasks, such as the works by Lee et al. (2012); de Souza et al. (2017) and Möller et al. (2016); Chao et al. (2019); Golob et al. (2021), respectively. In the current section, we describe both methods as well as their known limitations when applied to our classification problem.

#### Gaussian mixture models

GMMs (Fraley and Raftery, 2002), as used in the work by Bailer-Jones et al. (2019) for the supervised classification of extragalactic sources in Gaia, are defined in this work as our baseline classifier. In the training phase, for each class of the training set, the GMMs fit the distribution of the data as a sum of $M$ Gaussians defined in a multi-dimensional feature space by maximum likelihood. In the prediction phase, for an unclassified object, the trained classifier computes a probability density function normalised for each class to provide posterior class probabilities, which nominally is equivalent to adopting an equal class prior. The final class prediction is obtained from the highest posterior probability across all classes.

GMMs are known to reach their limitations when dealing with overlapping classes and sparse data. To prevent such limitations, we introduced an adjustment to the likelihood by setting a fraction $n$ of the lowest values to zero, which sets the final densities computed by the GMMs to zero. By forcing the Gaussian distributions to truncate to zero, sources at the boundaries of a class distribution (the potential overlap between classes) should be directly assigned to the most prevalent class (for our purposes, the star class), resulting in an increase in the purity and completeness of the rarer classes. We considered four different values for the threshold value $n$ of {1,5,20,50} applied to all models trained on the different input configurations (i.e. with and without infrared features). We found that the purity in the quasar and galaxy classes marginally improves (an increase of 0.02) when using $n = 50$ for the model trained without the infrared features compared to a standard GMM. However, this correction does not induce any improvement for the models trained on the dataset including infrared features. Furthermore, we find that the GMMs subject to the likelihood trimming method perform better than the standard GMM classifier, but attain lower performances compared to the boosted decision tree methods. We therefore do not consider the correction via likelihood trimming any further.

### Gradient boosting methods

Gradient boosting is a popular and powerful ensemble technique within supervised machine learning, where the ensemble technique refers to building a model from a collection of weaker learners. There are additional ensemble methods such as bagging, which splits the dataset into $N$ subsets with replacement, builds a model on each subset in parallel, and finally combines their individual predictions to compute the final class. Bagging is the basis of the random forest method (Breiman, 2001). By contrast, gradient boosting builds a model by sequentially fitting the weak learners in order to correct the residual errors at each iteration. The algorithm re-weights the data towards the most difficult cases at each training step, such that subsequent learners prioritise solving them. Typically, the learners used in gradient-boosting methods are decision trees, and the method is known as gradient-boosting decision trees (GBDT;Friedman (2001)).

The extreme gradient-boosting (XGBoost) method is a boosting algorithm presented by Chen and Guestrin (2016) that refers to one of the fastest implementations of GBDT. In particular, XGBoost improves upon GBDT in that it includes the second derivative of the loss function, which provides complementary information on the direction of gradients essential to solving the optimisation problem. Furthermore, the XGBoost method includes L1 and L2 regularisation used to prevent the model from overfitting.

A second gradient-boosting method, used in our work, is categorical boosting (CatBoost;Dorogush et al. (2017)). The key properties of CatBoost (which are lacking in XGBoost) are balanced (symmetric) trees and ordered boosting. Balanced trees, by definition, are built such that, at every step, the trees are split using the same feature criterion. By using a balanced tree architecture, CatBoost runs more efficiently and controls for overfitting as the balanced tree serves as regularisation. In general, classic boosting methods are prone to overfitting and CatBoost circumvents this limitation via ordered boosting, which refers to the process of training a model on a subset of the data and computing the residuals on a different subset.

In the following, we train two classifiers, using XGBoost and CatBoost, with the same set of input features used to train the GMM in order to assess the classification performances of all classifiers. We select the optimal hyperparameters by performing a five-fold grid search cross-validation that minimises the cross entropy loss function for finding the best hyperparameters. We then maximise the evaluation metric, the F1 score, when fitting the model with the best hyperparameters on the validation set. The hyperparameters we choose to optimise are *max_depth* or *depth* (in the case of CatBoost), *learning_rate*, and *n_estimators*, with the remaining set at their default values. Here, *max_depth* represents the maximum number of nodes allowed on a tree and is used to control for over-fitting, as a higher depth will make the model more complex and representative of the training dataset and thus more likely to be overfitted. The *max_depth* parameter ranges from zero to infinity and we consider the values of 3, 6, 8, and 10. The *learning_rate* is the step size taken by the model at each iteration to reach the minimum of the loss function; it takes a value of between 0 and 1, and is used to control for overfitting by modifying the weights of new trees added to the model. We consider the values 0.01, 0.03, 0.05, and 0.1. The last hyperparameter we consider tuning is the number of trees, specified by *n_estimators*. There is often a point of diminishing returns once there is a large number of trees, and each subsequent tree barely reduces the loss function. We considered the values of 100, 500, and 1000 for the *n_estimators* in our testing. The optimal values for the hyperparameters obtained from our grid search are a *max_depth* of 8, a *learning_rate* of 0.1, and a total number of trees *n_estimators* of 100 for XGBoost, and a *depth* of 6, a *learning_rate* of 0.03, and a total number of trees *n_estimators* of 1000 for CatBoost.

### 4.3.3 Prior

The class imbalance problem, that is, when the class distributions are highly skewed and we are interested in the least frequent class, is not unique to classification within astronomy. The problem is often encountered in various areas such as credit fraud detection where fraud is considerably less frequent than regular transactions. Multiple classification algorithms in this context attain low predictive accuracy for the rare class. Several data augmentation methods have been developed to address the imbalance problem from oversampling the rare classes, undersampling the most prevalent class, and generating synthetic observations using techniques such as SMOTE (Chawla et al., 2002). In this work, we attempt to correct for the class imbalance by applying a model correction exploiting prior knowledge that can be physically attributed, as introduced in Bailer-Jones et al. (2019). Lake and Tsai (2022) offer a similar approach which likewise proposes replacing the implicit prior of the classifier with one representative of the target population. The model correction is applied in two phases of the modelling process. First by adjusting the posterior probabilities by the class prior, as described in Eq. 4.5, and second via the modification of the confusion matrix by an adjustment factor, $\lambda_k$, shown in Eq. 4.6. The approach is thoroughly explained in section 3.4 of Bailer-Jones et al. (2019), but we summarise the key points in the following section for convenience.

First, the prior adjustment is done by re-weighting the posterior probabilities using a prior distribution to reflect the expected real class distribution:

$$P(C_k|x,\Theta) = \frac{1}{Z}\pi_k P(x,|C_k), \tag{4.5}$$

where $\Theta$ refers to any prior information, $Z = \sum_k \pi_k P(x,|C_k)$, and $\pi_k$ is the class prior for class $k$.

Second, when applying the model to a test dataset, the confusion matrix is modified using the adjustment factor in Eq. 4.6. This approach ensures that the results reflect the expected (prior) distribution of all classes, in particular the larger number of potential star contaminants to the quasar and galaxy classes. This step is necessary because the actual test dataset generally does not portray the class distribution expected in reality; in particular, it will tend to have too few stellar contaminants. The factor $\lambda_k$ scales the actual number of objects in each row to the number of objects expected within a dataset. Given the definition of the adjustment factor, the correction only affects the purity and not the completeness estimated from the confusion matrix.

$$\lambda_k = \frac{\pi_k}{\alpha_k} \left( \sum_{k'} (\frac{\pi_{k'}}{\alpha_{k'}}) \right)^{-1}, \tag{4.6}$$

where $\alpha_k$ is the class fraction within a dataset.

In this work, we describe three different priors and apply two of them: first the global prior reflecting a general class distribution, then a joint prior dependent upon latitude and magnitude, and lastly a mixed prior that combines the two aforementioned priors.

### The global prior

The global prior of $\{\pi^{GP}_{star}, \pi^{GP}_{quasar}, \pi^{GP}_{galaxy}\}$, which was introduced in the work by Bailer-Jones et al. (2019), outlines the scarcity of quasar and galaxy objects compared to stars across the sky. The prior sets the probabilities of observing a quasar or a galaxy to $1/1000$ and $1/5000$, respectively, from a sample of extragalactic sources with parallaxes and proper motion measurements. However, as discussed in Sect. 4.2.1, the majority of the galaxies observed in Gaia lack reported parallaxes and proper motions. To define our global prior, we count the number of sources across each class in the Stripe82 region from SDSS DR16, and extrapolate the distribution across the entire sky. The SDSS region Stripe 82 was chosen given the large sample of spectroscopic observations available for the majority of sources, thus providing a more complete count of identified targets. We find twice the number of galaxies compared to quasars and based on this we define our global prior as $1/1000$ for quasars and re-adjusted to $1/500$ for galaxies.

### The joint latitude and magnitude prior

Extragalactic sources are expected to have an intrinsic uniform distribution across the sky, but will not be observed due to dust extinction in the disk. At low latitudes closer to the Galactic plane, we would expect a higher number density of stars in comparison to galaxies and quasars. As the (absolute) latitude increases, the number density of galaxies and quasars with respect to stars increases. This information can be used to generate a latitude-based prior derived from densities at different latitudes.

We can also construct a prior based on apparent magnitude as we would expect the number of quasars and galaxies to increase towards the fainter brightness end. The G-band magnitude distribution in Fig. 4.2 supports this expectation. Exploiting such characteristics in the latitude and the G magnitude distributions, we have the functionality to represent what we consider to be true variations in latitude and magnitude as a prior to improve the performance of our classifier over a two-dimensional (2D) latitude and magnitude space.

To construct the joint class prior, we chose the overlapped region $50^\circ <= l <= 200^\circ$ in Gaia and SDSS DR16 in order to ensure that we are counting sources over the same area of the sky. We assume that SDSS DR16 includes all galaxies and quasars in this region, and that Gaia includes all stars within. Here the denomination 'all' refers to a randomly generated application data set (i.e. our randomly selected 50 million GDR3 sources). Using the compiled list of sources, we now further define bins in both $sinb$ and $G - mag$, count the number of stars, quasars, and galaxies and finally normalise in order to compute frequencies. The distributions for the different class priors can be seen in Fig. A.7. The top panel shows a large number of stars within the plane and a lower density of stars at higher latitudes and towards lower magnitudes. The middle panel reports the distribution observed for the quasars, for which the lowest density is identified within the lowest latitude bin, and a majority of quasar sources at $G = 18\,\mathrm{mag}$ and higher latitudes. For galaxy sources, the bottom panel reports the majority of sources at the highest magnitudes and those uniformly distributed across latitudes excluding the lowest latitude regions.

### The mixed prior

The mixed prior refers to the latitude- and magnitude-dependent prior that accounts for the overall sky distribution of classes represented by the global prior. We define the mixed prior as follows.

1. $g_S, g_Q, g_G \simeq (1, \frac{1}{1000}, \frac{1}{500})$, the (un-normalised) target global prior.

2. $F_S, F_Q, F_G$ are the measured fractions of sources by star, quasar, and galaxy class in the overlap of SDSS and Gaia over the region $50^\circ <= l <= 200^\circ$, over all latitudes and magnitudes.

3. In a specific latitude and magnitude bin, the number of sources from each class is counted to be $n_S, n_Q, n_G$.

4. The number of sources we should have in each latitude and magnitude bin according to our target prior are therefore
$n_S^{'} = n_S \frac{g_S}{F_S}$
$n_Q^{'} = n_Q \frac{g_Q}{F_Q}$
$n_G^{'} = n_G \frac{g_G}{F_G}$ .

5. Normalising these across the classes gives the target prior for each latitude and magnitude bin:
$n_S^{''} = \frac{n_S^{'}}{n_S^{'}+n_Q^{'}+n_G^{'}}$
$n_Q^{''} = \frac{n_Q^{'}}{n_S^{'}+n_Q^{'}+n_G^{'}}$
$n_G^{''} = \frac{n_G^{'}}{n_S^{'}+n_Q^{'}+n_G^{'}}$.

The distribution of this prior across latitude and magnitude is shown in Fig. 4.4. We see the dominance of stars in the lower latitudes and a gradual increase in the prevalence of quasars and galaxies at higher latitudes and fainter magnitudes. As the prior is discontinuous in magnitudes $G$ and latitudes $b$, we expect discontinuities in the classification probabilities and counts.

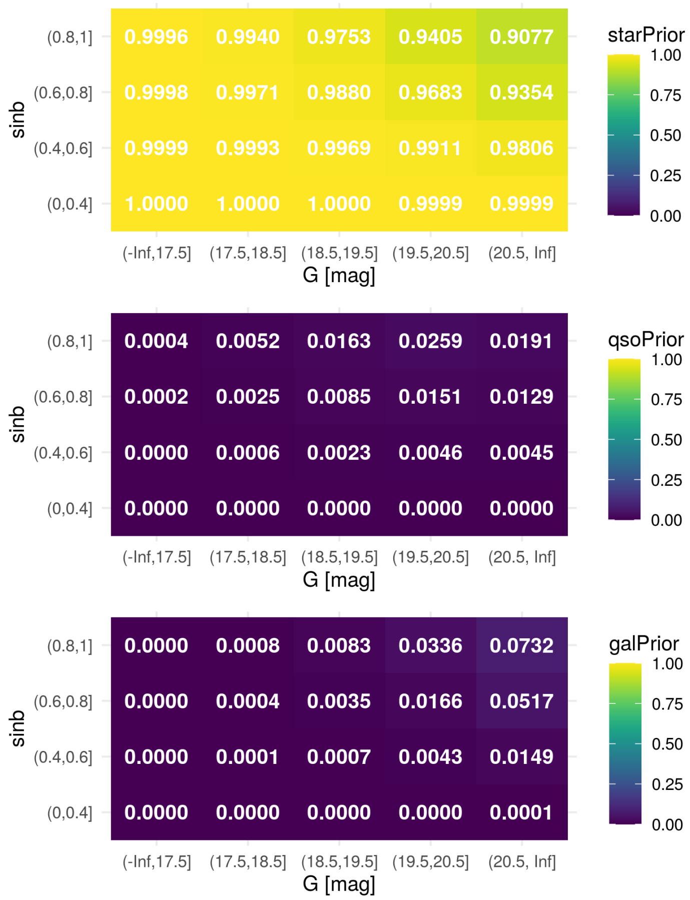


Figure 4.4: Heat map of the mixed prior distribution. In this representation, the number of stars at lower latitudes exceeds the number of observed quasars and galaxies. Whereas, at higher latitudes and fainter magnitudes, the number of quasars and galaxies surpasses the number of stars. Values of '0.0000' are not necessarily exactly zero, but could be below the numerical precision shown.

## 4.4 Results of different models and feature combinations on the test set

Section 4.4.1 presents the results of classification obtained with four different feature combinations using the GMM, XGBoost, and CatBoost methods applied to the balanced data set (for training and testing). We identify the best feature combination and method for the classification of extragalactic sources. Section 4.4.2 shows the results of the chosen model and feature combination fitted and assessed on the larger imbalanced training and test datasets, respectively. The effect of applying the priors to the model probabilities is discussed in Sect. 4.4.3. The selected classifier is applied to our application datasets in Sect. 4.5. Our tests were run on an Ubuntu server, with 344GB of RAM and an Intel Xeon CPU E5-2695 v3 at 2.30 GHZ with 30 threads.

Table 4.1: Classification performances obtained for different balanced classifiers using different algorithms and input features. Completeness and purity are shown for each class. From our tests, the best performing model is the XGBoost algorithm trained on the Gaia_f features supplemented with the infrared CatWise2020 colours (*Feature Set 4*)

| | | Completeness | | | Purity | | |
|---|---|---|---|---|---|---|---|
| | Features | Star | Quasar | Galaxy | Star | Quasar | Galaxy |
| GMM | Gaia_f | 0.9330 | 0.9580 | 0.9886 | 0.9532 | 0.9405 | 0.9860 |
| | Gaia_f + W1-W2 | 0.9714 | 0.9850 | 0.9906 | 0.9810 | 0.9784 | 0.9875 |
| | Gaia_f + W2 + G-W1 | 0.9766 | 0.9871 | 0.9919 | 0.9853 | 0.9837 | 0.9866 |
| | Gaia_f + W1-W2 + G-W1 | 0.9778 | 0.9859 | 0.9919 | 0.9840 | 0.9846 | 0.9869 |
| XGBoost | Gaia_f | 0.9418 | 0.9623 | 0.9922 | 0.9603 | 0.9489 | 0.9871 |
| | Gaia_f + W1-W2 | 0.9728 | 0.9878 | 0.9933 | 0.9857 | 0.9798 | 0.9885 |
| | Gaia_f + W2 + G-W1 | 0.9793 | 0.9896 | 0.9932 | 0.9878 | 0.9859 | 0.9884 |
| | Gaia_f + W1-W2 + G-W1 | 0.9793 | 0.9908 | 0.9936 | 0.9891 | 0.9857 | 0.9889 |
| CatBoost | Gaia_f | 0.9411 | 0.9619 | 0.9919 | 0.9593 | 0.9484 | 0.9872 |
| | Gaia_f + W1-W2 | 0.9720 | 0.9876 | 0.9930 | 0.9854 | 0.9787 | 0.9885 |
| | Gaia_f + W2 + G-W1 | 0.9785 | 0.9883 | 0.9927 | 0.9862 | 0.9847 | 0.9886 |
| | Gaia_f + W1-W2 + G-W1 | 0.9786 | 0.9905 | 0.9934 | 0.9886 | 0.9850 | 0.9888 |

Table 4.2: Classification performances adjusted by the global prior and adjustment factor for different balanced classifier models using the XGBoost algorithm and different input features.

| | | Completeness | | | Purity | | |
|---|---|---|---|---|---|---|---|
| | Features | Star | Quasar | Galaxy | Star | Quasar | Galaxy |
| XGBoost | 1: Gaia_f | 0.9992 | 0.0844 | 0.3625 | 0.9978 | 0.5295 | 0.4915 |
| | 2: Gaia_f + W1-W2 | 0.9987 | 0.2653 | 0.4261 | 0.9981 | 0.4209 | 0.4768 |
| | 3: Gaia_f + W2 + G-W1 | 0.9986 | 0.3402 | 0.4464 | 0.9982 | 0.4197 | 0.4924 |
| | 4: Gaia_f + W1-W2 + G-W1 | 0.9986 | 0.3289 | 0.4650 | 0.9983 | 0.3944 | 0.5054 |

### 4.4.1 Classifier trained on a balanced set

By considering a balanced class distribution across the training and test datasets with 200 000 sources in each class, as defined in Sect. 4.2.2, the intrinsic performances of each method applied to different input feature combinations are higher compared to those obtained when we subsequently apply the appropriate prior and allow for a higher level of contamination from stellar objects.

Table 4.1 reports the different methods GMM, XGBoost, and CatBoost, where each model is tested using four combinations of input features: *Feature Set 1*: Gaia_f, *Feature Set 2*: Gaia_f + W1-W2, *Feature Set 3*: Gaia_f + W2 + G-W1, and *Feature Set 4*: Gaia_f + W1-W2 + G-W1.

When performing model fitting, we search for the best input configuration with the highest purity and completeness in the quasar and galaxy classes. We add the colour difference of

Table 4.3: Classifier using XGBoost with two feature set configurations as applied to the imbalanced test dataset. Classification performances of the model trained on an imbalanced dataset show lower completeness but higher purity compared to the balanced classifier. Classification performances increase when the infrared data are incorporated as input features.

| | Completeness | | | Purity | | |
|---|---|---|---|---|---|---|
| Features | Star | Quasar | Galaxy | Star | Quasar | Galaxy |
| Gaia_f | 0.9714 | 0.9040 | 0.9914 | 0.9766 | 0.9091 | 0.9759 |
| Gaia_f + W1-W2 + G-W1 | 0.9858 | 0.9799 | 0.9922 | 0.9937 | 0.9705 | 0.9784 |

Table 4.4: Confusion matrix on the test set predictions using an XGBoost classifier trained on *Feature Set 4*. The right half of the table has been modified by the adjustment factor.

| | | Predicted | | | | | |
|---|---|---|---|---|---|---|---|
| | | Star | Quasar | Galaxy | Star | Quasar | Galaxy |
| Actual | Star | 887262 | 5127 | 7611 | 199119.6 | 100.6979 | 181.4556 |
| | Quasar | 3529 | 195980 | 491 | 133.8076 | 65.5892 | 0.0050 |
| | Galaxy | 2076 | 820 | 367104 | 213.3360 | 0.0260 | 185.4417 |

W1-W2 and G-W1 for long colour wavelength span to the original Gaia features. We do not consider colours such as BP-W2, as W2 is less sensitive than W1 and G has a higher signal-noise ratio than BP. From Table 4.1, we can see that across all feature combinations, the GMM has a lower classification performance for the two extragalactic classes compared to the gradient boosted methods. Moreover, the addition of infrared-derived features increases the purity and completeness for the quasar and galaxy classes. In Table 4.2, we apply a global prior and the adjustment factor, and report the purity and completeness for each class. We find that Feature sets 3 and 4 give comparable performances, and are better than sets 1 and 2. Given Table 4.1 and Table 4.2 we choose an XGBoost model with Gaia features and the W1-W2 and G-W1 infrared colours (*Feature Set 4*).

### 4.4.2 Classifier trained on an imbalanced set

Using the statistical model and features identified in Sect. 4.4.1, we now train a classifier using all available sources. This enables the design of a classifier that is more representative of the true class distribution, as discussed in Sect. 4.2.2. As introduced in Sect. 4.3.3, the global prior is set to 1, 1/1000, or 1/500, for star, quasar, and galaxy targets, respectively, which differs from the class fractions in the training and test sets. Given the available data, it would be infeasible to use this prior and have a representative number of objects in the extragalactic classes. Furthermore, the intrinsic prior of a model is not necessarily equal to the class fractions in the data initially used for training. Application of the adjustment to the posterior probabilities is discussed in Sect. 4.4.3.

The results of our classifier trained using *Feature Set 4* are reported in Table 4.3. We compare the final model with an XGBoost model trained exclusively on *Feature Set 1* and find a significant improvement in the completeness and purity for the quasar class, from 0.9040 to 0.9799 in the completeness, and from 0.9091 to 0.9705 in the purity. However, for the galaxy class, only an insignificant improvement is seen in the classification metrics, from 0.9914 to 0.9922 in the completeness and from 0.9759 to 0.9784 in the purity. Compared to the balanced classifier in Sect. 4.4.1, the current classifier exploits a larger dataset, and therefore the decrease in the classification performances is to be expected, particularly in terms of purity, which is due to the fact that the larger dataset likely has more contaminants. For the remainder of this work, we retain the classifier trained on the imbalanced dataset using *Feature Set 4* to assess the use of different priors applied to the models and apply the classifier to the application sets in Sect. 4.5.

### 4.4.3 Classifier adjusted using the priors

We now consider the effect of applying different prior probability distributions to the posterior probabilities estimated by the classifier in Sect. 4.4.2. In the figures discussed in this section, the left panels represent results obtained for the *Feature Set 1* model, whereas the right-panels report the results associated to the *Feature Set 4*, both with XGBoost.

The results using the global prior are reported in Table 4.5. The top half of the table ('Adj') shows results for a realistic level of stellar contamination by using the adjustment factor $\lambda_k$ in Eq 4.6; the bottom half shows raw unadjusted results, that is, with the lower level of contamination seen in the test set ('Unadj'). Using the global prior gives a lower completeness overall in comparison to the results obtained before applying the prior in Table. 4.3. On average, similar results are observed in the purity for the unadjusted case. Applying the adjustment factor results in lower purities across both the quasar and galaxy classes; however, the addition of infrared colour information clearly results in a better classifier in terms of performance.

Having assessed the impact of the global prior on the final classification. We now consider a more highly tuned prior, namely the 'mixed' prior introduced in Sect. 4.3.3, and assess the performance as a function of latitude and magnitude, while also applying the adjustment of the confusion matrix in order to incorporate the expected class fractions at each latitude and magnitude into the performance metrics. In Fig. 4.5, the completeness for the quasar class improves with higher latitudes and most significantly when we add infrared colour information as an input feature. However adding infrared data and moving to higher latitudes marginally improves the completeness in the galaxy class. As an illustration, there is an 18% increase in completeness for very faint quasars at high latitudes (top-right bin) and only a 0.7% increase in completeness for galaxies in the equivalent bin when adding infrared data. The purities for the quasar and galaxy classes are shown in Figs. 4.6 and 4.7, respectively. We would like to point out that the exact values of 1 and 0 are due to a rounding precision. The effect of the adjustment factor is reported in the top panels. For the quasar class, we observe a significant improvement in purity when adding the infrared colours, and as a function of latitude. For the galaxy class, the addition of infrared colours has only a marginal improvement on the purities as a function of latitude and magnitude. The application of the adjustment factor induces an expected decrease in purity for both the quasar and galaxy target classes.

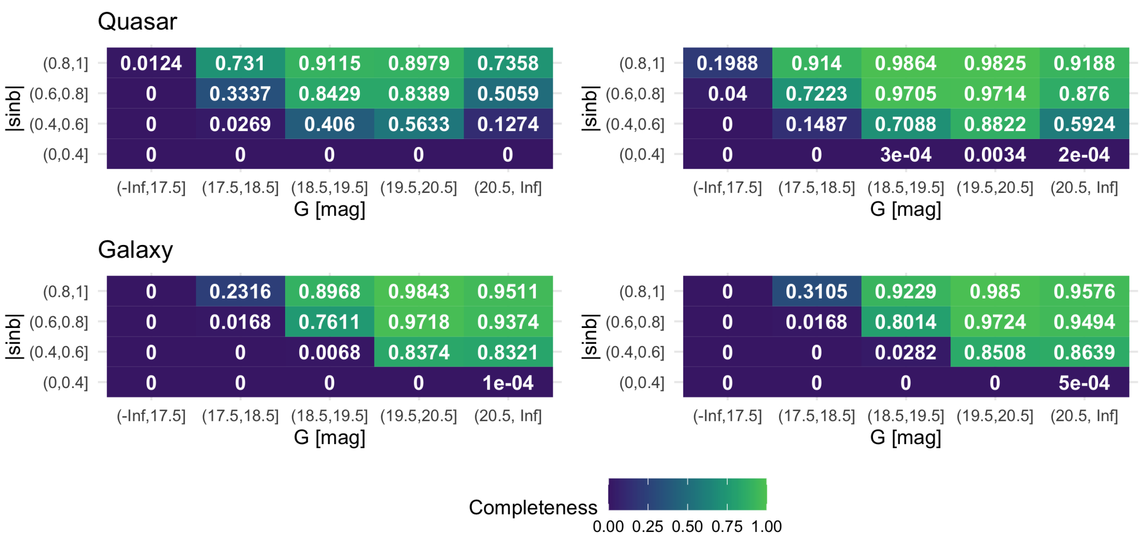


Figure 4.5: Imbalanced Classifier Mixed Prior: Completeness evaluated for the three target classes in the test set from predictions obtained by the best-performing models, i.e. XGBoost trained on *Feature Set 1* (left panel) and *Feature Set 4* (right panel).

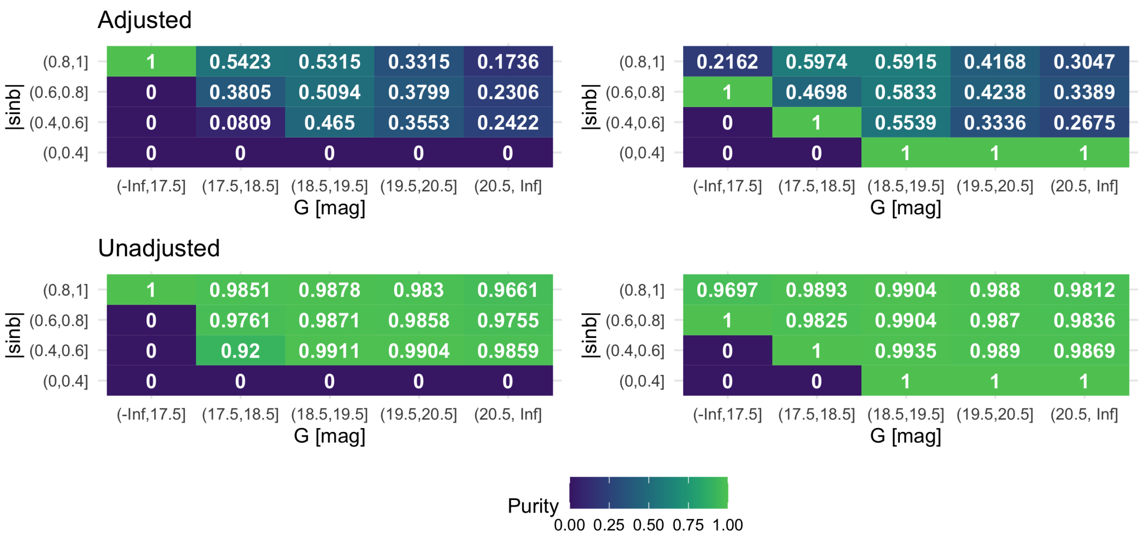


Figure 4.6: Imbalanced Classifier Mixed Prior: Purity evaluated for the quasar class in the test set from predictions obtained by the best-performing models, i.e. XGBoost trained on *Feature Set 1* (left panel) and *Feature Set 4* (right panel). Top panels show the classification performances modified by the adjustment factor. The near unit purity at low latitudes in the right panels is not meaningful as there are very few objects, as shown in Fig. 4.4

Table 4.5: Imbalanced Classifier Global Prior: Completeness and purity using the global prior as applied to the test dataset using the imbalanced classifier. 'Adj' is defined as adjusted using the adjustment factor, $\lambda_k$, in Eq. 4.6 and 'Unadj' is without such an adjustment.

| | | Completeness | | | Purity | | |
|---|---|---|---|---|---|---|---|
| | Features | Star | Quasar | Galaxy | Star | Quasar | Galaxy |
| Adj | Gaia_f | 0.9995 | 0.0897 | 0.3054 | 0.9977 | 0.4621 | 0.5958 |
| | Gaia_f + W1-W2 + G-W1 | 0.9993 | 0.2131 | 0.3790 | 0.9980 | 0.5694 | 0.6036 |
| Unadj | Gaia_f | 0.9995 | 0.0897 | 0.3054 | 0.6721 | 0.9946 | 0.9967 |
| | Gaia_f + W1-W2 + G-W1 | 0.9993 | 0.2131 | 0.3790 | 0.6991 | 0.9962 | 0.9968 |

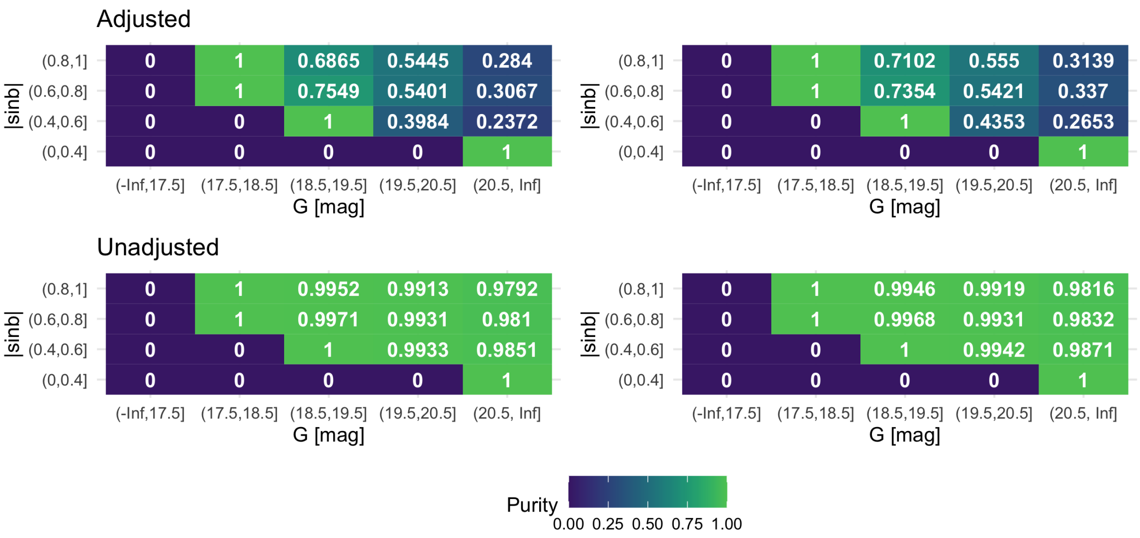


Figure 4.7: Similar to Fig. 4.6 for the galaxy target class in the test set.

## 4.5 Results of the best-performing model and feature combination on the application sets

To evaluate how our selected classifier performs and what distribution of the predicted classes is obtained on datasets with representative distributions, we apply the classifier to three datasets selected from the 1.8 billion sources observed in Gaia at the intersection between GDR3 and the CatWISE2020 catalogue. With our first application, we aim to predict the classes for a randomly selected subset of 50 million sources, without prior information on the target classes or their distribution. However, this application set has the distribution that our global and mixed priors are designed for. The second dataset is constructed from the GDR3 quasar and galaxy candidate tables defined in Gaia Collaboration et al. (2022), which are quoted as having purities of 0.52 and 0.69, respectively. The third data set is the purer subsample of the GDR3 quasar and galaxy candidate tables defined in Gaia Collaboration et al. (2022), which are quoted as having purities of 0.95 for the quasar class and 0.94 for the galaxy class. In addition to assessing the accuracy of our classifier, we wish to identify whether

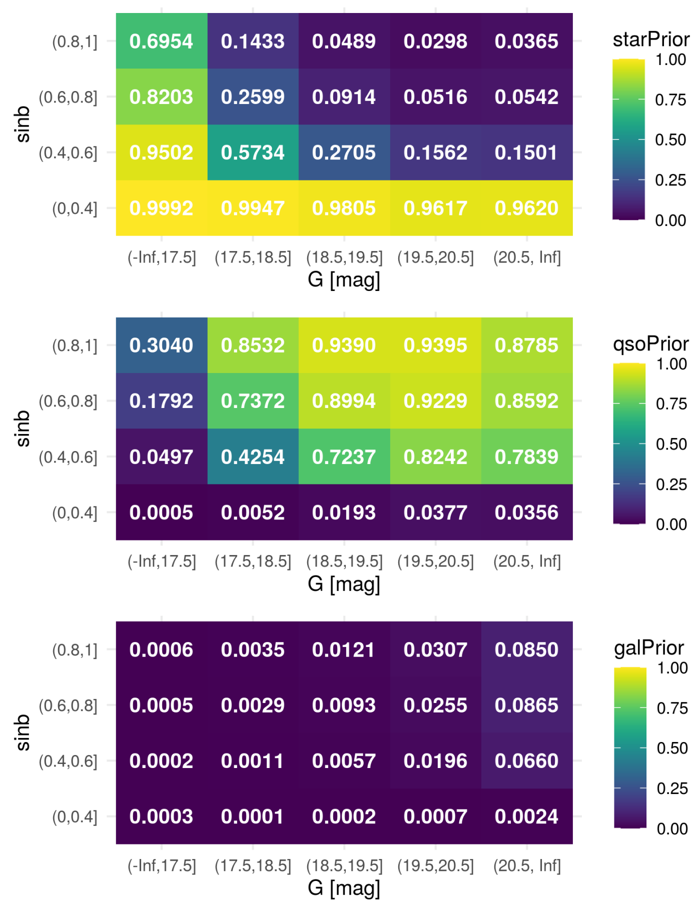


Figure 4.8: Heat map of the mixed prior distribution for the GDR3 Quasar Candidate Table. In this representation, the number of stars at the lowest latitude exceeds the number of observed quasars and galaxies.

adding infrared colours to Gaia data improves the reliability of these candidate tables, despite having removed parallax and proper motion as features.

Our priors, both global and mixed, are designed for a sample of sources drawn at random from the Gaia/CatWISE2020 all-sky sample. These priors are not appropriate for the classification of the GDR3 extragalactic tables in Sects. 4.5.2 and 4.5.3, where we have 50%–95% extragalactic objects, rather than 0.1%–0.2% as expected by the prior. For application to these, we redefine our global priors by taking the purity of each GDR3 extragalactic table as defined in (Gaia Collaboration et al., 2022), which we denote $p$. Considering the case of the quasar table, the normalised global prior becomes $(1-p-e, p, e)$, where $e$ is an estimation of the contamination from the galaxy class. The prior would be defined as $(1-p-e, e, p)$ in the case of the galaxy class. The normalised global priors are $(0.454, 0.520, 0.026)$ and $(0.274, 0.036, 0.690)$, with the re-adjusted mixed priors shown in Figs. 4.8 and 4.9 for the GDR3 quasar and galaxy candidate tables, respectively.

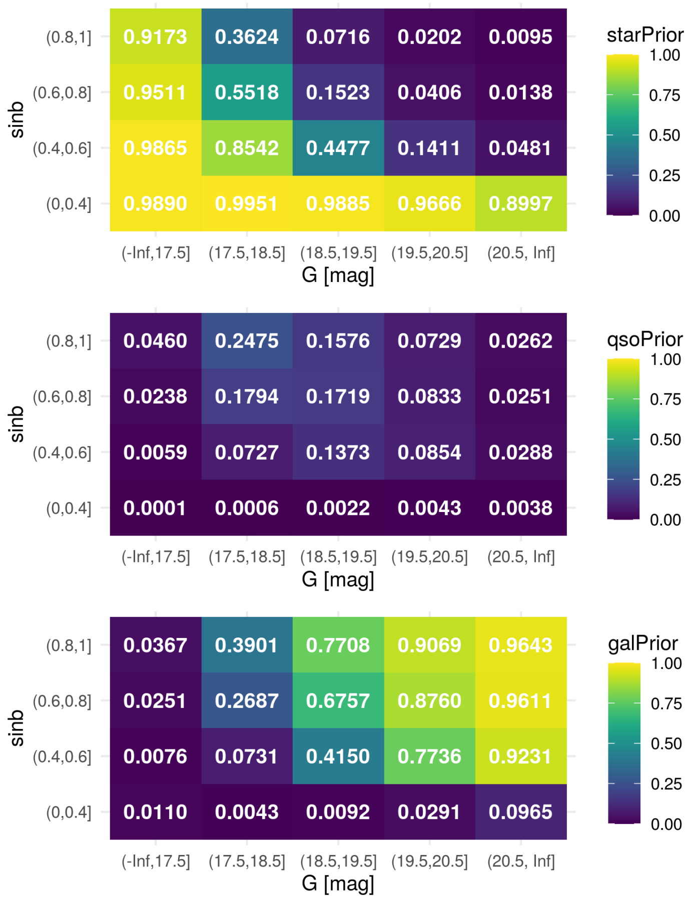


Figure 4.9: Heat map of the mixed prior distribution for the GDR3 Galaxy Candidate Table. In this representation, the number of stars at the lowest latitude exceeds the number of observed quasars and galaxies.

### 4.5.1 Application on a random subset of the overlap of GDR3 and CatWISE2020

For the 50 million sources at the intersection of GDR3 and CatWISE2020, the true class of the source is unknown and therefore reliable performance metrics cannot be computed. However, we can compare the number of sources classified with the different priors, and compare the counts to expectations. We find 12607 quasars and 41153 galaxies, or 1/4000 and 1/1200, using the global prior. When compared to the global prior values of 1/1000 for quasars and 1/500 for galaxies, we find that our results give a factor of 4 fewer quasars and a factor of 2 fewer galaxies. Using the mixed prior, we find 97294 quasars and 192231 galaxies, or 1/500 and 1/300. The mixed prior finds nearly eight times as many quasars (97294/12607 = 7.7) and roughly five times more galaxies (192231/41153 = 4.7) than the global prior. This may be attributed to the mixed prior being very non-uniform in magnitude and latitude, similar to the true distribution. Furthermore, by construction the mixed prior is better matched to the data.

In Fig. 4.10, we show the sky distributions of the sources by assigned class. For both priors, in a random sample of 50 million sources observed by Gaia, we classify less than 1% of the sample as either a galaxy or a quasar, highlighting the scarcity of the extragalactic sources.

It is interesting to compare our results with those used from the DSC-Combmod classifier, which was used to identify many quasars and galaxies published in the GDR3 extragalactic candidate tables (Gaia Collaboration et al., 2022). Combmod is the combination of the class posterior probabilities from two classifiers, DSC-Specmod and DSC-Allosmod Delchambre et al. (2022). Specmod classifies objects using BP/RP spectra, whereas Allosmod uses a GMM to classify objects using several astrometric and photometric features (the features being our Gaia_f set plus parallax and proper motion; see also Bailer-Jones et al. 2019). We use the quasar and galaxy class probabilities from Combmod, but take the star class probabilities to be one minus the sum of the quasar and galaxy probabilities (because Combmod reports more than three classes), and assign the class label to the class with the largest probability. When applying the global prior, we identify 7% of the Combmod quasars as quasars with the remaining 92.9% identified as stars and 0.1% as galaxies. We identify 21% of the Combmod galaxies as galaxies, with the remaining 78.9% identified as stars and 0.1% as galaxies. Using the mixed prior, we classify 40% of the Combmod quasars as quasars with the remaining 59.8% identified as stars and 0.2% as galaxies. For the Combmod galaxies using the mixed prior, we find 56% to be galaxies with the remaining fraction being 43.6% stars and 0.4% quasars.

We now refine the 50 million sources by considering those that are classified as a quasar or a galaxy in the pure samples defined in the GDR3 quasar and galaxy candidate tables, respectively. We aim to see whether the proportion of identified quasars and galaxies increases when the sample is refined. We find that our classifier identifies 12% of the quasars in the pure quasar candidate table using the global prior, an improvement of 5% compared to quasars classified in Combmod. Using the mixed prior, we identify 69% of the quasars in the pure quasar candidate table, which is over 25% better than the Combmod quasars. When considering the pure galaxy candidate table, we identify 18% as galaxies using the global prior which is a reduction of 3% when compared to the Combmod galaxies. A 2% reduction is seen when applying the mixed prior, where we identify 54% of galaxies in the pure galaxy candidate table. Using the three different classifications —DSC-Combmod, the pure subsample from the GDR3 candidate table, and our classifier— we illustrate the density of the predicted sources in colour–colour diagrams and a colour-magnitude diagram, with the contours representing the classifications from DSC-Combmod and the purer subsamples. Figure 4.11 shows the sources classified as quasars using the global prior in our classifier. We see that considerably fewer sources are classified as extragalactic when using the global prior compared to the mixed prior in Fig. 4.12, but are focused towards the redder magnitude. In contrast to the global prior, the mixed prior allows for more freedom in the identification of sources that are quasars, closely resembling the contours of the pure sample. Figures 4.13 and 4.14 represent the density of the galaxy class with the global prior and mixed prior adjustment, respectively. The global prior results are a subset of the mixed prior, with the mixed prior extending to bluer G-RP

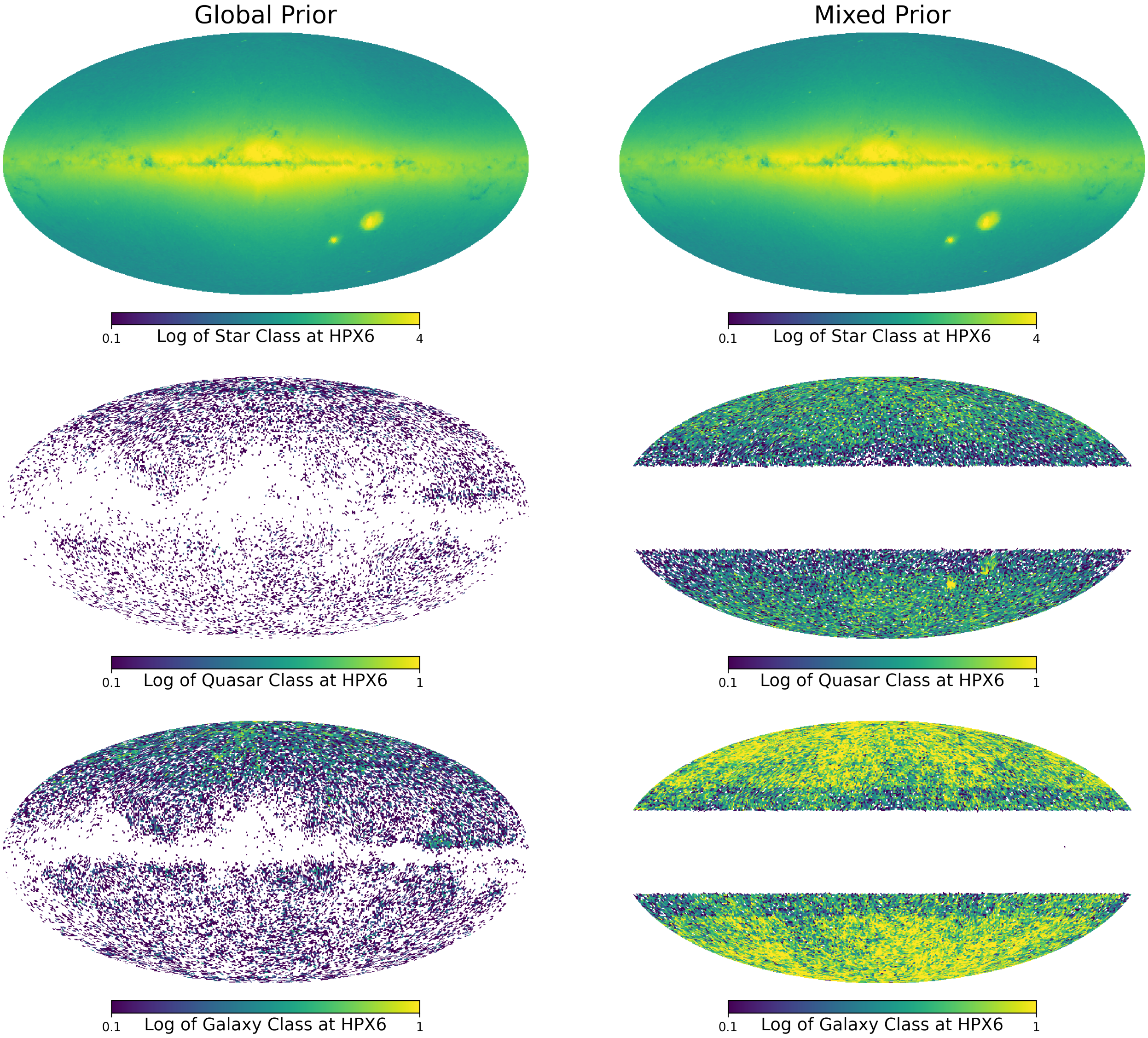

Global Prior
Mixed Prior
0.1
Log of Star Class at HPX6
4
0.1
Log of Star Class at HPX6
4
0.1
Log of Quasar Class at HPX6
1
0.1
Log of Quasar Class at HPX6
1
0.1
Log of Galaxy Class at HPX6
1
0.1
Log of Galaxy Class at HPX6
1

Figure 4.10: Log10 of counts for sources classified on the random Gaia DR3 sample on a healpix at level 6 (HPX6). As described in Sect. 4.3.3, the mixed prior is discretised by latitude and magnitude, this giving rise to the banded structure in the right panels. The white colour indicates a source density below the scale and anything above the scale is yellow.

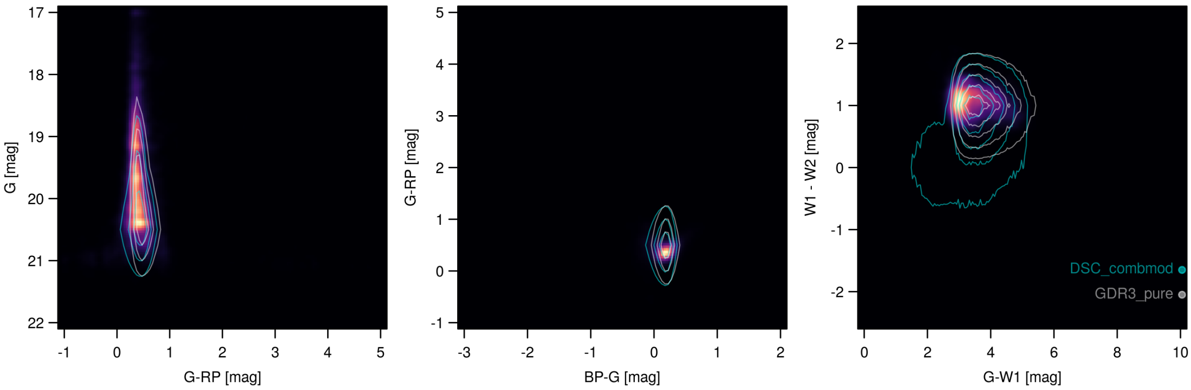


Figure 4.11: Results of the classifier when applied to the randomly selected set of 50 million Gaia DR3 and catWISE2020 sources using the global prior: Colour–magnitude and colour–colour diagrams for the quasars. Sources from the classifier adjusted by the global prior are given by the density scale where black is zero density and yellow is high density. DSC-Combmod sources are identified by the cyan contours and the GDR3-defined pure quasar sample by the white contours. This colour code is used for all subsequent colour–magnitude and colour–colour diagrams.

but the global prior not extending redder.

### 4.5.2 Application to quasar candidates from GDR3

The GDR3 quasar candidate table defined in Gaia Collaboration et al. (2022) contains 6.6 million potential quasars with a purity of 52%, and is further refined into a pure subsample containing 1.9 million quasars with a purity of 95%. The overlap with CatWISE2020 results in 4 048 626 GDR3 quasars and 1 822 922 pure subsample quasars.

We applied our trained classifier from Sect. 4.4.2 to the GDR3 quasar candidates that overlap with CatWISE2020 and estimated the probabilities of the three classes. We assess the classification performance of our model by considering the proportion of quasars identified by our classifier using the global prior and the mixed prior redefined for this application dataset (as explained at the beginning of this section), on the assumption that the quasar candidate overlap is entirely quasars.

The results are shown in Table. 4.6, where in the global prior case we identify 55% of quasars in the GDR3 candidate table. If we further constrain the sample by considering the pure subsample only, or the pure subsample and the SDSS16 quasar table together, we see the proportion of quasars identified by our classifier is considerably higher than the GDR3 candidate sample, reaching 99.8%. A similar trend is reported in the mixed prior case, this time identifying 58% of quasars in the GDR3 candidate table and 99.9% when restricting the sample to the pure subsample or the pure subsample plus the SDSS16 quasar table. Given

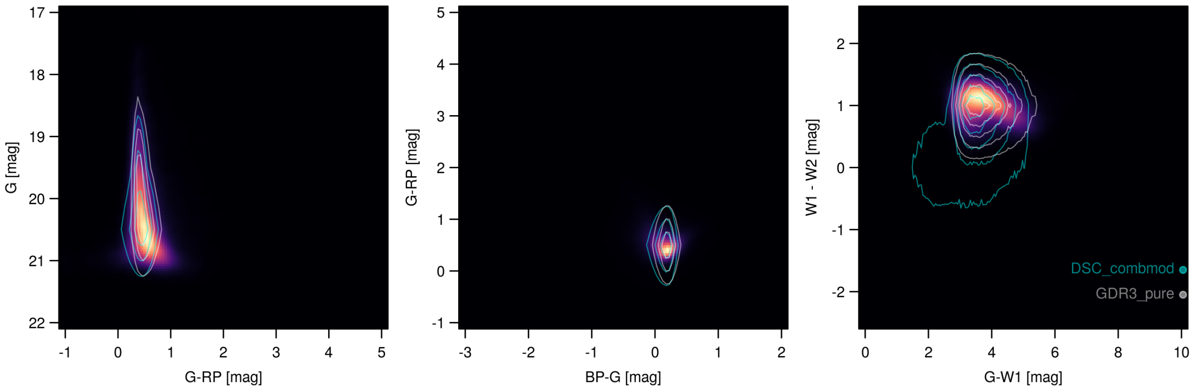


FIGURE 4.12: Results of the classifier when applied to the randomly selected set of 50 million Gaia DR3 and catWISE2020: As in Fig. 4.11, but using the mixed prior in our classifier.

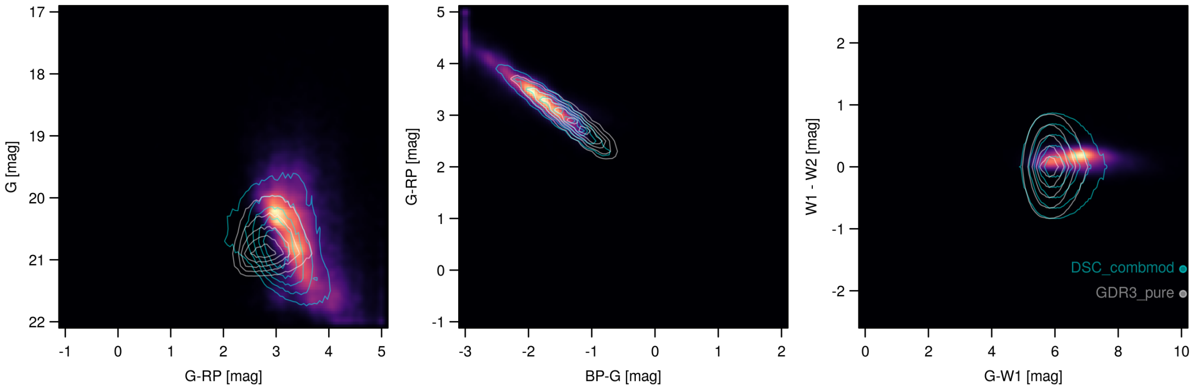


FIGURE 4.13: Results of the classifier when applied to the randomly-selected set of 50 million Gaia DR3 and catWISE2020: Colour–magnitude and colour–colour diagrams for the galaxies derived from DSC-Combmod, the GDR3-defined pure galaxy sample, and from the classifier adjusted by the global prior.

that both global prior and mixed prior have the same global prior behind them, adding the highly non-uniform distribution of the latitude and G dependencies to the prior makes it slightly more suited to finding quasars and galaxies where we expect to find them. We can deconstruct this results table further by considering the entire GDR3 quasar candidate sample in Fig. 4.15 for the global prior and Fig. 4.16 for the mixed prior. Comparing the two priors, we see a higher proportion of quasars identified in the fainter and higher magnitude end in the mixed prior case than the global prior, but the distribution on average is quite similar.

We visualise the application of the mixed prior to the quasar candidate table in Fig. 4.17, and the considerable overlap between the pure sample contours in the most dense region of the mixed prior classifier is evident. The same distribution can be seen in the case of the global prior.

By splitting the sample into two subsets based on the availability of parallax or proper motions in Fig. 4.18, we observe a higher density distribution for sources with parallaxes compared to the sources without parallax measurements. Furthermore, for the sources classified

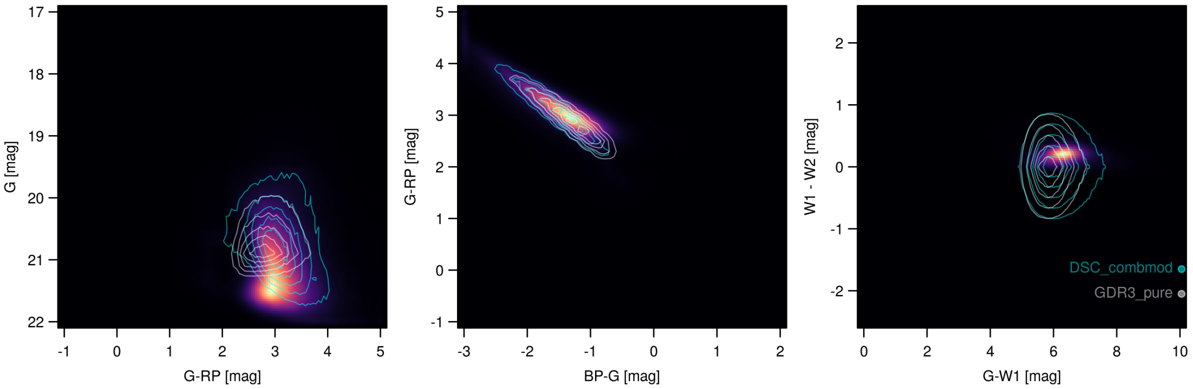
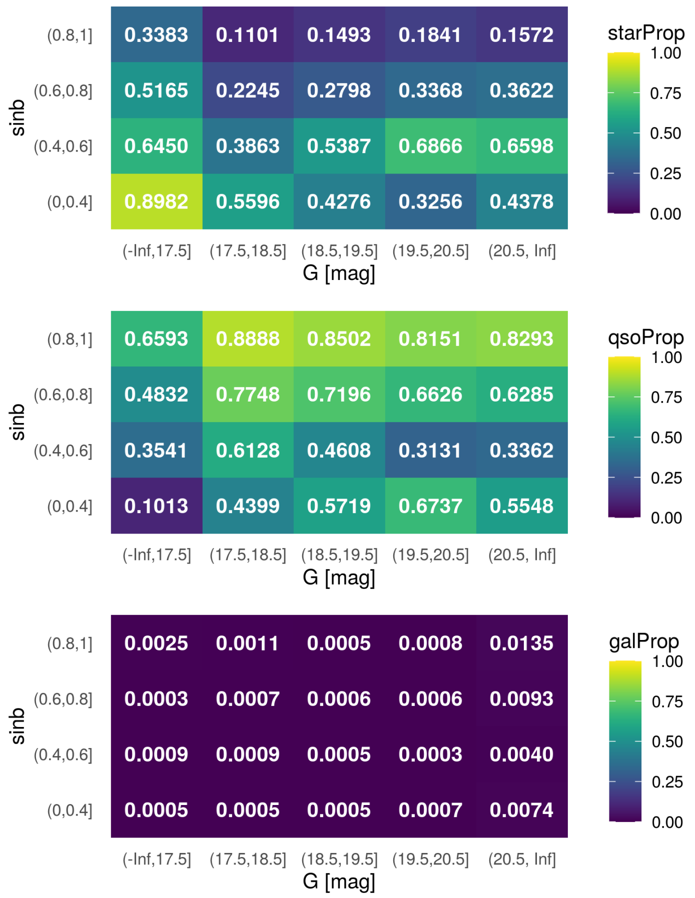


Figure 4.14: Results of the classifier when applied to the randomly-selected set of 50 million Gaia DR3 and catWISE2020: Colour-magnitude & Colour-colour diagrams for the galaxies using DSC-Combmod, the GDR3 defined pure galaxy sample and the classifier adjusted by the mixed prior.



| sinb | (-Inf,17.5] | (17.5,18.5] | (18.5,19.5] | (19.5,20.5] | (20.5, Inf] |
|---|---|---|---|---|---|
| (0.8,1] | 0.3383 | 0.1101 | 0.1493 | 0.1841 | 0.1572 |
| (0.6,0.8] | 0.5165 | 0.2245 | 0.2798 | 0.3368 | 0.3622 |
| (0.4,0.6] | 0.6450 | 0.3863 | 0.5387 | 0.6866 | 0.6598 |
| (0,0.4] | 0.8982 | 0.5596 | 0.4276 | 0.3256 | 0.4378 |

G [mag]

| sinb | (-Inf,17.5] | (17.5,18.5] | (18.5,19.5] | (19.5,20.5] | (20.5, Inf] |
|---|---|---|---|---|---|
| (0.8,1] | 0.6593 | 0.8888 | 0.8502 | 0.8151 | 0.8293 |
| (0.6,0.8] | 0.4832 | 0.7748 | 0.7196 | 0.6626 | 0.6285 |
| (0.4,0.6] | 0.3541 | 0.6128 | 0.4608 | 0.3131 | 0.3362 |
| (0,0.4] | 0.1013 | 0.4399 | 0.5719 | 0.6737 | 0.5548 |

G [mag]

| sinb | (-Inf,17.5] | (17.5,18.5] | (18.5,19.5] | (19.5,20.5] | (20.5, Inf] |
|---|---|---|---|---|---|
| (0.8,1] | 0.0025 | 0.0011 | 0.0005 | 0.0008 | 0.0135 |
| (0.6,0.8] | 0.0003 | 0.0007 | 0.0006 | 0.0006 | 0.0093 |
| (0.4,0.6] | 0.0009 | 0.0009 | 0.0005 | 0.0003 | 0.0040 |
| (0,0.4] | 0.0005 | 0.0005 | 0.0005 | 0.0007 | 0.0074 |

G [mag]

Figure 4.15: Heat map of the distribution for quasars identified using the global prior for the GDR3 Quasar Candidates as function of magnitude and latitude. Each mag/lat cell is normalised across the three classes.

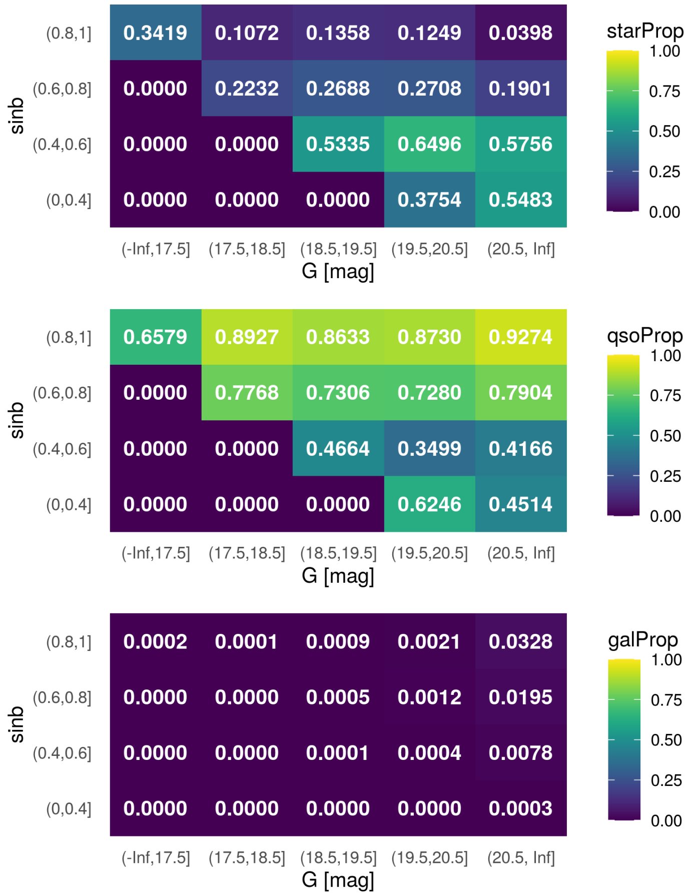


Figure 4.16: Heat map of the distribution for quasars as in Fig. 4.15 but using the mixed prior in our classifier.

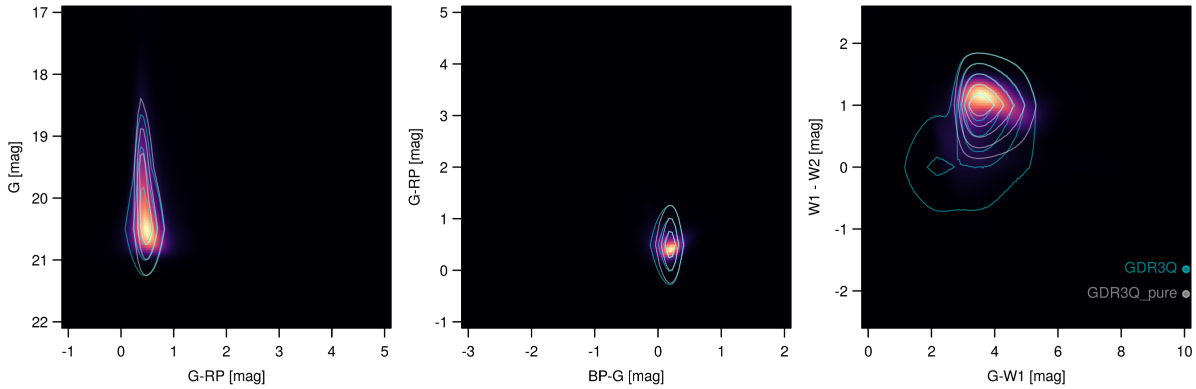


Figure 4.17: GDR3 Quasar Candidate Table Mixed Prior: Colour-magnitude & Colour-colour diagrams for the quasars using the mixed prior. The sources identified by the classifier are represented by the density scale, where black is zero density and yellow is high density. GDR3 quasar sources are identified by the cyan contours and the GDR3 pure quasar sample by the white contours.

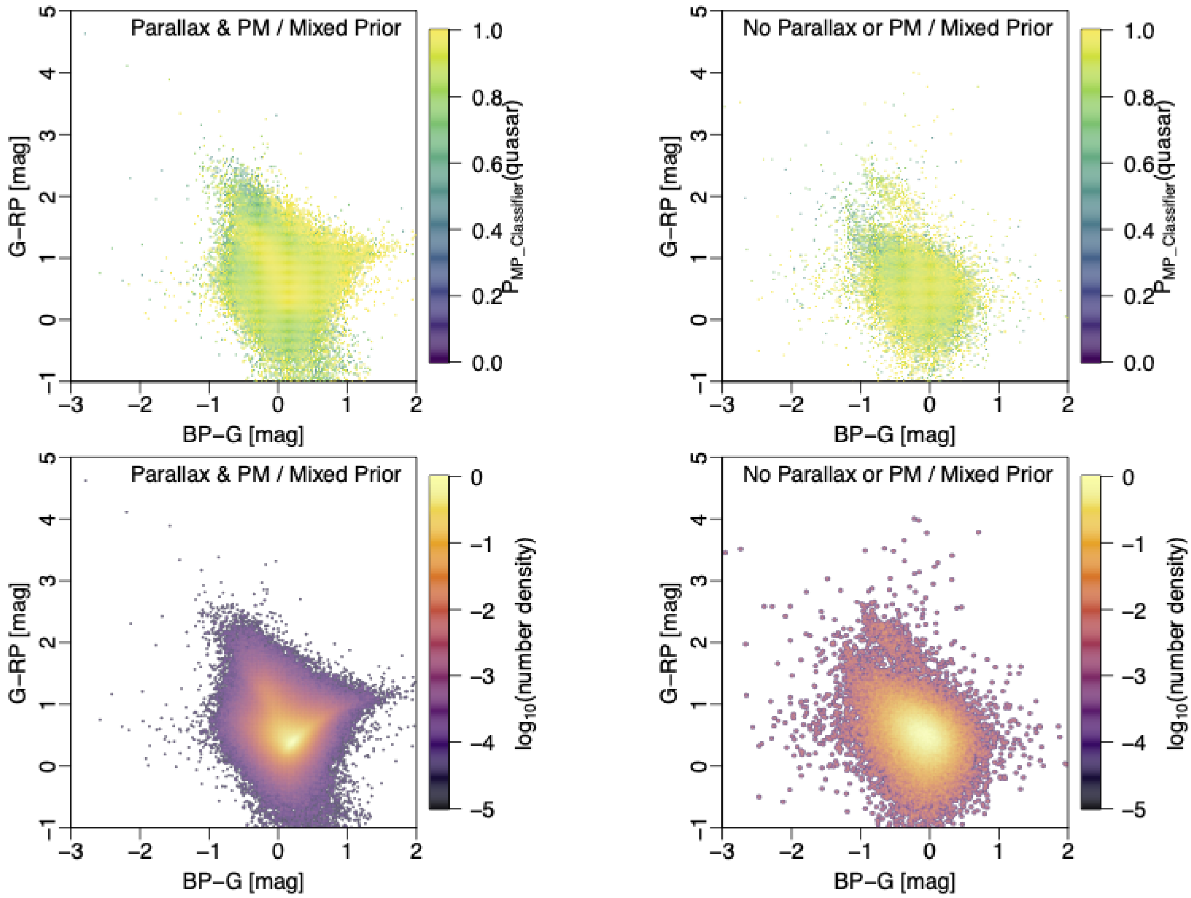


Figure 4.18: GDR3 Quasar Candidate Table Mixed Prior: Probability and density distributions for sources classified as a quasar. The left hand side panels correspond to sources with parallax while the right hand side panels represent the distribution for sources without parallax.

Table 4.6: Quasar candidates: Counts of objects in the predicted classes and the proportion identified as quasars using the extragalactic-table-tuned prior defined in Sect. 4.5. GP and MP refer to the global prior and mixed prior, respectively.

| | | Predicted | | | |
|---|---|---|---|---|---|
| | | Star | Quasar | Galaxy | Quasar proportion |
| GP | GDR3 Quasar | 1826019 | 2211696 | 10911 | 0.5463 |
| | Pure GDR3 Quasar | 54006 | 1768694 | 222 | 0.9703 |
| | SDSS16 Quasar + GDR3 | 753 | 401104 | 5 | 0.9981 |
| | SDSS16 Quasar + Pure GDR3 | 725 | 394418 | 5 | 0.9982 |
| MP | GDR3 Quasar | 1656379 | 2372430 | 19817 | 0.5860 |
| | Pure GDR3 Quasar | 68680 | 1753917 | 325 | 0.9621 |
| | SDSS16 Quasar + GDR3 | 491 | 401369 | 2 | 0.9988 |
| | SDSS16 Quasar + Pure GDR3 | 466 | 394680 | 2 | 0.9988 |

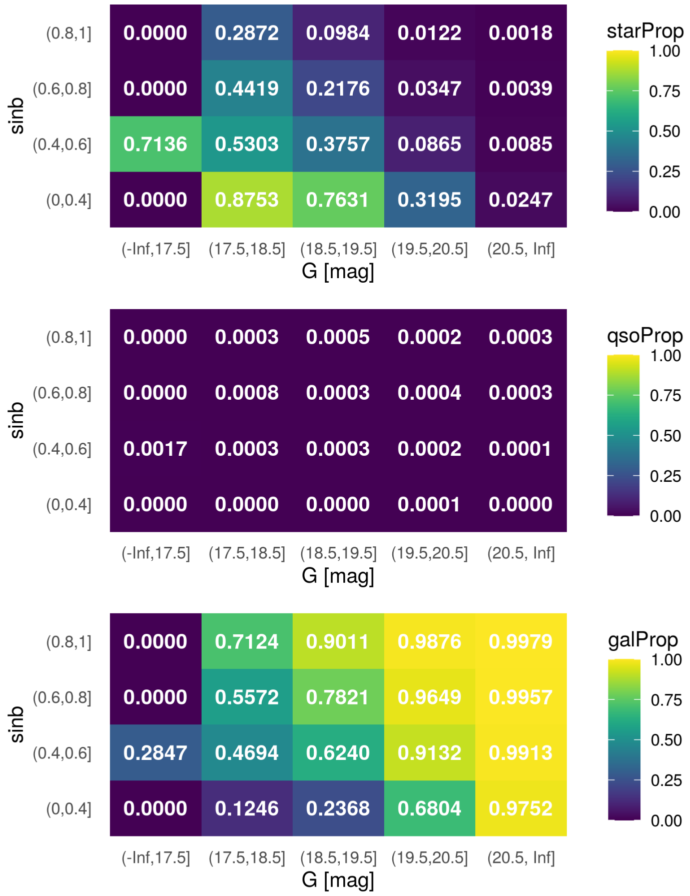
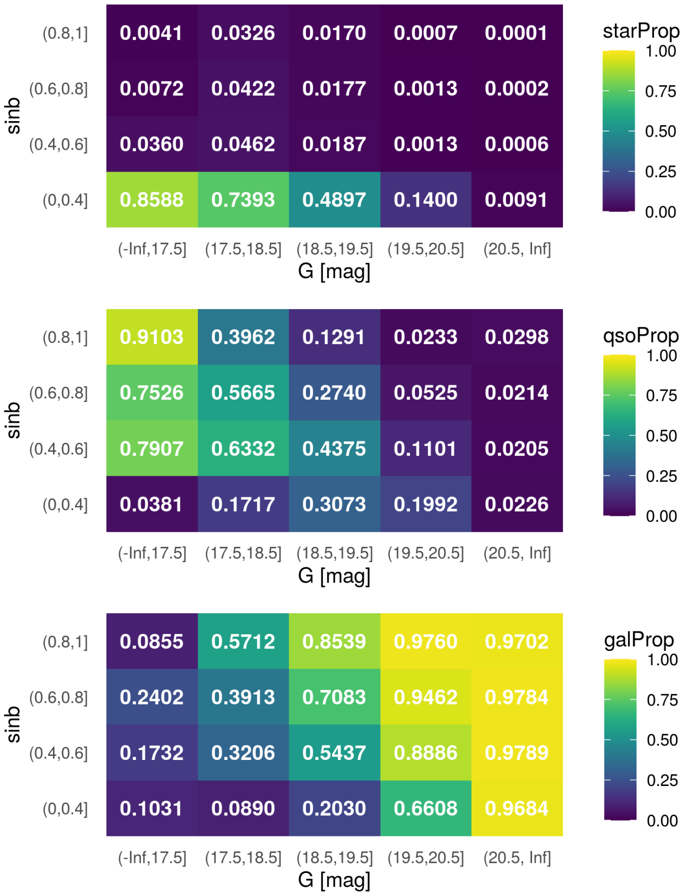


Figure 4.19: Heat map of the distribution for galaxies identified using the global prior for the GDR3 Galaxy Candidates by magnitude and latitude.

with parallax we see a shift in the density of the colour distribution, with more sources extending to $BP - G = 1$, whereas the sources without parallax and proper motions are centred around $BP - G = 0$ with a few outliers when $G - RP > 2$ for sources without parallax or proper motions in the case of the global prior. For the mixed prior, the distribution in colour space is similar; however, in the top-left panel for sources with $G - RP > 2$ the probabilities are lower than in the case of the global prior. Overall, the application of our classifier to the quasar candidates from GDR3 identifies 96%–97% of the pure quasar subsample as quasars. Moreover, when requiring the source to have a SDSS16 quasar classification, we identify 99.9% of them as quasars, irrespective of whether the source was in the GDR3 pure subsample or not.

### 4.5.3 Application to galaxy candidates from GDR3

In analogy to Sect. 4.5.2, here we apply our classifier to the galaxy candidate table in GDR3, which comprises 4.8 million candidates with a purity of 69%, and includes a purer subsample of 2.8 million candidates with a purity of 94%. The overlap with CatWISE2020 reduces the counts to 4 194 100 and 2 824 570, respectively.



| sinb | (-Inf,17.5] | (17.5,18.5] | (18.5,19.5] | (19.5,20.5] | (20.5, Inf] |
|---|---|---|---|---|---|
| (0.8,1] | 0.0041 | 0.0326 | 0.0170 | 0.0007 | 0.0001 |
| (0.6,0.8] | 0.0072 | 0.0422 | 0.0177 | 0.0013 | 0.0002 |
| (0.4,0.6] | 0.0360 | 0.0462 | 0.0187 | 0.0013 | 0.0006 |
| (0,0.4] | 0.8588 | 0.7393 | 0.4897 | 0.1400 | 0.0091 |

| sinb | (-Inf,17.5] | (17.5,18.5] | (18.5,19.5] | (19.5,20.5] | (20.5, Inf] |
|---|---|---|---|---|---|
| (0.8,1] | 0.9103 | 0.3962 | 0.1291 | 0.0233 | 0.0298 |
| (0.6,0.8] | 0.7526 | 0.5665 | 0.2740 | 0.0525 | 0.0214 |
| (0.4,0.6] | 0.7907 | 0.6332 | 0.4375 | 0.1101 | 0.0205 |
| (0,0.4] | 0.0381 | 0.1717 | 0.3073 | 0.1992 | 0.0226 |

| sinb | (-Inf,17.5] | (17.5,18.5] | (18.5,19.5] | (19.5,20.5] | (20.5, Inf] |
|---|---|---|---|---|---|
| (0.8,1] | 0.0855 | 0.5712 | 0.8539 | 0.9760 | 0.9702 |
| (0.6,0.8] | 0.2402 | 0.3913 | 0.7083 | 0.9462 | 0.9784 |
| (0.4,0.6] | 0.1732 | 0.3206 | 0.5437 | 0.8886 | 0.9789 |
| (0,0.4] | 0.1031 | 0.0890 | 0.2030 | 0.6608 | 0.9684 |

Figure 4.20: Heat map of the distribution for galaxies as in Fig. 4.19 but using the mixed prior in our classifier.

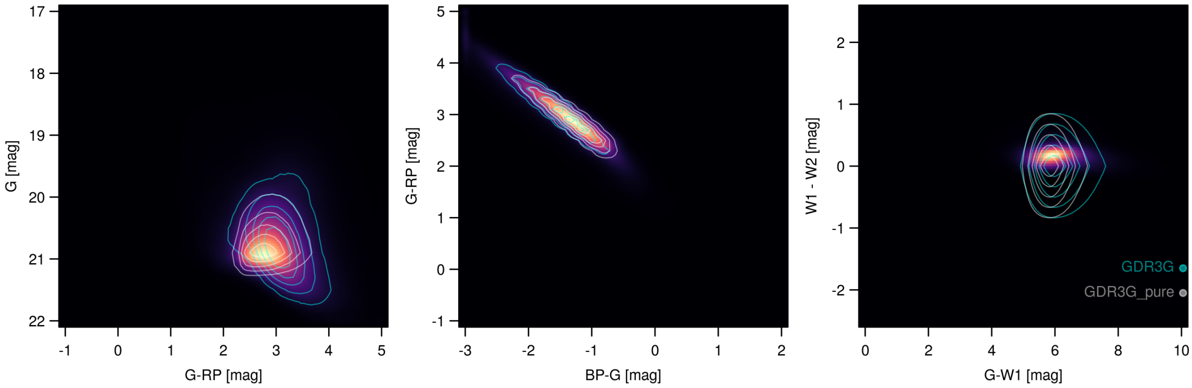


Figure 4.21: GDR3 Galaxy Candidate Table Mixed Prior: Colour-magnitude & Colour-colour diagrams for the galaxies using the mixed prior. The sources identified by the classifier are represented by the density scale, where black is zero density and yellow is high density. GDR3 galaxy sources are identified by the cyan contours and the GDR3 pure galaxy sample by the white contours.

Table 4.7: Galaxy candidates: Counts by predicted class and proportion identified as galaxies using the extragalactic-table-tuned prior defined in Sect. 4.5. GP and MP refer to the global prior and mixed prior respectively.

| | | Predicted | | | |
|---|---|---|---|---|---|
| | | Star | Quasar | Galaxy | Galaxy proportion |
| GP | GDR3 Galaxy | 306073 | 913 | 3887114 | 0.9268 |
| | GDR3 Pure Galaxy | 1834 | 20 | 2822716 | 0.9993 |
| | SDSS16 Galaxy + GDR3 | 27 | 3 | 514735 | 0.9999 |
| | SDSS16 Galaxy + Pure GDR3 | 1 | 0 | 393043 | 1.0000 |
| MP | GDR3 Galaxy | 128638 | 264534 | 3800928 | 0.9062 |
| | GDR3 Pure Galaxy | 710 | 41149 | 2782711 | 0.9852 |
| | SDSS16 Galaxy + GDR3 | 3 | 323 | 514439 | 0.9994 |
| | SDSS16 Galaxy + Pure GDR3 | 0 | 151 | 392893 | 0.9996 |

From Table 4.7, we find the proportion of galaxies identified by our classifier in the full galaxy candidate table to be 93% when using the global prior and if we apply the mixed prior to this table we find 91%. If we further constrain the sample by considering the pure subsample or the pure subsample plus the SDSS16 galaxy table, we see the proportion of galaxies identified by our classifier is higher, at 99% for both priors. Exploring the entire GDR3 galaxy candidate sample further in Fig. 4.19 for the global prior and Fig. 4.20 for the mixed prior. Comparing the two priors, we see a higher proportion of galaxies identified in the fainter and higher magnitude end in the global prior case than with the mixed prior, but the distribution on average is quite similar. Furthermore the mixed prior considers more galaxy sources to be quasars, particularly at the bright end and at higher latitudes, whereas the global prior considers more galaxy sources to be stars at the bright end but at lower latitudes.

We can see this distribution for the mixed prior results in Table 4.7 and in Fig. 4.21. We see closer contours for the GDR3 pure sample centred around the highest density region when using the mixed prior classifier and wider contours for the GDR3 sample as expected. A similar result is seen when applying the global prior. In contrast to the work by Bailer-Jones et al. (2019), our classifier was fit without using parallax or proper motions in order to retain as many galaxy sources as possible. We assess in Fig. 4.22 whether our classifier has a different distribution in either the count or probability spaces for sources with parallax and proper motions and for those that do not. We see a tendency towards redder magnitudes for the sources classified using the mixed prior without parallax and proper motions. The probability distributions are unperturbed and follow a similar trend with higher probabilities towards the lower magnitudes.

## 4.6 Conclusions

Building large catalogues of well-classified extragalactic sources is useful for large-scale statistical analyses in astronomy. In this paper, we look at how adding infrared colour information improves the classification of extragalactic sources compared to simply using Gaia. Our results

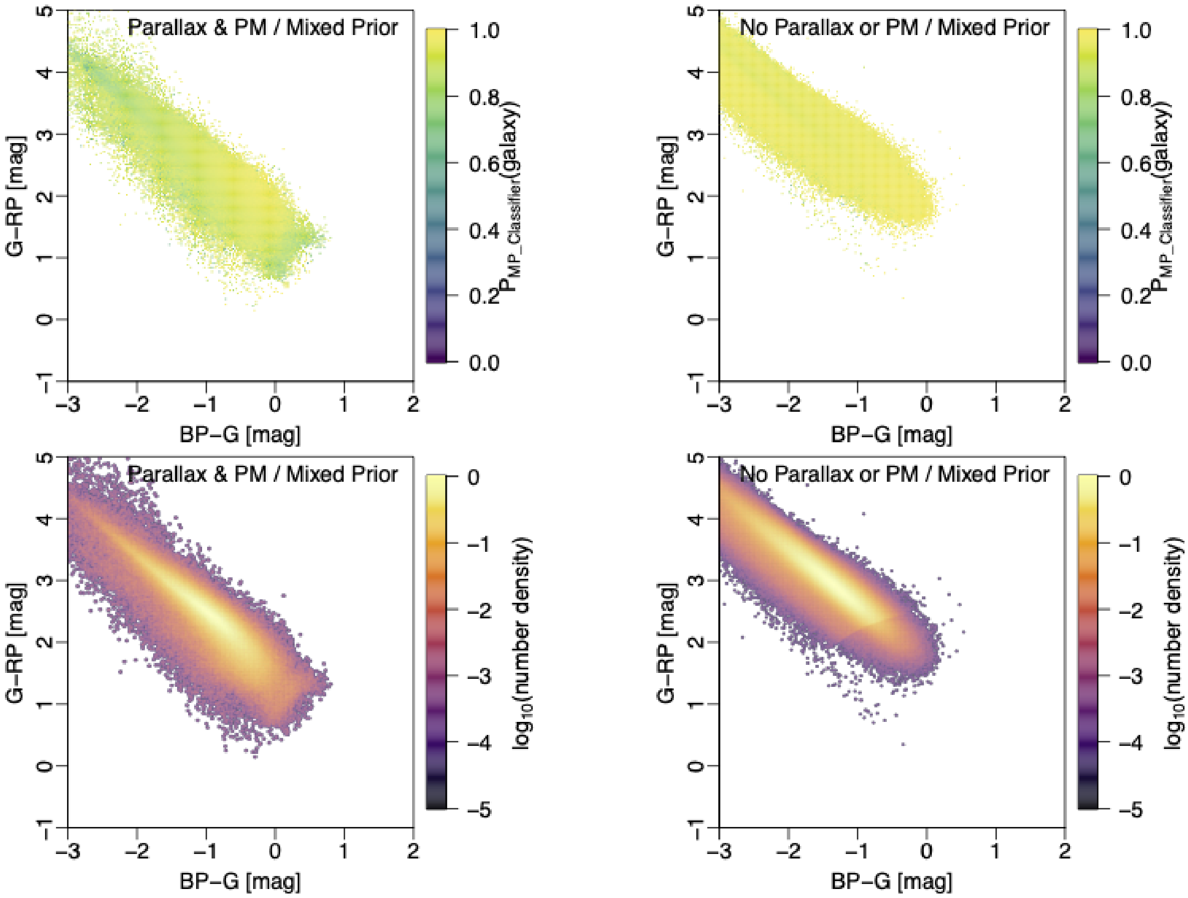


Figure 4.22: GDR3 Galaxy Candidate Table Mixed Prior: Probability and density distributions for sources classified as a galaxy The right hand side panels correspond to sources with parallax while the left hand side panels represent the distribution for sources without parallax.We find in the bottom-right panel a similar colour excess factor locus at BP-G =-0.5 and G-RP=2, as in figure 3 of Bailer-Jones et al. (2019) and figure 31 of Gaia Collaboration et al. (2022). This locus is however not evident in the case which has parallax and proper motions.

Table 4.8: Quasar candidates: A subset of the table of mixed prior probabilities as calculated on the quasar candidate table from GDR3. The full tables of probabilities as calculated on the quasar candidate table from GDR3 and on the galaxy candidate table from GDR3 are available upon request.

| source_id | isQSO_pure | isQSO_SDSS | pStar | pQSO | pGAL |
|---|---|---|---|---|---|
| 3470333738112 | 1 | 1 | 0.0001936 | 0.9998064 | 0.0000000 |
| 5944234902272 | 1 | 1 | 0.0001846 | 0.9998154 | 0.0000000 |
| 6459630980096 | 1 | 0 | 0.0009402 | 0.9969757 | 0.0020841 |
| 9517648372480 | 1 | 0 | 0.0001880 | 0.9998120 | 0.0000000 |
| 10655814178816 | 1 | 0 | 0.0001485 | 0.9998457 | 0.0000058 |
| ... | ... | ... | ... | ... | ... |

indicate an improved classification performance when adding the infrared colour information from CatWISE2020. The purities of the quasar and galaxy class improve from 0.9091 and 0.9759 to 0.9705 and 0.9784, respectively. We discuss how using a prior and adjusting the confusion matrix to reflect the expected (high) level of stellar contamination in a real application are necessary steps in ensuring that the reported results are representative of the performance of the classifier when test or application datasets do not reflect the true class distribution. Significantly, we find that using a prior that varies with latitude and magnitude gives higher purity *and* completeness for extragalactic objects: Looking at Fig. 4.6 in the adjusted case, and taking the bin where sinb $= (0.6, 0.8]$ and G $= (18.5, 19.5]$, we observe an improvement in the purity of the quasar class from 0.51 to 0.58. This result is coupled with a higher completeness seen in Fig. 4.5, from 0.84 to 0.97 in this bin. The published probabilities for the mixed prior classifier applied to the quasar and galaxy extragalactic candidate tables are available upon request. Table 4.8 illustrates the format of the tables. Exploiting the results of our classifications would be useful to scientific studies focusing on extragalactic sources as well as investigating stellar populations in the Milky Way as observed by Gaia and CatWise2020. Finally, when testing different statistical models, we find that decision-tree-based methods, in particular XGBoost, are more effective than Gaussian mixture models for this type of classification task.

# 5

# 3-Dimensional Solid Body Rotation of Clusters in Gaia DR3: An Application to Open Clusters

Still utilising Gaia, but specifically Gaia DR3, which provided a treasure trove of refinements and additional data products for the astronomical community, we turn our attention away from the machine learning approaches applied in Chapter 3 and 4 to focus on a more traditional statistical methodology of maximum likelihood estimation. In this Chapter we look at open clusters (OCs), whose 3D kinematics on a large scale would not have been accessible before the advent of Gaia. We estimate their internal rotational parameters, which we consider to be the angular velocity of the rotation axis, the inclination of the rotation axis relative to the line of sight, and the position angle of the rotation axis, using a list of probable members of numerous OCs compiled by Cantat-Gaudin (priv communication 2022). The goal would be to publish a list of rotational parameters for these OCs, and assess whether their rotational properties are linked with inherent characteristics of the Galaxy, from position to age distribution. However as discussed in this chapter, additional data are required to achieve this objective.

## 5.1 Introduction

Star formation occurs within hierarchically structured giant molecular clouds (e.g., Lada and Lada, 2003), with dense regions known as "clumps" being the birthplaces of star clusters (e.g., Shu et al., 1987). However, despite being gravitationally bound, only a small proportion of initial stellar groups observed in young stellar object distributions *remain* bound star clusters after gas dispersal. The process by which these surviving groups transform into bound clusters remains poorly understood, despite efforts to model it through direct N-body simulations (e.g., Kroupa et al., 2001). Once formed, bound star clusters are influenced by both internal and external factors, including stellar and binary evolution, mass segregation, two-body relaxation, Galactic tides, and gravitational interactions with passing molecular clouds (Brinkmann et al., 2017). The complex interplay of these factors can lead to the eventual dissolution of many star clusters.

Star clusters can be broadly categorised into two types: globular and open clusters. Globular clusters (GCs) are relatively stable and densely packed groups of stars, comprising tens of thousands to millions of stellar objects. They differ from open clusters (OCs) in their larger size and greater gravitational compactness. The strong gravitational attraction between the closely packed stars gives GCs their characteristic spherical shape. They predominantly consist of older and redder stars compared to OCs, which typically disperse before their stars reach old age. Due to their significant gravitational binding, GCs exhibit remarkable stability and can persist for billions of years. They are found in various types of environments, including the halo and bulge of our Milky Way. Studying GCs provides valuable insights into stellar processes, as their constituent stars form and evolve together. Significant work on the properties of GCs has been conducted by Baumgardt and Hilker (2018), among others. GCs have long been considered to be spherical, non-rotating stellar systems (Bianchini et al., 2013) given that internal relaxation processes would naturally dissipate, but in recent years several studies have shown evidence for significant internal rotation in many Milky Way GCs (Koch et al., 2018; Bianchini et al., 2018; Kacharov et al., 2014; Sollima et al., 2019; van Leeuwen et al., 2000). These studies used a variety of methods from in the plane of the sky rotation to an analysis of their stars' proper motions. Furthermore, signatures of rotation have also been found for young massive clusters and nuclear star clusters, indicating rotation is a common property across different compact stellar systems. Theoretically, the presence of internal rotation in GCs raises interesting questions regarding their formation and evolution. Firstly, rotation is known to influence the long-term dynamical evolution of GCs, with various studies suggesting that rotation would accelerate this process (Hong et al., 2013) and shape their current morphology (van den Bergh, 2008; Bianchini et al., 2013). Even a relatively small amount of angular momentum observed in modern GCs may reflect the remnants of strong primordial rotation in proto-GCs (Tiongco et al., 2018). Lastly, the formation mechanism of multiple stellar populations in GCs is still an unsolved puzzle, one that could potentially be solved by understanding the rotational properties of GCs (Mastrobuono-Battisti and Perets, 2013; Cordero et al., 2017)

On the other hand, OCs are loosely bound systems, typically composed of tens to a few hundred stars. In contrast to GCs, OCs are smaller and less densely populated. They exhibit a range of ages, from a few million to several billion years, although few OCs are known at the older end of that range; due to their open and dispersed structure, OCs lack significant stability, typically allowing their constituent stars to disperse over millions of years. Consequently, OCs are primarily located within galaxies undergoing active star formation, such as spiral and irregular galaxies. In contrast, elliptical galaxies, which lack significant star formation activity, no longer host OCs, as they likely dissipated long ago. Within our Milky Way galaxy, OCs are distributed throughout the spiral arms and inter-arm regions, providing valuable insights into the structure and evolution of our Galaxy. Despite the identification of numerous OCs in the Milky Way (e.g. Cantat-Gaudin et al., 2019; Cantat-Gaudin and Anders, 2020), their internal kinematics, particularly rotation, which is closely tied to their formation and evolution, remain poorly documented.

A few studies have focused on specific open clusters. Hao et al. (2022) and Healy et al. (2021) studied Praesepe, where they found the rotation value to be $0.2 \pm 0.05\,\mathrm{km\,s^{-1}}$and $0.132 \pm 0.027\,\mathrm{km\,s^{-1}}$, respectively. Kamann et al. (2019) looked at NGC 6791, and NGC 6819 with the values for rotation being $0.40 \pm 0.18\,\mathrm{km\,s^{-1}}$and $0.05 \pm 0.05\,\mathrm{km\,s^{-1}}$by considering rotation in the plane of the sky, rather than considering the cluster to be a solid body. This work aims to address this significant knowledge gap by estimating the rotational parameters, including theta, inclination, and rotational velocity, for a selection of open clusters identified in Gaia DR3 by Cantat-Gaudin (priv. communication, 2022).

## 5.2 Methodology

The following section outlines the methodology employed to identify evidence of rotation within a star cluster. This technique draws upon the approach presented by Sollima et al. (2019), in which the cluster is treated as a three-dimensional solid body.

When a cluster undergoes rotation, the average velocities of its stars exhibit variations based on their respective position angles. The position angle refers to the angular position of a star within the cluster, typically measured relative to a reference point or direction. As the cluster rotates, stars located at different position angles exhibit distinct average velocities. This is attributed to the gravitational interactions among the stars within the cluster; the gravitational forces exerted by neighbouring stars vary as a function of position, leading to differences in the stars' motion and hence in their average velocities.

To determine the mean velocity of the cluster, this method considers three velocity components, namely $v_Z$, $v_\parallel$ and $v_\perp$. $v_Z$ represents radial velocity, and $v_\parallel$, $v_\perp$ are the velocity components in the directions parallel and perpendicular to the rotation axis, respectively. These velocity components are influenced by a number of parameters, including the angular velocity of the rotation axis, the inclination of the rotation axis relative to the line of sight, and the position angle of the rotation axis.

In the following section, we derive the equations that describe a cluster exhibiting solid-body rotation, drawing inspiration from the work of Sollima et al. (2019).

### 5.2.1 3-Dimensional Solid Body Rotation

Consider a reference frame defined such that a cluster rotates clockwise in the $x - y$ plane, with the $z$-axis directed inward in the direction of the angular momentum. The systemic velocities along the three components can be written as

$$\begin{aligned} v_x &= \omega y \\ v_y &= -\omega x \\ v_z &= 0 \end{aligned} \tag{5.1}$$

where $\omega \equiv \omega(x, y, z)$ is the angular velocity, which is constant in the case of solid body rotation.

Now consider an observer in the frame defined by $(X, Y, Z)$. The velocity components measured by this observer looking at the cluster from an inclined perspective $(v_X, v_Y, v_Z)$ can be obtained by sequentially applying two rotations along the $x$ - and $z$-axes by angles $i$ and $\theta_0$, respectively:

$$\begin{bmatrix} v_X \\ v_Y \\ v_Z \end{bmatrix} = \begin{bmatrix} \cos\theta_0 & -\sin\theta_0\cos i & \sin\theta_0\sin i \\ \sin\theta_0 & \cos\theta_0\cos i & -\cos\theta_0\sin i \\ 0 & \sin i & \cos i \end{bmatrix} \begin{bmatrix} v_x \\ v_y \\ v_z \end{bmatrix} = \begin{bmatrix} \omega(x\sin\theta_0\cos i + y\cos\theta_0) \\ -\omega(x\cos\theta_0\cos i - y\sin\theta_0) \\ -\omega x\sin i \end{bmatrix} \tag{5.2}$$

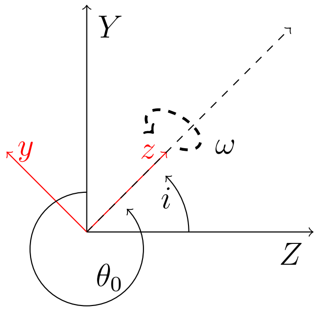


FIGURE 5.1: Diagram illustrating the coordinate frames and angles. Here the axes defined in red represent the frame of the cluster, and those in black the perspective of the observer. $\theta_0$ is defined anti-clockwise and inclination from line of sight, $\omega$ defines the rotational velocity around the dashed line representing the axis of rotation.

Defining the position angle $\theta$ anticlockwise from the $Y$-axis we have

$$X = -R\sin\theta \quad Y = R\cos\theta$$

where $R = \sqrt{X^2 + Y^2}$ is the projected distance from the cluster centre. The coordinate transformations between the two reference systems are

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \begin{bmatrix} \cos\theta_0 & -\sin\theta_0\cos i & \sin\theta_0\sin i \\ \sin\theta_0 & \cos\theta_0\cos i & -\cos\theta_0\sin i \\ 0 & \sin i & \cos i \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} \tag{5.3}$$

or

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} \cos\theta_0 & \sin\theta_0 & 0 \\ -\sin\theta_0\cos i & \cos\theta_0\cos i & \sin i \\ \sin\theta_0\sin i & -\cos\theta_0\sin i & \cos i \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \begin{bmatrix} -R\sin(\theta-\theta_0) \\ R\cos(\theta-\theta_0)\cos i + Z\sin i \\ -R\cos(\theta-\theta_0)\sin i + Z\cos i \end{bmatrix} \tag{5.4}$$

Consider the projections of the velocity vector in the plane of the sky in the directions parallel and perpendicular to the rotation axis:

$$\begin{bmatrix} v_\parallel \\ v_\perp \end{bmatrix} = \begin{bmatrix} -\sin\theta_0 & \cos\theta_0 \\ \cos\theta_0 & \sin\theta_0 \end{bmatrix} \begin{bmatrix} v_X \\ v_Y \end{bmatrix} \tag{5.5}$$

Combining 5.2, 5.4 and 5.5 we finally find

$$\begin{bmatrix} v_Z \\ v_\parallel \\ v_\perp \end{bmatrix} = \begin{bmatrix} \omega R\sin(\theta-\theta_0)\sin i \\ \omega R\sin(\theta-\theta_0)\cos i \\ \omega[R\cos(\theta-\theta_0)\cos i + Z\sin i] \end{bmatrix} \tag{5.6}$$

### 5.2.2 Systemic cluster velocities

To estimate the rotation of a cluster, we first need an estimate of the systemic motion of the cluster. We initially convert the celestial coordinates (RA,Dec) into projected distances from the cluster centre (X,Y), using equation 1 of van de Ven et al. (2006), and adopting the centres of each cluster from either Sollima et al. (2019) for GCs or Cantat-Gaudin and Anders (2020) for OCs.

We then maximise the likelihood function

$$
\begin{aligned}
\ln L = -\frac{1}{2}\sum_i \Bigg[ & \frac{\left(v_{\mathrm{Z},i} - \langle v_{\mathrm{Z}} \rangle\right)^2}{s^2_{\mathrm{Z},i}} + \ln\left(s^2_{\mathrm{Z},i}\left(1-\tilde{\rho}_i^2\right)\right) \\
& + \sum_{j=\mathrm{RA,\ Dec.}} \left( \frac{\left(v_{j,i} - \langle v_j \rangle\right)^2}{\left(1-\tilde{\rho}_i^2\right) s^2_{j,i}} + \ln\left(s^2_{j,i}\right) \right) \\
& - \frac{2\tilde{\rho}_i \left(v_{\mathrm{RA},i} - \langle v_{\mathrm{RA}} \rangle\right)\left(v_{\mathrm{Dec.}\ ,i} - \langle v_{\mathrm{Dec.}\ } \rangle\right)}{\left(1-\tilde{\rho}_i^2\right) s_{\mathrm{RA},i} s_{\mathrm{Dlc},i}} \Bigg],
\end{aligned}
\tag{5.7}
$$

where

$$
\begin{aligned}
s^2_{j,i} &= \sigma^2_{j,i} + \epsilon^2_{j,i} \quad j = \mathrm{Z}, \mathrm{RA}, \mathrm{Dec} \\
\tilde{\rho}_i &= \rho_i \frac{\epsilon_{\mathrm{RA}}\epsilon_{\mathrm{Dec}}}{s_{\mathrm{RA}} s_{\mathrm{Dec}}}
\end{aligned}
$$

as defined in Sollima et al. (2019), and subtract this systemic motion from each star in the cluster. In the above, $\sigma_{j,i}$ and $\epsilon_{j,i}$ are the velocity dispersion and error. The velocity dispersion is defined to be intrinsic to the cluster, and is assumed to be the same in all components of Z, RA and Dec.

### 5.2.3 Rotational parameters

In observed clusters, the angular velocity is a function of the distance from the rotation axis. This means that a rigorous model needs to be fitted to the data, but doing so can introduce a dependence on the assumptions of the model. To avoid this, an average projected velocity $\omega R$, defined as $A$, is used, which is assumed to be independent of distance. This approximation does not introduce any bias in the estimate of the position angle and inclination of the rotation axis.

For each cluster of our sample we searched for the values of $\theta$,$i$, and $A$ which maximised the following likelihood

$$
\begin{aligned}
\ln L = -\frac{1}{2}\sum_i \Bigg[ & \frac{\left(v_{Z,i} - \bar{v}_{Z,i}\right)^2}{s^2_{Z,i}} + \ln\left(s^2_{\mathrm{Z},i}\left(1-\bar{\rho}_i^2\right)\right) \\
& + \sum_{j=\parallel,\perp} \left( \frac{\left(v_{j,i} - \overline{v_j}\right)^2}{\left(1-\bar{\rho}_i^2\right) s^2_{j,i}} + \ln\left(s^2_{j,i}\right) \right) \\
& - \frac{2\bar{\bar{i}}_i \left(v_{\parallel,i} - \overline{v_{\parallel,i}}\right)\left(v_{\perp,i} - \overline{v_{\perp,i}}\right)}{\left(1-\bar{\rho}_i^2\right) s_{\parallel,i} s_{\perp,i}} \Bigg]
\end{aligned}
\tag{5.8}
$$

where

$$\begin{aligned}
s^2_{j,i} &= \sigma^2_{j,i} + \epsilon^2_{j,i} \quad j = \mathrm{LOS}, \parallel, \perp \\
\epsilon^2_{\parallel,i} &= \epsilon^2_{\mathrm{RA},i} \sin^2\theta_0 + \epsilon^2_{\mathrm{Dec},i} \cos^2\theta_0 - 2\rho_i \epsilon_{\mathrm{RA},i} \epsilon_{\mathrm{Dec},i} \sin\theta_0 \cos\theta_0 \\
\epsilon^2_{\perp,i} &= \epsilon^2_{\mathrm{RA},i} \cos^2\theta_0 + \epsilon^2_{\mathrm{Dec},i} \sin^2\theta_0 + 2\rho_i \epsilon_{\mathrm{RA},i} \epsilon_{\mathrm{Dec},i} \sin\theta_0 \cos\theta_0 \\
\bar{\rho}_i &= \frac{1}{s_{\parallel,i} s_{\perp,i}} \left[ \frac{\left(\epsilon^2_{\mathrm{Dec},i} - \epsilon^2_{\mathrm{RA},i}\right.}{2} \sin 2\theta_0 + \rho_i \epsilon_{\mathrm{RA},i} \epsilon_{\mathrm{Dec},i} \cos 2\theta_0 \right] \\
\overline{v_{Z,i}} &= A \sin\left(\theta_i - \theta_0\right) \sin i \\
\overline{v_{\parallel,i}} &= A \sin\left(\theta_i - \theta_0\right) \cos i \\
\overline{v_{\perp,i}} &= A \cos\left(\theta_i - \theta_0\right) \cos i
\end{aligned}$$

In the given notation, $A$ is unbounded and a positive value signifies clockwise rotation. Inclination is defined in the range of $0^\circ < i < 90^\circ$ and represents the angle with respect to the line of sight; for example, $i = 90^\circ$ is in the plane of the sky. $\theta_0$ is increases anti-clockwise from North ($0^\circ$) and is defined in the range of $0^\circ < \theta_0 < 360^\circ$.

### 5.2.4 Maximum Likelihood Estimation

Maximum Likelihood Estimation (MLE) is a statistical method used to estimate the parameters of a probability distribution that best fit a given set of observed data, by finding the parameter values that maximise the likelihood of observing the data under the assumed model.

Markov chain Monte Carlo (MCMC) is a tool that leverages the power of Markov chains to generate samples from complex probability distributions, and is a general tool for the simulation of stochastic processes. It however may be applied to the area of likelihood inference, and is particularly useful when the likelihood function is intractable or when an exact calculation of the maximum likelihood estimate is not possible (Geyer, 1991).

MCMC algorithms, such as the Metropolis-Hastings algorithm and the Gibbs sampler, iteratively generate a sequence of samples that asymptotically converge to the target distribution (Chib and Greenberg, 1995). The key idea is to construct a Markov chain with a stationary distribution equal to the desired target distribution. By running the chain for a sufficient number of iterations, the generated samples provide an approximation of the underlying distribution.

The MCMC algorithm starts with an initial set of parameter values and iteratively proposes new parameter values based on a proposal distribution. The acceptance or rejection of these proposals depends on the likelihood function and a Metropolis-Hastings acceptance ratio. One advantage of MCMC methods is their ability to handle complex and high-dimensional parameter spaces. They can efficiently explore the parameter space, allowing for robust estimation even when dealing with models with numerous parameters. Additionally, MCMC methods provide estimates of the uncertainty associated with the parameter estimates through the analysis of the sampled posterior distributions.

Differential evolution is another optimisation technique that can be combined with MCMC (DE) to enhance parameter estimation (Braak, 2006). DE is a population-based stochastic optimisation algorithm that mimics the process of natural selection. It operates by maintaining a population of candidate solutions and iteratively evolving the population by applying mutation, crossover, and selection operations. In the context of MLE, differential evolution with MCMC combines the global search capabilities of DE with the local exploration capabilities of MCMC. DE helps in efficiently exploring the parameter space by maintaining multiple candidate solutions, while MCMC refines the estimates by focusing on local regions of the parameter space. This hybrid approach can improve the convergence rate and robustness of the estimation process. The integration of DE with MCMC involves using DE to generate initial parameter values or to propose new parameter values during the MCMC iterations. The advantages of utilising DE over conventional MCMC are simplicity, speed of calculation and convergence, even for nearly collinear parameters and multimodal densities (Braak, 2006). The performance of DE combined with MCMC, which we will refer to as DE for the remainder of this work, depends on the appropriate choice of mutation strategies, crossover operators, and population sizes.

The likelihood function for MLE is typically defined as the product of the probability density function (PDF) evaluated at each data point. In the case of independent and identically distributed (i.i.d.) data, the likelihood function can be expressed as:

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(x_i|\boldsymbol{\theta}) \tag{5.9}$$

where $\boldsymbol{\theta}$ [1]represents the vector of parameters to be estimated, $n$ is the number of data points, and $f(x_i|\boldsymbol{\theta})$ is the PDF of the distribution.

In MCMC, the Metropolis-Hastings acceptance ratio is given by:

$$\alpha = \min\left(1, \frac{f(\boldsymbol{\theta}')q(\boldsymbol{\theta}|\boldsymbol{\theta}')}{f(\boldsymbol{\theta})q(\boldsymbol{\theta}'|\boldsymbol{\theta})}\right) \tag{5.10}$$

where $q(\boldsymbol{\theta}'|\boldsymbol{\theta})$ is the proposal distribution that generates a new parameter value $\boldsymbol{\theta}'$, given the current parameter value $\boldsymbol{\theta}$, and $f(\boldsymbol{\theta})$ and $f(\boldsymbol{\theta}')$ represent the likelihood function evaluated at $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$, respectively.

DE operates by iteratively updating the candidate solutions in the population. The mutation operation generates new candidate solutions by combining multiple existing solutions, and the crossover operation combines the mutated solutions with the original solutions to create offspring. The selection operation determines which candidate solutions survive to the next generation based on their fitness.

The mutation operation in differential evolution can be expressed as:

$$\boldsymbol{\theta}_{\text{mut}} = \boldsymbol{\theta}_a + F \cdot (\boldsymbol{\theta}_b - \boldsymbol{\theta}_c) \tag{5.11}$$

where $\boldsymbol{\theta}_a$, $\boldsymbol{\theta}_b$, and $\boldsymbol{\theta}_c$ are randomly selected solutions from the population, and $F$ is the scaling factor that controls the amplification of the difference between $\boldsymbol{\theta}_b$ and $\boldsymbol{\theta}_c$.

The crossover operation combines the mutated solution $\boldsymbol{\theta}_{\text{mut}}$ with the original solution $\boldsymbol{\theta}$ to generate offspring:

$$\boldsymbol{\theta}_{\text{off}}(i) = \begin{cases} \theta_{\text{mut}}(i) & \text{if} \quad \text{rand}(0,1) \leqslant CR \quad \text{or} \quad i = \text{rand}(1,D) \\ \theta(i) & \text{otherwise} \end{cases} \tag{5.12}$$

where $\theta_{\text{off}}(i)$ represents the $i$-th component of the offspring solution, $D$ is the dimensionality of the parameter space, and $\text{rand}(a,b)$ generates a random number between $a$ and $b$. $CR$ is the crossover rate that determines the probability of each component being replaced by the corresponding component of the mutated solution.

[1] Please note this $\boldsymbol{\theta}$ is not the same as the $\theta_0$ defining the axis of rotation.

The MCMC implementation that we use is the `Emcee` Python package (Foreman-Mackey et al., 2013), which provides an efficient and robust method for MCMC sampling. `Emcee`, or "the MCMC Hammer," offers a user-friendly interface and efficient algorithms, and has a few key components that distinguish it from other implementations. `Emcee` is an affine-invariant ensemble sampler, which is a variant of the MCMC method introduced by Goodman and Weare (2010); unlike the simple MCMC approach, which struggles with slow convergence in highly correlated parameter spaces, the ensemble sampler implemented in `Emcee` efficiently explores the parameter space. `Emcee` employs an ensemble of "walkers," where each walker represents a different point in the parameter space. The walkers evolve simultaneously, performing a random walk. At each step, the proposal is generated based on the positions of all the walkers, enabling better exploration of the parameter space and faster convergence. One of the significant advantages of `Emcee` is its ability to use parallelisation, by distributing the walkers across multiple cores or machines, thereby accelerating the sampling process. This feature is particularly useful when dealing with models that possess large parameter spaces or computationally intensive likelihood functions.

## 5.3 Data

Investigating the internal kinematics of stellar clusters and associations provides valuable insights into their formation and evolution. However, accurate astrometric measurements, including positions, velocities, and membership determinations, are essential for such studies. In 2022, the European Space Agency's Gaia mission released Gaia DR3 (Gaia Collaboration et al., 2022), the third data release from the project, which provides highly precise positions, distances, and proper motions for more than 1.8 billion stars. It is important to note that, prior to Gaia, this work on the parameters of OCs would not have been feasible.

To evaluate our methodology's performance under a variety of cluster parameterisations, we conduct tests on simulated clusters, as will be described in Section 5.4.1. In creating a simulated cluster of stars, we randomly sample two angles and a radius, all taken to be uniformly distributed: one angle between $0 - 2\pi$, the other between $0 - \pi$ and the radius between $0 - 5$ pc, and convert to Cartesian $x$, $y$, and $z$ coordinates within a sphere, centred at [0, 0, 0]. These coordinates define the cluster's frame, and velocities ($v_x$, $v_y$, $v_z$ in $\mathrm{km\,s^{-1}}$) are generated using Eqn. 5.1, with a specified angular velocity (omega). Eqn. 5.3 is then used to transform $x$, $y$, $z$ (arcmin) into $X$, $Y$, $Z$ (arcmin), considering inclination $i$ (degrees) and $\theta$ (degrees), along with subsequent velocities ($v_X$, $v_Y$, $v_Z$ in $\mathrm{km\,s^{-1}}$) using Eqn. 5.2 . Since our focus lies solely on the velocities within this $X$, $Y$, $Z$ frame, we do not extend the transformation into *RA*, *Dec* space, as the conversion between *RA*, *Dec*, and Cartesian coordinates by van de Ven et al. (2006) is well-established. We simulate different combinations of $\theta_0$, $i$, and $\omega$ to investigate and define our simulated dataset.

Regarding our dataset of GCs, we obtain their centres in *RA* and *Dec* from a comprehensive compiled database of GCs by Baumgardt and Hilker (2018). By querying the Gaia DR3 catalogue within a radius of 3 arcminutes and applying specific filtering criteria, we selected our cluster members. Notably, our cluster member selection may not be the same as that of Sollima et al. (2019), potentially yielding different results.

Finally, we applied our methodology to the list of clusters compiled by Cantat-Gaudin.

## 5.4 Results

The results section of this study presents findings related to the rotation of star clusters, encompassing three distinct aspects: rotation in simulated clusters, rotation in GCs, and rotation in OCs. These investigations shed light on the rotational dynamics and behaviours exhibited by different types of stellar groupings, which, as noted above, can provide valuable insights into their formation and evolution. To explore the phenomenon of rotation in star clusters, we initiated our analysis with simulated clusters in Section 5.4.1. By constructing simulated clusters using random generation of coordinates within a defined sphere, we were able to investigate the impact of various parameters on the rotational properties. These simulations served as a crucial foundation for understanding the underlying mechanisms and constraints governing rotational dynamics in real-world star clusters. In Section 5.4.2, our investigation delved into the realm of GCs. These tightly bound and gravitationally compact clusters are characterised by their spherical shape. By studying well-known GCs, we aimed to estimate the rotational parameters specific to these clusters. Utilising Gaia data and then supplementary radial velocity information from the catalogue of Baumgardt and Hilker (2018), we analysed the rotational signatures and compared our results with previous findings, acknowledging the potential influence of cluster membership selection on derived parameters. By applying our method to simulated and known GCs we were able to assess the validity of our method and whether having more radial velocities improve the accuracy. Unlike their GC counterparts, OCs are loosely bound systems consisting of a smaller number of stars. In Section 5.4.3, we estimated the rotational parameters for selected OCs, with the aim of contributing to a broader understanding of their formation and dynamical evolution.

### 5.4.1 Test on simulated clusters

The performance of the methodology described in Chapter 5.2 is evaluated by conducting analyses on simulated clusters with diverse sets of parameter values. By employing our methodology on simulated clusters, we can acquire insights into the performance of the technique under various specifications of cluster rotational parameters. This allows us to identify cases where the method excels and where it encounters limitations or challenges, thus revealing patterns and commonalities. Furthermore, we aim to demonstrate the performance trend from noiseless simulations to simulations that incorporate noise derived from expected data uncertainties. To accomplish these analyses, we consider the following simulations:

- **Simulation 1** We utilise noise-free simulated data with fixed standard deviations in the likelihoods corresponding to the expected error in the transverse velocity measurements in Gaia, estimated to be $[v_{Xe}]$, $[v_{Ye}] = 0.05\,\mathrm{km\,s^{-1}}$, and $[v_{Ze}] = 2\,\mathrm{km\,s^{-1}}$. The transverse velocity errors are taken to be 0, as are the $\rho_i$ (correlation between the proper motions) and $\sigma^2_{j,i}$ (dispersion) terms.

- **Simulation 2** We simulate a spherical open cluster by using expected properties of an open cluster, such as the number of stars representative of a median open cluster, and the transverse velocity errors, correlation between the proper motions from real Gaia data, and with the $\sigma^2_{j,i}$ dispersion set to $1\,\mathrm{km\,s^{-1}}$. No noise is added to the transverse velocities.

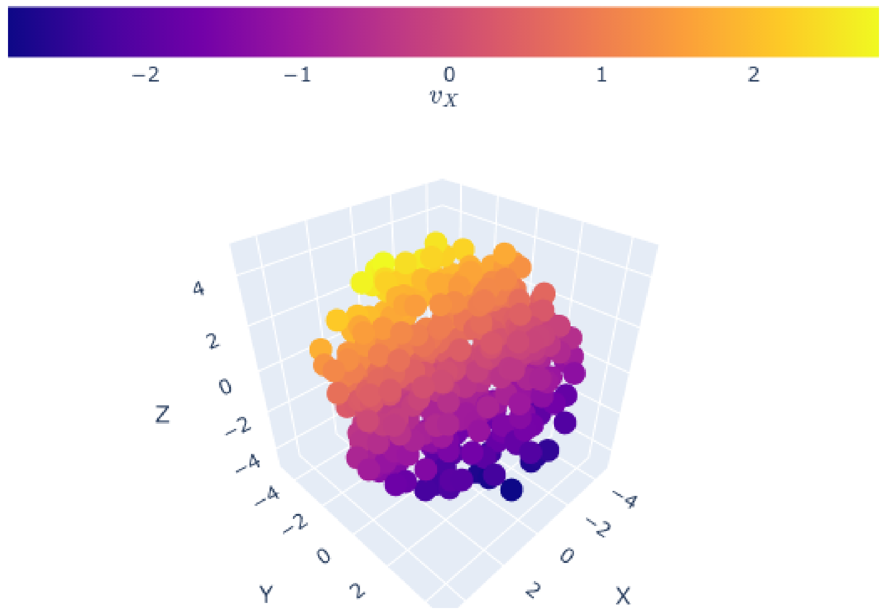


FIGURE 5.2: Velocity in the X coordinate for a simulated cluster with $\theta = 181^\circ$, $i = 49.0^\circ$, and $A = -0.6\,\mathrm{km\,s^{-1}}$. For an interactive view of each velocity component please click the following link: interactive plots

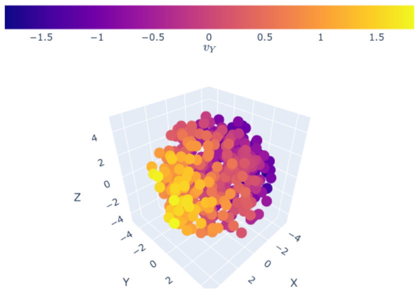


FIGURE 5.3: Velocity in the Y coordinate for the same simulated cluster.

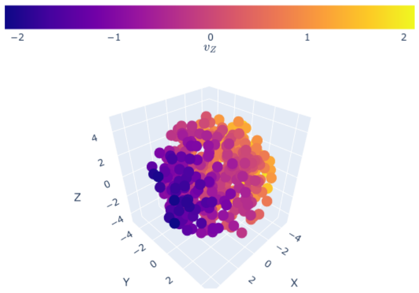


FIGURE 5.4: Velocity in the Z coordinate for the same simulated cluster.

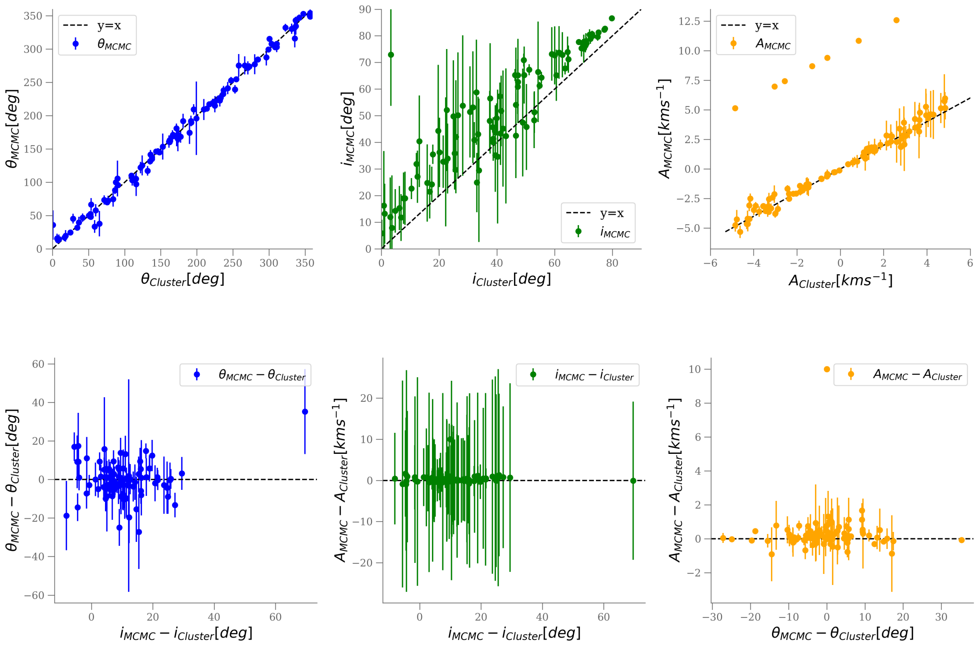


Figure 5.5: Simulation 1: Comparing the MCMC method's estimation of cluster parameters with their actual (input) values. The results show that the MCMC method performs well for parameter $\theta$, closely aligning with the identity line and exhibiting minimal errors. However, the parameter for inclination consistently displays an offset of approximately 10, indicating that the chains have not deviated significantly from their initial values. As for parameter A, the majority of estimates exhibit a well-distributed pattern, although larger errors are observed for values further from 0. Notably, there are significant outliers at a constant value of 10, similar to the inclination parameter, suggesting limited exploration of the parameter space by the chains.

- **Simulation 3** We consider the exact parameterisation as in Simulation 2, however this time we add some noise to the velocities sampled from a normal distribution with $\mu = 0$ and $\sigma = 0.05\,\mathrm{km\,s^{-1}}$for $v_X$ and $v_Y$, and $\sigma = 2\,\mathrm{km\,s^{-1}}$for $v_Z$.

In each of the simulations, we generate 100 spherical clusters, where the number of stars in Simulation 1 is 500 per cluster and in Simulations 2 and 3 is set to be 267 (described further below), each with radial velocity measurements, and the rotational parameters between the ranges of $0 \leqslant \theta \leqslant 360$, $0 \leqslant$ inclination $\leqslant 90$ and $-5 \leqslant A \leqslant 5$. Through these simulations, we aim to evaluate the performance of our methodology across varying rotational parameters and noise levels, contributing to a deeper understanding of its efficacy in analysing and characterising clusters with a range of properties.

## Simulation 1

In Simulation 1, our main objective is to evaluate the methodology's performance under ideal conditions without any noise, error or dispersion. We specifically focus on addressing the potential degeneracy between the parameters $\theta$ and $i$, as well as comparing the results obtained

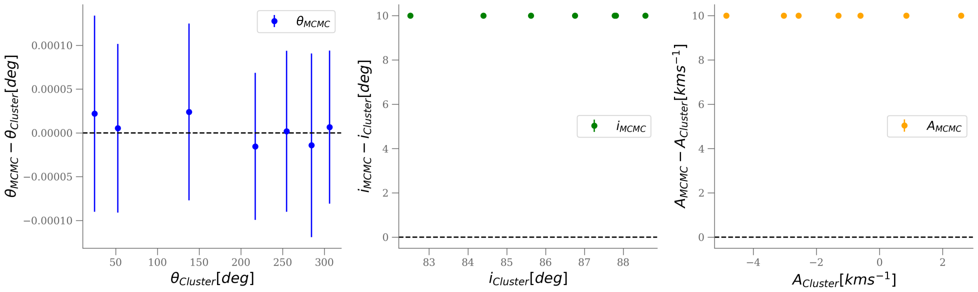


FIGURE 5.6: Simulation 1: Upon closer examination of the outliers in parameter A, it becomes evident that both inclination and A consistently exhibit an offset of 10. However, for these simulations, the estimation of parameter $\theta$ shows excellent agreement.

using two different sampling methods: MCMC and differential evolution with MCMC (DE).

For the MCMC analysis, we initialise the parameters as follows: $\theta = \bar{\theta} + \epsilon$, $i = i + 10 + \epsilon$, and $A = A + 10 + \epsilon$, where $\epsilon$ represents small Gaussian noise. Adding 10 to $\epsilon$ exaggerates the difference between the starting point and the true value, allowing for thorough exploration of the parameter space by the MCMC algorithm. For this simulation we use 100 walkers, with the numbers of steps set to 1000 with a 500 burn-in for both the MCMC and DE methods.

Table 5.1 summarises the considered parameter combinations and their corresponding estimated values from the analysis. Figure 5.5 illustrates the discrepancies between the estimated values and the true cluster parameters. The MCMC method performs well for the parameter $\theta$, closely following the identity line with minimal errors. However, the inclination parameter consistently exhibits an offset of approximately 10, suggesting limited exploration of the parameter space by the chains. Similarly, the parameter $A$ shows significant outliers at a constant value of 10, indicating deviations from the true value. On average, the differences between the estimated and true cluster parameters using MCMC are $-0.48^\circ$ for $\theta$, $10.25^\circ$ for inclination, and $0.88\,\mathrm{km\,s^{-1}}$for $A$.

To further investigate the issue of significant outliers, we examine Fig. 5.6, which reveals that while the estimate for $\theta$ aligns well, the offsets for both inclination and $A$ remain constant at 10 (degrees and $\mathrm{km\,s^{-1}}$, respectively). This suggests that the MCMC chains exhibit flat behaviour for these specific tests and fail to deviate from the initial values, as evident in Fig. A.8.

TABLE 5.1: This table presents a subset of the various combinations considered and the corresponding estimated values of the cluster parameters for 100 simulated clusters. For the full table see Table A.2 in the Appendix

| $\theta_C$ | $i_C$ | $A_C$ | $\theta_M$ | $\theta_{eM}$ | $i_M$ | $i_{eM}$ | $A_M$ | $A_{eM}$ | $\theta_D$ | $\theta_{eD}$ | $i_D$ | $i_{eD}$ | $A_D$ | $A_{eD}$ | $A_{snrM}$ | $A_{snrD}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.66 | 3.31 | -0.04 | 35.95 | 44.14 | 72.92 | 38.44 | -0.12 | 0.20 | 170.18 | 261.85 | 54.95 | 57.14 | -0.07 | 0.14 | 1.21 | 1.00 |
| 6.23 | 49.44 | 4.30 | 15.72 | 14.05 | 65.28 | 21.43 | 4.60 | 4.13 | 11.09 | 14.80 | 58.49 | 21.01 | 3.67 | 1.78 | 2.23 | 4.13 |
| 6.54 | 38.73 | -4.00 | 15.91 | 9.36 | 41.39 | 16.08 | -3.47 | 0.87 | 13.30 | 37.84 | 43.56 | 23.45 | -3.58 | 1.46 | 7.95 | 4.92 |
| 7.49 | 46.31 | 4.84 | 13.23 | 10.40 | 65.27 | 4.60 | 5.97 | 1.06 | 18.50 | 346.04 | 50.03 | 38.40 | 3.90 | 3.31 | 11.25 | 2.36 |

Continued on next page

Table 5.1 – continued from previous page

| $\theta_C$ | $i_C$ | $A_C$ | $\theta_M$ | $\theta_{eM}$ | $i_M$ | $i_{eM}$ | $A_M$ | $A_{eM}$ | $\theta_D$ | $\theta_{eD}$ | $i_D$ | $i_{eD}$ | $A_D$ | $A_{eD}$ | $A_{snrM}$ | $A_{snrD}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | |

For the DE approach, the average differences between the estimated and true cluster parameters are 2.08° for $\theta_0$, 4.96° for inclination, and 0.02 $\mathrm{km\,s^{-1}}$for $A$. Figure 5.5 and 5.7 illustrate the disparities between the estimated values obtained using both MLE methods and the true cluster parameters. Comparing the estimated and actual cluster parameters using the MCMC methodand the DE method, we observe that the DE method follows the expected cluster trend. However, it consistently overestimates inclination across the entire range. In terms of the parameter $A$, the DE method overestimates when $A < 0$ and underestimates when $A > 0$. Similar to the MCMC method, the errors for parameter $A$ are smaller when the values are closer to 0.

We further compare the two methods by considering the signal-to-noise ratio, estimate over error, in Fig. 5.8. Both methods exhibit an upward linear-like trend in parameter $\theta$, indicating that errors grow proportionally with increasing $\theta$ values. Inclination demonstrates an exponential trend, with the MCMC method showing more pronounced errors for lower inclination values. Parameter $A$ exhibits a quadratic shape, indicating a similar trend to that of $\theta$ due to the possibility of negative values.

Despite slight discrepancies (particularly the significant outlier for $\theta$ observed in the differential evolution case), we select the differential evolution with MCMC method (DE) as the preferred approach due to its effectiveness in exploring the parameter space, especially when the initial conditions significantly deviate from the true parameter value. We will still consider both sampling methods in Simulation 2 and Simulation 3, to further test this conclusion.

We present the example of a specific cluster with $\theta = 181°$, $i = 49.0\circ$, and $A = -0.6$ $\mathrm{km\,s^{-1}}$, using both MCMC and DE in Fig. 5.10 and Fig. 5.11, respectively. The errors obtained using DE are considerably lower for this simulated cluster compared to those using MCMC alone.

### Simulation 2

To further prepare for the application of our method on open clusters (OCs), we utilise parameters from known OCs in Gaia to generate a simulated cluster similar to an OC. Our selection process involves choosing a cluster representative of the median cluster type in Cantat-Gaudin, which then defines the number of stars and the velocity errors. For this purpose, we consider the OC Alessi 24, which has a median number of stars (267) and a radial velocity error of 4.61 $\mathrm{km\,s^{-1}}$, based on the available 94 stars with radial velocity measurements.

To simulate the cluster, we randomly sample from each transverse velocity component and use the observed Gaia errors for Alessi 24 as our simulated cluster errors, as well as the correlation between the proper motions ($\rho_i$). The dispersion term ($\sigma^2_{j,i}$) is set to 1 $\mathrm{km\,s^{-1}}$and there is no measurement noise added to the velocities. This allows us to assess the performance of the MCMC and DE sampling methods under more realistic conditions.

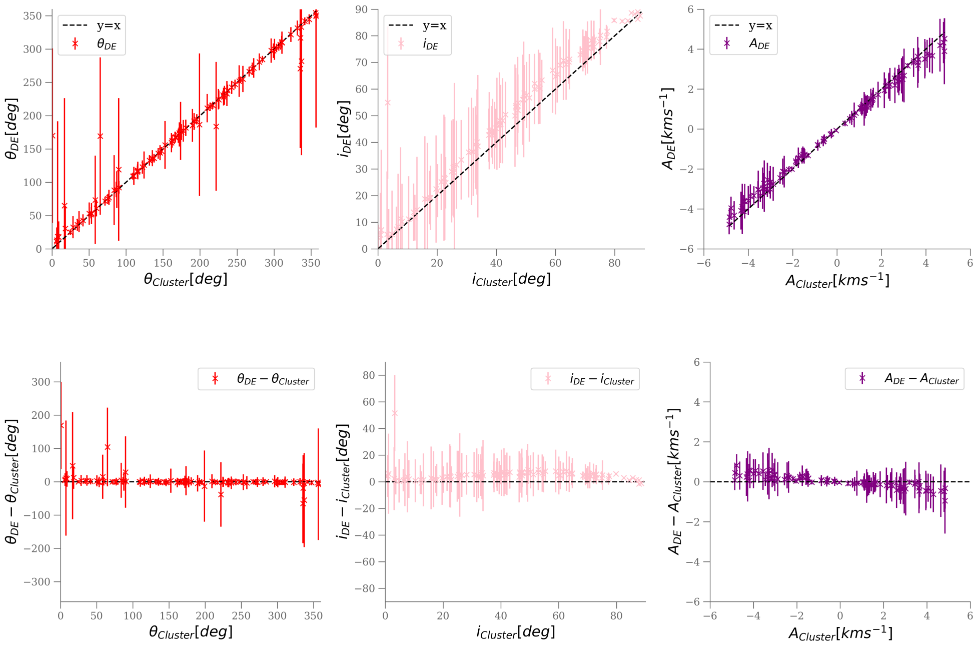


FIGURE 5.7: Simulation 1: Comparing the estimated and actual cluster parameters using the DE method. The results demonstrate that the DE method aligns with the expected cluster trend. However, it consistently overestimates the inclination across the entire range. In terms of parameter $A$, the DE method overestimates when $A < 0$ and underestimates when $A > 0$. Similar to the MCMC method, the errors for parameter $A$ are smaller when the values are closer to 0.

In Fig. 5.12, we compare the estimated and actual cluster parameters using the MCMC method. When incorporating real errors, we find that the MCMC method performs accurately for the parameter $\theta$, with minimal dispersion in errors for most points. However, a few points show significantly larger errors. The inclination parameter is consistently overestimated, indicating a potential issue with the initialisation process as the difference tends to be around 10. The inclination difference panel follows a similar pattern to Fig. 5.5. Parameter $A$ clearly shows the points that fail initialisation, as they all have a value of 10, which is the constant added to the initialisation value.

Next, we consider the DE method, shown in Fig. 5.13. We observe significantly larger errors for more points in the parameter $\theta$ compared to the MCMC method. The DE method consistently overestimates inclination across the entire range. Similar to the noise-free scenario, parameter $A$ is underestimated when $A < 0$ and overestimated when $A > 0$, with smaller errors observed when the values are closer to 0.

We analyse the distribution of errors and the trend in SNR in Fig. 5.14 and 5.15, respectively. The errors for $\theta$ in both methods do not exhibit any obvious trend but show a large spread relative to the range of $\theta$. Inclination shows a slightly negative linear trend, with the methods favouring an inclination closer to 90°. Parameter $A$ is randomly distributed with a low range, indicating that the methods might be accurately estimating errors. When looking at SNR, we would expect the angular parameters to follow a linear trend and $A$ to exhibit a positive definite parabolic shape. Both methods show the same expected trend across all

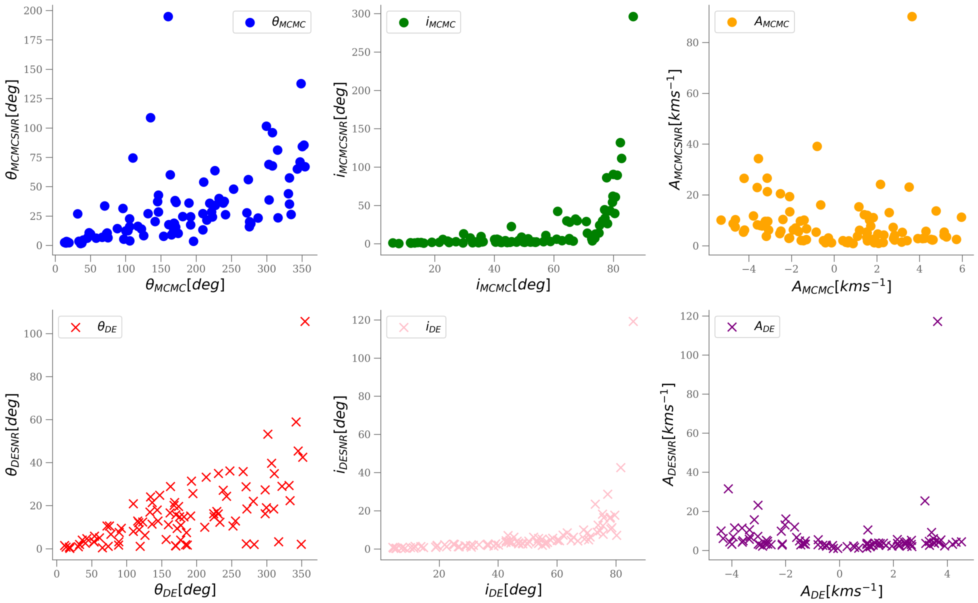


Figure 5.8: Simulation 1: The signal-to-noise ratio is examined for each parameter using two different sampling methods. Both methods display an upward linear-like trend in the parameter $\theta$, indicating that errors grow proportionally with increasing $\theta$ values. Inclination demonstrates an exponential trend, more pronounced in the MCMC method, suggesting higher errors for lower inclination values and lower errors for higher values. As the parameter $A$ can be negative, a quadratic shape is observed, indicating a similar trend to that of $\theta$.

parameters and highlight the preference for an inclination close to 90°.

Comparing with Simulation 1, it is interesting to note that the estimated errors are much smaller when errors and dispersion are added. This could be due to the likelihood model expecting errors outside the fixed values given in Simulation 1, preventing exploration of the entire likelihood.

To illustrate the performance of the methods, we focus on one specific simulated cluster with parameter values $\theta = 31.0^\circ$, $i = 63.5^\circ$, and $A = -2.3\,\mathrm{km\,s^{-1}}$. This example provides a visual representation of how the estimated parameters compare to the true values and highlights the differences between the MCMC and DE methods. The corner plots for this example cluster, simulated with real errors using both MCMC and differential evolution with MCMC (DE), are shown in the Appendix in Figures A.9 and A.10 respectively.

The simulations with real errors emphasise the challenges encountered when estimating rotational parameters using a solid body rotational model. It highlights the importance of carefully considering the initialisation process, error and dispersion terms, and understanding the impact of different methods on parameter estimation.

## Simulation 3

Our final simulation is an attempt to be a true representation of an observed cluster, parameterised in exactly the same configuration as Simulation 2, but with an added noise term to the transverse velocity components sampled from a normal distribution with $\mu = 0$ and

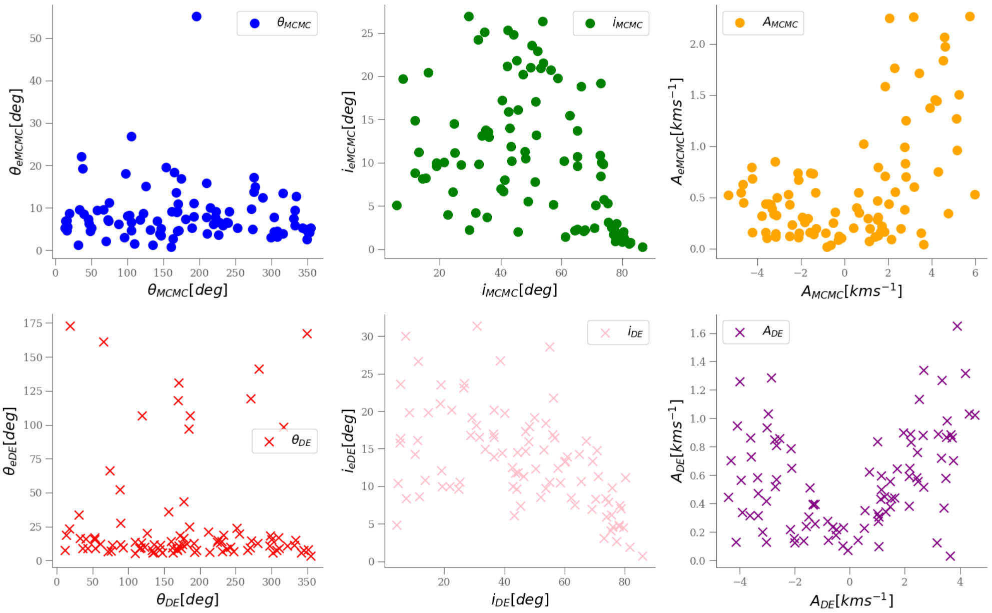


FIGURE 5.9: Simulation 1: The error distribution is examined for each parameter using two different sampling methods. $\theta$ has a random error distribution in both methods, and the errors on A get larger as the magnitude of A increases. For inclination however there is a clear negative linear-like trend illustrating that a cluster which is inclined closer to 90 has a lower error estimate than face on.

$\sigma = 0.05\,\mathrm{km\,s^{-1}}$for $v_X$ and $v_Y$, and $\sigma = 2\,\mathrm{km\,s^{-1}}$for $v_Z$. These sigma values are the same as in Simulation 1, and are representative of Gaia measurement errors.

Following the same structure of analysis as described in Simulation 2, the noticeable difference in Fig. 5.16, 5.17, 5.18 and 5.19 in comparison with the corresponding figures for Simulation 2 is that the errors are marginally larger, but the rotational parameters still follow the same expected trends. This suggests that applying the methodology to real data, which has inherent measurement errors, intrinsic dispersion and noise, may yield sensible parameter and error estimates.

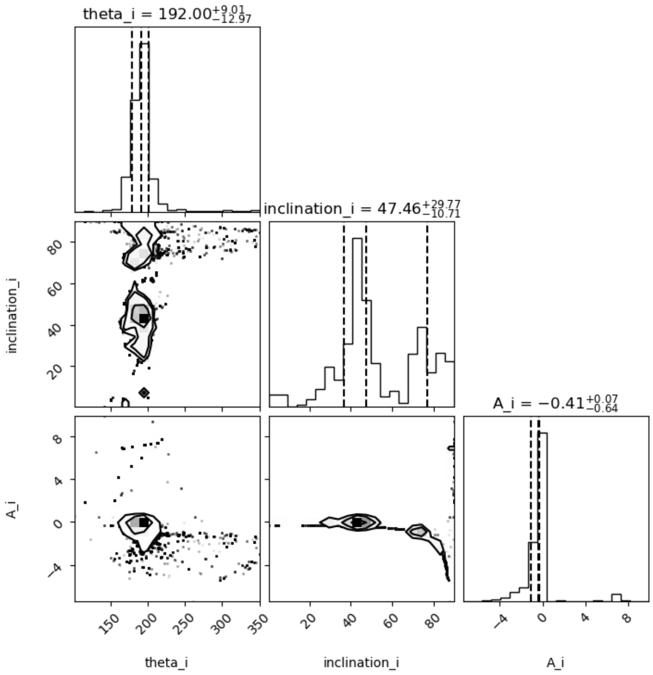


Figure 5.10: Corner plot for the simulated cluster with $\theta = 181^\circ$, $i = 49.0^\circ$, and $A = -0.6\,\mathrm{km\,s^{-1}}$, estimated using MCMC.

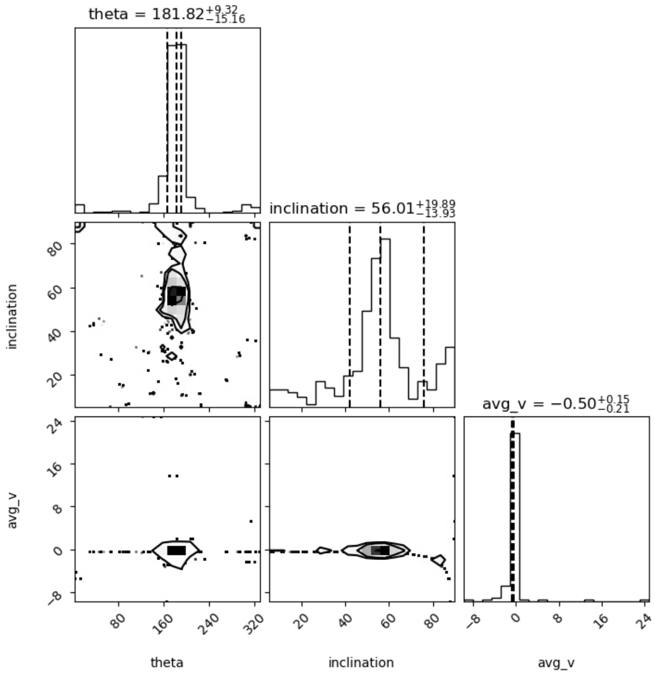


Figure 5.11: Corner plot for the simulated cluster with $\theta = 181^\circ$, $i = 49.0^\circ$, and $A = -0.6\,\mathrm{km\,s^{-1}}$, estimated using differential evolution and MCMC.

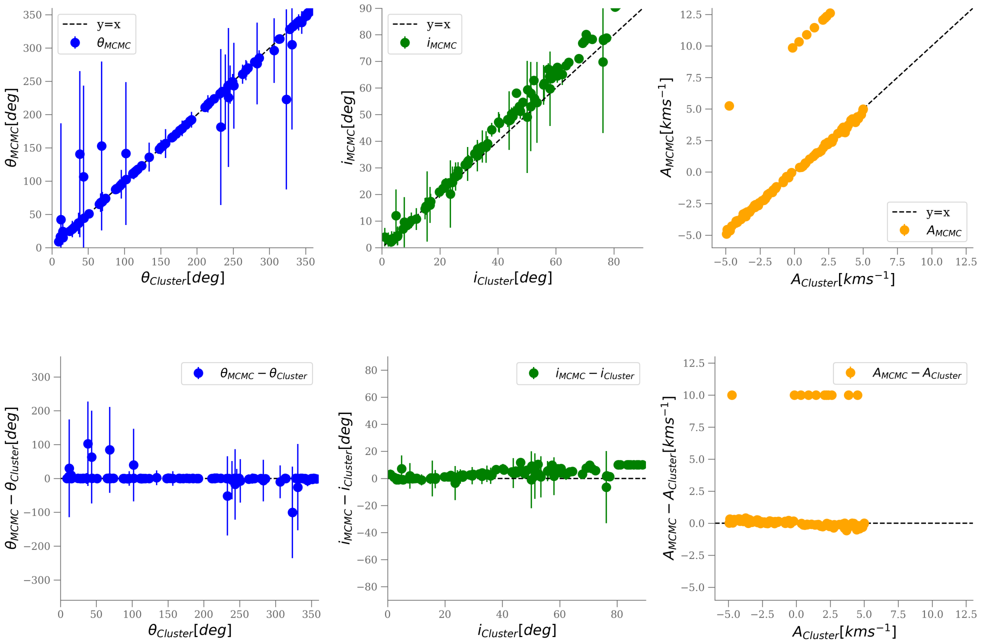


FIGURE 5.12: Simulation 2: We compare the estimated and actual cluster parameters using the MCMC method. Upon incorporating real errors, the MCMC method demonstrates accurate performance for the parameter $\theta$ with minimal error and dispersion for a majority of points. However, inclination is consistently overestimated, implying an issue with the initialisation, and exhibit substantial errors. The inclination difference panel follows a pattern akin to Fig. 5.5. Notably, the parameter $A$ reveals the points that have an issue with initialisation, while most points align well with the true cluster value.

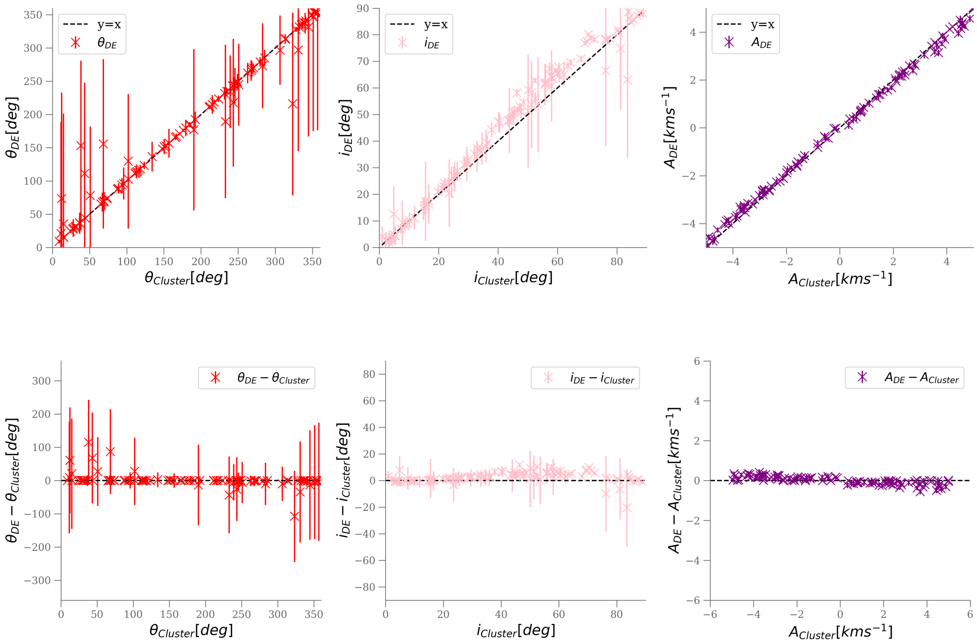


Figure 5.13: Simulation 2: The comparison between estimated and actual cluster parameters using the DE method. $\theta$ has as similar distribution of errors to the MCMC method. DE consistently overestimates inclination across the majority of the range. Similar to the noise-free scenario, parameter $A$ is underestimated when $A < 0$ and overestimated when $A > 0$, and exhibits lower errors across the range.

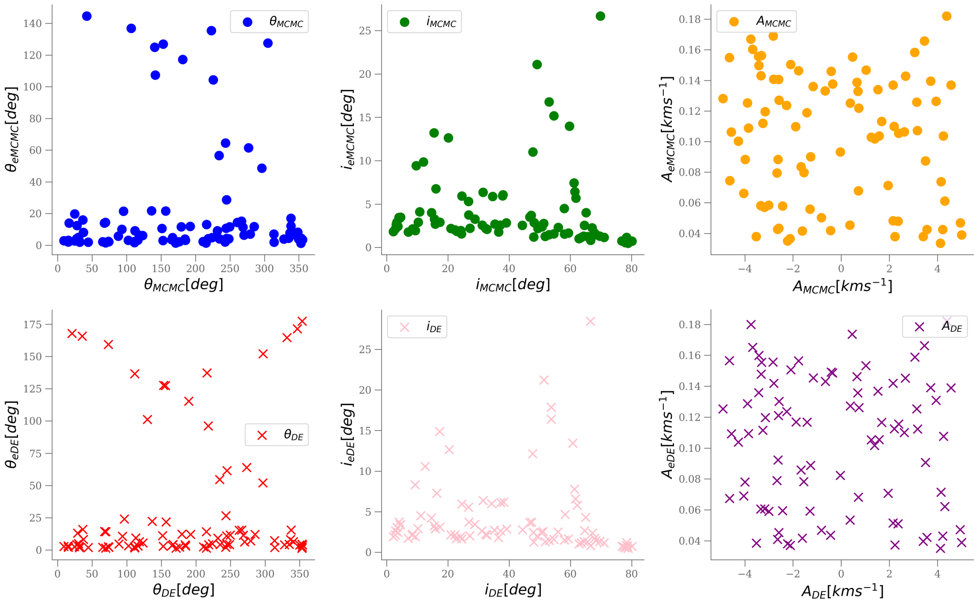


FIGURE 5.14: Simulation 2: The errors as a function of parameter for the MCMC and DE methods. The angular parameters show a few points with large errors with no trend seen in $\theta$ but a slight negative linear trend in inclination is noticeable, indicating that the method favours an axis of rotation close to 90°. $A$ has a random distribution of errors in both methods, which are relatively small in comparison to the expected range of $A$.

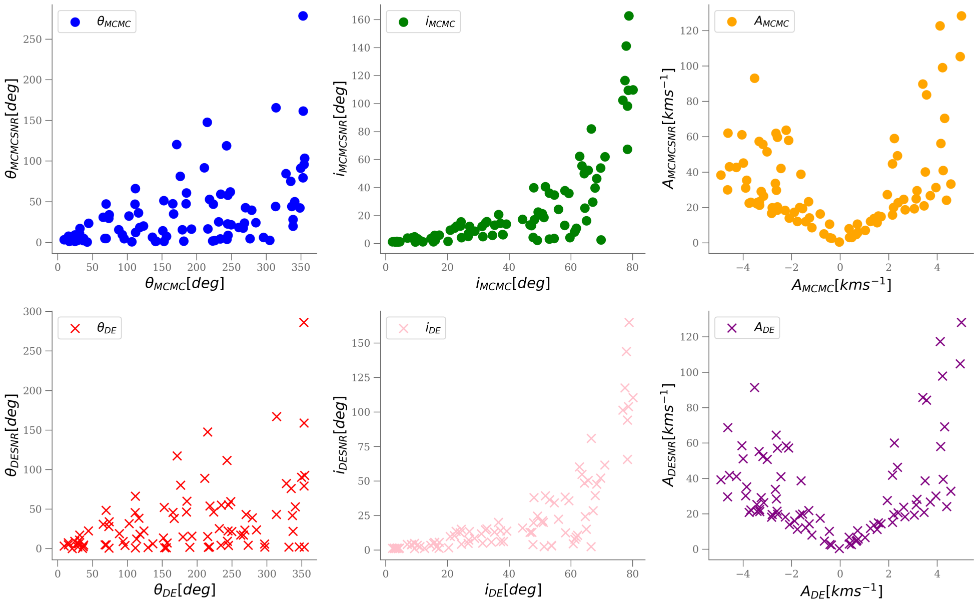


FIGURE 5.15: Simulation 2: The signal-to-noise ratio as a function of each parameter. As expected we see $\theta$ follows a roughly linear trend, inclination is favoured for higher values, and $A$ follows a positive parabolic distribution.

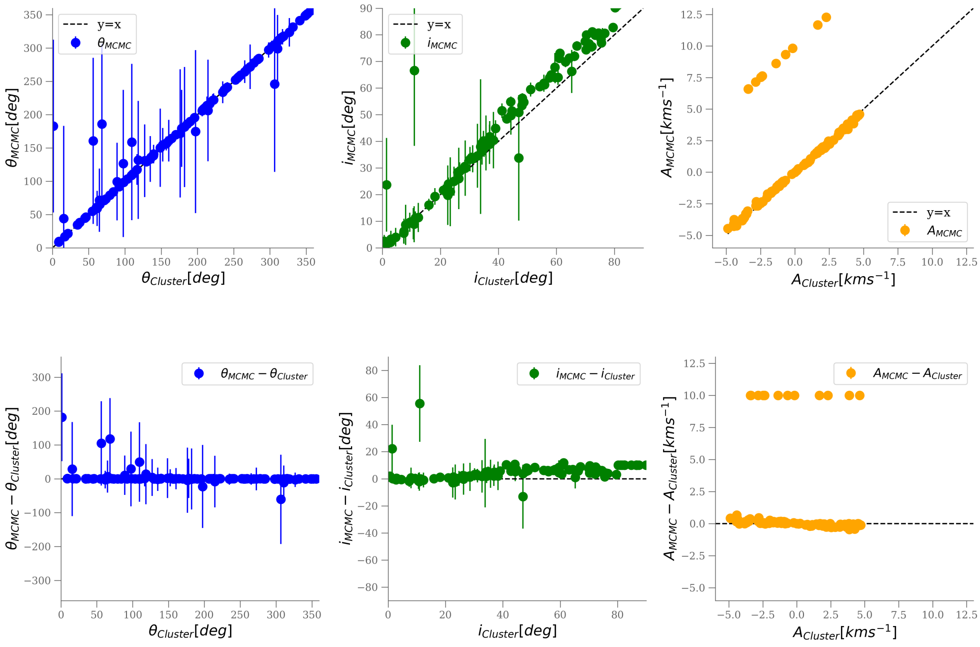


Figure 5.16: Simulation 3: The results for the MCMC method are similar to those of Fig. 5.12 in Simulation 2, with slightly larger errors for a few more points, but generally following the same trends.

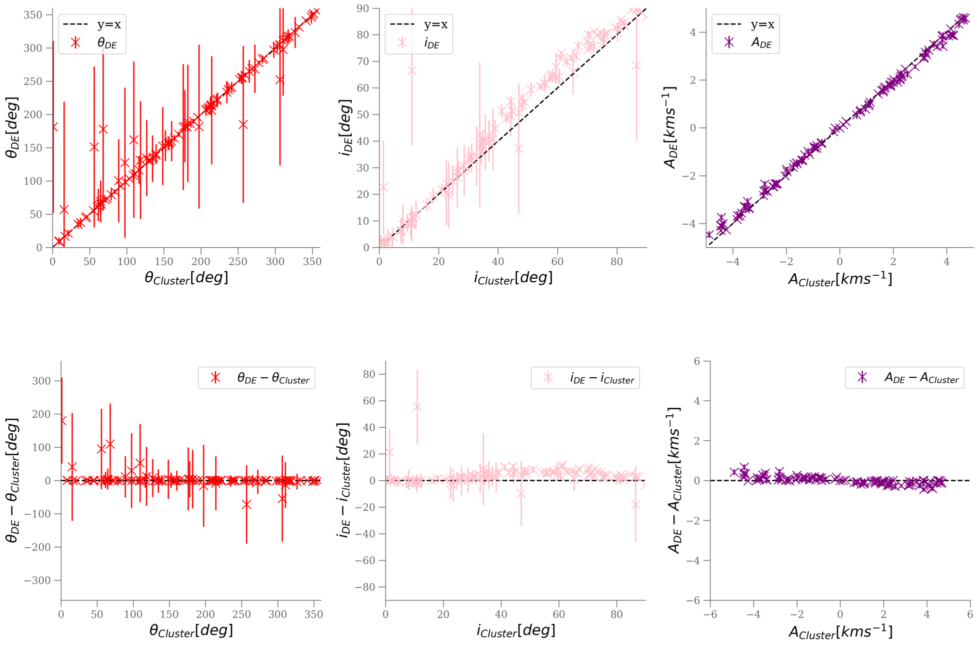


Figure 5.17: Simulation 3: The comparison between estimated and actual cluster parameters using the DE method is similar to that of Fig. 5.13, except for the outlier at an inclination of 10, with a large spread in errors.

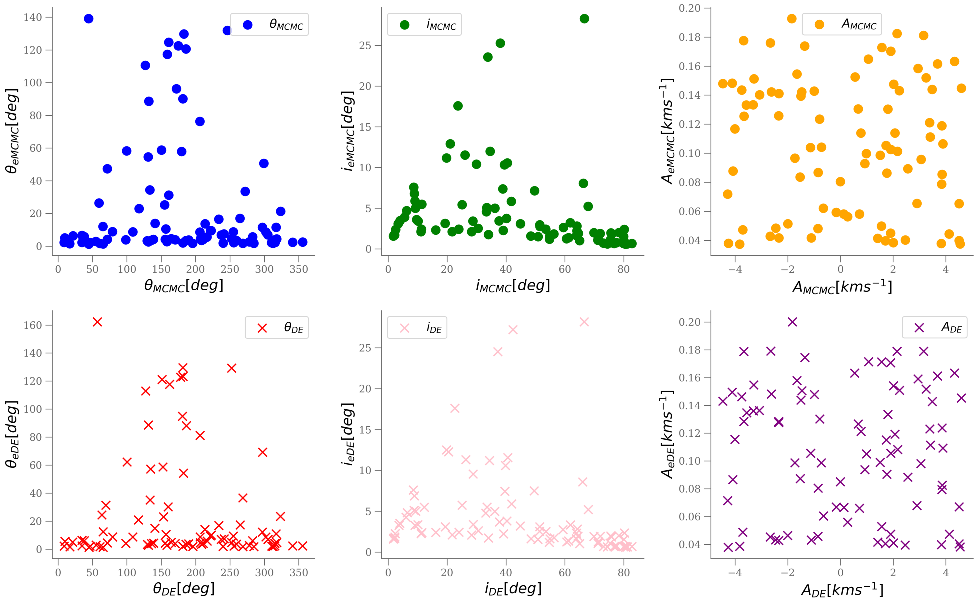


FIGURE 5.18: Simulation 3: Similar to Fig. 5.14, the errors are distributed as expected.

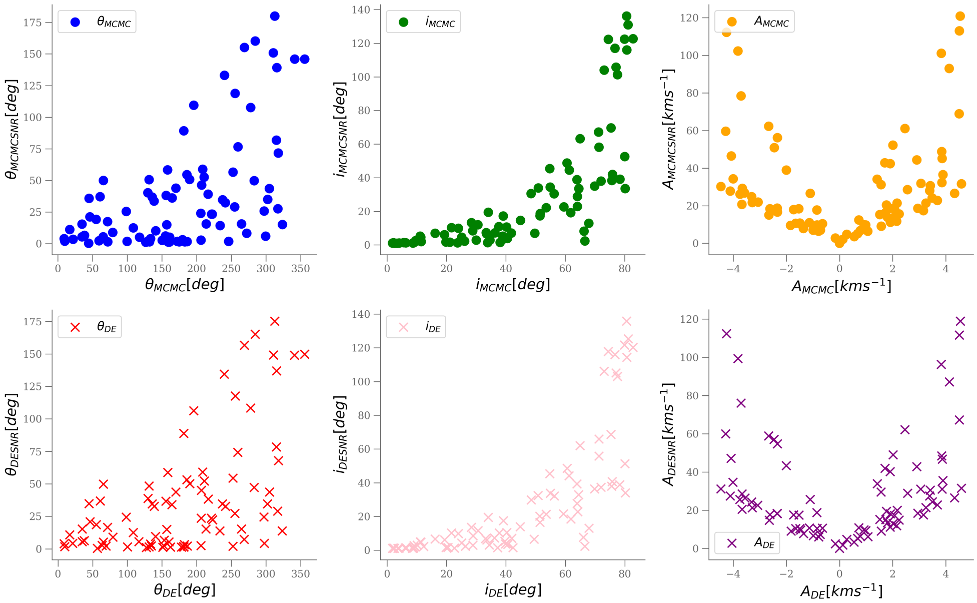


FIGURE 5.19: Simulation 3: As expected we see $\theta$ follow a slightly linear trend, inclination is favoured for higher values, and $A$ follows a positive parabolic distribution like that of Fig. 5.15.

Table 5.2: Table of known GCs with rotation, as measured by Sollima et al. (2019).

| ClusterName | $\theta_S$ | $\theta_{eS}$ | $i_S$ | $i_{eS}$ | $A_S$ | $A_{eS}$ |
|---|---|---|---|---|---|---|
| NGC104 | 224.30 | 4.60 | 33.60 | 1.80 | -5.00 | 0.32 |
| NGC2808 | 36.10 | 8.40 | 88.50 | 10.30 | -2.25 | 0.56 |
| NGC5139 | 170.20 | 7.60 | 39.20 | 4.40 | 4.27 | 0.52 |
| NGC5904 | 221.60 | 6.00 | 42.60 | 3.20 | 4.11 | 0.42 |
| NGC6205 | 165.50 | 14.20 | 85.90 | 11.60 | -1.53 | 0.61 |
| NGC6266 | 104.20 | 46.10 | 15.00 | 12.80 | 6.22 | 1.53 |
| NGC6273 | 56.90 | 13.20 | 41.90 | 7.10 | 4.19 | 1.12 |
| NGC6397 | 8.60 | 15.60 | 72.80 | 11.90 | -0.48 | 0.17 |
| NGC6541 | 83.20 | 18.30 | 65.40 | 13.90 | -3.73 | 1.15 |
| NGC6553 | 237.70 | 38.40 | 75.60 | 29.50 | 2.33 | 0.82 |
| NGC6626 | 28.60 | 17.70 | 83.50 | 13.30 | -2.42 | 1.08 |
| NGC6656 | 252.80 | 9.20 | 62.10 | 6.30 | 3.38 | 0.71 |
| NGC7078 | 52.60 | 28.80 | 15.40 | 5.40 | 3.29 | 0.51 |
| NGC7089 | 346.60 | 12.10 | 52.90 | 11.20 | -3.01 | 0.70 |
| Ter5 | 260.40 | 48.50 | 26.90 | 34.60 | 7.97 | 2.38 |

### 5.4.2 Test on Known Globular Clusters

This section examines the application of the method to globular clusters, for which parameters were previously estimated by Sollima et al. (2019), and assesses whether the addition of more or alternative sources of radial velocities other than Gaia improves the estimates of the rotational parameters. Table 5.2 shows the GCs considered and their respective rotational parameters as determined in Sollima et al. (2019). Similar to Sollima et al. (2019), we identify the cluster members from Gaia DR3 by selecting a radius of 3 arcminutes around the center of each cluster. The cluster centers are determined based on the Baumgardt and Hilker (2018) catalogue.

In the case of Gaia, it is worth noting that crowding in the central regions of globular clusters can affect the acquisition of radial velocity values. To mitigate this, we test supplementing the radial velocities obtained from Gaia with those from the Baumgardt and Hilker (2018) catalogue (henceforward referred to as the BH catalogue). It is important to acknowledge that the derived values may not match exactly with those of Sollima et al. (2019) due to differences in the membership list;it is to be expected that the various parameters calculated for clusters are sensitive to the selection of cluster members considered.

The section is organised as follows: we first analyse the clusters with known rotations from Sollima et al. (2019) using Gaia-only radial velocities, then we repeat the analysis using radial velocities from Gaia supplemented by the BH catalogue (Baumgardt and Hilker, 2018). After having done the general assessment we consider a more in-depth exploration of two clusters, NGC104 and NGC5139.

The simulations in Section 5.4.1 did not require the calculation of a systemic velocity, as we took the centre of the cluster to be at $(0, 0, 0)$. Hence in the following sections, which use real clusters, a calculation of the systemic velocities using the likelihood function in Eqn. 5.7 is required. We will discuss this process when we consider the two clusters NGC104 and NGC5139.

### Gaia-only Radial Velocities

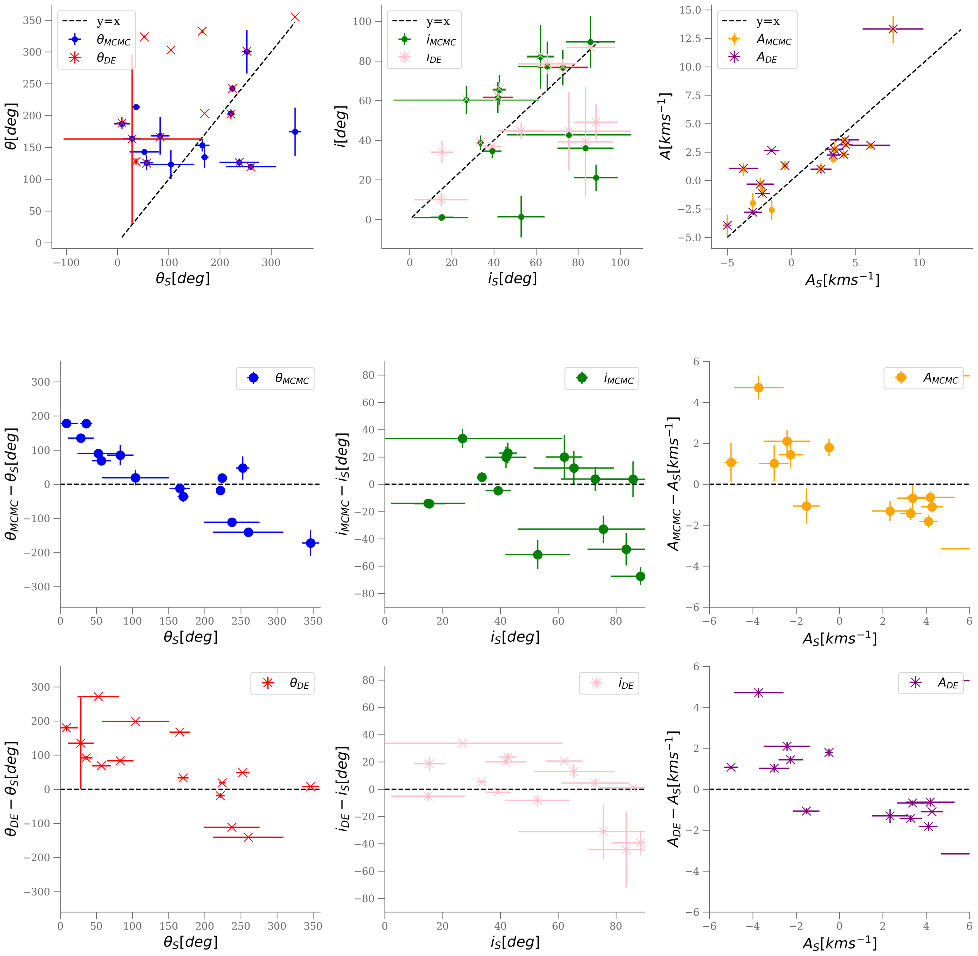


Figure 5.20: Globular clusters - Gaia only: Panels are denoted a-i going from top-left to bottom-right. Panel a) The estimate of $\theta_{MCMC}$ is more spread than for $\theta_{DE}$ with many more points being linearly aligned in the latter. The variation between the methods in panels b) and c) is minimal, however looking at the difference panels for inclination the spread is much less for the DE method.

To evaluate the effectiveness of applying a 3D solid body rotation model for determining rotational parameters of clusters, we first investigate whether using radial velocities from Gaia alone is sufficient, by comparing the results with known cluster properties. In this analysis, we employ the MCMC and DE sampling methods, initialising the parameters as $\theta_0 = \bar{\theta} + \epsilon$, $i = i_S + 10 + \epsilon$, and $A = A_S + 10 + \epsilon$, where $\epsilon$ represents a small amount of Gaussian noise. The values of $i_S$ and $A_S$ are from Sollima et al. (2019) and are listed in Table 5.2.

Figure 5.20 presents a comparison of the estimates obtained using MCMC and DE methods with the known parameters. We observe that the estimate of $\theta$ in the MCMC method exhibits a wider spread compared to the DE method, which demonstrates a more linear alignment of points. The variation between the methods in the inclination and $A$ panels is minimal.

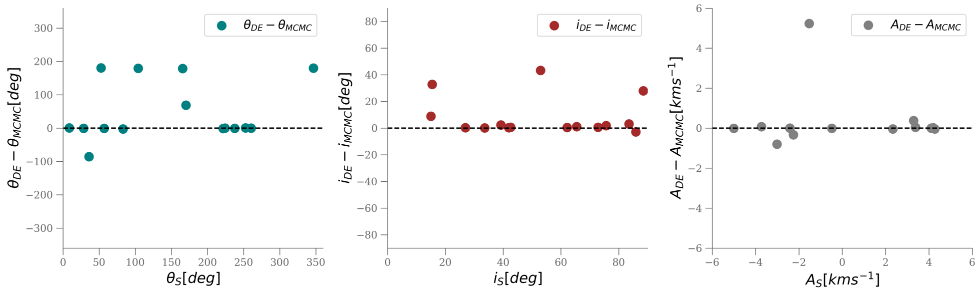


Figure 5.21: Globular clusters - Gaia only: The spread between the MCMC and DE methods. There are no noticeable trends in $\theta$, but with regards to inclination the DE method tends to have higher estimates, and DE tends to overestimate $A$.

However, upon examining the difference panels for inclination, we find that the spread is significantly narrower for the DE method. Both $\theta$ and $A$ exhibit similar spreads across both methods, but the DE method yields lower error estimates.

Further analysis of the differences between the MCMC and DE methods is presented in Fig. 5.21. We observe no noticeable trends in $\theta$, while for inclination, the DE method tends to produce higher estimates compared to MCMC. In terms of the parameter $A$, the DE method generally overestimates its value. These findings align with our expectations based on the signal-to-noise ratio plots observed in our simulations. Specifically, we anticipate a linear distribution for $\theta$ and inclination, and a quadratic shape for $A$. Figure 5.22 confirms these expectations, with the MCMC method exhibiting a randomly spread distribution for $\theta$, and the DE method being influenced by smaller errors. Inclination shows a similar but non-linear distribution in both methods, with the DE method demonstrating a higher signal-to-noise ratio. As for parameter $A$, the DE method follows the anticipated shape, exhibiting an overall higher signal-to-noise ratio.

Finally, we evaluate the accuracy of our estimates in relation to the number of radial velocities utilised for estimating the rotational parameters, as depicted in Fig. 5.23. It is evident that the estimates obtained using a higher number of radial velocities are more closely aligned with the known values documented in the literature. This observation highlights the importance of having a larger number of radial velocities, as it leads to improved estimation accuracy.

### Gaia and BH Catalogue Radial Velocities

After successfully demonstrating the feasibility of obtaining reasonable estimates for the rotational parameters through the utilisation of Gaia radial velocities, and establishing a clear correlation between the number of radial velocities and the accuracy of the estimates, our current focus shifts towards testing this idea, by supplementing the Gaia radial velocities with those obtained from the BH catalogue. This augmentation aims to explore the potential improvement in the estimation process. To evaluate the efficacy of this approach, we conduct a diagnostic analysis, similar to the one described in Section 5.4.2.

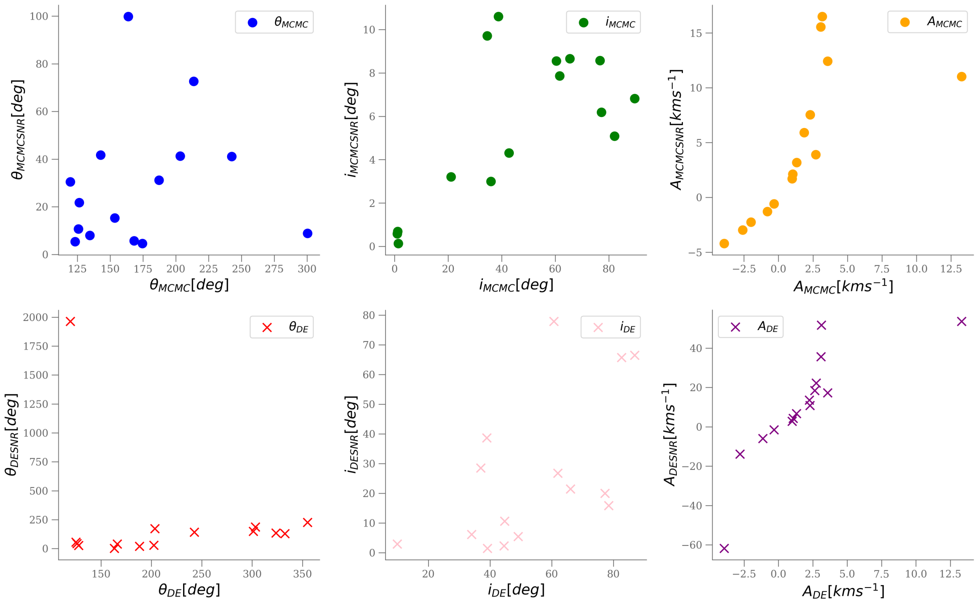


FIGURE 5.22: Globular clusters - Gaia only: The MCMC method displays a scattered distribution for $\theta$, while the DE method is affected by smaller errors, resulting in a more concentrated distribution. Inclination also exhibits a similar but non-linear distribution in both methods, with the DE method showing a higher signal-to-noise ratio. The parameter $A$ follows the expected pattern in the DE method, with a generally higher signal-to-noise ratio compared to the MCMC method.

Figure 5.24 reveals a substantially more prominent alignment with the identity line when compared to the results obtained solely based on Gaia radial velocities. This enhanced alignment is reflected in the difference panels, which exhibit a more concentrated distribution around the zero difference line. Notably, the addition of more radial velocities contributes to minimising the discrepancies between the estimates obtained through the MCMC and the DE methods, as depicted in Fig. 5.25.

Furthermore, Fig. 5.26 serves to reinforce the notion that a greater number of radial velocities leads to more accurate estimates. The plot clearly illustrates the correlation between the number of radial velocities utilised and the improved alignment of the estimates with the known values. This finding highlights the significance of having a dataset with as many radial velocities as possible when estimating rotational parameters. The combination of Gaia radial velocities and supplementary data from the BH catalogue demonstrates considerable potential for improving the accuracy of calculated rotational parameters.

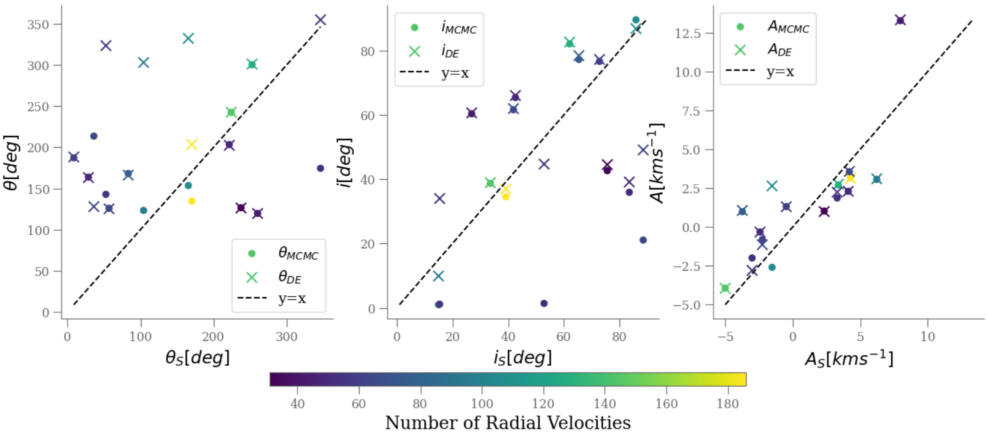


Figure 5.23: Globular clusters - Gaia only: The following plots show estimates compared with literature values, but coloured by the number of radial velocity values. On average, the more accurate estimates of the rotational parameters (i.e., those in closest agreement with the literature) have a higher number of radial velocities.

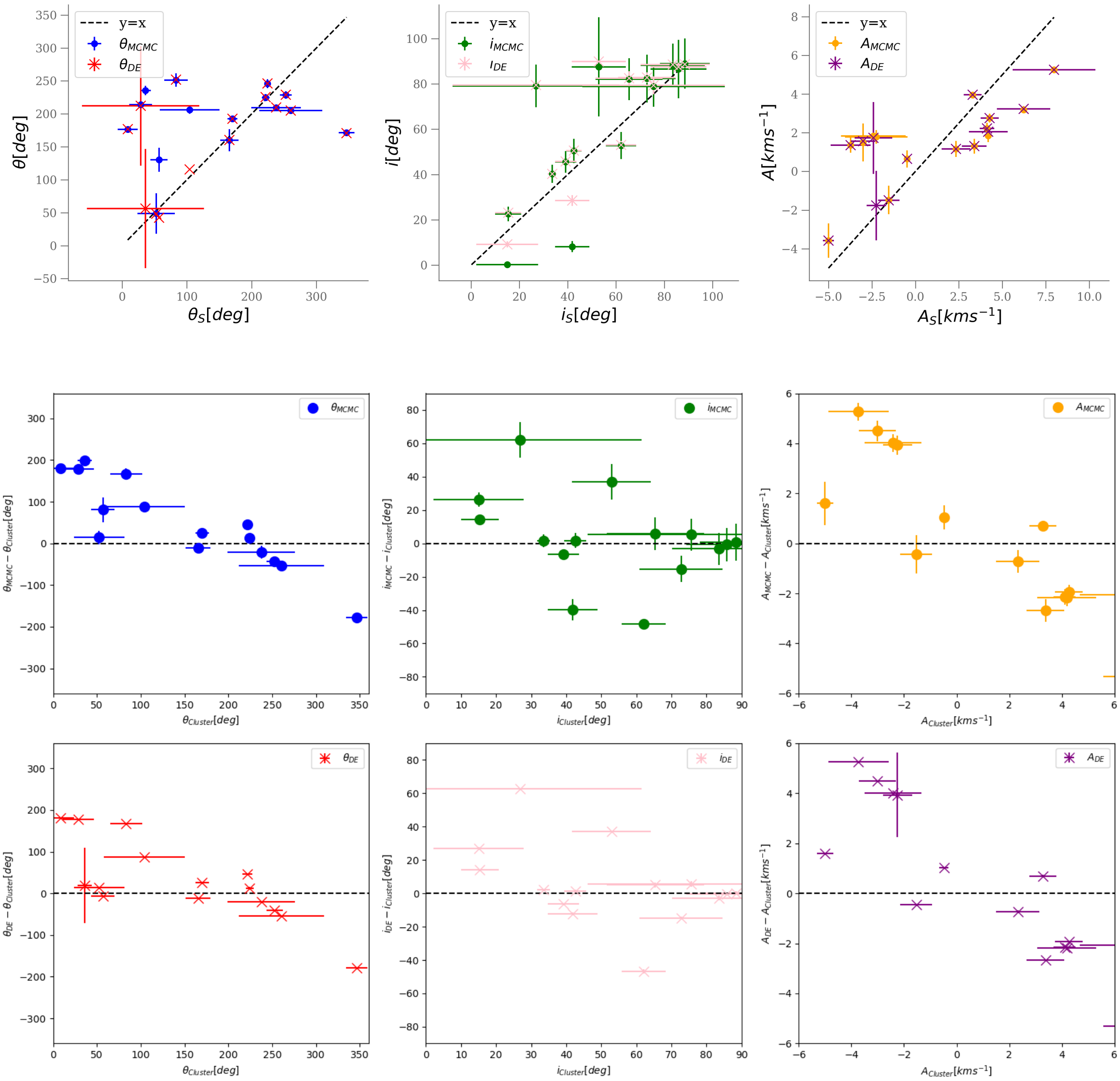


FIGURE 5.24: Globular clusters - Gaia and BH Catalogue: In line with the arrangement depicted in Fig. 5.20, we observe a distinct linear alignment of the estimates for panels a-c, which is even more pronounced than in the Gaia-only scenario. This alignment is further reflected in the subsequent difference panels, as they exhibit a tighter distribution centered around the zero-difference line.

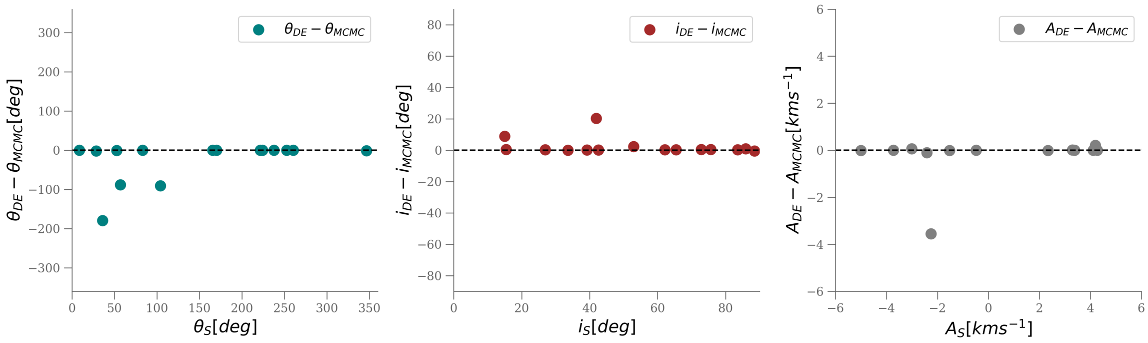


Figure 5.25: Globular clusters - Gaia and BH Catalogue: The addition of more radial velocities yields a minimal difference between the estimates from MCMC and DE methods.

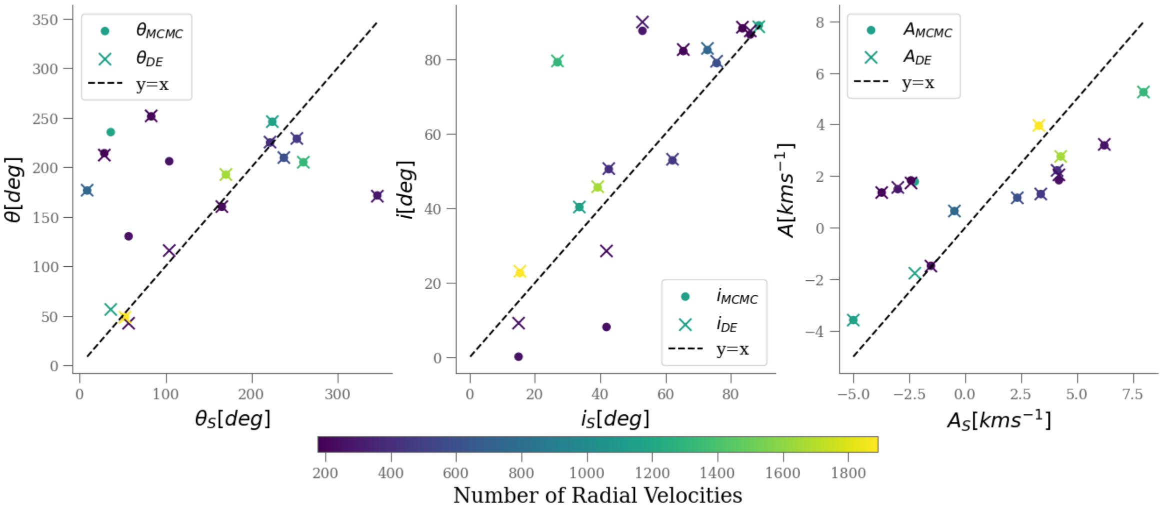


Figure 5.26: Globular clusters - Gaia and BH catalogue: Here we see a tighter distribution in the estimates of the rotational parameters to Fig 5.23 , and similarly the more radial velocities when estimating rotational parameters the better.

## NGC104

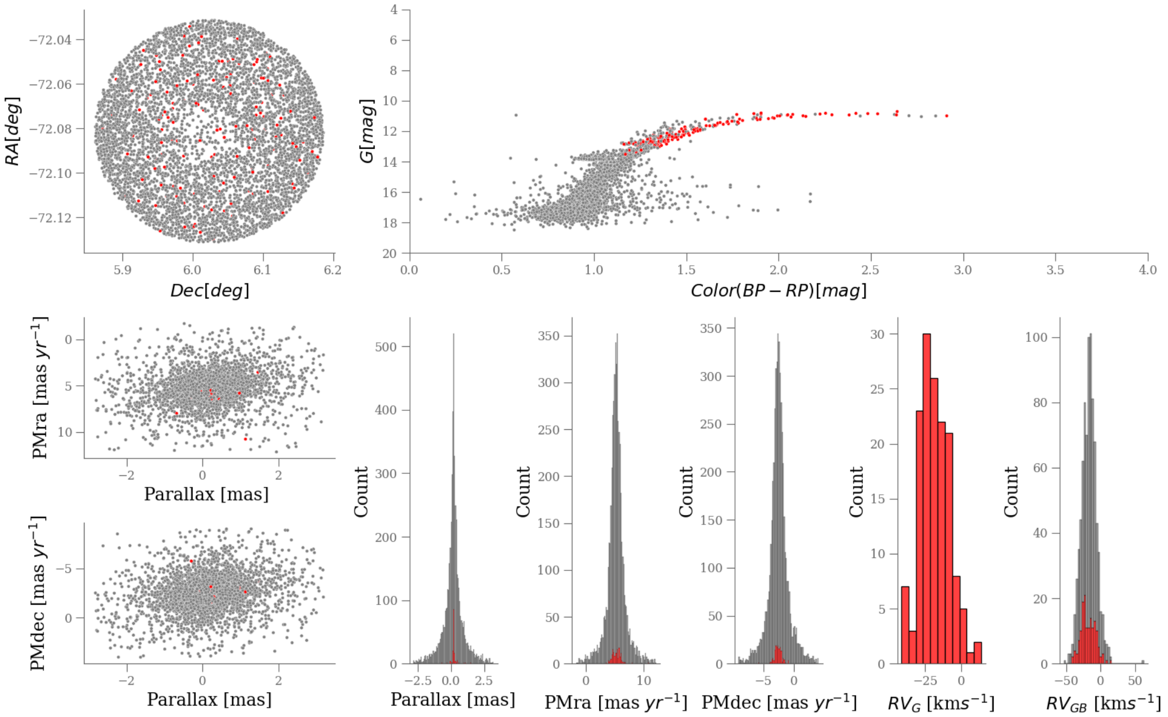


FIGURE 5.27: Globular clusters: Properties of the parameters for the member stars (grey) of NGC104, as well as those for the members with radial velocity measurements (red) in Gaia. Panel (1): distribution on the sky. Panels (2) and (3): diagrams of parallax vs. proper motions. Panel (4): color magnitude diagram. Panels (5), (6), (7), (8) and (9): histograms of the parallaxes, proper motions, radial velocity in Gaia, radial velocity in Gaia and the supplementary BH catalogue, respectively.

NGC 104, also known as 47 Tucanae or simply 47 Tuc, is a prominent GC located in the constellation Tucana in the southern hemisphere. It is at a distance of 4.45 ± 0.01 kpc, with an RA of 305.895333 and a Dec of -44.889114, has an estimated age of 13 Gyr and is the second brightest GC after Omega Centauri. The rotational parameters of NGC 104 as calculated in Sollima et al. (2019) are $A = -5.00 \pm 0.32$, $\theta = 224.3 \pm 4.6$ and $i = 33.6 \pm 1.8$. We highlight this cluster because of its low estimated errors relative to other clusters, and the prior work that has been done to understand its rotation (Sollima et al., 2019; Anderson and King, 2003).

Figure 5.27 provides an overview of the properties of NGC 104 observed in Gaia, with the radial velocities shown in red on the colour-magnitude diagram. It is evident that Gaia lacks reliable radial velocity estimates for many cluster members. Given that radial velocity plays a crucial role in calculating rotational parameters, identifying clusters with an adequate number of radial velocities becomes pivotal. This is particularly relevant for OCs, as they generally have fewer members and a lower density of stars compared to GCs.

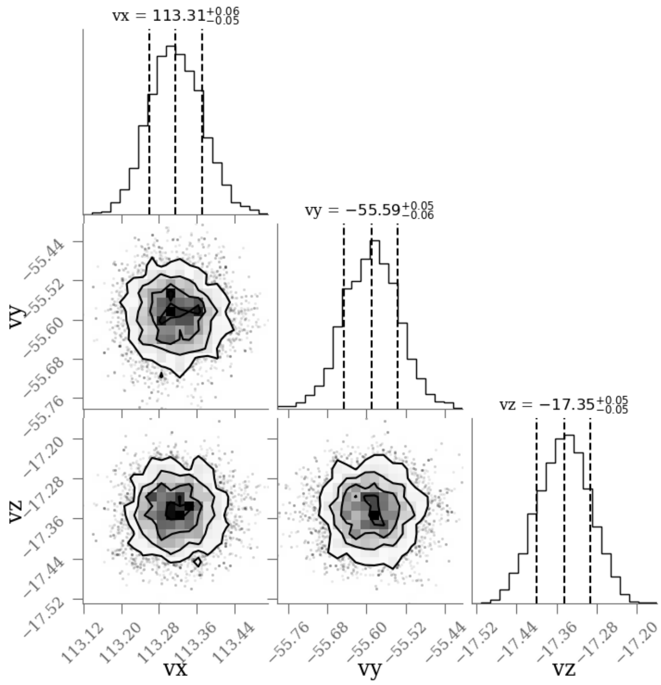


Figure 5.28: Globular clusters: The corner plot for the estimates of the systemic (central) velocities for NGC104.

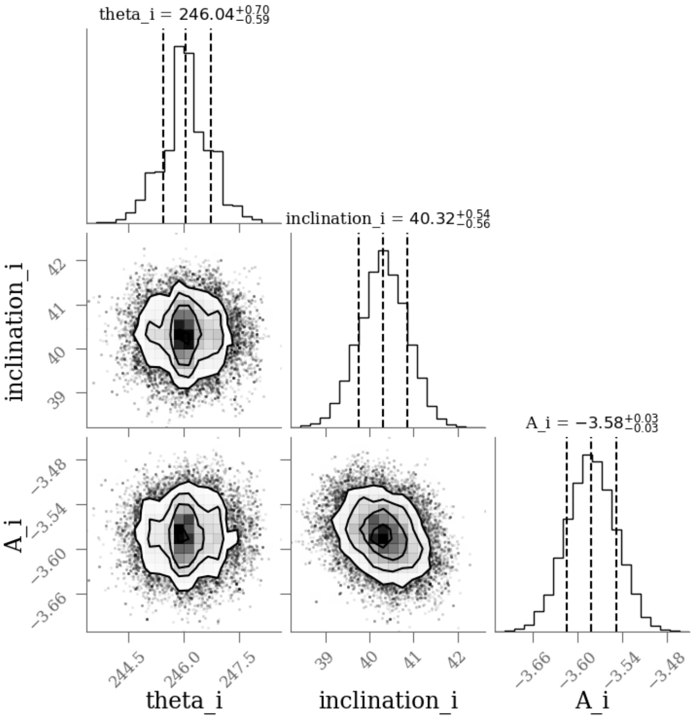


Figure 5.29: Globular clusters: The corner plot for the estimates of the rotational parameters for NGC104 using MCMC.

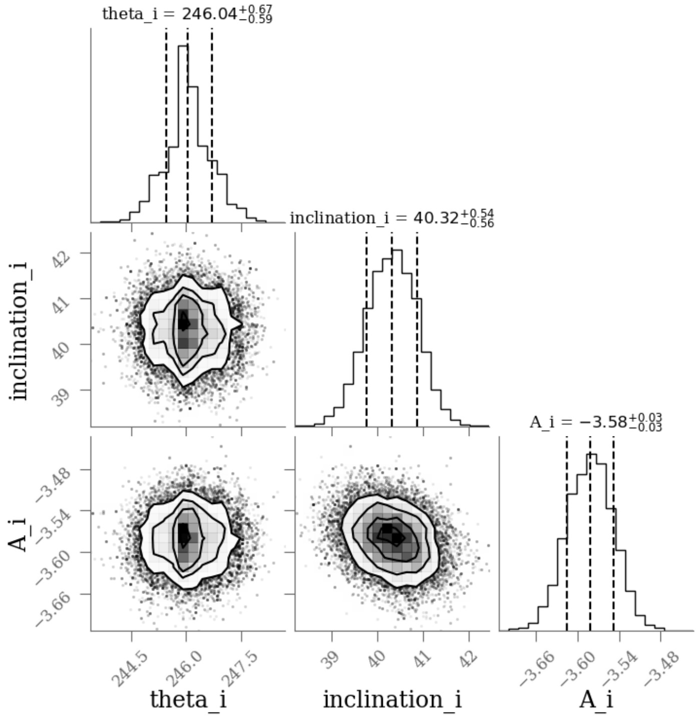


FIGURE 5.30: Globular clusters: The corner plot for the estimates of the rotational parameters for NGC104 using DE.

Unlike the case with simulated clusters, we need to estimate the systemic velocities of the cluster by maximising the likelihood of equation 5.7. The systemic velocity values obtained are $< v_X >= 112.84 \pm 0.03\,\mathrm{km\,s^{-1}}$, $< v_Y >= -55.83 \pm 0.03\,\mathrm{km\,s^{-1}}$, and $< v_Z >= -17.12 \pm 0.01\,\mathrm{km\,s^{-1}}$, as illustrated in Fig. 5.28. These values are then subtracted to obtain $v_X$, $v_Y$, and $v_Z$, respectively.

Next, the rotational parameters are estimated using the likelihood equation 5.8, employing both MCMC and DE methods, as shown in Figs. 5.29 and 5.30. Both methods yield consistent results for the rotational velocity $A$ of $-3.58$ and the inclination of 40.32, although the value of $\theta$ is less accurate at 246.04. Notably, a comparison with the values reported in Sollima et al. (2019), $A = -5.00 \pm 0.32$, $\theta = 224.3 \pm 4.6$, and $i = 33.6 \pm 1.8$, reveals a difference with our estimate for $\theta$ and rotational velocity within the error bounds. This is expected, as previously mentioned, due to differences in our selection of cluster members. Nevertheless, it is evident that the methodology performs adequately and provides valuable insights into the rotational parameters of NGC 104.

**NGC5139**

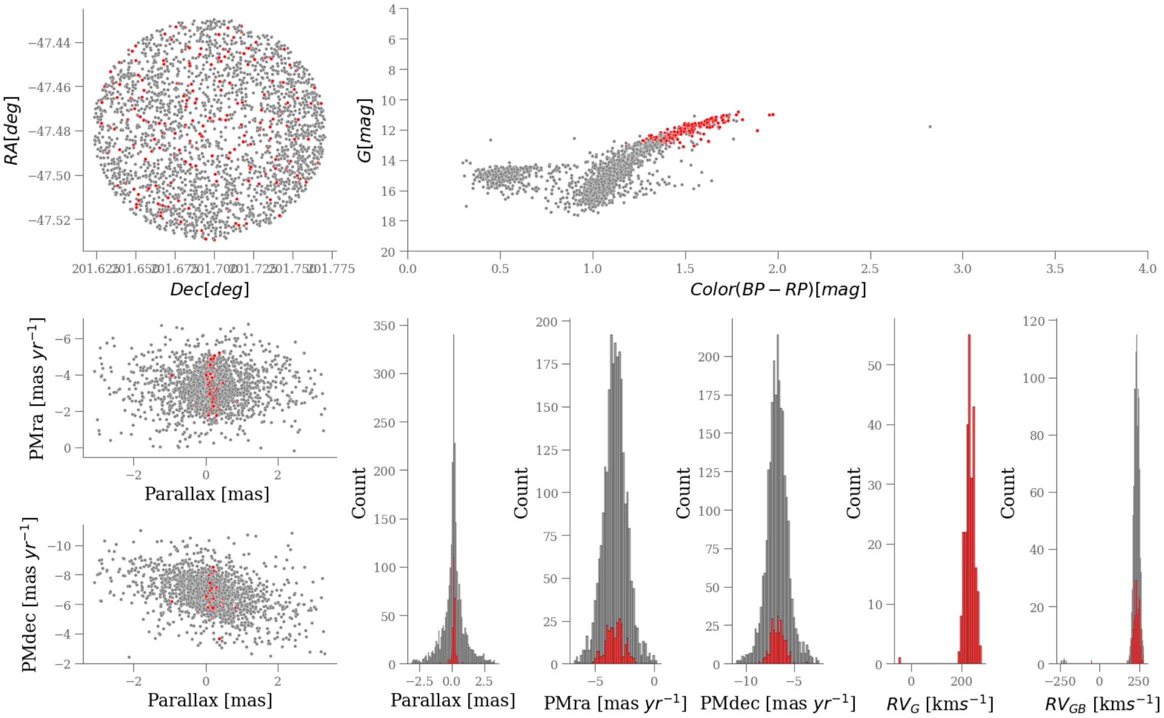


Figure 5.31: Globular clusters: Properties of the parameters for the member stars (grey) of NGC5139, as well as those for the members with radial velocity measurements (red) in Gaia. Panel (1): distribution on the sky. Panels (2) and (3): diagrams of parallax vs. proper motions. Panel (4): color magnitude diagram. Panels (5), (6), (7), (8) and (9): histograms of the parallaxes, proper motions, radial velocity in Gaia, radial velocity in Gaia and the supplementary BH catalogue, respectively.

NGC 5139, also known as Omega Centauri, is a GC situated in the constellation Centaurus. It is one of the Milky Way's most massive and luminous GCs. Positioned at a distance of approximately 5,240 parsecs from Earth at an RA of 309.10202 and Dec of 14.96833, NGC 5139 harbours an extraordinary population of around 10 million ancient stars. The cluster exhibits a distinct and intricate stellar distribution, indicating the presence of diverse stellar populations. Multiple generations of stars have been observed, suggesting a complex history of star formation. Notably, the cluster displays a prominent blue horizontal branch, indicative of a substantial number of evolved stars. These distinctive characteristics render Omega Centauri a vital target for investigating the formation, evolution, and dynamics of GCs and dense stellar systems in the Milky Way. For the purposes of this study, it has been chosen as a verification cluster alongside NGC 104 on account of its low error estimates and positive rotational velocity value, and because it is such a prominent and well-studied globular cluster (van Leeuwen et al., 2000). The analysis presented in Fig. 5.31 illustrates the limited number of radial velocities available from Gaia (269), especially in comparison to the BH catalogue (1660).

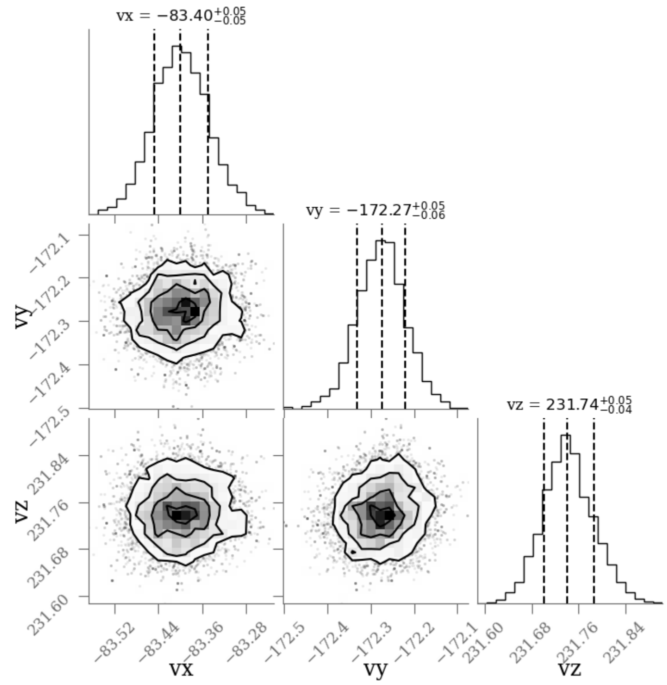


FIGURE 5.32: Globular clusters: Corner plot for the estimation of the systemic velocities for NGC5139.

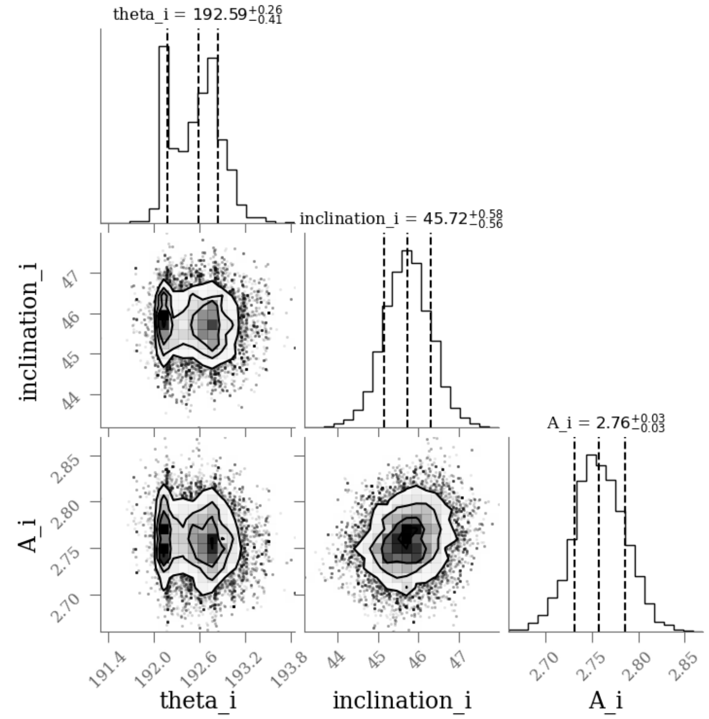


FIGURE 5.33: Globular clusters: The corner plot for the estimates of the rotational parameters for NGC5139 using MCMC.

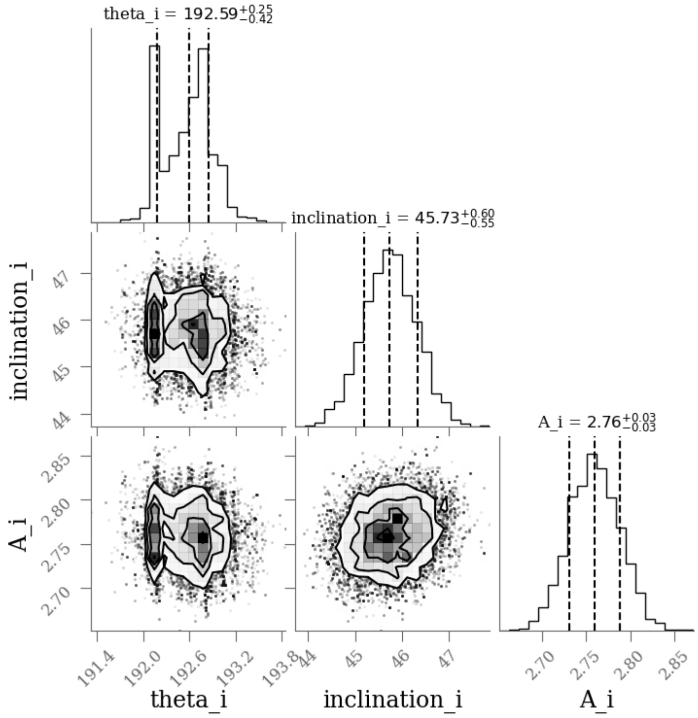


Figure 5.34: Globular clusters: The corner plot for the estimates of the rotational parameters for NGC5139 using DE.

Applying the same methodology outlined in Section 5.4.2, the systemic velocities for Omega Centauri, as illustrated in Fig. 5.32, are determined to be $< v_X >= -82.94 \pm 0.03$, $< v_Y >= -171.92 \pm 0.04$, and $< v_Z >= 230 \pm 0.01$. Comparing the rotational parameter values shown in Fig. 5.33 and 5.34 ($A$ of 2.76, the inclination at 45.73 and $\theta$ at 192.59) with the values reported in (Sollima et al., 2019) ($A = 4.27 \pm 0.52$, $\theta = 170.20 \pm 7.6$, and $i = 39.20 \pm 4.4$), we find our error estimates to be considerably smaller than estimated in (Sollima et al., 2019), and our parameter estimates close but not exact matches. A larger discrepancy is observed compared to NGC 104, however it is worth noting that this discrepancy could be at least partially due to differences in sample size.

### 5.4.3 Application to Open Clusters

OCs represent a major focus of astronomical research, facilitated by the advent of Gaia, which enables comprehensive exploration and characterisation of these clusters. After first employing simulated data and then comparing the efficacy of different methods in determining the rotational parameters of real GCs, we can now extend our analysis to OCs.

Cantat-Gaudin (priv. communication) offers a comprehensive compilation of 2000+ known OCs in Gaia DR3, including fundamental parameters such as Cartesian coordinates, and age estimates, alongside a potential member list. To refine our analysis, we consider only those clusters with more than 100 stars possessing radial velocities in Gaia, resulting in a reduced list of 131 clusters. Subsequently, we query the Gaia archive utilising the Gaia IDs obtained

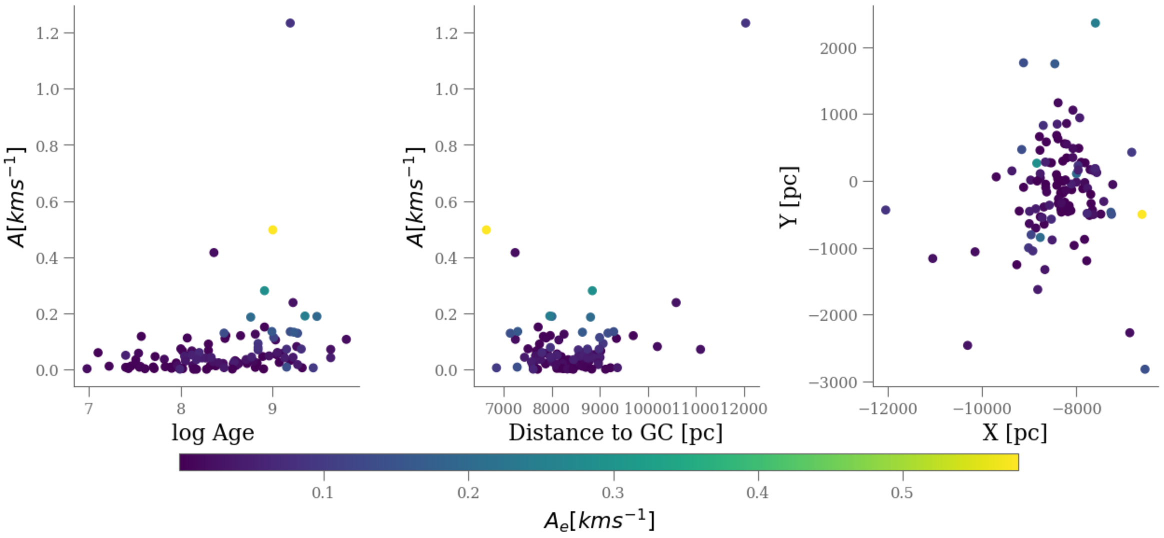


FIGURE 5.35: Open clusters: Parameters coloured by the error estimate for rotation. The relationship between the error estimate against age, distance to the Galactic centre and in the X-Y plane appears to be random; this may be a function of selection bias.

from Cantat-Gaudin's member list to obtain the necessary proper motion, error, photometric, and parallax data essential for calculating the rotational parameters.

To setup the analysis, we consider the initialisation of each parameter to be as follows: $\theta$ is set to its mean, inclination is set to 10 and rotation set to 5, with some small epsilon added. For the DE sampler, we consider 1000 steps with a burn-in of 500, and use 100 walkers. Our intrinsic cluster dispersion term is set to 1 $\mathrm{km\,s^{-1}}$, as in our simulations. This term is a potential source of error that may need to be tuned for each cluster individually – given OCs can have quite distinct morphologies in comparison to GCs – rather than being assumed to be the same for all clusters.

The results of the analysis are presented in Table 5.3, where the table is ordered based on candidates with the highest absolute rotational velocity. The rotational velocities for these top 5 clusters are $> 0.2\,\mathrm{km\,s^{-1}}$, with low errors. We present a brief analysis of the rotational parameters, looking at diagnostics and potential science cases, with the general cluster parameters taken from Cantat-Gaudin.

We evaluate the estimated rotational parameters for the open clusters, by distinguishing those estimates which have larger errors from the rest to see if there are any characteristics that drive the difference. We may expect clusters located on the periphery of the X-Y plane to have lower error estimates and more radial velocities measured, due to not being affected by the issue of crowding and its impact on radial velocity estimates in Gaia.

In Fig. 5.35, we plot rotation as a function of age of the cluster, distance to the Galactic centre and on the X- Y plane, coloured by the error estimate of $A$, which we label panels a), b) and c) respectively. In panel a) there is a slight tendency towards older clusters having higher rotational errors than the rest, in panel b) there is no discernible trend and similarly in panel c) there is no easily identifiable relationship. This analysis may be dominated by exceedingly small errors ($< 0.01$) not revealing the expected error structure in the X-Y plane.

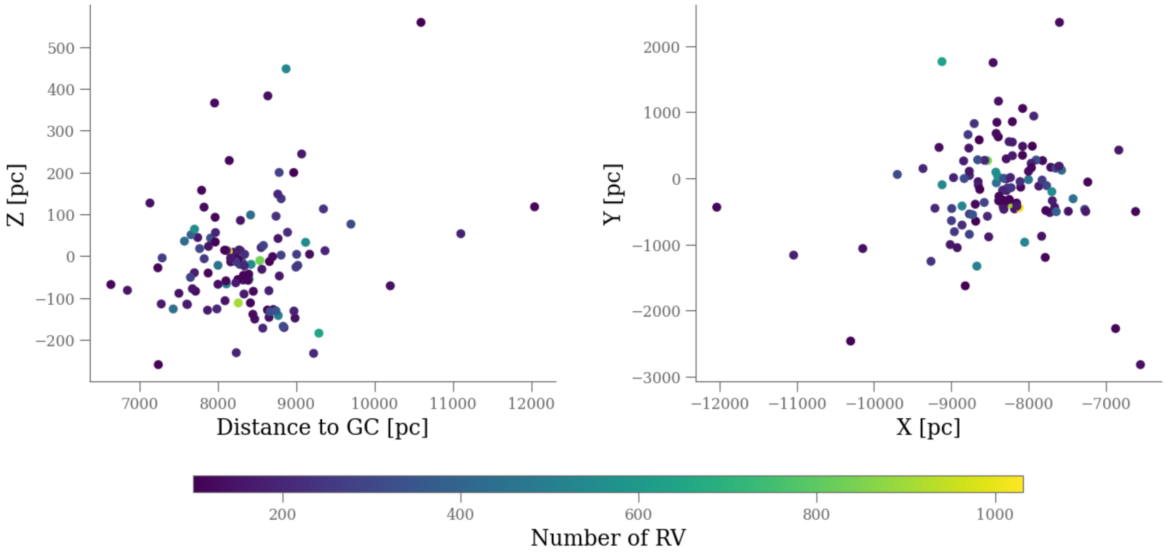


Figure 5.36: Open clusters: The spatial distribution of our open clusters, with colour representing the number of radial velocities used in the estimation. Our hypothesis that clusters with a higher number of known radial velocities in Gaia being on the edges of the X-Y plane is proven incorrect, but the accuracy of the radial velocities may be higher than in the central region.

In Fig. 5.36 we consider the spatial distribution of the OCs coloured by the number of radial velocities for each cluster. Unfortunately, the hypothesis of clusters located on the periphery of the X-Y plane having lower error estimates and more radial velocities is shown to be incorrect as clusters on the outer edges have a lower number of radial velocities than those more centrally located. However there is still some indication that the accuracy of the radial velocities is better on the extremities as can be seen in Fig. 5.37. There are still many alternative reasons other than crowding to explain the observed distributions, such as Gaia being poor at calculating radial velocities for young hot stars, but the following figures show that more accurate radial velocities are required to better constrain the rotational parameters.

We now turn our attention to the uncertainties estimated from the method shown in Fig 5.38 and consider each parameter's error against parameter value, but coloured by the rotational error in Fig. 5.39. The errors are quite discretised, and $\theta$ and inclination overall have large errors, which as shown in the tests in Section 5.4.1 makes sense due to their degeneracy. Relative to their actual parameters, Fig. 5.39 sees quite a random spread in the errors for $\theta$ and rotation even when compared with rotation error, but interestingly errors for inclination seem to asymptotically decrease when approaching 90°; at an inclination of 90°, the dominating motion would be radial velocity.

OCs serve as valuable tracers of the spatial structure of the young stellar population in the Galactic disk; they offer insights into reconstructing spiral arms and determining spiral pattern speeds (Castro-Ginard et al., 2021). However, the age-distribution analysis of OCs by Castro-Ginard et al. (2021) suggests a flocculent Milky Way with transient spiral arms, rather than the predicted age gradient of density wave or bar-driven spiral arms (Quillen, 2002). In

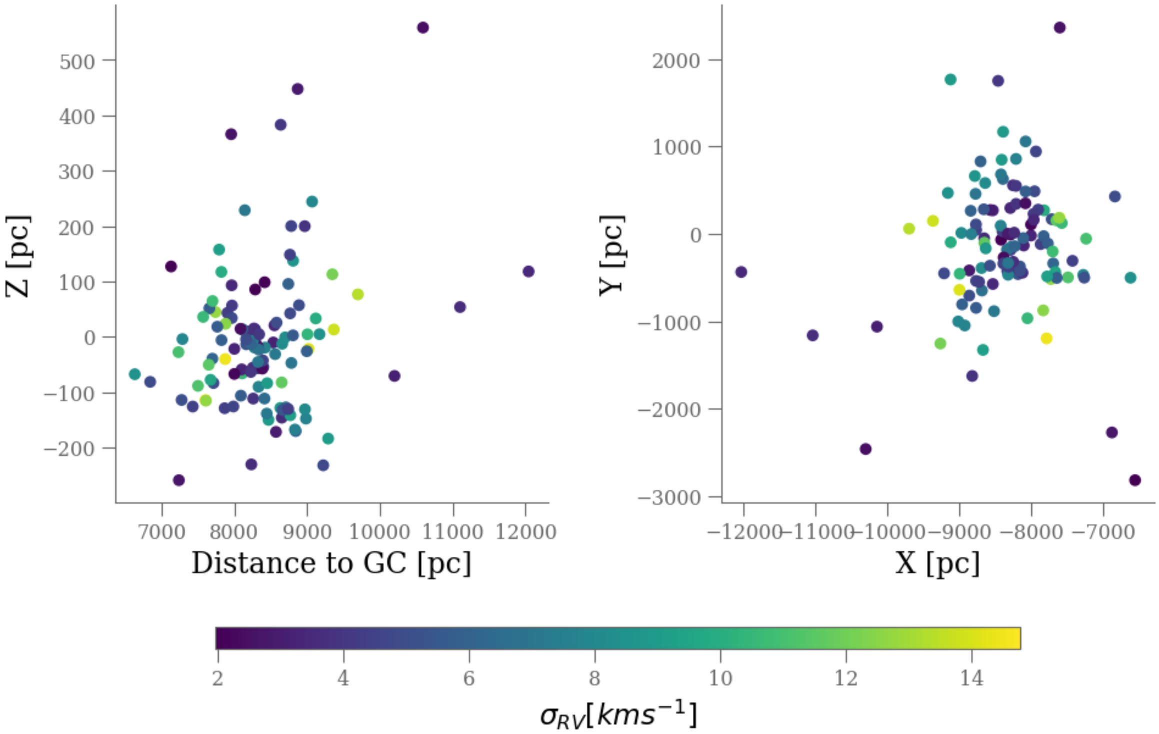


FIGURE 5.37: Open clusters: Similar to Fig. 5.36 but this time colour represents the uncertainty of the radial velocity. Noticeably the accuracy of the radial velocities decrease closer to the plane of the Galaxy as can be seen in the first panel, and towards the central region in the X-Y plot. This suggests that perhaps more accurate radial velocity measurements will be the driver of better rotational parameter estimates.

fact, Quillen (2002) identified multiple spiral features, each with a distinct pattern speed that decreases with Galactocentric radius. To gain a better understanding of the underlying processes driving these relationships, it becomes crucial to consider the internal rotation of OCs; and, as shown in Fig 5.40 and 5.41, a more extensive and accurate calculation of the rotational properties is necessary to identify any verifiable trends in age and position.

As an example, the Praesepe cluster is an intermediate-age open cluster (Gossage et al., 2018) located in the constellation Cancer at a distance of 1.7-1.9 kpc. Praesepe stands out due to its large proper motions (Kraus and Hillenbrand, 2007). Previous studies by Hao et al. (2022) and Healy et al. (2021) investigated Praesepe and reported rotation values of $0.2 \pm 0.05\,\mathrm{km\,s^{-1}}$ within its tidal radius and $0.132 \pm 0.027\,\mathrm{km\,s^{-1}}$, respectively. Our findings yield rotation parameters of $A = 0.023 \pm 0.0018$, with $\theta = 9.27 \pm 3.75$ and inclination $= 87.286 \pm 1.636$.

Another notable OC is NGC 6819, which has an age of 2.4 Gyr and is located at a distance of 2.4 kpc. Kamann et al. (2019) calculated its rotational velocity to be $0.05 \pm 0.05\,\mathrm{km\,s^{-1}}$. In our analysis, we determine its rotational properties to be $-0.18 \pm 0.49$. The literature values for both of these clusters are markedly different to what we obtained, but this may be attributed to the different rotation models considered, as here we consider 3D solid body and in the previous studies the model considered is plane of sky rotation. Additionally, as a quick test we changed our input velocity dispersion value to $2\,\mathrm{km\,s^{-1}}$ from $1\,\mathrm{km\,s^{-1}}$, to

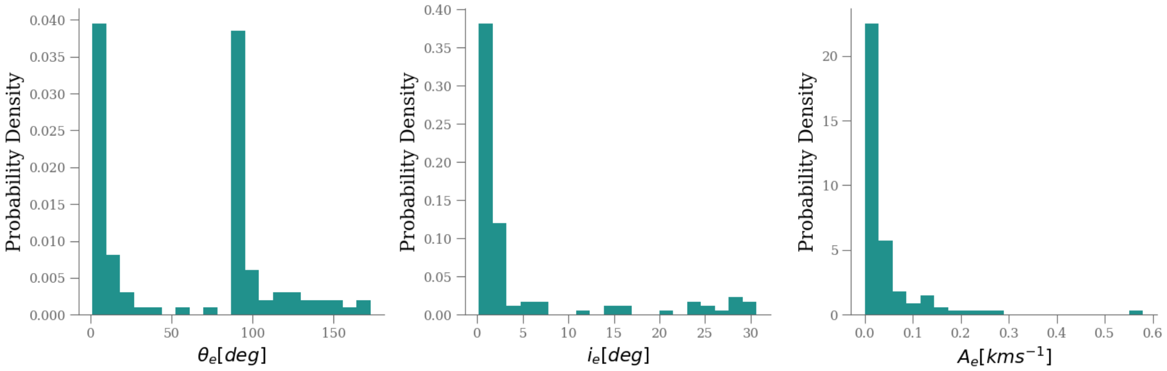


Figure 5.38: Open clusters: The density distribution of the estimated uncertainties for each rotational parameter. The range for the axis orientation parameters are both large relative to the expected range of the parameters, with $\theta$ ranging from 0-160 and inclination from 0-30. Interestingly the error for $\theta$ is double peaked at 0 and 90; this 90 degree offset is to be expected. The error on A is peaked at 0, which is to be expected given the low number of open clusters exhibiting rotation.

see if our results would get any closer to the literature values for these two clusters, but we found no significant improvement: for Praesepe we obtained $0.023277 \pm 0.002786\,\mathrm{km\,s^{-1}}$, and $-0.31726 \pm 0.809717\,\mathrm{km\,s^{-1}}$ for NGC6819.

To further investigate the general parameters of OCs and identify trends within the Galaxy based on their internal rotational velocities, additional research is required. This planned follow-up work involves constructing dispersion profiles and searching for and acquiring additional radial velocity information from past or ongoing ground-based surveys such as RAVE or GALAH, as well as future surveys like 4MOST and WEAVE. These efforts will enable us to refine our estimates of rotational velocities and deepen our understanding of the dynamics of OCs within the Galaxy.

Table 5.3: Estimated rotational parameters of open clusters

| Cluster | $n_{\text{pre}}$ | $n_{\text{post}}$ | $\theta$ | $\theta_e$ | $i$ | $i_e$ | $A$ | $A_e$ | $\theta_{\text{DE}}$ | $\theta_{\text{eDE}}$ | $i_{\text{DE}}$ | $i_{\text{eDE}}$ | $A_{\text{DE}}$ | $A_{\text{eDE}}$ | $A_{\text{snrMC}}$ | $A_{\text{snrDE}}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| NGC2158 | 113 | 113 | 146.06 | 21.47 | 85.39 | 25.99 | 1.21 | 0.92 | 145.36 | 5.49 | 85.84 | 2.39 | 1.23 | 0.17 | 2.63 | 14.17 |
| LP145 | 115 | 115 | 123.62 | 73.77 | 88.90 | 27.51 | -0.52 | 1.22 | 295.28 | 183.57 | 89.23 | 1.45 | 0.50 | 1.16 | 0.85 | 0.86 |
| NGC6416 | 111 | 111 | 190.95 | 9.50 | 89.02 | 23.59 | 0.42 | 1.00 | 190.97 | 2.88 | 89.13 | 0.94 | 0.42 | 0.04 | 0.83 | 22.20 |
| NGC1342 | 174 | 174 | 211.35 | 21.06 | 89.68 | 25.15 | -0.28 | 1.15 | 210.95 | 180.85 | 89.75 | 0.51 | -0.28 | 0.58 | 0.50 | 0.97 |
| NGC2506 | 101 | 101 | 135.90 | 60.83 | 88.25 | 34.05 | 0.22 | 1.04 | 315.25 | 3.56 | 82.48 | 6.30 | -0.24 | 0.06 | 0.42 | 8.64 |
| … | … | … | … | … | … | … | … | … | … | … | … | … | … | … | … | |

## 5.5 Conclusions

This Chapter has demonstrated the usefulness of applying a solid body rotation model to open clusters in order to gain insights into their rotational behavior. However, it is crucial to acknowledge the limitations of this approach, primarily arising from data constraints,

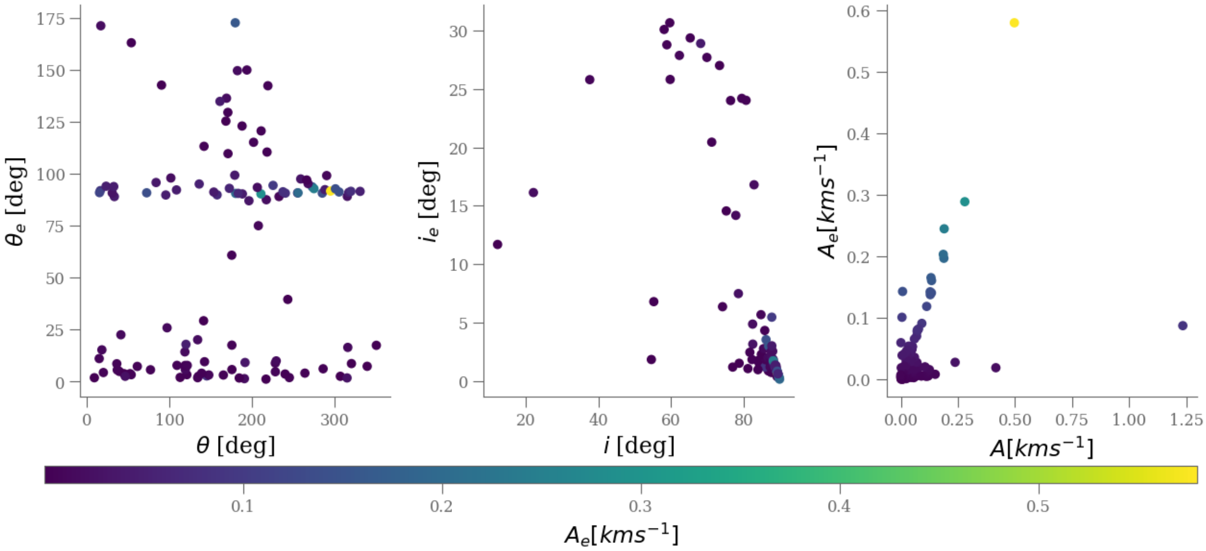


Figure 5.39: Open clusters: Patterns observed in the errors associated with $\theta$ and rotation. While both parameters exhibit a seemingly random spread of errors, it is worth noting that the errors for inclination display a distinctive trend. As the inclination approaches 90°, the errors asymptotically decrease.

particularly the lack of radial velocity measurements for many cluster members in the Gaia dataset. This limitation can be effectively addressed through the use of supplementary ground-based spectroscopy, which allows for the collection of additional radial velocity data, especially for clusters where rotation is inferred solely based on Gaia radial velocities. In additional an analysis of velocity dispersions may yield a more accurate picture of a cluster's profile than just considering a fixed value for each cluster, leading to more representative rotation parameter values.

Furthermore, it is recommended to explore alternative rotational models that may better capture the complexities of cluster rotation. By considering different models, we can gain a more comprehensive understanding of the rotational dynamics within OCs. Additionally, employing a variety of statistical estimation techniques can help improve the accuracy and reliability of the results obtained from the solid body rotation model.

By addressing these limitations and exploring alternative approaches, future work can refine and enhance our understanding of rotation in OCs. This will contribute to a more thorough comprehension of the formation and evolution processes of these stellar systems.

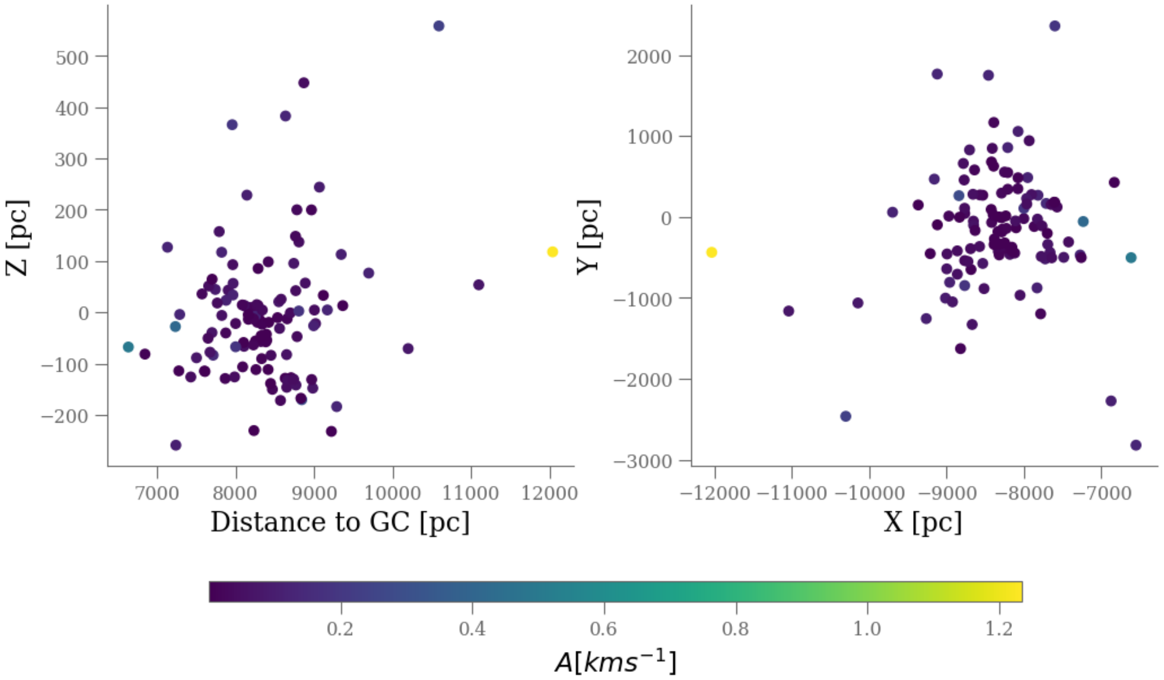


Figure 5.40: Open clusters: Looking at the spatial distribution of our OCs, there is no dominant trend with regard to the magnitude of rotational velocity and a cluster's position.

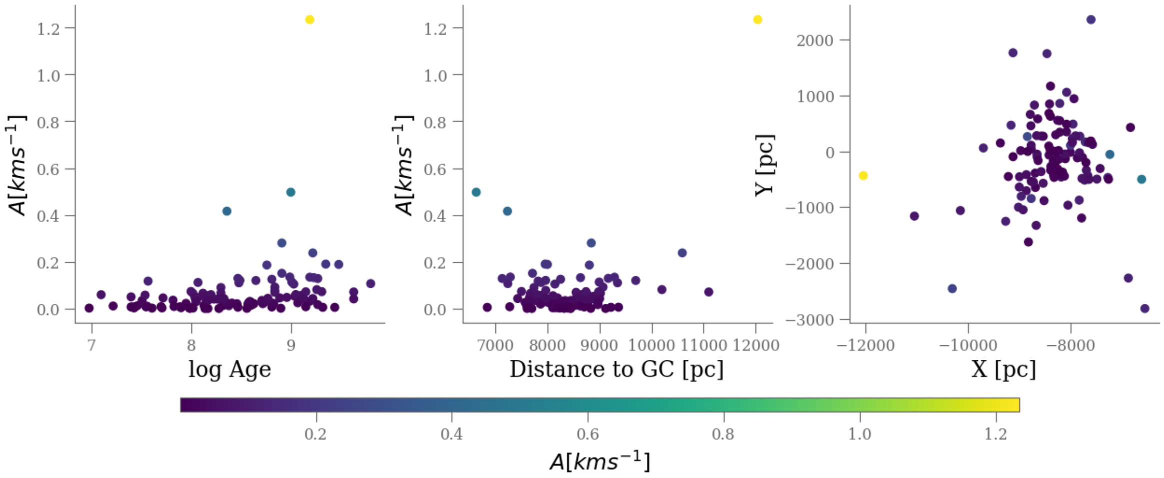


Figure 5.41: Open clusters: Considering the age of a cluster, the distance away from the GC and distribution in X-Y there is no clear relationship between these parameters and the magnitude of the rotational velocity.

# 6 Conclusion and Outlook

## 6.1 Conclusion

As the preceding work demonstrates, the development of astrostatistical methodology is of great importance to the advancement of future astronomy. By advancing and refining statistical techniques specific to astronomy, we greatly enhance our ability to extract meaningful insights from the vast amount of astronomical data at both our current and future disposal. Through leveraging advanced statistical techniques, we can analyse data for the presence of rare objects, accurately classify distinct astronomical sources, and effectively model complex astrophysical phenomena. Astrostatistical methods moreover enable the identification and quantification of uncertainties associated with various measurements, facilitating robust and reliable scientific conclusions.

This thesis has explored promising areas in astrostatistics to achieve more efficient extraction of astronomical results and their applications to future astronomical surveys. The overview provided in Chapter 2 delved into the history and future prospects of astrostatistics, as well as the significance of large astronomical surveys. A variety of surveys and the astronomical objects considered in this thesis were described.

In Chapter 3, a novel semi-supervised approach was presented for identifying different stellar types in large spectroscopic surveys, specifically when only a limited number of stellar types are known. This approach leveraged t-SNE, a dimensionality reduction technique that enables visualising object similarity in a 2D space. By overlaying unknown objects near known objects, the identification process is significantly improved, with close proximity indicating similarity. The application of this technique to over 600,000 high-resolution stellar spectra obtained from the GALAH survey successfully identified 54 potential rare EMP star candidates. EMP stars offer a valuable opportunity to investigate the early stages of chemical enrichment in the Milky Way galaxy. Among the 54 stars with estimated [Fe/H] below -3.0, there are 6 candidates with [Fe/H] below -3.5. Notably, our sample exhibits a higher proportion (approximately 20%) of main sequence EMP candidates compared to previous surveys targeting EMP stars. This Chapter highlighted the effectiveness of the statistical technique of clustering to identify rare objects with large spectroscopic datasets.

Shifting the focus to the Gaia survey, Chapter 4 introduced a statistical method for classifying extragalactic sources, including stars, galaxies, and quasars, based on their positions and photometry. We investigated the potential improvement in identifying extragalactic sources by combining Gaia photometry and astrometry with infrared data from CatWISE2020, compared to using Gaia data alone. The study involved a comprehensive analysis of different configurations of input features and prior functions, with the aim of developing a classification methodology that incorporated prior knowledge based on realistic class distributions in the Universe. To achieve this, we compared various classifiers, including Gaussian mixture models (GMMs), XGBoost, and CatBoost, in a supervised approach. The classification task involved categorising sources into three classes: stars, quasars, and galaxies. The labels for the quasar and galaxy classes are obtained from the Sloan Digital Sky Survey Data Release 16 (SDSS16), while the star labels come from Gaia EDR3. In our approach, we adjusted the posterior probabilities to account for the intrinsic distribution of extragalactic sources in the Universe using a prior function. We introduced two priors: a global prior that reflects the overall rarity of quasars and galaxies, and a mixed prior that incorporates the distribution of extragalactic sources based on Galactic latitude and magnitude. The best classification performance, in terms of completeness and purity of the extragalactic classes (galaxy and quasar), is achieved using the mixed prior for sources at high latitudes and within the magnitude range G = 18.5 to 19.5. We applied the identified best-performing classifier to three application datasets from GDR3 and observe that the global prior is more conservative in classifying sources as quasars or galaxies compared to the mixed prior. Specifically, when applied to the quasar and galaxy candidate tables from GDR3, the classifier using the global prior achieves purities of 55% for quasars and 93% for galaxies. The mixed prior yields purities of 59% for quasars and 91% for galaxies. When comparing these results to the performance on the GDR3 pure quasar and galaxy candidate samples, we achieve higher purities of 97% for quasars and 99.9% for galaxies using the global prior, and purities of 96% and 99%, respectively, using the mixed prior. By refining the GDR3 candidate tables through a cross-match with SDSS DR16 confirmed quasars and galaxies, the classifier attains purities of 99.8% for quasars and 99.9% for galaxies using the global prior, and 99.9% for both quasars and galaxies using the mixed prior. This Chapter highlighted the effectiveness of using tree-based statistical models instead of mixture models, the significance of applying adjusted priors to represent the realistic class distributions in the Universe and the incorporation of infrared data as ancillary inputs in the identification of extragalactic sources.

Finally, Chapter 5 investigated a model of 3D solid-body rotation within star clusters using Gaia data. Drawing inspiration from the work of Sollima et al. (2019), a method was implemented to identify and quantify three rotational parameters: the angular velocity of the rotation axis ($A$), the inclination of the rotation axis relative to the line of sight, and the position angle of the rotation axis ($\theta$). In this Chapter, simulations were conducted to assess the efficacy of two distinct sampling methods, namely MCMC and DE, for estimating the parameters from a likelihood function. Additionally, the simulations aimed to uncover potential degeneracies among the rotational parameters. Following the evaluation of the method on simulated data, its validity was tested on real data from extensively studied globular clusters (GCs), with our estimates compared against those of Sollima et al. (2019). Notably, despite differences in cluster membership, the results demonstrated the possibility of achieving similar outcomes. Finally, the methodology was applied to a compilation of open clusters (OCs) by Cantat-Gaudin, revealing modest indications of rotation in a few clusters, most notably

NGC2158. However, this list would greatly benefit from further ground-based spectroscopy to obtain additional radial velocities, thereby enhancing the precision of the obtained estimates.

## 6.2 Outlook

By exploring these areas of research, this thesis has demonstrated the potential for more efficient exploitation – that is, the extraction of scientific results – from large astronomical datasets. The methodologies and techniques presented here help lay the foundation for future studies and applications in the field of astrostatistics, paving the way for further discoveries and insights in astronomy.

There will be many forthcoming large surveys, such as the Euclid mission, that will greatly benefit from the development of new astrostatistics methodologies. The Euclid survey has the primary goal of investigating the nature of dark energy and dark matter by mapping the large-scale structure of the universe and studying the evolution of cosmic structures over time. Euclid will carry out a comprehensive survey of galaxies and clusters of galaxies using a combination of imaging and spectroscopic observations. It will cover a large area of the sky and observe billions of galaxies, providing precise measurements of their positions, shapes, and redshifts. With Euclid's high-quality data, astrostatistics can be used to understand the intricate relationships between different astrophysical parameters and to infer cosmological parameters with improved accuracy, and the development of robust statistical techniques will ensure the accurate and efficient extraction of science results. Moreover, astrostatistics will contribute to addressing challenges related to data processing, image analysis, and catalogue extraction in the Euclid survey, for example in the search of high-z quasars.

I am eagerly looking forward to actively contributing to the future of the astrostatistical methodology, specifically in developing and applying innovative methods that allow us to extract scientific insights from the forthcoming era of large-scale astronomical surveys.

# A
# Appendix

## A.1 Deriving metallicities for metal-poor candidates with GALAH spectra

This Section describes a set of simulation outputs that illustrate (1) the sensitivity of using only GALAH spectra to estimate metallicity at low metallicities ([Fe/H] $\sim -4.5$ is possible for cool giants) and (2) refine the line list we use for the low metallicity fits (the three bluer channels are preferred slightly over a few strong Fe lines).

The simulations work by adding realistic noise to synthetic stellar templates with known parameters, and attempting to recover [Fe/H] using the same templates. We fix [$\alpha$/H]$= 0.4$ and consider a limited range of Carbon enhancements ([C/H]$= 0.0, 0.5, 1.0$).

It is important to note that the following simulation results are only for fitting [Fe/H]. We assume that $\log g$ and $T_{\text{eff}}$ are already well-constrained (see Section 3.2.1 for how we do this with GALAH spectra) so we can limit the number of templates we fit to. The simulation results below did consider different carbon enhancements, in the sense that we explored whether carbon enhancement impacts the metallicity sensitivities. This means the current simulations were not meant to test our ability to constrain Carbon abundance in an individual spectrum.

As shown in Section 3.4, the current sample of candidate EMP stars have a median S/N of 35 and two rough sub-populations: cool giants ($T_{\text{eff}} = 5000$, $\log g \sim 2$) and hot main sequence stars ($T_{\text{eff}} = 6000$, $\log g \sim 4$).

To find the best regions of the GALAH spectra to fit for metallicity, we considered two different line lists as well as fitting to entire HERMES spectral channels. The first line list is from T. Nordlander and is highlighted shown in Figure A.1. The second is a list of 57 metal-sensitive (mostly iron) lines compiled from features found in synthetic spectra and observed stars around [Fe/H] $\sim -3$ from K. Venn (priv. comm.).

Figure A.2 shows the output from one run of the simulation on the strongest features. Parameters for the input spectra are given in the title and the number of simulated stars

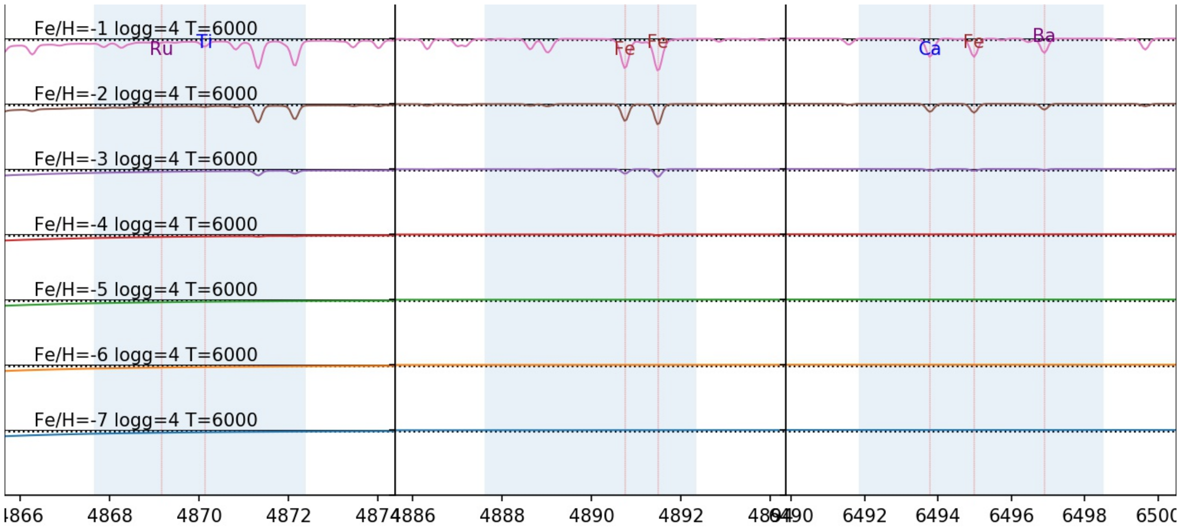


FIGURE A.1: A plot of synthetic spectra of a hot main sequence star for a range of metallicities. The blue shaded areas are the wavelength regions with metal-sensitive absorption features.

for each input [Fe/H] was set to N$_{sims}$ = 10000. The results show that for this spectral line that metallicity is well-constrained until [Fe/H] $\sim -4.5$. For lower metallicities the simulation indicates S/N=35 spectral data over this line cannot distinguish between [Fe/H] $\sim -5$ to $-7$ values.

Figure A.3 shows simulation runs where multiple spectral lines were simultaneously fit to illustrate improvement in metallicity constraints. While the second strongest set of lines improves the fitting to lower metallicities, the third set of lines does not influence the measured metallicity significantly for S/N=35 spectra.

Figure A.4 shows metallicity sensitivity simulations for an additional line lists (K. Venn, priv. comm.) with 57 features as well as full fits to the spectra over the 3 HERMES channels. These are compared to the best combination of 2 lines from Figure A.3. Increased wavelength coverage appears to yield better metallicity constraints in the simulations, but only marginally so.

We show simulation results for hot main sequence stars in Figure A.5. At these temperatures, the metallicity sensitivity decreases, so that only metallicities of [Fe/H] $\sim -3.5$ or higher are measurable with S/N=35 GALAH spectra.

Finally, we show in Figure A.6 how the metallicity sensitivity is expected to improve for higher S/N (S/N $\sim$ 150) data: [Fe/H] $\sim -5.0$ and $\sim -4$ are expected for cool giants and hot main sequence stars, respectively.

To conclude, we find:

- Synthetic cool giant spectra with typical GALAH S/N= 35 over the GALAH spectral range are good ($\sim$ 9%) at recovering metallicities as low as [Fe/H] $\sim -5.5$.

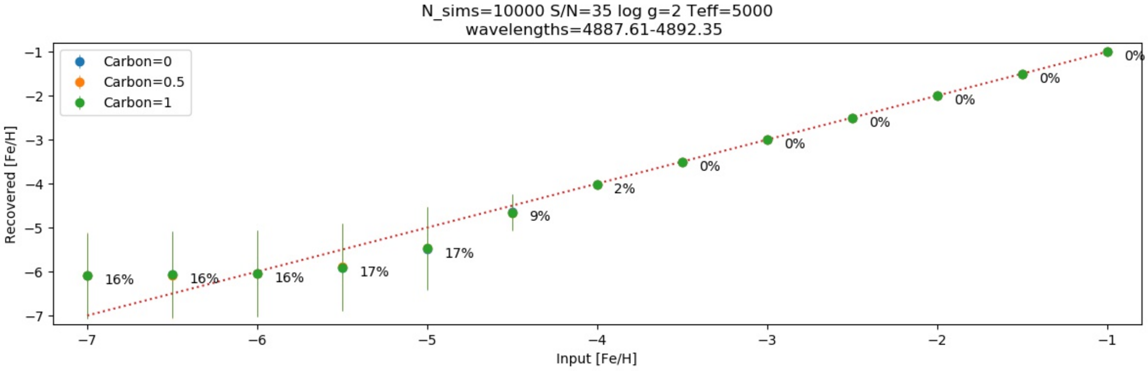
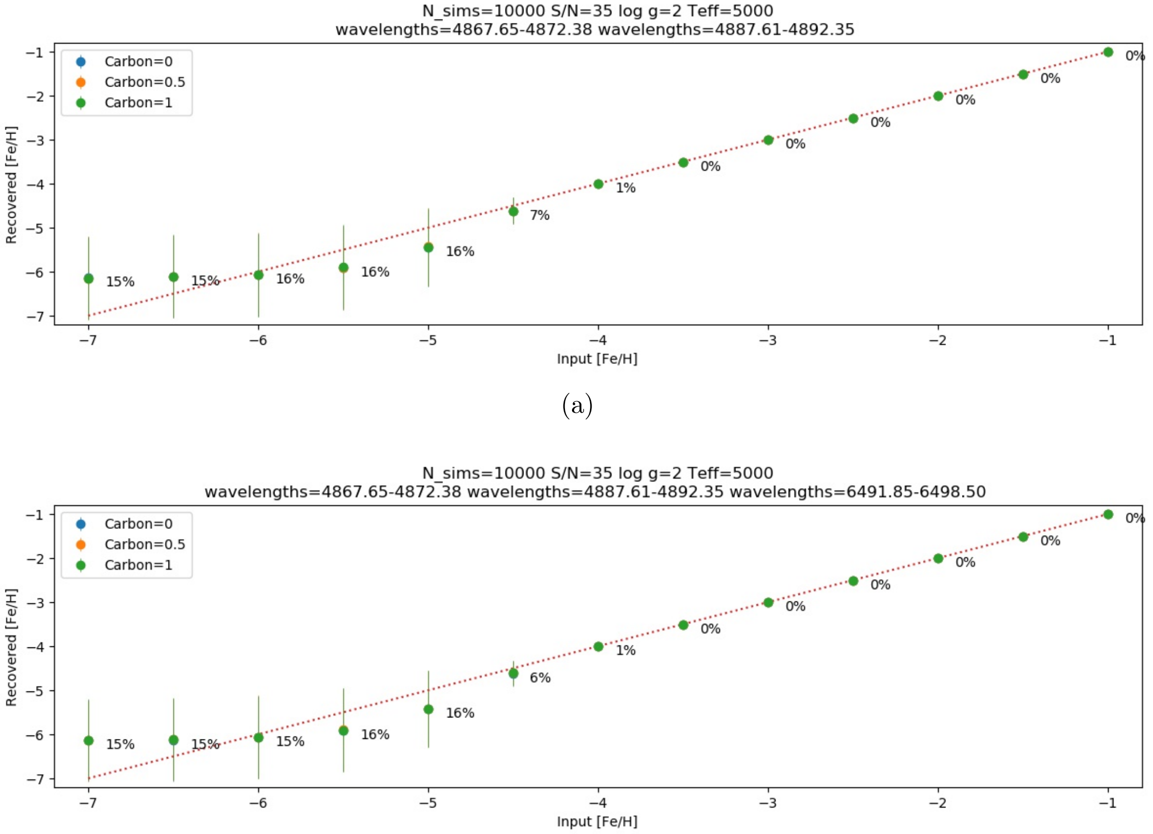


Figure A.2: Simulation results for recovering input metallicities of synthetic cool giant stellar spectra, with noise typical of the current sample. The points represent the recovered mean [Fe/H] with error bars reflecting the standard deviation in individual recovered [Fe/H] values. Percentages are the fractional uncertainty on the recovered metallicity. The data suggest we can constrain metallicity to within $\sim 8\%$ down to [Fe/H] $\sim -4.5$ for cool giants and data with S/N= 35. For lower input metallicities, the simulation results indicate the spectra have essentially no constraint on metallicities [Fe/H] $< -4.5$ at S/N= 35. The carbon abundances of the input spectra do not appear to impact the metallicity sensitivity as can be seen by the carbon abundance of 1 (green) overlaying both the carbon abundance of 0 and 0.5 ( blue and orange respectively).

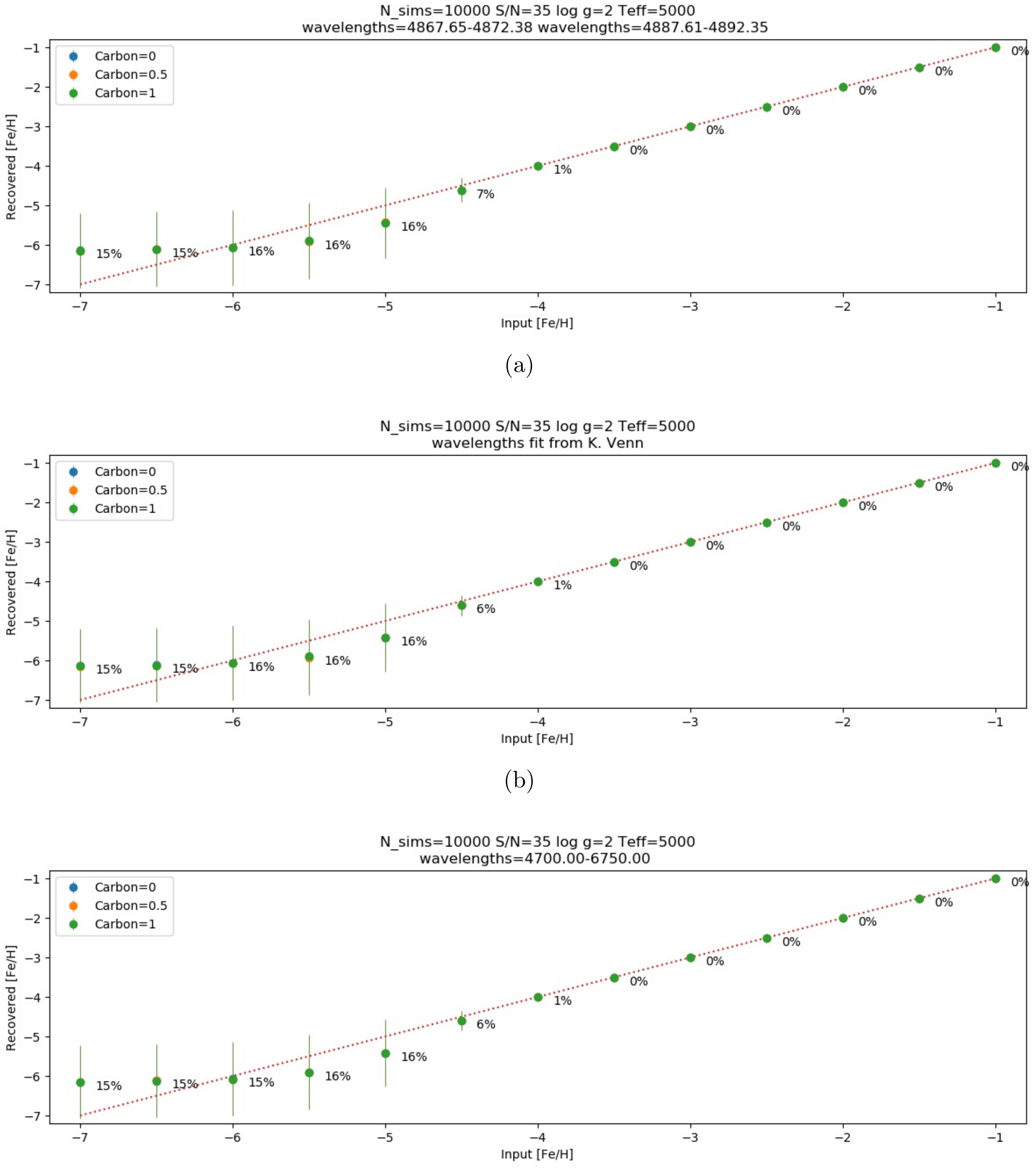
- At a certain metallicity the GALAH spectra are no longer sensitive to lower metallicities for S/N=35 spectra.
- With better S/N($\sim$ 100, or even $\sim$ 150), metallicities as low as [Fe/H] $\sim -4.5$ can be recovered to $\sim 9\%$.
- The metallicity sensitivity and fitting does not appear to be impacted by the level of carbon enhancement of the star within the wavelength coverage of the GALAH spectral channels.



(a)

(b)

FIGURE A.3: Same format as Figure A.2, now with panels showing two different combinations of spectral lines shown in Figure A.1, as indicated in the panel's subtitle. While the joint constraint of the spectral regions with the strongest features yields a better [Fe/H] constraint compared to a single line (*cf*. Figure A.2), the third spectral region does not improve the fits for this stellar type and assumed S/N, likely because its features are weaker.

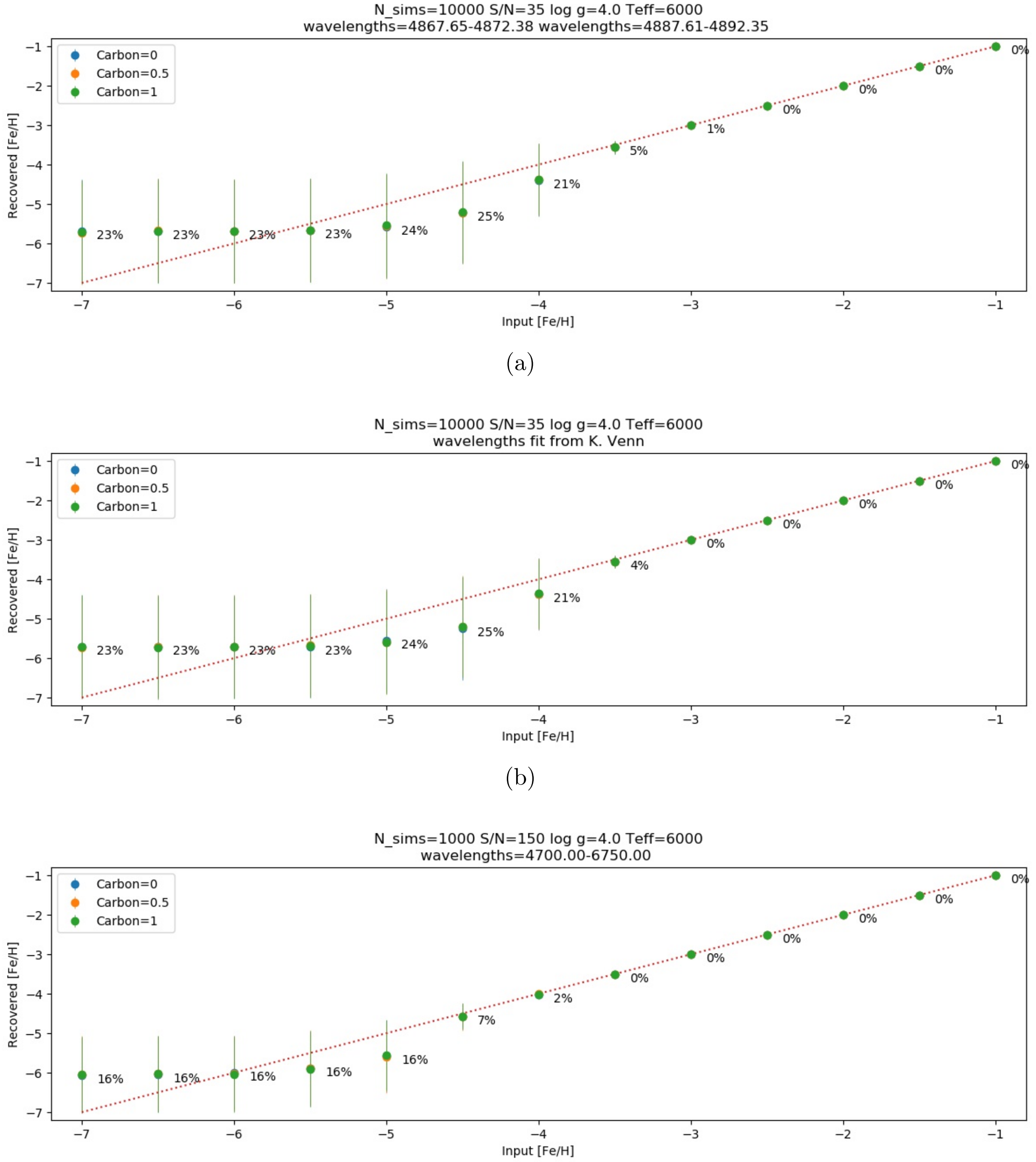


(a)

(b)

(c)

Figure A.4: Same format as Figure A.2, now with panels showing three different combinations of spectral lines: top is the Nordlander best features (Figure A.3), middle is fit to 57 metal-sensitive features (K. Venn, priv. comm) and the bottom plot shows fitting results to the first 3 HERMES Channels. The fit to 3 spectral channels yields a marginally better constraint at [Fe/H] $\sim -4.5$ for cool giants with S/N=35 than fits for [Fe/H] to the other spectra regions.

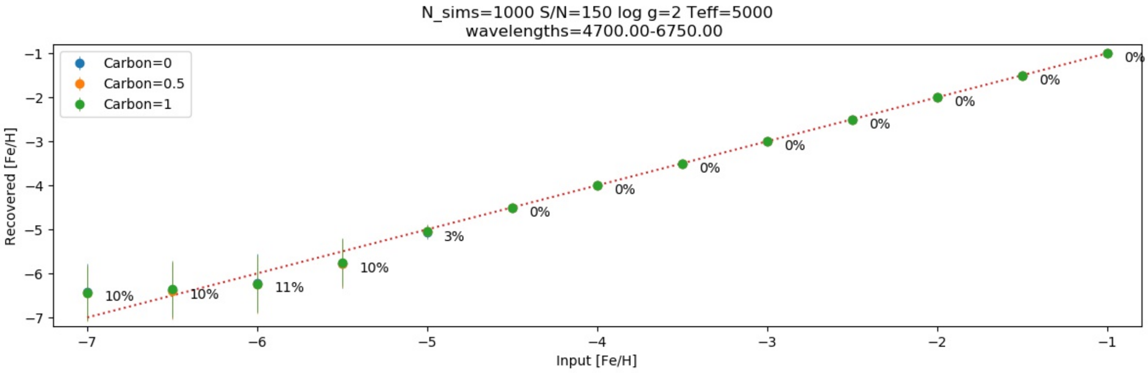


(a)

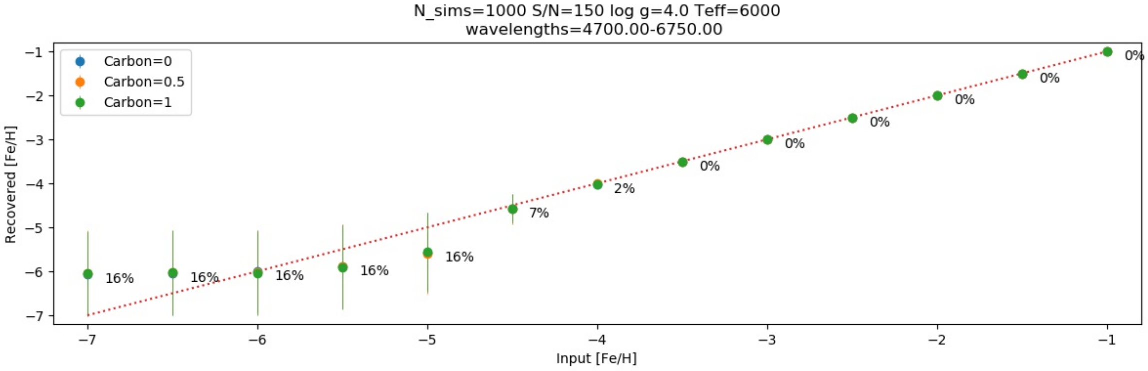
(b)

(c)

Figure A.5: Same as Figure A.4 now for hot ($T_{\mathrm{eff}} = 6000$) main sequence ($\log g = 4$) stellar templates. Overall the metallicity sensitivity decreases such that we may only expect to make measurements to [Fe/H] $\sim -3.5$.



(a)



(b)

Figure A.6: A few spectra in the GALAH sample reach S/N=150. These simulations show how much better the low metallicity constraints can be with higher S/N data and fits over the entire first 3 GALAH Channels.

## A.2 Prior counts, and Galactic latitude and magnitude priors

The following section shows the counts of sources in SDSS DR16 as a function of Galactic latitude and magnitude as well as the distribution of the prior.

| cutSinb | cutgMag | starN | qsoN | galN |
|---|---|---:|---:|---:|
| (0,0.4] | (17.5,18.5] | 67338 | 184 | 81 |
| (0,0.4] | (18.5,19.5] | 107892 | 1112 | 277 |
| (0,0.4] | (19.5,20.5] | 159394 | 3274 | 1320 |
| (0,0.4] | (20.5, Inf] | 130105 | 2526 | 3839 |
| (0,0.4] | (-Inf,17.5] | 81827 | 22 | 250 |
| (0.4,0.6] | (17.5,18.5] | 5644 | 2196 | 133 |
| (0.4,0.6] | (18.5,19.5] | 8077 | 11332 | 2060 |
| (0.4,0.6] | (19.5,20.5] | 11479 | 31767 | 17322 |
| (0.4,0.6] | (20.5, Inf] | 10052 | 27529 | 53069 |
| (0.4,0.6] | (-Inf,17.5] | 9921 | 272 | 21 |
| (0.6,0.8] | (17.5,18.5] | 2910 | 4328 | 390 |
| (0.6,0.8] | (18.5,19.5] | 4147 | 21412 | 5063 |
| (0.6,0.8] | (19.5,20.5] | 6037 | 56641 | 35817 |
| (0.6,0.8] | (20.5, Inf] | 5946 | 49410 | 113906 |
| (0.6,0.8] | (-Inf,17.5] | 5508 | 631 | 40 |
| (0.8,1] | (17.5,18.5] | 1928 | 6022 | 571 |
| (0.8,1] | (18.5,19.5] | 2625 | 26410 | 7771 |
| (0.8,1] | (19.5,20.5] | 4227 | 69930 | 52335 |
| (0.8,1] | (20.5, Inf] | 5214 | 65828 | 145758 |
| (0.8,1] | (-Inf,17.5] | 3542 | 812 | 39 |

Table A.1: Prior Table Counts

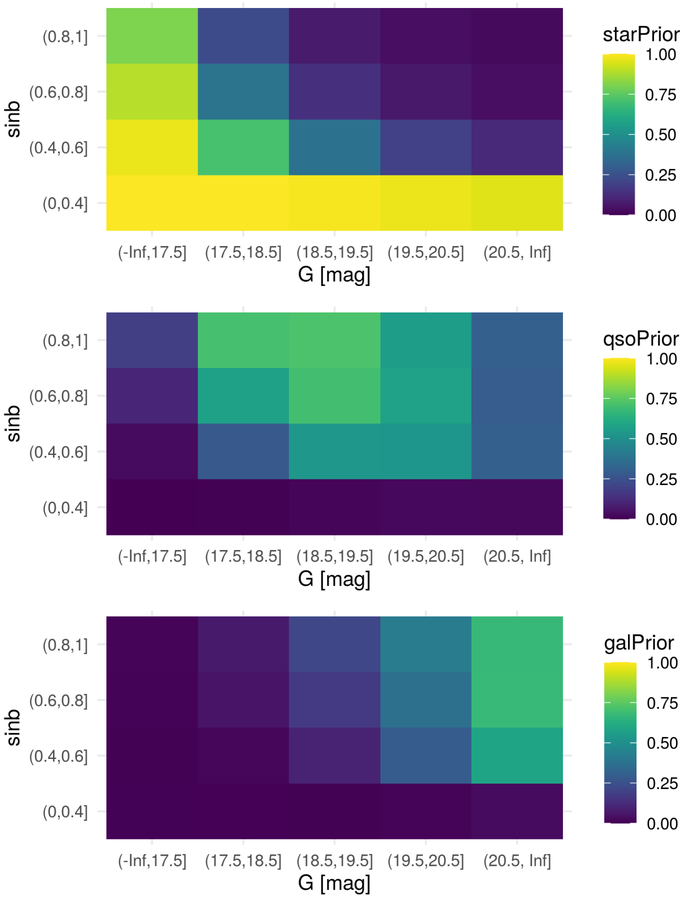


Figure A.7: Heat map of the joint latitude and magnitude prior for each class. The top panel refers to the star class, middle panel to the quasar class, and the lower panel to the galaxy class. A higher density of stars is noticeable at lower latitudes, while more quasars and galaxies clusters are seen at higher magnitudes.

## A.3 Simulated clusters

Table A.2: This table presents the various combinations considered and the corresponding estimated values of the cluster parameters.

| $\theta_C$ | $i_C$ | $A_C$ | $\theta_M$ | $\theta_{eM}$ | $i_M$ | $i_{eM}$ | $A_M$ | $A_{eM}$ | $\theta_D$ | $\theta_{eD}$ | $i_D$ | $i_{eD}$ | $A_D$ | $A_{eD}$ | $A_{snrM}$ | $A_{snrD}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.66 | 3.31 | -0.04 | 35.95 | 44.14 | 72.92 | 38.44 | -0.12 | 0.20 | 170.18 | 261.85 | 54.95 | 57.14 | -0.07 | 0.14 | 1.21 | 1.00 |
| 6.23 | 49.44 | 4.30 | 15.72 | 14.05 | 65.28 | 21.43 | 4.60 | 4.13 | 11.09 | 14.80 | 58.49 | 21.01 | 3.67 | 1.78 | 2.23 | 4.13 |
| 6.54 | 38.73 | -4.00 | 15.91 | 9.36 | 41.39 | 16.08 | -3.47 | 0.87 | 13.30 | 37.84 | 43.56 | 23.45 | -3.58 | 1.46 | 7.95 | 4.92 |
| 7.49 | 46.31 | 4.84 | 13.23 | 10.40 | 65.27 | 4.60 | 5.97 | 1.06 | 18.50 | 346.04 | 50.03 | 38.40 | 3.90 | 3.31 | 11.25 | 2.36 |
| 8.99 | 37.62 | 3.96 | 14.72 | 13.63 | 56.54 | 41.45 | 4.54 | 3.67 | 17.87 | 46.78 | 44.36 | 29.02 | 3.50 | 1.74 | 2.47 | 4.03 |
| 16.80 | 7.64 | 1.12 | 17.35 | 11.21 | 19.02 | 19.31 | 1.15 | 0.15 | 65.20 | 322.11 | 11.44 | 53.36 | 1.11 | 0.86 | 15.34 | 2.57 |
| 18.14 | 15.87 | 1.53 | 19.62 | 17.32 | 24.77 | 29.01 | 1.52 | 0.56 | 30.63 | 67.18 | 18.80 | 47.00 | 1.46 | 1.11 | 5.39 | 2.61 |
| 24.72 | 87.78 | -4.86 | 24.72 | 0.00 | 97.78 | 0.00 | 5.14 | 0.00 | 24.91 | 8.53 | 86.30 | 0.74 | -4.77 | 0.99 | * | 9.60 |
| 28.40 | 40.22 | 2.95 | 45.46 | 14.78 | 34.66 | 50.28 | 2.08 | 4.51 | 32.53 | 33.11 | 46.97 | 34.58 | 2.51 | 2.27 | 0.92 | 2.21 |
| 34.30 | 64.72 | -2.38 | 32.00 | 2.37 | 71.24 | 10.17 | -2.09 | 1.35 | 36.37 | 18.01 | 71.20 | 22.63 | -2.13 | 1.58 | 3.11 | 2.71 |
| 36.55 | 22.63 | 2.13 | 39.71 | 17.14 | 52.12 | 45.81 | 2.83 | 2.51 | 40.76 | 31.53 | 26.41 | 46.24 | 1.94 | 1.80 | 2.26 | 2.16 |
| 41.76 | 63.61 | -3.13 | 46.87 | 12.54 | 67.77 | 4.49 | -2.41 | 0.46 | 42.91 | 18.12 | 69.27 | 26.68 | -2.54 | 1.72 | 10.37 | 2.95 |
| 50.31 | 49.29 | 1.22 | 50.01 | 10.34 | 70.89 | 4.87 | 1.71 | 0.44 | 53.17 | 33.44 | 56.12 | 43.46 | 1.01 | 1.68 | 7.71 | 1.20 |
| 52.61 | 85.62 | 0.85 | 52.61 | 0.00 | 95.62 | 0.00 | 10.85 | 0.00 | 51.57 | 9.56 | 88.65 | 0.40 | 0.81 | 0.19 | * | 8.59 |
| 53.03 | 41.23 | -3.29 | 48.11 | 8.91 | 43.41 | 23.77 | -2.55 | 1.06 | 53.79 | 18.33 | 46.43 | 22.62 | -2.72 | 1.04 | 4.81 | 5.23 |
| 53.47 | 22.88 | -4.28 | 66.66 | 19.08 | 33.92 | 26.31 | -4.21 | 1.37 | 52.95 | 31.45 | 24.83 | 19.35 | -3.90 | 0.68 | 6.14 | 11.52 |
| 58.54 | 6.23 | -4.19 | 33.62 | 19.10 | 15.34 | 16.52 | -4.22 | 0.32 | 73.61 | 132.90 | 7.46 | 16.79 | -4.13 | 0.26 | 26.55 | 31.49 |
| 59.98 | 30.70 | 3.65 | 58.02 | 18.69 | 51.42 | 34.22 | 4.15 | 2.91 | 61.01 | 24.25 | 36.24 | 33.87 | 3.20 | 1.78 | 2.85 | 3.59 |
| 65.06 | 0.82 | 1.05 | 37.87 | 38.38 | 16.30 | 40.81 | 1.09 | 0.59 | 169.49 | 236.07 | 5.22 | 31.58 | 1.05 | 0.20 | 3.70 | 10.54 |
| 71.63 | 61.53 | 1.58 | 73.55 | 13.89 | 73.74 | 19.73 | 1.85 | 1.42 | 71.70 | 13.58 | 69.02 | 13.90 | 1.45 | 0.76 | 2.60 | 3.82 |
| 75.60 | 70.77 | 1.85 | 70.14 | 4.16 | 77.99 | 5.91 | 2.04 | 0.88 | 75.37 | 14.23 | 75.64 | 19.64 | 1.72 | 1.29 | 4.64 | 2.67 |
| 77.02 | 39.42 | -1.46 | 72.69 | 14.31 | 43.71 | 20.41 | -1.30 | 0.39 | 77.32 | 22.93 | 42.63 | 25.08 | -1.27 | 0.52 | 6.69 | 4.89 |
| 83.89 | 8.16 | -2.04 | 74.76 | 22.47 | 18.99 | 20.01 | -2.08 | 0.31 | 88.26 | 104.89 | 10.25 | 28.49 | -1.99 | 0.25 | 13.38 | 16.08 |
| 86.68 | 42.11 | 3.63 | 88.24 | 12.25 | 45.19 | 43.61 | 3.16 | 4.52 | 89.26 | 23.48 | 48.12 | 31.95 | 3.34 | 2.54 | 1.40 | 2.63 |
| 87.65 | 12.10 | -3.30 | 100.11 | 16.13 | 31.87 | 8.40 | -3.62 | 0.32 | 88.68 | 54.99 | 13.73 | 21.59 | -3.16 | 0.40 | 22.90 | 15.68 |
| 90.05 | 3.67 | 3.19 | 105.82 | 53.82 | 7.87 | 39.47 | 3.19 | 1.21 | 119.28 | 213.91 | 4.55 | 20.84 | 3.17 | 0.25 | 5.27 | 25.40 |
| 90.53 | 51.18 | -3.01 | 95.93 | 6.11 | 67.26 | 4.16 | -3.79 | 0.65 | 92.25 | 19.39 | 56.72 | 32.95 | -2.67 | 1.62 | 11.71 | 3.28 |
| 109.23 | 77.41 | -4.32 | 110.19 | 2.96 | 82.71 | 1.49 | -4.63 | 0.90 | 109.27 | 10.43 | 81.64 | 3.84 | -4.08 | 1.90 | 10.31 | 4.30 |
| 110.57 | 62.67 | -0.58 | 103.10 | 16.35 | 72.87 | 16.91 | -0.51 | 0.51 | 109.51 | 27.24 | 70.72 | 29.63 | -0.49 | 0.44 | 2.01 | 2.22 |
| 115.41 | 60.56 | -1.57 | 105.54 | 13.24 | 65.17 | 19.27 | -1.31 | 1.09 | 115.71 | 17.90 | 65.54 | 21.25 | -1.34 | 0.78 | 2.39 | 3.43 |
| 115.73 | 37.50 | -4.78 | 105.44 | 9.33 | 48.06 | 22.61 | -4.26 | 1.59 | 117.19 | 19.38 | 42.74 | 19.93 | -3.95 | 1.13 | 5.35 | 6.98 |
| 116.08 | 33.03 | -1.86 | 97.32 | 35.99 | 24.90 | 22.26 | -1.42 | 0.28 | 117.79 | 26.65 | 36.41 | 27.86 | -1.60 | 0.48 | 10.13 | 6.67 |
| 122.02 | 38.72 | -3.91 | 122.11 | 17.37 | 40.05 | 14.04 | -3.12 | 0.65 | 122.84 | 19.86 | 44.54 | 23.67 | -3.35 | 1.17 | 9.57 | 5.75 |
| 124.28 | 16.70 | 1.61 | 125.41 | 30.14 | 21.49 | 20.13 | 1.54 | 0.26 | 126.01 | 40.64 | 18.73 | 42.07 | 1.51 | 0.87 | 11.95 | 3.47 |
| 131.80 | 50.26 | 2.79 | 117.37 | 14.40 | 45.79 | 32.31 | 1.87 | 3.18 | 132.97 | 15.62 | 57.76 | 24.85 | 2.45 | 1.52 | 1.18 | 3.23 |
| 135.28 | 72.83 | -4.27 | 131.56 | 9.70 | 79.78 | 2.57 | -4.64 | 1.26 | 134.09 | 11.11 | 78.09 | 9.41 | -4.00 | 2.52 | 7.37 | 3.17 |
| 135.73 | 41.46 | -3.35 | 141.54 | 13.91 | 51.80 | 26.47 | -3.19 | 1.71 | 136.12 | 23.87 | 46.91 | 29.39 | -2.96 | 2.07 | 3.74 | 2.87 |
| 137.76 | 72.40 | 4.06 | 135.09 | 2.48 | 78.85 | 3.65 | 4.30 | 1.50 | 136.23 | 12.70 | 77.14 | 5.38 | 3.76 | 1.41 | 5.72 | 5.35 |
| 137.93 | 87.80 | -0.60 | 137.93 | 0.00 | 97.80 | 0.00 | 9.40 | 0.00 | 136.42 | 8.75 | 89.04 | 0.30 | -0.60 | 0.13 | * | 9.12 |
| 143.56 | 69.68 | -2.88 | 146.39 | 6.84 | 77.03 | 3.47 | -3.38 | 0.86 | 143.38 | 15.90 | 73.93 | 11.87 | -2.74 | 1.70 | 7.84 | 3.23 |

Continued on next page

Table A.2 – continued from previous page

| $\theta_C$ | $i_C$ | $A_C$ | $\theta_M$ | $\theta_{eM}$ | $i_M$ | $i_{eM}$ | $A_M$ | $A_{eM}$ | $\theta_D$ | $\theta_{eD}$ | $i_D$ | $i_{eD}$ | $A_D$ | $A_{eD}$ | $A_{snrM}$ | $A_{snrD}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 146.00 | 32.94 | 3.95 | 146.32 | 10.25 | 58.82 | 39.50 | 5.27 | 3.01 | 144.53 | 22.49 | 39.16 | 28.85 | 3.52 | 1.97 | 3.50 | 3.57 |
| 148.04 | 77.21 | 1.09 | 145.18 | 7.76 | 82.29 | 1.25 | 1.41 | 0.23 | 146.95 | 11.82 | 79.60 | 9.00 | 1.06 | 0.61 | 12.23 | 3.46 |
| 152.79 | 25.80 | -0.26 | 153.66 | 39.14 | 42.09 | 42.40 | -0.26 | 0.24 | 156.10 | 72.05 | 30.90 | 62.76 | -0.23 | 0.47 | 2.12 | 0.99 |
| 161.50 | 28.17 | 2.74 | 161.17 | 18.16 | 53.80 | 52.76 | 3.42 | 3.43 | 159.41 | 28.68 | 33.48 | 29.87 | 2.42 | 1.17 | 1.99 | 4.13 |
| 162.14 | 68.35 | 3.67 | 163.14 | 5.43 | 77.69 | 1.80 | 4.77 | 0.70 | 162.46 | 11.23 | 72.93 | 6.23 | 3.48 | 1.16 | 13.71 | 6.01 |
| 162.30 | 39.86 | -3.75 | 160.25 | 1.64 | 49.12 | 11.04 | -3.60 | 0.87 | 161.76 | 19.37 | 45.30 | 14.68 | -3.33 | 0.94 | 8.24 | 7.08 |
| 166.25 | 47.31 | 4.17 | 169.06 | 18.14 | 62.71 | 31.00 | 4.63 | 3.94 | 167.57 | 16.67 | 54.90 | 21.15 | 3.70 | 1.73 | 2.35 | 4.29 |
| 166.91 | 69.26 | -0.87 | 171.36 | 9.37 | 75.31 | 6.23 | -0.85 | 0.31 | 168.32 | 15.68 | 74.54 | 15.71 | -0.80 | 0.55 | 5.39 | 2.92 |
| 170.58 | 54.26 | -1.53 | 169.83 | 8.90 | 66.47 | 37.73 | -1.44 | 1.46 | 170.06 | 21.56 | 60.82 | 27.95 | -1.27 | 0.79 | 1.97 | 3.20 |
| 171.73 | 39.45 | -4.20 | 180.92 | 14.65 | 35.21 | 27.60 | -3.08 | 1.00 | 174.11 | 17.76 | 43.32 | 12.32 | -3.62 | 0.63 | 6.18 | 11.45 |
| 172.15 | 19.65 | 2.33 | 167.88 | 27.16 | 44.31 | 49.65 | 2.78 | 1.99 | 171.96 | 36.54 | 22.48 | 40.29 | 2.15 | 1.18 | 2.79 | 3.64 |
| 173.73 | 6.78 | 2.18 | 165.04 | 36.68 | 11.81 | 17.58 | 2.17 | 0.18 | 177.05 | 87.33 | 8.39 | 39.72 | 2.15 | 0.77 | 24.16 | 5.57 |
| 174.92 | 70.09 | -0.30 | 170.70 | 21.94 | 75.16 | 10.61 | -0.27 | 0.18 | 175.42 | 23.62 | 73.56 | 16.76 | -0.26 | 0.20 | 3.08 | 2.53 |
| 177.40 | 24.92 | -1.48 | 168.52 | 17.82 | 49.82 | 42.00 | -1.55 | 1.48 | 177.27 | 32.59 | 29.09 | 38.25 | -1.33 | 0.79 | 2.09 | 3.35 |
| 180.95 | 48.97 | -0.60 | 192.00 | 21.98 | 47.46 | 40.48 | -0.41 | 0.71 | 181.82 | 24.49 | 56.01 | 33.81 | -0.50 | 0.36 | 1.15 | 2.78 |
| 189.87 | 12.46 | -3.20 | 174.40 | 33.82 | 27.11 | 19.59 | -3.32 | 0.79 | 185.42 | 49.63 | 11.75 | 17.21 | -3.02 | 0.26 | 8.39 | 23.12 |
| 191.95 | 60.03 | 0.36 | 192.13 | 15.72 | 72.67 | 21.80 | 0.41 | 0.41 | 191.23 | 25.18 | 64.74 | 28.54 | 0.28 | 0.29 | 2.00 | 1.97 |
| 192.09 | 73.06 | 1.44 | 189.75 | 10.49 | 79.95 | 1.77 | 1.83 | 0.33 | 192.38 | 12.25 | 75.86 | 9.29 | 1.31 | 0.69 | 11.05 | 3.79 |
| 195.55 | 64.48 | 4.64 | 209.46 | 15.43 | 73.96 | 11.55 | 5.15 | 2.55 | 194.60 | 15.20 | 70.30 | 16.64 | 4.19 | 2.63 | 4.04 | 3.19 |
| 199.28 | 1.10 | -0.62 | 196.13 | 110.32 | 13.22 | 22.46 | -0.63 | 0.08 | 186.55 | 213.97 | 7.15 | 60.09 | -0.62 | 0.47 | 16.04 | 2.62 |
| 209.63 | 22.29 | 0.58 | 209.47 | 31.55 | 32.67 | 48.43 | 0.55 | 0.80 | 211.38 | 42.28 | 26.76 | 47.37 | 0.52 | 0.46 | 1.39 | 2.27 |
| 213.68 | 72.06 | 1.36 | 210.63 | 7.81 | 80.42 | 4.07 | 1.55 | 0.62 | 213.43 | 12.82 | 78.24 | 15.00 | 1.25 | 0.90 | 5.01 | 2.78 |
| 217.23 | 82.52 | -2.57 | 217.23 | 0.00 | 92.52 | 0.00 | 7.43 | 0.00 | 215.61 | 10.59 | 85.51 | 1.14 | -2.45 | 0.60 | * | 8.18 |
| 221.85 | 0.10 | 3.62 | 219.17 | 12.21 | 5.87 | 10.19 | 3.64 | 0.08 | 184.00 | 193.66 | 4.23 | 9.69 | 3.63 | 0.06 | 90.26 | 117.17 |
| 224.83 | 21.39 | -4.85 | 215.06 | 19.96 | 32.75 | 19.69 | -4.75 | 1.09 | 224.83 | 30.40 | 25.16 | 21.16 | -4.40 | 0.89 | 8.68 | 9.92 |
| 225.31 | 47.24 | 0.67 | 221.67 | 15.56 | 60.89 | 20.47 | 0.68 | 0.70 | 224.08 | 27.47 | 53.74 | 36.72 | 0.55 | 0.70 | 1.94 | 1.58 |
| 231.06 | 72.95 | 2.38 | 226.46 | 7.12 | 80.63 | 2.64 | 2.76 | 0.77 | 231.13 | 13.19 | 78.60 | 14.07 | 2.23 | 1.66 | 7.22 | 2.68 |
| 231.77 | 33.47 | 3.02 | 222.95 | 18.32 | 43.09 | 28.02 | 2.80 | 1.67 | 228.66 | 26.38 | 40.14 | 39.53 | 2.68 | 2.68 | 3.36 | 2.00 |
| 232.95 | 20.12 | -1.85 | 227.01 | 13.28 | 36.24 | 25.99 | -1.92 | 0.62 | 230.71 | 28.45 | 21.85 | 19.92 | -1.67 | 0.28 | 6.24 | 12.02 |
| 236.33 | 25.67 | -0.88 | 232.50 | 11.61 | 29.70 | 4.55 | -0.80 | 0.04 | 233.20 | 37.75 | 28.67 | 33.69 | -0.79 | 0.29 | 39.17 | 5.51 |
| 237.15 | 46.56 | -3.05 | 238.19 | 13.20 | 42.56 | 31.88 | -2.13 | 1.48 | 237.79 | 17.52 | 53.73 | 19.46 | -2.67 | 1.14 | 2.87 | 4.70 |
| 242.74 | 58.99 | 1.26 | 241.41 | 18.38 | 73.17 | 20.10 | 1.52 | 0.94 | 242.49 | 19.87 | 67.06 | 32.08 | 1.13 | 1.20 | 3.23 | 1.89 |
| 247.32 | 74.89 | 1.11 | 252.86 | 10.55 | 79.66 | 2.94 | 1.18 | 0.39 | 247.45 | 13.68 | 77.94 | 9.74 | 1.01 | 0.66 | 6.07 | 3.07 |
| 252.90 | 13.17 | 3.45 | 239.64 | 12.78 | 40.42 | 34.44 | 4.23 | 2.88 | 251.89 | 47.60 | 14.68 | 39.69 | 3.33 | 1.44 | 2.94 | 4.63 |
| 254.68 | 86.76 | -3.03 | 254.68 | 0.00 | 96.76 | 0.00 | 6.97 | 0.00 | 253.48 | 9.39 | 88.21 | 0.26 | -2.98 | 0.45 | * | 13.19 |
| 257.91 | 33.78 | 0.78 | 275.31 | 34.40 | 29.48 | 53.85 | 0.64 | 1.09 | 255.24 | 39.43 | 38.57 | 53.46 | 0.71 | 1.25 | 1.17 | 1.13 |
| 266.38 | 52.91 | -4.16 | 275.56 | 27.34 | 48.30 | 20.98 | -2.50 | 0.87 | 266.29 | 14.86 | 62.04 | 17.03 | -3.62 | 1.73 | 5.76 | 4.18 |
| 270.63 | 47.31 | 2.93 | 271.87 | 19.58 | 65.18 | 27.35 | 3.92 | 2.75 | 270.42 | 18.73 | 52.04 | 17.04 | 2.67 | 1.04 | 2.85 | 5.16 |
| 272.66 | 31.69 | 1.32 | 273.90 | 9.76 | 53.18 | 41.93 | 1.56 | 1.59 | 271.66 | 29.32 | 36.57 | 40.37 | 1.15 | 1.00 | 1.97 | 2.30 |
| 280.26 | 26.51 | 4.80 | 277.34 | 29.78 | 50.18 | 47.20 | 5.73 | 4.54 | 280.85 | 25.49 | 31.60 | 33.06 | 4.33 | 2.06 | 2.53 | 4.19 |
| 284.48 | 84.40 | 2.59 | 284.48 | 0.00 | 94.40 | 0.00 | 12.59 | 0.00 | 284.60 | 10.46 | 85.68 | 0.94 | 2.54 | 0.55 | * | 9.29 |
| 296.28 | 17.47 | -2.18 | 287.70 | 24.65 | 24.27 | 13.26 | -2.09 | 0.22 | 297.87 | 36.69 | 19.30 | 20.06 | -2.03 | 0.31 | 19.38 | 13.06 |
| 299.24 | 54.75 | -3.55 | 299.36 | 5.90 | 61.28 | 2.89 | -3.14 | 0.29 | 297.70 | 21.70 | 60.25 | 26.77 | -3.02 | 1.87 | 21.34 | 3.22 |
| 300.65 | 17.77 | -2.36 | 315.51 | 7.78 | 35.53 | 7.35 | -2.52 | 0.25 | 299.62 | 31.43 | 19.17 | 24.19 | -2.16 | 0.44 | 20.49 | 9.83 |
| 303.37 | 71.01 | 4.84 | 307.93 | 9.08 | 77.16 | 4.63 | 5.15 | 1.93 | 301.60 | 11.33 | 75.40 | 8.28 | 4.52 | 2.04 | 5.35 | 4.43 |

Continued on next page

Table A.2 – continued from previous page

| $\theta_C$ | $i_C$ | $A_C$ | $\theta_M$ | $\theta_{eM}$ | $i_M$ | $i_{eM}$ | $A_M$ | $A_{eM}$ | $\theta_D$ | $\theta_{eD}$ | $i_D$ | $i_{eD}$ | $A_D$ | $A_{eD}$ | $A_{snrM}$ | $A_{snrD}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 306.12 | 88.57 | -1.31 | 306.12 | 0.00 | 98.57 | 0.00 | 8.69 | 0.00 | 304.35 | 8.89 | 87.57 | 0.52 | -1.32 | 0.24 | * | 11.19 |
| 308.23 | 75.16 | -3.27 | 303.08 | 8.80 | 81.27 | 1.82 | -3.23 | 0.67 | 306.93 | 15.46 | 80.19 | 22.30 | -2.86 | 2.58 | 9.59 | 2.22 |
| 310.49 | 52.94 | -2.45 | 303.23 | 15.68 | 51.33 | 15.63 | -1.67 | 0.59 | 310.86 | 17.81 | 60.03 | 26.46 | -2.13 | 1.30 | 5.64 | 3.26 |
| 310.77 | 25.38 | 2.91 | 307.93 | 6.42 | 35.86 | 27.20 | 2.83 | 1.40 | 309.82 | 33.45 | 30.60 | 36.24 | 2.66 | 1.76 | 4.04 | 3.02 |
| 322.84 | 38.51 | -3.56 | 332.23 | 11.54 | 45.72 | 4.05 | -3.14 | 0.24 | 321.91 | 22.16 | 43.97 | 19.32 | -3.04 | 0.84 | 26.66 | 7.23 |
| 331.42 | 46.57 | 2.60 | 330.61 | 15.03 | 54.08 | 43.05 | 2.30 | 3.53 | 331.75 | 22.54 | 53.09 | 29.66 | 2.19 | 1.78 | 1.30 | 2.46 |
| 335.49 | 3.10 | 1.63 | 332.21 | 18.83 | 11.93 | 29.80 | 1.66 | 0.33 | 270.65 | 238.41 | 5.48 | 47.24 | 1.63 | 0.88 | 10.12 | 3.71 |
| 335.65 | 10.40 | -3.45 | 315.97 | 26.80 | 22.66 | 7.92 | -3.55 | 0.21 | 317.12 | 196.93 | 11.21 | 32.21 | -3.34 | 0.63 | 34.28 | 10.58 |
| 336.93 | 32.22 | -1.72 | 343.25 | 10.54 | 40.90 | 13.38 | -1.59 | 0.33 | 333.24 | 29.79 | 36.45 | 33.16 | -1.49 | 0.61 | 9.73 | 4.86 |
| 337.22 | 4.79 | 3.45 | 334.44 | 25.36 | 14.28 | 16.31 | 3.51 | 0.30 | 282.13 | 282.57 | 5.51 | 32.69 | 3.41 | 0.74 | 23.13 | 9.18 |
| 341.60 | 70.62 | 2.34 | 347.17 | 9.77 | 76.33 | 5.09 | 2.32 | 1.11 | 341.87 | 11.60 | 75.55 | 12.48 | 2.17 | 1.30 | 4.20 | 3.34 |
| 347.43 | 55.38 | 2.82 | 352.81 | 8.27 | 64.31 | 4.30 | 2.58 | 0.39 | 344.83 | 15.17 | 63.40 | 15.10 | 2.48 | 1.12 | 13.04 | 4.45 |
| 356.22 | 79.78 | -4.63 | 350.58 | 8.33 | 86.63 | 0.58 | -5.31 | 1.05 | 354.69 | 6.71 | 85.84 | 1.44 | -4.32 | 1.41 | 10.07 | 6.13 |
| 356.87 | 41.04 | -1.74 | 348.75 | 5.06 | 57.26 | 10.38 | -1.81 | 0.52 | 349.49 | 334.64 | 48.33 | 35.94 | -1.44 | 1.02 | 6.98 | 2.81 |
| 357.06 | 43.08 | 1.15 | 354.23 | 10.58 | 42.39 | 50.70 | 0.89 | 2.04 | 351.39 | 16.52 | 50.39 | 24.71 | 1.02 | 0.55 | 0.87 | 3.71 |

### A.3.1 Simulation 1

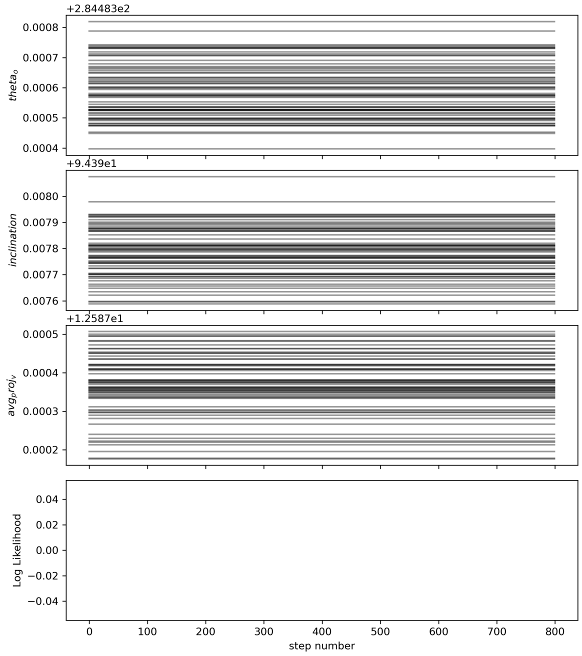


FIGURE A.8: Simulation 1: An example of a failed run with flat chains.

### A.3.2 Simulation 2

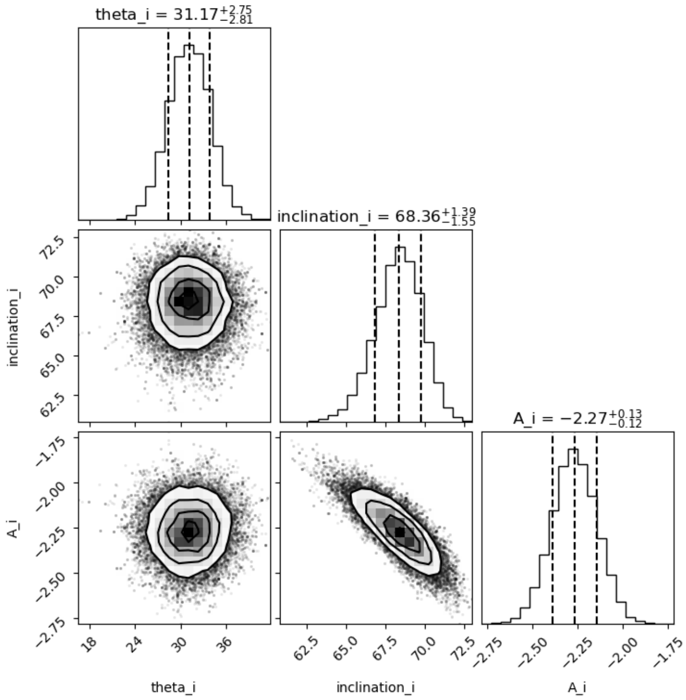


Figure A.9: Simulation 2: Corner plot for the simulated cluster with $\theta = 31^{\circ}$, $i = 63.5^{\circ}$, and $A = -2.3\,\mathrm{km\,s^{-1}}$ estimated using MCMC.

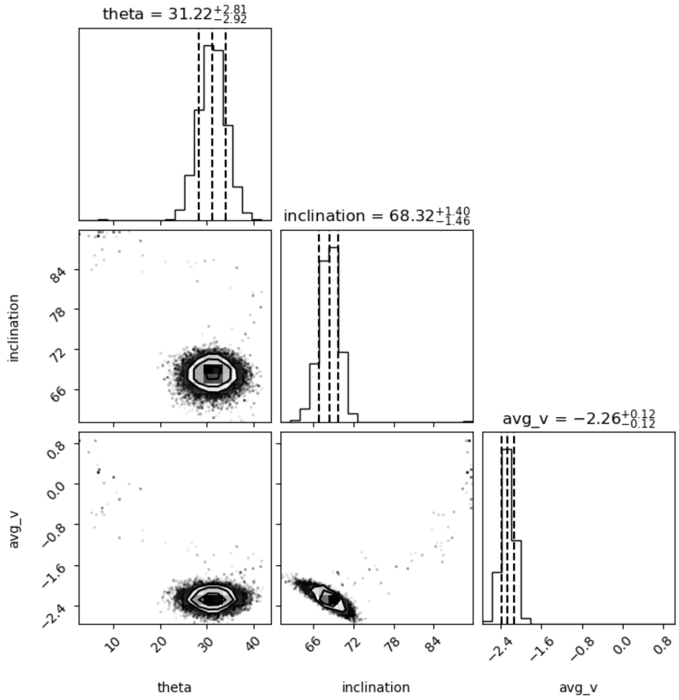


FIGURE A.10: Simulation 2: Corner plot for the simulated cluster with $\theta = 31.0^\circ$, $i = 64^\circ$, and $A = -2.3\,\mathrm{km\,s^{-1}}$ estimated using differential evolution and MCMC.

# References

K. Abazajian, J. K. Adelman-McCarthy, M. A. Agüeros, et al. The First Data Release of the Sloan Digital Sky Survey. *The Astronomical Journal*, 126:2081–2086, Oct. 2003. ISSN 0004-6256. doi: 10.1086/378165. URL https://ui.adsabs.harvard.edu/abs/2003AJ....126.2081A. ADS Bibcode: 2003AJ....126.2081A. 9

T. M. C. Abbott, F. B. Abdalla, S. Allam, et al. The Dark Energy Survey Data Release 1. *The Astrophysical Journal Supplement Series*, 239(2):18, Nov. 2018. ISSN 1538-4365. doi: 10.3847/1538-4365/aae9f0. URL http://arxiv.org/abs/1801.03181. arXiv:1801.03181 [astro-ph]. 14

R. Ahumada. The 16th Data Release of the Sloan Digital Sky Surveys: First Release from the APOGEE-2 Southern Survey and Full Release of eBOSS Spectra. *The Astrophysical Journal Supplement Series*, page 21, 2020. 43, 44

R. Ahumada et al. The 16th Data Release of the Sloan Digital Sky Surveys: First Release from the APOGEE-2 Southern Survey and Full Release of eBOSS Spectra. *Astrophys. J. Suppl.*, 249(1):3, 2020. doi: 10.3847/1538-4365/ab929e. 18, 43

A. M. Amarsi, K. Lind, Y. Osorio, et al. The GALAH Survey: non-LTE departure coefficients for large spectroscopic surveys. *A&A*, 642:A62, Oct. 2020. doi: 10.1051/0004-6361/202038650. 20

D. An, T. C. Beers, J. A. Johnson, et al. The Stellar Metallicity Distribution Function of the Galactic Halo from SDSS Photometry. *The Astrophysical Journal*, 763(1):65, Jan. 2013. ISSN 0004-637X, 1538-4357. doi: 10.1088/0004-637X/763/1/65. URL http://arxiv.org/abs/1211.7073. 12

F. Anders, C. Chiappini, I. Minchev, et al. Red giants observed by CoRoT and APOGEE: The evolution of the Milky Way's radial metallicity gradient. *Astronomy and Astrophysics*, 600: A70, Apr. 2017. ISSN 0004-6361. doi: 10.1051/0004-6361/201629363. URL https://ui.adsabs.harvard.edu/abs/2017A&A...600A..70A. ADS Bibcode: 2017A&A...600A..70A. 14

F. Anders, C. Chiappini, B. X. Santiago, et al. Dissecting stellar chemical abundance space with t-SNE. *A&A*, 619:A125, Nov. 2018. doi: 10.1051/0004-6361/201833099. 18

J. Anderson and I. R. King. The Rotation of the Globular Cluster 47 Tucanae in the Plane of the Sky. *The Astronomical Journal*, 126:772–777, Aug. 2003. ISSN 0004-6256. doi: 10.1086/376480. URL https://ui.adsabs.harvard.edu/abs/2003AJ....126..772A. ADS Bibcode: 2003AJ....126..772A. 104

C. A. L. Bailer-Jones, J. Rybizki, M. Fouesneau, et al. Estimating Distance from Parallaxes. IV. Distances to 1.33 Billion Stars in Gaia Data Release 2. *AJ*, 156(2):58, Aug. 2018. doi: 10.3847/1538-3881/aacb21. 31

C. A. L. Bailer-Jones, M. Fouesneau, and R. Andrae. Quasar and galaxy classification in Gaia Data Release 2. *Monthly Notices of the Royal Astronomical Society*, 490(4):5615–5633, Dec. 2019. ISSN 0035-8711. doi: 10.1093/mnras/stz2947. URL https://doi.org/10.1093/mnras/stz2947. xvi, 1, 13, 41, 43, 44, 46, 47, 50, 52, 53, 63, 72, 73

N. M. Ball and R. J. Brunner. Data Mining and Machine Learning in Astronomy. *International Journal of Modern Physics D*, 19:1049–1106, Jan. 2010. ISSN 0218-2718. doi: 10.1142/S0218271810017160. URL https://ui.adsabs.harvard.edu/abs/2010IJMPD..19.1049B. ADS Bibcode: 2010IJMPD..19.1049B. 4

P. S. Barklem, N. Christlieb, T. C. Beers, et al. The Hamburg/ESO R-process enhanced star survey (HERES). II. Spectroscopic analysis of the survey sample. *A&A*, 439(1):129–151, Aug. 2005. doi: 10.1051/0004-6361:20052967. 19

H. Baumgardt and M. Hilker. A catalogue of masses, structural parameters, and velocity dispersion profiles of 112 Milky Way globular clusters. *Monthly Notices of the Royal Astronomical Society*, 478(2):1520–1557, Aug. 2018. ISSN 0035-8711. doi: 10.1093/mnras/sty1057. URL https://doi.org/10.1093/mnras/sty1057. 76, 82, 83, 97

M. Bayes and M. Price. An Essay towards Solving a Problem in the Doctrine of Chances. By the Late Rev. Mr. Bayes, F. R. S. Communicated by Mr. Price, in a Letter to John Canton, A. M. F. R. S. *Philosophical Transactions*, 53:370–418, 1763. doi: 10.1098/rstl.1763.0053. URL http://rstl.royalsocietypublishing.org/content/53/370.short. 4

E. Bañados, B. P. Venemans, C. Mazzucchelli, et al. An 800-million-solar-mass black hole in a significantly neutral Universe at a redshift of 7.5. *Nature*, 553(7689):473–476, Jan. 2018. ISSN 1476-4687. doi: 10.1038/nature25180. URL https://www.nature.com/articles/nature25180. Number: 7689 Publisher: Nature Publishing Group. xv, 5

T. C. Beers and N. Christlieb. The Discovery and Analysis of Very Metal-Poor Stars in the Galaxy. *ARA&A*, 43(1):531–580, Sept. 2005. doi: 10.1146/annurev.astro.42.053102.134057. 11, 18, 39

T. C. Beers, V. M. Placco, D. Carollo, et al. Bright Metal-Poor Stars from the Hamburg/ESO Survey. II. A Chemodynamical Analysis. *ApJ*, 835(1):81, Jan. 2017. doi: 10.3847/1538-4357/835/1/81. 19

K. Bekki. Rotation and Multiple Stellar Population in Globular Clusters. *The Astrophysical Journal*, 724:L99–L103, Nov. 2010. ISSN 0004-637X. doi: 10.1088/2041-8205/724/1/L99. URL https://ui.adsabs.harvard.edu/abs/2010ApJ...724L..99B. ADS Bibcode: 2010ApJ...724L..99B. 15

P. Bianchini, A. L. Varri, G. Bertin, and A. Zocchi. Rotating Globular Clusters. *The Astrophysical Journal*, 772:67, July 2013. ISSN 0004-637X. doi: 10.1088/0004-637X/772/1/67. URL https://ui.adsabs.harvard.edu/abs/2013ApJ...772...67B. ADS Bibcode: 2013ApJ...772...67B. 76

P. Bianchini, R. P. van der Marel, A. del Pino, et al. The internal rotation of globular clusters revealed by Gaia DR2. *Monthly Notices of the Royal Astronomical Society*, 481 (2):2125–2139, Dec. 2018. ISSN 0035-8711. doi: 10.1093/mnras/sty2365. URL https://doi.org/10.1093/mnras/sty2365. 76

M. R. Blanton. Sloan Digital Sky Survey IV: Mapping the Milky Way, Nearby Galaxies, and the Distant Universe. *The Astronomical Journal*, page 35, 2017. 44

H. E. Bond, E. P. Nelan, D. A. VandenBerg, et al. HD 140283: A Star in the Solar Neighborhood that Formed Shortly after the Big Bang. *The Astrophysical Journal*, 765:L12, Mar. 2013. ISSN 0004-637X. doi: 10.1088/2041-8205/765/1/L12. URL https://ui.adsabs.harvard.edu/abs/2013ApJ...765L..12B. ADS Bibcode: 2013ApJ...765L..12B. 12

C. J. F. T. Braak. A Markov Chain Monte Carlo version of the genetic algorithm Differential Evolution: easy Bayesian computing for real parameter spaces. *Statistics and Computing*, 16(3):239–249, Sept. 2006. ISSN 1573-1375. doi: 10.1007/s11222-006-8769-1. URL https://doi.org/10.1007/s11222-006-8769-1. 80, 81

J. Brehmer, S. Mishra-Sharma, J. Hermans, et al. Mining for Dark Matter Substructure: Inferring Subhalo Population Properties from Strong Lenses with Machine Learning. *The Astrophysical Journal*, 886:49, Nov. 2019. ISSN 0004-637X. doi: 10.3847/1538-4357/ab4c41. URL https://ui.adsabs.harvard.edu/abs/2019ApJ...886...49B. ADS Bibcode: 2019ApJ...886...49B. 5

L. Breiman. Random Forests. *Machine Learning*, 45(1):5–32, Oct. 2001. ISSN 1573-0565. doi: 10.1023/A:1010933404324. URL https://doi.org/10.1023/A:1010933404324. 51

N. Brinkmann, S. Banerjee, B. Motwani, and P. Kroupa. The bound fraction of young star clusters. *Astronomy & Astrophysics*, 600:A49, Apr. 2017. ISSN 0004-6361, 1432-0746. doi: 10.1051/0004-6361/201629312. URL https://www.aanda.org/articles/aa/abs/2017/04/aa29312-16/aa29312-16.html. Publisher: EDP Sciences. 75

S. Buder, S. Sharma, J. Kos, et al. The GALAH+ Survey: Third Data Release. *arXiv e-prints*, art. arXiv:2011.02505, Nov. 2020. 18

S. Buder, S. Sharma, J. Kos, et al. The GALAH+ survey: Third data release. *MNRAS*, 506 (1):150–201, Sept. 2021. doi: 10.1093/mnras/stab1242. 20

T. Cantat-Gaudin and F. Anders. Clusters and mirages: cataloguing stellar aggregates in the Milky Way. *Astronomy and Astrophysics*, 633:A99, Jan. 2020. ISSN 0004-6361. doi: 10.1051/0004-6361/201936691. URL https://ui.adsabs.harvard.edu/abs/2020A&A...633A..99C. ADS Bibcode: 2020A&A...633A..99C. 76, 78

T. Cantat-Gaudin, P. Donati, A. Vallenari, et al. Abundances and kinematics for ten anticentre open clusters. *Astronomy and Astrophysics*, 588:A120, Apr. 2016. ISSN 0004-6361. doi: 10.1051/0004-6361/201628115. URL https://ui.adsabs.harvard.edu/abs/2016A&A...588A.120C. ADS Bibcode: 2016A&A...588A.120C. 14

T. Cantat-Gaudin, C. Jordi, L. Balaguer-Nú{ñ}ez, and A. Castro-Ginard. *A Gaia DR2 view of the open cluster population in the Milky Way*. Mar. 2019. URL https://ui.adsabs.harvard.edu/abs/2019hsax.conf..401C. Conference Name: Highlights on Spanish Astrophysics X Pages: 401-401 ADS Bibcode: 2019hsax.conf..401C. 14, 15, 76

R. Carballo, J. I. González-Serrano, C. R. Benn, and F. Jiménez-Luján. Use of neural networks for the identification of new z 3.6 QSOs from FIRST–SDSS DR5. *Monthly Notices of the Royal Astronomical Society*, 391(1):369–382, Nov. 2008. ISSN 0035-8711. doi: 10.1111/j.1365-2966.2008.13896.x. URL https://doi.org/10.1111/j.1365-2966.2008.13896.x. 5

A. Castro-Ginard, P. J. McMillan, X. Luri, et al. Milky Way spiral arms from open clusters in Gaia EDR3. *Astronomy & Astrophysics*, 652:A162, Aug. 2021. ISSN 0004-6361, 1432-0746. doi: 10.1051/0004-6361/202039751. URL https://www.aanda.org/articles/aa/abs/2021/08/aa39751-20/aa39751-20.html. Publisher: EDP Sciences. 111

L. Chao, Z. Wen-hui, and L. Ji-ming. Study of Star/Galaxy Classification Based on the XGBoost Algorithm. *Chinese Astronomy and Astrophysics*, 43(4):539–548, Oct. 2019. ISSN 0275-1062. doi: 10.1016/j.chinastron.2019.11.005. URL https://www.sciencedirect.com/science/article/pii/S0275106219300815. 50

N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16(1):321–357, June 2002. ISSN 1076-9757. 52

T. Chen and C. Guestrin. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA, Aug. 2016. Association for Computing Machinery. ISBN 978-1-4503-4232-2. doi: 10.1145/2939672.2939785. URL https://doi.org/10.1145/2939672.2939785. 43, 51

S. Chib and E. Greenberg. Understanding the Metropolis-Hastings Algorithm. *The American Statistician*, 49(4):327–335, 1995. ISSN 0003-1305. doi: 10.2307/2684568. URL https://www.jstor.org/stable/2684568. Publisher: [American Statistical Association, Taylor & Francis, Ltd.]. 80

N. Christlieb, M. S. Bessell, T. C. Beers, et al. A stellar relic from the early Milky Way. *Nature*, 419:904–906, Oct. 2002. ISSN 0028-0836. doi: 10.1038/nature01142. URL https://ui.adsabs.harvard.edu/abs/2002Natur.419..904C. ADS Bibcode: 2002Natur.419..904C. 12

J.-F. Claeskens, A. Smette, L. Vandenbulcke, and J. Surdej. Identification and redshift determination of quasi-stellar objects with medium-band photometry: application to Gaia. *Monthly Notices of the Royal Astronomical Society*, 367(3):879–904, Apr. 2006. ISSN 0035-8711. doi: 10.1111/j.1365-2966.2006.10024.x. URL https://doi.org/10.1111/j.1365-2966.2006.10024.x. 5

F. Combes, S. Leon, and G. Meylan. N-body simulations of globular cluster tides. *Astronomy and Astrophysics*, 352:149–162, Dec. 1999. ISSN 0004-6361. doi: 10.48550/arXiv.astro-ph/9910148. URL https://ui.adsabs.harvard.edu/abs/1999A&A...352..149C. ADS Bibcode: 1999A&A...352..149C. 15

A. J. Connolly and A. S. Szalay. A Robust Classification of Galaxy Spectra: Dealing with Noisy and Incomplete Data. *AJ*, 117(5):2052–2062, May 1999. doi: 10.1086/300839. 22

M. J. Cordero, V. Hénault-Brunet, C. A. Pilachowski, et al. Differences in the rotational properties of multiple stellar populations in M13: a faster rotation for the 'extreme' chemical subpopulation. *Monthly Notices of the Royal Astronomical Society*, 465:3515–3535, Mar. 2017. ISSN 0035-8711. doi: 10.1093/mnras/stw2812. URL https://ui.adsabs.harvard.edu/abs/2017MNRAS.465.3515C. ADS Bibcode: 2017MNRAS.465.3515C. 76

G. Cordoni, G. S. Da Costa, D. Yong, et al. Exploring the Galaxy's halo and very metal-weak thick disc with SkyMapper and Gaia DR2. *MNRAS*, 503(2):2539–2561, May 2021. doi: 10.1093/mnras/staa3417. 12, 36, 40

S. M. Croom, S. Fine, and t. S. S. Team. Quasar and Supermassive Black Hole Evolution. *Proceedings of the International Astronomical Union*, 5(S267):223–230, Aug. 2009. ISSN 1743-9221, 1743-9213. doi: 10.1017/S1743921310006320. URL https://www.cambridge.org/core/journals/proceedings-of-the-international-astronomical-union/article/quasar-and-supermassive-black-hole-evolution/B4DB326C8516289B3A9AFB60550B7CB3. Publisher: Cambridge University Press. 13, 42

G. S. Da Costa, M. S. Bessell, A. D. Mackey, et al. The skymapper dr1.1 search for extremely metal-poor stars. *Monthly Notices of the Royal Astronomical Society*, 489(4):5900–5918, Sep 2019. ISSN 1365-2966. doi: 10.1093/mnras/stz2550. URL http://dx.doi.org/10.1093/mnras/stz2550. 30, 31, 35, 36, 38, 39

G. S. Da Costa, M. S. Bessell, A. D. Mackey, et al. The SkyMapper DR1.1 search for extremely metal-poor stars. *MNRAS*, 489(4):5900–5918, Nov. 2019. doi: 10.1093/mnras/stz2550. 12, 32

G. Dalton, S. Trager, and et al. Project overview and update on weave: the next generation wide-field spectroscopy facility for the william herschel telescope. *Ground-based and Airborne Instrumentation for Astronomy V*, Jul 2014. doi: 10.1117/12.2055132. URL http://dx.doi.org/10.1117/12.2055132. 18, 39

G. S. Da Costa, M. S. Bessell, T. Nordlander, et al. Spectroscopic follow-up of statistically selected extremely metal-poor star candidates from GALAH DR3. *Monthly Notices of the Royal Astronomical Society*, 520(1):917–924, Mar. 2023. ISSN 0035-8711. doi: 10.1093/mnras/stad170. URL https://doi.org/10.1093/mnras/stad170. 17, 39

F. De Angeli, M. Weiler, P. Montegriffo, et al. Gaia Data Release 3: Processing and validation of BP/RP low-resolution spectral data, June 2022. URL http://arxiv.org/abs/2206.06143. arXiv:2206.06143 [astro-ph]. 44

R. S. de Jong, O. Agertz, A. A. Berbel, et al. 4most: Project overview and information for the first call for proposals. *The Messenger*, 175:3–11, 2019. 18, 39

G. M. De Silva, K. C. Freeman, J. Bland-Hawthorn, et al. The GALAH survey: scientific motivation. *Monthly Notices of the Royal Astronomical Society*, 449(3):2604–2617, 2015. ISSN 0035-8711. doi: 10.1093/mnras/stv327. URL https://doi.org/10.1093/mnras/stv327. _eprint: https://academic.oup.com/mnras/article-pdf/449/3/2604/9376648/stv327.pdf. 10, 20

R. S. de Souza, M. L. L. Dantas, M. V. Costa-Duarte, et al. A probabilistic approach to emission-line galaxy classification. *Monthly Notices of the Royal Astronomical Society*, 472(3):2808–2822, Dec. 2017. ISSN 0035-8711, 1365-2966. doi: 10.1093/mnras/stx2156. URL http://academic.oup.com/mnras/article/472/3/2808/4091443/A-probabilistic-approach-to-emissionline-galaxy. 50

L. Delchambre, C. A. L. Bailer-Jones, I. Bellas-Velidis, et al. Gaia DR3: Apsis III – Non-stellar content and source classification. *Astronomy & Astrophysics*, June 2022. ISSN 0004-6361, 1432-0746. doi: 10.1051/0004-6361/202243423. URL http://arxiv.org/abs/2206.06710. arXiv:2206.06710 [astro-ph]. 63

A. V. Dorogush, V. Ershov, and A. Gulin. CatBoost: gradient boosting with categorical features support. page 7, 2017. 43, 51

G. Efstathiou, R. S. Ellis, and B. A. Peterson. Analysis of a complete galaxy redshift survey - II. The field-galaxy luminosity function. *Monthly Notices of the Royal Astronomical Society*, 232(2):431–461, May 1988. ISSN 0035-8711, 1365-2966. doi: 10.1093/mnras/232.2.431. URL https://academic.oup.com/mnras/article-lookup/doi/10.1093/mnras/232.2.431. 4

M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A Density-based Algorithm for Discovering Clusters a Density-based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, pages 226–231, Portland, Oregon, 1996. AAAI Press. URL http://dl.acm.org/citation.cfm?id=3001460.3001507. 21

Euclid Collaboration, R. Scaramella, J. Amiaux, et al. Euclid preparation. I. The Euclid Wide Survey. *Astronomy and Astrophysics*, 662:A112, June 2022. ISSN 0004-6361. doi: 10.1051/0004-6361/202141938. URL https://ui.adsabs.harvard.edu/abs/2022A&A...662A.112E. ADS Bibcode: 2022A&A...662A.112E. 4

E. D. Feigelson. Statistics in astronomy. *arXiv preprint arXiv:0903.0416*, 2009. URL https://arxiv.org/abs/0903.0416. 3

R. A. Fisher. On the Mathematical Foundations of Theoretical Statistics. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 222(594-604):309, Jan. 1922. doi: 10.1098/rsta.1922.0009. URL http://rsta.royalsocietypublishing.org/content/222/594-604/309.abstract. 4

D. Foreman-Mackey, D. W. Hogg, D. Lang, and J. Goodman. emcee: The MCMC Hammer. *Publications of the Astronomical Society of the Pacific*, 125:306, Mar. 2013. ISSN 0004-6280. doi: 10.1086/670067. URL https://ui.adsabs.harvard.edu/abs/2013PASP..125..306F. ADS Bibcode: 2013PASP..125..306F. 82

C. Fraley and A. E. Raftery. Model-Based Clustering, Discriminant Analysis, and Density Estimation. *Journal of the American Statistical Association*, 97(458):611–631, June 2002. ISSN 0162-1459. doi: 10.1198/016214502760047131. URL https://doi.org/10.1198/016214502760047131. 43, 50

A. Frebel and J. E. Norris. Near-Field Cosmology with Extremely Metal-Poor Stars. *ARA&A*, 53:631–688, Aug. 2015. doi: 10.1146/annurev-astro-082214-122423. 11, 18, 39

J. H. Friedman. Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics*, 29(5):1189–1232, 2001. ISSN 0090-5364. URL https://www.jstor.org/stable/2699986. Publisher: Institute of Mathematical Statistics. 51

Gaia Collaboration, T. Prusti, J. H. J. de Bruijne, et al. The *Gaia* mission. *Astronomy & Astrophysics*, 595:A1, Nov. 2016. ISSN 0004-6361, 1432-0746. doi: 10.1051/0004-6361/201629272. URL http://www.aanda.org/10.1051/0004-6361/201629272. 10, 43

Gaia Collaboration, A. G. A. Brown, A. Vallenari, et al. Gaia Data Release 2. Summary of the contents and survey properties. *A&A*, 616:A1, Aug. 2018. doi: 10.1051/0004-6361/201833051. 20, 31

Gaia Collaboration, A. G. A. Brown, A. Vallenari, et al. *Gaia* Early Data Release 3: Summary of the contents and survey properties. *Astronomy & Astrophysics*, 649:A1, May 2021. ISSN 0004-6361, 1432-0746. doi: 10.1051/0004-6361/202039657. URL https://www.aanda.org/10.1051/0004-6361/202039657. 43

Gaia Collaboration, C. A. L. Bailer-Jones, and al. Gaia Data Release 3. The extra-galactic content. *Astronomy & Astrophysics*, June 2022. ISSN 0004-6361, 1432-0746. doi: 10.1051/0004-6361/202243232. URL https://www.aanda.org/10.1051/0004-6361/202243232. xvi, 10, 43, 46, 60, 61, 63, 65, 73, 82

C. F. Gauss and G. Stewart. *Theory of the Combination of Observations Least Subject to Errors, Part One, Part Two, Supplement.* SIAM, 1995. 3

C. J. Geyer. Markov Chain Monte Carlo Maximum Likelihood. Interface Foundation of North America, 1991. URL http://conservancy.umn.edu/handle/11299/58440. Accepted: 2010-02-24T20:38:06Z. 80

S. Giacalone, C. D. Dressing, E. L. N. Jensen, et al. Vetting of 384 TESS Objects of Interest with TRICERATOPS and Statistical Validation of 12 Planet Candidates. *The Astronomical Journal*, 161:24, Jan. 2021. ISSN 0004-6256. doi: 10.3847/1538-3881/abc6af. URL https://ui.adsabs.harvard.edu/abs/2021AJ....161...24G. ADS Bibcode: 2021AJ....161...24G. 5

A. Golob, M. Sawicki, A. D. Goulding, and J. Coupon. Classifying stars, galaxies, and AGNs in CLAUDS + HSC-SSP using gradient boosted decision trees. *Monthly Notices of the Royal Astronomical Society*, 503(3):4136–4146, May 2021. ISSN 0035-8711. doi: 10.1093/mnras/stab719. URL https://doi.org/10.1093/mnras/stab719. 50

J. Goodman and J. Weare. Ensemble samplers with affine invariance. *Communications in Applied Mathematics and Computational Science*, 5:65–80, Jan. 2010. doi: 10.2140/camcos.2010.5.65. URL https://ui.adsabs.harvard.edu/abs/2010CAMCS...5...65G. ADS Bibcode: 2010CAMCS...5...65G. 82

S. Gossage, C. Conroy, A. Dotter, et al. Age Determinations of the Hyades, Praesepe, and Pleiades via MESA Models with Rotation. *The Astrophysical Journal*, 863:67, Aug. 2018. ISSN 0004-637X. doi: 10.3847/1538-4357/aad0a0. URL https://ui.adsabs.harvard.edu/abs/2018ApJ...863...67G. ADS Bibcode: 2018ApJ...863...67G. 112

R. G. Gratton, E. Carretta, and A. Bragaglia. Multiple populations in globular clusters. *The Astronomy and Astrophysics Review*, 20(1):50, Feb. 2012. ISSN 1432-0754. doi: 10.1007/s00159-012-0050-3. URL https://doi.org/10.1007/s00159-012-0050-3. 14

B. Gustafsson, B. Edvardsson, K. Eriksson, et al. A grid of MARCS model atmospheres for late-type stars - I. Methods and general properties. *Astronomy & Astrophysics*, 486(3):951–970, Aug. 2008. ISSN 0004-6361, 1432-0746. doi: 10.1051/0004-6361:200809724. URL https://www.aanda.org/articles/aa/abs/2008/30/aa09724-08/aa09724-08.html. Number: 3 Publisher: EDP Sciences. 21

C. J. Hao, Y. Xu, S. B. Bian, et al. On the Nature of Rotation in the Praesepe Cluster. *The Astrophysical Journal*, 938:100, Oct. 2022. ISSN 0004-637X. doi: 10.3847/1538-4357/ac92fc. URL https://ui.adsabs.harvard.edu/abs/2022ApJ...938..100H. ADS Bibcode: 2022ApJ...938..100H. 76, 112

C. M. Harrison, T. Costa, C. N. Tadhunter, et al. AGN outflows and feedback twenty years on. *Nature Astronomy*, 2(3):198–205, Mar. 2018. ISSN 2397-3366. doi: 10.1038/s41550-018-0403-6. URL https://www.nature.com/articles/s41550-018-0403-6. 42

K. Hawkins, G. Zeimann, C. Sneden, et al. The Stars of the HETDEX Survey. I. Radial Velocities and Metal-poor Stars from Low-resolution Stellar Spectra. *ApJ*, 911(2):108, Apr. 2021. doi: 10.3847/1538-4357/abe9bd. 18

B. F. Healy, P. R. McCullough, and K. C. Schlaufman. Stellar Spins in the Pleiades, Praesepe, and M35 Open Clusters. *The Astrophysical Journal*, 923(1):23, Dec. 2021. ISSN 0004-637X. doi: 10.3847/1538-4357/ac281d. URL https://dx.doi.org/10.3847/1538-4357/ac281d. Publisher: The American Astronomical Society. 76, 112

J. K. Hollek, A. Frebel, I. U. Roederer, et al. The Chemical Abundances of Stars in the Halo (CASH) Project. II. A Sample of 14 Extremely Metal-poor Stars. *ApJ*, 742(1):54, Nov. 2011. doi: 10.1088/0004-637X/742/1/54. 19

J. Hong, E. Kim, H. M. Lee, and R. Spurzem. Comparative study between N-body and Fokker-Planck simulations for rotating star clusters - II. Two-component models. *Monthly Notices of the Royal Astronomical Society*, 430:2960–2972, Apr. 2013. ISSN 0035-8711. doi: 10.1093/mnras/stt099. URL https://ui.adsabs.harvard.edu/abs/2013MNRAS.430.2960H. ADS Bibcode: 2013MNRAS.430.2960H. 76

E. Hubble. A relation between distance and radial velocity among extra-galactic nebulae. *Proceedings of the National Academy of Sciences*, 15(3):168–173, Mar. 1929. doi: 10.1073/pnas.15.3.168. URL http://www.pnas.org/content/15/3/168.short. 4

A. Hughes. Needles in a haystack: advanced statistical techniques and large stellar spectroscopic datasets, 2017. URL http://hdl.handle.net/1959.14/1284747. 18, 37

A. C. N. Hughes, C. A. L. Bailer-Jones, and S. Jamal. Quasar and galaxy classification using Gaia EDR3 and CatWise2020. *Astronomy and Astrophysics*, 668:A99, Dec. 2022a. ISSN 0004-6361. doi: 10.1051/0004-6361/202244859. URL https://ui.adsabs.harvard.edu/abs/2022A&A...668A..99H. ADS Bibcode: 2022A&A...668A..99H. 41

A. C. N. Hughes, L. R. Spitler, D. B. Zucker, et al. The GALAH Survey: A New Sample of Extremely Metal-poor Stars Using a Machine-learning Classification Algorithm. *The Astrophysical Journal*, 930(1):47, May 2022b. ISSN 0004-637X. doi: 10.3847/1538-4357/ac5fa7. URL https://dx.doi.org/10.3847/1538-4357/ac5fa7. Publisher: The American Astronomical Society. 17

Z. Ivezic, S. M. Kahn, J. A. Tyson, et al. LSST: From Science Drivers to Reference Design and Anticipated Data Products. *The Astrophysical Journal*, 873:111, Mar. 2019. ISSN 0004-637X. doi: 10.3847/1538-4357/ab042c. URL https://ui.adsabs.harvard.edu/abs/2019ApJ...873..111I. ADS Bibcode: 2019ApJ...873..111I. 4

H. R. Jacobson, S. Keller, A. Frebel, et al. High-Resolution Spectroscopic Study of Extremely Metal-Poor Star Candidates from the SkyMapper Survey. *ApJ*, 807(2):171, July 2015. doi: 10.1088/0004-637X/807/2/171. 19

P. Jofré, G. Traven, K. Hawkins, et al. Climbing the cosmic ladder with stellar twins in RAVE with Gaia. *MNRAS*, 472(3):2517–2533, Dec. 2017. doi: 10.1093/mnras/stx1877. 18

N. Kacharov, P. Bianchini, A. Koch, et al. A study of rotating globular clusters. The case of the old, metal-poor globular cluster NGC 4372. *Astronomy and Astrophysics*, 567:A69, July 2014. ISSN 0004-6361. doi: 10.1051/0004-6361/201423709. URL https://ui.adsabs.harvard.edu/abs/2014A&A...567A..69K. ADS Bibcode: 2014A&A...567A..69K. 76

S. Kamann, N. J. Bastian, M. Gieles, et al. Linking the rotation of a cluster to the spins of its stars: the kinematics of NGC 6791 and NGC 6819 in 3D. *Monthly Notices of the Royal Astronomical Society*, 483(2):2197–2206, Feb. 2019. ISSN 0035-8711. doi: 10.1093/mnras/sty3144. URL https://doi.org/10.1093/mnras/sty3144. 76, 112

S. C. Keller, M. S. Bessell, A. Frebel, et al. A single low-energy, iron-poor supernova as the source of metals in the star SMSS J031300.36-670839.3. *nature*, 506(7489):463–466, Feb. 2014. doi: 10.1038/nature12990. 12, 21, 36

E. N. Kirby, P. Guhathakurta, J. D. Simon, et al. Multi-element Abundance Measurements from Medium-resolution Spectra. II. Catalog of Stars in Milky Way Dwarf Satellite Galaxies. *ApJS*, 191(2):352–375, Dec. 2010. doi: 10.1088/0067-0049/191/2/352. 19

A. Koch, M. Hanke, and N. Kacharov. Kinematics of outer halo globular clusters: M 75 and NGC 6426. *Astronomy and Astrophysics*, 616:A74, Aug. 2018. ISSN 0004-6361. doi: 10.1051/0004-6361/201833110. URL https://ui.adsabs.harvard.edu/abs/2018A&A...616A..74K. ADS Bibcode: 2018A&A...616A..74K. 76

J. A. Kollmeier, G. Zasowski, H.-W. Rix, et al. Sdss-v: Pioneering panoptic spectroscopy, 2017. 18

J. Kos, J. Lin, T. Zwitter, et al. The GALAH survey: the data reduction pipeline. *MNRAS*, 464(2):1259–1281, Jan. 2017. doi: 10.1093/mnras/stw2064. 20

J. Kos, J. Bland-Hawthorn, K. Freeman, et al. The GALAH survey: chemical tagging of star clusters and new members in the Pleiades. *MNRAS*, 473(4):4612–4633, Feb. 2018. doi: 10.1093/mnras/stx2637. 18

A. L. Kraus and L. A. Hillenbrand. The Stellar Populations of Praesepe and Coma Berenices. *The Astronomical Journal*, 134:2340–2352, Dec. 2007. ISSN 0004-6256. doi: 10.1086/522831. URL https://ui.adsabs.harvard.edu/abs/2007AJ....134.2340K. ADS Bibcode: 2007AJ....134.2340K. 112

J. H. Krijthe. *Rtsne: T-Distributed Stochastic Neighbor Embedding using Barnes-Hut Implementation*, 2015. URL https://github.com/jkrijthe/Rtsne. R package version 0.15. 24

P. Kroupa, S. Aarseth, and J. Hurley. The formation of a bound star cluster: from the Orion nebula cluster to the Pleiades. *Monthly Notices of the Royal Astronomical Society*, 321:699–712, Mar. 2001. ISSN 0035-8711. doi: 10.1046/j.1365-8711.2001.04050.x. URL https://ui.adsabs.harvard.edu/abs/2001MNRAS.321..699K. ADS Bibcode: 2001MNRAS.321..699K. 75

M. Kunimoto and J. M. Matthews. Searching the Entirety of Kepler Data. II. Occurrence Rate Estimates for FGK Stars. *The Astronomical Journal*, 159:248, June 2020. ISSN 0004-6256. doi: 10.3847/1538-3881/ab88b0. URL https://ui.adsabs.harvard.edu/abs/2020AJ....159..248K. ADS Bibcode: 2020AJ....159..248K. 5

A. Kurcz, M. Bilicki, A. Solarz, et al. Towards automatic classification of all WISE sources. *Astronomy & Astrophysics*, 592:A25, Aug. 2016. ISSN 0004-6361, 1432-0746. doi: 10.1051/0004-6361/201628142. URL https://www.aanda.org/articles/aa/abs/2016/08/aa28142-16/aa28142-16.html. 42

C. J. Lada and E. A. Lada. Embedded Clusters in Molecular Clouds. *Annual Review of Astronomy and Astrophysics*, 41:57–115, Jan. 2003. ISSN 0066-4146. doi: 10.1146/annurev.astro.41.011802.094844. URL https://ui.adsabs.harvard.edu/abs/2003ARA&A..41...57L. ADS Bibcode: 2003ARA&A..41...57L. 75

S. E. Lake and C. W. Tsai. An exploration of how training set composition bias in machine learning affects identifying rare objects. *Astronomy and Computing*, 40:100617, July 2022. ISSN 2213-1337. doi: 10.1016/j.ascom.2022.100617. URL https://www.sciencedirect.com/science/article/pii/S2213133722000440. 52

K. J. Lee, L. Guillemot, Y. L. Yue, et al. Application of the Gaussian mixture model in pulsar astronomy - pulsar classification and candidates ranking for the Fermi 2FGL catalogue: Application of the Gaussian mixture model. *Monthly Notices of the Royal Astronomical Society*, 424(4):2832–2840, Aug. 2012. ISSN 00358711. doi: 10.1111/j.1365-2966.2012.21413.x. URL https://academic.oup.com/mnras/article-lookup/doi/10.1111/j.1365-2966.2012.21413.x. 50

Y. S. Lee, T. C. Beers, T. Masseron, et al. Carbon-enhanced Metal-poor Stars in SDSS/SEGUE. I. Carbon Abundance Estimation and Frequency of CEMP Stars. *AJ*, 146(5):132, Nov. 2013. doi: 10.1088/0004-6256/146/5/132. 12, 39

H.-N. Li, G. Zhao, N. Christlieb, et al. SPECTROSCOPIC ANALYSIS OF METAL-POOR STARS FROM LAMOST: EARLY RESULTS. *The Astrophysical Journal*, 798(2):110, jan 2015. doi: 10.1088/0004-637x/798/2/110. URL https://doi.org/10.1088/0004-637x/798/2/110. 19

J. Li, B. P. Venemans, F. Walter, et al. Spatially Resolved Molecular Interstellar Medium in a z = 6.6 Quasar Host Galaxy. *The Astrophysical Journal*, 930(1):27, May 2022. ISSN 0004-637X, 1538-4357. doi: 10.3847/1538-4357/ac61d7. URL https://iopscience.iop.org/article/10.3847/1538-4357/ac61d7. 42

M. Lochner, J. D. McEwen, H. V. Peiris, et al. Photometric Supernova Classification with Machine Learning. *The Astrophysical Journal Supplement Series*, 225:31, Aug. 2016. ISSN 0067-0049. doi: 10.3847/0067-0049/225/2/31. URL https://ui.adsabs.harvard.edu/abs/2016ApJS..225...31L. ADS Bibcode: 2016ApJS..225...31L. 4

L. B. Lucy. An iterative technique for the rectification of observed distributions. *The Astronomical Journal*, 79:745, June 1974. ISSN 00046256. doi: 10.1086/111605. URL http://adsabs.harvard.edu/cgi-bin/bib_query?1974AJ.....79..745L. 4

B. W. Lyke, A. N. Higley, J. N. McLane, et al. The Sloan Digital Sky Survey Quasar Catalog: Sixteenth Data Release. *The Astrophysical Journal Supplement Series*, 250(1):8, Aug. 2020. ISSN 1538-4365. doi: 10.3847/1538-4365/aba623. URL https://iopscience.iop.org/article/10.3847/1538-4365/aba623. 44

N. Lützgendorf, M. Kissler-Patig, K. Gebhardt, et al. Central kinematics of the globular cluster NGC 2808: upper limit on the mass of an intermediate-mass black hole. *Astronomy & Astrophysics*, 542:A129, June 2012. ISSN 0004-6361, 1432-0746. doi: 10.1051/0004-6361/201219375. URL https://www.aanda.org/articles/aa/abs/2012/06/aa19375-12/aa19375-12.html. Publisher: EDP Sciences. 15

L. v. d. Maaten and G. Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(Nov):2579–2605, 2008. URL http://www.jmlr.org/papers/v9/vandermaaten08a.html. 18, 24

M. K. Mardini, V. M. Placco, A. Taani, et al. Metal-poor stars observed with the automated planet finder telescope. II. chemodynamical analysis of six low-metallicity stars in the halo system of the milky way. *The Astrophysical Journal*, 882(1):27, aug 2019. doi: 10.3847/1538-4357/ab3047. URL https://doi.org/10.3847/1538-4357/ab3047. 19

F. Marocco, P. R. M. Eisenhardt, J. W. Fowler, et al. The CatWISE2020 Catalog. *The Astrophysical Journal Supplement Series*, 253(1):8, Mar. 2021. ISSN 0067-0049, 1538-4365. doi: 10.3847/1538-4365/abd805. URL http://arxiv.org/abs/2012.13084. 11, 43

S. L. Martell, S. Sharma, S. Buder, et al. The GALAH survey: observational overview and Gaia DR1 companion. *Monthly Notices of the Royal Astronomical Society*, 465(3):3203–3219, 2017. URL http://mnras.oxfordjournals.org/content/465/3/3203.short. 10, 21

T. Masseron, J. A. Johnson, B. Plez, et al. A holistic approach to carbon-enhanced metal-poor stars. *A&A*, 509:A93, jan 2010. doi: 10.1051/0004-6361/200911744. 19

A. Mastrobuono-Battisti and H. B. Perets. Evolution of Second-generation Stars in Stellar Disks of Globular and Nuclear Clusters:  Centauri as a Test Case. *The Astrophysical Journal*, 779:85, Dec. 2013. ISSN 0004-637X. doi: 10.1088/0004-637X/779/1/85. URL https://ui.adsabs.harvard.edu/abs/2013ApJ...779...85M. ADS Bibcode: 2013ApJ...779...85M. 76

G. Matijevič, C. Chiappini, E. K. Grebel, et al. Very metal-poor stars observed by the RAVE survey. *A&A*, 603:A19, July 2017. doi: 10.1051/0004-6361/201730417. 18

A. McWilliam, G. W. Preston, C. Sneden, and L. Searle. Spectroscopic Analysis of 33 of the Most Metal Poor Stars. II. *AJ*, 109:2757, June 1995. doi: 10.1086/117486. 19

A. M. Meisner, D. Lang, E. F. Schlafly, and D. J. Schlegel. unWISE Coadds: The Five-year Data Set. *Publications of the Astronomical Society of the Pacific*, 131:124504, Dec. 2019. ISSN 0004-6280. doi: 10.1088/1538-3873/ab3df4. URL https://ui.adsabs.harvard.edu/abs/2019PASP..131l4504M. ADS Bibcode: 2019PASP..131l4504M. 10

Meléndez, Jorge, Placco, Vinicius M., Tucci-Maia, Marcelo, et al. 2mass j18082002-5104378: The brightest (v = 11.9) ultra metal-poor star. *A&A*, 585:L5, 2016. doi: 10.1051/0004-6361/201527456. URL https://doi.org/10.1051/0004-6361/201527456. 19

P. Márquez-Neila, C. Fisher, R. Sznitman, and K. Heng. Supervised machine learning for analysing spectra of exoplanetary atmospheres. *Nature Astronomy*, 2:719–724, June 2018. ISSN 2397-3366. doi: 10.1038/s41550-018-0504-2. URL https://ui.adsabs.harvard.edu/abs/2018NatAs...2..719M. ADS Bibcode: 2018NatAs...2..719M. 5

A. Möller, V. Ruhlmann-Kleider, C. Leloup, et al. Photometric classification of type Ia supernovae in the SuperNova Legacy Survey with supervised learning. *Journal of Cosmology and Astroparticle Physics*, 2016(12):008–008, Dec. 2016. ISSN 1475-7516. doi: 10.1088/1475-7516/2016/12/008. URL https://doi.org/10.1088/1475-7516/2016/12/008. 50

T. Nordlander, A. M. Amarsi, K. Lind, et al. 3D NLTE analysis of the most iron-deficient star, SMSS0313-6708. *A&A*, 597:A6, Jan. 2017a. doi: 10.1051/0004-6361/201629202. 21

T. Nordlander, A. M. Amarsi, K. Lind, et al. 3D NLTE analysis of the most iron-deficient star, SMSS0313-6708. *A&A*, 597:A6, Jan. 2017b. doi: 10.1051/0004-6361/201629202. 19

T. Nordlander, M. S. Bessell, G. S. Da Costa, et al. The lowest detected stellar Fe abundance: the halo star SMSS J160540.18-144323.1. *Monthly Notices of the Royal Astronomical Society*, 488:L109–L113, Sept. 2019. ISSN 0035-8711. doi: 10.1093/mnrasl/slz109. URL http://adsabs.harvard.edu/abs/2019MNRAS.488L.109N. 12, 21

M. Ntampaka, D. J. Eisenstein, S. Yuan, and L. H. Garrison. A Hybrid Deep Learning Approach to Cosmological Constraints from Galaxy Redshift Surveys. *The Astrophysical Journal*, 889:151, Feb. 2020. ISSN 0004-637X. doi: 10.3847/1538-4357/ab5f5e. URL https://ui.adsabs.harvard.edu/abs/2020ApJ...889..151N. ADS Bibcode: 2020ApJ...889..151N. 5

C. A. Onken, C. Wolf, M. S. Bessell, et al. SkyMapper Southern Survey: Second data release (DR2). *PASA*, 36:e033, Aug. 2019. doi: 10.1017/pasa.2019.27. 30

F. Pedregosa, G. Varoquaux, A. Gramfort, et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, Oct. 2011. doi: 10.48550/arXiv.1201.0490. URL https://ui.adsabs.harvard.edu/abs/2011JMLR...12.2825P. ADS Bibcode: 2011JMLR...12.2825P. xv, 8

N. Piskunov and J. A. Valenti. Spectroscopy Made Easy: Evolution. *A&A*, 597:A16, Jan. 2017. doi: 10.1051/0004-6361/201629124. 20

V. M. Placco, A. Frebel, T. C. Beers, et al. Metal-poor Stars Observed with the Magellan Telescope. II. Discovery of Four Stars with [Fe/H] <= -3.5. *ApJ*, 781(1):40, Jan. 2014a. doi: 10.1088/0004-637X/781/1/40. 19

V. M. Placco, A. Frebel, T. C. Beers, and R. J. Stancliffe. Carbon-enhanced Metal-poor Star Frequencies in the Galaxy: Corrections for the Effect of Evolutionary Status on Carbon Abundances. *ApJ*, 797(1):21, Dec. 2014b. doi: 10.1088/0004-637X/797/1/21. 36

V. M. Placco, R. M. Santucci, T. C. Beers, et al. The R-Process Alliance: Spectroscopic Follow-up of Low-metallicity Star Candidates from the Best & Brightest Survey. *ApJ*, 870 (2):122, Jan. 2019. doi: 10.3847/1538-4357/aaf3b9. 12, 19

R. L. Plackett. Studies in the History of Probability and Statistics: VII. The Principle of the Arithmetic Mean. *Biometrika*, 45(1/2):130–135, 1958. ISSN 0006-3444. doi: 10.2307/2333051. URL https://www.jstor.org/stable/2333051. Publisher: [Oxford University Press, Biometrika Trust]. 3

B. Plez. Turbospectrum: Code for spectral synthesis. *Astrophysics Source Code Library*, page ascl:1205.004, May 2012. URL http://adsabs.harvard.edu/abs/2012ascl.soft05004P. 21

A. C. Quillen. Prospecting for Spiral Structure in the Flocculent Outer Milky Way Disk with Color-Magnitude Star Counts from the Two Micron All Sky Survey. *The Astronomical Journal*, 124:924–930, Aug. 2002. ISSN 0004-6256. doi: 10.1086/341379. URL https://ui.adsabs.harvard.edu/abs/2002AJ....124..924Q. ADS Bibcode: 2002AJ....124..924Q. 111, 112

A. C. Quillen, E. Nolting, I. Minchev, et al. Migration in the shearing sheet and estimates for young open cluster migration. *Monthly Notices of the Royal Astronomical Society*, 475:4450–4466, Apr. 2018. ISSN 0035-8711. doi: 10.1093/mnras/sty125. URL https://ui.adsabs.harvard.edu/abs/2018MNRAS.475.4450Q. ADS Bibcode: 2018MNRAS.475.4450Q. 14

P. Re Fiorentin, C. A. L. Bailer-Jones, Y. S. Lee, et al. Estimation of stellar atmospheric parameters from SDSS/SEGUE spectra. *A&A*, 467(3):1373–1387, June 2007. doi: 10.1051/0004-6361:20077334. 22

G. T. Richards, R. P. Deo, M. Lacy, et al. Eight-Dimensional Mid-Infrared/Optical Bayesian Quasar Selection. *The Astronomical Journal*, 137:3884–3899, Apr. 2009a. ISSN 0004-6256. doi: 10.1088/0004-6256/137/4/3884. URL https://ui.adsabs.harvard.edu/abs/2009AJ....137.3884R. ADS Bibcode: 2009AJ....137.3884R. 13

G. T. Richards, A. D. Myers, A. G. Gray, et al. Efficient Photometric Selection of Quasars from the Sloan Digital Sky Survey. II. ~1,000,000 Quasars from Data Release 6. *The Astrophysical Journal Supplement Series*, 180:67–83, Jan. 2009b. ISSN 0067-0049. doi: 10.1088/0067-0049/180/1/67. URL https://ui.adsabs.harvard.edu/abs/2009ApJS..180...67R. ADS Bibcode: 2009ApJS..180...67R. 5

J. W. Richards, D. L. Starr, N. R. Butler, et al. On Machine-learned Classification of Variable Stars with Sparse and Noisy Time-series Data. *The Astrophysical Journal*, 733:10, May 2011. ISSN 0004-637X. doi: 10.1088/0004-637X/733/1/10. URL https://ui.adsabs.harvard.edu/abs/2011ApJ...733...10R. ADS Bibcode: 2011ApJ...733...10R. 4

M. Riello, F. De Angeli, D. W. Evans, et al. *Gaia* Early Data Release 3: Photometric content and validation. *Astronomy & Astrophysics*, 649:A3, May 2021. ISSN 0004-6361, 1432-0746. doi: 10.1051/0004-6361/202039587. URL https://www.aanda.org/10.1051/0004-6361/202039587. 10

C. M. Sakari, V. M. Placco, E. M. Farrell, et al. The R-Process Alliance: First Release from the Northern Search for r-process-enhanced Metal-poor Stars in the Galactic Halo. *ApJ*, 868(2):110, Dec. 2018. doi: 10.3847/1538-4357/aae9df. 12, 19

N. E. Sanders, A. M. Soderberg, S. Gezari, et al. Toward Characterization of the Type IIP Supernova Progenitor Population: A Statistical Sample of Light Curves from Pan-STARRS1. *The Astrophysical Journal*, 799:208, Feb. 2015. ISSN 0004-637X. doi: 10.1088/0004-637X/799/2/208. URL https://ui.adsabs.harvard.edu/abs/2015ApJ...799..208S. ADS Bibcode: 2015ApJ...799..208S. 4

E. F. Schlafly, A. M. Meisner, and G. M. Green. The unWISE Catalog: Two Billion Infrared Sources from Five Years of *WISE* Imaging. *The Astrophysical Journal Supplement Series*, 240(2):30, Feb. 2019. ISSN 1538-4365. doi: 10.3847/1538-4365/aafbea. URL https://iopscience.iop.org/article/10.3847/1538-4365/aafbea. 10

K. C. Schlaufman and A. R. Casey. THE BEST AND BRIGHTEST METAL-POOR STARS. *The Astrophysical Journal*, 797(1):13, nov 2014. doi: 10.1088/0004-637x/797/1/13. URL https://doi.org/10.1088/0004-637x/797/1/13. 19

D. P. Schneider, P. B. Hall, G. T. Richards, et al. The Sloan Digital Sky Survey Quasar Catalog. IV. Fifth Data Release. *The Astronomical Journal*, 134:102–117, July 2007. ISSN 0004-6256. doi: 10.1086/518474. URL https://ui.adsabs.harvard.edu/abs/2007AJ....134..102S. ADS Bibcode: 2007AJ....134..102S. 13

F. Sestito, N. Longeard, N. F. Martin, et al. Tracing the formation of the Milky Way through ultra metal-poor stars. *MNRAS*, 484(2):2166–2180, Apr. 2019. doi: 10.1093/mnras/stz043. 12

F. Sestito, N. F. Martin, E. Starkenburg, et al. The Pristine survey - X. A large population of low-metallicity stars permeates the Galactic disc. *MNRAS*, 497(1):L7–L12, Sept. 2020. doi: 10.1093/mnrasl/slaa022. 36, 40

C. J. Shallue and A. Vanderburg. Identifying Exoplanets with Deep Learning: A Five-planet Resonant Chain around Kepler-80 and an Eighth Planet around Kepler-90. *The Astronomical Journal*, 155:94, Feb. 2018. ISSN 0004-6256. doi: 10.3847/1538-3881/aa9e09. URL https://ui.adsabs.harvard.edu/abs/2018AJ....155...94S. ADS Bibcode: 2018AJ....155...94S. 5

S. Sharma, D. Stello, S. Buder, et al. The TESS-HERMES survey data release 1: high-resolution spectroscopy of the TESS southern continuous viewing zone. *MNRAS*, 473(2):2004–2019, Jan. 2018. doi: 10.1093/mnras/stx2582. 37

F. H. Shu, F. C. Adams, and S. Lizano. Star formation in molecular clouds: observation and theory. *Annual Review of Astronomy and Astrophysics*, 25:23–81, Jan. 1987. ISSN 0066-4146. doi: 10.1146/annurev.aa.25.090187.000323. URL https://ui.adsabs.harvard.edu/abs/1987ARA&A..25...23S. ADS Bibcode: 1987ARA&A..25...23S. 75

J. D. Simpson, P. L. Cottrell, and C. C. Worley. Spectral matching for abundances and clustering analysis of stars on the giant branches of $\omega$ Centauri. *MNRAS*, 427(2):1153–1167, Dec. 2012. doi: 10.1111/j.1365-2966.2012.22012.x. 19

J. D. Simpson, G. M. De Silva, J. Bland-Hawthorn, et al. The GALAH survey: relative throughputs of the 2dF fibre positioner and the HERMES spectrograph from stellar targets. *Monthly Notices of the Royal Astronomical Society*, 459:1069–1081, June 2016. ISSN 0035-8711. doi: 10.1093/mnras/stw746. URL https://ui.adsabs.harvard.edu/abs/2016MNRAS.459.1069S. ADS Bibcode: 2016MNRAS.459.1069S. 10

V. V. Smith, N. B. Suntzeff, K. Cunha, et al. The Chemical Evolution of the Globular Cluster Centauri (NGC 5139). *The Astronomical Journal*, 119:1239–1258, Mar. 2000. ISSN 0004-6256. doi: 10.1086/301276. URL https://ui.adsabs.harvard.edu/abs/2000AJ....119.1239S. ADS Bibcode: 2000AJ....119.1239S. 15

A. Sollima, H. Baumgardt, and M. Hilker. The eye of Gaia on globular clusters kinematics: internal rotation. *Monthly Notices of the Royal Astronomical Society*, 485:1460–1476, May 2019. ISSN 0035-8711. doi: 10.1093/mnras/stz505. URL https://ui.adsabs.harvard.edu/abs/2019MNRAS.485.1460S. ADS Bibcode: 2019MNRAS.485.1460S. xxii, 1, 76, 77, 78, 79, 82, 97, 98, 104, 106, 109, 118

J. Southworth. Homogeneous studies of transiting extrasolar planets – IV. Thirty systems with space-based light curves. *Monthly Notices of the Royal Astronomical Society*, 417(3):2166–2196, Nov. 2011. ISSN 0035-8711. doi: 10.1111/j.1365-2966.2011.19399.x. URL https://doi.org/10.1111/j.1365-2966.2011.19399.x. xv, 5, 6

F. Spite, M. Spite, B. Barbuy, et al. Abundance patterns of the light neutron-capture elements in very and extremely metal-poor stars. *A&A*, 611:A30, Mar. 2018. doi: 10.1051/0004-6361/201732096. 12, 19

E. Starkenburg, N. Martin, K. Youakim, et al. The Pristine survey – I. Mining the Galaxy for the most metal-poor stars. *Monthly Notices of the Royal Astronomical Society*, 471(3):2587–2604, 05 2017. ISSN 0035-8711. doi: 10.1093/mnras/stx1068. URL https://doi.org/10.1093/mnras/stx1068. 33, 38

M. Steinmetz, G. Matijevič, H. Enke, et al. The sixth data release of the radial velocity experiment (rave). i. survey description, spectra, and radial velocities. *The Astronomical Journal*, 160(2):82, Jul 2020. ISSN 1538-3881. doi: 10.3847/1538-3881/ab9ab9. URL http://dx.doi.org/10.3847/1538-3881/ab9ab9. 18

T. Suda, Y. Katsuta, S. Yamada, et al. Stellar Abundances for the Galactic Archeology (SAGA) Database — Compilation of the Characteristics of Known Extremely Metal-Poor Stars. *PASJ*, 60:1159, Oct. 2008. doi: 10.1093/pasj/60.5.1159. 18

A. Szalay and J. Gray. The world-wide telescope. *Science*, 293(5537):2037–2040, 2001. URL http://science.sciencemag.org/content/293/5537/2037.short. 4

Y.-S. Ting, K. C. Freeman, C. Kobayashi, et al. Principal component analysis on chemical abundances spaces. *MNRAS*, 421(2):1231–1255, Apr. 2012. doi: 10.1111/j.1365-2966.2011.20387.x. 22

M. A. Tiongco, E. Vesperini, and A. L. Varri. The complex kinematics of rotating star clusters in a tidal field. *Monthly Notices of the Royal Astronomical Society*, 475:L86–L90, Mar. 2018. ISSN 0035-8711. doi: 10.1093/mnrasl/sly009. URL https://ui.adsabs.harvard.edu/abs/2018MNRAS.475L..86T. ADS Bibcode: 2018MNRAS.475L..86T. 76

G. Traven, G. Matijevič, J. Kos, et al. The Galah Survey: Classification and diagnostics with t-SNE reduction of spectral information. *arXiv preprint arXiv:1612.02242*, 2017. URL https://arxiv.org/abs/1612.02242. 20, 24

G. Traven, S. Feltzing, T. Merle, et al. The GALAH survey: multiple stars and our Galaxy. I. A comprehensive method for deriving properties of FGK binary stars. *A&A*, 638:A145, June 2020. doi: 10.1051/0004-6361/202037484. 18

J. Tumlinson. Chemical Evolution in Hierarchical Models of Cosmic Structure II: The Formation of the Milky Way Stellar Halo and the Distribution of the Oldest Stars. *The Astrophysical Journal*, 708(2):1398–1418, Jan. 2010. ISSN 0004-637X, 1538-4357. doi: 10.1088/0004-637X/708/2/1398. URL http://arxiv.org/abs/0911.1786. 11

D. G. Turner. The Power of Archival Astronomy. *Journal of the American Association of Variable Star Observers (JAAVSO)*, 31:160–170, Dec. 2003. ISSN 0271-9053. URL https://ui.adsabs.harvard.edu/abs/2003JAVSO..31..160T. ADS Bibcode: 2003JAVSO..31..160T. 9

J. A. Valenti and N. Piskunov. Spectroscopy made easy: A new tool for fitting observations with synthetic spectra. *A&AS*, 118:595–603, Sept. 1996. 20

A. Vallenari, A. G. A. Brown, and T. Prusti. Gaia Data Release 3. Summary of the content and survey properties. *Astronomy & Astrophysics*, June 2022. ISSN 0004-6361, 1432-0746. doi: 10.1051/0004-6361/202243940. URL https://www.aanda.org/articles/aa/abs/forth/aa43940-22/aa43940-22.html. Publisher: EDP Sciences. 46

G. van de Ven, R. C. E. van den Bosch, E. K. Verolme, and P. T. de Zeeuw. The dynamical distance and intrinsic structure of the globular cluster \omega Centauri. *Astronomy and Astrophysics*, 445:513–543, Jan. 2006. ISSN 0004-6361. doi: 10.1051/0004-6361:20053061. URL https://ui.adsabs.harvard.edu/abs/2006A&A...445..513V. ADS Bibcode: 2006A&A...445..513V. 78, 82

S. van den Bergh. The Flattening of Globular Clusters. *The Astronomical Journal*, 135:1731–1737, May 2008. ISSN 0004-6256. doi: 10.1088/0004-6256/135/5/1731. URL https://ui.adsabs.harvard.edu/abs/2008AJ....135.1731V. ADS Bibcode: 2008AJ....135.1731V. 76

F. van Leeuwen, R. S. Le Poole, R. A. Reijns, et al. A proper motion study of the globular cluster  Centauri. *Astronomy and Astrophysics*, 360:472–498, Aug. 2000. ISSN 0004-6361. URL https://ui.adsabs.harvard.edu/abs/2000A&A...360..472V. ADS Bibcode: 2000A&A...360..472V. 15, 76, 107

E. Vasiliev. Proper motions and dynamics of the Milky Way globular cluster system from Gaia DR2. *Monthly Notices of the Royal Astronomical Society*, 484(2):2832–2850, Apr. 2019. ISSN 0035-8711. doi: 10.1093/mnras/stz171. URL https://doi.org/10.1093/mnras/stz171. 15

F. Villaescusa-Navarro, D. Anglés-Alcázar, S. Genel, et al. The CAMELS Project: Cosmology and Astrophysics with Machine-learning Simulations. *The Astrophysical Journal*, 915:71, July 2021. ISSN 0004-637X. doi: 10.3847/1538-4357/abf7ba. URL https://ui.adsabs.harvard.edu/abs/2021ApJ...915...71V. ADS Bibcode: 2021ApJ...915...71V. 5

M. Wattenberg, F. Viégas, and I. Johnson. How to Use t-SNE Effectively. *Distill*, 2016. doi: 10.23915/distill.00002. URL http://distill.pub/2016/misread-tsne. 26

R. A. Wittenmyer, S. Sharma, D. Stello, et al. The K2-HERMES Survey. I. Planet-candidate Properties from K2 Campaigns 1-3. *AJ*, 155(2):84, Feb. 2018. doi: 10.3847/1538-3881/aaa3e4. 37

E. L. Wright, P. R. M. Eisenhardt, A. K. Mainzer, et al. THE WIDE-FIELD INFRARED SURVEY EXPLORER (WISE): MISSION DESCRIPTION AND INITIAL ON-ORBIT PERFORMANCE. *The Astronomical Journal*, 140(6):1868–1881, Dec. 2010. ISSN 0004-6256, 1538-3881. doi: 10.1088/0004-6256/140/6/1868. URL https://iopscience.iop.org/article/10.1088/0004-6256/140/6/1868. 42

C. W. Yip, A. J. Connolly, D. E. Vanden Berk, et al. Spectral Classification of Quasars in the Sloan Digital Sky Survey: Eigenspectra, Redshift, and Luminosity Effects. *AJ*, 128(6):2603–2630, Dec. 2004. doi: 10.1086/425626. 22

D. Yong, B. W. Carney, and M. L. Teixera de Almeida. Elemental Abundance Ratios in Stars of the Outer Galactic Disk. I. Open Clusters. *The Astronomical Journal*, 130:597–625, Aug. 2005. ISSN 0004-6256. doi: 10.1086/430934. URL https://ui.adsabs.harvard.edu/abs/2005AJ....130..597Y. ADS Bibcode: 2005AJ....130..597Y. 14

D. Yong, J. E. Norris, M. S. Bessell, et al. THE MOST METAL-POOR STARS. II. CHEMICAL ABUNDANCES OF 190 METAL-POOR STARS INCLUDING 10 NEW STARS WITH [fe/h] $\leqslant -3.5$, ,. *The Astrophysical Journal*, 762(1):26, dec 2012. doi: 10.1088/0004-637x/762/1/26. URL https://doi.org/10.1088/0004-637x/762/1/26. 19

D. Yong, J. E. Norris, M. S. Bessell, et al. The Most Metal-poor Stars. II. Chemical Abundances of 190 Metal-poor Stars Including 10 New Stars with [Fe/H] <= -3.5. *ApJ*, 762(1):26, Jan. 2013. doi: 10.1088/0004-637X/762/1/26. 39

D. Yong, G. S. Da Costa, M. S. Bessell, et al. High-resolution spectroscopic follow-up of the most metal-poor candidates from SkyMapper DR1.1. *MNRAS*, 507(3):4102–4119, Nov. 2021. doi: 10.1093/mnras/stab2001. 32, 37

D. G. York, J. Adelman, J. John E. Anderson, et al. The Sloan Digital Sky Survey: Technical Summary. *The Astronomical Journal*, 120(3):1579, Sept. 2000. ISSN 1538-3881. doi: 10.1086/301513. URL https://iopscience.iop.org/article/10.1086/301513/meta. Publisher: IOP Publishing. 9, 42

K. Youakim, E. Starkenburg, N. F. Martin, et al. The Pristine Survey - VIII. The metallicity distribution function of the Milky Way halo down to the extremely metal-poor regime. *MNRAS*, 492(4):4986–5002, Mar. 2020. doi: 10.1093/mnras/stz3619. 32

C. Yèche, P. Petitjean, J. Rich, et al. Artificial neural networks for quasar selection and photometric redshift determination. *Astronomy and Astrophysics*, 523:A14, Nov. 2010. ISSN 0004-6361. doi: 10.1051/0004-6361/200913508. URL https://ui.adsabs.harvard.edu/abs/2010A&A...523A..14Y. ADS Bibcode: 2010A&A...523A..14Y. 13

Y. Zhang and Y. Zhao. Classification in Multidimensional Parameter Space: Methods and Examples. *Publications of the Astronomical Society of the Pacific*, 115(810):1006–1018, 2003. ISSN 0004-6280. doi: 10.1086/376847. URL https://www.jstor.org/stable/10.1086/376847. Publisher: [The University of Chicago Press, Astronomical Society of the Pacific]. 5