How to Accurately Map the Milky Way with a Billion Sources

A Journey into the *Gaia*-verse

Andrew Everall

Institute of Astronomy
University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy

Hughes Hall

10th August 2021

Declaration

This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the preface and specified in the text. It is not substantially the same as any work that has already been submitted before for any degree or other qualification except as declared in the preface and specified in the text.

It does not exceed the prescribed word limit of 60,000 words for the Faculty of Physics & Chemistry Degree Committee.

The vast majority of the content of this thesis has either been published or submitted for publication. The papers which the thesis draws material from are

Ch. 2 Published in *Monthly Notices of the Royal Astronomical Society* as
The Tilt of the Local Velocity Ellipsoid as seen by *Gaia*
Everall, Evans, Belokurov & Schönrich 2019

Ch. 3 Published in *Monthly Notices of the Royal Astronomical Society* as
seestar: Selection Functions for Spectroscopic Surveys of the Milky Way
Everall & Das 2020

Ch. 4 Accepted for publication in *Monthly Notices of the Royal Astronomical Society* as
Completeness of the *Gaia*-verse V: Astrometry and Radial Velocity sample selection functions in *Gaia* EDR3
Everall & Boubert 2021
This Chapter also includes material from Boubert & Everall 2020

Ch. 5 Published in *Monthly Notices of the Royal Astronomical Society* as
Completeness of the *Gaia*-verse IV: The Astrometry Spread Function of *Gaia* DR2
Everall, Boubert, Koposov, Smith & Holl 2021c

Ch. 6 Submitted to *Monthly Notices of the Royal Astronomical Society* as
The Photo-Astrometric Vertical Tracer Density of the Milky Way:
I – The Method.
Everall, Evans, Belokurov, Boubert & Grand 2021a,
II – Results from *Gaia*
Everall, Belokurov, Evans, Boubert & Grand 2021b.

Andrew Everall
10th August 2021

How to Accurately Map the Milky Way with a Billion Sources

by Andrew Everall

Abstract

Accurately modelling the phase-space distribution of Milky Way sources relies on two vital components: unbiased distance estimators and survey selection functions. Without either, models are susceptible to significant systematic uncertainties.

My case study of the tilt of the local velocity ellipsoid demonstrates this. Well-constructed distances for the *Gaia* DR2 RVS sample return a velocity ellipsoid broadly consistent with spherical alignment. Using the reciprocal parallax distance estimator significantly alters the conclusions.

I produce selection functions for catalogues needed to model the phase-space structure of the Galaxy. My spectrograph selection function method is generalisable to many multi-fibre observatories. I supplement this with tools to combine selection functions for unions of samples and transform from observable to intrinsic coordinates. I produce selection functions for *Gaia* catalogues including astrometry and RVS samples. My model fits the complex behaviour of the *Gaia* spacecraft impressively well.

To enhance our understanding of the published *Gaia* astrometry, I introduce the Astrometric Spread Function, the expected covariance for a simple point source in *Gaia*. This reproduces the mean behaviour of published observations to degree level resolution.

This is brought together to model the vertical distribution of Milky Way sources. Systematics are minimized by marginalising over parallax uncertainties and regulating the likelihood with *Gaia* EDR3 selection functions. The veracity of the method is demonstrated on a *Gaia*-like mock population. Applying to *Gaia* EDR3, I infer a north-south asymmetry weaker than previously reported and provide updated parameter values for the vertical scale heights of the thin and thick disks, the halo power-law exponent, local stellar mass density and surface density of the Milky Way.

My thesis demonstrates the potential of *Gaia* when distances are well modelled and incompleteness is accounted for. My tools will be invaluable for answering further questions about the Milky Way using future *Gaia* data.

Acknowledgements

My PhD journey started in earnest with Payel inviting me back to the Rudolf Peierls Center for Theoretical Physics in Oxford to continue my Masters' research. Without her support, confidence and encouragement I would not have even thought to undertake a PhD so it's no understatement to say that she changed the direction of my life. This was enabled by the generous financial support of the Oxford Physics department and James Binney's ERC grant.

This thesis was only possible thanks to the Studentship funding of the Science and Technology Facilities Council of the United Kingdom. I am extremely lucky to have had this level of accessible funding available to me.

I am enormously privileged to have attended the Cambridge Institute of Astronomy, both a highly prestigious institute and one that values a friendly, inclusive atmosphere. In particular, huge thanks go to my supervisors, Professors Wyn and Vasily who have shown me how to keep up with an incredibly high paced field and even push ahead of it. The entire streams group have been incredibly supportive and I thank all of them for interesting and exciting discussions and for including me in some of their awesome projects. I'll particularly highlight Sergey and Eugene whose inputs motivated significant improvements to my work. I've massively enjoyed collaborating with fellow PhD students Zephyr and Peter. I'm also very grateful to Paul whose open-minded advice inspired me towards my next step, I'll let you know how I get on in Biostatistics.

At least half of my thesis was only possible thanks to the *Completeness of the Gaia-verse* collaboration. Working with Douglas over the last two years has taught me more about statistical modelling than I thought was possible to know. I hope the ethos of *CoG* is maintained in *GaiaUnlimited*.

I also want to thank Victoria and my parents for enduring my horrific spelling and grammar when proof reading many of my papers and reports. I hope you learned something about my "wobbly stars" along the way.

Over the last year, as for so many, I have almost entirely worked from home. I would be remiss not to acknowledge the influence that Jake, Steve, Adam, Matthew and Regor (otherwise known as "Hefty House") had on me. Thank you for keeping me sane, or sending me insane, it made the last 18 months bearable, even enjoyable. I speculated.

Andrew Everall
10th August 2021

Scientific Acknowledgements

My work was made possible due to the publicly available data and software which I have been able to use. I will acknowledge the key resources here.

Gaia This work makes use of data from the European Space Agency (ESA) mission *Gaia* (https://www.cosmos.esa.int/gaia), processed by the *Gaia* Data Processing and Analysis Consortium (DPAC, https://www.cosmos.esa.int/web/gaia/dpac/consortium). Funding for the DPAC has been provided by national institutions, in particular the institutions participating in the *Gaia* Multilateral Agreement.

LAMOST Chapter 2 uses data from Guoshoujing Telescope (the Large Sky Area Multi-Object Fiber Spectroscopic Telescope LAMOST) is a National Major Scientific Project built by the Chinese Academy of Sciences. Funding for the project has been provided by the National Development and Reform Commission. LAMOST is operated and managed by the National Astronomical Observatories, Chinese Academy of Sciences.

Galaxia A code for generating synthetic models of the Galaxy (Sharma et al., 2011). I use Galaxia in Chapter 3 for generating mock Milky Way samples.

`selectionfunctions` I used and contributed to the development of `selectionfunctions`. The `selectionfunctions` package aspires to provide a uniform interface to the selection functions of the major astronomical surveys. This package is entirely derivative of the truly excellent dustmaps package created by Green 2018. The selectionfunctions package is a product of the Completeness of the Gaia-verse collaboration (CoG, https://www.gaiaverse.space).

`scanninglaw` I have developed the package `scanninglaw`. `scanninglaw` provides and easy-to-use portal to the *Gaia* scanning law. This is also based on `dustmaps` (Green, 2018).

`NumPy` All of my work is performed in PYTHON and is made hugely more effective and efficient thanks to the `NumPy` code repository.

`SciPy` I use numerical optimization recipes from `SciPy` for much of the model fitting I performed.

`emcee` I use `emcee` (Foreman-Mackey et al., 2013), an MIT licensed pure-PYTHON implementation of Goodman & Weare's Affine Invariant Markov chain Monte Carlo (MCMC) Ensemble sampler.

`Astropy` I makes use of Astropy (http://www.astropy.org) a community-developed core PYTHON package for Astronomy (Astropy Collaboration et al., 2013, 2018).

`multiprocessing` Many of the problems I have produced solutions for are highly computationally intensive. I use `multiprocessing` to parallelize across the multi-core computers I have had access to through the Institute of Astronomy.

`Numba` `Numba` translates Python functions to optimized machine code at runtime using the industry-standard LLVM compiler library. `Numba`-compiled numerical algorithms in PYTHON can approach the speeds of C or FORTRAN. I use `Numba` in particular in Chapter 6 to significantly improve numerical integration run times.

`matplotlib` All figures are generated using `matplotlib`.

`corner.py` Some figures are generated with my own adaptations of `corner.py`, a plotting package built around `matplotlib`.

Table of contents

List of figures

List of tables

1

Introduction

"Hey Google, what's your favourite star?"

"Favourite star? There are too many to choose from."
A short conversation with Google Home.

There are approximately one hundred billion stars in our Galaxy, the Milky Way. Some are like our Sun, others are very different. Dispersed between the stars is dust and gas which has been accreted into the Galaxy or expelled from dying stars. There is an additional, dominant mass component, dark matter, which we think is composed of invisible particles permeating the visible Galaxy and extending out far beyond.

We have been trying to answer many questions about the Milky Way ever since we learned that our Sun is just one of billions of stars in the Galaxy. What is the shape of the Galaxy? Where are all of the stars located and how many of them are there? Where have these stars come from and where are they going to? What are their properties? How bright and hot are they and exactly what are they made of?

Whilst these questions are important to understand our place in the Milky Way, their answers will also help solve more fundamental questions about the Universe. How do Galaxies form and evolve and what will become of them as the Universe continues to age? Where is this mysterious dark matter and what is it even made of? Is it a modification to our understanding of gravity or is it truly a new unclassified type of matter? What are the physical processes shaping our Galaxy, from the dynamics of stellar populations to fusion within individual star cores?

Correctly answering these questions requires groundbreaking observations and the tool kits to analyse their information content. In this thesis I provide vital tools for obtaining unbiased inference from the latest pioneering astronomical surveys. I apply these tools to answer questions about the spatial and velocity structure of stars in the Milky Way.

Before delving into the main content of the thesis, I provide a brief history of Milky Way maps, introducing core concepts of astrometry, photometry and spectroscopy along the way. I introduce the missions dominating the modern era of Milky Way astronomy. Leveraging the data produced by these missions requires using statistical techniques which are appropriate for the processes occurring in data collecting. I explain some of the key statistical concepts which arise recurrently throughout the thesis such as Bayes' theorem, selection functions and statistical and systematic uncertainties.

If you learn something new along the way, I will consider my PhD to have been a worthwhile endeavour. If you learn nothing, then you will be relieved to know I am moving into Biology[1].

1.1 A Census of the Sky

Astrometry is the science of measuring precise positions of sources on the sky and how these positions change with time.

Human interest in positions and motions of celestial bodies has influenced our cultures for millennia. From Stonehenge, aligned for solstices and still holding unresolved secrets to this day[2], to the entirely independent Aboriginal cultures using a sophisticated understanding of the sky for calendars and navigation (Norris, 2016), many of the developments in ancient history still influence how astronomy is studied today.

The ancient Sumarians, dating to around 2000BCE, gave us the sexagesimal (base sixty) counting system which we still use for angles today. We typically divide the circle into 360 degrees and further still with 60 arcminutes to a degree and 60 arcseconds to an arcminute. The Greek astronomer Erathosthenes developed the system of Equatorial coordinates which is still the standard astronomical convention for on-sky position used today.

Equatorial Coordinates:

> Equatorial coordinates are aligned with the orbital axis of the Earth. Declination provides the latitude of a source above the plane aligned with the Earth's equator whilst Right Ascension is the longitudinal angle measured Eastward from the Solar position at the vernal equinox.[3] I will use the standard notation for equatorial right ascension, α, and declination, δ.

Our on-sky coordinates can be rotated for convenience depending on the research being undertaken. Take the map of the World. Conventionally latitude is measured from the Equator and zero longitude is at the Prime Meridian. However, one could equally define a coordinate system with zero latitude defined as the geodesic connecting Beijing and New York and longitude zero'd on Alexandria (where Erathosthenes died c. 194 BCE). We use the Equator to define our coordinates as it is aligned with the Earth's rotation such that the northern hemisphere has it's longest day of the year around the 21st June and southern hemisphere near the 21st December.

I will tend to provide figures in Galactic coordinates with longitude l and latitude b where the Galactic plane has $b = 0$ and the center of the Milky Way is at the coordinates $l = 0, b = 0$. In Fig. 1.1 I show Mollweide projections of the apparent magnitude of stars across the sky in Equatorial – aligned with the Earth's equator, Ecliptic – aligned with

[1] As of October 2021, I will no longer be working in Astronomy but in Biostatistics for cancer research. However I will be taking all of the statistical techniques I've learned with me.

[2] https://www.smithsonianmag.com/history/what-lies-beneath-Stonehenge-180952437/

[3] The Earth's rotation axis is not stationary but slowly oscillates so in modern day astronomy we instead use stars measured by *Hipparcos* (defining ICRS) or extragalactic radio sources (quasars) from *Gaia* (defining *Gaia*-CRF2) to fix the equatorial coordinate system.

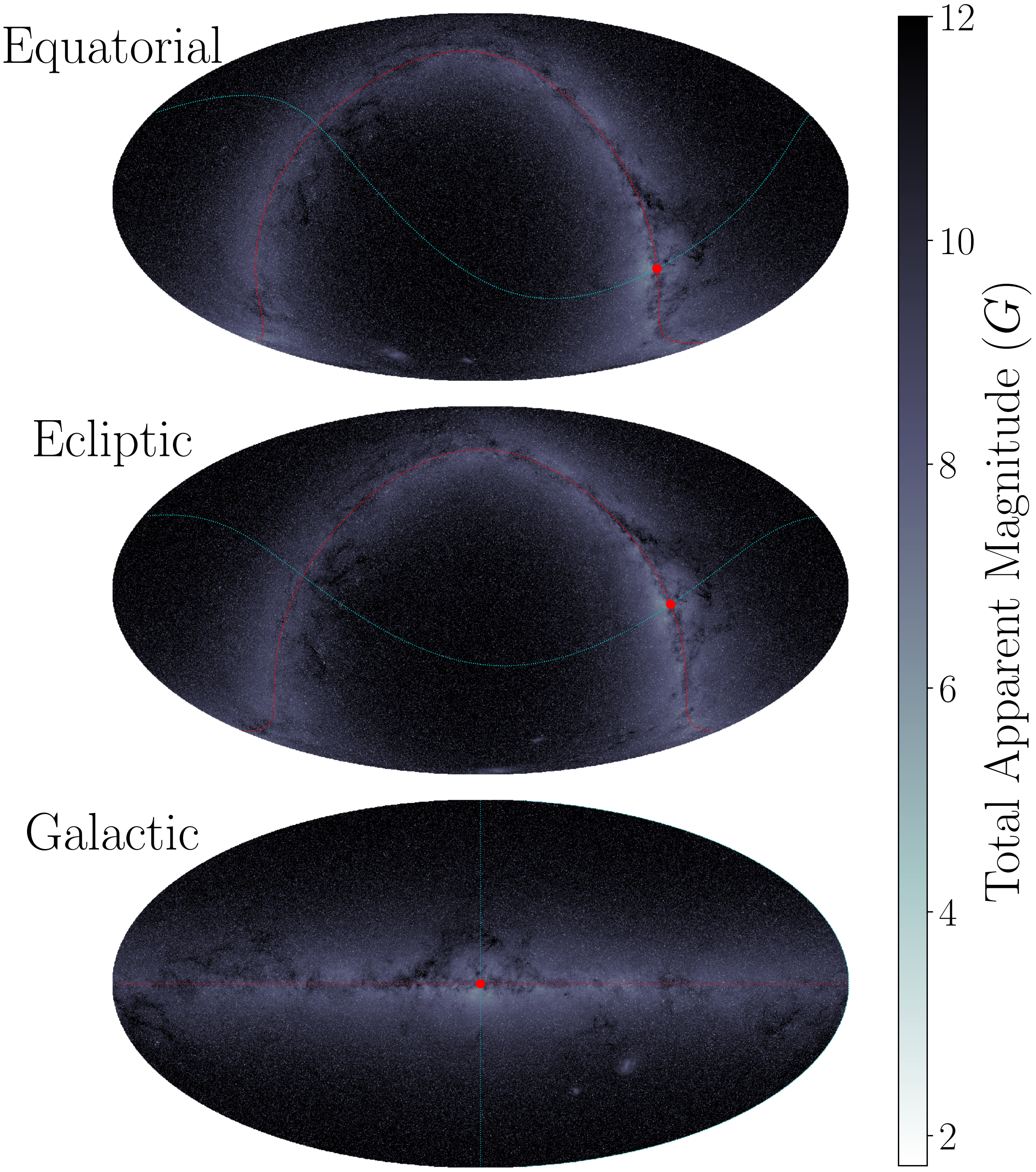


Fig. 1.1 The apparent magnitude of the sky, calculated by summing G-band flux of *Gaia* EDR3 sources in pixels and converting this to apparent magnitude. This is what the sky might look like to the naked eye in perfect observing condition (e.g. from the International Space Station). This is shown in Equatorial (top), Ecliptic (middle) and Galactic coordinates (bottom). For reference, I also show the Galactic plane (white dotted line) and great circle at 0 and 180 longitude (cyan dotted line) which cross at the Galactic centre (red dot) and anticentre.

Earth's orbit around the Sun and Galactic – aligned with the Sun's orbit around the Milky Way.

Hipparchus produced one of the earliest known cosmic censuses which was developed and published in Ptolomy's Almagest (c.100 – c.170 CE, Toomer, 1984). This catalogue measured star positions accurate to ~ 1 degree across the sky.

Greek astronomers are credited with many more astronomical developments such as determining the periodic motions of celestial bodies, inferring that the Earth is spherical and measuring size of the Sun and Moon, however their legacy was that of a geocentric Universe. In their view, the celestial spheres moved around the Earth with epicyclic motions, a system which would take nearly two millennia to overcome across astronomical research.

1.1.1 Renaissance of Astronomy

The astronomical perspective shifted with Copernicus' explanation of planetary epicycles through a Sun-centred Solar system. Kepler completed this explanation with the contribution of elliptical planetary orbits. Kepler's astronomical achievements were only possible in large part due to the precise catalogue of astrometric measurements made by Tycho Brahe and published by Kepler as the Rudolphine Tables (Kepler et al., 1627). These measurements achieved ~ 20 arcsecond accuracy on the sky and are considered the last developments in observational astronomy without using a telescope.

Copernicus' discovery significantly opened the field of astrometry. The Earth was no longer considered at the centre of the Universe and was free to move with respect to other celestial bodies. This raised many possibilities for observing apparent motions of other objects due to the Earth's relative motion.

Halley 1717 noticed that stars were in different positions to those observed and measured by ancient Greek astronomers. He used this to make the first measurement of stellar proper motion, the apparent linear motion of a star across the sky due to its velocity relative to the Solar system barycentre[4]. In Fig. 1.2, I show an image of the sky taken with the Pan-STARRS1 survey in 2012 around Vega. The red dashed track in the first inset shows the position of Vega as a function of time over the last 180 years which has an approximately linear track due to the star's high proper motion.

Proper Motion:

This is the apparent rate of change of position of a star on the sky due to its velocity relative to the Solar system[5]. A star 30 light years (one ly is 9.46×10^{12} km) away moving at 100km/s will track 2 arcseconds per year across the sky[6]. If the star were 300 ly away, it would appear to move at a slower 0.2 arcseconds/y. That's like traversing the width of a hair placed at the opposite end of a football pitch each

[4] The barycentre is the centre of mass of the Solar system. The motion of the barycentre is unaffected by the dynamics of planets and so it can be used to define the position of the Solar system.

[5] More specifically relative to the Solar system barycentre, the centre of mass of the Solar system.

[6] Proper motion: $\mu = \frac{100\mathrm{km/s}}{30\times 9.46\times 10^{12}\mathrm{km}} \times \frac{3\times 10^{7}\mathrm{s/y}\times(360\times 3600\mathrm{arcseconds})}{2\pi\mathrm{radians}} = 2\mathrm{arcseconds/y}$

year. Proper motion in the equatorial right ascension and declination directions is denoted by μ_α and μ_δ.

Bradley 1727 discovered the effect of *aberration* on source apparent positions. The apparent direction of light reaching an observer changes from the observer's perspective depending on their velocity. The apparent position of stars on the sky likewise changes due to the Earth's changing velocity as we orbit the Sun. In addition to this, Bradley 1748 also discovered *nutation*, the oscillation of the Earth's orbital axis relative to that expected from precession alone.

However, the holy grail of astrometry at the time was to measure parallax.

Parallax:

The Earth's motion around the Sun means that we observe the sky from different positions at different times of year. Hold your thumb in front of you at arm's length and see what it covers in the background. Then shift your head one way or the other whilst keeping your thumb stationary and notice how it appears to move relative to the background. If you bring your thumb to a few inches in front of your face you'll notice the apparent motion is much larger. This is the case with parallax motion. Stars move with respect to the background of distant objects. The closer the star, the greater that apparent motion (or parallax). The apparent position of a star 3.09×10^{13} km (or ~ 3 ly) away will change by 1 arcsecond when observed from the Earth relative to being observed from the Solar system centre. This distance is called a *parsec* ($1\text{pc} \equiv 3.09 \times 10^{13}\text{km}$). Parallax is denoted as ϖ.

However, it wasn't until Bessel 1838, Henderson 1840 and von Struve 1840 that parallax was finally measured for any stars. The reason for this is simple, the distance between stars is vast.

If the Sun were the size of a tennis ball on Jesus Green in Cambridge, the Earth would be no larger than a grain of sand. The nearest star system, Alpha Centauri, for which Henderson measured the parallax, would be as far away as Bulgaria. Meanwhile, Bessel's star, 61 Cygni, would be the same distance as Qatar and von Struve's, Vega, would be in Indonesia. To measure the parallax shifts of these stars requires precision beyond one tenth of an arcsecond. That's 0.00003 degrees!

This level of astrometric precision was only possible thanks to the invention of the telescope two centuries earlier and even then was incredibly challenging.

The second inset of Fig. 1.2 shows the position of Vega from the 25th July 2014 to the 28th May 2017, the time frame used for data published in the latest data release of the *Gaia* mission. Proper motion causes the star to travel from the bottom left to the top right of the panel and parallax motion produces the loops. The size of these loops gives the parallax from which the distance to Vega can be inferred.

1.1.2 Beyond the Human Eye

Another revolution in astronomy came with the invention of photography with John Whipple and George Bond taking the first photograph of Vega in 1850. The ability to

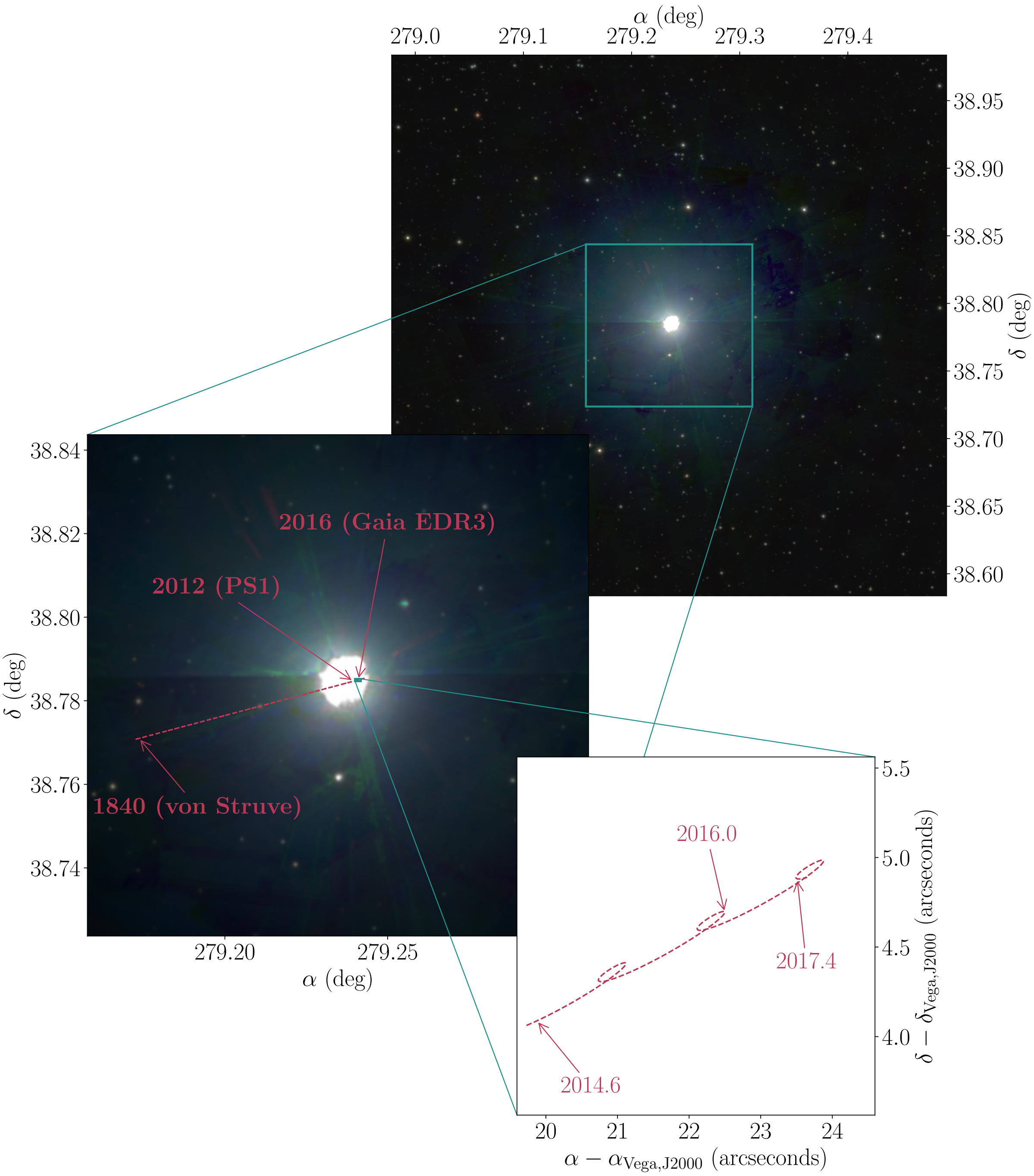


Fig. 1.2 The top image, taken from the Pan-STARRS1 survey (PS1, Chambers et al., 2016), shows a quarter of a square degree of the sky with Vega at the centre, approximately the same as a finger nail placed at arms length. The red dashed line in the first inset shows Vega's position from 1840, when von Struve observed the star, to the present day where the PS1 image was taken in 2012. Zooming in much further, the second inset shows the motion of Vega over the *Gaia* time frame from the 25th July 2014 to 28th May 2017 where the axes are in arcseconds relative to the position of Vega in 2000. The loops are the result of apparent parallax motion of Vega due to the Earth's orbit of the Sun changing our perspective. This is the motion which von Struve measured in 1840 for Vega and which *Gaia* measures for 1.5 billion sources on the sky.

precisely record stellar positions which could then be analysed by a team of people led to the production of large catalogues of stars. A noteable example being the *Carte du Ceil* which recorded photographic measurements for over five million stars over nearly 60 years of observations with the last exposure taken in 1950.

Photographic surveys led to a boom in parallax measurements with large catalogues being produced. Key works include Schlesinger et al. 1935, Jenkins 1952 and van Altena et al. 1995 which eventually recorded parallaxes for eight thousand sources.

Developments in electronics throughout the 20th century eventually led to the invention of the charge-coupled device (CCD Boyle & Smith, 1970). An array of capacitors is integrated onto a silicon sheet. Photons incident on the silicon liberate electrons which are accumulated by the capacitors. A voltage applied across the array causes the charge to be transferred between capacitors until it reaches the output node where the signal is amplified, measured and digitized.

Several key advantages of CCDs are particularly relevant to astronomy. Firstly they have high quantum efficiency, this is the proportion of incident photons which liberate electrons and so contribute to the observation. This is extremely important for observing incredibly faint sources. OGLE-1's Ford-Loral 2048x2048 pixel CCD (Udalski et al., 1992) had $\sim 37\%$ peak efficiency whilst SDSS-I used an array of Tk2048E CCDs (Gunn et al., 1998) with closer to 80%[7].

Another key advantage is the digitization of the signal which makes the readout easy to process. CCDs are also ideally designed for continuous scanning observing strategies. In time-delay integration mode, the CCD charge is accumulated across the panel in-phase with the image using integrated clocks. This has been a key aspect of SDSS and, as we will see, is also a central feature of *Gaia*.

1.1.3 Escape from Earth

However big you make your telescope, there are several limitations with ground-based astrometry measurements which are difficult to avoid.

The atmosphere distorts apparent positions of sources on the sky. Think about looking at a fish swimming in a pond. If you're looking at the pond from the side, a fish will appear to be much closer to the surface than it actually is. This is because the change in medium from the air to the pond refracts light. For the fish peering up out of the water, we would appear to be higher up in the air than we actually are. The Earth's atmosphere causes similar distortions. We are like the fish, peering out of the atmosphere trying to estimate the positions of celestial bodies. The temperature of the Earth also varies daily and throughout the seasons. This causes the ground and any instruments to expand and contract which flexes mirrors and marginally alters the orientation of the entire observatory. We also have the issue that our field of view is limited. An observatory on the surface of the Earth can only observe a limited portion of the sky at any point in time making it more challenging to measure angles relative to reference stars.

[7] https://www.ing.iac.es/PR/wht_info/opticalccds.html

These problems can all be solved by using a satellite. In 1967, Pierre Lacroute proposed the *Hipparcos* mission (HIgh Precision PARallax COllecting Satellite, Høg, 2011). This was to be a continuously rotating satellite with two fields of view separated by a wide angle enabling simultaneous measurements of separate regions of the sky. The mission was launched in 1989 and proved to be incredibly successful producing measurements for 100,000 sources at better than milliarcsecond precision. This was a 50× improvement on the previous best astrometry catalogues (Perryman et al., 1997).

Shortly after *Hipparcos*, the Hubble Space Telescope (HST) was launched in 1990 with its impressive 2.4m mirror, an order of magnitude larger than the 29cm mirror on-board *Hipparcos*. Hubble is capable of reaching 0.1 arcsecond resolution (Burrows et al., 1991), however, Hubble only has a single field of view and is not designed to generate all-sky astrometric catalogues.

The next major leap in astrometry would come 24 years later with the launch of *Gaia*, a satellite capable of Hubble-like astrometry precision whilst scanning the entire sky (Lindegren & Perryman, 1996). I will come on to *Gaia* a bit later but for now I will introduce the two other core astronomy observations, photometry and spectroscopy.

1.1.4 Photometry

Photometry is the practice of measuring the apparent brightness of a star. Ptolomy developed the first recorded brightness measurement system placing stars in six magnitude categories measured by the time at which they ceased to be visible to the naked eye at twilight (Miles, 2007).

Pickering used the visual photometer, which enabled the observer to compare sources using a rotating Nicol prism which would attenuate polarised light according to Malus' law, to publish photometries for 9110 sources (Pickering, 1908). Over the course of the 19th century, brightness measurements were standardised leading to the definition of apparent magnitude.

Apparent Magnitude:

Apparent magnitude is the logarithm of the relative intensity of light from one source relative to a reference source

$$m = -2.5 \log_{10}\left(\frac{I}{I_{\mathrm{ref}}}\right) \tag{1.1}$$

(note that higher intensity means lower apparent magnitude). Intensity scales with inverse square of distance, therefore doubling the distance to a source will increase the apparent magnitude by $2.5 \log_{10}(4) = 1.5$.

Human eyes operate logarithmically such that repeatedly doubling a source brightness results in a linear change in the apparent brightness to us. Therefore the magnitude scale was developed to reflect this with apparent magnitude as the logarithm of source intensity. A scaling of 2.5 was introduced to fit measurements at the time which results in a similar magnitude scaling to Ptolemy's Almagest (Miles, 2007). In Fig. 1.1 I've estimated the visible apparent magnitude as a function of position on the sky by summing the flux of individual sources in pixels and transforming this to apparent magnitude. Human sight is limited to $G \lesssim 6.5$ so I've scaled these plots to be approximately what a human would see under perfect observing conditions, for example from the International Space Station. You can see that the Galactic plane becomes a single band of light for which it got the name 'Milky Way'.

Photometry is a vital ingredient of Galactic astronomy. A key example is Henrietta Leavitt's discovery of the period-luminosity relation for Cepheid variables, incredibly bright stars which radially pulsate (Leavitt & Pickering, 1912). This enabled significant developments in our understanding of the Galaxy including Harlow Shapley's measurement of the size and shape of the Milky Way (Shapley, 1918) and Edwin Hubble's resolution of 'The Great Debate' with the observation that 'spiral nebulae' are in fact distant galaxies with properties similar to our own (Hubble, 1925).

1.1.5 Spectroscopy

When we look at stars in the sky, most appear white although some may be bluer or redder than others.

In the mid 17th century, Newton discovered that apparently 'white' light from the Sun could be split into a rainbow of colour, a spectrum, by passing it through a prism[8]. Light from a star is made up of contributions from photons across the electromagnetic spectrum. The relative contributions from different parts will determine the star's colour.

- Hotter stars will have more flux at higher energies (shorter wavelengths) and so will appear bluer (e.g. Vega).
- Cooler stars will have a larger contribution from lower energies (longer wavelengths) and will appear redder (e.g. Betelgeuse).

Wollaston and Fraunhofer independently found that, buried within the continuum of light in the Solar spectrum, narrow regions were missing. Photons generated in the star pass through a cooler surface before escaping as starlight. Some photons are absorbed by atoms in this surface layer, exciting them to higher energy states. This occurs when the photon energy matches the atom's transition energy. The removal of these photons from the starlight leaves gaps in the spectrum characteristic of the atoms on the star's surface. In 1859, Kirchoff and Bunsen discovered the correspondence between these 'Absorption lines' and the 'Emission lines' found by heating up elements in the laboratory.

The depth of an absorption line (the amount of flux that is absorbed) tells us the abundance of the element responsible on the star's surface. This mapping of chemical abundances for stars throughout the Milky Way is hugely valuable for Galactic archaeology - studying the history of the Galaxy through the footprints of stars. However, mapping of chemical abundances is not the focus of my thesis. There is more we can learn from spectral lines.

[8] This is indeed exactly what a rainbow is, although, in the case of a rainbow the light passes through and internally reflects in the spherical water droplets of rain rather than a triangular prism.

When a fire engine is driving towards you on the road with its siren blaring, it will sound high pitched. As the fire engine passes, the pitch of the siren seems to drop. This is the due to the Doppler Effect and you could use this to measure the fire engine's speed (radar speed guns – often used by police for monitoring vehicle speeds – use the Doppler Effect in radio waves).

Doppler Effect:

Sound from the siren is emitted at a wavelenth λ_0 with speed c. Therefore the time between successive wavefronts is $\Delta t = \lambda_0/c$. In that time the fire engine moving at speed v has moved $\Delta x = v\Delta t$ away from us so it seems that the separation between consecutive wavefronts is longer $\lambda = \lambda_0 + \Delta x = \lambda_0 + v\lambda_0/c$ which produces the lower pitch. This is the Doppler Effect. The observed frequency of a wave emitted with frequency f_0 is

$$f = \frac{f_0}{\left(1 + \frac{v}{c}\right)} \tag{1.2}$$

in the observer's frame of reference. The same applies to photons emitted from stars where v is the relative velocity between the star and observer and c is the speed of light[9].

Absorption lines are be at characteristic wavelengths in the rest frame of the star which we can predict from the elements which have produced them. Therefore, we can measure the star's velocity relative to our own from the wavelength at which we observe the absorption lines

$$v = \left(\frac{\lambda - \lambda_0}{\lambda_0}\right) c. \tag{1.3}$$

If this is positive, the star is moving away from us and the light is red-shifted. If this is negative, the star is moving towards us and the light is blue shifted.

Throughout this chapter I have introduced *equatorial coordinates* providing 2D position on the sky, *parallax* providing the distance to source and *proper motion* providing 2D transverse motion of a star across the sky. Doppler shifting of absorption spectroscopy provides the source velocity in the *radial* direction. This is the last piece of the kinematic puzzle. With these six pieces of data we can measure the 3D position and 3D velocity of a star. I will use all of these six components in Chapter 2 when measuring the velocity distribution of stars in the Solar neighbourhood.

1.2 Modern Observatories

My research has focused on understanding our Galaxy using large catalogues of stars. In this section I introduce the key observatories which are the focus of my research.

[9] There are also relativistic corrections (see the relativistic Doppler Effect) however we don't discuss these here for brevity.

1.2.1 Multi-fibre Spectrographs

As I've just discussed, spectroscopy is incredibly valuable for studying stellar dynamics in the Milky Way, but how do we record stellar spectra in practice these days?

There are two challenges which make spectroscopy significantly harder than photometry and astrometry, which both rely on 2D imaging of the sky.

We measure flux from stars by effectively counting photons arriving on a CCD pixel. As we can project onto a 2D plane, this allows us to take 2D images, but spectra introduce a new dimension, wavelength, which can be projected into space using a prism for example. So the first problem is that we cannot simultaneously take 2D images and spectra in the same way as we do for photometry and astrometry. In general, photographic surveys are used to detect new sources on the sky which contribute to our map of the Galaxy. Spectrographs are then targeted at a subset of sources-of-interest from the photographic catalogue for follow-up.

Secondly, the precision with which we can measure a source's properties is dependent on the number of photons which are incident on a pixel. Fewer photons means larger 'shot noise'. By dispersing a star's light into a spectrum, we divide the photons up among far more bins and the signal per bin is much weaker. Therefore we're more limited to bright sources requiring longer exposures in order to get a significant signal.

The most common solution to these challenges is a multi-fibre spectrograph installed on a large telescope. The telescope observes a small region of the sky. Light from the aperture is projected onto the focal plane which contains an array of fibre heads strategically positioned on the image locations of selected sources. Photons pass along the fibre-optic cables and are dispersed into a spectrum which is projected onto a CCD panel.

The SDSS spectrograph (York et al., 2000), installed on the Apachee Point Observatory, New Mexico, included 'plug plates' – $\sim$ 1m diameter alluminium sheets with holes drilled at source locations – into which the fibre heads were inserted by hand by a team of technicians throughout the day. Each plate could hold up to 640 fibres and each night up to 9 of these plates would be used for their respective fields on the sky. Over the spectrograph's 14 year lifetime, thousands of plates were produced for SDSS observations.

By contrast 4MOST (de Jong et al., 2012), installed on the VISTA telescope in Chile's Atacama desert, employs 2436 fibres which can be robotically positioned in under two minutes with a precision of 0.2 arcseconds on the sky. 4MOST recently saw first light and, as one of many scientific goals, it will be tasked with collecting 18 million spectra of stars in the Milky Way to complement the *Gaia* mission.

Between these two extremes are a plethora of historical missions (e.g. SEGUE Yanny et al. 2009), ongoing projects (APOGEE Prieto et al. 2008, RAVE Steinmetz et al. 2006, LAMOST Zhao et al. 2012, Gaia-ESO Gilmore et al. 2012a, GALAH De Silva et al. 2015) and spectrographs soon to begin (WEAVE Dalton et al. 2012 , MOONS Cirasuolo et al. 2014).

In Chapter 2, I use the LAMOST DR4 value-added catalogue (Xiang et al., 2017) to help model the velocity distribution of stars in the Solar neighbourhood. Chapter 3 is devoted to developing a method for evaluating selection functions for any multi-fibre spectrograph. The vast majority of this thesis, however, is concentrated on data from the *Gaia* mission.

1.2.2 Gaia

The *Gaia* satellite was launched in 2013 and started operating seven months later (Gaia Collaboration et al., 2016). Since then it has been collecting data on two billion sources within and beyond the Milky Way resulting in the biggest leap forward in our understanding of the Galaxy since *Hipparcos*.

The *Gaia* mission was designed with several key scientific objectives:

- Determine the structure and dynamics of the Milky Way by modelling the distribution of stars throughout the Galaxy.
- Discover stellar companions (such as binaries and exoplanets) from the acceleration they induce in their host stars from the centre of mass of the system (e.g. Belokurov et al., 2020b; Penoyre et al., 2020).
- Constrain the star formation history of the Milky Way.
- Improve stellar models such as evolution tracks.
- Produce catalogues of variable stars and improve their luminosity calibration (e.g. Riello et al., 2018).

The mission has many other notable objectives such as Solar system astrophysics, extragalactic source distributions and constraints on fundamental physics parameters. Full details on *Gaia*'s objectives are provided in Perryman et al. 2001. My thesis is primarily focused on using properties of the *Gaia* satellite in order to enhance our models of the structure and kinematics of sources in the Milky Way. However, the tools I develop are also valuable for binary systems, stellar models and any projects which involve modelling distributions of sources.

The *Gaia* satellite started recording observations on 25th July 2014. Since then, three data releases have been published spanning observation windows up to 16th September 2015 (DR1 Brown et al., 2016), 23rd May 2016 (DR2 Gaia Collaboration et al., 2018a) and 28th May 2017 (DR3 Gaia Collaboration et al., 2021a). My thesis is based on both DR2 (which was published 6 months before I started at Cambridge) and EDR3 (the early part of the third data release, the rest of which will be published in 2022).

The Scanning Law

The *Gaia* satellite orbits the Earth-Sun second Lagrange point (L2) 1.5 million kilometers from Earth. It is composed of a 3m diameter cylinder containing the optical equipment mounted on a 10m Sun shield underneath which an antenna transmits data back to the ground stations on Earth. I provide a brief explanation of how the space craft operates focusing on aspects which are important to this thesis.

The satellite has three dominant components of rotation:

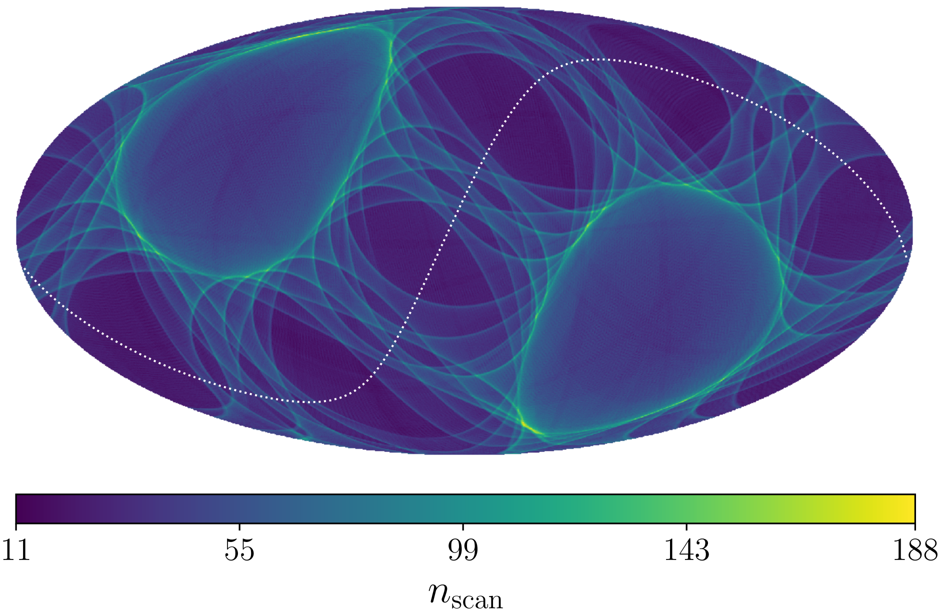


Fig. 1.3 The number of times *Gaia* scans each position on the sky produces a complex intricate pattern due to the spinning, precessing and orbiting motion of the satellite. Here we show the number of scans received in Galactic coordinates. The white dotted line shows the Ecliptic plane. *Gaia*'s orbit around the Solar system means that its motion is aligned in Ecliptic coordinates and the scanning law displays symmetries in this coordinate system.

- Spin: The spacecraft spins around its axis of symmetry on a 6 hour period.
- Precess: Throughout *Gaia*'s operational lifetime, the spin axis is always at a 45 degree Solar aspect angle (angle relative to the line connecting the Sun, Earth and L2). However, the spin axis precesses around the *Gaia*-Sun vector on a 63 day period. Imagine *Gaia* as a spinning top which has started to wobble. It is always at 45 deg to the table but the orientation precesses whilst still spinning at a much higher frequency.
- Orbit: *Gaia* remains within $\sim 340,000$km of L2 and as a result orbits the Solar sytem in phase with the Earth on a 365.25 day period. This means that *Gaia*-Sun vector around which the satellite precesses is continuously changing direction throughout the year.

For a full description of the satellite's position and dynamics see Section 5.5 of Gaia Collaboration et al. 2016.

On the spacecraft's cylinder are two windows referred to as the 'fields of view' (FoV) which look out approximately perpendicular to the spin axis and at a basic angle of 106.5 degrees from one another. This enables two vastly different regions of the sky to be simultaneously observed allowing a significant improvement in astrometric precision (the motivation behind this is detailed in Section 3 of Gaia Collaboration et al., 2016).

Due to the satellite's spin, the FoVs approximately follow one another with the preceding FoV (pFoV) scanning a position on the sky followed by the following FoV (fFoV) around 2 hours later then pFoV again 4 hours after that. Due to the precession, the scans don't exactly coincide but usually overlap.

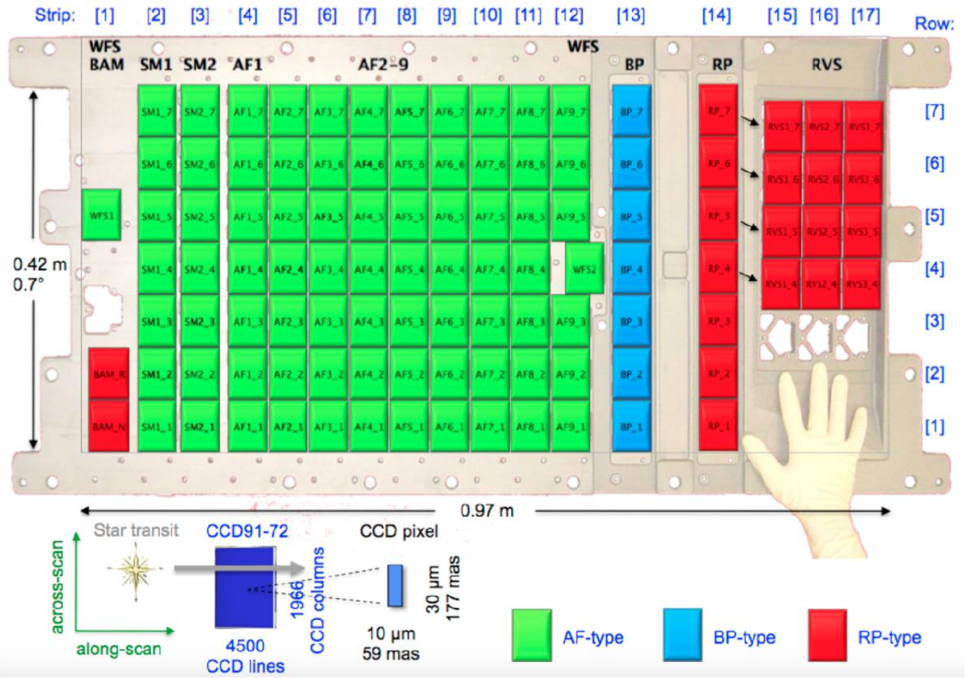


Fig. 1.4 (Fig. 4, Gaia Collaboration et al., 2016) The *Gaia* focal plane consists of an array of 106 CCDs. The majority of these are in the Astrometric Field (AF) and are used to measure the position and G-band apparent magnitude of sources. As *Gaia* spins, the telescope scans over the sky such that star images appear to travel across the focal plane from left to right crossing a CCD in 4.2 seconds. Each CCD takes its own independent source measurement such that, for example, three RVS spectra can be taken by the three columns of RVS CCDs on a single scan.

The primary mirror for each FoV is a 1.45×0.5m rectangle with both windows projecting onto a single panel of CCDs. The physical width of 0.5m corresponds to an angular with of 0.7 degrees on the sky (Crowley et al., 2016). The position on the sky being observed by *Gaia* as a function of time is the *scanning law*. In Fig. 1.3 I show the number of times each position of the sky was scanned by *Gaia* in the DR3 time frame which produces an intricate pattern on the sky. The times and directions of scans significantly affect whether a source is published in the *Gaia* catalogues (which I analyse in detail in Chapter 4) and the quality of the source's astrometry (Chapter 5).

Data Processing

The light from both FoV projects onto the *Gaia* focal plane, an array of 106 CCDs shown in Fig. 1.4 (Fig.4 of Gaia Collaboration et al., 2016). In this projection, stars enter the focal plane from the left side as the satellite spins and are tracked across the entire CCD panel.

I will briefly go through the main components and their importance.

- Sky Mapper (SM): The first two columns are the SM CCDs. Each SM column only takes light from one of the FoVs and is used to provide an initial source detection. All sources are measured in 2D and this is used to produce an initial G-band apparent magnitude estimate.

- Astrometric Field (AF): The following 7row × 9column panel of CCDs (minus the 4th row 9th column CCD) are all of the same type as the SM. The first column has a special assignment of confirming detections made by the SM. The AF CCDs all act in the same way, each providing an independent measurement of a star's position and apparent magnitude. This means that a single scan of a source can produce up to nine G-band apparent magnitude and position measurements.
- Blue and Red Photometers (BP & RP): The next two columns of CCDs provide apparent magnitude measurements on the blue (BP) and red (RP) side of the G-band. This is achieved through coatings on the CCDs and adjustments to the electrical properties such as resistivity. These CCDs actually record low resolution spectra but these won't be published until next year so we treat BP & RP as pure photometric instruments.
- Radial Velocity Spectrograph (RVS): the final 4row × 3column panel records spectra in the wavelength $847 - 874$nm capturing the Calcium triplet for most sources. Before being projected onto the CCDs, light passes through a slitless spectrograph composed of a filter and two prisms sandwiching a diffraction grating (Section 4 of Cropper et al., 2018).

There are also four CCDs, the Basic Angle Monitors (BAM) and Wavefront Sensors (WFS), monitoring the performance of the satellite.

CCDs work optimally in Time Delay Integration (TDI) which is ideal for *Gaia*'s observing strategy. As the image of a source travels across the CCD panel, the charge accumulated under the image is transferred along the pixels. When the source reaches the end of a CCD, the charge is accumulated and read out by the computer.

To avoid saturation (where pixels in the CCD panel reach their electron capacity whilst the source is still being observed) the satellite introduces a set of different observation types dependent on the apparent magnitude of the source measured by the SM CCDs. These are Window Classes and Gates.

Three Window Classes are used by the satellite (Table 2 Carrasco et al., 2016):

- WC0: $G \leq 13$ - 2D observation window
- WC1: $13 < G \leq 16$ - 18 pixel 1D observation window
- WC2: $16 < G$ - 12 pixel 1D observation window.

Gates introduce a potential barrier on the CCD. Charge accumulated to that point is drained away separately. Only charge accumulated after the barrier is used to generate the source observation. All gate activations occur at $G < 13$ with a separate set of configurations for AF, BP and RP. This is shown by the saw-tooth measurement error curve in Figure 14 of Riello et al. 2021.

As with any observatory, the *Gaia* satellite is not perfectly well behaved and a huge amount of work is required to analyse, correct and calibrate for systematic issues in the data, impressively performed by the Data Processing and Analysis Consortium (DPAC). Some challenges relate to the orientation (attitude) of the satellite which can be perturbed by micro clanks (structural changes in the spacecraft which can be generated by tempera-

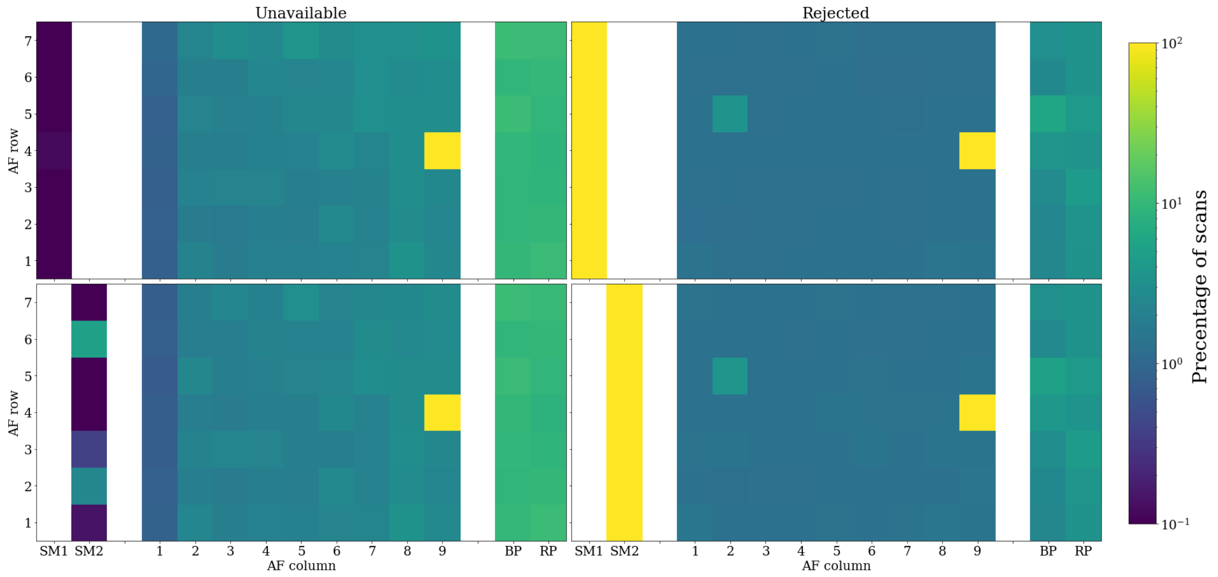


Fig. 1.5 The percentage of observations unavailable from the satellite (left) and rejected in data processing (right) for pFoV (top) and fFoV (bottom) across the focal plane where I've laid CCDs out in the same way as Fig. 1.4. The WFS CCD and SM1/2 CCD observations were not used for source modelling which is why these appear as 100% rejection. Even so, there is a reasonable amount of variation between the performance of CCDs particularly. AF2_5 clearly has some issues as a significant fraction of its data was rejected.

ture changes or mechanical features of the optical system) and micro-meteoroids which continuously bombard the satellite (Section 3.3 of Lindegren et al., 2018).

Decontamination and refocusing processes are also regularly required in order to keep the optical system at optimal performance. During these events and for a myriad of other reasons, there are periods of time when *Gaia* is unable to record usable data. I discuss these more in Chapters 4 and 5.

Some CCDs even perform worse than others. In Fig. 1.5 I show the fraction of observations in the two FoV which were not published in the epoch photometry (Holl et al., 2018) because they were not available from the spacecraft or they were rejected due to low quality in data processing. The variation in the astrometric field (AF), sky mappers (SM1/2) and BP/RP panels demonstrates that not all CCDs are created equal.

Where data is available, 2D observations (WC0) are fit with a calibrated point spread function (PSF) whilst 1D observations are fit with a line spread function (LSF). The fits provide parameters of the images which describe the source flux and precise position. This process, called Image Parameter Determination (IPD), is detailed in Rowell et al. 2021.

Finally a decision has to be made as to whether each source should be published in the output catalogue which is a vital component of Chapter 4. A source is included if it has a 2D position measurement from the astrometry pipeline, AGIS (Astrometric Global Iterative Solution Lindegren et al., 2012). For *Gaia* EDR3 AGIS determines whether to provide a 2D position measurement on two criteria[10] (Section 4.4 of Lindegren et al., 2021a):

[10] In DR2, an additional criterion, ASTROMETRIC_EXCESS_NOISE < 20mas, was used (Section 4.3 of Lindegren et al., 2018).

(i) ASTROMETRIC_MATCHED_TRANSITS ≥ 5
(ii) $\sigma_{\mathrm{pos,max}} < 100$ mas

A further cut is made on the closeness of neighbours. In DR2, if two sources are within 0.4 arcseconds of one another, only one is kept in the sample (this was lowered to 0.18 arcseconds in EDR3, Section 2.1 of Gaia Collaboration et al., 2021a). The choice of source is based on three priorities:

(i) Keep source included in *Gaia* reference frame.
(ii) Keep source with five parameter solution with smallest ASTROMETRIC_SIGMA5D_MAX.
(iii) Keep source with two parameter solution with smallest ASTROMETRIC_SIGMA5D_MAX.

A source with a two parameter astrometric solution only has position on the sky (α, δ) published in the *Gaia* catalogue while sources with five parameter solutions also have published parallax and proper motion $(\varpi, \mu_\alpha, \mu_\delta)$. ASTROMETRIC_SIGMA5D_MAX relates to the uncertainty of the astrometric fit which I discuss in much more detail in Chapter 5. Sources which had a nearby neighbour which was removed due to proximity are flagged with the DUPLICATED_SOURCE flag.

The result is a catalogue with 1,811,709,771 sources published in the *Gaia* EDR3 source catalogue (1,692,919,135 sources were published in DR2). Of the sources in *Gaia* EDR3 1,467,744,818 have published parallax and proper motion and 7,209,831 have measured radial velocities (although these are from DR2 data and the sample will be significantly expanded in the full DR3 in 2022). In Chapter 4 I provide a full description of the selection criteria used to select the subset of sources with astrometry and radial velocity. The method used to evaluate the astrometry of sources in *Gaia* is discussed and modelled in detail in Chapter 5.

1.3 The Milky Way from the Inside

The primary focus of this thesis is developing tools and techniques which enable the structure and kinematics of the Milky Way to be modelled with all available data without introducing biases. I provide a brief overview of our current understanding of the structure and evolution of the Milky Way and the observations which have enabled this.

1.3.1 In Theory

From models of the cosmic microwave background, we understand that energy content of the Universe may be decomposed into dark energy (69%), dark matter (26%) with only the final 5% as baryonic matter which makes up all that we can see (Planck Collaboration et al., 2020). It is through observations of baryonic content that we infer everything else.

As the Universe formed, regions collapsed under gravity to form nodes and filaments including halos on a spectrum of scales. This initiated the process of hierarchical structure formation where low mass halos collapse into high mass halos (Davis et al., 1985). This is a highly nonlinear process but is well modelled by N-body simulations such as the Millenium simulation suite (Springel et al., 2005) for which I've shown a snapshot in Fig. 1.6. This

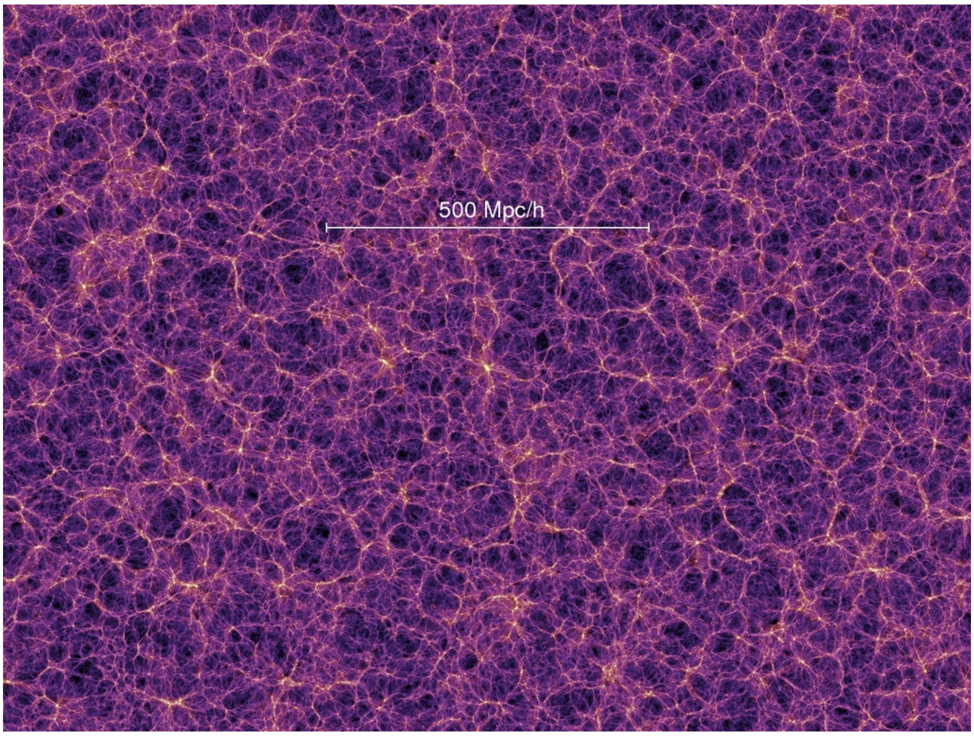


Fig. 1.6 A snapshot of the Millenium dark matter-only simulation (Springel et al., 2005) at $z = 0$ (present day equivalent) shows the node and filament structure of the Universe formed from gravitational collapse.

implies that the Milky Way would have been formed through a combination of initial gravitational collapse followed by a rich history of galaxy mergers (Bullock & Johnston, 2005).

Dark matter is unable to efficiently radiate away its energy and so maintains its structure in large spheroidal halos. The dust and gas in the baryonic content collapses further as particles collide and lose energy. By maintaining the total angular momentum, thin disks are formed in the cores of dark matter halos. It is from this cool, high density gas that stars are formed producing classical stellar galaxy disks like the Milky Way's.

Whilst gas and dust are highly collisional, stars are well approximated as collisionless particles. Dynamical processes between stars such as radial migration (Schönrich & Binney, 2009) mix the stellar populations. Whilst this is happening, satellite galaxies which have formed their own stellar populations continue to merge into the Milky Way halo. The gravitational effects of these accretion events also heats and thickens the disk (Kazantzidis et al., 2008; Laporte et al., 2019).

The result is a thin disk of stars which have formed inside the Milky Way, a thick disk which has been excited by dynamical interactions, a spheroidal halo of stars which have been accreted through galaxy mergers and a spheroidal dark matter halo extending to much larger distances.

So how do we work out the structure of the Milky Way?

1.3.2 Positions

The simplest way to model the structure of the Galaxy is by counting stars. The number density of stars as a function of 3D position throughout the Milky Way is a direct measure of the structure of the Galaxy at a snapshot in time.

This method has enabled us to model the vertical structure of the Milky Way disk (Ak et al., 2008; Bilir et al., 2006a,b; Jurić et al., 2008; de Jong et al., 2010). Picking out only young Cepheid variable stars demonstrates that the Galactic disk is not well constrained to a plane but is warped due to interactions with satellites such as the LMC (Skowron et al., 2019).

Simple maps of the sky have shown that the Milky Way halo is composed of an abundance of substructure as demonstrated by Belokurov et al. 2006's 'Field of Streams', consistent with the theoretical picture of hierarchical structure formation. Other models average out the substructure and consider the halo as a smooth distribution determining how the halo stellar content declines with distance from the Galactic centre (Fukushima et al., 2019; Mateu & Vivas, 2018).

The spatial structure of the Milky Way is discussed in more detail in Chapter 6 where I model the vertical tracer density of stars in the Galactic disk and halo using the tools I develop throughout the thesis with *Gaia* data.

1.3.3 Velocities

Whilst the spatial model describes a snapshot of the Milky Way, kinematic information describes how the stellar population dynamically evolves over time. Assuming that the Galaxy is in dynamical equilibrium, i.e. the position-velocity (phase space) distribution of sources doesn't change with time, we can relate this distribution to the gravitational potential and therefore to the mass distribution of the Milky Way.

Stars in the disk orbit the Milky Way centre (Kapteyn, 1922b) and the rotation curve can be used to determine the mass distribution of the Galaxy (e.g. Dehnen & Binney, 1998). Rotation curves of galaxies, including the Milky Way, provided some of the earliest evidence of dark matter (see de Swart et al., 2017, for full discussion).

The Collisionless Boltzmann Equation provides a more general relation between the position-velocity distribution function of stars and the gravitational potential they experience (Binney & Tremaine, 2008). Integrating over velocity components, this provides the Jeans equations (Jeans, 1922) which relate moments of the velocity distribution to the potential. This has been used to evaluate the potential throughout the halo (Wegg et al., 2019) and the local Solar neighbourhood (Nitschai et al., 2020). By subtracting off the mass contribution from local baryonic matter, the local density of dark matter can be estimated (Read, 2014; Silverwood et al., 2016; Sivertsson et al., 2018) with significant implications for direct detection experiments.

These dynamical models all assume steady state equilibrium (i.e. the phase-space distribution is constant as a function of time). However, within the Milky Way this turns out to be a poor assumption. The structure of stars in the Solar neighbourhood in velocity space is complex and scarred by dynamical perturbations (Dehnen, 1998; Kawata et al., 2018). Ripples in the velocity distribution of stars in the Solar neighbourhood describe an oscillating Milky Way disk (Antoja et al., 2018; Widrow et al., 2012).

As I've already mentioned, the halo is also not a smooth distribution of stars but a continually evolving mixture of merger remnants. A significant fraction of the halo stellar population is dominated by a single merger event which occurred some eight billion years ago (Belokurov et al., 2018; Helmi et al., 2018). This also splashed a significant amount of material out of the disk into the halo (Belokurov et al., 2020a).

The dynamical models discussed are heavily dependent on unbiased source distance estimates. In Chapter 2, I evaluate the velocity distribution of stars in the Solar neighbourhood demonstrating the importance of distance systematics in the process. The Jeans equations also require the tracer density of stars in the Milky Way which is the focus of Chapter 6.

1.4 Statistical Methodology

My thesis is all about the application of sophisticated statistical techniques to model large volumes of data. I briefly introduce some of the key concepts which come up regularly throughout the chapters. Some more specific statistical models and derivations are provided in Appendix A.

1.4.1 Statistical or Systematic Uncertainty?

A core theme of this thesis is the handling and minimization of systematic biases. In particular I draw attention to the use of parallaxes for distance estimation and modelling source distributions using incomplete catalogues. But what specifically does systematic uncertainty mean and why is it a problem?

Statistical Uncertainty:

Statistical uncertainty is the *random noise* in taking measurements. For example, in order to measure the position of a source, the *Gaia* collaboration is searching for the peak of the flux distribution on the CCD panel. However, each independent measurement finds a slightly different position depending on shot noise with photon counts in pixels and of course limited by the width of a pixel. The error distribution associated with statistical uncertainty is well understood and often quoted as the measurement error.

Systematic Uncertainty:

Systematic uncertainty includes all sources of error which aren't known statistical uncertainty. Since the *Gaia* mirrors are not perfectly achromatic, depending on the colour of a source, the position of the CCD image may shift. This introduces a systematic error which is unknown at the point of measuring the source position. An additional treatment is included the source astrometry fitting to account for chromatic effects on-board *Gaia*.

We typically refer to measurements with low *statistical* uncertainty as being *precise*. Measurements with both low *statistical* **and** *systematic* uncertainty are *accurate*.

One of the ways in which *Gaia* has propelled astronomy research forward is to provide an astrometry sample with unprecedented numbers of sources. This significantly reduces the statistical uncertainty making *Gaia* a survey with unprecedented precision. As a result, underlying systematic uncertainties become significant and these need to be modelled in order to make *Gaia* optimally accurate.

1.4.2 Bayes' Theorem

"If you did a rapid lateral flow test at a test site and the result was positive [...] get a PCR test to confirm your result as soon as possible."

NHS Advice, March 2021

The advice given by the NHS on lateral flow tests for COVID-19 may have led you to wonder how effective these tests really are. Studies have shown that they have a very low false positive rate so if I test positive, does this not mean it is highly likely that I have the virus? Not quite, it depends on the prevalence of the virus.

Lateral flow devices have a 99.5% specificity (0.5% chance of testing positive despite not being infected) and 72% sensitivity (28% chance of testing negative despite being infected) (Griffin, 2021). In May 2021, around 1/1000 people in the UK had the virus at any point in time $\mathrm{P}(C) = 0.1\%$[11]. Take a random group of 100 000 people in May 2021, on average 100 would have been infected. Of those 100, 72 would correctly test positive. Of the remaining 99 900, 500 would incorrectly test positive. If you've tested positive, there's a 72/572 = 13% chance that you're one of the people actually infected.

This brings me on to a simple piece of mathematics which can quantify your infection chances, Bayes' theorem.

The probability of two events, A *and* B, happening is the product of A given B times the probability of B:

$$\mathrm{P}(A, B) = \mathrm{P}(A \mid B)\,\mathrm{P}(B). \tag{1.4}$$

For example, the probability of Team GB coming fourth *and* collecting 65 medals in Tokyo2020 is the probability of coming fourth given that they collect 65 medals times the probability that they collect 65 medals.

The ordering of events can be switched without loss of generality so I can equally show that

$$\mathrm{P}(A, B) = \mathrm{P}(B \mid A)\,\mathrm{P}(A). \tag{1.5}$$

I can equate Eq. 1.4 and 1.5 and rearrange to get

$$\mathrm{P}(A \mid B) = \frac{\mathrm{P}(B \mid A)\,\mathrm{P}(A)}{\mathrm{P}(B)}. \tag{1.6}$$

[11] ONS COVID-19 Infection Surveys: https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/bulletins/coronaviruscovid19infectionsurveypilot/previousReleases

This is Bayes' theorem.

I want to know the probability that I had COVID (C) given that I received a positive test (✓) in May 2021. The probability of testing positive, P(✓) is the sum of the probability that I test positive *and* I have the virus (C) and the probability that I do not ($\overline{C}$)

$$\begin{aligned} \mathrm{P}(\checkmark) &= \mathrm{P}(\checkmark, C) + \mathrm{P}(\checkmark, \overline{C}) \\ &= \mathrm{P}(\checkmark \,|\, C)\, \mathrm{P}(C) + \mathrm{P}(\checkmark \,|\, \overline{C})\, \mathrm{P}(\overline{C}). \end{aligned} \tag{1.7}$$

The test sensitivity is $\mathrm{P}(\checkmark \,|\, C) \approx 72\%$. The false positive rate is $\mathrm{P}(\checkmark \,|\, \overline{C}) \approx 0.5\%$. I can use Bayes' theorem to work out the probability of having the virus

$$\begin{aligned} \mathrm{P}(C \,|\, \checkmark) &= \frac{\mathrm{P}(\checkmark \,|\, C)\, \mathrm{P}(C)}{\mathrm{P}(\checkmark)} \\ &= \frac{\mathrm{P}(\checkmark \,|\, C)\, \mathrm{P}(C)}{\mathrm{P}(\checkmark \,|\, C)\, \mathrm{P}(C) + \mathrm{P}(\checkmark \,|\, \overline{C})\, \mathrm{P}(\overline{C})} \\ &= \frac{0.72 \times 0.001}{0.72 \times 0.001 + 0.005 \times 0.999} \\ &= 0.13. \end{aligned} \tag{1.8}$$

There's only a 13% chance that I would have had the virus given a single positive lateral flow test in May 2021. This is despite the impressively low 0.5% false positive rate.

By the start of August, the virus prevalence was closer to 1.3% in the UK. By substituting this in to Eq. 1.8, I can work out that positive lateral flow test at that time would correspond to a 65% chance of having the virus. This is a huge shift in the probability despite the fact that the test I've used has not changed.

Bayes' theorem comes up several times in this thesis. Usually it is in the form

$$\mathrm{P}(\psi \,|\, d) = \frac{\mathrm{P}(d \,|\, \psi)\, \mathrm{P}(\psi)}{\mathrm{P}(d)}. \tag{1.9}$$

where d is the data and ψ are the parameters of my model. The four terms are:

- $\mathrm{P}(d \,|\, \psi)$: The *likelihood* (also written as $\mathcal{L}$) is the probability that I would measure the data given my hypothesis (or model). E.g. what is the probability that I test positive given that I have COVID?
- $\mathrm{P}(\psi)$: The *prior* provides our *a priori* knowledge of whether the model is probable or not. e.g. what is the probability that I have COVID?
- $\mathrm{P}(d)$: The *evidence* is the probability of obtaining my data, independent of the model. Since this is independent of the model parameters, ψ, when trying to determine the best model the evidence only acts as a normalisation constant which I can neglect.
- $\mathrm{P}(\psi \,|\, d)$: The *posterior* is what I want to know. This tells me how good my model is given the observed data.

An application of Bayes' theorem appears in the next section where I discuss a core concept of this thesis, selection functions.

1.4.3 Selection Functions

As observational astronomers, we use data from observatories such as the *Gaia* satellite (Gaia Collaboration et al., 2016) to test our physical theories about stars, galaxies and the Universe. However, to reliably test these theories we must understand our data and control for any significant systematics.

No survey is 100% complete. Despite containing 1.8 billion sources, *Gaia* is only capturing $\sim 1\%$ of the stars in the Milky Way. Why is this a problem?

To demonstrate the issue let's assume that the absolute magnitude distribution of stars in the Milky Way is Gaussian distributed with mean M_0 and variance $\sigma^2 = 1$ and we want to evaluate M_0. We measure the apparent magnitude of *all* stars to some brightness limit m_{lim}. Assuming we know the exact distance to these stars (which is often not the case) we could calculate the mean absolute magnitude $\langle M \rangle$ of our sample. If the Milky Way has a uniform density of stars our mean absolute magnitude would be biased by $\langle M \rangle - M_0 = 1.382$ mag. This is the Malmquist bias due to incompleteness of our survey (Malmquist, 1922). To put this in context, the apparent magnitude uncertainty of the faintest sources in *Gaia* EDR3 is only 0.01 mag (Riello et al., 2021), far smaller than the bias.

This is just one example of how survey incompleteness can bias measurements but will change depending on the nature of the sample limits (e.g. Lutz-Kelker bias related to removing sources with negative measured parallax Lutz & Kelker, 1973) or the type of measurement we're trying to make (e.g. the luminosity function, shape of the Milky Way, local density of stars etc.)

Understanding and accounting for incompleteness is critical for unbiased modelling of populations. The significance of this problem increases with increasing volumes of data. There are several ways in which a survey can be incomplete:

- *Apparent magnitude*: As per my example, sources are too dim to detect above background noise (e.g. atmospheric noise for ground based surveys, photon count noise, CCD readout noise...)
- *Colour*: Stars have different brightness in different wavebands according to their colour so observatories in different wavebands select differently in colour. Spectrographs also typically only have spectral templates in a limited effective temperature range (e.g. Sartoretti et al., 2018).
- *Position*: Telescopes have a limited observation angle on the sky and cannot observe anything outside this window. Since *Gaia* is continually scanning, this leads to a particularly complex dependence of completeness on sky position. This was also the case for *Hipparcos* and resulted in similar challenges when working with that catalogue (Hernandez et al., 2000).
- *Spatial resolution*: The ability of a telescope to resolve individual sources is limited by their spatial resolution. For a circular telescope, the minimum separation is $\sim 1.22\lambda/D$ where D is the diameter of the telescope collecting area.

- *Data rate*: For both instrumental and data processing reasons, most observatories have a limit on the rate at which they can take data. Multi-fibre spectrographs have limited numbers of fibres to observe stars with at a given time and the *Gaia* satellite has limited on-board data storage capacity.

Given that our source catalogues are not complete, we need to know what kinds of sources could have been observed and what could not. This is achieved through a selection function.

Selection Function:

The selection function of a catalogue is a probability which gives the chances of an object with known parameters $\mathbf{y}$ being included in the catalogue ($\mathcal{S}$)

$$\mathrm{P}(\mathcal{S}\,|\,\mathbf{y}). \tag{1.10}$$

For any catalogue of observed sources the selection function tells us "Given a real or hypothetical object, what is the probability that the object would have been successfully observed and included in the catalogue?". A detailed exposition of selection functions and their applications is provided in (Rix et al., 2021).

The choice of observable parameters, $\mathbf{y}$, needs to be made. This depends on our prior understanding of the observatory and how the catalogue has been constructed and the types of problems we want to use the selection function to solve. More practically, it also depends on the information we have available to model the selection function. I discuss parameter choices when producing selection functions in Chapters 3 and 4.

There are three core approaches to estimating the survey selection functions depending on the information available.

When the true distribution of sources is known, the selection function may be modelled by comparing the observed number of sources in the catalogue to the number expected. This true distribution may be theoretical (e.g. a simulation or mock Milky Way) or empirical (e.g. comparing to a larger catalogue which is known to be more complete). This is commonly used for spectrograph surveys by comparing with a larger photometric surveys (e.g. Chen et al., 2018; Das & Binney, 2016; Mints & Hekker, 2019; Nandakumar et al., 2017; Wojno et al., 2017) or assuming a true population (e.g. Schönrich et al., 2019).

The second method is when the observatory limitations are precisely known. It is extremely rare that the behaviour of an observatory is well enough understood that this may be done reliably for the entire population of observed stars. However, it is common to take a limited sample of the full catalogue which occupies a region of observable parameter space which is believed to be complete for the given observatory (Reid, 1982). Under certain assumptions, such as a Gaussian absolute magnitude distribution, the effects of selection biases can then be directly corrected (e.g. Reid & Gilmore, 1982).

The final method is a hybrid approach where information from the observatory is used to provide some information on the selection function such as magnitude limits however the selection probability is also empirically measured against a complete sample (e.g. Bovy et al., 2014; Stonkutė et al., 2016).

In Chapter 3 I present my method for generating selection functions for multi-fibre spectrographs which is a true distribution comparison approach complemented with known observatory field pointings. Chapter 4 includes the selection function I have developed for the full *Gaia* catalogue which is an observatory limitations based method along with the selection functions for subsamples of *Gaia* relative to the full *Gaia* catalogue.

Once we have evaluated the sample selection function, how do we use it? The aim of my thesis is to model the distribution of sources in the Milky Way and I focus on this application. To use the selection function to model a source distribution, one would be tempted to take observed distribution of sources and divide through by the selection function

$$\nu(\vec{x}) = \frac{\lambda(\vec{x})}{\mathcal{S}(\vec{x})} \tag{1.11}$$

where λ is the observed distribution, $\mathcal{S}$ the selection function and ν the inferred underlying distribution. However this breaks down whenever $\mathcal{S}$ is small. If the selection function in a particular region of parameter space is zero, we don't know what the true distribution of sources is as we have no empirical data to work with.

The more self-consistent method is a forward modelling approach. The algorithm would go as follows:

(i) Generate a model $\nu(\vec{x}\,|\,\vec{\phi})$ with model parameters $\vec{\phi}$.

(ii) Apply the selection function to the model to generate the observable distribution $\lambda(\vec{x}\,|\,\vec{\phi}) = \nu(\vec{x}\,|\,\vec{\phi})\mathcal{S}(\vec{x})$.

(iii) Compare the observable distribution to the observed data, e.g. in the form of a likelihood function.

(iv) Update model parameters based on their posterior probability evaluated from Bayes' theorem.

(v) Repeat steps 1-4 until the parameters converge on a solution.

In step (iv), Bayes' theorem is applied to generate the model posterior probability

$$\mathrm{P}(\vec{\phi}\,|\,d) = \frac{\mathrm{P}(d\,|\,\vec{\phi})\,\mathrm{P}(\vec{\phi})}{\mathrm{P}(d)} \tag{1.12}$$

where d represents the data, $\mathrm{P}(d\,|\,\vec{\phi})$ is the likelihood evaluated in step (iii). $\mathrm{P}(\vec{\phi})$ is our prior model and $\mathrm{P}(d)$ is the 'evidence' which provides the probability of measuring the given data for any model parameters and acts as a normalisation constant as it is independent of $\vec{\phi}$.

I apply this approach in Chapter 6 to estimate the vertical distribution of stars in the Milky Way using the selection functions I have produced for *Gaia* and its subsamples.

1.5 My Journey into the *Gaia*verse

Over the course of my PhD, I have produced selection function models and developed and applied methods to minimize the systematic uncertainty in fitting the distribution of stars in the Milky Way with data from *Gaia* and complementary spectrographs. The chapters of my thesis are composed of my submitted and published papers. For each chapter I give a brief outline here and reference the relevant paper(s).

Chapter 2: **The Tilt of the Local Velocity Ellipsoid.**

My journey is motivated by a case study in modelling the distribution of the velocity dispersions in the local ∼ 1 kpc of the Milky Way. I demonstrate how incorrect use of distances leads to artefacts in the velocity distribution which significantly change the inferred conclusions. Using distance distributions from Schönrich et al. 2019 I produce a new and improved model of the local velocity ellipsoid.

Everall, Evans, Belokurov & Schönrich 2019

Chapter 3: **Selection Functions for Spectroscopic Surveys of the Milky Way.**

Multifibre spectrographs, such as LAMOST, RAVE and APOGEE, are hugely complementary to the *Gaia* data with several more seeing first light in the next few years. I present my Bayesian Hierachical method to infer selection functions of Milky Way multi-fibre spectrographs supplemented with tools to combine multiple independent selection functions for unions of samples. I also demonstrate how to transform selection functions from observable coordinates ($\vec{q}$) to intrinsic coordinates using isochrones.

Everall & Das 2020

Chapter 4: **Completeness of the Gaiaverse.**

Gaia itself has a myriad of samples with different data and classifications. I introduce the full *Gaia* sample selection function which I've developed for DR2 in collaboration with Dr Douglas Boubert and generalised to EDR3. I then present the selection functions for *Gaia* EDR3 subsamples which have been constructed relative to the full *Gaia* catalogue.

Everall & Boubert 2021

Chapter 5: **The Astrometry Spread Function.**

The Astrometry Spread Function (ASF) provides the expected astrometric covariance for a simple point source observed by *Gaia*, a concept akin to point/line spread functions. I reconstruct the ASF of *Gaia* DR2 which reproduces the mean behaviour of published *Gaia* observations strikingly well.

Everall, Boubert, Koposov, Smith & Holl 2021c

Chapter 6: **The Photo-Astrometric Tracer Density of the Milky Way.**

I pull all components together to model the vertical distribution of Milky Way sources. The likelihood is regulated by the integral over observable coordinate space using the *Gaia* EDR3 astrometry selection function described in Chapter 4 and I marginalise over parallax uncertainty to avoid distance biases discussed in Chapter 2. Testing the method on a mock population sampled from the *Gaia* selection function with errors drawn from the ASF, I demonstrate unbiased inference of the input parameters for the mock model showing no significant systematic uncertainty.

Everall, Evans, Belokurov, Boubert & Grand 2021a
Everall, Belokurov, Evans, Boubert & Grand 2021b

2

The Tilt of the Velocity Ellipsoid

"I know of no more depressing thing in the whole domain of astronomy than to pass from the consideration of the accidental errors of our star places to that of their systematic errors."

Kapteyn 1922a

Professor Jacobus Kapteyn is referring to the dominance of statistical vs systematic errors. As a prominent researcher in a revolutionary era of astronomy, with the introduction of photographic measurements, Kapteyn witnessed and contributed to vast improvements in observational data. However, even at this time few stars had measured parallaxes and those that did were hugely uncertain. Any spatial and kinematic data of stars in the Milky Way would have been dominated by statistical uncertainties on distance.

Given his sentiments, Kapteyn may have been horrified to witness the *Gaia* era. In this Chapter I will use the case study of the tilt of the local velocity ellipsoid to demonstrate the significance of systematic uncertainties in parallax measurement with *Gaia* observations.

2.1 Spherical or Not?

Understanding the distribution of mass in the Milky Way is of great interest for constraining our Galaxy's formation history. Unfortunately, the majority of the mass does not emit detectable electromagnetic radiation and so we are forced to use indirect methods. One such method is to analyse the velocity dispersion of stars, as this is related to the Galactic potential through the Jeans equations.

The sample of 7 224 631 stars seen by the *Gaia* Radial Velocity Spectrometer (hereafter RVS, Gaia Collaboration et al., 2018a; Katz et al., 2019) provides a tempting dataset to study the behaviour of the velocity dispersion tensor. A recent attempt to do so was conducted by Hagen et al. (2019, henceforth H19). By augmenting the dataset with multiple spectroscopic surveys, including LAMOST Data Release 4 (DR4, Cui et al., 2012), APOGEE DR14 (Abolfathi et al., 2018) and RAVE DR5 (Kunder et al., 2017), H19 generated a sample of the Solar neighbourhood in excess of 8 million stars. They found that the velocity ellipsoids of their sample were close to spherically aligned within the Solar radius, but became cylindrically aligned at larger radii.

The results of H19 show comparable total misalignment to Binney et al. 2014 using RAVE DR5 (Kunder et al., 2017). Both studies find that the tilt of the ellipsoids of their thin disc dominated samples deviate significantly from spherical alignment in the Solar neighbourhood. The mismatch is significantly greater than found by Büdenbender et al. 2015 using SEGUE G dwarfs (Yanny et al., 2009). The disagreement is more striking when compared to the halo population. A number of studies using Sloan Digital Sky Survey data (Adelmam-McCarthy et al., 2008) found an almost spherically aligned velocity ellipsoid for halo stars (Bond et al., 2010; Evans et al., 2016; Smith et al., 2009b). This seems to be confirmed by the recent study of Wegg et al. 2019, who used a set of RR Lyrae extracted from *Gaia* Data Release 2 to conclude that the potential of the halo is spherical. This necessarily implies that the velocity ellipsoid is spherically aligned (An & Evans, 2016; Smith et al., 2009b). This is contrary to the results from H19, where the ellipsoid is cylindrically aligned at large distances from the Galactic centre and high above the plane.

Here, I analyse the behaviour of the local velocity ellipsoid using the *Gaia* RVS, complemented with LAMOST. I introduce the datasets in Section 2.2, paying careful attention to distance errors and biases. I provide my algorithm in Section 2.3 and present my results in Sections 2.4 and 2.5. I find that simple use of the reciprocal of parallax as a distance estimator is dangerous and can produce misleading results. The local velocity ellipsoid is always close to spherical alignment, and this remains true even for the thin disc and halo populations separately. The only substantial misalignment occurs for star samples at low latitudes and close to the Galactic centre, where the potential is strongly disc dominated.

2.2 Data

2.2.1 The *Gaia* DR2 RVS Sample

The *Gaia* DR2 RVS sample is a subset of the main DR2 catalogue with radial velocities derived from the on-board spectrograph (Gaia Collaboration et al., 2018a; Katz et al., 2019). Although this provides six-dimensional phase space data for over 7 million stars, the information on the distance is of course encoded as the parallax (an introductory discussion how to infer distances from *Gaia* parallaxes can be found in Bailer-Jones, 2015a; Luri et al., 2018). To recover the tilt of the velocity ellipsoid, special care needs to be taken with the inferred distances. To convert the proper motions into the tangential velocities requires the distance, and so poorly computed and noisy distances can overwhelm calculations of the tilt. We thus face two central problems: i) the parallaxes of *Gaia* can be biased, and ii) the method of inferring distances can be biased.

Concerning the parallax bias, Lindegren et al. 2018 used a sample of known quasars to determine a zero-point parallax offset of $\delta_\varpi = -29\,\mu$as, while they also showed that the parallax uncertainties are underestimated by about $\delta\sigma_\varpi = 43\,\mu$as, which are to be added in quadrature. The offset is known to depend on colour and apparent magnitude, might also depend on the object type and parallax, and hence is likely inappropriate for stellar objects in the RVS catalogue. More appropriate to the RVS catalogue, but still restricted to a particular subsets of stars, are a series of papers which found different parallax offsets: Riess et al. 2018 constrained $\delta_\varpi = -46 \pm 13\,\mu$as for Cepheids, whilst Xu et al. 2019 found a value of $-75 \pm 29\,\mu$as using VLBI astrometry of YSOs and pulsars. Zinn et al. 2019 and Khan et al. 2019 use asteroseismology for (mostly) red giants in the Kepler fields to get parallax offsets $\sim -50\,\mu$as, depending on the field position. For the full *Gaia* DR2 RVS sample, the parallax offset was calculated by Schönrich et al. (2019, henceforth S19) using their statistical distance method. They find an average parallax offset of $-54 \pm 0.06\,\mu$as, where the uncertainty comprises their systematic uncertainty with a negligible statistical error.

The literature contains in principle four approaches to infer stellar distances:

(i) Simply setting the distance $s = 1/\varpi$, as done by Hagen et al. 2019. This approach should only be used in situations where the parallax uncertainty is negligible for the problem, since it produces a three-fold bias: neglect of the selection function, neglect of the spatial distribution of stars, and ignorance of the fact that $1/\varpi$ is not the expectation value of the probability distribution function $\mathrm{P}(s)$. The latter bias was already identified by Strömberg 1927 and, along with the cut required to remove sources with negative measured parallax, became later well-known as the Lutz-Kelker bias (Lutz & Kelker, 1973).

(ii) Performing Bayesian distance estimates with a set of generic assumptions about the sample and the underlying Galactic density distribution, which eliminates the major problems of $s = 1/\varpi$, but leaves some uncertainties concerning the selection function. A good example of this approach is Bailer-Jones et al. 2018.

(iii) Doing a full Bayesian estimate involving stellar models, such as the Anders et al. 2019 distances.

(iv) Doing a full Bayesian approach with a self-informed prior that estimates the selection function from the data directly (Schönrich & Aumer, 2017; Schönrich et al., 2019).

Speaking generally, approaches (iii) and (iv) yield the most trustworthy results, though they involve greater expenditure of effort.

The mean bias between the different δ_ϖ estimates, and distance estimators is shown in Fig. 2.1. The S19 distances deviate from a simple parallax reciprocal $1/\varpi$ for distances beyond ~ 1 kpc. They also show substantially greater offset than would be accounted for by the $29\,\mu$as correction. I also note that the distance deviation is smaller than if one were to use the $54\,\mu$as offset and naively use $1/\varpi$. Fig. 2.1 underscores the point that the crude calculation of $1/\varpi$ overestimates the distance.

Tangential velocities are calculated by multiplying the proper motion by the distance whilst the spectroscopically determined radial velocities are independent of distance. If the true distance is underestimated (or overestimated), then so will be the tangential velocities. When inferring the velocity ellipsoid using spectroscopic radial velocities, this will tend to lead to heliocentrically aligned velocity ellipsoids, i.e., the velocity ellipsoids will become elongated (or compressed) towards the solar position. Fig. 2.1 demonstrates that using $s = 1/\varpi$ overestimates distances therefore will enhance the tangential velocities and cause the velocity ellipsoids to circularise around the Solar position. Notably, the result would be a flattening of the tilt of the velocity ellipsoids at the Solar radius as observed by H19.

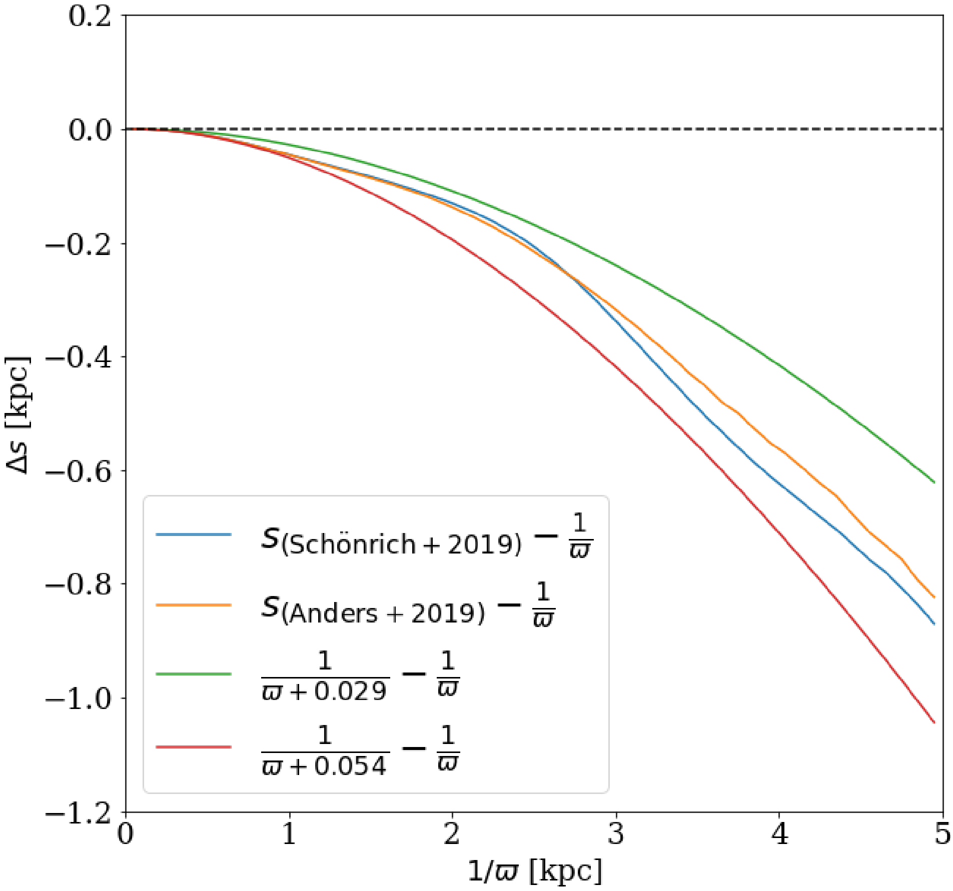


Fig. 2.1 Running median of the distance offset from a naive parallax reciprocal. The green curve is generated by corrections using the 29 μas parallax offset suggested by Lindegren et al. 2018, whilst the red curve uses the 54 μas parallax offset suggested by S19. Finally, the blue and orange curves show the difference between the parallax reciprocal and the Bayesian distance estimates from S19 and A19 respectively. Using the reciprocal of the parallax as a distance estimator is unwise beyond heliocentric distances $s \sim 1$ kpc.

I use the Bayesian distance estimates derived by S19 for my RVS sample. The data set includes corrected parallaxes and parallax uncertainties, which were also revised upwards by S19, and which I use to make quality cuts when applying to this data. Following common practice for parallax-based distance sets, I use $\varpi/\sigma_\varpi > 5$ [1]. I select only stars within 5 kpc of the Sun ($\varpi > 200\,\mu$as or $s < 5\,$kpc for the Bayesian distances). To remove spurious line-of-sight velocity outliers, I apply $\sigma_{v_r} < 20\,\mathrm{km\,s^{-1}}$ as well as $|v_r| < 500\,\mathrm{km\,s^{-1}}$ and further follow the recommendations of Boubert et al. 2019, which remove stars with less than 4 RVS transits and bright neighbours that can contaminate the measurements.

A concern with S19 distance estimates is that the kinematic model prior used to calculate the distances assumed a spherically aligned velocity ellipsoid. If this assumption was dominant in the distance inference, my results would be heavily biased towards finding a spherically aligned velocity ellipsoid. I address this concern in two ways. First, I compare S19 distance estimates with those found by Anders et al. (2019, henceforth A19) from photo-astrometric distances using the StarHorse pipeline (Queiroz et al., 2018). The potential biases between the S19 and A19 distances are very different. The latter set profits in precision from stellar model priors, while it may also inherit biases from the stellar models and have less well-defined distance uncertainties. These two datasets provide an excellent mutual control for remaining biases on either side. To correct for the parallax offset, A19 linearly interpolate as a function of G-band magnitude between the Lindegren et al. 2018 value of $29\,\mu$as at $G = 16.5$ and the $50\,\mu$as offset found by Zinn et al. 2019 at $G = 14$. I place the same cuts to the dataset using A19 distances as described earlier, but using a signal-to-noise cut of $s/\sigma_s > 5$ on heliocentric distance rather than parallax. The A19 distance estimates are also in Fig. 2.1. The estimates are similar to S19 within 3 kpc where the inference in both methods is dominated by parallax information with low uncertainties. Outside 3 kpc, the distance estimates of A19 are systematically larger by ~ 0.1 kpc. It is unclear where this disagreement originates from, however, I find it to be a small enough shift that my results are not significantly affected. In Section 2.4, I calculate the tilt for RVS data from StarHorse distances and find it to be consistent with that measured with S19.

Secondly, to truly quench any remaining uncertainty and to reinforce the use of $54\,\mu$as offset, I test the effect of the velocity ellipsoid correction terms on distance bias found in S19. This is shown in Fig. 2.2, where I plot the measured average distance bias versus distance for the S19 distances calculated with and without the parallax offset. The dashed lines show the "measured" distance bias, when I completely remove the velocity ellipsoid correction (which is equivalent to the wrong assumption that the velocity ellipsoid has a perfect cylindrical alignment).

Two things are obvious: i) Even with such a drastic error in assumptions, the change to the distance statistics is less than a third of the overall correction. As a result, the uncertainty in the velocity ellipsoid correction term is more than an order of magnitude smaller than the measured value of the parallax offset in S19. This is also reflected in the

[1] S19 helpfully provide a ϖ/σ_ϖ parameter with revised σ_ϖ which I use to cut on parallax signal-to-noise when applying their distance estimates.

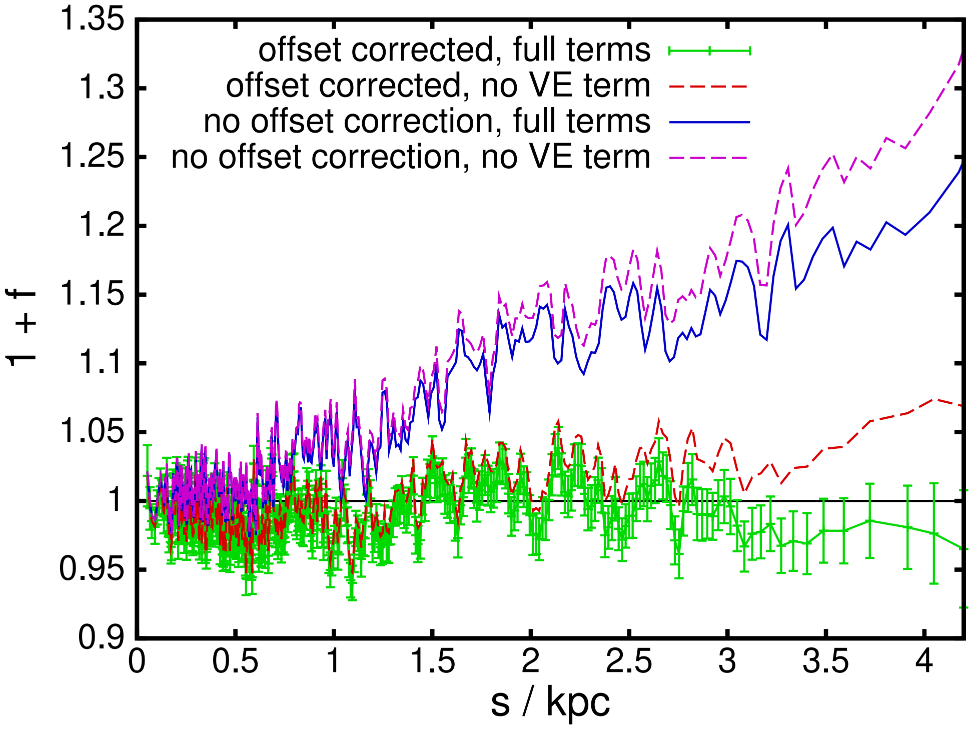


Fig. 2.2 A scan of the *Gaia* RVS for the fractional distance error $1 + f$ versus distance s with the quality cuts described in S19. Just as in S19, we move a mask of 12000 stars in steps of 4000 stars over the sample. The green error bars show the distance statistics after the distance correction, while the solid line shows the statistics when no parallax offset correction is applied. For both values of δ_{ϖ}, we show with dashed lines the same statistics when we completely remove the velocity ellipsoid correction term, which is equivalent to the wrong assumption that the velocity ellipsoid is cylindrically aligned. The resulting difference overestimates the actual uncertainty, but is still comparably small.

systematic uncertainty budget provided by S19. ii) When neglecting the velocity ellipsoid correction term, I actually require a **larger** correction for the parallax offset. As subsequent analysis will show, larger parallax offsets tend to flatten ellipsoids towards the Galactic centre and increase the tilt of ellipsoids around and outside the Solar radius. Hence this only strengthens my conclusion that the flattening of the tilt at the solar radius reported by H19 is driven by biased distance estimates.

2.2.2 The LAMOST DR4 and *Gaia* DR2 Cross-match

I separately analyse the velocity ellipsoids generated from the combination of 5D phase space information from *Gaia* DR2 (Gaia Collaboration et al., 2018a), together with radial velocities from the LAMOST DR4 value added catalogue (Cui et al., 2012; Xiang et al., 2017). This enables me to analyse the velocity ellipsoids with an independent catalogue of stars. LAMOST also provides metallicity estimates, which I use to produce halo and thin disc samples by cutting on [Fe/H] < −1.5 and [Fe/H] > −0.4 respectively, as done in H19.

I apply the same cuts to this dataset as for RVS, namely $\varpi/\sigma_{\varpi} > 5$, $\varpi > 200\,\mu$as, $\sigma_{v_r} < 20\,\mathrm{km\,s^{-1}}$ and $v_r < 500\,\mathrm{km\,s^{-1}}$. In the region of overlap between *Gaia* RVS and LAMOST, I use the radial velocity estimate with the least uncertainty.

I should be cautious of the radial velocities in LAMOST due to the statistical analysis performed by Schönrich & Aumer 2017. They determined that the LAMOST radial velocities were offset high by $\sim 5\,\mathrm{km\,s^{-1}}$. Assuming this offset is global throughout the dataset, it would shift my mean velocities without significantly impacting the velocity dispersions. Hence, I do not include this offset in my analysis.

2.3 Method

To transform from heliocentric to Galactocentric coordinates, I need to fix some Galactic constants. I assume a Solar position [2] in cylindrical polar coordinates of $(R_\odot, z_\odot) = (8.27, 0.014)$ kpc (e.g., Binney et al., 1997). The circular velocity of the Local Standard of Rest is taken as $v_c(R_\odot) = 238\,\mathrm{km\,s^{-1}}$ (Schönrich, 2012), whilst the Solar peculiar motion is $(U_\odot, V_\odot, W_\odot) = (11.1, 12.24, 7.25)\,\mathrm{km\,s^{-1}}$ (Schönrich et al., 2010).

I determine the velocity ellipsoid parameters using maximum likelihood estimation on the bivariate Gaussian likelihood function convolved with Gaussian measurement uncertainties similar to previous works (e.g., Bond et al., 2010; Evans et al., 2016, H19). I resolve the velocities into Galactocentric spherical polar coordinates (v_r, v_θ, v_ϕ) and use a likelihood function

$$\log \mathcal{L} = -\frac{1}{2}\log|2\pi\Lambda| - \frac{1}{2}\sum_i (\boldsymbol{x}_i - \boldsymbol{\mu})^{\mathrm{T}}\Lambda^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}). \tag{2.1}$$

Here, $\boldsymbol{x}_i = (v_{r.i}, v_{\theta,i})$ are the velocity components of the i^{th} star, and $\Lambda = \Sigma + \boldsymbol{C}$, where Σ is the velocity covariance matrix in (v_r, v_θ) and $\boldsymbol{C}$ the measurement uncertainty covariance matrix of the data. The data are binned in a 20×20 grid of Galactocentric cylindrical polar coordinates (R, z), such that each bin is approximately 500×500 pc. For every bin, I analytically calculate the means and covariances of the contained populations without measurement uncertainties. These parameters are used to initialise my likelihood optimization in order to calculate a best fit model with the uncertainties. The algorithm proceeds by optimizing the means and covariances for each bin independently.

For the measurement errors in the RVS sample, I take the standard deviation and correlation parameters for parallaxes, radial velocities and proper motions from the *Gaia* DR2 dataset. The challenge here is that my likelihood function is inherently Gaussian, whilst, assuming parallax uncertainties are Gaussian, the distance uncertainty distribution is inherently non-Gaussian. When using $1/\varpi$ as my distance estimator, the parallax uncertainty is propagated so I do not assume Gaussian distance uncertainties. However, I do assume Gaussian velocity uncertainties when calculating the likelihood function. When using distance estimates from S19, it is important to use the correct uncertainty distribution. For the purposes of this work, I assume Gaussian distance uncertainties using the second moment of distance given by S19 as the variance. For future work, it will be important to understand the impact of the third and fourth moments of distance on my velocity ellipsoids. I also assume here that the distance is uncorrelated with the remaining

[2] The effect of changing the Solar position is investigated in Section 2.5.1

astrometric parameters. For the LAMOST cross-matched with *Gaia* sample, I assume that radial velocities are uncorrelated with all the *Gaia* astrometric parameters.

I determine the parameter posteriors by using the MCMC python package `emcee` (Foreman-Mackey et al., 2013). I find that initialising walkers in a small ball around my analytically determined parameters allows the chains to converge within 50 iterations. I run 20 walkers for 300 iterations and use the last 150 to calculate my posteriors.

2.4 Results

2.4.1 The *Gaia* DR2 RVS sample

For my analysis of the *Gaia* RVS sample, I compute the velocity ellipsoids for three different assumptions to show the effects of distance errors:

(i) without any parallax correction and using $s = 1/\varpi$,
(ii) with a parallax correction of 29 μas and using $s = 1/\varpi$,
(iii) with the Bayesian distance estimates from S19, which use a parallax correction of 54 μas.

my total sample sizes after applying cuts are 5 375 902, 5 499 054, and 5 221 912 respectively. The velocity ellipsoids produced using assumptions (i) and (ii) are shown in Fig. 2.3, whilst those produced using (iii) are given greater prominence in Fig. 2.4. I only show ellipsoids in bins with greater than 30 stars, as these still provide clean results and allow us to view the distribution out to greater distances.

In the top panel of Fig. 2.3, I recover Figure 2 of H19. I see the same transition from approximate spherical to cylindrical alignment across the Solar radius. I note that my results are somewhat more noisy, since I have not augmented my data-set with spectroscopic catalogues and so my sample is about 75% of the size of H19. This effect is consistent with overestimates of the distances, and hence tangential velocities, as already discussed in Section 2.1. The bottom panel of Fig. 2.3 shows the same results with a 29 μas correction. The behaviour of the velocity ellipsoid is now much more consistent throughout the meridional plane, without the awkward transition from spherical to cylindrical alignment at the Solar circle. However, of course this correction is conservative and not physically motivated for stars in the RVS sample.

Fig. 2.4 uses the Bayesian distance estimates from S19 and is the centrepiece of my results. I note that the ellipsoids do not extend out as far as in the previous plots. The reason for this is that S19 also revise the parallax uncertainty upwards. As a consequence, when cutting on parallax uncertainty $\varpi/\sigma_\varpi > 5$, I remove more stars, particularly at large distances. Those bins which are no longer included do not contain a requisite number of stars for us to plot the ellipsoids. I do observe a slight deviation of the spherical alignment of the velocity ellipsoids at low elevation towards inner radii, tending to cylindrical alignment. This is likely the effect of the contribution of the baryonic disc to the gravitational potential. The same effect can be seen in the velocity ellipsoids of RR Lyrae in the halo in Wegg

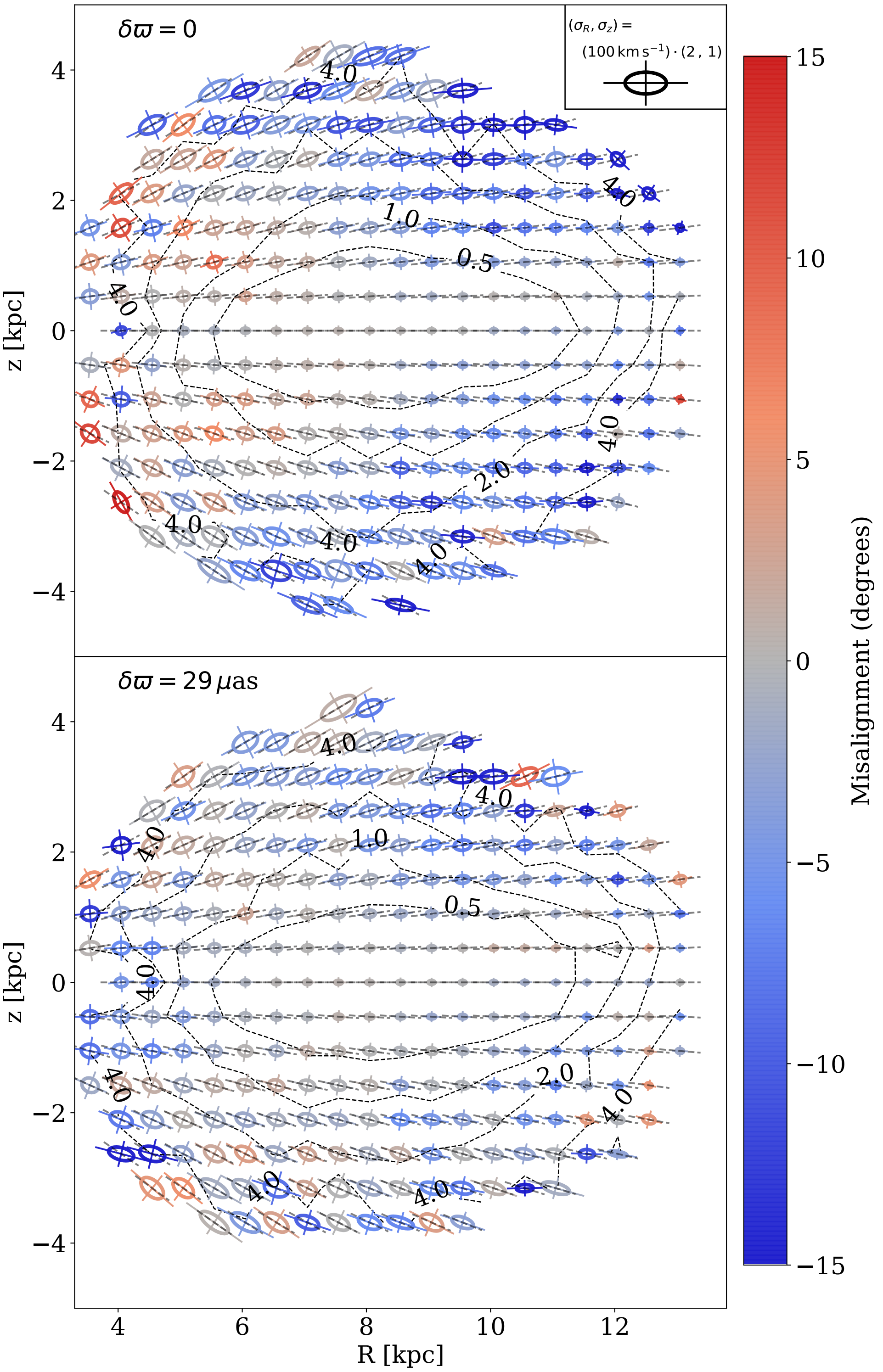


Fig. 2.3 Velocity ellipsoids generated from the *Gaia* RVS DR2 dataset with different treatments of parallax bias. The size of the ellipsoid is proportional to the value of the velocity dispersion in each bin. The short-dashed lines correspond to the orientation of a spherically aligned velocity dispersion, while the colour bar gives the deviation in degrees of the velocity ellipsoid orientation from this spherical alignment, with blue indicating a flattening and red an over-tilting towards the disc. The black dashed lines show contours of misalignment uncertainty. Top: Using distance as $1/\varpi$ with no parallax offset correction. Bottom: Distance as $1/\varpi$ with $29\,\mu$as parallax correction.

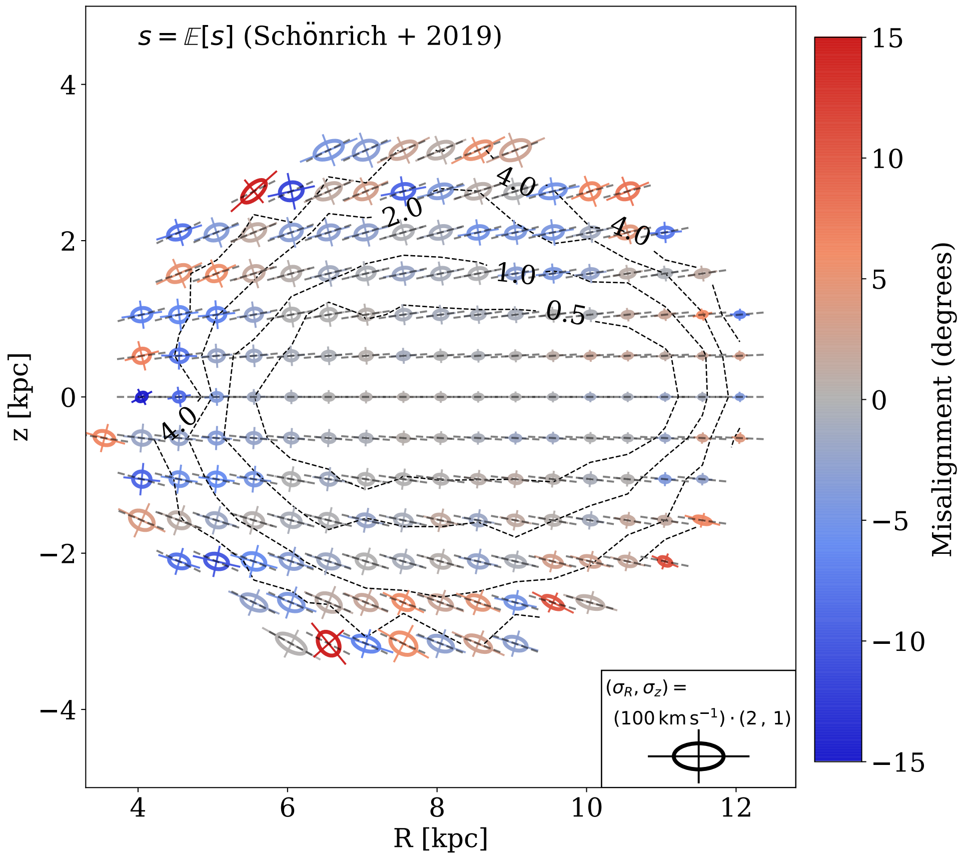


Fig. 2.4 Velocity ellipsoids generated from the *Gaia* RVS DR2 with Bayesian distance estimates from S19 which include a parallax offset correction of 54 μas. This figure can be compared to Fig. 2.3 which make inferior assumptions as to the distance estimates. Black dashed contours give the ellipsoid orientation uncertainty for 0.5°, 1°, 2° and 4° respectively. Note that the artificial transition from spherical to cylindrical alignment at the Solar circle visible in the upper panel of Fig. 2.3 has been removed.

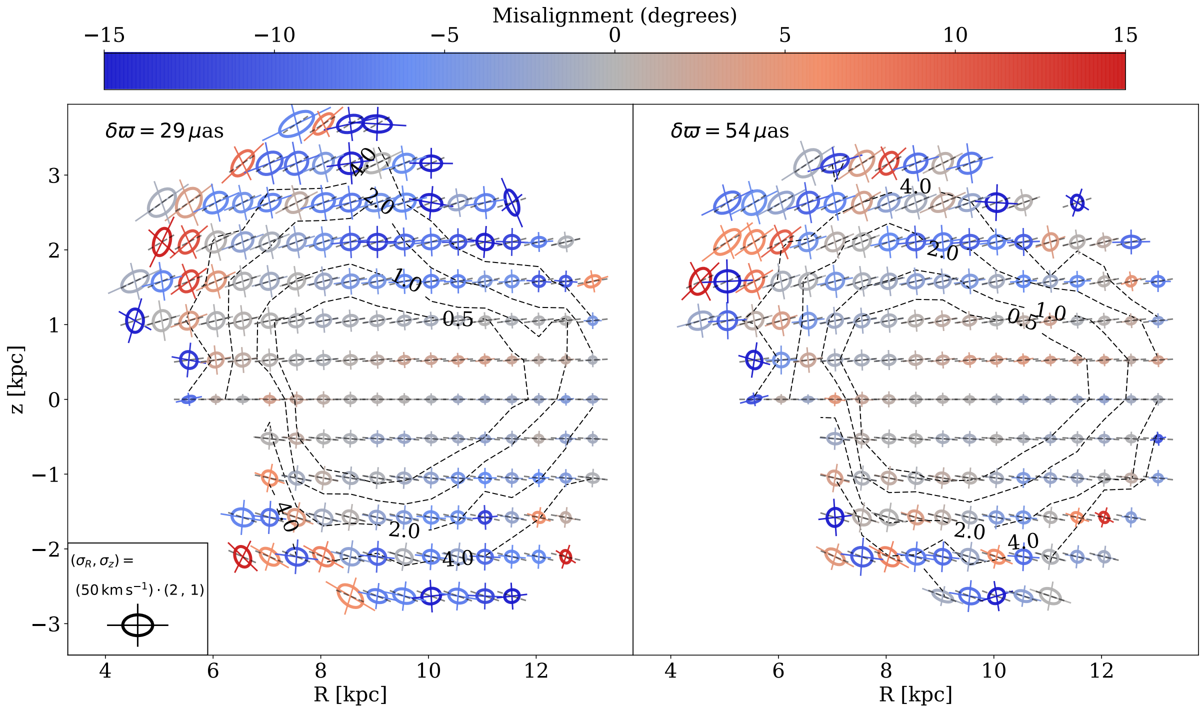


Fig. 2.5 Velocity ellipsoids generated from *Gaia* DR2 cross-matched with LAMOST with [Fe/H] > −0.4 producing a thin disc sample. As usual, the size of the ellipse is related to the value of the velocity dispersion in the given spatial bin. The short dashed lines correspond to the direction of spherical alignment. The colour corresponds to the deviation in degrees of the velocity ellipsoid orientation from spherical alignment. In other words, grey implies spherical alignment whilst blue implies tending towards cylindrical alignment. The black dashed contour shows the misalignment uncertainty. I use $1/\varpi$ as a distance estimator but with 29 μas correction (left panel) and 54 μas correction (right panel). These bracket the range of possibilities, as the former overestimates and the later underestimates the true distances.

et al. 2019, although most of the effect in their analysis occurs within 4 kpc of the Galactic centre, outside of which the velocity ellipsoids appear to be spherically aligned.

Notice that the size of the velocity ellipsoids increases with elevation above and below the plane. This is caused by the inclusion of three populations of stars, belonging to the thin disc, thick disc and halo. It is interesting to look at the populations separately, and for this I turn to the LAMOST and *Gaia* cross-matched sample, which has spectroscopic metallicities.

2.4.2 The LAMOST DR4 and *Gaia* DR2 Crossmatch

Without Bayesian distances for this sample, I use $s = 1/\varpi$ as my estimator with parallax corrections 29μas and 54μas. I expect these to overestimate and underestimate distances respectively, as indicated by Fig. 2.1. Therefore, my results on the tilt of the velocity ellipsoid merely bracket the range of possibilities.

I split the sample into two separate populations, [Fe/H] > −0.4 as a thin disc sample and [Fe/H] < −1.5 as a halo sample. Neither sample is completely pure, as the metallicity cuts only approximately separate populations. After applying the cuts, my halo samples

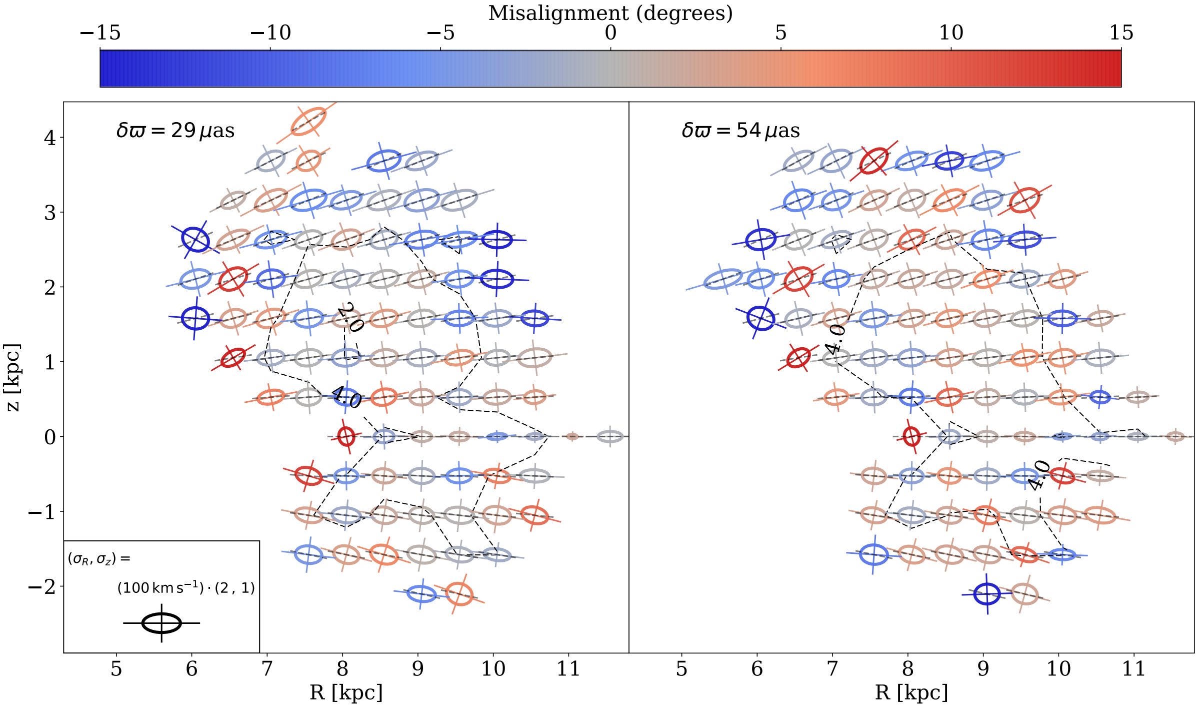


Fig. 2.6 As Fig. 2.5, but for the halo sample obtained from *Gaia* DR2 cross-matched with LAMOST with [Fe/H] < −1.5.

contain 18 424 and 19 661 stars for 29 μas and 54 μas corrections respectively and the thin disc samples contain 2 286 528 and 2 306 729 stars.

In Fig. 2.5, I present results for the thin disc sample. In the left plot, the flattening of the tilt is still strong for the 29 μas correction, with cylindrical alignment particularly prevalent at elevations above 2 kpc from the plane. In the right plot, with a 54 μas correction, the majority of this signal has been removed. However, there appears to be a small but significant deviation from spherical alignment remaining for heights $|z| \sim 2.5$ kpc. It is suggestive that there the thin disc population may not be exactly spherically aligned.

The results for the low metallicity halo sample are given in Fig. 2.6. This contains a much smaller number of stars, which allows fewer bins and causes the results to appear more noisy. However, in the left plot, with the conservative 29 μas correction, almost cylindrical alignment can be seen for $R \sim 10$ kpc and $z \sim 2$ kpc which is completely removed in the right hand plot for the 54 μas over-correction. I also note here that the scales of the velocity dispersions are much more consistent across elevations which demonstrates the effect of selection of the halo sample with only small impurities.

2.5 The Tilt of the Velocity Ellipsoid

Binney et al. 2014 and Büdenbender et al. 2015 introduced and exploited a compact way to summarize results on the tilt of the velocity ellipsoids. They used a model in which the angle between the Galactic plane and the direction of the longest axis of the velocity ellipsoid is

$$\alpha = \alpha_0 \arctan |z|/R. \tag{2.2}$$

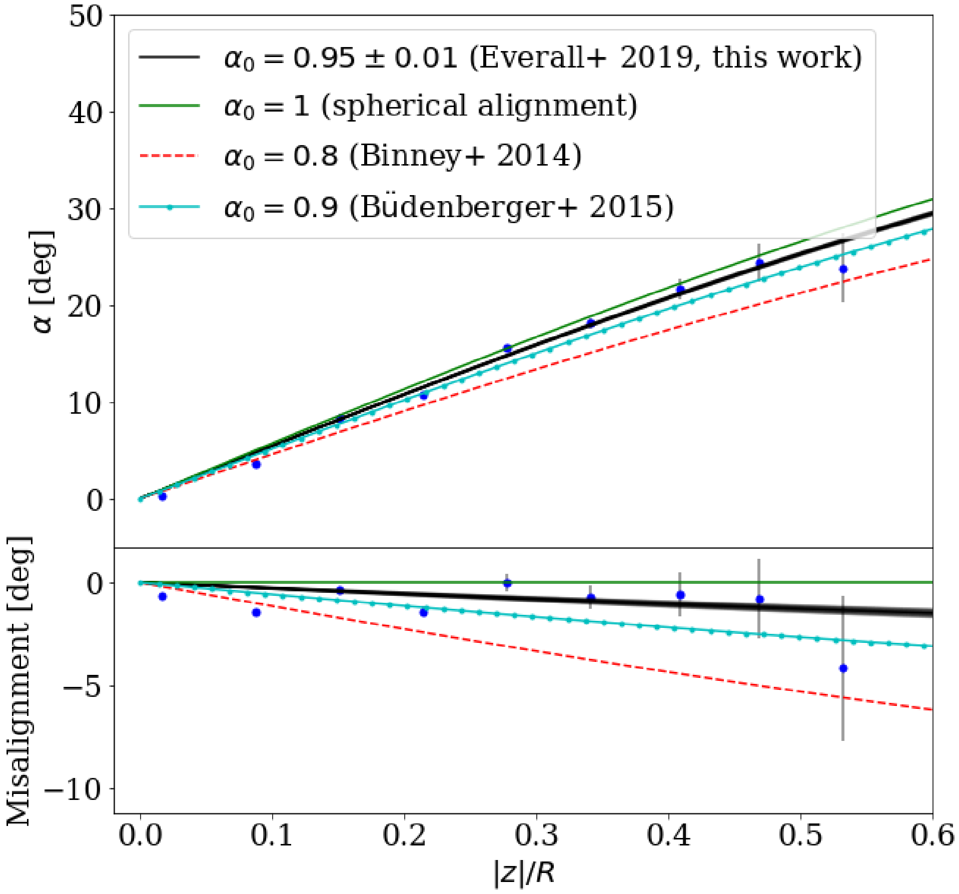


Fig. 2.7 The upper panel shows fits for the tilt of the velocity ellipsoid using Eq. (2.2). The blue points provide the posterior means and uncertainties of ellipsoids aggregated in in $|z|/R$ bins as evaluated in Eq. 2.3. Perfect spherical alignment corresponds to the green line, whereas the black line is my result from the *Gaia* RVS sample with distances from Schönrich et al. 2019. For comparison, I also show recent fits from Binney et al. 2014 (red) and Büdenbender et al. 2015 (pale blue). Notice that the binned datapoints show a transition from below to above the best fit line, as the disc potential becomes less dominant. The lower panel shows the deviation from spherical alignment.

They fitted the binned data to the model to determine the best fit α_0 parameter. A result of $\alpha_0 = 1$ implies exact spherical alignment, whilst $\alpha_0 < 1$ means that the ellipsoids are tilted towards cylindrical alignment.

I perform a least squares regression on all bins with $n_{\text{stars}} > 5$ as these still contain valuable information about ellipsoid alignment although with large uncertainties[3]. Bins with fewer stars are almost randomly aligned. For the *Gaia* RVS sample with S19 distances, I acquire a tilt value of $\alpha_0 = 0.952 \pm 0.007$. This is in significant disagreement with $\alpha_0 \sim 0.8$ determined in Binney et al. 2014 from the local RAVE stars (Steinmetz et al., 2006). It is in reasonable agreement with Büdenbender et al. 2015, who found a value of 0.90 ± 0.04 using the Segue G dwarf sample. As discussed in Section 2.2, I also calculate this parameter for the distance estimates of A19 with the RVS sample and retrieve $\alpha_0 = 0.956 \pm 0.006$, in remarkably good agreement with the estimate from S19 distances.

I see no physical reason why this parameter should be constant across all populations of stars and in all parts of the Galaxy. Under the hypothesis that tilt of the velocity ellipsoids is controlled at least in part by the contribution of the baryonic disc to the potential, I anticipate that α_0 should be lowest near the plane and tend towards 1 at high elevation. I also suggest that the flattening of the tilt should be more extreme in the inner radii. To test this hypothesis, I compute α_0 for subsets of my velocity ellipsoids. I find that for $|z| < 2\,\text{kpc}$, $\alpha_0 = 0.950 \pm 0.007$ whilst for $|z| > 2\,\text{kpc}$, $\alpha_0 = 0.966 \pm 0.018$. I also find that at $R < 7\,\text{kpc}$, $\alpha_0 = 0.917 \pm 0.013$ whilst for $R > 7\,\text{kpc}$, $\alpha_0 = 0.963 \pm 0.007$. This is consistent with the hypothesis that the effects of the disc potential are driving much of the deviation from spherical alignment.

I also look at the tilt at large radii and high elevation. For $|z| > 2\,\text{kpc}$ and $R > 7\,\text{kpc}$, I retrieve the result $\alpha_0 = 0.986 \pm 0.020$, which is consistent with spherical alignment. This is in good agreement with a number of studies of the velocity ellipsoids of halo stars in SDSS (Bond et al., 2010; Evans et al., 2016; Smith et al., 2009b), as well as the recent work of Wegg et al. 2019 who determined that the kinematics of the RR Lyrae in the halo, extracted from *Gaia* DR2, imply a spherically symmetric halo potential.

In Fig. 2.7, I show the fit of the tilt of the velocity ellipsoids as a function of $|z|/R$. The green solid line shows the expected trend for spherical alignment ($\alpha_0 = 1$). I plot my best fit, as well as the earlier results from Binney et al. 2014 and Büdenbender et al. 2015. I also group my velocity ellipsoids in $|z|/R$ bins, where each bin is a cone annulus viewed from the Galactic center, and estimate the posterior tilt of all ellipsoids within that bin. I do this by assuming Gaussian uncertainties on individual ellipsoid inclinations in which case the the posterior uncertainty and mean are given by

$$\sigma_\alpha = \sqrt{\frac{1}{\sum_i \sigma_{\alpha i}^{-2}}}, \qquad \mu_\alpha = \sigma_\alpha^{-2} \sum_i \alpha_i \sigma_{\alpha i}^2 \tag{2.3}$$

[3] In Section 2.4, I only use bins with $n_{\text{stars}} > 30$ because the scatter in less populated bins make the ellipsoid plots appear untidy and muddied the trends in behaviour.

where the sums are over all ellipsoids in a given bin and $\alpha_i, \sigma_{\alpha i}$ are the mean and uncertainty of tilt measured for ellipsoid i. These means, μ_α, and standard deviations, σ_α are given by the blue data points and error bars in Fig. 2.7. Notice that the binned datapoints show an interesting pattern with respect to the best fit. The datapoints with high $|z|$ mostly lie just above the best fit, those with low $|z|$ lie just below. This trend suggests that the deviations from spherical alignment are induced by the disc potential.

I also compare ellipsoids above and below the plane. I find that above the disc $\alpha_0 = 0.964 \pm 0.009$, whilst below the plane, $\alpha_0 = 0.940 \pm 0.009$, showing 2σ disagreement. However, this asymmetry is far more stark when separating in-plane from high elevation contributions. Considering only ellipsoids within 1 kpc of the plane, I find that $\alpha_0 = 0.989 \pm 0.014$ above and 0.888 ± 0.013 below which has a 5σ difference. Conversely outside 1kpc, $\alpha_0 = 0.94 \pm 0.01$ and 0.99 ± 0.01 above and below respectively, in 3σ disagreement and opposite to the in-plane difference.

For an axisymmetric equilibrium that is reflexion symmetric about the Galactic plane, results above and below the plane should be consistent. This apparent discrepancy particularly in the disc may be caused by substructure and streams, buckling of the Galactic bar (Saha et al., 2013), or by the effects of bending modes in the disc (e.g. Gómez et al., 2013; Laporte et al., 2019; Williams et al., 2013; Xu et al., 2015), or by unrecognized systematics in the data.

I analyse the thin disc and halo samples generated from the *Gaia*-LAMOST cross-match. For the disc sample, I recover $\alpha = 0.909 \pm 0.008$ for the 29 μas correction, which becomes $\alpha = 1.038 \pm 0.008$ for the 54 μas correction. As anticipated, this straddles the RVS results demonstrating the effect of overestimating and underestimating the distances. The same effect is present in the halo sample with $\alpha = 0.927 \pm 0.035$ and $\alpha = 1.063 \pm 0.036$ for corrections of 29 μas and 54 μas respectively.

2.5.1 The Solar Position

In the analysis, I assumed a Solar distance to the Galactic centre of $R_\odot = 8.27$ (Binney et al., 1997) and neglected uncertainties on this estimate. This is mainly to ease comparison with earlier work, especially H19. Recently, the Gravity Collaboration et al. 2018 reported a high precision distance to Sagittarius A* of 8.127 ± 0.031 kpc, which is smaller than my assumed value.

Adjusting the Solar position with respect to the Galactic centre does not change the properties of velocity ellipsoids in Cartesian coordinates. The only impact is that I now calculate the misalignment with respect to a new central point in the Galaxy.

For this change in $R_\odot$, the shift in misalignment is small. In the most extreme cases of velocity ellipsoids at ($|z| \sim 2$, $R \sim 4$) kpc, the misalignment is reduced by 0.84° which falls well within my uncertainties. On average, across all my ellipsoid positions, the induced flattening is 0.33°. The effect on any individual ellipsoid is negligible.

However, a change in $R_\odot$ induces a coherent shift in all ellipsoid misalignments, and so there is a somewhat larger effect on my inference of the tilt normalization parameter, α_0. I find that using $R_\odot = 8.127$ kpc, the full RVS sample generates a tilt parameter of

$\alpha_0 = 0.953\pm0.007$. This shift is still within the original uncertainties. Similar calculations for sub-samples of the ellipsoids prove even less significant due to their increased uncertainties.

2.6 Round Up

The tilt of the velocity ellipsoid of local stars is important for several reasons. First, determinations of the local dark matter density are usually based on the vertical kinematics of stars. The gravitational potential is inferred from the Jeans equations or distribution functions, a calculation known to be sensitive to the tilt of the velocity ellipsoid (e.g. Silverwood et al., 2016; Sivertsson et al., 2018). Secondly, the heating processes that thicken discs include scattering by in-plane spiral arms and by giant molecular clouds. These scattering processes can produce different signatures in the tilt of the thin disc velocity ellipsoid (e.g., Sellwood, 2014). Thirdly, the alignment can give direct information on the potential in some instances (e.g., Binney & McMillan, 2011; Eddington, 1915). For example, the halo stars are believed to be close to spherical alignment, as judged by a number of earlier studies of SDSS star samples (e.g., Bond et al., 2010). Exact spherical alignment implies a spherically symmetric force field (An & Evans, 2016; Smith et al., 2009b).

The *Gaia* Radial Velocity Spectrometer (RVS) sample comprises 7 224 631 stars with full phase space coordinates. The main hurdle to overcome in exploiting this dataset to study the tilt is the accurate and unbiased conversion of parallaxes ϖ to heliocentric distances s. I find that the Bayesian distances of Schönrich et al. 2019, which incorporate a parallax offset of 54 μas, give reliable results. I have checked that substitution of photo-astrometric distances from Anders et al. 2019 using the StarHorse pipeline gives consistent results.

The *Gaia* RVS sample is consistent with nearly spherical alignment. The tilt is accurately described by the relation $\alpha = (0.952 \pm 0.007)\arctan(|z|/R)$. If the normalising constant were unity, then this would imply exact alignment with spherical polars. My result is pleasingly close to that found by Büdenbender et al. 2015 from the Segue G dwarf stars in the Solar neighbourhood. If the sample is restricted to stars at large Galactocentric radii, or great distances above or below the plane, then the alignment becomes still closer to spherical. The data support the conjecture that any deviation from spherical alignment of the velocity ellipsoids is caused by the gravitational potential of the disc. Such deviations occur at low $|z|$ and close to the Galactic centre, whilst at $|z| > 2$ kpc and $R > 7$ kpc the ellipsoids are consistent with spherical alignment.

With subsamples from *Gaia* DR2 cross-matched with LAMOST, I study the disc and halo populations separately. Even though Bayesian distances are not available for all these stars, I can bracket the tilt of the velocity ellipsoids by making assumptions that underestimate and overestimate the heliocentric distances. For thin disc stars, I find $\alpha = (0.909-1.038)\arctan(|z|/R)$ and for halo stars $\alpha = (0.927-1.063)\arctan(|z|/R)$. Both populations are close to spherical alignment, with the only real deviations occurring in the inner Galaxy near the Galactic plane.

2.7 Systematic Dominance

The use of reciprocal parallax as a distance estimator produces artefacts in the local velocity ellipsoid which can completely change the conclusions which one draws from the data. In this instance, I have demonstrated that the dominant effect was the zero point parallax offset in the *Gaia* data.

However, I have also discussed the importance of Bayesian distance estimates and propagation of non-Gaussian uncertainties. Evaluating Bayesian distances require a selection function for the given catalogue which can be used to estimate the observable distribution of sources as a prior. In Chapters 3 and 4 I introduce methods for evaluating selection functions of several different surveys and produce the selection functions for the *Gaia* source catalogue and scientific subsets.

3

Spectrograph Survey Selection Functions

"One who knows and knows that he knows...
...his horse of wisdom will reach the skies.
One who knows, but doesn't know that he knows...
...he is fast asleep, so you should wake him up!
One who doesn't know, but knows that he doesn't know...
...his limping mule will eventually get him home.
One who doesn't know and doesn't know that he doesn't know...
...he will be eternally lost in his hopeless oblivion!"
Ibn Yamin, Persian poet, c. 14thC

This quote, controversially popularised by Donald Rumsfeld in 2002, highlights a challenge we face in astronomy. We are well practiced at discussing topics which we known we know, and a huge amount of our research effort goes into transforming known unknowns to known knowns. A much more challenging task is discovering the unknown unknowns. For example, until 2007 the transient events referred to as fast radio bursts (FRBs) were not known to exist. They were unknown unknowns, an observable which we didn't know we did not understand because there had been no confirmed detection. Observations of the Lorimer burst (Lorimer et al., 2007) transformed these transient objects into known unknowns, objects which we knew we didn't understand. Since then vast effort has gone into understanding the causes of these events.

Selection functions are hugely important for differentiating between what we know or do not know from our observations. You are a football fan at the 2020 Euros final at Wembley watching England play Italy. You want to estimate the number of people in the stadium. This is easy in your block, you just turn around and count individuals. As you start to look further away though it gets tough, on the other side of the stadium people are all blurred together and with some standing whilst others sit, there are individuals who are barely visible. The solution is simple, you count the people in your own block and assume all other blocks have a similar number so you multiply by the number of blocks.

In this scenario, you observed all people in your block so the selection function here is 1, these are the known knowns. In all other blocks you have not counted any individuals so the selection function is 0, the known unknowns. Because you understand your selection function, you can achieve a reasonable estimate about the number of fans in the stadium.

However, what if you had just tried to count as many people as possible without keeping track of how much of the stand you were trying to count? Suddenly you do not know your selection function and you only have a count for some unknown fraction of the stadium. There are parts of the stadium where you did not count anyone but you do not know if that's because there truely is no-one there or because you were not able to count that stand, there are unknown unknowns and you are lost in hopeless oblivion!

In this chapter, I demonstrate a method for estimating selection functions of multi-fibre spectrographs. As I show in Chapter 6, selection functions are invaluable for estimating the population and distribution of stars in the Milky Way, just as a football fan estimates the number of spectators at Wembley.

3.1 Spectroscopic Surveys

Within the Milky Way, we are able to resolve stars out to large distances, but with that privilege comes a bias. The most comprehensive spectroscopic surveys can typically measure quantities for only a fraction of a percent of the stars in the Milky Way. The survey or 'observed' selection function is the probability of a star being included in the survey given its sky positions, colour and apparent magnitude. To make inferences regarding the intrinsic chemodynamical structure of the Milky Way, knowledge is required of the intrinsic selection function, i.e. the probability of a star being included in the survey given its intrinsic coordinates: Galactic location, metallicity, mass, and age.

The second data release (Gaia Collaboration et al., 2018a) of the ESA's *Gaia* mission has pushed us far further than ever before, measuring photometry, positions, and proper motions for over 1.3 billion stars. In parallel, several ground-based spectroscopic surveys are measuring spectra for millions of these stars. Despite an often relatively simple nominal selection in colour and apparent magnitude, taking cross-matches with other surveys or selecting stellar type sub-samples result in a selection function that is no longer simple. Furthermore, to convert the observed colour-magnitude selection function into an intrinsic one that depends on distance, metallicity, mass, and age, one needs to engage with stellar isochrones. In order to understand the bias generated by selecting subsamples, we must calculate the selection functions for these surveys.

Many studies have developed survey selection functions, in which the completeness along a line of sight is given by the ratio of the number of stars in the spectroscopic survey to the number of stars in a photometric survey in the same region of colour and apparent magnitude. Das & Binney (2016) and Das et al. (2016) use this method to construct selection functions for halo blue horizontal branch stars and K giants in Sloan Extension for Galactic Understanding and Exploration-2 (SEGUE-II, Xue et al., 2011). Similar methods are used in determining the selection function for the Radial Velocity Experiment (RAVE, Kordopatis et al., 2013) by Wojno et al. 2017 and for the Large Sky Area Multi-Object Fibre Spectroscopic Telescope (LAMOST, Cui et al., 2012) Spectroscopic Survey of the Galactic Anticentre (LSS-GAC, Xiang et al., 2017) by Chen et al. 2018. Stonkutė et al.

2016 perform a similar analysis of the *Gaia*-ESO survey (Gilmore et al., 2012a), taking into account the survey's observing strategy.

Vickers & Smith 2018 generate a combined selection function for LAMOST (Zhao et al., 2012), RAVE (Kunder et al., 2017) and the Tycho-Gaia Astrometric Solution (TGAS, Michalik et al., 2015a) by binning the fields in colour, apparent magnitude, and distance using a synthetic galaxy catalogue as the assumed complete sample. Bovy et al. 2012b and Bovy et al. 2014 consider the dependence of the selection function on apparent magnitude for G-type dwarfs in the Sloan Extension for Galactic Understanding and Exploration (SEGUE, Yanny et al., 2009) survey and the Apache Point Observatory Galaxy Evolution Experiment (APOGEE, Majewski et al., 2017). Their selection function along any line of sight is assumed to be uniform in apparent magnitude with limits defined either by the faintest star observed in a field or by the survey's nominal magnitude limit. Bovy et al. 2014 also present a selection function in distance that takes into account the dust extinction along the line of sight.

Finally, Nandakumar et al. 2017 and Mints & Hekker 2019 determine selection biases for a large number of spectroscopic surveys by binning the sample in colour-magnitude space, either using a regular grid or a specialised median binning algorithm, and comparing with the 2-Micron All Sky Survey (2MASS, Skrutskie et al., 2006).

Amongst all previous works, fundamental aspects of the methodology remain the same. The dependence of the selection function on colour and magnitude is treated as a ratio of number counts of stars between the spectroscopic and a given photometric survey or a synthetically generated population, which is assumed to be complete. Uncertainties in measurements in colour and apparent magnitude are not used in the calculation. The effects of overlapping coordinate fields on the selection function are not considered. Selection functions are largely constructed as a function of colour and magnitude and not converted to intrinsic coordinates, with the exception of Das & Binney (2016), Das et al. (2016), and Sanders & Binney (2015). With the exception of Vickers & Smith 2018, no methods are presented for combining selection functions.

In this chapter, I build on previous work to create `seestar`[1], a Python code that can be applied to any multi-fibre spectroscopic survey, independent of its footprint on the sky, the number of stars observed, and the selection criteria of the survey. I construct an algorithm to treat the limitations of Poisson noise when bins have small numbers of stars by using a Poisson point process, whose parameters I determine using maximum likelihood estimation. I propose how the uncertainties in colour and magnitude measurements may be incorporated into this. I include a method for calculating the union of overlapping field probabilities, which also enables selection functions of independent surveys to be combined. I use isochrones to convert the selection function depending on colour and apparent magnitude to one that depends on distance, metallicity, mass, and age. This is an essential component of chemodynamical models of the Milky Way (Das & Binney, 2016; Sanders & Binney, 2015; Schönrich & Binney, 2009).

[1] https://github.com/aeverall/seestar

In Section 3.2, I demonstrate how to calculate the selection function in observable coordinates (colour and apparent magnitude) and intrinsic coordinates (distance, metallicity, mass, and age), and how to determine the dependence on sky position. The results of tests on a mock catalogue are presented in Section 3.3. Finally, I discuss my results and potential future developments and applications of this method in Section 3.4. Section 3.2 is rather technical in nature, and I therefore advise readers only interested in the performance of the selection function package to skip to the results.

3.2 Method

The selection function $\mathrm{P}(\mathcal{S} \mid \mathbf{x})$ of a stellar survey is the probability of a star being in the survey, $\mathcal{S}$, given the star's coordinates $\mathbf{x}$. The coordinates may be observed (sky positions, apparent magnitude, and colour) or intrinsic quantities (galactic location, metallicity, mass, and age). I assume that for the spectroscopic survey, there is a photometrically selected catalogue of stars which is complete in the region of colour and apparent magnitude observable by the spectrograph. Throughout this chapter, I use the superscript 'spec' to refer to the spectroscopic sample and 'phot' to refer to the photometric sample. By Bayes' theorem

$$\mathrm{P}(\mathcal{S} \mid \mathbf{x}) = \frac{\mathrm{P}(\mathbf{x} \mid \mathcal{S})\mathrm{P}(\mathcal{S})}{\mathrm{P}(\mathbf{x})}. \tag{3.1}$$

$\mathrm{P}(\mathbf{x})\mathrm{d}^n x \equiv f(\mathbf{x})\mathrm{d}^n x$ is the probability that a star chosen at random has coordinates within the volume element, $\mathrm{d}^n x$, where $f(\mathbf{x})$ is the distribution function of stars in the Milky Way. $\mathrm{P}(\mathbf{x} \mid \mathcal{S})\mathrm{d}^n x \equiv f^{\mathrm{spec}}(\mathbf{x} \mid \mathcal{S})\mathrm{d}^n x$ is the probability that a star picked at random from the survey has coordinates in $\mathrm{d}^n x$, where $f^{\mathrm{spec}}(\mathbf{x})$ is the distribution function of observed stars. Finally, $\mathrm{P}(\mathcal{S}) = \mathbb{E}\left[N^{\mathrm{spec}}\right] / \mathbb{E}\left[N\right]$ is the probability of a star entering the survey and is given by the ratio of the expected number of stars in the survey to the expected total number of stars in the Milky Way. Under the assumption that the photometric sample is complete within the given range of observable coordinates, I denote $f^{\mathrm{phot}}(\mathbf{x}) = f(\mathbf{x})$ and $N^{\mathrm{phot}} = N$ and substitute into Equation (3.1).

$$\mathrm{P}(\mathcal{S} \mid \mathbf{x}) = \frac{f^{\mathrm{spec}}(\mathbf{x})\mathbb{E}\left[N^{\mathrm{spec}}\right]}{f^{\mathrm{phot}}(\mathbf{x})\mathbb{E}\left[N^{\mathrm{phot}}\right]} = \frac{\mathbb{E}\left[n^{\mathrm{spec}}(\mathbf{x})\right]}{\mathbb{E}\left[n^{\mathrm{phot}}(\mathbf{x})\right]}, \tag{3.2}$$

where $\mathbb{E}\left[n^{\mathrm{spec}}(\mathbf{x})\right]$ and $\mathbb{E}\left[n^{\mathrm{phot}}(\mathbf{x})\right]$ are the expected number densities of stars in the spectroscopic and photometric surveys respectively, at coordinates $\mathbf{x}$.

The observation of a star is dependent on the star's observable properties, $\mathbf{x} = (l, b, c, m)$ where (l, b) are the star's coordinates on the sky and (c, m) the colour and apparent magnitude of the star. For ease of notation I group these into positional coordinates, $\boldsymbol{\theta} = (l, b)$ and photometric properties, $\mathbf{v} = (c, m)$.

The positional coordinates indicate which region of the sky or 'patch' the star belongs to. The best method to characterize the dependence of the selection function on these patches depends on the survey design. This is described in Section 3.2.2. On each patch, the selection function is calculated as a function of $\mathbf{v}$ as described in Section 3.2.3.

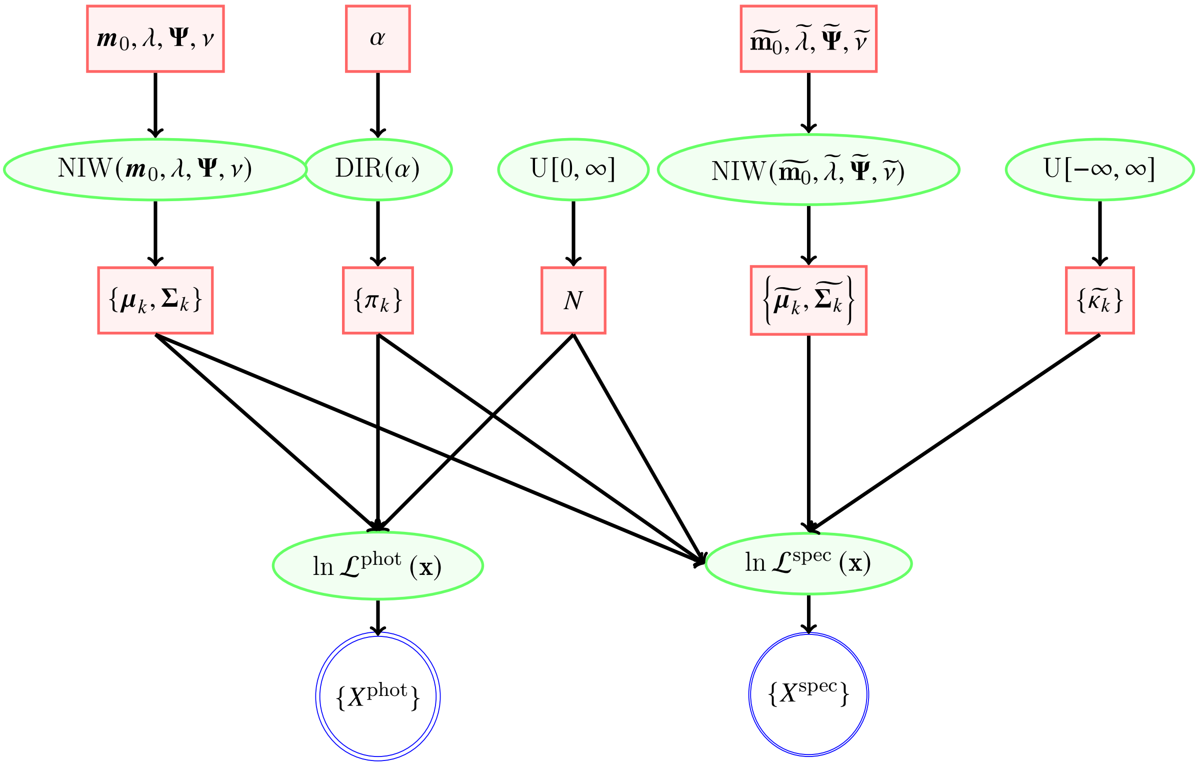

$\boldsymbol{m}_0, \lambda, \boldsymbol{\Psi}, \nu$
α
$\widetilde{\mathbf{m}_0}, \widetilde{\lambda}, \widetilde{\boldsymbol{\Psi}}, \widetilde{\nu}$
$\mathrm{NIW}(\boldsymbol{m}_0, \lambda, \boldsymbol{\Psi}, \nu)$
$\mathrm{DIR}(\alpha)$
$\mathrm{U}[0, \infty]$
$\mathrm{NIW}(\widetilde{\mathbf{m}_0}, \widetilde{\lambda}, \widetilde{\boldsymbol{\Psi}}, \widetilde{\nu})$
$\mathrm{U}[-\infty, \infty]$
$\{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}$
$\{\pi_k\}$
N
$\left\{\widetilde{\boldsymbol{\mu}_k}, \widetilde{\boldsymbol{\Sigma}_k}\right\}$
$\{\widetilde{\kappa_k}\}$
$\ln \mathcal{L}^{\mathrm{phot}}(\mathbf{x})$
$\ln \mathcal{L}^{\mathrm{spec}}(\mathbf{x})$
$\{X^{\mathrm{phot}}\}$
$\{X^{\mathrm{spec}}\}$

Fig. 3.1 Bayesian network (directed acyclic graph) of the model described in this chapter as a function of colour and apparent magnitude for a single field. The network describes a method for determining the posterior of photometric density and selection function parameters where both are parameterised GMMs.

In Section 3.2.4, the observed coordinates are transformed into intrinsic coordinates, $\mathbf{v}(s, [\mathrm{M/H}], \mathcal{M}_{\mathrm{ini}}, \tau)$, where $s, [\mathrm{M/H}], \mathcal{M}_{\mathrm{ini}}, \tau$ are distance, metallicity, initial mass, and age respectively. These coordinates underpin the description of the chemodynamical distribution of stars in the Milky Way.

3.2.1 Model Parameters

In Table 3.1 I provide short descriptions of the notation followed in this chapter to help with following the method.

Fig. 3.1 is a Bayesian network (or acyclic directed graph) representation of the method for calculating the selection functions for each field as a function of colour and apparent magnitude. Red boxes contain parameters and hyperparameters of the model. Green ellipses are conditional probabilities from which parameters and data are sampled. The blue double circles are the observable data.

In summary, the method consists of maximising the product off all probability distributions in green ellipses. The best fit hyperparameters then define the posterior parameters of the photometric density and selection function GMMs. In my method this is achieved as a two stage process, first fitting the photometric data to determine posterior GMM parameters for the photometric density. Subsequently the photometric GMM parameter posteriors are provided as priors to the spectroscopic likelihood function.

$\mathbf{x} = (\boldsymbol{\theta}, \mathbf{v})$	Full coordinates of the star.
$\{X_\mathrm{phot}\}$ $(\{X_\mathrm{spec}\})$	Sample of photometric (spectroscopic) data.
$\boldsymbol{\theta} = (l, b)$	Galactic coordinates.
$\mathbf{v} = (m, c)$	Observable coordiantes: colour and apparent magnitude.
τ, [M/H], $\mathcal{M}_\mathrm{ini}, s$	Intrinsic coordinates: age, metallicity, initial mass and distance
$\widetilde{\mathcal{M}}_\mathrm{ini}$	Initial mass scaled to the range $\widetilde{\mathcal{M}}_\mathrm{ini} \in [0, 1]$.
Θ_i	Event that a star is on field i.
$\mathcal{S}$	Event that a star is selected in the spectroscopic catalogue.
$\mathcal{S}_i$	Event that a star is selected on field i of the spectrosopic catalogue.
$f_\mathrm{phot}(f_\mathrm{spec})$	Distribution function (spectroscopic distribution).
$n_\mathrm{phot}(n_\mathrm{spec})$	Density of stars in the photometric (spectrosopic) sample.
$N_\mathrm{phot}(N_\mathrm{spec})$	Number of stars in the photometric (spectrosopic) sample.
$g_\mathrm{DF}(g_\mathrm{SF})$	GMM for the distribution function (selection function).
$\epsilon(\widetilde{\epsilon})$	Parameters of the distribution function (selection function).
α_k	Dirichlet concentration parameters for the photometric density GMM.
$w_{ik}, \boldsymbol{\mu}_{ik}, \boldsymbol{\Sigma}_{ik}$	Parameters of field i, component k of the photometric density GMM.
π_{ik}	Normalised component loadings of the photometric density GMM.
N	Normalisation of the photometric density GMM.
$\widetilde{w_{ik}}, \widetilde{\boldsymbol{\mu}_{ik}}, \widetilde{\boldsymbol{\Sigma}_{ik}}$	Parameters of field i, component k of the selection function GMM.
$\widetilde{\kappa_{ik}}$	logit transform of the component weights of the selection function GMM.
$\mathcal{L}^\mathrm{phot}(\mathcal{L}^\mathrm{spec})$	Likelihood function for the photometric density (spectroscopic density) fit.
M	Number of fields or patches.
K	Number of GMM components in photometric density.
$\widetilde{K}$	Number of GMM components in selection function.
GMM	Gaussian Mixture Model, Eq. 3.11.
NIW	Normal Inverse Wishart distribution, Eq. 3.14.
DIR	Dirichlet distribution, Eq. 3.15.
U	Uniform distribution.

Table 3.1 Notation followed in this chapter.

3.2.2 Dependence of Selection Function on Sky Positions

To determine the dependence of the selection function on positional coordinates, I bin the sky into independent regions called patches across which the selection function does not vary directly as a function of sky position. Therefore I can denote $\mathbf{\Theta}_i$ as the event that the star's coordinates are located on patch $\boldsymbol{i}$. Using this notation, I can define the probability of a star being located on a field given the star's sky positions as

$$\mathrm{P}(\Theta_i \mid \boldsymbol{\theta}) = \begin{cases} 1 & \text{if } \boldsymbol{\theta} \text{ is located on patch } \boldsymbol{i}, \\ 0 & \text{otherwise.} \end{cases} \tag{3.3}$$

Defining a Patch

I assume that the selection function given $\mathbf{v}$ is independent of $\boldsymbol{\theta}$ across any patch. This motivates us to divide the sky into smaller patches particularly in locations where I expect rapid changes in the selection function with respect to sky positions. However, using smaller patches reduces the number of stars per patch, which amplifies the effects of Poisson noise. A sufficient coarseness is required such that the effects of Poisson noise are limited but fine enough such that the dependence of the selection function on sky positions is small.

Before continuing, I should briefly explain what I mean by Poisson noise and why it is so central to this work. Detections and recordings of stars in stellar catalogues can be considered as independent events which are randomly sampled from a distribution over observed coordinates. Thus the detection of stars by stellar surveys is well modelled by a Poisson point process. The signal-to-noise ratio for a Poisson point process scales with the square root of the mean signal strength, $\frac{\mathrm{S}}{\mathrm{N}} \sim \sqrt{\lambda}$ where λ is the expected number of events per interval of observed coordinate space. Therefore the larger the number of events per interval, the stronger the signal-to-noise ratio. This is why reducing the number of stars per patch in order to increase the spatial resolution also leads to a reduction in the signal-to-noise ratio.

A multi-fibre spectroscopic survey is constructed by placing fibres on a plate so that each fibre is at the image of a star. The plate covers a solid angle on the sky. The solid angle of sky observed by each plate is here referred to as a *field*, and the coordinates of the centre of the field define the *field pointing*. The distributions of fields for APOGEE and RAVE in sky positions are shown in Fig. 3.2 where each coloured circle represents the extent of a different field, and the field pointings are located at the centre of each circle. For multi-fibre spectroscopic surveys I define a patch as the region of sky covered by a field. From here on I refer to patches as fields.

For a point on a single field, the selection function is the probability that the star is observed by that field given the star's photometric properties, $\mathbf{v}$. If a coordinate does not lie in a field, the selection function is 0.

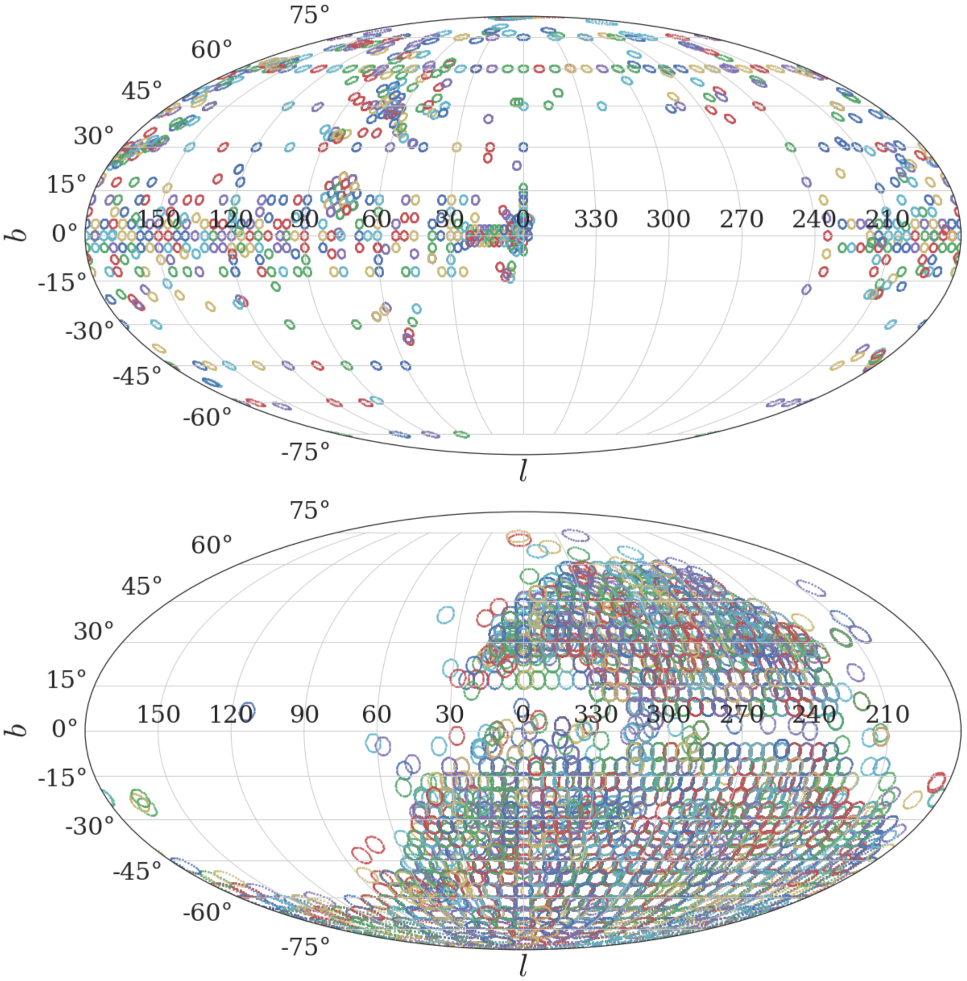


Fig. 3.2 Individual fields in APOGEE DR14 (top) and RAVE DR5 (bottom), shown here in Galactic coordinates, have substantial overlap within each survey. Despite being northern and southern hemisphere surveys respectively, their footprints overlap particularly in the equatorial plane.

The selection function for field i is given by

$$\mathrm{P}(\mathcal{S}_i \mid \boldsymbol{\theta}, \mathbf{v}) = \mathrm{P}(\mathcal{S}_i \mid \Theta_i, \mathbf{v})\mathrm{P}(\Theta_i \mid \boldsymbol{\theta}) \tag{3.4}$$

where $\mathcal{S}_i$ is the event that a star is selected by field i.

In many spectrograph surveys, as can be seen for APOGEE and RAVE in Fig. 3.2, fields may heavily overlap leading to a single coordinate being observed multiple times on different fields. The selection function is the probability that *at least* one field selects the star. This is given by the union of the event of each field selecting the star

$$\mathrm{P}(\mathcal{S} \mid \boldsymbol{\theta}, \mathbf{v}) = \mathrm{P}\left(\bigcup_{i=1}^{M} \mathcal{S}_i \middle| \boldsymbol{\theta}, \mathbf{v}\right) \tag{3.5}$$

where M is the total number of fields employed. This is calculated using all observations made on field i, even if the same star is observed on another field. The probability of the union of being on either of the two fields is the probability that one or the other occurs. I expand this in terms of the selection by individual fields

$$\mathrm{P}\left(\bigcup_{i}^{M} \mathcal{S}_i \middle| \boldsymbol{\theta}, \mathbf{v}\right) = \sum_{k=1}^{M} (-1)^{k+1} \left[\sum_{1 \leq i_1 < ... < i_k \leq M} \mathrm{P}(\mathcal{S}_{i_1}, \mathcal{S}_{i_2} ... \mathcal{S}i_k) \right] \tag{3.6}$$

where

$$\mathrm{P}(\mathcal{S}_1, \mathcal{S}_2, ...) = \prod_{i=1,2,...} \mathrm{P}(\mathcal{S}_i \mid \Theta_i, \mathbf{v})\mathrm{P}(\Theta_i \mid \boldsymbol{\theta}). \tag{3.7}$$

assuming the events $\mathcal{S}_1, \mathcal{S}_2...$ are independent. Appendix A.1 provides a detailed explanation of this expansion.

As mentioned earlier, many surveys contain multiple observations of the same positional coordinates. These can be observations taken on separate days to different magnitude depths. For Equation (3.7) to be appropriate, different observations of the same field should only be considered as separate fields if the observations are independent, i.e. if the probability of a star being selected by one observation is independent of whether the star is selected by the other observation. If the observations are dependent, as is the case if stars in one observation are chosen deliberately to be exclusive or inclusive of those observed in another observation, then the observations should be combined to form a single field in the selection function. In my method I combine all observations with the same field pointing as a single field.

In the following sections, I examine the dependence of the selection function on $\mathbf{v}$. These are calculated for a given field, i. Having calculated the selection function for each field as a function of $\mathbf{v}$, they are combined using the method above to achieve the full selection function for the entire survey.

3.2.3 Selection function in Observable Coordinates

The probability of a star being selected by field i given $\mathbf{v}$, and given that the star lies on the field (i.e. $\mathrm{P}(\Theta_i \mid \boldsymbol{\theta}) = 1$) is

$$\mathrm{P}(\mathcal{S}_i \mid \mathbf{v}, \Theta_i) = \frac{\mathrm{P}(\mathbf{v} \mid \mathcal{S}_i, \Theta_i)\mathrm{P}(\mathcal{S}_i \mid \Theta_i)}{\mathrm{P}(\mathbf{v} \mid \Theta_i)}. \tag{3.8}$$

$\mathrm{P}(\mathbf{v} \mid \Theta_i)\mathrm{d}\mathbf{v} \equiv f(\mathbf{v} \mid \Theta_i)\mathrm{d}\mathbf{v}$ is the probability that a star chosen at random on field i has coordinates within the d$\mathbf{v}$ volume element, where $f(\mathbf{v} \mid \Theta_i)$ is the distribution of stars in the Milky Way inside the cone projecting onto field i.

$\mathrm{P}(\mathbf{v} \mid \mathcal{S}_i, \Theta_i)\mathrm{d}\mathbf{v} \equiv f^{\mathrm{spec}}(\mathbf{v} \mid \Theta_i)\mathrm{d}\mathbf{v}$ is the probability that a star observed on field i of the spectroscopic survey has coordinates within the d$\mathbf{v}$ volume element. Finally, $\mathrm{P}(\mathcal{S}_i) = \mathbb{E}\left[N_i^{\mathrm{spec}}\right]/\mathbb{E}\left[N_i\right]$ is the probability of a star on field i entering the survey, and is given by the ratio of the number of stars in the survey on field i to the total number of stars in the Milky Way inside the cone projecting onto field i. Assuming that the photometric sample is complete within the given range of observable coordinates $f^{\mathrm{phot}}(\mathbf{v} \mid \Theta_i) = f(\mathbf{v} \mid \Theta_i)$ and $N_i^{\mathrm{phot}} = N_i$. Substituting into Equation (3.8)

$$\mathrm{P}(\mathcal{S}_i \mid \Theta_i, \mathbf{v}) = \frac{f_i^{\mathrm{spec}}(\mathbf{v} \mid \Theta_i)\mathbb{E}\left[N_i^{\mathrm{spec}}\right]}{f_i^{\mathrm{phot}}(\mathbf{v} \mid \Theta_i)\mathbb{E}\left[N_i^{\mathrm{phot}}\right]} = \frac{\mathbb{E}\left[n_i^{\mathrm{spec}}(\mathbf{v} \mid \Theta_i)\right]}{\mathbb{E}\left[n_i^{\mathrm{phot}}(\mathbf{v} \mid \Theta_i)\right]}, \tag{3.9}$$

where $n_i^{\mathrm{spec}}(\mathbf{v} \mid \Theta_i) = f_i^{\mathrm{spec}}(\mathbf{v} \mid \Theta_i)N_i^{\mathrm{spec}}$ is the number density of stars observed by the survey on field i and $n_i^{\mathrm{phot}}(\mathbf{v} \mid \Theta_i) = f_i^{\mathrm{phot}}(\mathbf{v} \mid \Theta_i)N_i^{\mathrm{phot}}$ is the number density of stars in the Milky Way on the cone projecting onto field i.

Number density of photometric sample

I start by calculating the expected number density of stars in the Milky Way on field i, $\mathbb{E}\left[n_i^{\mathrm{phot}}(\mathbf{v} \mid \Theta_i)\right]$. The choice of photometric survey is discussed in Section 3.4.

The stars in the photometric survey represent a Poisson realisation of the smooth underlying number density function, n^{phot}. The aim is to use the observed stars to estimate the true smooth number density function. I assume this function can be parameterised as a bivariate Gaussian Mixture Model (GMM) for each field i. Each bivariate Gaussian component is given by

$$G(\mathbf{v} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{\mid 2\pi\boldsymbol{\Sigma} \mid}} \exp\left(-\frac{(\mathbf{v}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{v}-\boldsymbol{\mu})}{2}\right), \tag{3.10}$$

where $\boldsymbol{\mu} \in \mathbb{R}^2$ is the mean of the bivariate Gaussian in colour and magnitude and $\boldsymbol{\Sigma} \in \mathbb{R}^{2\times2}$ is the symmetric covariance matrix. Using this, I can write

$$n_i^{\mathrm{phot}}(\mathbf{v} \mid \boldsymbol{\epsilon}_i, \Theta_i) = \sum_{k=1}^{K} w_{ik} G(\mathbf{v} \mid \boldsymbol{\mu}_{ik}, \boldsymbol{\Sigma}_{ik}), \tag{3.11}$$

where $\boldsymbol{\epsilon}_i \in \mathbb{R}^{K\times6}$ for the K components of the GMM, each with six parameters defining the bivariate Gaussian component w, μ^0, μ^1, Σ^{00}, Σ^{11} and Σ^{01} ($\boldsymbol{\Sigma}$ is symmetric so $\Sigma^{10} = \Sigma^{01}$).

I then find the parameters $\boldsymbol{\epsilon}_i$ by maximising the likelihood of the photometric catalogue stars. The Poisson likelihood is derived from Poisson count probabilities to give

$$\ln(\mathcal{L}^{\mathrm{phot}}(X_i^{\mathrm{phot}}|\boldsymbol{\epsilon}_i)) = -\int \mathrm{d}\mathbf{v}\, n_i^{\mathrm{phot}}(\mathbf{v} \mid \boldsymbol{\epsilon}_i, \Theta_i) + \sum_{j=1}^{N_i^{\mathrm{phot}}} \log\left(n_i^{\mathrm{phot}}(\mathbf{v}_j \mid \boldsymbol{\epsilon}_i, \Theta_i)\right), \tag{3.12}$$

where $\mathbf{v}_j$ are the colour and apparent magnitude of star j in the photometric catalogue on field i (see Appendix A.2).

I reparameterise the weights as normalised 'loadings', $\pi_{ik} = w_{ik}/N$ where $N = \sum_{k=1}^{K}[w_{ik}]$.

$$\ln(\mathcal{L}^{\mathrm{phot}}(X_i^{\mathrm{phot}}|\boldsymbol{\epsilon}_i)) \propto -N + N_i^{\mathrm{phot}}\ln(N) + \sum_{j=1}^{N_i^{\mathrm{phot}}} \ln\left(\hat{n}_i^{\mathrm{phot}}(\mathbf{v}_j \mid \boldsymbol{\epsilon}_i, \Theta_i)\right), \tag{3.13}$$

where $\hat{n}_i^{\mathrm{phot}} = n_i^{\mathrm{phot}}/N$. The normalisation, N is now a parameter of the model whilst the loadings provide $K-1$ free parameters under the constraint $\sum_{k=1}^{K}\pi_{ik} = 1$.

To obtain the posterior probabilities on each of my parameters I need priors. The prior on the normalisation N is uniform, $N \sim U[0,\infty]$. For the mean and covariance of the Gaussian components, I employ a Normal Inverse Wishart (NIW) prior.

$$\mathrm{NIW}(\boldsymbol{\mu},\boldsymbol{\Sigma} \mid \mathbf{m}_0, \lambda, \boldsymbol{\Psi}, \nu) \propto \mathcal{N}\left(\boldsymbol{\mu} \mid \mathbf{m}_0, \frac{1}{\lambda}\boldsymbol{\Sigma}\right)|\boldsymbol{\Sigma}|^{-\frac{(\nu+p+1)}{2}} \exp\left[-\frac{1}{2}\mathrm{Tr}\left(\boldsymbol{\Psi}\boldsymbol{\Sigma}^{-1}\right)\right], \tag{3.14}$$

where p is the number of degrees of freedom of the system, in my case two (colour and apparent magnitude). The Normal and Inverse Wishart distributions are the conjugate priors for the mean and covariance of a multivariate Gaussian distribution respectively. As a result, the posterior on $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ is also represented by a NIW distribution. The chosen hyperparameters $(\mathbf{m}_0, \lambda, \boldsymbol{\Psi}, \nu)$ are dependent on the prior information about the data. Choosing small values of λ and setting $\nu = p$ provides a least informative prior on the parameters. I discuss a good choice of hyperparameters in my mock tests in Section 3.3.

The loadings are evaluated with a Dirichlet prior,

$$\mathrm{DIR}(\pi_1, \pi_2 ... \pi_K | \alpha_1, \alpha_2 ... \alpha_K) \propto \prod_{k=1}^{K} \pi_k^{\alpha_k - 1} \tag{3.15}$$

which satisfies the constraint $\sum_{k=1}^{K}[\pi_{ik}] = 1$. The Dirichlet distribution is the conjugate prior for membership probability of a categorical variable. Setting concentration hyperparameters $\alpha_i = \alpha = 1/K \,\forall\, i$ provides a least informative prior.

For a given number of components, K I determine the posteriors on parameters of the normalised GMM using the BayesianGaussianMixture module from scikit-learn (Pedregosa et al., 2011) with parameters initialised by k-means clustering. This module iterates using a Variational Inference algorithm which alternates between assigning component membership probabilities and fitting the individual components. I recover the best fit hyperparameters of the posterior Dirichlet and NIW distributions $\alpha'_k, \mathbf{m}'_{0k}, \lambda'_k, \mathbf{\Psi}'_k$ and ν'_k where k is the Gaussian mixture component. These hyperparameters define the posterior distribution of parameters for the GMM. The posterior on the normalisation, N is given by the first two terms of the right hand side of Equation (3.13) where the uniform prior doesn't provide a contribution. This is proportional to a Poisson distribution with mean N_i^{phot}.

To determine the number of components I fit my GMM with $K = 1$ up to a maximum of $K = 20$ components and calculate the Bayesian information criterion (BIC, Schwarz, 1978) for the best fit parameters (taken as the expected values of the posteriors)

$$\mathrm{BIC} = \ln(N_i^{\mathrm{phot}})d - 2\ln(\mathcal{L}_{\mathrm{max}}^{\mathrm{phot}}) \tag{3.16}$$

where $d = K \times 6$ (the number of parameters in my model) and $\mathcal{L}_{\mathrm{max}}^{\mathrm{phot}}$ is the likelihood corresponding to the best fit parameters. My final model for the photometric sample colour-apparent magnitude function is the model which minimizes the BIC.

Selection function in observed coordinates

The selection function in observed coordinates for field i is $\mathrm{P}(\mathcal{S}_i \mid \mathbf{v}, \Theta_i)$. The stars in the spectroscopic catalogue represent a Poisson realisation of the product of the distribution function and the selection function

$$n_i^{\mathrm{spec}}(\mathbf{v} \mid \Theta_i) = n_i^{\mathrm{phot}}(\mathbf{v} \mid \Theta_i) \times \mathrm{P}(\mathcal{S}_i \mid \mathbf{v}, \Theta_i). \tag{3.17}$$

Therefore, I can once more use the Poisson likelihood in Equation (3.12), replacing $n^{\mathrm{phot}}(\mathbf{v} \mid \Theta_i)$ with $n^{\mathrm{spec}}(\mathbf{v} \mid \Theta_i)$.

I parameterise the selection function $\mathrm{P}(\mathcal{S} \mid \mathbf{v}, \Theta_i)$ also as a bivariate GMM in colour-magnitude space with parameters $\widetilde{\boldsymbol{\epsilon}_i}$

$$g_i^{\mathrm{SF}}(\mathbf{v} \mid \widetilde{\boldsymbol{\epsilon}_i}, \Theta_i) = \sum_{k=1}^{\widetilde{K}} \widetilde{w_{ik}} G(\mathbf{v} \mid \widetilde{\boldsymbol{\mu}_{ik}}, \widetilde{\mathbf{\Sigma}_{ik}}). \tag{3.18}$$

The log likelihood is then given by

$$\ln(\mathcal{L}^{\mathrm{spec}}(X_i^{\mathrm{spec}}|\boldsymbol{\epsilon}_i,\widetilde{\boldsymbol{\epsilon}_i})) \propto -\int \mathrm{d}\mathbf{v}\, n_i^{\mathrm{phot}}(\mathbf{v}\mid\boldsymbol{\epsilon}_i,\Theta_i)\, g_i^{\mathrm{SF}}(\mathbf{v}\mid\widetilde{\boldsymbol{\epsilon}_i},\Theta_i) \\ + \sum_{j=1}^{N_i^{\mathrm{spec}}} \log\left(n_i^{\mathrm{phot}}(\mathbf{v}_j\mid\boldsymbol{\epsilon}_i,\Theta_i)\, g_i^{\mathrm{SF}}(\mathbf{v}_j\mid\widetilde{\boldsymbol{\epsilon}_i},\Theta_i)\right). \tag{3.19}$$

As the selection function is a probability distribution, it must fall in the range $g_i^{\mathrm{SF}}(\mathbf{v}\mid\widetilde{\boldsymbol{\epsilon}_i},\Theta_i)\in[0,1]$ for all $\mathbf{v}$. I reparametrise the selection function component weights as $\widetilde{\kappa_{ik}}=\mathrm{logit}\left(\frac{\widetilde{w_{ik}}}{\sqrt{|2\pi\widetilde{\boldsymbol{\Sigma}_{ik}}}}\right)$ with a uniform prior $\widetilde{\kappa_{ik}}\sim\mathrm{U}[-\inf,\inf]$. This constrains the component weights to the range $\widetilde{w_{ik}}\in\left[0,\sqrt{|2\pi\widetilde{\boldsymbol{\Sigma}_{ik}}|}\right]$ such that the maxima of any Gaussian component is less than or equal to one. The sum of selection function Gaussian mixture components must also be less than or equal to one everywhere. Since the maxima of a GMM is non-analytic, I find the roots of the gradient of the GMM using the *hybr* method from MINPACK-1 (More et al., 1980) implemented in SCIPY initialising at the mean of each GMM component. If the value of the GMM at any root is greater than one, the posterior probability is set to zero.

A NIW prior is used for $\widetilde{\boldsymbol{\mu}_{ik}},\widetilde{\boldsymbol{\Sigma}_{ik}}$ of the selection function.

I fit simultaneously for both the selection function parameters, $\widetilde{\boldsymbol{\epsilon}_i}$ and photometric density parameters, $\boldsymbol{\epsilon}_i$ where the prior distributions on $\boldsymbol{\epsilon}_i$ are the posteriors of the fit to the photometric sample in Section 3.2.3.

Combining Equation (3.19) with the priors on all parameters the posterior is given by

$$\ln(\mathrm{P}(\boldsymbol{\epsilon}_i,\widetilde{\boldsymbol{\epsilon}_i}|X_i^{\mathrm{spec}})) = \ln(\mathcal{L}^{\mathrm{spec}}(X_i^{\mathrm{spec}}|\boldsymbol{\epsilon}_i,\widetilde{\boldsymbol{\epsilon}_i}))-N+N_i^{\mathrm{phot}}\ln(N) \\ +\mathrm{DIR}(\pi_1,\pi_2...\pi_K|\alpha'_1,\alpha'_2...\alpha'_K) \\ +\sum_{k=1}^{K}\left[\mathrm{NIW}(\boldsymbol{\mu}_{ik},\boldsymbol{\Sigma}_{ik}|\mathbf{m}'_{0ik},\lambda'_{ik},\boldsymbol{\Psi}'_{ik},\nu'_{ik})\right] \\ +\sum_{k=1}^{\widetilde{K}}\left[\mathrm{NIW}(\widetilde{\boldsymbol{\mu}_{ik}},\widetilde{\boldsymbol{\Sigma}_{ik}}|\widetilde{\mathbf{m}_0},\widetilde{\lambda},\widetilde{\boldsymbol{\Psi}},\widetilde{\nu})\right]. \tag{3.20}$$

To fit the selection function, I need to find $\mathrm{argmin}_{\boldsymbol{\epsilon}_i,\widetilde{\boldsymbol{\epsilon}_i}}\,(-\ln\mathrm{P})$. All terms in the posterior probability are analytically differentiable so I employ the 'Truncated Newton' method in SCIPY which uses the gradient of the probability distribution to converge on the minima. Parameters of the photometric density, $\boldsymbol{\epsilon}_i$ are initialised at the mean values of the prior distribution. The selection function parameters are initialised using two methods:

- k-means clustering with each spectroscopic sample star providing a weighted contribution of $\frac{1}{n_i^{\mathrm{phot}}(\mathbf{v}_j|\boldsymbol{\epsilon}_i,\Theta_i)}$. The reweighting is akin to approximating the selection function as $n^{\mathrm{spec}}/n^{\mathrm{phot}}$.

- Fit a GMM directly to the spectroscopic data using BayesianGaussianMixture with least informative Dirichlet and NIW priors.

The parameters are optimised for both initialisations and the one which leads to the largest posterior probability is used as the best fit. This helps to avoid some local optima in the posterior distribution.

In order to determine the optimal number of Gaussian components to use for the selection function, the algoritm is run for $\widetilde{K} = [1, K]$ components. I do not allow more components than the photometric density GMM as this would likely result in overfitting of the selection function. I use the number of components which minimizes the BIC

$$\mathrm{BIC} = \ln(N_i^{\mathrm{spec}})d - 2\ln(\mathcal{L}_{\mathrm{max}}^{\mathrm{spec}}) \tag{3.21}$$

where the number of degrees of freedom $d = (K + \widetilde{K}) \times 6$ and $\mathcal{L}_{\mathrm{max}}^{\mathrm{spec}}$ is the likelihood corresponding to the best fit parameters. To generate a full posterior on the parameters of the photometric density and selection function, I run a set of $6 \times (K + \widetilde{K}) \times 2$ chains for 2000 iterations with `emcee` (Foreman-Mackey et al., 2013) initialising from a small ball around the best fit parameters. For every optima of the posterior distribution, there are a set of $\widetilde{K}!$ degenerate optima generated by reordering components. For finding the best fit parameters, it is unimportant which of the optima I sample around, however the posterior on the parameters should not be degenerate. I constrain the means of the photometric density and selection function components to maintain their respective orders in colour and apparent magnitude throughout the `emcee` iterations.

The models, their parameters and the prior distributions are summarised in Table 3.2 for reference. The choice of model hyperparameters is discussed further in Section 3.3.3.

3.2.4 Intrinsic Coordinates

Chemodynamical models of the Milky Way make predictions for the metallicities, masses, and ages of stars at different positions. In order to test these models against observations, the selection function is required to account for the observation biases of the survey. To use the selection function, I require a transformation between observable coordinates (colour and apparent magnitude) and intrinsic coordinates (distance, metallicity, mass, and age). The selection function in terms of the intrinsic coordinates of the stars is

$$\mathrm{P}(\mathcal{S} \mid s, [\mathrm{M/H}], \mathcal{M}_{\mathrm{ini}}, \tau, \boldsymbol{\theta}), \tag{3.22}$$

where s is the heliocentric distance, and [M/H], $\mathcal{M}_{\mathrm{ini}}$, and τ are the star's metallicity, initial mass, and age respectively.

Here I follow a similar method for transforming between observable and intrinsic coordinate systems as described in Sanders & Binney (2015) and Das & Binney (2016). Any combination of [M/H], $\mathcal{M}_{\mathrm{ini}}$, and τ maps to a single set of coordinates, (c, M), which are the colour and absolute magnitude of the star. Any values of M and s uniquely define m, the apparent magnitude of the star. Therefore any intrinsic coordinates, $(s, [\mathrm{M/H}], \mathcal{M}_{\mathrm{ini}}, \tau)$

Data	Component	Model	Parameters	Prior	Hyperparameters
Photometric	Distribution Function	GMM	N_i	$\mathrm{U}[0, \mathrm{inf}]$	
			π_{ik}	Dirichlet	$\alpha = 1/K$
			$w_{ik} = N_i \, \pi_{ik}$		
			$\boldsymbol{\mu}_{ik}, \, \boldsymbol{\Sigma}_{ik}$	NIW	$\mathbf{m}_0 = \mathrm{med}[\mathbf{v}_{\mathrm{min}}, \mathbf{v}_{\mathrm{max}}], \; \lambda = 10^{-4}$ $\boldsymbol{\Psi} = \left(\frac{\mathrm{diag}(\mathbf{v}_{\mathrm{max}} - \mathbf{v}_{\mathrm{min}})/2}{5}\right)^2, \; \nu = 2$
Spectroscopic	Distribution Function	GMM	N_i	Poisson	$\mathrm{mean} = N_i^{\mathrm{phot}}$
			π_{ik}	Dirichlet	α'_{ik}
			$w_{ik} = N_i \, \pi_{ik}$		
			$\boldsymbol{\mu}_{ik}, \, \boldsymbol{\Sigma}_{ik}$	NIW	$\mathbf{m}'_{0ik}, \; \lambda'_{ik}$ $\boldsymbol{\Psi}'_{ik}, \; \nu'_{ik}$
	Selection Function	GMM	$\widetilde{\kappa_{ik}}$	$\mathrm{U}[-\mathrm{inf}, \mathrm{inf}]$	
			$\widetilde{w_{ik}} = \sqrt{\lvert 2\pi \widetilde{\boldsymbol{\Sigma}_{ik}} \rvert} \, \mathrm{logit}^{-1}(\widetilde{\kappa_{ik}})$		
			$\widetilde{\boldsymbol{\mu}_{ik}}, \, \widetilde{\boldsymbol{\Sigma}_{ik}}$	NIW	$\widetilde{\mathbf{m}_0} = \mathrm{med}[\mathbf{v}_{\mathrm{min}}, \mathbf{v}_{\mathrm{max}}], \; \widetilde{\lambda} = 10^{-4}$ $\widetilde{\boldsymbol{\Psi}} = \left(\frac{\mathrm{diag}(\mathbf{v}_{\mathrm{max}} - \mathbf{v}_{\mathrm{min}})/2}{5}\right)^2, \; \widetilde{\nu} = 2$

Table 3.2 Parameters and hyperparameters of the GMMs and their respective priors to be fit to the photometric and spectroscopic catalogues. i refers to the field on which the model fit is performed with every field fit independently.

map to observable coordinates, (c, m) and to a single value of the selection function, $\mathrm{P}(\mathcal{S} \mid \mathbf{v}, \boldsymbol{\theta})$. The selection function in intrinsic coordinates for a star on field i is therefore given by

$$\mathrm{P}(\mathcal{S} \mid s, \tau, [\mathrm{M/H}], \mathcal{M}_{\mathrm{ini}}, \boldsymbol{\theta}) = \mathrm{P}(\mathcal{S} \mid \mathbf{v}(s, \tau, [\mathrm{M/H}], \mathcal{M}_{\mathrm{ini}}), \boldsymbol{\theta}) \tag{3.23}$$

where $\mathbf{v} = (c, m)$. The maps are given by

$$m = M + 5\log_{10}\left(\frac{s}{10\mathrm{pc}}\right) \tag{3.24}$$

and

$$(c, M) = F^{\mathrm{iso}}(\tau, [\mathrm{M/H}], \mathcal{M}_{\mathrm{ini}}), \tag{3.25}$$

where F^{iso} is the mapping introduced by the isochrones.

There are a variety of methods for generating the F^{iso} map. The most common method is to adopt the nearest isochrone to the values of age and metallicity provided. For this work, I linearly interpolate between isochrones, which improves the accuracy of the transformation.

I use the PARSEC isochrones (Bressan et al., 2012) on a grid of 353 ages and 57 metallicities. Ages in the range $-2.40 \leq \log_{10}(\tau/\mathrm{Gyr}) \leq 1.12$ with a spacing of 0.01 dex, and metallicities in the range $-2.192 \leq [\mathrm{M/H}] \leq 0.696$ with a spacing of 0.051 dex, were considered.

Every isochrone has a maximum initial mass above which a star of the given age and metallicity cannot exist. I generate a scaled initial mass coordinate so that each isochrone varies from $\widetilde{\mathcal{M}}_{\mathrm{ini}} \in [0, 1]$,

$$\widetilde{\mathcal{M}}_{\mathrm{ini}} = \frac{\mathcal{M}_{\mathrm{ini}} - \mathcal{M}_{\mathrm{ini}}^{\mathrm{min}}}{\mathcal{M}_{\mathrm{ini}}^{\mathrm{max}} - \mathcal{M}_{\mathrm{ini}}^{\mathrm{min}}}, \tag{3.26}$$

where $\mathcal{M}_{\mathrm{ini}}^{\mathrm{max}}(\tau, [\mathrm{M/H}])$ is the maximum initial mass value for a star of a given age and metallicity which I determined by linearly interpolating between maximum initial mass of isochrones as a function of τ, [M/H]. Likewise for $\mathcal{M}_{\mathrm{ini}}^{\mathrm{min}}(\tau, [\mathrm{M/H}])$.

The scaled mass is linearly interpolated along each isochrone. A single set of $\widetilde{\mathcal{M}}_{\mathrm{ini}}$ values is drawn for all isochrones which samples most heavily where the curvature of the isochrone in colour-apparent magnitude space is greatest. This parameterisation enables us to interpolate colour and absolute magnitude on a regular $\tau - [\mathrm{M/H}] - \widetilde{\mathcal{M}}_{\mathrm{ini}}$ grid. The interpolation is the mapping $F^{\mathrm{iso}}(\tau, [\mathrm{M/H}], \widetilde{\mathcal{M}}_{\mathrm{ini}}(\mathcal{M}_{\mathrm{ini}}))$.

3.3 Mock Tests

To test the performance of the method presented in Section 3.2, I apply it to mock samples with known selection functions.

3.3.1 `Galaxia` Mock Catalogue

I generate my mock catalogue using the `Galaxia` code (Sharma et al., 2011). The thin disk is generated from the analytic Besançon model (Robin et al., 2003). Non-circular motions in the plane of the disk are introduced through the Shu 1969 DF, and the Bullock & Johnston 2005 N-body models simulate any substructure in the halo. The code synthesises a population of stars with coordinates, trajectories and intrinsic properties including age, metallicity and mass. Given the age, metallicity, initial mass and distance, `Galaxia` also calculates the colour and apparent magnitude of stars as observed from the Sun by employing the nearest Padova isochrone (Marigo et al., 2008) from a grid of 182 ages and 34 metallicities. The PARSEC (Bressan et al., 2012) isochrones however present a significant update on the Padova isochrones in terms of revisions to major input physics such as the equation of state, opacities, nuclear reaction rates, and inclusion of the pre-main sequence phase. The nearest isochrone calculation method is also less accurate than the interpolation method I employ with PARSEC isochrones discussed in Section 3.2.4. For these reasons, I recalculate the `Galaxia` apparent magnitudes and colours using my method.

I sample from `Galaxia` with an H-band magnitude limit, $m_{\mathrm{H}} < 15$, similar to the limitations of the 2MASS survey (Skrutskie et al., 2006). The H-band magnitude limit was placed using `Galaxia` such that stars were filtered out based on the built-in Padova magnitudes. The difference between this and my own magnitude calculation leads to some stars dimmer than the magnitude cut being included and some brighter stars being excluded. This does not significantly affect my tests.

I refer to this catalogue as the photometric sample.

3.3.2 Imposed Selection Function

The selection function is applied across two fields with coordinates

$$l, b = \begin{cases} 30, 60 & \text{(Field 1)}, \\ 30, 61 & \text{(Field 2)}. \end{cases} \tag{3.27}$$

Each field has a half-opening angle of 2 degrees and hence a solid angle of 12.6 degrees^2. The locations of the fields in galactic coordinates are given in Fig. 3.3 where the red and green shaded regions are fields 1 and 2 respectively and the greyscale background shows the number density of stars in the photometric sample.

For each field I apply three different selection functions:

- Flat: Selection boundaries of $m_H < 13.5$ and $J - K > 0.5$. The value of the selection function is 0.1 within the boundaries and 0 outside.
- tanh:

$$\begin{aligned} \mathrm{P}(S|m_H, J-K) = 0.1 \times &\left(1 - \tanh\left(\frac{m_H - 13.5}{e^{-2}}\right)\right)/2 \\ \times &\left(1 - \tanh\left(\frac{J - K - 0.5}{e^{-2}}\right)\right)/2. \end{aligned} \tag{3.28}$$

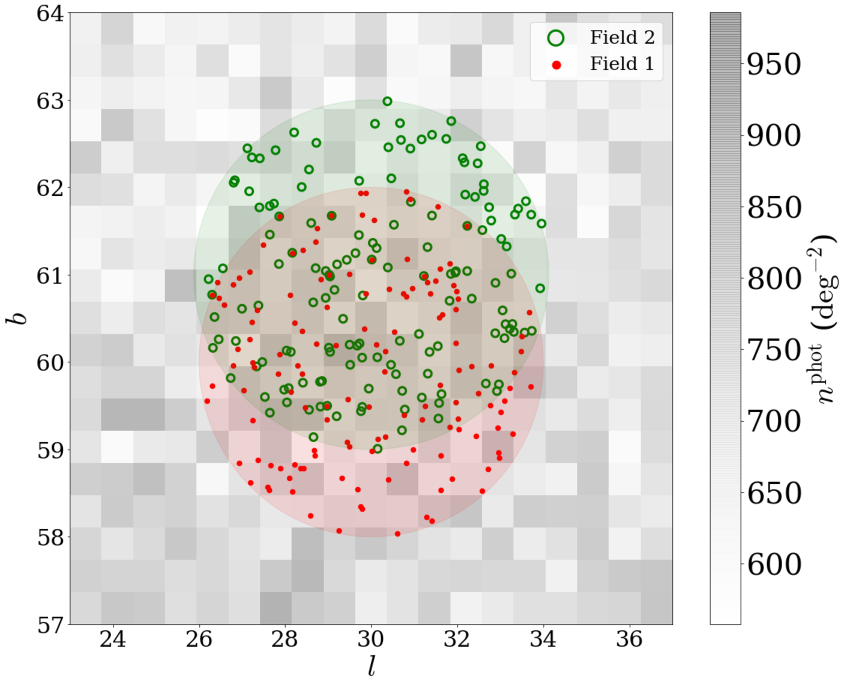


Fig. 3.3 The mock `Galaxia` model uses two fields offset by 1 deg latitude represented here by the red and green shaded regions. The grey scale bins give the number density of stars in the photometric sample which systematically decreases further from the plane. Applying the Flat selection function returns the stars represented by red points and green circles for fields 1 and 2 respectively. In the region of overlap there are cases where stars are selected on both fields.

This is of a similar form to the selection function proposed for SEGUE G-dwarfs by Bovy et al. 2012b.

- RAVE: I take 10 randomly selected fields from the RAVE survey (Kordopatis et al., 2013) along with any 2MASS (Skrutskie et al., 2006) stars on the same fields, bin the stars in $m_{H,2\mathrm{MASS}} - (J-K)_{2\mathrm{MASS}}$ and use the ratio of RAVE to 2MASS stars in each bin as the selection function probability.

The three applied selection functions are plotted in top panels of Fig. 3.4. The red and green scatter points in Fig. 3.3 are the stars selected by the Flat selection function to demonstrate the setup.

In the region of overlap of the fields, stars may be selected by either or both of the fields. I include every star selected by at least one field however I also record which fields each star was selected by as this is important for calculating the selection function. For instance, when deriving the selection function for field 1, I must use all stars which were selected by field 1 even if they were also selected by field 2. Likewise for field 2, I must include all stars selected by field 2 even if they were also selected by field 1. The effective double counting of selection functions in overlapping regions is accounted for by the union calculation as described in Section 3.2.2.

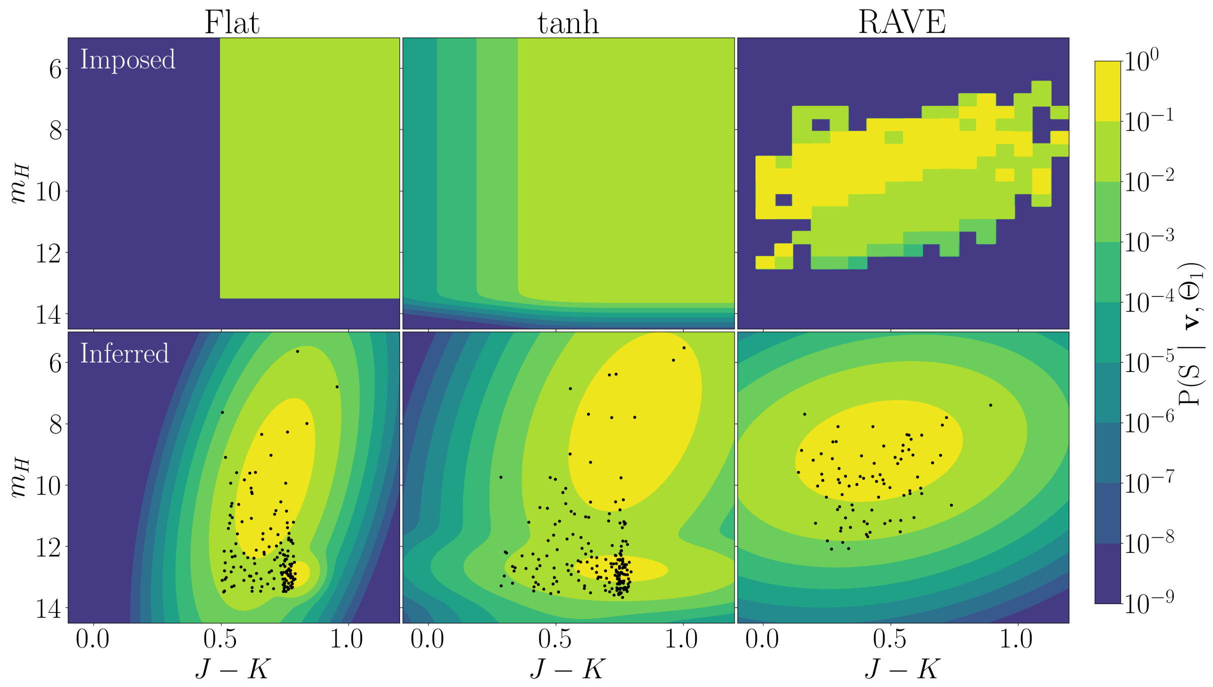


Fig. 3.4 I apply three different selection function models to the `Galaxia` mock sample, Flat (left), tanh (middle) and RAVE (right). Top: The models are defined as a function of colour and apparent magnitude. Bottom: Using the BIC as my model selection criteria, two GMM components are fit for each model shown by the contours. The model broadly recovers the form of the imposed selection function in regions occupied by the selected samples which are shown as black points.

3.3.3 Results

The `Galaxia` photometric and spectroscopic catalogues are used to calculate the selection functions in observable coordinates, $\mathrm{P}(\mathcal{S}\mid\boldsymbol{\theta},c,m)$, and intrinsic coordinates, $\mathrm{P}(\mathcal{S}\mid\boldsymbol{\theta},s,[\mathrm{M/H}],\mathcal{M}_{\mathrm{ini}},\tau)$ following the methods described in Section 3.2.

I first discuss the values of hyperparameters used for the GMM priors. I present and test the posterior for field 1 as a function of colour and apparent magnitude. I then show the fit to both fields in observable coordinates and position on the sky which tests how the model handles overlapping fields. Finally I present the selection function in intrinsic coordinates.

Prior Hyperparameters

For the prior distributions on the photometric density and selection function I require hyperparameter values which encode my prior knowledge about the distribution. I assume that I know all data falls within the apparent magnitude and colour ranges $m_H\in[4,15]$ and $J-K\in[-0.1,1.2]$.

For the NIW hyperparameters on the photometric density fit I use $\mathbf{m}_0=\mathrm{med}\,[\mathbf{v}_{\mathrm{min}},\mathbf{v}_{\mathrm{max}}]=(9.5,0.55)$ which is at the centre of my prior range. I set $\lambda=10^{-4}$. By choosing a small value for λ I ensure that the prior is extremely uninformative on the mean. For the covariance I choose $\mathbf{\Psi}=\mathrm{diag}((\frac{\mathbf{v}_{\mathrm{max}}-\mathbf{v}_{\mathrm{min}}}{2})/5)^2)=\mathrm{diag}(1.21,0.0169)$ such that the standard deviation is 1/5 the prior range in m_H, $J-K$. The precise choice of

1/5 is somewhat arbitrary and chosen as it leads to reasonable numbers of Gaussian components in the photometric density and selection function GMMs in all cases. The degrees of freedom is chosen as $\nu = 2$ which is the most uninformative choice. For the Dirichlet prior I choose a concentration prior of $\alpha_k = 1/K$ for all components which is the least informative prior.

I use the same hyperparameters for the NIW prior on the selection function as those used for the photometric density fit. This is because I have no more information about the selection function than I did about the photometric density other than that it is only important in the given colour and magnitude range. Having found the best fit parameters, I also run a set of `emcee` chains as discussed in Section 3.2.3 to generate posterior samples on all of the photometric density and selection function parameters.

The choices of hyperparameter values are summarised in the right-hand column of Table 3.2.

Observable Coordinates

In this section I demonstrate the selection function fit to field 1 of the mock sample in colour-apparent magnitude space. To minimize the BIC, the optimal number of photometric density GMM components was 14 whilst the Flat, tanh and RAVE selection functions were fit with 2, 2 and 1 components.

For the RAVE mock, the distribution of photometric and spectroscopic points and the GMM distribution fits are given in Fig. 3.5. For a successful fit I would expect the centres of 68% of histogram bins to fall within the Poisson uncertainties of the model and I see that this appears to be the case for my fit to the RAVE selection function against both colour and apparent magnitude. The same number density distributions for all three mock selection functions are shown in Fig. 3.6. For the Flat model (top panel) the smooth GMM selection function fails to correctly fit the sharp cuts at $m_H = 13.5$ and $J - K = 0.5$. This is a predictable limitation of attempting to fit a discontinuous model with a smooth function. In all other cases the histograms of the spectroscopic sample mostly fall within the Poisson noise uncertainties of the products between my fits to the photometric density and selection function. This provides a qualitative demonstration that the method is providing reasonable results.

For a more quantitative assessment I use the Kolmogorov-Smirnov (KS) test. I randomly draw 1000 samples from the last 500 iterations of each `emcee` fit to the spectroscopic sample. The KS probability (p-value of the KS test) is the probability that the spectroscopic sample is consistent with being drawn from the given probability distribution, in this case the product of the photometric and selection function GMMs which is itself a GMM with $K \times \widetilde{K}$ components.

The KS statistics are computed in 1D as a function of colour and apparent magnitude separately. For a good fit to the data, the KS probabilities are uniformly distributed in the range $[0, 1]$. If the probability is skewed towards 0, the method has underfit or failed to fit the data. If the probability is skewed towards 1, the method has overfit the data and hence the selection function would not generalise well to a new dataset. The most extreme

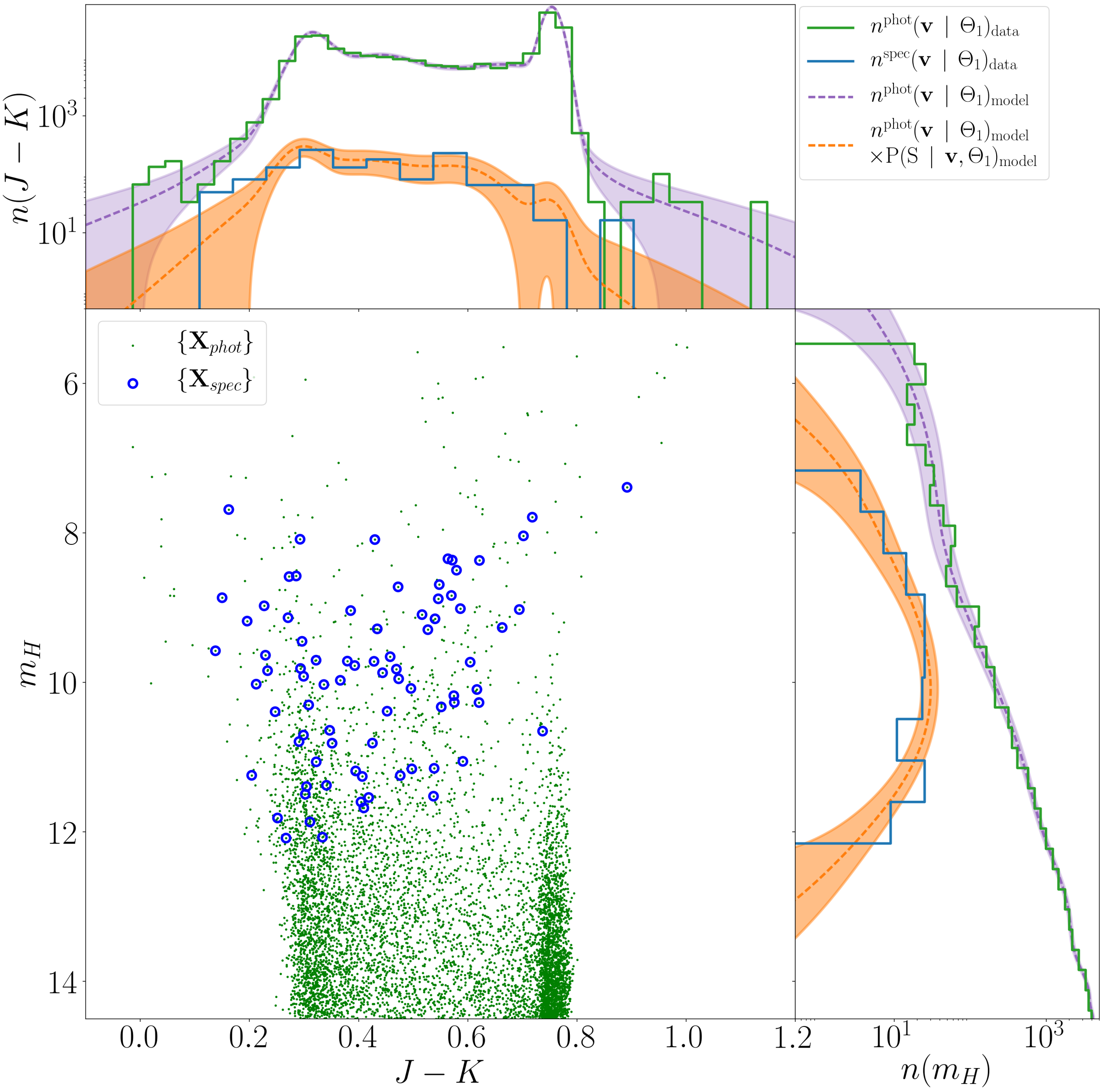


Fig. 3.5 The selection function selects a subset (blue circles) of the photometric sample (green points) to be included in the spectroscopic catalogue where in this case I am showing the RAVE mock applied to field 1. The model fits 14 GMM components to the photometric sample for which purple dashed lines show the 1D projection. The photometric sample distribution (green solid histograms) mostly fall within Poisson noise uncertainty (purple shaded region) of the GMM fit. The selection function is fit with 2 GMM components such that the spectroscopic sample is modelled by the 28 component product of the two GMMs, represented by the orange dashed line. The spectroscopic subsample (blue solid histogram) largely falls within the Poisson noise uncertainty (orange shaded region) of the model fit.

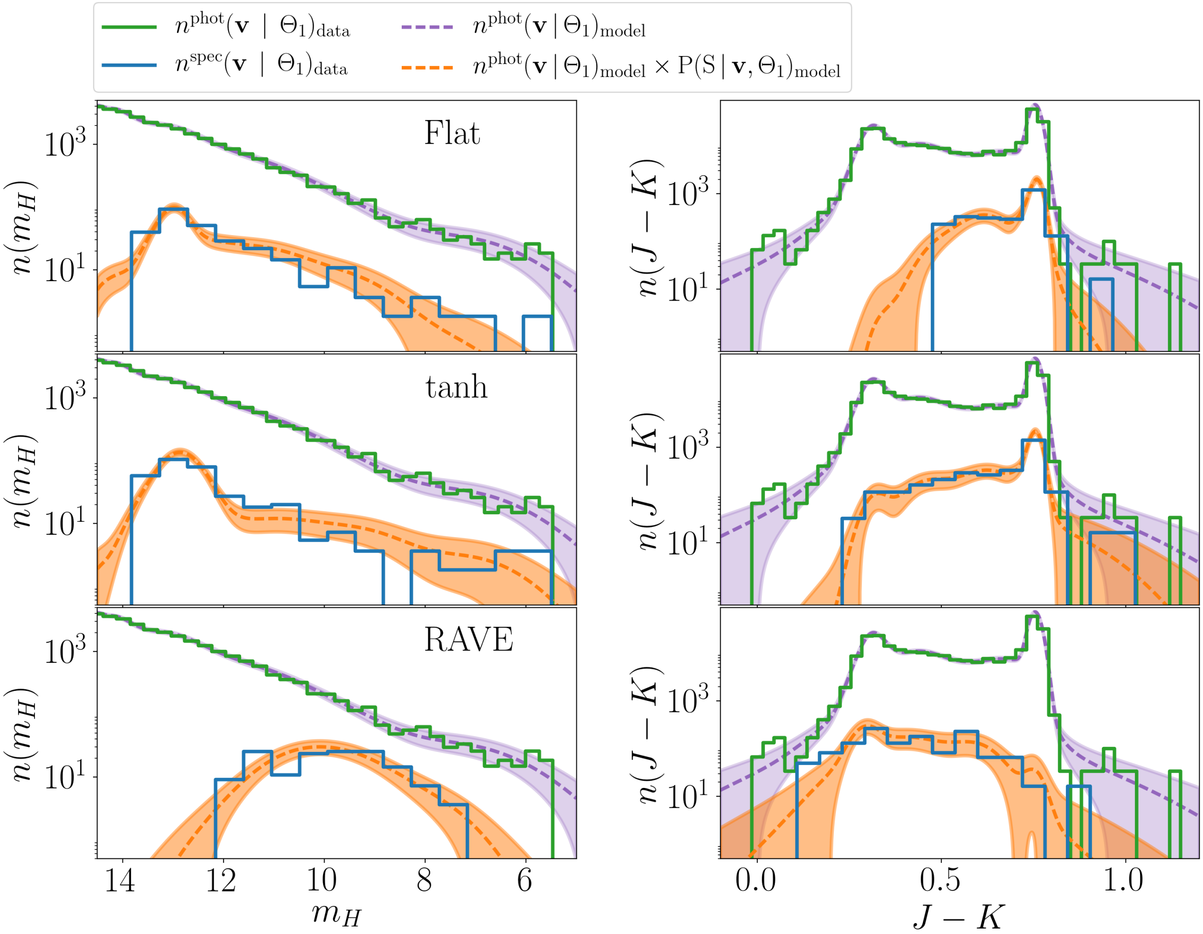


Fig. 3.6 The same as the side panels of Fig. 3.5 for the Flat (top), tanh (middle) and RAVE (bottom) models applied to field 1 of the `Galaxia` mock. In the majority of cases the spectroscopic model, given by the product of the photometric sample and selection function GMMs, traces the spectroscopic samples within Poisson noise uncertainty. The noticeable limitation is in the Flat model (top panels) where the smooth GMM cannot reproduce the sharp cut-off in colour ($J-K = 0.5$) and apparent magnitude ($m_H = 13.5$).

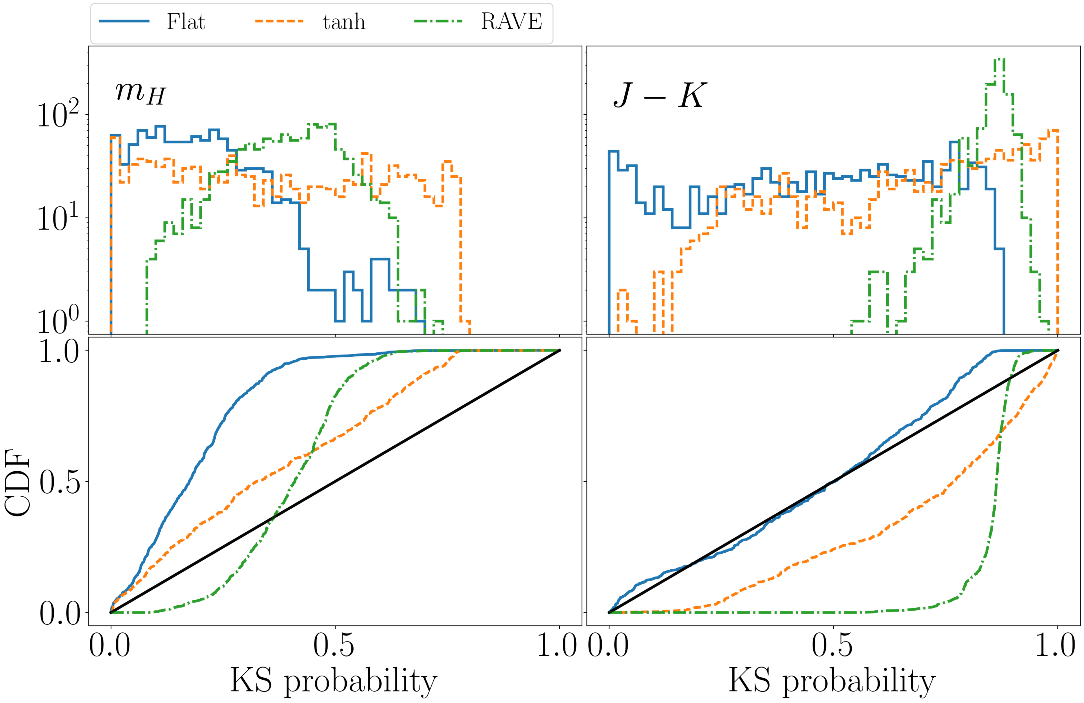


Fig. 3.7 The 1D one sample KS probability is evaluated for the field 1 spectroscopic sample against the GMMs of 1000 random draws from the `emcee` chains. In apparent magnitude (left) the Flat selection function shows underfitting (blue solid) as the distribution of probabilities (top) is biased towards 0 and the CDF (bottom) sits well above a uniform distribution (black solid). In colour (right) the RAVE sample has overfit the data (green dot-dashed) as the KS probabilities are biased towards 1. The remainder of the tests are closer to uniformly distributed and demonstrate reasonable fits to the data against the given coordinates.

example of this would be placing delta functions on every star which would achieve a KS probability of 1.

I expect that the KS statistic should demonstrate some overfitting since the test does not consider the prior information provided by the photometric sample.

Histograms and cumulative distributions of the KS tests for field 1 against colour and apparent magnitude are shown in Fig. 3.7. Due to the sharp cut off in the Flat model, the KS probabilities demonstrate that the GMM underfits the selection function as a function of magnitude. Curiously, the same problem is not seen as a function of colour where the Flat model is almost perfectly fit. KS tests are only weakly sensitive to the wings of the distribution and I postulate that this is why the sharp cutoff issue is not picked up in colour space. The RAVE and tanh models provide reasonable fits as a function of apparent magnitude however the RAVE model is heavily overfit as a function of colour. The cause of this overfitting is unclear.

To test the normalisation of the photometric density and selection function I compare the integral over colour and apparent magnitude of the GMMs with the number of objects in the spectroscopic sample. For large samples, I would expect the difference between the integral and true count to be Gaussian distributed with variance equal to the size of the sample. Fig. 3.8 shows the distribution of these integrals from the 1000 `emcee`

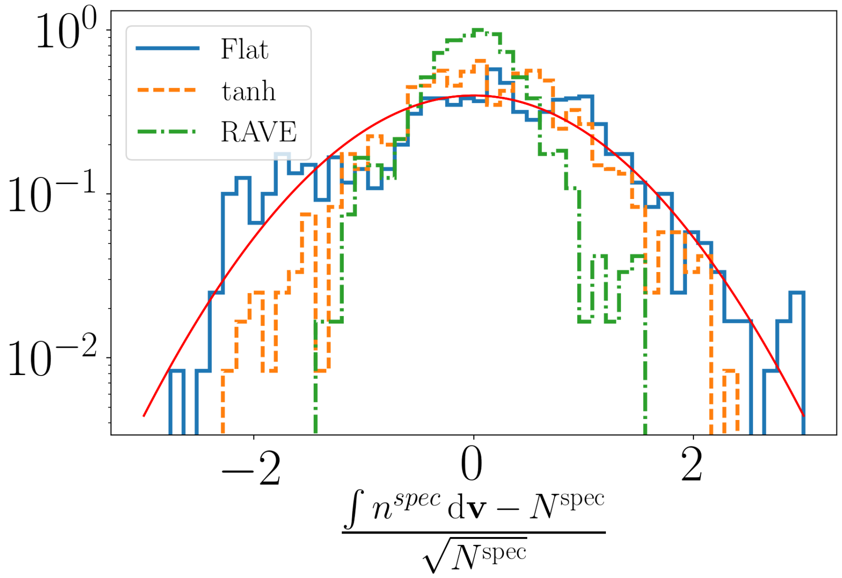


Fig. 3.8 The integrals over the GMM fits are compared to the number of stars in the spectrograph subsample. A well-fit model would be normally distributed around $\int n^{\mathrm{spec}}\mathrm{d}\mathbf{v} = \sqrt{N^{\mathrm{spec}}}$ with dispersion $\sqrt{N^{\mathrm{spec}}}$ demonstrated by the red solid line. The Flat model fit (blue solid) accomplishes this whilst the tanh (orange dashed) and RAVE (green dot-dashed) fits show a tighter distribution indicative of overfitting.

parameter draws re-centred and re-scaled by the spectroscopic sample number counts. All distributions are centred around zero which implies that the fits are well normalised. The fits to the RAVE model show a tight distribution which is another indication of overfitting to the spectroscopic data.

Overlapping Fields

On field 2 the photometric density is fitted with 14 GMM components and the Flat, tanh and RAVE selection functions with 2, 3 and 1 components respectively to minimize the BIC. I now only consider the selection function resulting from the combination of fields 1 and 2.

The selection function is now a superposition of two GMMs as a function of galactic coordinates. As such the KS one-sample statistic can no longer be used as I do not have an analytic distribution to fit. Instead I draw random samples from the photometric catalogue with the probability of inclusion given by the model selection function as a function of galactic coordinates, colour and apparent magnitude. I compute the two-sample KS probability against the spectroscopic catalogue which gives the probability that samples are drawn from the same probability distribution.

In Fig. 3.9 I show the distribution of KS probabilities for fields 1 and 2 with all combinations of the Flat, tanh and RAVE selection functions applied. The Flat+Flat model (blue solid line) systematically underfits the data in both apparent magnitude and colour which, as discussed earlier, is caused by the sharp cuts in the selection function which a smooth GMM cannot correctly reproduce. All other combinations of fields produce

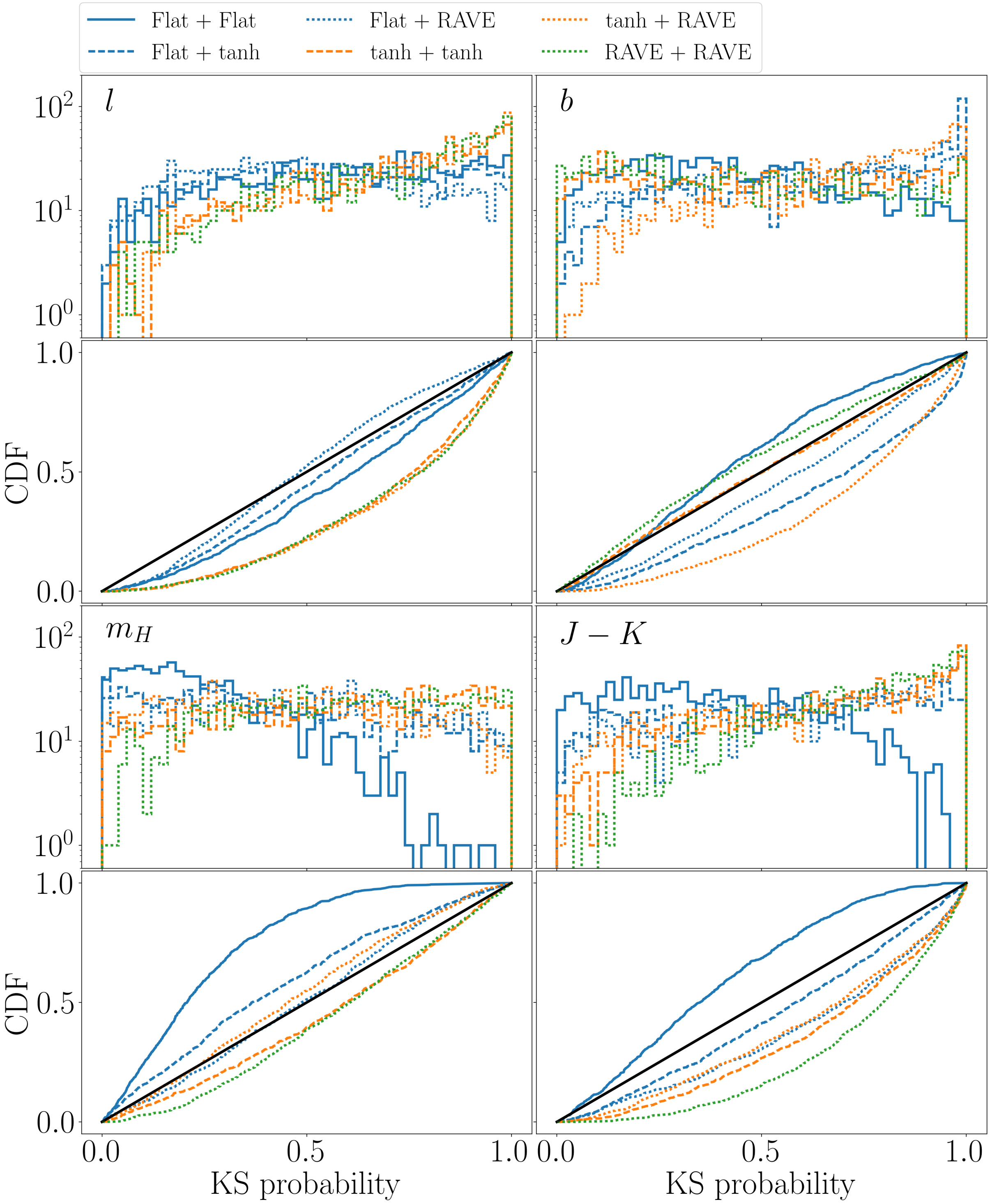


Fig. 3.9 Two sample KS tests are used for fields 1 and 2 combined as the distributions are non-analytic. 1000 sets of selection function parameters are drawn from the `emcee` chains and applied to generate subsamples of the photometric catalogue. Six combinations of the three selection functions are used for the two fields and most show close to uniform probability distributions with some weak overfitting against galactic longitude (top left) and latitude (top right). The most significant deviation is the Flat+Flat model (blue solid) which is underfit in apparent magnitude (bottom left) and colour (bottom right).

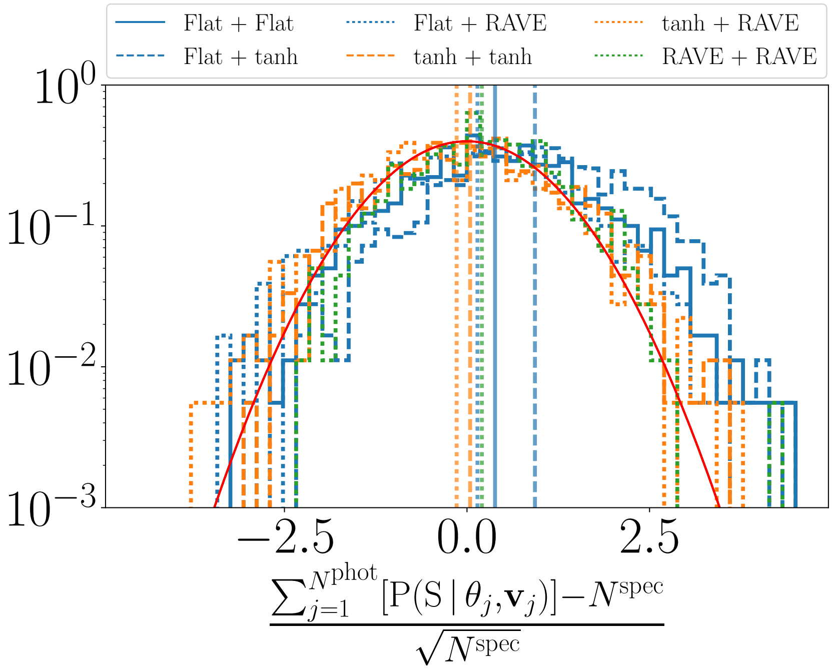


Fig. 3.10 By taking the sum of probability of selection of all stars in the photometric sample for a given selection function, I obtain an estimate of the integral over the spectroscopic distribution. All models are normally distributed with widths similar to that expected by Poisson noise of the spectrograph sample and the means of the distribution (vertical lines) are correctly centred on zero. The red solid line is a normal distribution with zero mean unit variance for comparison. Only the Flat+tanh model (blue dashed) systematically overestimates the normalisation by approximately one standard deviation.

close to uniform distributions of KS probabilities with weak under and overfitting against different coordinates.

To test the normalisation of the selection function I compare the sum of the selection function over the stars in the photometric sample with the number in the spectroscopic sample. This sum is the mean sample size which would be produced by drawing many selection function weighted samples from the photometric catalogue. The re-centred and re-scaled distributions for all field combinations are shown in Fig. 3.10. I see significantly less overfitting here with all distributions reproducing the expected Poisson noise. Only the Flat+tanh model (blue dashed) systematically overestimates the normalisation however the offset is only at the one standard deviation level so I do not consider this to be a significant issue.

These tests demonstrate the effectiveness of the union method described in Section 3.2.2 for evaluating the selection function of overlapping fields of the spectroscopic survey. This also shows that the method provides good fits to selection functions with very different properties. A caveat is that the GMM is not perfectly suited to fitting selection functions with discontinuous changes however even in these cases reasonable fits can still be achieved.

I also show here the power of my method for calculating selection functions of combined surveys. For example I could apply this method to generate a single selection function for the combined APOGEE and RAVE catalogues given their individual selection functions.

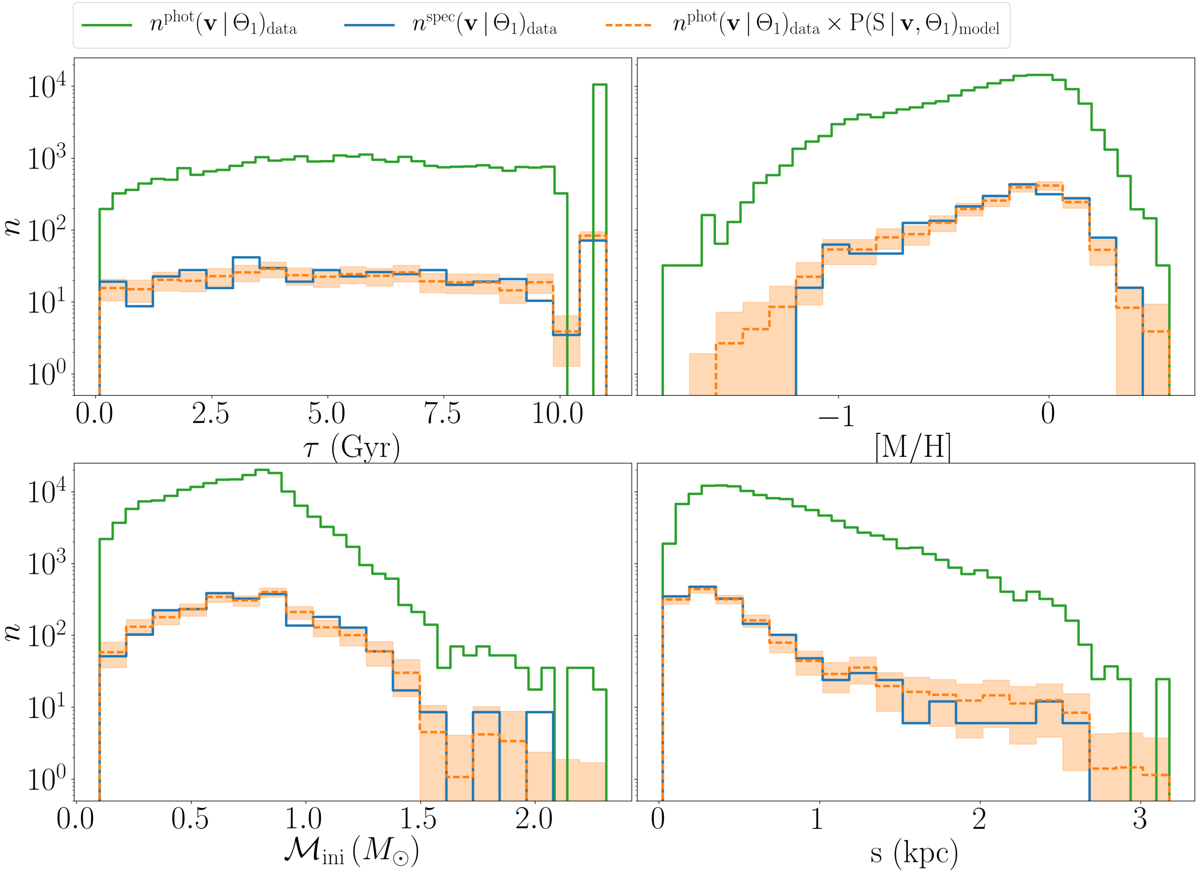


Fig. 3.11 Similar to Fig. 3.6 now for intrinsic coordinates, the distributions of stars in the photometric sample (green solid) and spectroscopic sample (blue solid) compared to the sum of selection function probabilities of the photometric sample in bins (orange dashed). For this example, field 1 and 2 have been used with tanh and RAVE selection functions applied respectively. For the majority of bins the spectroscopic sample falls within Poisson noise uncertainty (orange shaded) of the model fit.

Intrinsic Coordinates

My final test is on the selection function as a function of intrinsic coordinates (age, metallicity, mass and distance) using the mapping laid out in Section 3.2.4. I test the selection function for the two field sample with the tanh and RAVE selection functions applied to fields 1 and 2 respectively. A histogram for the model fit is generated by taking the binned sum of the selection function probabilities of the photometric sample and is shown by the orange dashed histogram in Fig. 3.11. The blue solid line shows the histogram for the spectroscopic sample and this falls within Poisson noise uncertainties of the model in the majority of bins.

Similar to Section 3.3.3, I generate 1000 mock samples from the photometric catalogue weighted by the inferred selection function and calculate the goodness of fit from the two-sample KS statistic. The distribution of KS probabilities is given in Fig. 3.12. As I saw in observable coordiantes, the distribution of KS probabilities is near uniform with very weak under fitting against distance and overfitting in initial mass and metallicity. This demonstrates that my method is also well suited to determining selection functions in intrinsic coordinates.

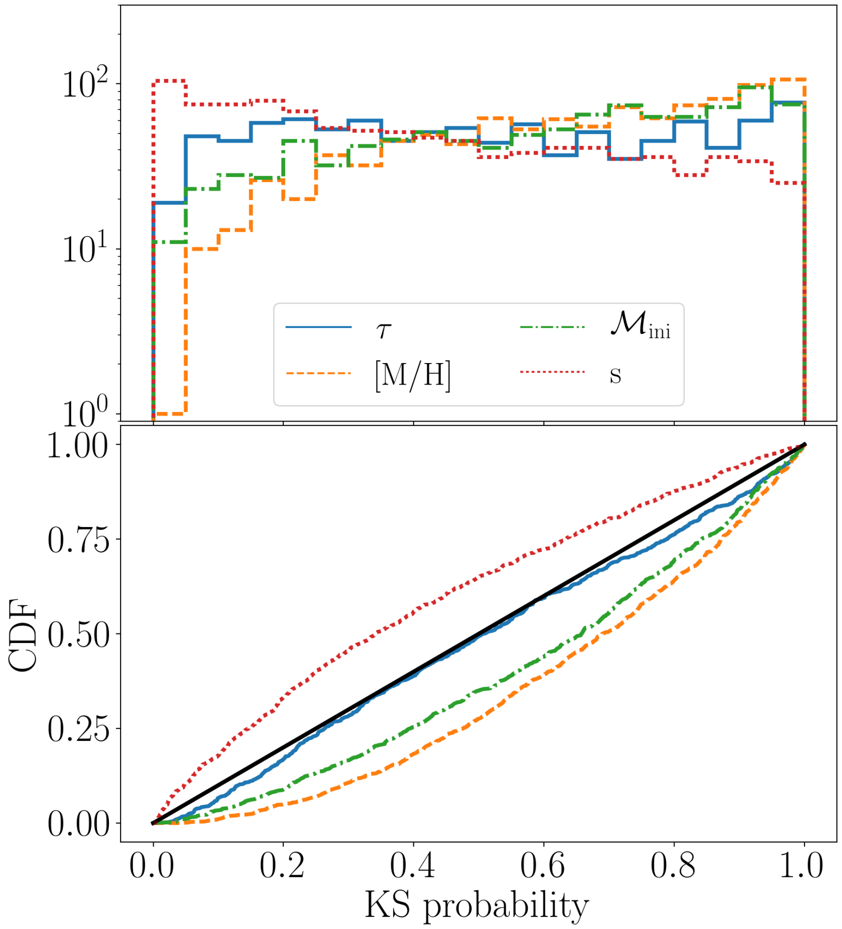


Fig. 3.12 As in Fig. 3.9, the two sample KS test is applied but here as a function of each intrinsic coordinate with the tanh and RAVE selection functions applied for fields 1 and 2. Much of the over and underfitting seen in Fig. 3.9 is averaged out in these coordinates and the models demonstrate extremely good fits to the data.

3.4 Discussion

In this chapter, I have introduced a novel method for deriving selection functions of spectroscopic surveys. I have also demonstrated the success of this method on a set of test cases using `Galaxia`. In this section, I discuss some key points to consider when applying the method, and detail potential improvements to come.

3.4.1 Choice of Photometric Catalogue

When calculating the selection function of any spectroscopic survey, a photometric survey needs to be specified which may be assumed complete in the region of the colour-apparent magnitude space explored. The choice of photometric catalogue is dependent on the characteristics of the survey. The photometric catalogue should cover the whole footprint of the spectrograph or else a combination of photometric catalogues should be used. It is beneficial to use a photometric catalogue with observing bands closely matching the spectrograph wavelength range as this enables prior information on the spectrograph's selection limitations to be applied more easily.

That said, particularly for low-latitude observations, dust attenuation is a significant factor, which suggests that infrared photometric surveys may be more appropriate. I discuss the inclusion of dust attenuation to the intrinsic selection function in Section 3.4.3.

Gaia DR2 (Gaia Collaboration et al., 2018a) represents the largest survey of the Milky Way to date and is a complete photometric survey for the magnitude ranges of many spectrographs, particularly in high latitude fields. The selection of *Gaia* DR2 as a function of l, b, G is the subject of (Boubert & Everall, 2020) and can be used to test whether it is complete in the required magnitude range.

3.4.2 Error Convolution

The selection function I have derived here is the probability of selection given measured properties of the stars. By convolving the likelihood with the measurement uncertainty of the photometries of each star, I can in principle derive the selection function given true properties of the stars. By virtue of defining the selection function as a GMM in colour-apparent magnitude space the convolution is analytic and as such the calculation is computationally feasible.

I do not present this here, but consider it as a potential avenue to pursue in the future.

3.4.3 Dust Attenuation

In this work I have not considered the impact of interstellar dust on the derived selection function. The observable selection function is dependent on dust attenuation only through its effect on colour and apparent magnitude. My inferred selection function given colour and apparent magnitude is unchanged.

Dust attenuation changes the mapping from intrinsic coordinates to observable coordinates discussed in Section 3.2.4. The effect of dust on the transformation from absolute to apparent magnitude of the star is given by

$$m_x = M_x + 5\log_{10}\left(\frac{s}{10\mathrm{pc}}\right) + A_x(l, b, s), \tag{3.29}$$

where x represents the observation band being used for apparent magnitude in the selection function. Likewise the colour will also need to be corrected for dust reddening in the mapping from intrinsic coordinates to observables.

This will be an especially important consideration for low-latitude fields. To do this, I require an adequate 3D all-sky dust map.

Bovy et al. 2016a construct a composite map that patches together maps of Marshall et al. 2006, Sale & Magorrian 2014 and Green et al. 2015. A significant improvement on this is provided by Green et al. 2019 for dec > -30 degrees.

I can include these maps in my model as an extinction and reddening term in the intrinsic to observable coordinate mapping described in Section 3.2.4.

3.5 Summary of Spectrograph Selected Sources.

I have developed a Bayesian model for empirically determining the selection function of multi-fibre spectrographs, where there exists a complete photometric survey in the same region of observable (colour-apparent magnitude) parameter space.

The method improves on previous works by modelling the selection function with a Gaussian mixture model that is fit to the data through a Poisson likelihood function. This generates a selection function which accounts for Poisson noise in low-count data. This approach also allows us to define the uncertainties in my selection function by analysing the posterior distribution on the model parameters.

I further incorporate a union calculation which allows the selection function to be calculated in regions of sky where fields partially or fully overlap. This can be applied to merged catalogues of independent surveys to produce a combined selection function. In an era where large amounts of spectroscopic data are becoming available from many independent observatories, each with their own observational limitations, combining surveys can hugely enhance our understanding of the Milky Way. I can also apply my method to subsamples of any catalogue if analysis is being done on a more specific or constrained stellar population.

Finally, I present a method of translating selection functions from observable colour-apparent magnitude coordinates into intrinsic coordinates of age, metallicity, mass and distance using the PARSEC isochrones (Bressan et al., 2012). This allows a deeper insight into the effects of the selection function on stellar parameters.

I have demonstrated the effectiveness of my method on a mock catalogue generated using the `Galaxia` (Sharma et al., 2011) population synthesis code. I am successful in reproducing the applied selection function within Poisson noise uncertainty. Using KS tests in one dimension I show that my method produces good fits to the data only struggling where the model selection functions undergo large discontinuous transitions.

My code is made publicly available as a Python repository at https://github.com/AndrewEverall/seestar.git.

3.6 End of story?

Betteridge's law states that "Any headline that ends in a question mark can be answered by the word no", that is indeed the case here. If you have made it through this chapter taking in every word and equation, I commend you, this is no mean feat. However, as I have discovered to my own pain, this is not the end of the story, nor is it the most effective way to reach a satisfactory conclusion.

For one, there is no sample which is complete in all regions of parameter space that *Gaia* observes so this method cannot be directly applied to the *Gaia* source catalogue. Fortunately *Gaia*'s observing strategy provides a far more satisfyingly simple approach which I will introduce in the next chapter.

My method of hierarchical Poisson modelling is also not necessarily the most effective way to model the selection function of a subset of a more complete catalogue. This is because I haven't used a valuable piece of information – all objects in the spectroscopic survey's catalogue are also in the complete catalogue. This is not two independent samples, one of which is smaller than the other. The smaller sample is a *subset drawn from the larger sample*. In the next chapter I will use this additional information to evaluate the selection functions of science subsets of the *Gaia* EDR3 source catalogue.

4

Completeness of the *Gaia*-verse

"There's a star map waiting in the sky... we would like you to come and meet it... it is stored in the Gaia Archive...
Let the scientists search it... let the scientists use it... let the scientists boogie..."
Gaia website

The *Gaia* archive indeed provides a map of 1,811,709,771 sources in the sky. As I discussed in previous chapters, knowing the observing limits of our data is hugely important for modelling the spatial structure of the Milky Way. In this chapter I explain how I evaluate the selection function for the *Gaia* EDR3 source catalogue and several science subsets.[1]

4.1 *Gaia* Source Catalogue Selection Function

The *Gaia* DR2 source catalogue selection function has been modelled in (Boubert & Everall, 2020). I give a brief summary of the method and explain how I have updated this selection function for *Gaia* EDR3.

4.1.1 Rolling a Six, Fifty Billion Times.

A maths teacher wants to do a class on probability distributions so she acquires a box of loaded dice. She instructs the students in the class to roll their dice as many times as they can in ten minutes and record whether each roll was a six or not.

Some industrious students manage one hundred rolls, others managed only a few. She decides that those who recorded fewer than five sixes have clearly not had enough attempts and tells them they will need to report to detention later where they can continue rolling the dice. She is left with the dice of more diligent (or luckier) pupils who managed at least five sixes. To her astonishment, over all of the dice, 95% of the rolls came out as a six. These dice really are loaded.

[1] Chapters 4 and 5 are composed of research form the *"Completeness of the Gaia-verse"* project for which I have had the privilege of working along side Dr Douglas Boubert as co-PI. Section 4.1 includes content from Boubert & Everall 2020 whilst the rest of the chapter is taken from Everall & Boubert 2021. Chapter 5 is from Everall et al. 2021c.

Assuming that all dice are equally loaded with some probability, θ of landing a six, the number of sixes each student gets would be drawn from a Binomial distribution

$$k \sim \mathrm{Binomial}(n, \theta) \tag{4.1}$$

where n is the number of rolls the student managed in the ten minute window. Therefore the likelihood of getting the k sixes measured by the student is

$$\mathrm{P}(k \,|\, n, p) = \binom{k}{n} \theta^k \, (1-\theta)^{n-k}. \tag{4.2}$$

However the teacher now realises an issue. She has removed any dice with fewer than five sixes which will bias the result, so this needs to be correct for. This is achieved by renormalising the likelihood

$$\mathrm{P}(k \,|\, n, p, k \geq 5) = \begin{cases} \frac{1}{\mathrm{P}(k \geq 5 \,|\, n, \theta)} \binom{k}{n} \theta^k \, (1-\theta)^{n-k} & k \geq 5 \\ 0 & \text{otherwise.} \end{cases} \tag{4.3}$$

This scenario is entirely analogous to the *Gaia* source catalogue. Each time a star is scanned by *Gaia* there is some probability of recording a detection[2]. A source needs least 5 detections to be included in the *Gaia* source catalogue (Lindegren et al., 2021a). I refer to the probability of a transit successfully producing an observation which is used in the astrometric solution as the *efficiency*, θ. The number of successful scans is recorded in the *Gaia* catalogue as ASTROMETRIC_MATCHED_TRANSITS and this provides the key to the *Gaia* source catalogue selection function.

4.1.2 A Thunderstorm over Madrid

ASTROMETRIC_MATCHED_TRANSITS is the number of successful transits for each source analogous to the number of sixes recorded for each dice, k. The other piece of the puzzle I need is the total number of transits, like the total number of dice rolls, n. This is helpfully provided by the scanning law which I introduced in Chapter 1.

DPAC provide the nominal scanning law for *Gaia* DR2[3] and *Gaia* EDR3 however this is only accurate to 30 arcsec. Furthermore, the satellite is not always taking data. There are three types of missing data which are important for my purposes.

- *Gaps* are stretches of time where the satellite produced no observations which were subsequently used in the astrometric solution. Causes of these include mirror decontamination, refocusing, micro meteoroid impacts and many more technical satellite issues (see Table 1 Lindegren et al., 2021a). Gaps are source-independent, no data was taken for any source in these time periods.

[2] 58,217,094,919 transits were used for the astrometric solutions in *Gaia* EDR3. That might have taken a while for our maths class rolling dice.

[3] DR2 nominal scanning law: https://www.cosmos.esa.int/web/gaia/scanning-law-pointings

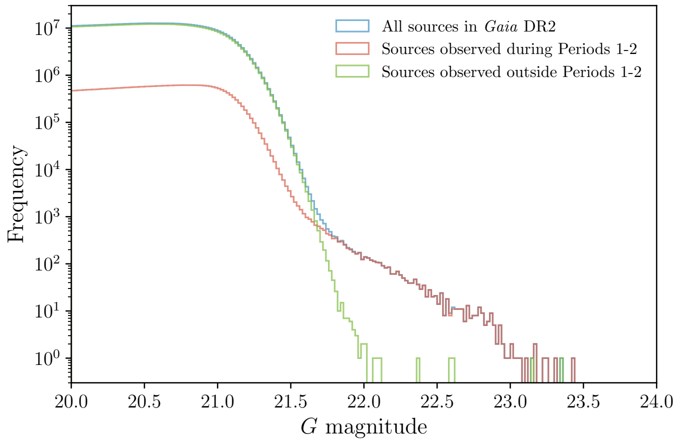
- *Deletions* take place due to memory limitations on-board the spacecraft. When the satellite continuously scans high source density regions of the sky and is not able to transmit the data to Earth fast enough, the on-board memory can run out. In this circumstance data is deleted according to a scheme which tends to prioritize brighter objects but also specific calibration magnitude bins (Section 1.3.3 of de Bruijne et al., 2018). This produces time periods with magnitude-dependent missing data.
- There were also periods of time where poor data was recorded and was less likely to be used in the final solution. These are not empty gaps but rather low efficiency periods. I discuss one of these and the cause in this section.

In Boubert et al. 2020 we used the *Gaia* DR2 epoch photometry to find the gaps as well as calibrating the scanning law using source astrometry. We performed a more detailed analysis in Boubert et al. 2021b to model the deletion periods and further calibrate scanning law to sub arcsec accuracy. Here I briefly detail one problematic time window in *Gaia* DR2 which demonstrates the challenges associated with management of such an impressive mission.

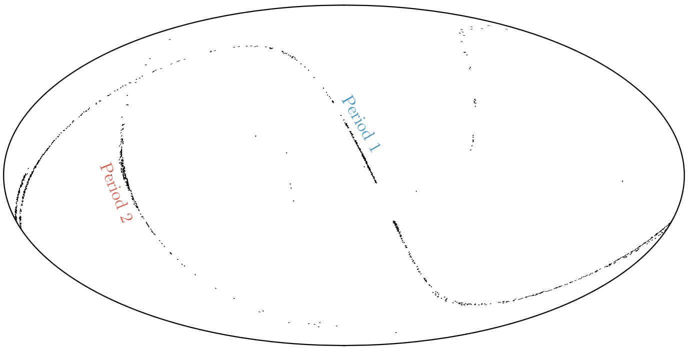
99.865% of the sources in *Gaia* DR2 are brighter than $G = 21.3$ (Gaia Collaboration et al., 2018a), but there is a tail of fainter sources out to $G = 23.5$ which is shown in the top panel of Fig. 4.1. These sources are likely to be spurious because *Gaia* is not sufficiently sensitive to detect sources this faint in the short time in which sources transit the focal plane. The drop-off in the number of sources at each magnitude shows a change in behaviour at around $G = 21.7$ and I conjecture that the magnitudes of most of the sources fainter than $G = 22$ are likely to be spurious. The on-sky distribution of the 1 869 stars fainter than $G = 22$ in Galactic coordinates are shown in the middle panel of Fig. 4.1. Almost all of these sources lie along two narrow strips and thus can be attributed to specific periods of the *Gaia* scanning law. Using a new tool developed in Holl et al. (in prep.) for this use-case, I identified two rough time ranges which I label Period 1 (OBMT = 1388−1392 rev) and Period 2 (OBMT = 2211−2215 rev), and I illustrate where *Gaia* was scanning during these periods in the bottom panel of Fig 4.1. Curiously, Period 1 aligns with the drop in the colour photometry efficiency mentioned in Section 2.3 of Boubert et al. 2020.

We queried the *Gaia* Helpdesk about this period and the following italicised text is an abridgement of their response.

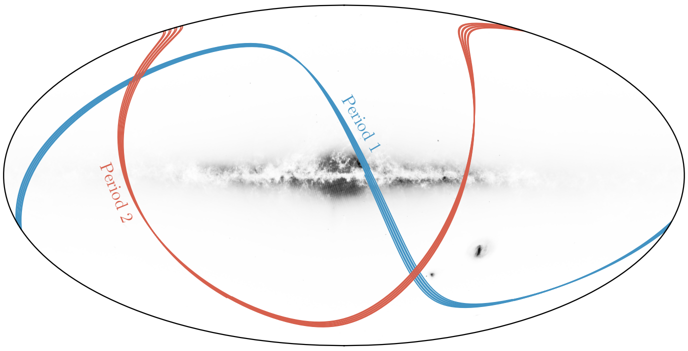
On Saturday evening 11 October 2014, around 18:55 UTC, while Gaia was transmitting data to the Cebreros ground station near Madrid, a thunderstorm developed over the Madrid sky-line and heavy rain started falling. As a result, contact with the spacecraft was lost until 19:04 UTC. However, during these 9 minutes, Gaia kept on transmitting its data to ground not knowing it would not be recorded. Whereas the bulk science data transmitted during this short interval was permanently lost, so-called critical auxiliary science data (ASD) packets that were lost were re-transmitted to ground the following day. This, however, was too late to use these packets in the regular, semi-live initial data treatment (IDT), which forms the first step in the astrometric and photometric (pre-)processing chains. As a result of the missing data, critical background information has been absent in the Gaia DR2 photometric



(a) Faint magnitude distribution of all sources in *Gaia* DR2 (blue), of sources predicted to have been observed at least once during the two time periods identified in the main text (red), and of those sources with no predicted observation during those periods (green). Almost all of the extremely faint sources can be traced to those time periods.



(b) Galactic coordinate distribution of stars fainter than $G > 22$. Almost all of the stars lie along two narrow tracks which correspond to the strips of sky observed by *Gaia* during small time windows.



(c) I identified the time windows corresponding to each of the two strips and show where *Gaia* was looking during these windows.

Fig. 4.1 There are sources in *Gaia* with reported mean G-band magnitudes as faint as $G = 23.5$. I identified that an overwhelming majority of these sources were observed during two narrow time windows and so are likely due to missing calibration data packets preventing an accurate magnitude determination.

processing for faint, one-dimensional windows ($G \geq 13$) for short stretches of time. The affected intervals (in OBMT revolutions) are 1389.7-1391.7 for row 2, 1389.2-1391.7 for row 3, 1389.2-1391.7 for row 4, 1389.2-1391.7 for row 5, 1389.2-1391.7 for row 6, and 1389.2-1391.3 for row 7; row 1 was not affected. The Gaia DR2 photometric calibration has "solved" the absence of background information by linearly interpolating between existing data (see Section 4.2 in Riello et al., 2018). This interpolation has, in this case, not worked perfectly and has failed to catch several straylight-induced peaks in the background. As a result, the photometry collected during these few revolutions is systematically biased and not reliable. In fact, the entire stretch from OBMT 1388.0 to 1392.0, which corresponds to the relevant calibration time interval, is indirectly affected by this issue. On the bright side: for Gaia (E)DR3, there is hope that this issue will be gone. Not only will gaps at IDT level have been fixed by the raw data reprocessing that has been undertaken, there has also been an update to the computation of the local background and this new feature should perform significantly better in periods with missing data.

In summary, the miscalibration is caused by a break of communication with the spacecraft which resulted in the loss of some scientific data and the delay in transmission of critical auxiliary science data (ASD) which was then not used in *Gaia* DR2 processing. There were several stray light peaks within the down-time which were therefore missed by the calibration and instead interpolated over. This directly affected measured fluxes of many of the sources observed in the range OBMT=1389.2-1391.7, which resulted in G_{BP} and G_{RP} observations not being included in the epoch photometry as seen in Fig. 8 of Boubert et al. 2020. *Gaia*'s photometric calibration occurs in one day time intervals which in this case corresponds to the interval OBMT = 1388.0−1392.0 rev. Those observations within the range OBMT = 1389.2−1391.7 rev which were directly affected received overestimated fluxes due to the unobserved stray-light peaks. Observations within the calibration interval but not directly affected by the down-time received underestimated fluxes due to the calibration process effectively averaging out the observed error. It is these stars which appear as extremely dim sources in Period 1. Period 2 is also likely a result of background interpolation issues however I do not have an explanation of the exact cause of this particular event.

4.1.3 Selection Function Probability

I now have all I need to evaluate the source catalogue selection function. For each source the number of transits can be estimated by the number of times *Gaia* scanned the source's position on the sky. The number of these transits which are successful observations, k is given by ASTROMETRIC_MATCHED_TRANSITS.

Equation 4.3 gives the likelihood of k observations but I want the probability distribution for observation efficiency, θ. So I apply Bayes' theorem

$$\mathrm{P}(\theta \,|\, k, n, k \geq 5) \propto \mathrm{P}(k \,|\, \theta, n, k \geq 5)\, \mathrm{P}(\theta). \tag{4.4}$$

In Boubert & Everall 2020 we show that the Beta prior is appropriate for this task and show the posterior efficiency as a function of apparent magnitude in the top panel of Fig. 5 in that paper.

Given θ, evaluation of the selection function is incredibly simple. The probability of a source being included in the *Gaia* source catalogue is the probability that at least five detections are made

$$\mathrm{P}(\mathcal{S}_{\mathrm{source}}\,|\,n,\theta) = \sum_{k=5}^{n} \mathrm{P}(k\,|\,n,\theta). \tag{4.5}$$

In Boubert & Everall 2020 we make this more sophisticated by modelling the parameters of the Beta distribution which gives the distribution of θ rather than directly modelling θ. The final selection probability is then marginalised over θ for the given apparent magnitude. The results are shown in the lower panel of Fig. 5 in Boubert & Everall 2020.

One final complexity to add is that I neglect the impact of crowding on *Gaia* observations. In crowded regions, *Gaia* is less likely to detect sources due to the challenges in assigning independent windows on-board the space craft. We applied the model described above at multiple crowding levels in Section 4 of Boubert & Everall 2020. The impact on the inferred efficiency is shown in Fig. 7 and the final results as a function of position on the sky are given in Fig. 8.

4.1.4 *Gaia* EDR3 Source Catalogue

Whilst 2020 will be remembered for some events more than others, one thing the year did provide us with was the early third data release of the *Gaia* satellite, EDR3.

The method presented for the *Gaia* DR2 source catalogue selection function required precise predictions of the number of occasions on which the source could have been detected from the scanning law. The scanning law was calibrated using the *Gaia* DR2 epoch photometry (Boubert et al., 2020, 2021b). However epoch photometry for the EDR3 baseline will not be published until the full data release in 2022 which prevents me from calibrating the EDR3 scanning law and repeating the method previously used for DR2.

Work is currently underway to produce a significantly more impressive selection function for the EDR3 source catalogue (Boubert et al., 2021a; Fraser et al., 2021). For now, I provide a simple estimate for the EDR3 selection function based on Boubert & Everall 2020 but emphasise that this will be superseded when the new results are published.

The *Gaia* EDR3 source catalogue contains about 7% more sources than *Gaia* DR2. Fig 4.2 shows the ratio of source density across the sky between EDR3 and DR2. A striking feature of this is that the increase in source density follows patterns of the scanning law. Regions of the sky which were under-scanned in DR2 but received more transits in the extra 14 months of data to EDR3 have the biggest increase in content.

Motivated by this, I propose a very simple adjustment to evaluate the EDR3 selection function. I assume that the detection efficiency *Gaia* EDR3 is the same as in DR2. Therefore I use the results from Fig. 7 of Boubert & Everall 2020 to evaluate the detection efficiency as a function of apparent magnitude and source density. The number of times a

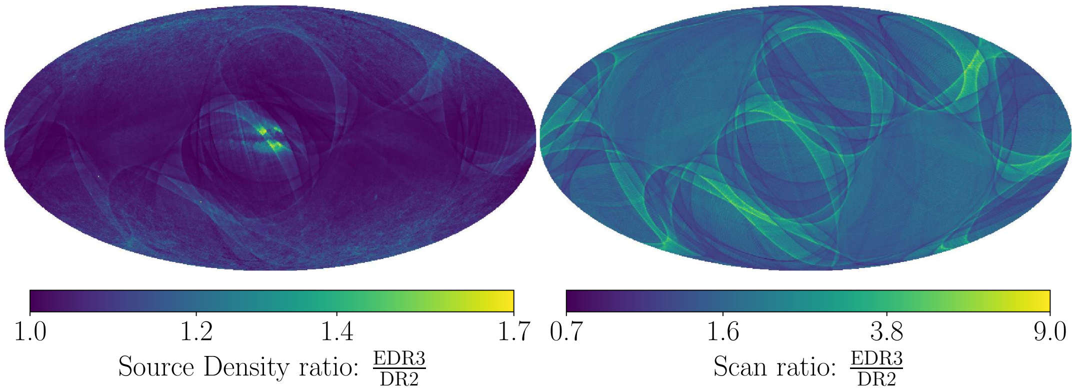

1.0
1.2
1.4
1.7
Source Density ratio: $\frac{\text{EDR3}}{\text{DR2}}$
0.7
1.6
3.8
9.0
Scan ratio: $\frac{\text{EDR3}}{\text{DR2}}$

Fig. 4.2 The increase of sources in the *Gaia* source catalogue between DR2 and EDR3 is driven by new scans. In the top panel I show the ratio of source density in HEALPix pixels between EDR3 and DR2 showing clear structure of the scanning law. The bottom panel is the approximate ratio of number of scans between the data releases using the calibrated DR2 scanning law and nominal scanning law for EDR3. Regions which received few observations in DR2 with many more in EDR3 saw large increase in number of objects in the *Gaia* source catalogue.

source was transited is evaluated from the uncalibrated *Gaia* EDR3 nominal scanning law. Therefore the selection function is (Eq. C1, Boubert & Everall, 2020)

$$\mathrm{P}(\mathcal{S}_{\mathrm{source}}\,|\,G,l,b) = 1 - \sum_{m=0}^{4} \binom{n}{m} \frac{\mathrm{Beta}(A+m, B+n-m)}{\mathrm{Beta}(A,B)} \tag{4.6}$$

where A, B are the DR2 Beta distribution parameters of the efficiency as a function of source density and apparent magnitude. n is the number of times a source was transited in EDR3 according to the EDR3 scanning law.

The numbers of transits in DR2 and EDR3 from (Boubert et al., 2021b) and the EDR3 nominal scanning law respectively are shown in the top panels of Fig. 4.3. In the bottom panels I show the selection function probability at $G = 21$. The selection probability has increased across the majority of the sky but most significantly in the ecliptic plane regions where there were few transits in DR2 such as the West side of the Galactic bulge.

This result provides a first estimate of the *Gaia* EDR3 selection function but should be used with caution. There are three limitations to this model.

- I have assumed that the pipeline of DR2 and EDR3 will lead to the same source detection efficiency in both catalogues. With each new data release, the data processing pipeline is rerun with improved source calibration which can change whether individual source observations are used in the astrometric solution.
- In *Gaia* DR2, when two sources were observed to be within 0.4 arcseconds of one another, the lower priority source was removed from the source catalogue. In EDR3, this threshold was reduced to 0.18 arcseconds (Gaia Collaboration et al., 2021a; Torra et al., 2021) so EDR3 will be more complete in crowded regions independent of the scanning law.

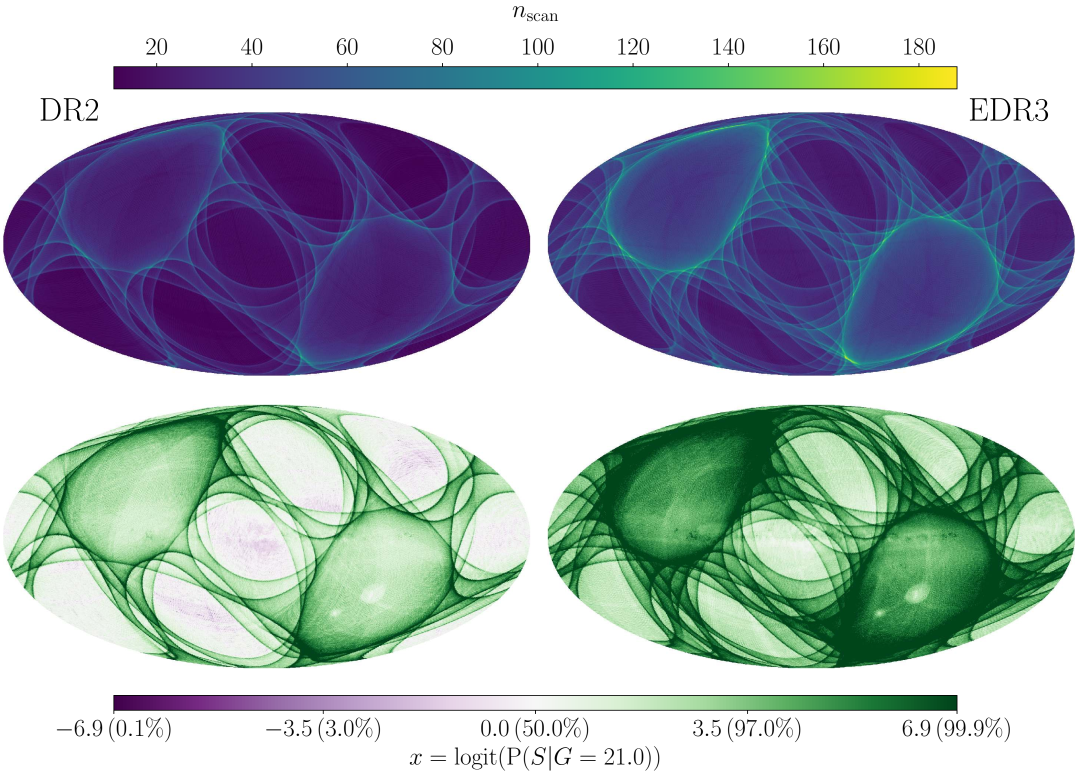


Fig. 4.3 A simple estimate of the *Gaia* EDR3 selection function is achieved by updating the DR2 selection function with the nominal scanning law for EDR3. In the top row I show the number of scans received as a function of position on the sky in Galactic coordinates for DR2 (left) and EDR3 (right). The inferred *Gaia* source catalogue selection function probability at $G = 21.0$ is shown in the bottom panel where regions which have received more scans in EDR3 have significantly higher selection probability.

- The EDR3 nominal scanning law has not been calibrated directly against the data and will deviate from *Gaia*'s true scanning history by up to 30 arcseconds at any point in time. Therefore the predicted observations in this model will be marginally off the true numbers over small regions of sky.

Details on how to access the *Gaia* EDR3 source catalogue selection function are provided in Section 4.8.

4.2 Science Subsets

The *Gaia* DR2 selection function, modelled in Boubert & Everall 2020, has enabled us to answer the question "What is the probability that a source was observed and recorded in the *Gaia* DR2 source catalogue?". However, additional information is provided for subsets of objects in the *Gaia* source catalogue which are vital for answering certain scientific questions. For example, the sample with measured parallax and proper motion (Lindegren et al., 2018, 2021a) is enabling us to study and understand the structure and kinematics of the Milky Way, whilst the sample with radial velocities (Gaia Collaboration et al., 2018b) includes the final dimension of kinematic data for a smaller, but no less impressive, set of objects.

Usually selection functions are estimated by comparing the sample to a more complete source catalogue. For the *Gaia* source catalogue, there is no more complete sample to compare against. Therefore we had to use our understanding of the observing strategy of *Gaia* to model the selection function. However, for subsets of *Gaia*, the source catalogue provides a sample which we know is more complete in all areas of parameter space. Previous attempts at selection functions for samples where a more complete catalogue is available involve taking the ratio of objects in the subset to source catalogue in carefully chosen on-sky position-colour-apparent magnitude bins (Bovy et al., 2012b; Mints & Hekker, 2019; Wojno et al., 2017, e.g.). This requires a selection function which can be defined discretely with enough objects in each bin of the subset and source catalogues such that the ratio is approximately equal to the expected completeness. This is not the case for subsets of *Gaia* which have complex selection criteria within the processing pipeline relating to satellite performance and the scanning law (Boubert et al., 2020, 2021b).

A strong example of this is the *Gaia* DR2 radial velocity catalogue (which I will address in this chapter). Rybizki et al. 2021b attempted to evaluate the selection function for this sample relative to the *Gaia* source catalogue using number count ratios in colour-magnitude-sky bins. However, as they demonstrate, the selection function has a strong dependence on sky position over small scales due to the scanning law, crowding limitations and the Initial *Gaia* Source List (IGSL, Smart & Nicastro, 2014). Poisson noise prevented them from going to sufficient resolution to model this structure without running out of objects in *Gaia*.

In Chapter 3 I went beyond simple count ratios by modelling the source catalogue and subset CMDs as continuous Poisson count processes with Gaussian Mixture Models.

Whilst this model was well-suited to multi-fibre spectrographs with well defined fields, a significant improvement was needed to fit the all-sky selection function of *Gaia* catalogues.

This major step forward has come from Boubert & Everall 2021 which presents a general approach to estimating selection functions for subsets of catalogues. We model the selection function as a sum of well-localised spherical needlets on the sky with weights drawn from a Gaussian Process across colour and apparent magnitude. We also use the formally correct Binomial likelihood function to fit the model.

In this work I will apply the Boubert & Everall 2021 model to three subsets of *Gaia* EDR3:

(i) Astrometry: the sample with published parallax and proper motion.[4]

(ii) RUWE: Renormalised unit weight error specifies the goodness of fit of an astrometric solution and is often used to filter out poorly fit sources. I provide selection functions for the sample with RUWE < 1.4.

(iii) RVS: The sample with published radial velocities from the radial velocity spectrograph (RVS).

I provide a brief review of the model developed by Boubert & Everall 2021 in Section 4.3 and explain how binomial statistics are used to fit the likelihood function and test the results in Section 4.4. Readers primarily interested in the results may wish to skip this and go directly to Section 4.5 where the selected samples are described. I present my results in Section 4.6. Before concluding I also discuss other *Gaia* catalogues which have not been investigated here due to additional complexities but which should be a key goal for future work in this field.

4.3 Methodology

An in-depth description of the method for constructing sub-sample selection functions for astronomical surveys is given in Boubert & Everall 2021. Here I provide a brief summary of the key points which I will apply in this chapter.

My aim is to estimate the selection probabilities of subsets of the *Gaia* EDR3 relative to the source catalogue. e.g. ‘What is the probability that a source in *Gaia* EDR3 has parallax and proper motion?’. This selection function may be written down as

$$\mathrm{P}(\mathcal{S}_{\mathrm{subset}}\,|\,\mathcal{S}_{\mathrm{source}},\mathbf{y}) \tag{4.7}$$

where $\mathbf{y}$ are the observables (or functions of observables) over which the selection function is defined and ‘source’ refers to the more complete catalogue, in this case the *Gaia* source catalogue. The probability of any object being included in the subset is

$$\mathrm{P}(\mathcal{S}_{\mathrm{subset}}\,|\,\mathbf{y}) = \mathrm{P}(\mathcal{S}_{\mathrm{subset}}\,|\,\mathcal{S}_{\mathrm{source}},\mathbf{y})\;\cdot\;\mathrm{P}(\mathcal{S}_{\mathrm{source}}\,|\,\mathbf{y}) \tag{4.8}$$

[4] In *Gaia* EDR3 this is the combined sample with 5D or 6D astrometric solutions. I note that sources without parallax and proper motions in *Gaia* still received 2D astrometric solutions however I refer to the sample with 5/6D solutions as the astrometry sample for brevity.

where the source catalogue selection function, $\mathrm{P}(\mathcal{S}_{\mathrm{source}}\,|\,\mathbf{y})$, is the probability that a source is included in the *Gaia* source catalogue which I introduced in Section 4.1.4.

All selection functions estimated here will be over the variables $\mathbf{y} = (l, b, G)$ with the RUWE and RVS selection functions additionally a function of $G - G_{\mathrm{RP}}$. These are the dominant variables in selecting sources to enter the samples. The scanning law produces complex selection patterns across the sky generating a heavy dependence on (l, b). Whether individual observations are used in the *Gaia* data is dependent on the on-board estimated apparent magnitude of the source (scientific measurements of a source are only made if $G_{\mathrm{onboard}} < 20.7$, Gaia Collaboration et al., 2016), for which G is a reasonable proxy. The publication of measured radial velocity for any source in *Gaia* is contingent on an estimated RVS magnitude (G_{RVS}) calculated using the IGSL (See Section 2.1 in Sartoretti et al., 2018). As will be discussed later, G_{RVS} is more similar to G_{RP} than G which means the magnitude limit in G will be a function of $G - G_{\mathrm{RP}}$. Finally, the selection criteria for the RVS sample is explicitly dependent on source temperature and RUWE is implicitly dependent on colour. Making my selection function dependent on $G - G_{\mathrm{RP}}$ allows me to capture these dependencies. I choose to use $G - G_{\mathrm{RP}}$ rather than $G_{\mathrm{BP}} - G_{\mathrm{RP}}$ due to calibration issues at the faint end of G_{BP} (Riello et al., 2021). The drawback of this is that the G_{RP} uses larger spatial windows than G which means that our colours will be extremely red for extended sources with high excess flux (see Section 9.4 Riello et al., 2021).

The selection function, as described in Boubert & Everall 2021, is composed of a sum of spherical needlets across the sky with coefficients subject to a Gaussian Process prior in apparent magnitude and colour. This is fit to the data using a Binomial likelihood function in HEALPix-apparent magnitude-colour bins. In the following subsections, I briefly describe the maths of each of these components.

4.3.1 Logit Probability

My model will estimate the probability, $q \in [0, 1]$ that an object will be included in the subset given the observables and that it is in the source catalogue. My model is defined in an infinite domain so I directly model $x \in [-\infty, \infty]$, the logit-transformed probability

$$x = \mathrm{logit}(q) = \log\left(\frac{q}{1-q}\right). \tag{4.9}$$

This is then inverse transformed to retrieve the selection probability

$$q = \mathrm{logit}^{-1}(x) = \frac{1}{1+\exp(-x)} \tag{4.10}$$

which is also referred to as the 'expit' function. All figures in this chapter are given in terms of x but I provide values of q in the axes to help with interpretation.

4.3.2 Needlets

Spherical harmonics provide an orthonormal basis for functions on the sphere and so are commonly-used to model distributions over a spherical surface. However, since individual spherical harmonics are not localised on the sky, a small change in the function at one location changes the coefficients of all the harmonics. This can be overcome by using a convolution of spherical harmonics called a 'needlet' (Marinucci et al., 2008)

$$\psi_{jk} = \sqrt{\lambda_{jk}} \sum_{\ell=0}^{\ell_{\max}} b_\ell(j) \frac{2\ell+1}{4\pi} P_\ell(\cos(\phi_{jk}(l,b))), \tag{4.11}$$

where P_l are the Legendre polynomials and $\phi_{jk}(l,b)$ is the great arc separation between the coordinates (l,b) and the Needlet centre. Localisation of the Needlets is achieved through the window function, $b(j)$. For this I use the 'Chi-square' Needlets[5] described in Geller & Mayeli 2010 and Scodeller et al. 2011

$$b_\ell(j|B,p) = \left(\frac{j}{B^j}\right)^{2p} \exp\left(-\frac{\ell^2}{B^{2j}}\right). \tag{4.12}$$

In order to satisfy the reconstruction formula for spherical Needlets, the window function must satisfy (Baldi et al., 2006)

$$\sum_{j=0}^{\infty} b(j)^2 \equiv 1. \tag{4.13}$$

This is not generally true for the window function in Eq. 4.12 so I numerically renormalise the window functions for each l summing for j up to 1000 to guarantee the relation is satisfied.

The needlets are centred on HEALPix pixels (Górski et al., 2005) such that

$$x = \sum_{j=0}^{j_{\max}} \sum_{k=0}^{N_j} \beta_{jk}\, \psi_{jk}(l,b) \tag{4.14}$$

where $j_{\max}$ is the maximum HEALPix level used and $N_j = 12 \times 2^{2j}$ is the number of pixels in a given HEALPix level.

The selection probabilities as a function of position on the sky are then given by $q = \text{logit}^{-1}(x)$.

4.3.3 Gaussian Process Prior

The free parameters β_{jk} are modelled as a function of apparent magnitude G and colour $C = G - G_{\text{RP}}$. This is performed in discrete colour-magnitude bins with a Gaussian Process prior placed on each dimension independently

$$\{\beta_{jk}\}_{mc} \sim \mathcal{GP}\left(\{G\}_m, \{C\}_c\right) \tag{4.15}$$

[5] Referred to in those works as 'Mexican' Needlets.

where I use an independent Squard Exponential kernel for each of apparent magnitude and colour,

$$K(G, C, G', C') = \sigma^2 \exp\left(-\frac{(G - G')^2}{2l_m^2}\right) \exp\left(-\frac{(C - C')^2}{2l_c^2}\right), \tag{4.16}$$

with apparent magnitude and colour length-scales l_m and l_c.

4.4 Binomial Statistics

The selection function fit is based on the Binomial likelihood. For each bin, the number of sources in the sub-sample k is assumed to be drawn from a Binomial distribution, given the number of sources which could have been selected from the source catalogue in the given bin n and the selection probability q.

In this section I briefly revise the likelihood function and describe the Beta-Binomial expected value and p-value test which will be used for analysing the results.

4.4.1 Binomial Likelihood

The source catalogue and subset are counted in HEALPix-colour-apparent magnitude bins. The overall likelihood is given by

$$\begin{aligned} \mathcal{L} &= \prod_{i=p,m,c} \mathrm{Binomial}(k_i \,|\, n_i, q_i) \\ &= \prod_{i=p,m,c} \binom{n_i}{k_i} q_i^{k_i} (1 - q_i)^{n_i - k_i} \end{aligned} \tag{4.17}$$

where p, m, c is the HEALPix-magnitude-colour bin index and q_i is the model probability at the bin centre (Eq. 4.14). Since the Binomial coefficient is independent of the selection probability, this can be dropped out as a constant and the log likelihood simplifies to

$$\log \mathcal{L} \sim \sum_{i=p,m,c} k_i \log(q_i) + (n_i - k_i) \log(1 - q_i) \tag{4.18}$$

To optimize the likelihood function in terms of model parameters, β_{jk}, I use the L-BFGS-B algorithm (Zhu et al., 1997) implemented in SCIPY. The boundaries are placed at $[-50, 50]$ for the unscaled parameters for which the prior distribution is a unit variance Gaussian. In other words, a parameter would have to be a 50-sigma outlier from the prior to reach the optimization boundaries.

4.4.2 Beta-Binomial Posterior

For each bin, I can independently estimate the posterior distribution of q which will be useful when testing the results. This is not used to fit my model but instead to help understand the results.

Consider the stars in one bin to be like identical marbles in a bag. You get handed the marbles one by one and choose whether keep the marble or give it away. Eventually, of the original bag of n marbles, you're left with k in your hand. What is the expected probability of any marble being selected?

The Binomial distribution gives the likelihood of choosing k marbles given the selection probability q,

$$\begin{aligned}\mathrm{P}(k \mid n, q) &= \mathrm{Binomial}(k|n, q) \\ &= \binom{n}{k} q^k (1-q)^{n-k}.\end{aligned} \tag{4.19}$$

To evaluate the posterior probability distribution of q, I apply Bayes theorem but first I need a prior. The Beta distribution is mathematically sensible as it is the conjugate prior of the Binomial distribution. Even more appealing, a Beta distribution with $\alpha_0 = \beta_0 = 1$ is equivalent to a uniform distribution, $\mathrm{U}[0, 1]$. The posterior probability distribution is

$$\begin{aligned}\mathrm{P}(q \mid n, k) &= \frac{\mathrm{Binomial}(k|n, q)\,\mathrm{Beta}(q|\alpha_0, \beta_0)}{\mathrm{P}(k \mid n)} \\ &= \mathrm{Beta}(q|k + \alpha_0, n - k + \beta_0).\end{aligned} \tag{4.20}$$

This is the Beta-Binomial distribution with an expected q of

$$\mathbb{E}[q] = \frac{\alpha}{\alpha + \beta} \tag{4.21}$$

$$= \frac{k + \alpha_0}{n + \alpha_0 + \beta_0}. \tag{4.22}$$

For a uniform prior this equates to $(k+1)/(n+2)$. This formula is more commonly known as the "rule of succession" and was written down by Laplace more than two hundred years ago (Laplace, 1921).

This might not be quite what one expects. Naively k/n is often used as the expected value of the selection probability given k objects drawn from a sample of n. It is worth noting that the expected value is $1/2$ when $k = n = 0$, which is the expected value of a uniform distribution. If there are no stars in a bin then I have no information to work with and the posterior reverts to the prior. This can be seen happening in the brighter bins in Figs. 4.9 and 4.12.

4.4.3 p-value Test

I can test the veracity of a selection probability model using a p-value test (in this case I use a one-tailed p-value). This answers the question 'Given the model, what is the probability that a measurement of an observable would not be larger than the given value?'. For a Binomial distribution, the question is 'What is the probability that less than k sources out of n are observed in this bin given the bin's selection probability?'. The Binomial p-value is given by

$$P \sim \mathrm{U}\left[\sum_{0}^{k'=k-1} \mathrm{Binomial}(k' \,|\, n, q), \sum_{0}^{k'=k} \mathrm{Binomial}(k' \,|\, n, q)\right]. \tag{4.23}$$

which I explain in more detail in Appendic A.3. Note that p-value is not a deterministic value but a stochastic variable. For example, if I have no data in a bin ($k = n = 0$) the p-value simplifies to $P \sim \mathrm{U}\,[0, 1]$.

As in any one-tailed p-value test, if the model has successfully reproduced the data, the set of p-values for all data points will be uniformly distributed between 0 and 1. This test will be used with all of my inferred selection functions to check that they have accurately captured the information in the bins.

4.5 Data

The selection function is estimated for three samples: Astrometry, RUWE and RVS. Here I provide a brief description of each sample and how it can be accessed from the *Gaia* archive.

4.5.1 Astrometry

The astrometry sample is the subset of *Gaia* EDR3 with published parallax and proper motions from the Astrometric Global Iterative Solution (AGIS, Lindegren et al., 2012). There are three criteria for a source having published parallax and proper motion (Lindegren et al., 2021a):

- $N_{\mathrm{VPU}} \geq 9$
- $G_{\mathrm{DR2}} < 21.0$
- $\sigma_{\mathrm{5Dmax}} < 1.2\mathrm{mas} \times \gamma(G)$.

The VPU (VISIBILITY_PERIODS_USED) cut depends on the number of observations a source has received and their distribution in time. This produces a strong dependence of the selection function on position on the sky. The apparent magnitude cut is performed on DR2 apparent magnitudes and all magnitudes have been recomputed in EDR3. Therefore the G_{DR2} cut doesn't produce a discontinuous change in the selection function with G however it will still lead to a distinct drop-off towards the faint end. Finally, the cut on σ_{5Dmax} (ASTROMETRIC_SIGMA5D_MAX) will depend on position on the sky and apparent magnitude, because both contribute to the expected astrometric uncertainty for any source (I discuss the individual effect of the σ_{5Dmax} cut in Chapter 5). The effects of these cuts will be empirically modelled in this chapter without directly considering the effects of each cut on their own.

The distribution of sources as a function of colour and apparent magnitude is shown in Fig. 4.4 and as a function of position on the sky in Fig. 4.5. The SQL query for accessing the data is as follows.

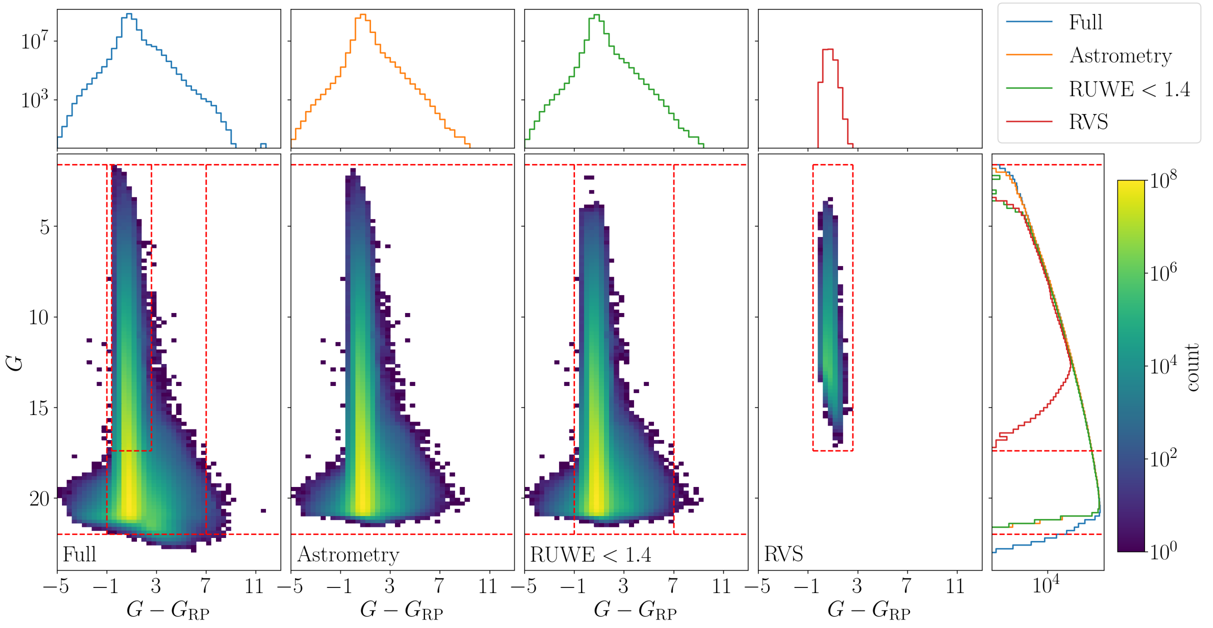


Fig. 4.4 The number of *Gaia* EDR3 sources in each colour-magnitude bin is shown for the full sample (left) astrometry sample (middle left) astrometry with RUWE < 1.4 sample (middle right) and sample with DR2 radial velocity pubilshed in the EDR3 catalogue (right). The astrometry selection removes sources from the dim end whilst RVS is cut on G_{RVS} producing an extended drop-off in G as can be sen from the histogram on the right. Red dashed lines show the region of parameter space used to fit the selection functions.

```
SELECT *
FROM gaiaedr3.gaia_source
WHERE
    astrometric_params_solved>3
AND phot_g_mean_mag between 1.6 and 22
```

There are no sources in *Gaia* EDR3 with apparent magnitude brighter than $G = 1.6$ and none with parallax or proper motion fainter than $G = 22$. This results in a sample of 1 465 211 050 sources out of a total 1 806 195 366 in *Gaia* EDR3 with $1.6 < G < 22$.

4.5.2 RUWE

When using *Gaia* astrometry, various cuts are often placed on the sample to generate a subset with maximal information and minimal systematics. One such cut recommended by DPAC is RUWE < 1.4 (Gaia Collaboration et al., 2021a). Sources with large RUWE have scatter between astrometric measurements which is poorly fit by the linear astrometry model. A common cause of this is binary motion (Belokurov et al., 2020b; Lindegren et al., 2018; Penoyre et al., 2020) however this can also be generated by source contamination in crowded regions and possibly even astrometric microlensing (McGill et al., 2020). Whilst these sources can be astrophysically interesting, they introduce systematic errors in the AGIS pipeline and it is recommended to remove the more extreme cases in order to clean the sample.

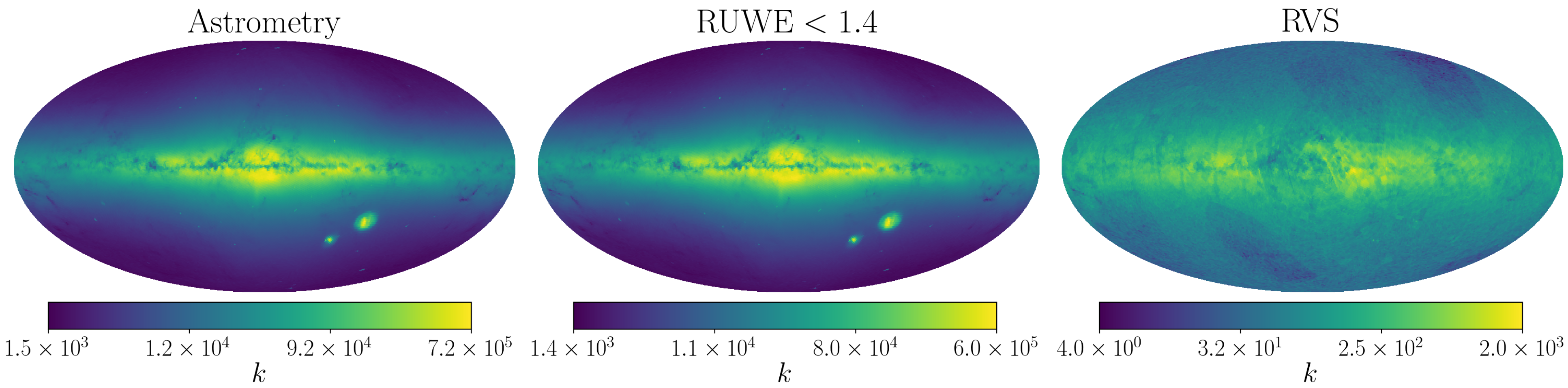


Fig. 4.5 The counts of each sample in HEALPix level 6 bins across the sky summed over the colour-apparent magnitude range used for the selection function. The RUWE sample shows slightly reduced star counts in the highest density regions compared with the Astrometry and RVS is significantly more spatially uniform due to the RVS spectrograph's inability to collect independent spectra for low separation sources. The RVS star counts also show features of the scanning law and vertical stripes in the East and squares in the Equatorial South deriving from the SDSS and Schmidt photographic plate contributions to the Initial *Gaia* Source List (Rybizki et al., 2021b).

There have been other recommended cuts for removing spurious astrometric solutions on IPD_GOF_HARMONIC_AMPLITUDE (Fabricius et al., 2021) or ASTROMETRIC_GOF_AL (Lindegren et al., 2021a) however RUWE < 1.4 is commonly used in the community so I keep to this single cut here.

Whilst an ideal telescope would have an entirely symmetric point spread function, it is slightly asymmetric on the CCD panel of *Gaia* introducing chromatic behaviour to the inferred centroid location (for full discussion see Lindegren et al., 2021a, Section 2.3). This propagates through to the astrometric solution where the unit weight error has a colour dependence. RUWE is the renormalised unit weight error where the renormalisation is to the 41st percentile of the unit weight error as a function of colour and apparent magnitude[6]. As a result, the 41st percentile is achromatic however the spread of RUWE through the observed population will still have a residual colour dependence as will the fraction of sources with RUWE>1.4. Therefore colour dependence is an important aspect of the RUWE<1.4 selection function.

The problem with modelling RUWE as a function of colour is that I require all sources to have published colour. Approximately 88% of sources in the *Gaia* EDR3 catalogue have published G_{RP}. Therefore the selection function I will actually be modelling is $\mathrm{P}(\mathrm{RUWE} < 1.4\,|\,\mathcal{S}_{Gaia}, \mathcal{S}_{G_{\mathrm{RP}}})$. If the event that a source has RUWE < 1.4 is entirely independent of the the event that it has published G_{RP}, this probability is the same as $\mathrm{P}(\mathrm{RUWE} < 1.4\,|\,\mathcal{S}_{Gaia})$.

Within the apparent magnitude range $1.6 < G < 22$, there are 1.81 billion sources in *Gaia*, 1.40 billion of which have RUWE < 1.4 ($\mathrm{P}(\mathrm{RUWE} < 1.4\,|\,\mathcal{S}_{Gaia}) \sim 77.6\%$). 1.55 billion sources in *Gaia* have published G_{RP}, 1.29 billion of which have RUWE < 1.4 ($\mathrm{P}(\mathrm{RUWE} < 1.4\,|\,\mathcal{S}_{Gaia}, \mathcal{S}_{G_{\mathrm{RP}}}) \sim 83.3\%$). Therefore, whether a source has published G_{RP} affects the probability that the source will have RUWE < 1.4.

I show this as a function of apparent magnitude in Fig. 4.6 where I give the median and $16^{\mathrm{th}} - 84^{\mathrm{th}}$ percentile ranges of the probability of a source having G_{RP} from the full sample or given the source satisfies the RUWE cut. For $16 < G < 21$, a source is significantly more likely to have G_{RP} if the source has published RUWE < 1.4.

Therefore I advise that the colour-dependent RUWE selection function should only be used in conjunction with the G_{RP} selection function

$$\begin{aligned}&\mathrm{P}(\mathcal{S}_{\mathrm{RUWE}}, \mathcal{S}_{G_{\mathrm{RP}}}\,|\,\mathcal{S}_{Gaia}, \mathbf{y}) \\ &\qquad = \mathrm{P}(\mathcal{S}_{\mathrm{RUWE}}\,|\,\mathcal{S}_{G_{\mathrm{RP}}}, \mathcal{S}_{Gaia}, \mathbf{y}) \cdot \mathrm{P}(\mathcal{S}_{G_{\mathrm{RP}}}\,|\,\mathcal{S}_{Gaia}, \mathbf{y}),\end{aligned} \tag{4.24}$$

which is the probability of a source in EDR3 having both RUWE < 1.4 and published G_{RP}. Since the selection function for G_{RP} is not known, and for cases where one wishes to fit a model to *Gaia* data without colour dependence, I also fit a magnitude-only selection function to the RUWE data including all sources with published G independent of whether G_{RP} was published. By evaluating this, I am implicitly marginalising over the colour

[6] http://www.rssd.esa.int/doc_fetch.php?id=3757412

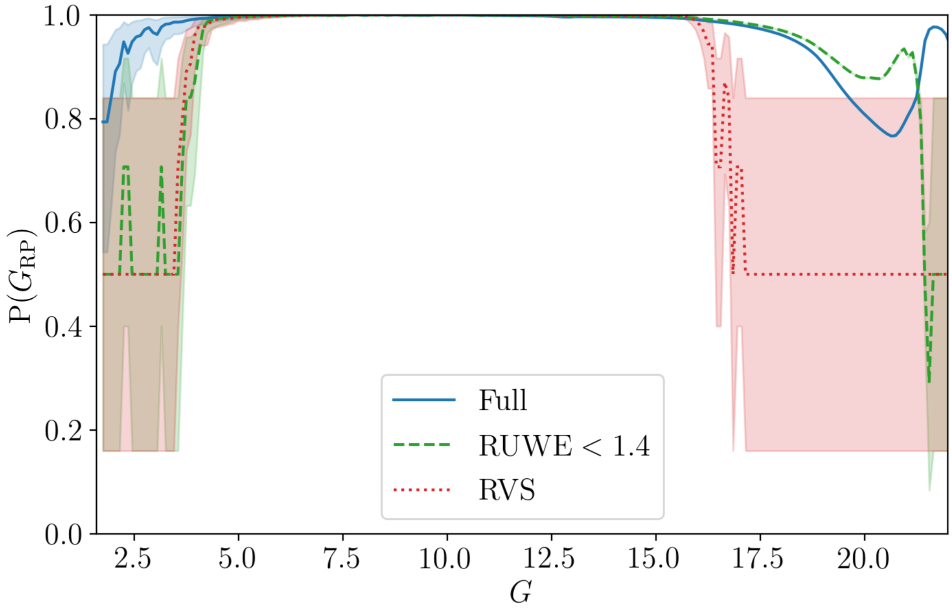


Fig. 4.6 The probability of a source having G_{RP} is evaluated from the Beta-Binomial posterior distribution in Equation 4.22. Beyond $G \sim 15$ there are few enough sources in RVS that this isn't significantly different to the full sample probability however for $16 \lesssim G \lesssim 21$ the RUWE G_{RP} probability changes strongly from the full sample which could lead to significant biases in the RUWE colour selection function if not used with a correct G_{RP} selection function.

distribution of sources

$$\mathrm{P}(\mathcal{S}_{\mathrm{RUWE}} \,|\, \mathcal{S}_{Gaia}, \mathbf{y}) = \int \mathrm{d}C\, \mathrm{P}(\mathcal{S}_{\mathrm{RUWE}} \,|\, \mathcal{S}_{Gaia}, \mathbf{y}, C) \cdot \mathrm{P}(C \,|\, \mathcal{S}_{Gaia}, \mathbf{y}) \tag{4.25}$$

where $\mathrm{P}(C \,|\, \mathcal{S}_{Gaia}, \mathbf{y})$ is the distribution of source colours in the *Gaia* source catalogue at the given position on the sky and apparent magnitude, $\mathbf{y}$.

The distributions of all *Gaia* sources and those with RUWE < 1.4 as a function of G and C are shown in the first and third panels of Fig. 4.4 respectively. Comparing with the *Gaia* Catalogue of Nearby Stars (Gaia Collaboration et al., 2021b) there are sources extending out to the blue ($G - G_{\mathrm{RP}} < -1$) and red ($G - G_{\mathrm{RP}} > 3$) extremes of the colour distribution. The red sources will likely be extended sources or stars in crowded regions with high excess flux as mentioned in Section 4.3. The blue sources are very low in numbers and only appear at the faint end suggesting they are driven by photometric measurement errors for faint sources. We will only fit the colour-dependent selection for the range $-1 < G - G_{\mathrm{RP}} < 7$ which contains sources with well measured colours including those with high excess flux.

```
SELECT *
FROM gaiaedr3.gaia_source
WHERE
    ruwe<1.4
AND phot_rp_n_obs>0
AND phot_g_mean_mag BETWEEN 1.6 AND 22
AND phot_g_mean_mag-phot_rp_mean_mag BETWEEN -1 AND 7
```

The RUWE sample has 1 292 152 210 sources out of a total 1 550 860 236 in the *Gaia* EDR3 source catalogue with colours and apparent magnitudes in the given range.

For the colour-independent selection function, the RUWE sample is composed of 1 400 803 102 sources out of 1 806 195 366 in the same magnitude range in the *Gaia* source catalogue which can be retrieved with the following query.

```
SELECT *
FROM gaiaedr3.gaia_source
WHERE
    ruwe<1.4
AND phot_g_mean_mag BETWEEN 1.6 AND 22
```

The source counts are shown as a function of position on the sky in Fig. 4.5 and predominantly trace the Milky Way stellar distribution.

4.5.3 RVS

The *Gaia* radial velocity sample is incredibly important to dynamical studies of the Milky Way (e.g. Nitschai et al., 2020). Radial velocities are measured by the *Gaia* satellite's on-board spectrograph using the calcium triplet. Radial velocities have not yet been released for *Gaia* DR3, however the DR2 sample of 7 million RVS stars is still by far the largest sample of stellar radial velocities available from any single observatory. The selection criteria used to produce the DR2 sample is given in Gaia Collaboration et al. 2018b. These are some of the main cuts:

- $G_{\rm RVS}^{\rm ext} < 12$ or $G_{\rm RVS}^{\rm int} < 14$
- $3550 < T_{\rm eff} < 6900$
- No double line spectroscopic binaries or emission line stars.

$G_{\rm RVS}^{\rm ext}$ is the apparent magnitude estimated through photometric transformations of observations from ground-based observatories in the Initial *Gaia* Source List (IGSL, Smart & Nicastro, 2014). $G_{\rm RVS}^{\rm int}$ is directly estimated from the *Gaia* spectroscopy data in the radial velocity pipeline. In either case, $G_{\rm RP}$ provides a better approximation to $G_{\rm RVS}$ than G. However since the *Gaia* source catalogue selection function is modelled as a function of G, I keep to that here in the interests of simplicity and usability.

Selection on IGSL-measured $G_{\rm RVS}$ produces substantial structure across the sky as shown in the right-hand panel of Fig. 4.7 with vertical lines in the East due to SDSS IGSL sources and a grid pattern around the South equatorial pole from the Guide Star Catalogue. This is demonstrated and discussed in more detail in Rybizki et al. 2021b.

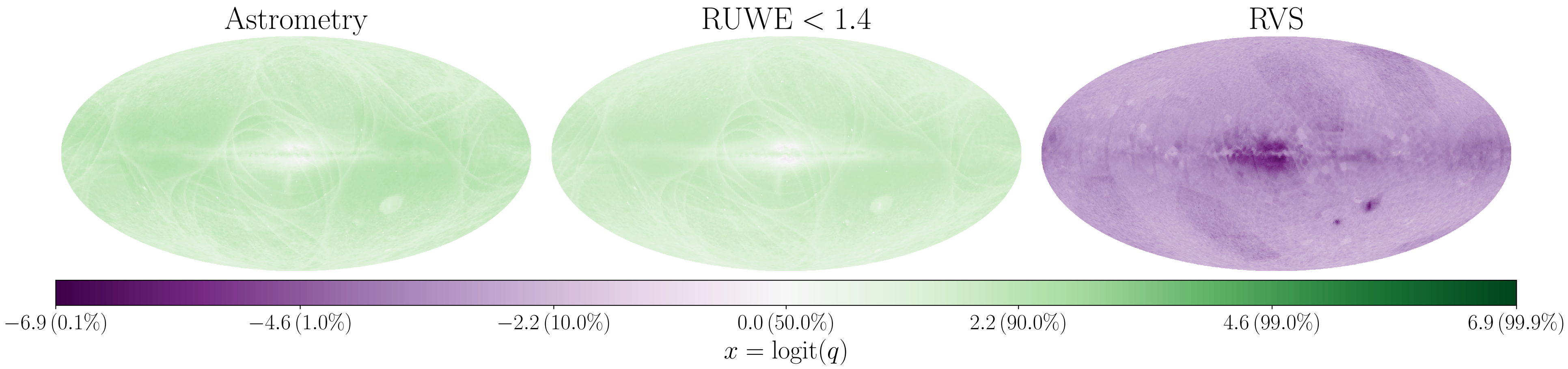


Fig. 4.7 $q = (k+1)/(n+2)$ gives the expected binomial probability as explained in Section 4.4.2. This is shown for each of the samples with k and n taken as the number of sources in each HEALPix level 6 bin summed over the colour and apparent magnitude range used for the selection function in the subsample and full samples respectively. In all three cases, the samples are limited relative to the source catalogue in crowded regions such as the bulge, LMC and SMC whilst there are also features of the scanning law which are apparent.

Motivated by Boubert et al. 2019, additional analysis and cleaning of the DR2 sample took place for the RVS sources published with EDR3[7]. 3 876 sources with incorrect radial velocities due to nearby neighbours were removed and 10 924 could not be successfully crossmatched with any source in *Gaia* EDR3. I also apply the cut RV_NB_TRANSITS ≥ 4 recommended by Boubert et al. 2019 to clean out spurious radial velocity measurements.

The distribution of RVS sources as a function of apparent magnitude and $G - G_{\mathrm{RP}}$ colour is shown in the fourth panel of Fig. 4.4. The RVS sample occupies a very narrow colour range and is heavily magnitude-limited at the dim end. Only the range of colour and apparent magnitude containing radial velocity sources is used for the RVS sample, namely $-0.6 < G - G_{\mathrm{RP}} < 2.6$ and $1.6 < G < 17.4$, where I have applied the same bright-end cut as in the astrometry sample.

Once again, I am faced with the same issue as in the RUWE selection where not all RVS sources have published G_{RP}. However the RVS selection function is only non-zero at brighter magnitudes where G_{RP} is much more complete. For $1.6 < G < 17.4$, 6.2 out of 206 million sources, or 3.00%, have published DR2 radial velocities which rises to 3.04% of sources with published G_{RP}. The RVS selection function has a much weaker dependence on G_{RP} selection than is the case for RUWE. By using the colour-dependent RVS selection function without accounting for the G_{RP} selection probability, a $\sim 1\%$ systematic uncertainty would be introduced to the results. This is also shown in Fig. 4.6 where a dotted blue line and shaded blue regions show that the G_{RP} probability is very high out to $G \sim 16$ at which point I start to run out of RVS sources so the uncertainties become large.

```
SELECT *
FROM gaiaedr3.gaia_source
WHERE
    dr2_rv_nb_transits>=4
AND phot_rp_n_obs>0
AND phot_g_mean_mag BETWEEN 1.6 AND 17.4
AND phot_g_mean_mag-phot_rp_mean_mag BETWEEN -0.6 AND 2.6
```

The RVS sample within the given colour-apparent magnitude range contains 6 186 950 out of a total 203 513 110 objects in the source catalogue within the same colour-magnitude range with published G_{RP}.

For RVS, the distribution on the sky (shown in the right panel of Fig. 4.5) is no longer solely dominated by the Milky Way source distribution. Aspects of the *Gaia* scanning law and features of the IGSL are visible in the on-sky distribution.

4.5.4 Power Spectrum

The Gaussian Process prior requires a dispersion parameter σ which needs to be well chosen for the problem. The expected variance will depend on the Needlet scale.

[7] https://gea.esac.esa.int/archive/documentation/GEDR3/Data_processing/chap_cu6spe/sec_cu6spe_intro/ssec_cu6spe_nonewdata.html

I start from the expected selection probability across the sky, $\mathbb{E}[q] = \frac{\alpha}{\alpha+\beta} = \frac{k+1}{n+2}$ (see Section 4.4.2). The distribution of logit $\left(\frac{k+1}{n+2}\right)$ is plotted across the sky for each of the samples in Fig. 4.7. A spherical harmonic model can be directly estimated for the distribution from the HEALPix values using

$$\hat{a}_{\ell m} \sim \frac{4\pi}{N_{\mathrm{pix}}} \sum_{p=0}^{N_{\mathrm{pix}}-1} Y^{*}_{\ell m}(l_p, b_p) x_p \tag{4.26}$$

where x_p are the pixel values. The power spectrum is then estimated by taking the square mean of mode amplitudes

$$\hat{C}_\ell = \frac{1}{2\ell+1} \sum_{m=-\ell}^{\ell} |\hat{a}^2_{\ell m}|. \tag{4.27}$$

This is done for each sample with the power-spectra shown by the black lines in Fig. 4.8.

The power spectrum gives the expected variance of spherical harmonic coefficients. I model the power spectra with a single power-law distribution

$$C_\ell = A(\ell+1)^\gamma \tag{4.28}$$

where γ describes how the power decays for smaller scales.

The sum of square spherical harmonic coefficients renormalised by their uncertainty is chi-square distributed with $2\ell+1$ degrees of freedom.

$$\sum_{m=-\ell}^{\ell} \frac{|\hat{a}^2_{\ell m}|}{\sigma^2_\ell} \sim \chi^2(2\ell+1) \tag{4.29}$$

where $\sigma^2_\ell = C_\ell$ is the expected coefficient variance. Therefore I can use this to derive the likelihood of the given power spectrum model

$$\mathrm{P}(\hat{C}_\ell \,|\, C_\ell) \propto x^{\frac{2\ell+1}{2}-1} \exp\left(-\frac{x}{2}\right) \tag{4.30}$$

where $x = \frac{(2\ell+1)\hat{C}_\ell}{C_\ell}$. The $16^{\mathrm{th}} - 84^{\mathrm{th}}$ percentiles of this distribution provide the shaded regions in Fig. 4.8.

This is maximized with respect to A and γ in Eq. 4.28 to determine the best fit parameters using gradient descent with the Newton Conjugate Gradient method as implemented in SCIPY. I only use data with $\ell < 135$ as this is the smallest scale spherical harmonic used in the Needlets and corresponds to a scale length on the sky ~ 1 degree which is the approximate pixel size of the data I will be using with HEALPix nside = 64. The best fit power law profiles are shown by the red dashed lines in Fig. 4.8 for the Astrometry, RUWE (magnitude-only) and RVS samples respectively with parameter values given in Table 4.1.

Given the power spectrum of spherical harmonics, I want to work out what this implies for the variance of Needlet coefficients. From Appendix B of Boubert & Everall 2021, the

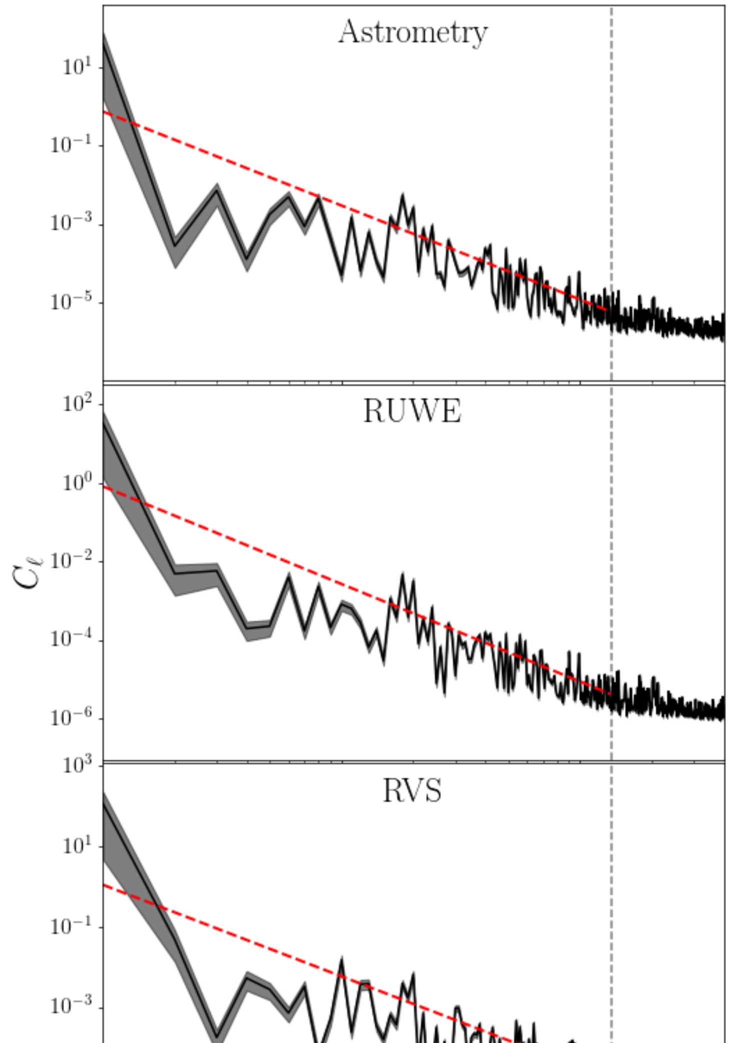


Fig. 4.8 The power spectrum for each sample, evaluated using the spatial $(k+1)/(n+2)$ distribution from Fig. 4.7 shown by the black lines with grey 16th – 84th percentile uncertainties, declines strongly with increasing ℓ in all samples. The red dashed line shows a power law fit for each sample out to $\ell = 135$ which is the maximum ℓ set of spherical harmonics used to construct the needlets.

		Astrometry	RUWE < 1.4 (G only)	RUWE < 1.4	RVS
Sample	$\sum_i n_i$	1,806,195,366	1,806,195,366	1,550,860,236	203,513,110
	$\sum_i k_i$	1,465,211,050	1,400,803,102	1,292,152,210	6,186,950
	G range	[1.6,22]	[1.6,22]	[1.6,22]	[1.6,17.4]
	$G - G_{\mathrm{RP}}$ range	-	-	[-1,7]	[-0.6,2.6]
Bins	HEALPix nside	64	64	64	64
	G bins	0.2	0.2	0.2	0.2
	$G - G_{\mathrm{RP}}$ bins	-	-	2.0	0.4
Power Spectrum	$\log A$	-0.2937	-0.6206	-0.2032	0.0937
	γ	-2.3985	-2.3552	-2.4845	-2.2761
Model	j_{max}	5	5	5	5
	l_G	0.3	0.3	0.3	0.6
	$l_{G-G_{\mathrm{RP}}}$	-	-	3.0	1.2
	B	2.0	2.0	2.0	2.0

Table 4.1 For each *Gaia* sub-sample selection function, I provide the key parameters describing the data and model.

variance of the Needlet coefficients as a function of the power spectrum is given by

$$\langle|\beta_{jk}^2|\rangle = \lambda_{jk} \sum_{\ell=0}^{\ell_{\max}} b_\ell^2(j) C_\ell \frac{(2\ell+1)}{4\pi} \tag{4.31}$$

where C_ℓ is taken from Eq. 4.28 using the best fit parameters and the normalisation constant is the area per pixel, $\lambda_{jk} = \frac{4\pi}{N_j}$.

The reader should be concerned that I have done something uncomfortably non-Bayesian. I used the data to determine the appropriate prior for my model. There are two reasons why I can get away with this. Firstly, I used the data aggregated over colour-apparent magnitude space and only used this to estimate two parameters of a simple power spectrum such that the vast amount of information is hidden from my prior model decision. Secondly, and more importantly, I am not attempting to infer a posterior distribution for my model. Using data to infer the prior would lead to an underestimation of posterior uncertainties as data has implicitly been double counted, however, I am only inferring the best fit selection function model.

The reason I used the power spectra as described is that it makes the optimization significantly more computationally efficient as I am using an informed start point.

4.6 Results

When estimating each of the selection functions I used a Needlet model with $j_{\max} = 5$ and fit to data at HEALPix level 6 (nside = 64) resolution. This gives a total of 16 381 Needlets with ~ 2 degree resolution fit to 49 152 pixels. Details about the samples, binning schemes and model parameters are listed in Table 4.1.

In this Section, I show the results of the fits as a function of sky position, apparent magnitude and colour. To test the veracity of the fits given the data, I use the Binomial p-value test explained in Section 4.4.3.

4.6.1 Astrometry

The astrometry is fit in 0.2 mag bins with a magnitude scale length of 0.3 mag - long enough that neighbouring bins are correlated, but short enough that the data can produce sharp changes in the model.

The astrometry selection function across the sky is shown in the middle row of Fig. 4.9. The top row shows $(k+1)/(n+2)$ for the given magnitude bins, which, as discussed in Section 4.4.2, is the expected selection probability in the given magnitude bin independent of all others. High source density regions show much stronger selection limitations, particularly at the dimmer magnitudes. This is apparent through the white spots at $G = 17.5$ which are centred on globular clusters such as ω Centauri and dwarf galaxies such as the LMC and SMC.

The bottom row of Fig. 4.9 shows the results of the p-value test discussed in Section 4.4.3. At brighter magnitudes, p-value is distributed completely randomly demonstrating that

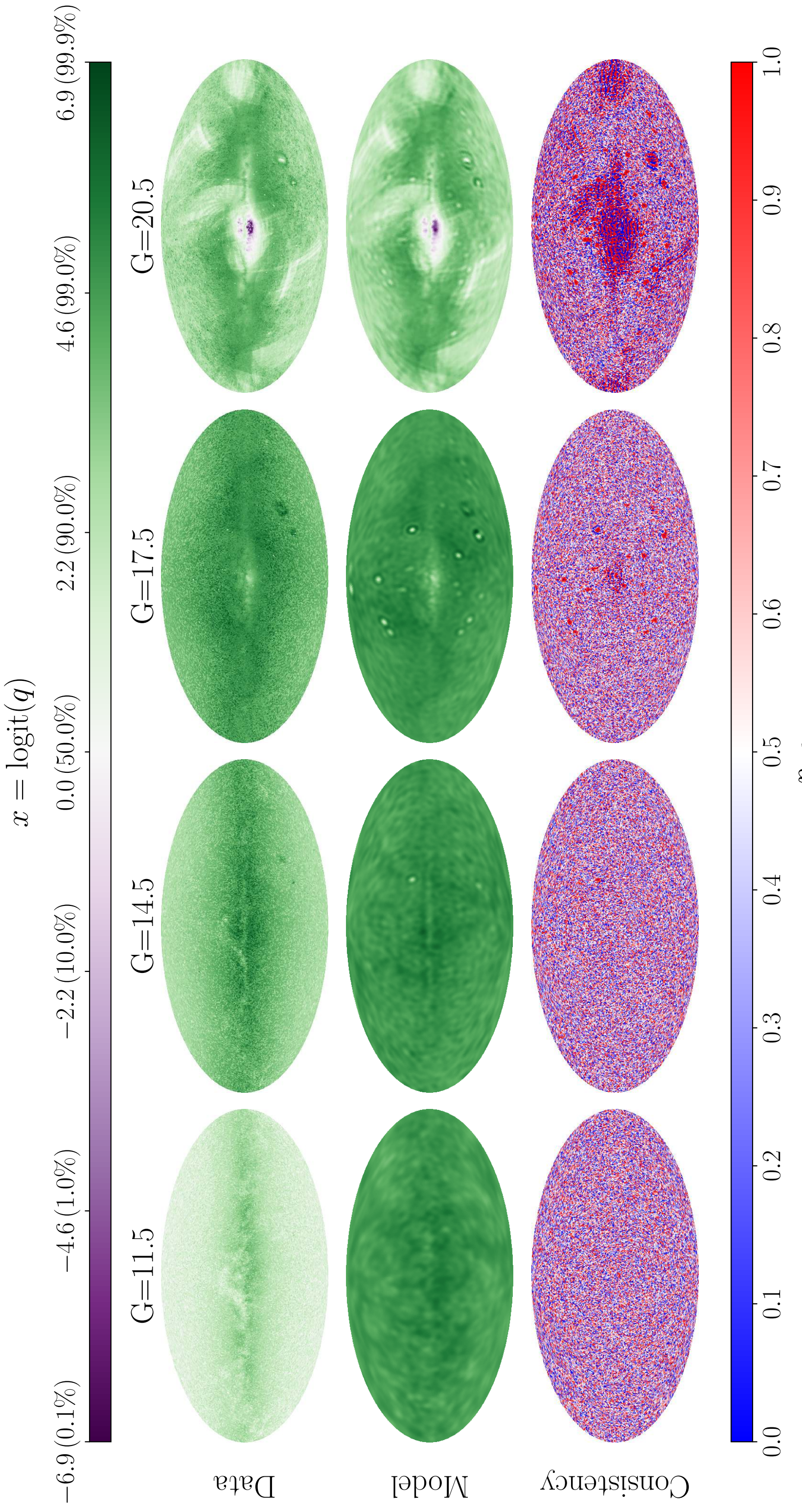


Fig. 4.9 The results of the astrometry selection function fit (middle row) show a striking drop in high source density regions, not just in the disk and Magellanic clouds but also in the Milky Way globular clusters. There is also scanning structure at the faint end which significantly reduces the selection function probability in under-scanned regions. The model shows strong agreement with the approximate expected distribution from the data (top row). This is demonstrated more precisely with the Binomial p-value tests which are dominated by random noise. In the faint bins, some structure arises due to lack of resolution in the model, as such my model should only be trusted to ∼ 2 degree scales.

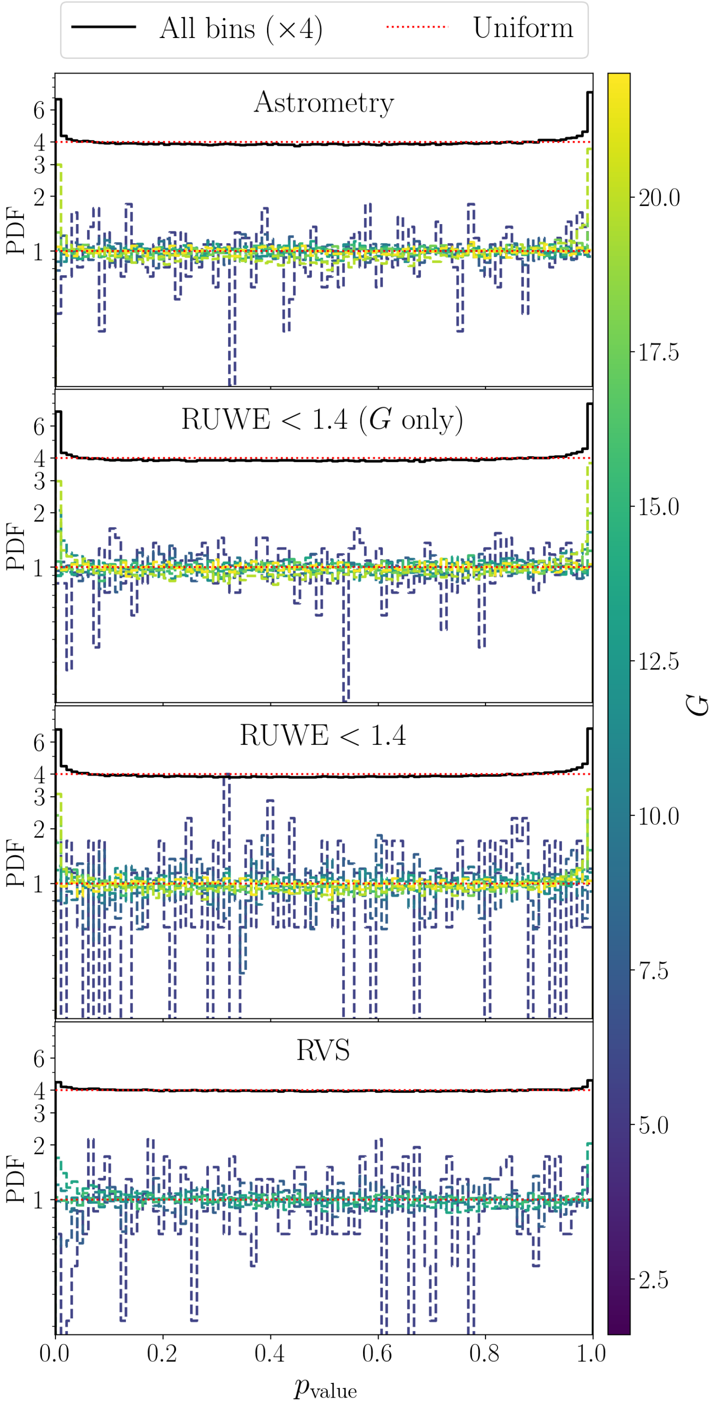


Fig. 4.10 For all four models, the Binomial p-values of HEALPix-colour-magnitude bins are uniformly distributed at almost all magnitudes. This demonstrates that the model has fit the data exceptionally well. The minor exceptions are some flaring of the wings in the faintest bins of each sample which is caused by the lack of spatial resolution of the model which isn't able to pick up structure on scales under 2 degrees.

the model has been successfully fit. Shifting to fainter magnitudes the selection function probability is reduced in regions of the sky with fewer scans. These ‘holes’ in the scanning law are prominent due to the selection cuts on $N_{\rm vpu}$ and $\sigma_{\rm 5Dmax}$ used for the EDR3 astrometry sample.

Structure also appears particularly around ω Centauri and the disk and LMC. This is where the selection function changes on scales smaller than the model can resolve. For example, the half-light radius of ω Centauri is ~ 4.8 arcminutes (van de Ven et al., 2006) however the Needlets can only resolve structure on 2 degree scales. Therefore the reduction in selection probability due to the globular cluster is spread out over a wider area by the Needlet. At the core, in the pixel which contains the globular cluster, the selection probability is overestimated whilst being underestimated in any neighbouring pixels. There is a further-out halo of overestimated probability due to the structure of the Needlets which go negative before returning to zero (see Fig.1 Boubert & Everall, 2021).

I provide a histogram of p-values in the top panel of Fig. 4.10 where I only include bins where $n > 0$ as bins with $n = 0$ have a uniformly distributed p-value independent of the model. The histogram for all bins is offset by a factor of 4 to make it clearer. For the vast majority of the data, I see well-behaved solutions with uniformly distributed p-values however, at the faint end, there are over-densities at $P = 0, 1$ where the resolution limitations become significant and I am under-fitting to the data.

The astrometry selection function as a function of magnitude is shown in the top panel of Fig. 4.11. I group level 6 HEALPix pixels by number of sources in the pixel in the *Gaia* source catalogue with $1.6 < G < 22$. The selection function is the count-weighted mean of the selection functions in the given pixels. For $G \lesssim 17$, the Astrometry sample is $\sim 99\%$ complete however this drops off quickly in high source density regions. Low density regions stay close to complete out to $G \sim 19$ before also falling rapidly. By $G \sim 20$, less than 1% of sources from the *Gaia* source catalogue are included in the astrometry sample.

For $G < 1.6$ I am not able to say anything informative about the selection function probability as there is no data in the source catalogue here to use. For $G > 22$ the selection function should be taken as zero.

For the astrometry sample I have shown that the selection function model reproduces the observed data down to Needlet scales of ~ 2 degrees. At high latitudes, the astrometry sample can be complete out to $G \sim 19$ and significantly drops at $G \sim 21$ due to the cut placed on DR2 apparent magnitude. However the selection probability declines much brighter for crowded regions or where there are very few scans in EDR3.

4.6.2 RUWE

As discussed in Section 4.5, I evaluate the RUWE selection function in terms of sky-position and apparent magnitude against all sources in *Gaia* EDR3 and separately also as a function of $C = G - G_{\rm RP}$ for sources where $G_{\rm RP}$ is published.

Both RUWE selection functions are fit to 0.2 mag bins in G with a magnitude scale length of 0.3. The colour dependent model includes four colour bins each 2 mag wide with

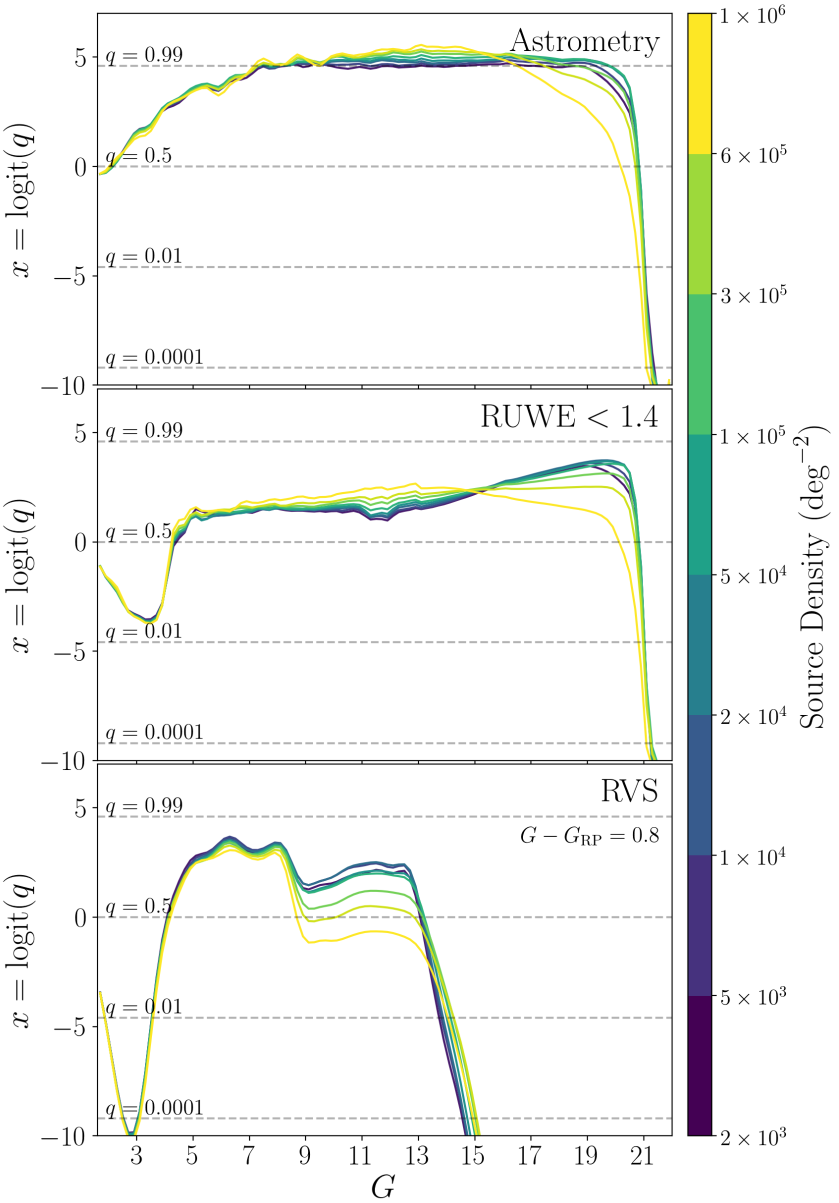


Fig. 4.11 The selection function probability is strongly dependent on source density in all samples. By splitting pixels according to the source density in the full *Gaia* sample, I evaluate the selection function probability using the count-weighted average of all pixels within the given density bin. For the astrometry sample (top) the selection function drops at brighter magnitudes for higher source density regions. The RUWE < 1.4 colour-independent selection function (middle panel), shows a slightly different behaviour as much brighter sources are more likely to be cut out due to excess noise however the faint end shows similar behaviour to the astrometry. The RVS sample (bottom panel) at $G - G_{\mathrm{RP}} = 0.5$ shows a more complicated pattern. At the bright end, as expected, the selection function is lower in high source density regions. However this flips at $G \sim 13$. This is because RVS is limited in $G_{\mathrm{RVS}} \sim G_{\mathrm{RP}}$ and hence is more complete for redder sources as shown by Fig. 4.15 which will largely be bulge fields due to dust extinction. In all samples, the model reverts to the prior mean ($x = 0$) for $G \lesssim 3$ due to a lack of data in the source catalogue.

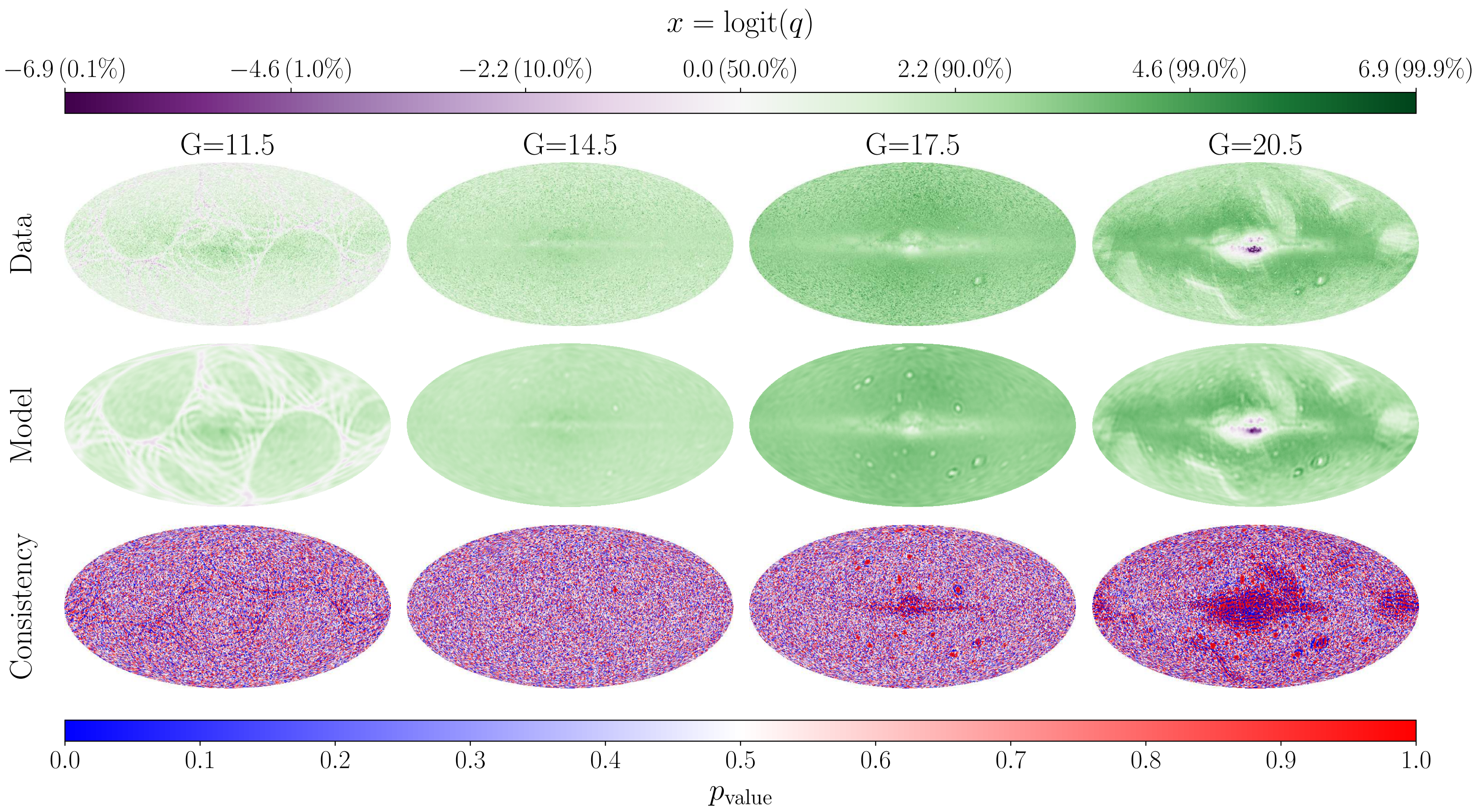


Fig. 4.12 The RUWE < 1.4 selection function probability across the sky (middle row) shows high density region of sources at faint magnitudes due to crowding and scanning law features at bright magnitudes where the most precisely measured sources are more likely to have high RUWE due to excess source noise such as binary motion. Much of this structure is clearly visible in the $(k+1)/(n+2)$ plots (top row). The p-value test (bottom row) shows a good fit across most magnitudes and regions of the sky but shows that the model cannot fully resolve the scanning law at the bright end and source density structure at the faint end.

a scale length of 3 mag which enables broad changes in the selection function probability as a function of C.

The results of the colour-independent model are shown in Fig. 4.12. At the faint end the behaviour is similar to the astrometry selection function and the binomial p-value also shows that the fits are struggling to resolve features in crowded regions at faint magnitudes. For $G = 17.5$ and brighter, however, the RUWE < 1.4 cut removes significant numbers of sources which have published astrometry. The selection probability is lower and for $G = 11.5$ the scanning law is highlighted as heavily scanned regions are more likely to be removed by the cut on RUWE.

The magnitude-dependent behaviour is shown more clearly in the middle panel of Fig. 4.11 where the RUWE < 1.4 selection probability actually peaks near the faint end before declining and shows the same crowding dependence as the astrometry selection. However, the bright end has significantly reduced selection probability particularly in non-crowded regions. My interpretation is that astrometric measurements have lower variance for bright sources and therefore sources with genuine intrinsic noise (such as binary systems) would have a more significant excess noise and be removed by the RUWE < 1.4 cut. For $G \lesssim 3$ the selection probability picks up again which is likely an artefact of the prior as there is very little data in the source catalogue at these magnitudes.

The colour-dependent selection function for RUWE < 1.4 across the sky for four colour and apparent magnitude bins is given in Fig. 4.13. Because RUWE is a measure of the astrometric error above the expected uncertainty, areas with high RUWE aren't necessarily those where *Gaia* performs worst but rather those where the satellite struggles unexpectedly. The selection function for $C = 0$ at the dim end shows a similar structure to the astrometry selection suggesting that the astrometry sample cuts are the limiting factor in this region of colour space. As I shift towards redder sources the picture changes significantly. For redder sources, the low RUWE sample is more complete in the galactic plane and significantly less so in un-crowded regions. This inversion appears to be more extreme at brighter magnitudes. A possible explanation is that the *Gaia* attitude error model either overestimates measurement uncertainties for bright sources in Galactic plane or underestimates those at high latitudes. This would result in higher RUWE for high latitudes and explain the observed selection pattern. I show a systematic issue with a similar structure in Chapter 5.

Fig. 4.14 shows the p-values in four magnitude bins across the sky for the $C \in [1.0, 3.0]$ colour bin. As in the astrometry selection, the bright bins have very successfully modelled the data however at the faint magnitudes, spatial resolution becomes important in the Galactic plane. Once again we see a rippling effect on scales which the Needlets aren't able to fully resolve. The second panel of Fig. 4.10 gives the distribution of all p-values from which we can see that the resolution issues are confined to dim magnitudes and overall the model still fits the data extremely well.

The selection function probability for the RUWE sample across the CMD is shown in Fig. 4.15. For the highly crowded bulge region (top panel) the selection probability peaks at $G \sim 18$ in the blue and $G \sim 11$ in the red before declining towards fainter magnitude

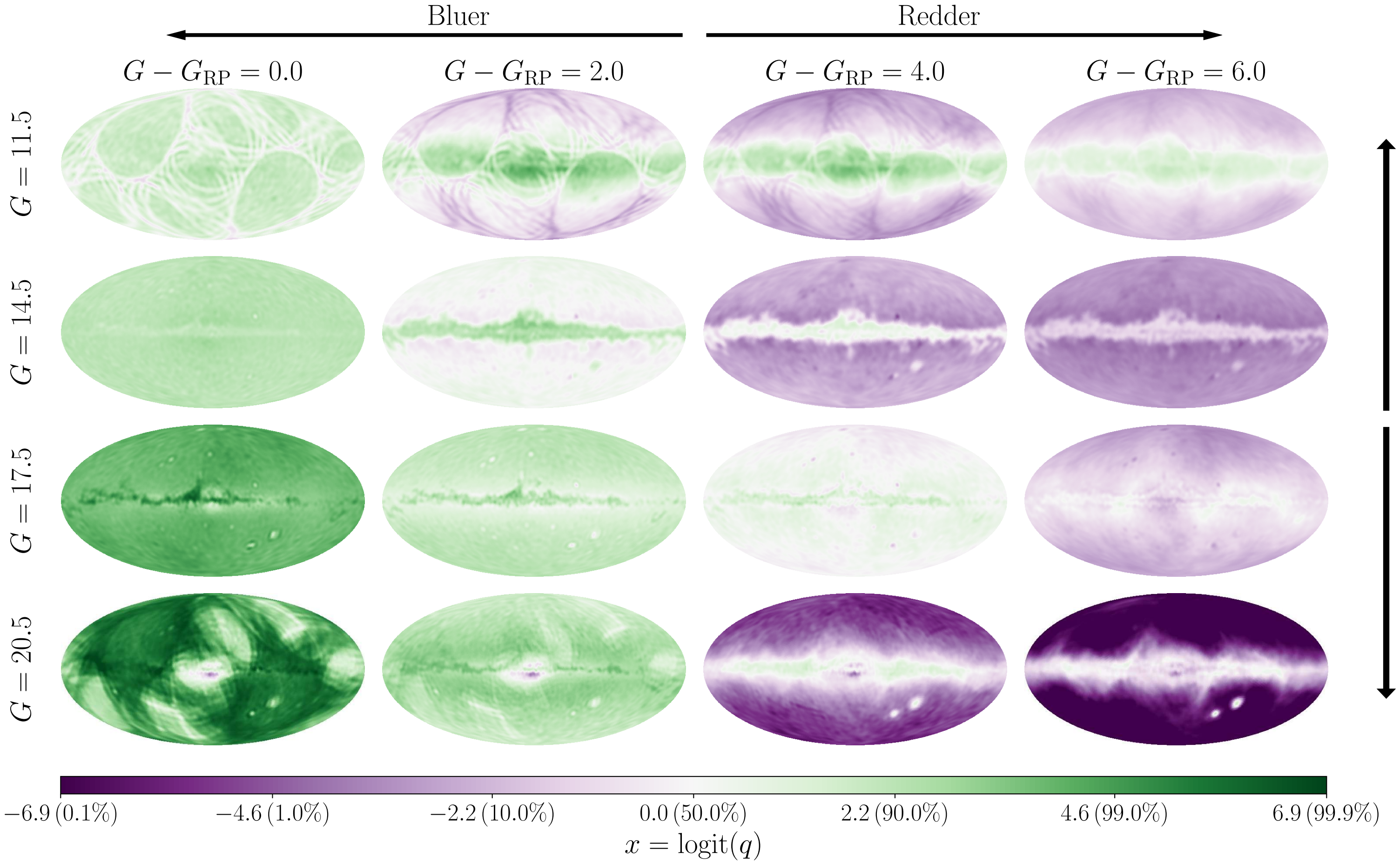


Fig. 4.13 The RUWE selection function shows expected behaviour for blue faint bins (bottom left) where the probability is lower in high source density regions. However moving to redder and brighter magnitudes, this trend distinctly reverses with the disk showing significantly higher selection probability than high latitudes. An explanation would be that sources at high latitudes are typically nearer due to the sharp vertical drop off of the disk and so binary systems are more likely to produce a significantly large RUWE due to their large angular scale. The binary main sequence sits above and to the red side of the single star main sequence due to the combination of stellar fluxes which might explain why this reversal is more prominent at redder colours.

particularly for redder sources before cutting off sharply at $G \sim 20.5$. In a low-density field (bottom), there is a stronger bimodality with selection function peaks at $G \sim 7$ and $G \sim 20$. This structure is likely connected to the different types of observations taken onboard *Gaia* with 1D windows for $G > 13$ sources and 2D windows $G < 13$ with window gating happening at still brighter magnitudes (Evans et al., 2018).

The overall trends with apparent magnitude for bluer sources are similar to the results from the colour-independent model shown in the middle panel of Fig 4.11. The cause of the drop-off in selection probability for faint red sources is unknown but may be related to chromaticity of the astrometric solution. An alternative explanation is that redder sources have lower G_{RP} and as such are more likely to have a measured G_{RP} in the source catalogue. This could push the selection function down at these colours. In this region of parameter space, one should make sure that the G_{RP} selection function is used in conjunction with RUWE < 1.4 as discussed in Section 4.5.

As in the magnitude-only model, the RUWE selection probability picks up for $G \lesssim 3$ due to lack of data in the source catalogue.

4.6.3 RVS

The RVS sample is binned in 0.2 mag bins in colour and 0.4 mag bins in $G - G_{\mathrm{RP}}$. I use a scale length of 0.6 in apparent magnitude and 1.2 in colour. I opt for higher colour resolution than I did for RUWE because RVS selection will be closely related to G_{RP} which will generate strong $C = G - G_{\mathrm{RP}}$ dependence in the selection function.

The radial velocity sample is dependent on CCD observations from the on-board spectrograph which occupies four out of seven CCD rows on the *Gaia* focal plane. Depending on the G_{RVS} apparent magnitude evaluated for the source, a variety of window classes will be used as described in Section 7.1 of Cropper et al. (2018). Sources with $G_{\mathrm{RVS}} < 7$ (Class 0) received a full 2D windows whilst fainter sources only received 1D windows (Class 1 or 2). In crowded regions, Class 1 and 2 sources with overlapping windows would have their windows truncated in the region of overlap often leading to non-rectangular windows. These observations were not used in DR2 data processing as explained in Sartoretti et al. 2018 leading to 40% of spectra being removed and a much higher percentage in crowded regions.

This process is manifested in the RVS selection function. The third panel in Fig. 4.11 shows that the RVS selection function for $G \lesssim 8$ (corresponding to approximately $G_{\mathrm{RVS}} \lesssim 7$) is source density independent whereas sources in crowded regions are much less likely to be selected with $8 \lesssim G \lesssim 12.5$.

For $G \gtrsim 12.5$, this behaviour changes significantly. At dimmer and redder magnitudes, the RVS population traces the distribution of dust with high completeness in regions of the sky with significant extinction. I do not know the precise cause of this however there are some plausible explanations. The RVS catalogue filters out stars cooler than $T_{\mathrm{eff}} = 3550\mathrm{K}$ which corresponds to $G - G_{\mathrm{RP}} \gtrsim 1.2$ (see Fig.3 Andrae et al., 2018). However, in regions with high dust extinction I will have hotter stars appearing with higher $G - G_{\mathrm{RP}}$ due to reddening. Therefore, for redder colours, the RVS cut on T_{eff} will be more strict where

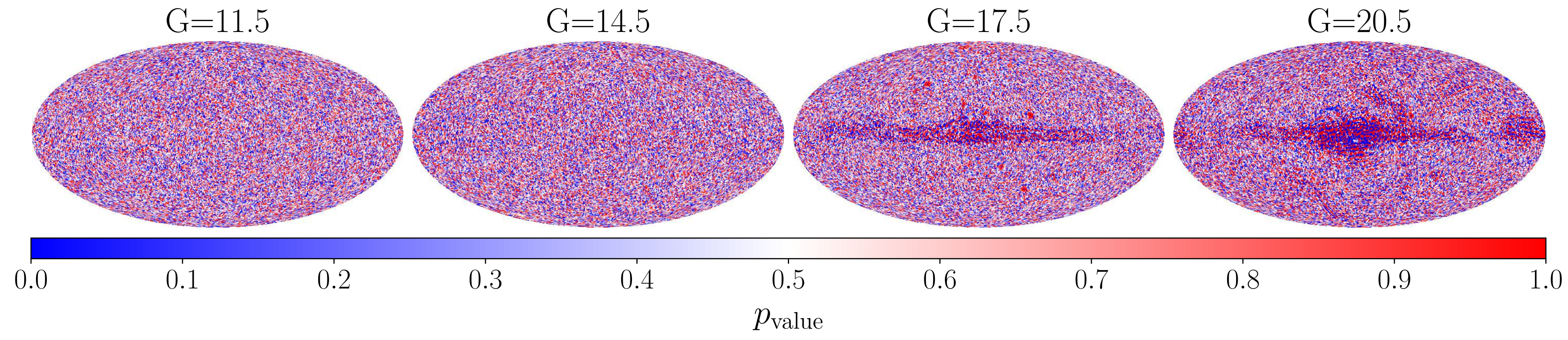


Fig. 4.14 The binomial p-value for colour-magnitude bins of the RUWE < 1.4 sample selection function with $\boldsymbol{G} - \boldsymbol{G}_{\mathrm{RP}} \in [1., 3.]$ shows no residual structure at the bright end. At faint magnitudes the small scale structure of the disk shows up in waves as the model averages the variations out on 2deg scales.

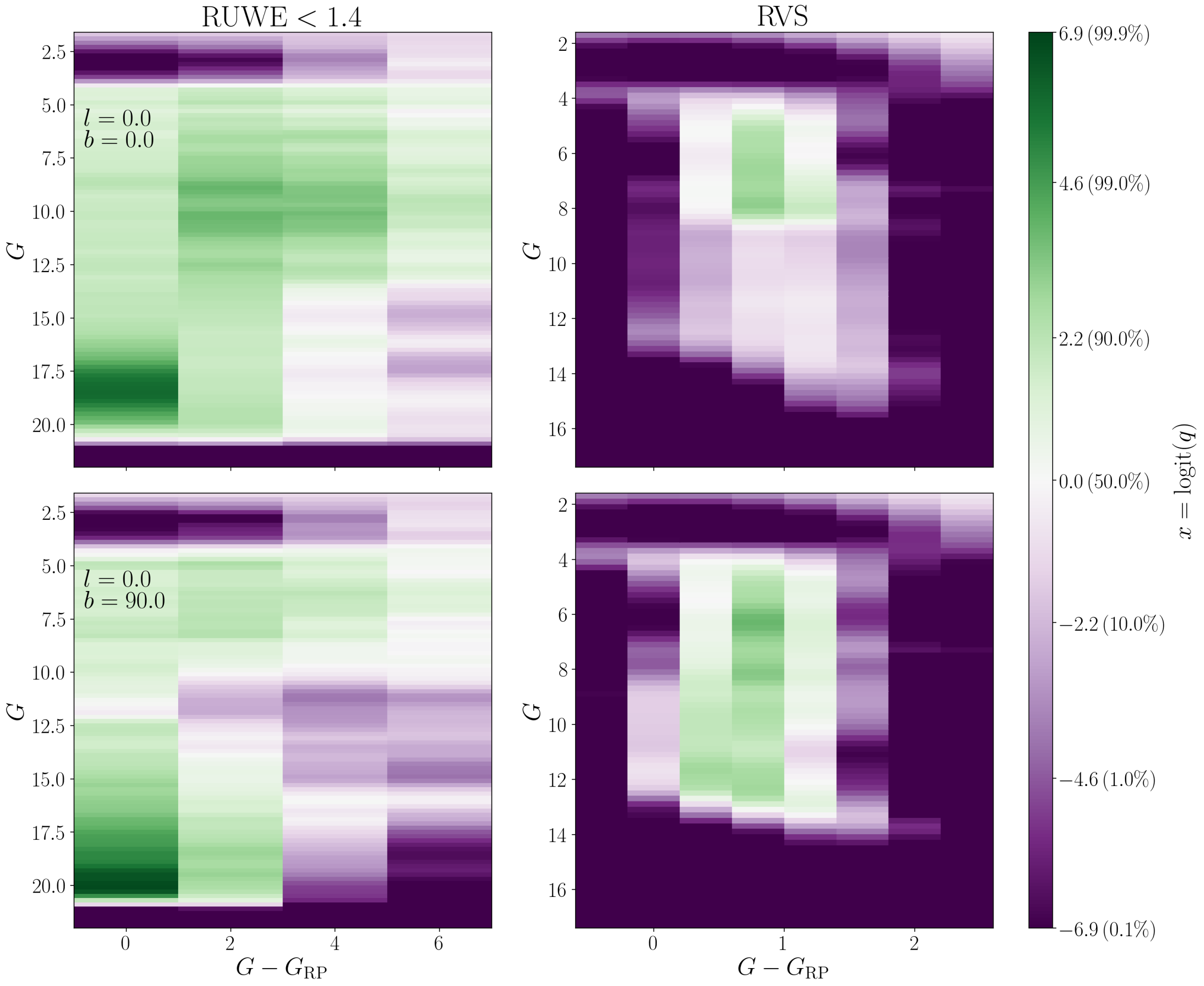


Fig. 4.15 The RUWE selection function (left) is highest at faint blue magnitudes ($G \sim 18-20$, $G - G_{\mathrm{RP}} \sim 0$) and drops significantly at both the brightest and faintest ends for the crowded bulge field (top) and uncrowded Galactic pole field (bottom). There is also a strong gradient to lower selection probability for redder sources. This may be related to chromaticity of astrometric fits or relatively high completeness of G_{RP} at these colours. RVS has a narrow range of colours with significant selection probability which is likely the result of the T_{eff} selection used in the RVS pipeline (Sartoretti et al., 2018). At the faint end, the probability rapidly reduces but correlated with colour such that redder sources are observed to fainter magnitudes. This is due to the fact that RVS observes in the red part of the *Gaia* waveband and so the true RVS limit is closer to a G_{RP} cut. This produces the nice diagonal cut in G vs $G - G_{\mathrm{RP}}$ corresponding to $G_{\mathrm{RP}} \sim 12.5$.

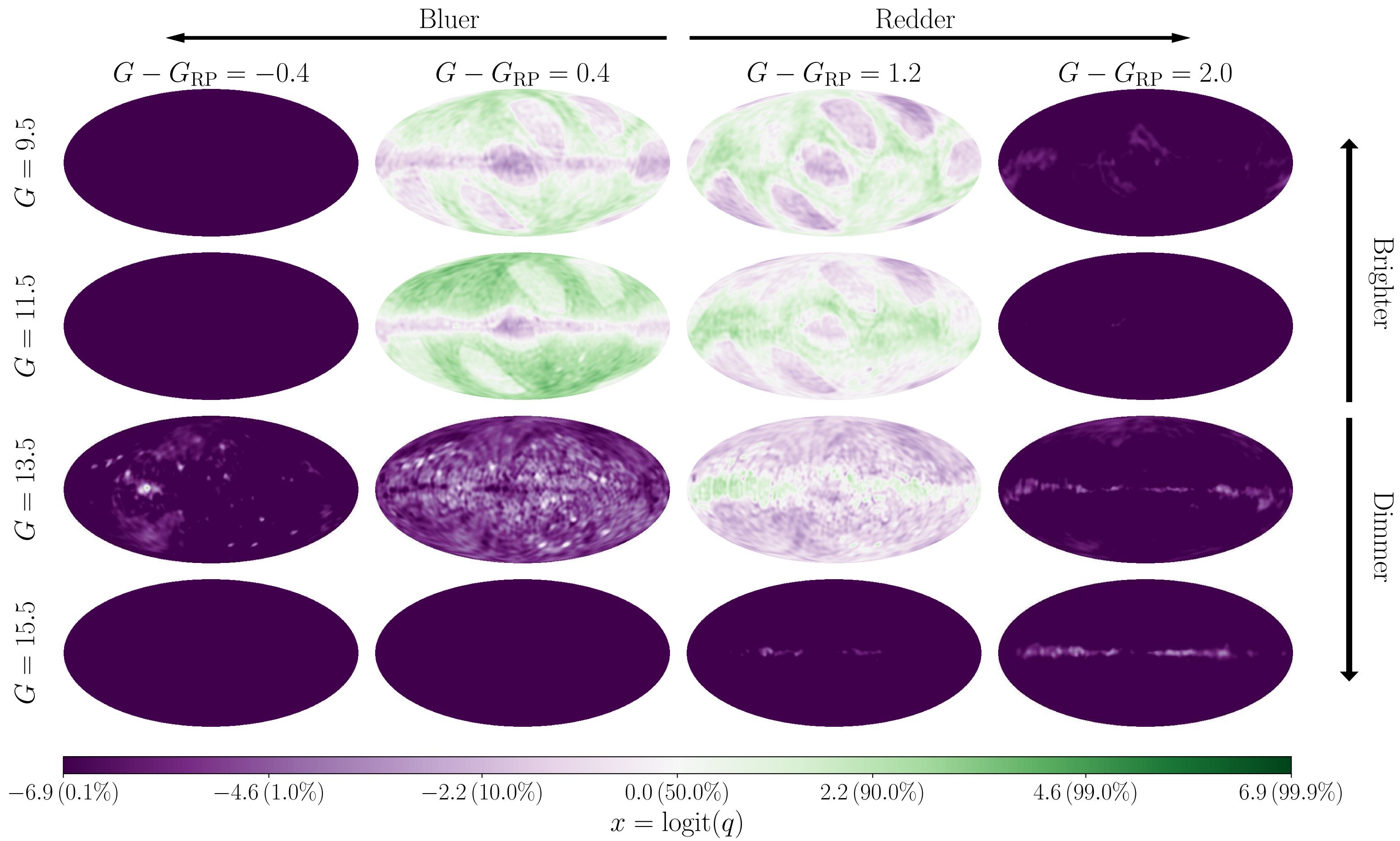


Fig. 4.16 At brighter magnitudes (top rows) and bluer colours (left columns), the RVS selection probability is significantly higher in regions of the sky with more scans and lower source density. As we move dimmer and redder, this shifts to a higher selection probability in the Milky Way disk where dust extinction causes sources to be systematically redder and therefore have be relatively brighter in G_{RP} (see Fig.4.15).

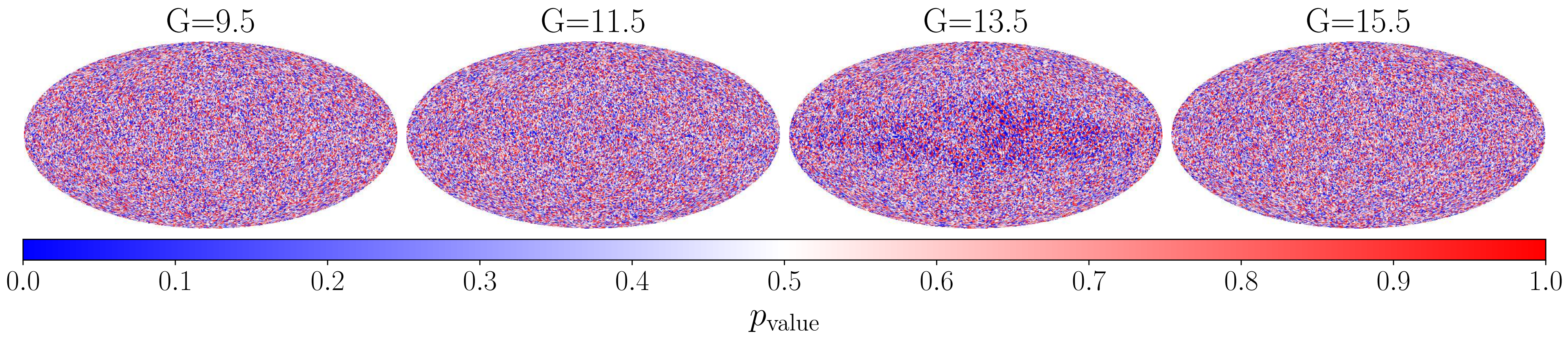


Fig. 4.17 The p-value test for RVS bins, shown here for $\boldsymbol{G} - \boldsymbol{G}_{\mathrm{RP}} \in [0.6, 1]$, is dominated by random noise with some small scale structure just visible at $\boldsymbol{G} = 13.5$. This shows that the model has provided an extremely good fit to the data.

there is less extinction and produce a higher selection probability in dusty regions of the sky.

As discussed earlier, the RVS sample is selected on proxies for the apparent magnitude in the RVS waveband, G_{RVS}. This is quite close to the G_{RP} waveband. A sharp cut in G_{RP} will lead to a correlated cut between G and $G - G_{\mathrm{RP}}$. This is exactly what we see at the faint end of Fig. 4.15. For the bulge line of sight, the selection probability also drops off at much brighter magnitudes due to the crowding limits of the spectrograph. The narrow range of colour with non-zero selection probability is reflective of the RVS cuts on source effective temperature. The effect of the G_{RVS} cut is most distinctive in the third row, second column of Fig. 4.16 where the complex structure of the IGSL is very prominent.

The p-value test, shown across the sky in Fig. 4.17 for $G - G_{\mathrm{RP}} = 0.8$, demonstrates that the model is correctly representing the data across most magnitudes. The third panel of Fig. 4.10 also shows this with only one magnitude bin showing any poorly fit pixels with $G \sim 14$. This may be the model struggling to reproduce sharp changes in the behaviour of the IGSL.

As I saw for the RUWE model, the RVS selection function shoots up at $G \lesssim 3$ which can be seen in both Fig. 4.11 and Fig. 4.15 due to a lack of objects in the source catalogue. Users should be wary of this when applying the selection functions at these magnitudes.

4.6.4 RVS and Astrometry or RUWE

As well as individual subsets, various science cases will also require the intersection of subsets. For example, if one wanted full 6D phase space information for all objects in their sample, they'd want the subset of *Gaia* which has both published radial velocities **and** proper motion and parallax.

The selection function for the intersection of two subsets is given by

$$\mathrm{P}(\mathcal{S}_{\mathrm{subset1}}, \mathcal{S}_{\mathrm{subset2}} \,|\, \mathcal{S}_{\mathrm{source}}, \mathbf{y}) = \mathrm{P}(\mathcal{S}_{\mathrm{subset1}} \,|\, \mathcal{S}_{\mathrm{subset2}}, \mathcal{S}_{\mathrm{source}}, \mathbf{y}) \cdot \mathrm{P}(\mathcal{S}_{\mathrm{subset2}} \,|\, \mathcal{S}_{\mathrm{source}}, \mathbf{y}). \tag{4.32}$$

The RUWE sample only contains sources with parallax and proper motion. Therefore $\mathrm{P}(\mathcal{S}_{\mathrm{Astrometry}} \,|\, \mathcal{S}_{\mathrm{RUWE}}) = 1$ and the selection function for Astrometry and RUWE is equal to the RUWE selection function as expected.

If subset 1 and subset 2 are independent

$$\mathrm{P}(\mathcal{S}_{\mathrm{subset1}} \,|\, \mathcal{S}_{\mathrm{subset2}}, \mathcal{S}_{\mathrm{source}}, \mathbf{y}) = \mathrm{P}(\mathcal{S}_{\mathrm{subset1}} \,|\, \mathcal{S}_{\mathrm{source}}, \mathbf{y}) \tag{4.33}$$

and I can take the selection function for the intersection to be the product of the two individual selection functions.

I do this for the RVS sample where parallax and proper motion are provided and where RUWE < 1.4. There are 6 162 273 in *Gaia* EDR3 with both RVS and parallax and proper motion where 5 303 693 of these also have RUWE < 1.4.

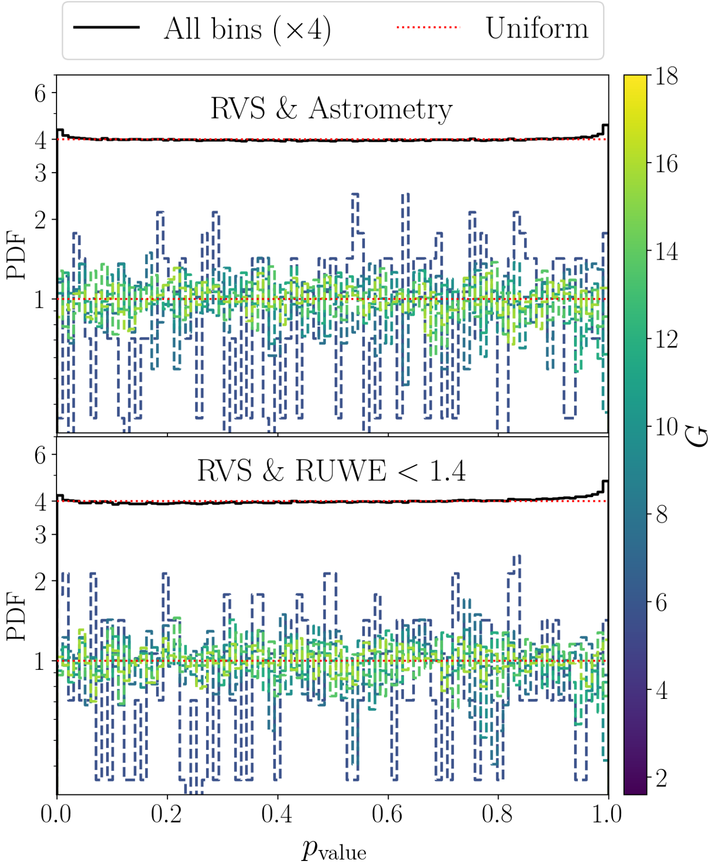


Fig. 4.18 Distribution of p-values for samples with both RVS and Astrometry (top) and RUWE < 1.4 (bottom) compared with the model assuming independent samples is close to uniform suggesting that this is a reasonable assumption. In this case I've only included bins where $n > 0$.

I take the product of the RVS position-colour-magnitude selection function defined in the previous section with the colour-independent selection functions for Astrometry

$$\mathrm{P}(\mathcal{S}_{\mathrm{RVS}}, \mathcal{S}_{\mathrm{Astrometry}} \,|\, \mathcal{S}_{Gaia}, \mathbf{y}) = \mathrm{P}(\mathcal{S}_{\mathrm{RVS}} \,|\, \mathcal{S}_{Gaia}, \mathbf{y}) \cdot \mathrm{P}(\mathcal{S}_{\mathrm{Astrometry}} \,|\, \mathcal{S}_{Gaia}, \mathbf{y}) \tag{4.34}$$

and RUWE

$$\mathrm{P}(\mathcal{S}_{\mathrm{RVS}}, \mathcal{S}_{\mathrm{RUWE}} \,|\, \mathcal{S}_{Gaia}, \mathbf{y}) = \mathrm{P}(\mathcal{S}_{\mathrm{RVS}} \,|\, \mathcal{S}_{Gaia}, \mathbf{y}) \cdot \mathrm{P}(\mathcal{S}_{\mathrm{RUWE}} \,|\, \mathcal{S}_{Gaia}, \mathbf{y}). \tag{4.35}$$

I can then test this against the data. Fig. 4.18 shows the distribution of p-values for this sample where I only include bins with $n > 0$ as bins with $n = 0$ give a uniformly distributed p-value independent of the model (see Appendix A.3). There is on significant sign of underfitting so I conclude that it is reasonable in this case to assume that RVS and Astrometry or RUWE samples are independent and therefore use the product of probabilities.

4.7 Other Samples

In this chapter I have produced selection functions for three scientifically important subsamples of the *Gaia* mission. However there are many catalogues in the *Gaia*-verse which are used by the astronomy community for a wide range of exciting research. Here I list some which were considered but which, due to additional complexity, I was not able to evaluate with the current method.

4.7.1 Variables

Variable sources are incredibly important as standard candles allowing us to measure precise distances to sources and systems across the Milky Way and in external galaxies (Muraveva et al., 2018; Riess et al., 2019). This has enabled detailed maps of the distribution of stars throughout the Milky Way disk and halo (Iorio & Belokurov, 2019; Skowron et al., 2019).

Holl et al. 2018 describes the *Gaia* DR2 variable sample following the processing detailed in Eyer et al. 2017. As discussed in Holl et al. 2018, overall completeness was not an aim of the DR2 variables sample. Considering only the lower limits placed on number of successful FoV transits required for variable processing they estimate that the average completeness is $\sim 80\%$ for the FoV ≥ 12 pipeline and as low as $\sim 51\%$ for FoV ≥ 20. This varies heavily across the sky due to the scanning law and crowding effects. Therefore a selection function for this sample would be hugely valuable for models of the Milky Way.

There are several reasons why the method I have applied here would not be directly applicable to variable samples. Firstly we need to consider what question we're asking. We want to know e.g. "Given that a source with l, b, G is an RRLyrae, what is the probability that *Gaia* would publish this source classified as an RRLyrae?". Written as a probability, this is

$$\mathrm{P}(\mathcal{S}_{\mathrm{RRLyrae}} \,|\, l, b, G, X_{\mathrm{RRLyrae}}) \tag{4.36}$$

where $X_{\rm RRLyrae}$ is the event that the source is an RRLyrae-type star. Were I to naively use the approach I've used in this chapter for other samples, I would actually evaluate the probability that a source is selected **and** it is an RRLyrae

$$\mathrm{P}(\mathcal{S}_{\rm RRLyrae}, X_{\rm RRLyrae}\,|\,l, b, G). \tag{4.37}$$

This is because I would be comparing the population of classified RRLyraes against the full sample of *Gaia* sources the vast majority of which are not RRLyrae stars. To model the selection function for variable samples, I would need to know which sources in *Gaia* **would have been classified** as a variable had they actually been one.

The second issue is in choosing the observables over which to define the selection function. The selection function for variable sources will be highly dependent on variability amplitude and period. *Gaia* observes stars with certain dominant frequencies due to the spin, precession and orbital dynamics of the satellite. This would make some variability periods harder to detect than others. The apparent G-band magnitude is also highly unreliable for variable sources (see Section 5.4 of Arenou et al., 2018). Improved apparent magnitudes may be determined from the time series however this adds another layer of complexity when comparing against the full *Gaia* dataset.

Finally, there are also many individual variable samples, such as RRLyrae, Cepheids and Long Period Variables, which are each used for separate science aims. A full study of the selection functions of variable sources in *Gaia* should aim to construct selection functions for each population.

The variables selection function for the *Gaia* mission is a significant, complex project worthy of dedicated further study.

4.7.2 Parallax SNR

$1/\varpi$ is often used as a distance estimate for sources with *Gaia* astrometry. However, for sources with low parallax signal-to-noise (SNR) the distance uncertainty distribution can be highly asymmetric and the measured parallax is negative for many sources in *Gaia* EDR3 (see Bailer-Jones, 2015b; Luri et al., 2018, for full discussion).

A common selection to avoid this is taking only sources with $\varpi/\sigma_\varpi > X$ where commonly used values for X are 5 or 10 depending on the scientific aims. This choice significantly affects the selection function as a function of distance and as a result, the inferred distance to sources can be heavily biased (see Schönrich & Aumer, 2017, for a detailed discussion).

One solution to this was demonstrated by Schönrich et al. 2019 who infer a distance-dependent selection function for the *Gaia* DR2 RVS sample. The limitations of the method used here is that they must assume an a-priori model for the distribution of sources in the Milky Way. As a result, their selection function is not model independent. Their selection function is also sky-averaged for the entire sample and doesn't look to quantify any of the complex variation of the selection function on small spatial scales.

I could use the method from Boubert & Everall 2021 directly on the $\varpi/\sigma_{\varpi} > 5$ sample as a subset of the *Gaia* catalogue and evaluate the selection function as a function of $G, G - G_{\mathrm{RP}}, l, b$. Indeed, σ_{ϖ} is a strong function of position on the sky due to the scanning law and a function of G due to CCD photon noise as I will show in Chapter 5.

However the selection is also directly dependent on ϖ so any selection function for a parallax SNR selected sample would need to feature measured ϖ or observables which are directly dependent on it. For this reason, I do not attempt to model the parallax SNR selection function here.

4.8 Accessing the Selection Functions

The selection functions are accessibly through the GitHub repository SELECTIONFUNCTIONS (https://github.com/gaiaverse/selectionfunctions). Here I provide a brief example of how to load the EDR3 source catalogue and RVS subsample selection functions to estimate the probability that S5-HVS1 (Koposov et al., 2020) could have been observed in *Gaia* EDR3 with radial velocity measure from DR2. The inferred probability is $\sim 3 \times 10^{-9}$ and indeed S5-HVS1 does not have published radial velocity.

```
import selectionfunctions.cog_v as CoGV
import selectionfunctions.cog_ii as CoGII
from selectionfunctions.source import Source

# Load DR3 selection function
dr3_sf = CoGII.dr3_sf(version='modelAB',crowding=False)

# Load RVS selection function
rvs_sf = CoGV.subset_sf(map_fname='rvs_cogv.h5')

# Introduce source
s5_hvs1 = Source('22h54m51.68s',
           '-51d11m44.19s',
           photometry={'gaia_g':16.02,
                       'gaia_bp_gaia_rp':-0.008},
           frame='icrs')

# Selection probability
print(dr3_sf(s5_hvs1) * rvs_sf(s5_hvs1))
```

4.9 Summary of an incomplete *Gaia*-verse

I have introduced the selection function for the *Gaia* source catalogue and provided a first-order estimate of the *Gaia* EDR3 selection function to be used with the subset selection functions.

I have evaluated the selection functions for three subsamples of the *Gaia* EDR3 data release: sources with published parallax and proper motion, the subset with RUWE < 1.4 and sources with DR2 radial velocities published in the EDR3 source catalogue.

The astrometry and RUWE selection functions are evaluated as a function of apparent magnitude and position on the sky showing a strong dependence on crowding and features of the *Gaia* scanning law. I also evaluated the RUWE and RVS selection probabilities as a function of $G - G_{\mathrm{RP}}$ colour relative to the *Gaia* sample with published G_{RP}. For RVS, the effects of temperature cuts on the sample and selection on G_{RVS} using IGSL magnitudes heavily impact the selection function.

In all samples the model is able to reproduce the data down to scales ~ 2 degrees. The only areas in which the model breaks down are where the selection probability varies on smaller scales and my spatial model doesn't have the flexibility to reproduce the data. This is an issue in the Milky Way bulge and around globular clusters.

In Chapter 6 I demonstrate the immense value of these selection functions for estimating the vertical structure of the Milky Way. Before that, there is one remaining piece of the puzzle to fit. I need to understand the principles and systematics of the astrometric solutions provided by *Gaia*.

5

The Astrometric Spread Function

"The Gaia astrometric processing must relate individual visits to positions, proper motions and parallaxes by linear regression, otherwise we're hosed."

David Hogg 2021

This was David Hogg's response when I mentioned that the astrometric solution used in the *Gaia*'s pipeline, was, in fact, not pure linear regression but rather iterative linear regression. In spite of this, we can make some interesting statistical predictions of the astrometric uncertainties under the assumption of pure linear regression, which is the subject of this chapter.

5.1 What is an Astrometry Spread Function?

Gaia has initiated an era of large scale Milky Way dynamical modelling by providing 5D astrometry (position, proper motion and parallax) for more than 1.3 billion stars (Gaia Collaboration et al., 2016, 2018a; Lindegren et al., 2018). The *Gaia* satellite measures source positions at multiple epochs over the mission lifetime. These epoch astrometry measurements are the inputs of the Astrometric Global Iterative Solution (AGIS Lindegren et al., 2012) which iteratively solves for the spacecraft attitude, geometric calibration of the instrument, global parameters, and 5D astrometry of each source: the right ascension and declination (α_0, δ_0), the proper motions (μ_α, μ_δ) and the parallax (ϖ). Alongside the source astrometry, *Gaia* also publishes the 5D astrometric measurement covariance and various statistics of the astrometric solution for all sources which meet the quality cuts. The 5-parameter astrometric model of AGIS assumes sources are point-like with apparent non-accelerating uniform motion relative to the solar system barycentre which I refer to as 'simple point sources'.

Both resolved and unresolved binary stars accelerate due to their orbits around the common centre of mass which shifts the centroid off a uniform motion trajectory. Using this shift, it is expected that *Gaia* can characterise the orbits of stars with brown dwarf companions out to 10 pc and black hole companions out to more than 1 kpc when considering tight constraints of the uncertainty on the mass function $M_2^3 M_{\mathrm{tot}}^{-2}$ (Andrews et al., 2019). When one is interested in (the less constraining) orbital parameter recovery with $\sim$ 10% precision, *Gaia* might detect a staggering 20 k brown dwarfs around FGK-stars out to many tens up to a few hundreds of pc for the longer period objects (100-3000 d), which could reach even 50 k out to several hundred pc when one is only interested in the detection of BD candidates e.g. for follow-up studies (Holl et al., 2021), with black hole companions being detectable out to several kpc.

Similarly exoplanet orbits pull their host stars away from uniform motion although with a much smaller amplitude due to the lower companion mass. From simulations it is expected that *Gaia* is capable of detecting 21,000 long-period, 1-15 Jupiter mass planets during the 5 year mission (Perryman et al., 2014), more than 4 times the number of currently known exoplanets. Ranalli et al. 2018 have further demonstrated that the 5 year *Gaia* mission will be able to find Jupiter-mass planets on 3 au orbits around $1M_\odot$ stars out to 39 pc and Neptune-mass planets out to 1.9 pc. Not only will the presence of planets be detectable but it is expected that ~ 500 planets around M-dwarfs will receive mass constraints purely from *Gaia* astrometry (Casertano et al., 2008; Sozzetti et al., 2014).

Microlensing occurs when the light from a background source is gravitationally lensed by a foreground lensing star causing a shift in the apparent position of the source, detectable by high precision astrometric surveys (Miralda-Escude, 1996). The deflection can be used as a direct measurement of the lens mass as demonstrated by Kains et al. 2017 using HST observations. A signficant amount of work has gone towards predicting microlensing events using *Gaia* proper motions (Bramich, 2018; Klüter et al., 2018; McGill et al., 2019) with 528 events expected in the extended *Gaia* mission ~ 39% of which pass astrometry quality cuts (McGill et al., 2020). For a small number of these events *Gaia* will be able to determine the lens mass to < 30% uncertainty (Klüter et al., 2020).

Extended sources such as galaxies will have a reduced astrometric precision from each *Gaia* observation due to the increased spread of flux. *Gaia* scans a source in many different directions over the mission lifetime from which the source shape can be reconstructed (Harrison, 2011). With the *Gaia* epoch astrometry for the 5 year mission, *Gaia* will be able to distinguish between elliptical and spiral/irregular galaxies with ~ 83% accuracy (Krone-Martins et al., 2013). These classifications would be incredibly valuable for galaxy morphology studies.

The *Gaia* epoch astrometry will be first released in DR4, several years from now. However, a (very) condensed form of this large amount of information is stored in the summary statistics of the astrometric solution currently published in *Gaia* DR2 and updated in EDR3. Binary stars, exoplanet hosts, microlensing events and extended sources will induce an excess noise in the astrometric solution as they are not well described by simple point sources. This excess noise has been modelled for binaries (Penoyre et al., 2020; Wielen, 1997) and already Belokurov et al. 2020b has found many binaries in *Gaia* DR2 using the renormalised unit weight error statistic, RUWE, that is the re-normalised square root of the reduced χ^2 statistic of the astrometric solution.

RUWE is a 1D summary statistic of the residuals of the 5-parameter astrometric solution of a source relative to the *Gaia* inertial rest frame. But we can glean even more information on the excess noise from the 5D uncertainty of the astrometric solution. The uncertainty in the 5D astrometric solution for a source in *Gaia* can be expressed as the convolution of *Gaia*'s astrometric measurement uncertainty expected for a simple point source and excess noise. I term *Gaia*'s expected astrometric measurement uncertainty the Astrometry Spread Function (ASF) defined as the probability of measuring a simple point source to have astrometry $\mathbf{r}' \in \mathbb{R}^5$ given the true source astrometry $\mathbf{r} \in \mathbb{R}^5$ and apparent magnitude G

$$\mathrm{ASF}(\mathbf{r}') = \mathrm{P}(\mathbf{r}' \,|\, \mathbf{r}, G). \tag{5.1}$$

The excess noise will be driven by un-modelled source characteristics such as binary motion, exoplanet host motion, microlensing or extended source flux as well as any calibration noise which is not accounted for in the ASF. In this work, I assume that all significant calibration effects are included in the ASF such that the excess noise is dominated by un-modelled source characteristics. However this assumption breaks down in some regimes, particularly for bright sources in crowded regions where CCD saturation becomes a significant issue. Possible un-accounted calibration effects should be considered when using *Gaia* astrometry to search for excess noise due to genuine un-modelled source characteristics.

Since the astrometric solution is evaluated using least squares regression, the ASF is Gaussian distributed

$$\mathrm{ASF}(\mathbf{r}') = \mathcal{N}(\mathbf{r}'\,;\, \mathbf{r}, \Sigma(l, b, G)) \tag{5.2}$$

where $\Sigma(l, b, G) \in \mathbb{R}^{5\times5}$ is the expected covariance for a simple point source with position l, b and apparent magnitude G as measured by *Gaia*. The astrometric calibration is also a function of source colour which was either estimated from $G_{\mathrm{BP}} - G_{\mathrm{RP}}$ or added as a sixth parameter of the astrometric solution, ASTROMETRIC_PSEUDO_COLOUR. As colours are only published for a subset of the *Gaia* catalogue and the astrometric correlation coefficients for pseudo-colour are not published in DR2, I neglect colour dependence of the astrometric solution in this work. For EDR3, all pseudo-colour correlation coefficients are published and it is worth considering how this impacts the ASF.

Given the ASF and published astrometric 5-parameter model uncertainties I can reconstruct the 5D excess noise and use it to characterise binary systems, exoplanet orbits, microlensing events and extended sources in *Gaia* without requiring the epoch astrometry. The focus of this chapter is to construct the ASF for *Gaia* DR2.

This builds on analysis of the scanning law from Boubert et al. 2020 and Boubert et al. 2021b and is used, in conjunction with the results of Boubert & Everall 2020 to determine the selection function for the subsample of *Gaia* DR2 with published 5D astrometry.

In Section 5.2 I provide a whistle-stop tour of the *Gaia* spacecraft, scanning law and how this translates to constraints on the position, proper motion and parallax of sources. This chapter is focused on *Gaia* DR2 for which I estimate the ASF, although I note that the method is be directly applicable to *Gaia* EDR3. The method for constructing the ASF of *Gaia* is derived in Section 5.3 and the results compared with the astrometry sample are shown in Section 5.4. I also use the ASF for an alternative derivation of the Unit Weight Error demonstrating the applicability of the method in Section 5.5.

As a secondary motivation, the *Gaia* DR2 5D astrometry sample is selected from the full catalogue with a cut on the parameter ASTROMETRIC_SIGMA5D_MAX which is a function of the astrometric covariance matrix. In predicting the astrometric covariance for simple point sources, I can also estimate the contribution from this cut to the astrometric selection function which I present in Section 5.6. Finally I discuss applications of the ASF in Section 5.7.

5.2 Astrometry with Gaia

5.2.1 The Scanning Law

The *Gaia* spacecraft is in orbit around the Lagrange point 2 (L2), orbiting the Sun in phase with the Earth. The spacecraft spins with a 6 hour period around a central axis which precesses with an aspect angle of 45 degrees around the pointing connecting the satellite and the Sun, with a 63 day period. This is similar to a spinning top which has been left long enough to wobble. The orbit of the spacecraft around the sun adds a third axis of rotation.

Perpendicular to the spin axis, two fields of view (FoV) observe in directions separated by 106.5 deg. The direction in which each FoV is pointing at any point in time throughout *Gaia*'s observing period is the scanning law.

The *Gaia* DR2 observing period runs from July 25 2014 (10:30 UTC) until May 23 2016 (11:35 UTC). The scanning law for DR2 is published by DPAC and refined by Boubert et al. 2020 and Boubert et al. 2021b. Whilst this tells us where *Gaia* was pointing, it doesn't tell us whether Gaia was obtaining useful scientific measurements that contributed to the published data products. Many time periods in the DR2 window did not result in measurements which contributed to the *Gaia* astrometry as discussed in Boubert et al. 2021b.

In this chapter I only include the scanning law in the OBMT[1] interval 1192.13–3750.56 rev (Lindegren et al., 2018, hereafter L18) which removes the Ecliptic Polar Scanning Law, an initial calibration phase of *Gaia* which contributed to the published photometry but not astrometry. DPAC have published a series of additional gaps in astrometry data taking[2]. I remove any time spans of the scanning law for which the gap is flagged as 'persistent'. Using the published Epoch Photometry for 550 737 variable sources (Evans et al., 2018; Holl et al., 2018; Riello et al., 2018), Boubert et al. 2020, 2021b constrained additional gaps which are persistent across all data products of *Gaia* DR2 which I also remove from the scanning law. Finally, Boubert et al. 2021b determines the probability of an observation being recorded and used in *Gaia* DR2 in 19 magnitude bins (the *deletions* discussed in Chapter 4). These observation probabilities are be used to weight observations in the ASF in Section 5.3.3.

[1] Onboard Mission Time (OBMT) is the timing system used in *Gaia* and is normalised such that OBMT is 0 in October 2013 and increments by 1 for every revolution of the *Gaia* satellite which corresponds to 6 hours

[2] https://www.cosmos.esa.int/web/gaia/dr2-data-gaps

5.2.2 Taking Observations

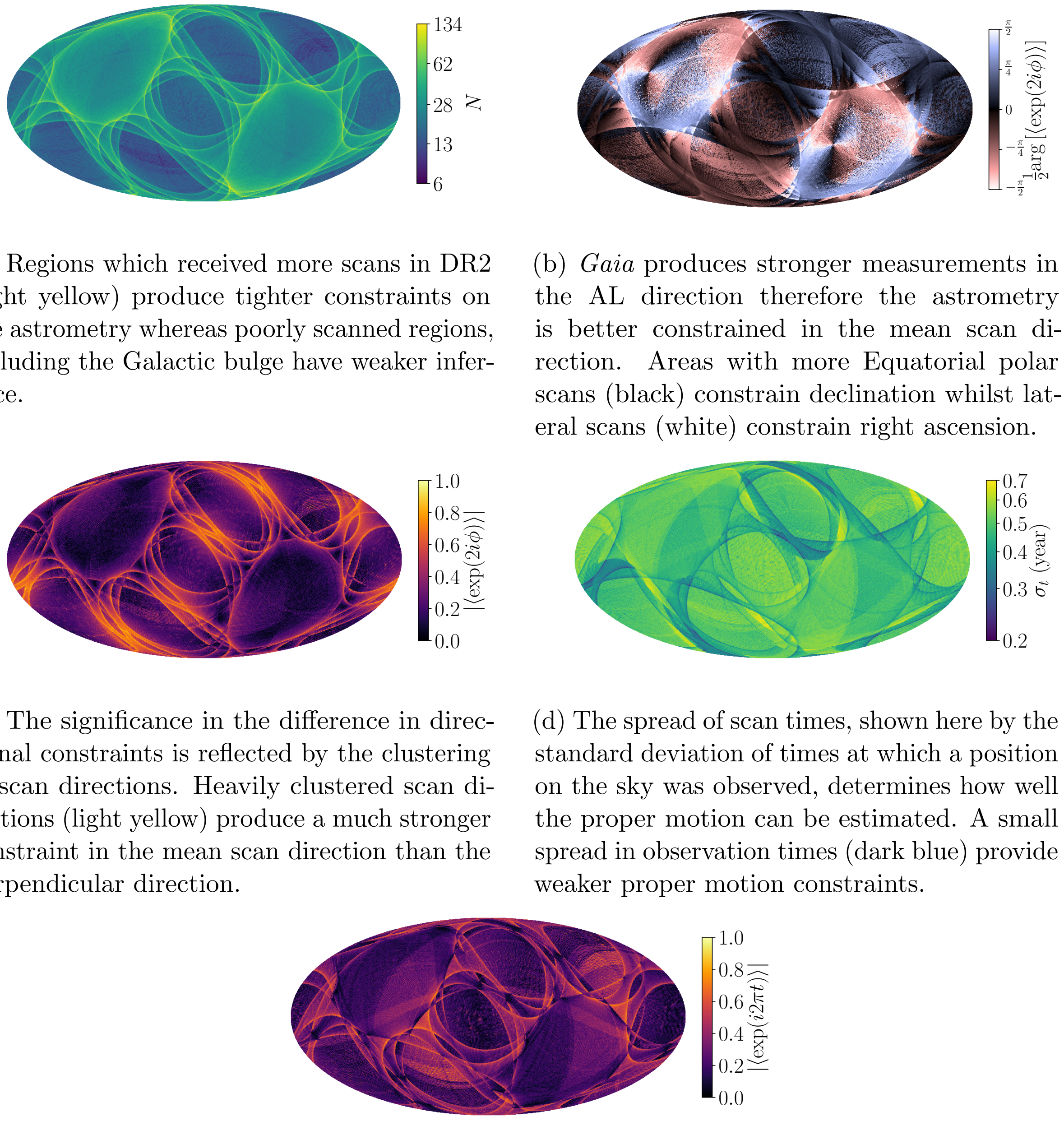
Both FoVs project source images onto a single panel of CCDs called the focal plane. On the focal plane there are 9 columns and 7 rows of CCDs, referred to as the astrometric field (AF), which measure the position of a source although the middle row only has 8 CCDs (because one of the 9 CCD positions is taken by a wave front sensor). As the spacecraft spins, stars track across the CCD panel in the along-scan direction and are observed with up to 9 astrometric CCDs during a single FoV transit (see e.g. Fig. 1 of Lindegren et al., 2016). Individual CCD measurements are referred to as observations whilst a full track across the CCD panel is a scan (also referred to as a FoV transit). Before the AF, sources pass over the 'Sky Mapper' (SM) CCD which triggers the initial detection and needs to be confirmed by the first AF CCD in order for any observations within the scan to successfully provide a measurement. Each observation records the position and apparent brightness of the source. If the source is recorded with $G < 13$ by the SM, a 2D observation window is assigned measuring position in the along-scan (AL) direction and orthogonal across-scan (AC) direction. For fainter stars with $G > 13$, only the AL position is recorded.

Observations are saved on-board *Gaia* in 'Star Packets' grouped by apparent magnitude in 19 bins (Table 1.10, Section 1.3.3 in de Bruijne et al., 2018). The majority of data is uploaded to Earth, however some can be lost or deleted (see Section 3.3 of Gaia Collaboration et al., 2016) changing the scanning law sampling for stars in different Star Packet magnitude bins.

After a first process of CCD signal level, background, and PSF/LSF calibration, the data is input to the AGIS pipeline (Lindegren et al., 2012) which uses an iterative linear regression algorithm to simultaneously fit the attitude of the spacecraft, a large number of calibration parameters, and the position, proper motion and parallax of all sources in *Gaia* DR2. I here provide a general description of how position, proper motion and parallax can be understood to depend on the nature of the observations a source receives, though in *Gaia* they are simultaneously solved from the offsets of all source observations with respect to its (iteratively improved) internal reference system.

The precision with which position, proper motion and parallax of a source can be measured is heavily dependent on magnitude (beyond $G > 13$ the uncertainties monotonically increase with magnitude), the number of observations taken, the scan directions of these observations and their distribution in time. More observations produce a greater precision therefore sources in regions of the sky with the most scans as shown in Fig. 5.1a have the best constrained astrometry. Notably, the Galactic centre in the middle of the plot has received only ~ 10 scans whilst the best observed regions of the sky are scanned over 100 times.

For the vast majority of sources, *Gaia* only measures position in the AL direction and even for 2D observations, the AL position constraint is much tighter than the AC measurement (Lindegren et al., 2012). Therefore North-South scans in equatorial coordinates constrain declination, δ whilst East-West scans constrain right ascension, α. Fig. 5.1b gives the mean direction of *Gaia* DR2 scans modulo π such that a North-South and South-North



(a) Regions which received more scans in DR2 (light yellow) produce tighter constraints on the astrometry whereas poorly scanned regions, including the Galactic bulge have weaker inference.

(b) *Gaia* produces stronger measurements in the AL direction therefore the astrometry is better constrained in the mean scan direction. Areas with more Equatorial polar scans (black) constrain declination whilst lateral scans (white) constrain right ascension.

(c) The significance in the difference in directional constraints is reflected by the clustering of scan directions. Heavily clustered scan directions (light yellow) produce a much stronger constraint in the mean scan direction than the perpendicular direction.

(d) The spread of scan times, shown here by the standard deviation of times at which a position on the sky was observed, determines how well the proper motion can be estimated. A small spread in observation times (dark blue) provide weaker proper motion constraints.

(e) Measuring parallax requires position measurements throughout the year. Scans clustered at one time of year (light yellow) produce a weaker parallax measurement than a spread of scans through the year (dark purple).

Fig. 5.1 The precision with which *Gaia* measures 5D astrometry is heavily dependent on the number of FoV transits, scan times and directions, obtained from the scanning law. The plots provide some central summary statistics of the scanning law as a function of position on the sky in Galactic coordinates on HEALPix level 7.

scan appear the same with $\langle\phi\rangle = 0$. The mean direction is estimated from the argument of the mean vector

$$\langle\phi\rangle = \frac{1}{2}\arg\left[\langle\exp(2i\phi)\rangle\right]. \tag{5.3}$$

This statistic is published for sources in *Gaia* EDR3 as SCAN_DIRECTION_MEAN_K2. Darker areas have stronger δ constraints whilst lighter areas constrain α more tightly. The difference in accuracy in right ascension and declination depends on the clustering of scan directions. The absolute value of the mean scan vector, $|\langle\exp(2i\phi)\rangle|$ which is ~ 1 for heavily clustered scans and ~ 0 for a spread of scan directions. This is shown in Fig. 5.1c where light areas strongly constrain position in the mean scan direction but only provide a weak constraint in the perpendicular direction whilst dark regions have a spread of scan directions and therefore won't show a strong direction preference. This statistic is also published in *Gaia* EDR3 as SCAN_DIRECTION_STRENGTH_K2.

Constraints on the source proper motion come from measuring the position of a source at multiple different epochs and estimating the rate of change. A larger spread of observation times produce a tighter proper motion constraint. Fig. 5.1d shows the standard deviation of observation times with light regions producing tighter proper motion constraints whilst dark regions produce a weaker constraint.

Finally, source parallax is estimated from the apparent motion of a source relative to the background of distant sources due to *Gaia*'s motion around the sun on a one year period. A larger spread of observations throughout the year produce a tighter constraint on the source parallax. The position of an observation in the yearly solar orbit is described by the complex vector $\exp(2\pi it)$. As with the scan direction, the clustering of observations in the year is estimated from the absolute value of the mean vector $|\langle\exp(2\pi it)\rangle|$. If observations are heavily clustered at one time of year the absolute mean is close to 1, shown by lighter areas of Fig. 5.1e, and only a weak constraint on parallax are achieved. Values close to 0 have well spread observations throughout the year and therefore provide a stronger constraint on parallax.

5.2.3 Data

The previous sections have provided a qualitative prediction of *Gaia*'s expected performance as a function of position on the sky. In the following sections I produce a quantitative estimate of the predicted precision with which *Gaia* can measure source astrometry as a function of position on the sky and apparent magnitude.

Gaia DR2 (Gaia Collaboration et al., 2016, 2018a) provides 5D astrometry for 1 331 909 727 of the 1 692 919 135 source in the full DR2 catalogue. To test my predictions, I use the full *Gaia* DR2 source catalogue and 5D astrometry sample.

5.3 Method

As input AGIS takes the 1D measurement of the position of each source in the AL and, for bright sources, also AC direction. For bright sources, *Gaia* produces a 2D observation however AGIS assumes the constraints in the AL and AC directions are uncorrelated treating them as independent 1D observations. I make the same assumption in this work. This is a gross simplification of all the steps which AGIS takes – for instance, calibrating the satellite attitude noise – however it allows for a very appealing and tractable derivation of the ASF from the available data.

I proceed with four key assumptions:

- The 1D position measurement uncertainty is Gaussian.
- Individual measurements, including AL and AC measurements from the same observation, are independent and uncorrelated. As the AGIS pipeline uses the same assumption, this produces no discrepancy between my predictions and the published *Gaia* astrometry.
- The position measurement uncertainty is a function of source apparent magnitude at the time of observation only. Any dependence of the observation precision of the satellite as a function of time for a given apparent magnitude is neglected which I justify in Section 5.3.1. This also assumes that the measurement uncertainty is colour independent.
- Astrometric parameters of different sources are assumed to be independent. In reality measurements of different sources can be considered independent, however due to the joint estimation of the attitude and geometric calibration from the same set of observations, the posterior astrometric parameters are correlated. Pre-launch estimates by Holl et al. 2010 predicted correlations of only a fraction of a percent for sources separated by less than one degree (in a fully calibrated AGIS solution dominated by photon noise). DR1 (see Section D.3 of Lindegren et al., 2016) seems to be well above that with correlation as high as perhaps 0.25 at separations up to 1 degree, though much smaller on longer scales. Studies on the quasar sample in DR2 (see Section 5.4 of Lindegren et al., 2018) show still very large covariances as small spatial scales (< 0.125 deg) and milder effects over larger spatial scales. With each successive data release it is expected that these spatial correlations will shrink, though they will never be zero, especially at small scales.

Throughout this chapter I also only consider sources with constant magnitude to keep the results simple and tractable however the formalism is easily generalisable to variable sources. Over this section I derive the ASF of *Gaia* DR2. As many different variables are introduced, I refer the reader to Table 5.1 to clarify my notation.

5.3.1 Time Dependence

To estimate the *Gaia* astrometric precision, I have assumed that all observations at the same apparent magnitude have the same precision. This assumption breaks down if the precision of *Gaia* is time dependent.

Without any epoch astrometry, it is challenging to assess the scale of the impact of this time dependence on the posterior precision. We are however provided epoch photometry for 550 737 sources. The astrometric uncertainty should scale with $\sigma_{\mathrm{AL}} \sim \sigma_f/f$ where f is the observed flux of the source as both are dominated by photon count noise. The centroid is actually mainly sensitive to the slope of the wings of the LSF/PSF while the

$\mathbf{x}$	$\mathbb{R}^N$	True AL position of source at observation time.
$\mathbf{x}'$	$\mathbb{R}^N$	Measured AL position of source at observation time.
$\mathbf{r}$	$\mathbb{R}^5$	Source astrometry.
M	$\mathbb{R}^{N\times5}$	Design matrix of astrometric solution.
K	$\mathbb{R}^{N\times N}$	Expected measurement covariance.
C	$\mathbb{R}^{5\times5}$	Published astrometry covariance.
Σ	$\mathbb{R}^{5\times5}$	Expected astrometry covariance ($\Sigma = \rho\,\Phi$).
ρ	$\mathbb{R}$	Magnitude dependence of ASF.
Φ	$\mathbb{R}^{5\times5}$	Spatial dependence of ASF from scanning law.

Table 5.1 Notation followed in linear regression.

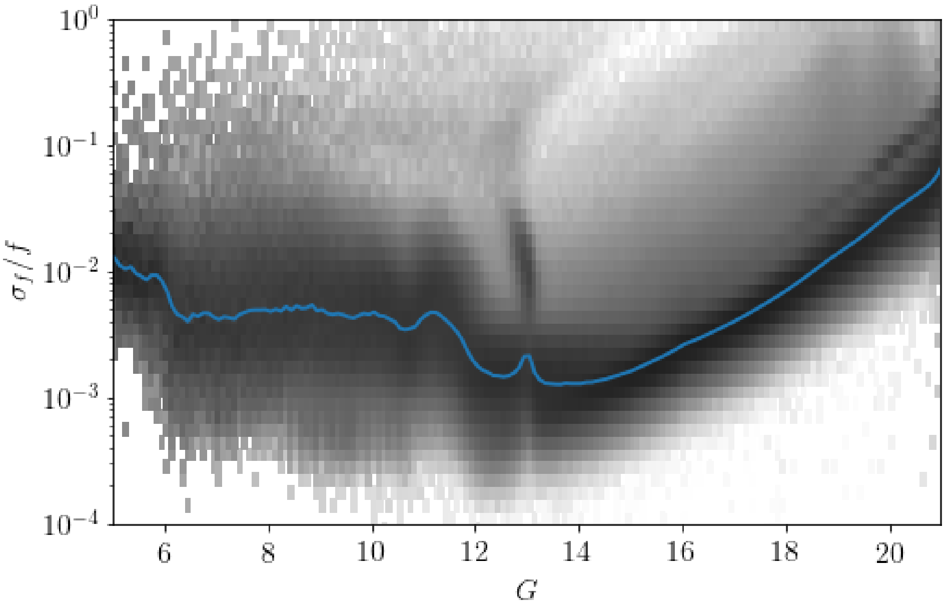


Fig. 5.2 The relative flux error as a function of magnitude for all observations in the *Gaia* epoch photometry (Evans et al., 2018; Holl et al., 2018; Riello et al., 2018) (black histograms, log normalised) shows complex structure due to changes in window class configurations (Riello et al., 2018). The running median (blue solid line) shows similar structure to σ_{AL} in Fig. 5.4 with both dominated by photon count noise.

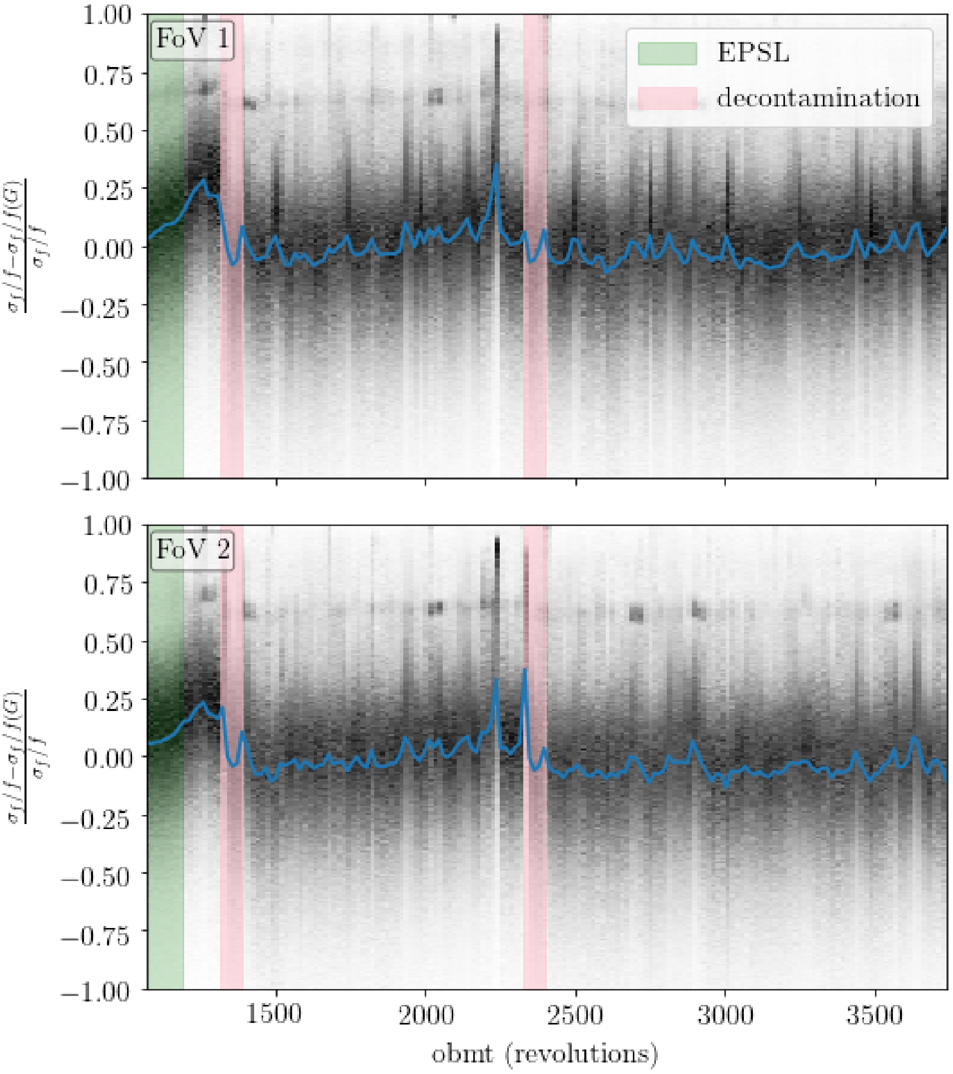


Fig. 5.3 The relative flux error, recentred and renormalised by the median as a function of magnitude from Fig. 5.2, varies as a function of time throughout *Gaia* DR2 due to mirror contamination, micro-meteoroid impacts and variations in background flux. Data taken during the Ecliptic Polar Scanning Law (green) and decontamination events (pink) are not used in the *Gaia* DR2 astrometry. The running median (blue solid) of the observations does not vary significantly from zero relative to the variance in individual measurements. This justifies the assumption that the astrometric measurement uncertainty is not a strong function of time.

flux measurement to the core (as there is the most signal), but to first degree this relation should hold to understand how the centroid uncertainty depends on magnitude and time.

I take all 17 712 391 observations in the epoch photometry and find the median σ_f/f in 0.2 mag bins shown in Fig. 5.2. I subtract the median off all data leaving the residuals. The distribution of residual errors against observation time is given in Fig. 5.3 for each *Gaia* field-of-view (FoV). As *Gaia* operates, material from the satellite condenses and accumulates on the mirrors scattering light and reducing the precision of observations. To mitigate this, the spacecraft was heated up to evaporate the condensation and clean the mirrors (see Section 4.2.1 Gaia Collaboration et al., 2016). These decontamination events (pink shaded regions) have taken place twice in Gaia DR2 (L18). The impact of the first decontamination event on the flux error is significant however at later times, the measurement precision does not degrade appreciably. In fact, any longer term trends are insignificant compared to the short term fluctuations on 10 rev timescales.

The epoch flux supports the conjecture that the measurement precision of Gaia does not significantly change with time.

5.3.2 Astrometry from Linear Regression

Gaia's goal for each source in the astrometry catalogue is to measure the five parameter astrometric solution, $\mathbf{r} \in \mathbb{R}^5$, consisting of the positions, proper motions and parallax. The AGIS pipeline estimates the astrometry of sources through linear regression on all observations of a single source in a step called source update (see Section 5.1 of Lindegren et al., 2012). I use the same technique to determine the expected precision for a simple point source as a function of apparent magnitude and position on the sky.

Take N observations of a source at times t_i with the scan direction ϕ_i where $i \in \{1, \ldots, N\}$. The on-sky positions of the source at times t_i in ICRS coordinates are (α_i, δ_i). The source position relative to the solar system barycentre at the reference epoch (J2015.5 for *Gaia* DR2 L18) is α_0, δ_0. The position at time t_i is a linear combination of the position at a reference epoch with the proper motion and parallax motion. The offset due to parallax motion is given by

$$\Delta\alpha_i \cos\delta_i = -\varpi\left(-X_i \sin\alpha_0 + Y_i \cos\alpha_0\right) = \varpi\Pi_{\alpha_i} \tag{5.4}$$

$$\Delta\delta_i = -\varpi\left(-X_i \cos\alpha_0 \sin\delta_0 - Y_i \sin\alpha_0 \sin\delta_0 + Z_i \cos\delta_0\right) = \varpi\Pi_{\delta_i} \tag{5.5}$$

where X_i, Y_i, Z_i are the barycentric coordinates of *Gaia* at time t_i and ϖ is the parallax of the source. I have assumed the parallax and proper motion are small enough such that the parallax ellipse is only dependent on the source reference epoch position which keeps the system of equations linear.

Therefore the position of the source at time t_i is given by

$$\alpha_i^* = \alpha_0^* + \mu_{\alpha^*} t_i + \varpi\Pi_{\alpha_i} \tag{5.6}$$

$$\delta_i = \delta_0 + \mu_\delta t_i + \varpi\Pi_{\delta_i} \tag{5.7}$$

where μ_α, μ_δ is the source proper motion, t_i is the time relative to the reference epoch and I use the notation $\alpha^* = \alpha\cos(\delta)$ and $\mu_{\alpha^*} = \mu_\alpha\cos(\delta)$.

Writing this set of linear equations out in matrix notation:

$$\begin{bmatrix} \alpha_1^* \\ \delta_1 \\ \vdots \\ \alpha_N^* \\ \delta_N \end{bmatrix} = \begin{bmatrix} 1 & 0 & \Pi_{\alpha_1} & t_1 & 0 \\ 0 & 1 & \Pi_{\delta_1} & 0 & t_1 \\ & & \vdots & & \\ 1 & 0 & \Pi_{\alpha_N} & t_N & 0 \\ 0 & 1 & \Pi_{\delta_N} & 0 & t_N \end{bmatrix} \begin{bmatrix} \alpha_0^* \\ \delta_0 \\ \varpi \\ \mu_{\alpha^*} \\ \mu_\delta \end{bmatrix}. \tag{5.8}$$

Our measurables are 1D positions in either the AL or AC direction of the *Gaia* focal plane. This is given by $x_i = \alpha_i^* \sin\phi_i + \delta_i \cos\phi_i$ where the scan position angle, ϕ_i is the scan direction of *Gaia* at the observation time for AL observations (shifted by $\pi/2$ for AC observations) and is defined such that $\phi = 0°$ in the direction of local Equatorial North, and $\phi = 90°$ towards local East[3].

Substituting x_i into Eq. 5.8:

$$\begin{aligned} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix} &= \begin{bmatrix} s_1 & c_1 & \Pi_{\alpha_1}s_1 + \Pi_{\delta_1}c_1 & t_1s_1 & t_1c_1 \\ s_2 & c_2 & \Pi_{\alpha_2}s_2 + \Pi_{\delta_2}c_2 & t_2s_2 & t_2c_2 \\ & & \vdots & & \\ s_N & c_N & \Pi_{\alpha_N}s_N + \Pi_{\delta_N}c_N & t_Ns_N & t_Nc_N \end{bmatrix} \begin{bmatrix} \alpha_0^* \\ \delta_0 \\ \varpi \\ \mu_{\alpha^*} \\ \mu_\delta \end{bmatrix} \\ &= \mathbf{Mr} \end{aligned} \tag{5.9}$$

where $c_i = \cos\phi_i$ and $s_i = \sin\phi_i$, $\mathbf{M} \in \mathbb{R}^{N\times5}$ is the design matrix for the linear equations and $\mathbf{r} \in \mathbb{R}^5$ is the vector of astrometric parameters.

Assuming Gaussian measurement uncertainty for both AL and AC measurements and assuming all observations are independent, the observed source positions are distributed $\mathbf{x}' \sim \mathcal{N}(\mathbf{x}, \mathbf{K})$ where the covariance matrix $\mathbf{K} = \mathrm{diag}\left[\sigma_1^2, \ldots, \sigma_N^2\right]$. This measurement covariance implicitly assumes all observations are independent and uncorrelated, one of my key assumptions also adopted in AGIS.

Following standard linear least squares regression (Hogg et al., 2010), the astrometric uncertainty covariance matrix of the inferred $\mathbf{r}$ is given by

$$\Sigma^{-1} = \mathbf{M}^{\mathrm{T}}\mathbf{K}^{-1}\mathbf{M}. \tag{5.10}$$

[3] https://www.cosmos.esa.int/web/gaia/scanning-law-pointings

Expanding this out in terms of all scan angles, the full inverse covariance matrix is given by

$$\Sigma^{-1} = \sum_{i=1}^{N} \frac{1}{\sigma_i^2} \mathbf{A}_i \tag{5.11}$$

with

$$\mathbf{A}_i = \begin{bmatrix} s_i^2 & s_i c_i & s_i \Pi_i & s_i^2 t_i & s_i c_i t_i \\ s_i c_i & c_i^2 & c_i \Pi_i & c_i s_i t_i & c_i^2 t_i \\ s_i \Pi_i & c_i \Pi_i & \Pi_i^2 & s_i t_i \Pi_i & c_i t_i \Pi_i \\ s_i^2 t_i & s_i c_i t_i & s_i t_i \Pi_i & s_i^2 t_i^2 & s_i c_i t_i^2 \\ s_i c_i t_i & c_i^2 t_i & c_i t_i \Pi_i & s_i c_i t_i^2 & c_i^2 t_i^2 \end{bmatrix} \tag{5.12}$$

where $\Pi_i = \Pi_{\alpha_i} s_i + \Pi_{\delta_i} c_i$.

Eq. 5.11 assumes that every scan of a source produces a detection which contributes to the astrometric solution. Even after removing gaps in the scanning law, there are periods of time and magnitudes which are less likely to result in good astrometric observations. I need to account for the efficiency of *Gaia* observations.

5.3.3 Scan Weights

As *Gaia* scans a source, up to 9 observations are taken with the 9 astrometric-field CCD columns. There are two ways in which observations may not be propagated to the astrometric solution. If a source is not detected and confirmed by the SM and first AF CCDs and allocated a window, none of the CCDs in the scan produce a successful detection. Secondly, an individual CCD observation may either not be taken or the measurement may be down-weighted in the astrometric solution. There are many reasons why this might happen such as stray background light, attitude calibration or the source simply passing through the small gaps between CCD rows. Accounting for these processes, the astrometric precision matrix may be approximated as

$$\Sigma^{-1} = \sum_{i=1}^{N} \frac{y_i}{\sigma_i^2} \mathbf{A}_i \tag{5.13}$$

where $y_i \sim \mathrm{Bernoulli}(\xi_i) \times \mathrm{Binomial}(9, \theta_i)$ where ξ_i is the fraction of scans used in the astrometric solution and θ_i is the probability of a CCD producing a successful observation. The binomial distribution assumes that a CCD observation is either successful or not therefore only allowing a full weight or zero weight. The weight formula in the AGIS pipeline (Eq. 66, Lindegren et al., 2012) does allow for non-discrete weights however

I anticipate that this have a small effect on my results. Assuming that the event of a successful scan or observation are independent events, the expected value of the weights is given by

$$\begin{aligned} w_i = \mathbb{E}\left[y_i\right] &= \mathbb{E}\left[\mathrm{Bernoulli}(\xi_i)\right] \times \mathbb{E}\left[\mathrm{Binomial}(9, \theta_i)\right] \\ &= \xi_i \times 9\theta_i. \end{aligned} \tag{5.14}$$

Therefore the expected astrometric precision is given by

$$\mathbb{E}\left[\Sigma^{-1}\right] = \sum_{i=1}^{N} \frac{w_i}{\sigma_i^2} \mathbf{A}_i. \tag{5.15}$$

For sources with $G > 13$, *Gaia* only measures a 1D position in the AL direction however for bright sources, $G < 13$ and AC measurement is also taken. Following the method in Lindegren et al. 2012, I treat the AL and AC observations as independent 1D measurements such that Eq. 5.15 expands out to

$$\mathbb{E}\left[\Sigma^{-1}\right] = \sum_{i=1}^{N_{\mathrm{AL}}} \frac{w_i^{\mathrm{AL}}}{\sigma_{\mathrm{AL},i}^2} \mathbf{A}_i + \sum_{i=1}^{N_{\mathrm{AC}}} \frac{w_i^{\mathrm{AC}}}{\sigma_{\mathrm{AC},i}^2} \mathbf{A}_i. \tag{5.16}$$

In Boubert et al. 2021b the fraction of scans, $f(t_i)$ which contribute to the *Gaia* photometry is estimated in Star Packet magnitude bins as a function of time in DR2. Due to their separate pipelines, the probability of an observation contributing to the astrometric solution differs from the photometry. To determine the astrometry weights, I renormalise the photometry scan fraction using the published number of astrometric detections used, ASTROMETRIC_N_GOOD_OBS_AL

$$\begin{aligned} w_i^{\mathrm{AL}} &= 9 \times \frac{62}{63} f(t_i) \left\langle \frac{\text{ASTROMETRIC\_N\_GOOD\_OBS\_AL}}{9 \times \frac{62}{63} \sum_{i=1}^{N_{\mathrm{scan}}} f(t_i)} \right\rangle_G \\ &= \upsilon\, f(t_i)\, f_{\mathrm{good}}(G). \end{aligned} \tag{5.17}$$

where w_i^{AL} is the weight for AL source observations since I have renormalised by the number of good AL observations used in the astrometry. The multiplication of the scan fraction by $\upsilon = 9 \times \frac{62}{63}$ converts the scan fraction to average number of observations. There are 9 columns and 7 rows of CCDs in the astrometric field of *Gaia* however one CCD is replaced by a wave front sensor hence only 62 are left. $f_{\mathrm{good}}(G)$, shown in the middle panel of Fig. 5.4, is above 90% across most magnitudes and only significantly deviates from 100% at the bright end.

For a given source, the number of AL and AC observations is published in *Gaia* DR2 as ASTROMETRIC_N_OBS_AL and ASTROMETRIC_N_OBS_AC. These statistics do not account for down-weighting of observations in the astrometry pipeline, however, assuming the AL and AC measurements of the same observations are equally likely to be down-weighted, the ratio between the numbers is unaffected. $R = \frac{\text{ASTROMETRIC\_N\_OBS\_AC}}{\text{ASTROMETRIC\_N\_OBS\_AL}}$ gives the fraction of observations which produce an AC measurement. The bottom panel of Fig. 5.4 shows that observations with $G < 13$ produce AC measurements whilst $G > 13$ do not. I use this fraction to relate the observation weights $w_i^{\mathrm{AC}} = R(G) w_i^{\mathrm{AL}}$. In truth, the scan fraction, R, may be a weak function of position on the sky at bright magnitudes due to crowding causing problems with window assignment. However, as we'll see in the following section, the contribution from AC observations to the astrometric precision is $\sim 3\%$ compared to the AL contribution and so any weak uncertainty in R has a small impact on the estimated precision.

5.3.4 Centroid Error

The centroid error in the AL and AC directions, $\sigma_i^{\mathrm{AL}}, \sigma_i^{\mathrm{AC}}$ is a function of the spacecraft instrumentation and apparent brightness of the source due to photon shot noise. For the remainder of this section, I assume that all CCDs in the astrometric field of the CCD panel have similar noise properties. I also assume that this performance is time independent and does not depend on the position of the source on the plane. Changes to the spacecraft such as mirror condensation and micrometeoroid impacts mean that the performance of the space craft is not perfectly time independent, however, as I demonstrated in Section 5.3.1 the dependence is small compared to the scatter of individual measurements using the epoch photometry.

Therefore I assume that all AL observations of a single source have the same precision and likewise for all AC observations such that Eq. 5.16 becomes

$$\mathbb{E}\left[\Sigma^{-1}\right] = \frac{1}{\sigma_{\mathrm{AL}}(G)^2} \sum_{i=1}^{N_{\mathrm{AL}}} w_i^{\mathrm{AL}} \mathbf{A}_i + \frac{1}{\sigma_{\mathrm{AC}}(G)^2} \sum_{i=1}^{N_{\mathrm{AC}}} R(G) w_i^{\mathrm{AL}} \mathbf{A}_i \tag{5.18}$$

Table 1 of Lindegren et al. 2012 gives the ratio of AC to AL error for bright sources as typically $\psi = 520/92$ such that $\sigma_{\mathrm{AC}} = \psi \sigma_{\mathrm{AL}}$, although I note that this was only a pre-launch estimate and the true calibrated uncertainty is likely marginally different. Substituting into the expected precision

$$\mathbb{E}\left[\Sigma^{-1}\right] = \frac{1}{\sigma_{\mathrm{AL}}(G)^2} \left(1 + \frac{R(G)}{\psi^2}\right) \sum_{i=1}^{N_{\mathrm{AL}}} w_i^{\mathrm{AL}} \mathbf{A}_i. \tag{5.19}$$

The final unknown in Eq. 5.22 is σ_{AL}, the astrometric centroid error of AL observations. σ_{AL} was estimated from the *Gaia* published astrometry by Belokurov et al. 2020b using the formula $0.53\sqrt{N}\sigma_{\varpi}$ where N was the number of AL observations used for the source astrometry published as ASTROMETRIC_N_GOOD_OBS_AL. 0.53 was used as this empirically matched the published distribution in Fig.9 of L18. However $\sqrt{N}\sigma_{\varpi}$ is a strong function of position on the sky depending on scan directions and spread of observations throughout the year. This means that the running median as a function of magnitude is heavily affected by where the given stars lie on the sky.

For this work, I find a more mathematically motivated route to the scan variance. By summing up the first two diagonal terms of the inverse covariance matrix from Eq. 5.12, the dependence on the scan angle ϕ_i disappears.

$$\begin{aligned}\mathbb{E}\left[\Sigma^{-1}\right]_{\alpha,\alpha} + \mathbb{E}\left[\Sigma^{-1}\right]_{\delta,\delta} =& \frac{1}{\sigma_{AL}(G)^2}\left(1+\frac{R(G)}{\psi^2}\right)\sum_{i=1}^{N} w_i^{\mathrm{AL}}(s_i^2+c_i^2) \\ =& \frac{1}{\sigma_{AL}(G)^2}\left(1+\frac{R(G)}{\psi^2}\right)\sum_{i=1}^{N} w_i^{\mathrm{AL}} \end{aligned} \tag{5.20}$$

Therefore the AL astrometric error can be determined independent of position on the sky by substituting Σ for the published covariance C and rearranging in terms of σ_{AL}.

$$\sigma_{AL}^2(G) = \left(1+\frac{R(G)}{\psi^2}\right)\left\langle\frac{\textsc{astrometric\_n\_good\_obs\_al}}{(\mathbf{C}^{-1})_{\alpha\alpha}+(\mathbf{C}^{-1})_{\delta\delta}}\right\rangle_G \tag{5.21}$$

where $\sum_{i=1}^{N} w_i^{\mathrm{AL}} = \textsc{astrometric\_n\_good\_obs\_al}$. As I discuss in more detail in Section 5.6, the selection of the *Gaia* 5D astrometry sample included a cut on ASTROMETRIC_SIGMA5D_MAX. Sources with large astrometric uncertainty would not receive 5D astrometry and therefore, particularly at the dim end, σ_{AL} would be biased low. To mitigate this, I calculate $\sigma_{\mathrm{AL}}(G)$ using all stars in *Gaia* DR2 with at least 6 VISIBILITY_PERIODS_USED. For sources without 5D astrometry I use the inverse of the published 2D astrometry covariance matrix as a proxy. This is a rough approximation and therefore I suggest that my results are only trusted out to $G \lesssim 20.5$ at which point the cut on ASTROMETRIC_SIGMA5D_MAX becomes significant (see Section 5.6).

The distribution of σ_{AL} is shown in the top panel of Fig. 5.4 demonstrating a relatively flat behaviour for $G < 13$ where 2D observations are taken and time windows are truncated to avoid saturation. For $G > 18$ the variance grows with magnitude due to photon shot noise. The red line gives the median value in 0.1mag and I linearly interpolate this as a function of magnitude to estimate $\sigma_{\mathrm{AL}}(G)$. For reference, the grey-scale histograms are the σ_{AL} for 5D astrometry sources where the truncation for $\sigma_{\mathrm{AL}} \sim 10$ mas is caused by the ASTROMETRIC_SIGMA5D_MAX cut. The blue line in the top panel of Fig. 5.4 is the blue line from Fig. 9 of L18. Across most of the magnitude range, my estimate is lower than L18 by ~ 10%. This is expected because we're actually calculating slightly different statistics. L18 used the residuals of all AL observations relative to the best fit astrometric solution. In calculating the source astrometry, observations are assigned weights as a function of their residuals which disfavoured observations with large residuals from being used in the astrometric solution. Therefore the value of σ_{AL} inferred by L18 is higher than mine which has implicitly ignored large outliers. As my task in this chapter is to predict the published 5D astrometry uncertainties, my formula for σ_{AL} is the appropriate one to use.

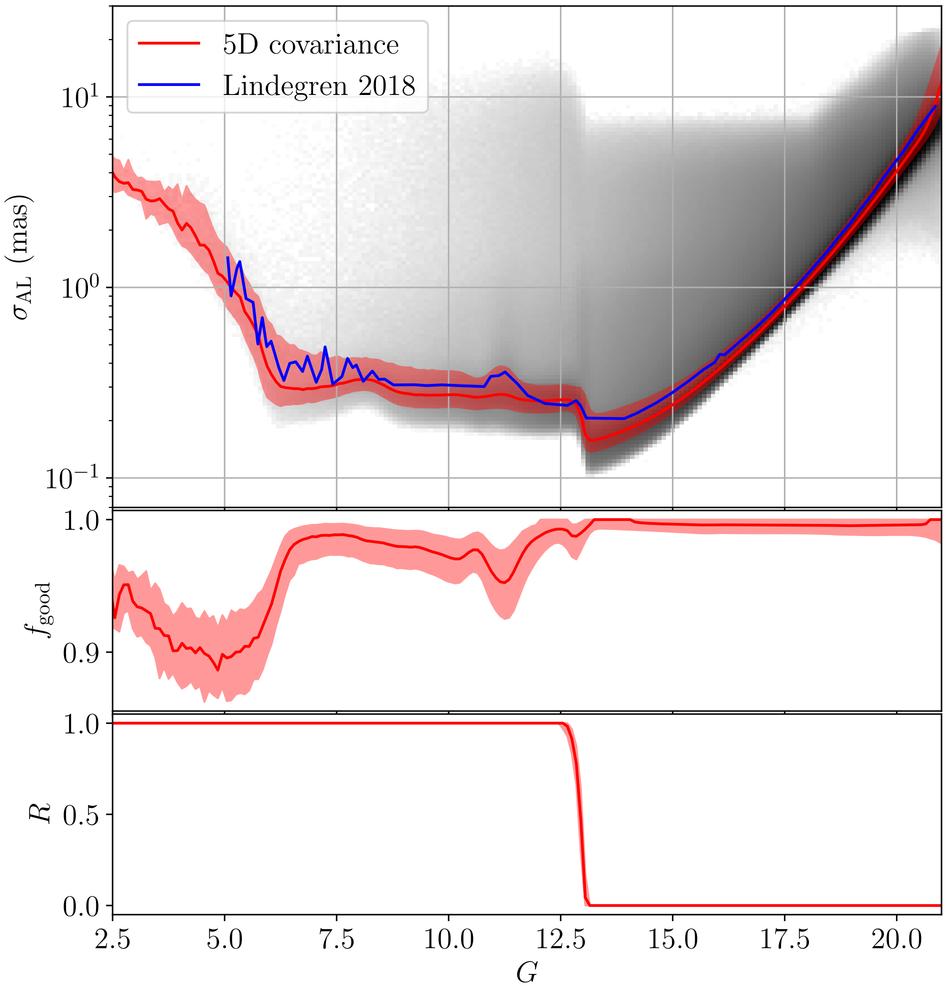


Fig. 5.4 The magnitude dependence of the ASF is a function of the AL astrometric uncertainty σ_{AL} (top), the fraction of photometric observations which generate good astrometric observation used in the AGIS pipeline f_{good} (middle) and the ratio of AC to AL observations $\boldsymbol{R}$ (bottom). In all cases the median and 16th to 84th percentiles of the *Gaia* DR2 astrometry sample are given by the red sold line and shaded area respectively. The distribution of σ_{AL} (black histograms, log normalised) extends high above the median due to source excess noise.

Finally, I can substitute in w_i^{AL} from Section 5.3.3.

$$\begin{aligned}\mathbb{E}\left[\Sigma^{-1}\right] =& \frac{1}{\sigma_{\mathrm{AL}}(G)^2}\left(1+\frac{R(G)}{\psi^2}\right) f_{\mathrm{good}}(G) \sum_{i=1}^{N_{\mathrm{AL}}} \upsilon f(t_i) \mathbf{A}_i \\ =& \rho(G)\, \Phi(l,b) \end{aligned} \tag{5.22}$$

where I have defined

$$\rho(G) \equiv \frac{1}{\sigma_{\mathrm{AL}}(G)^2}\left(1+\frac{R(G)}{\psi^2}\right) f_{\mathrm{good}}(G) \tag{5.23}$$

as the magnitude dependent normalisation and $\boldsymbol{\Phi}(l,b) \equiv \sum_{i=1}^{N_{\mathrm{AL}}} \upsilon f(t_i)\mathbf{A}_i$ as the scanning law dependent matrix. $\boldsymbol{\Phi}$ has a weak magnitude dependence as the fractions $f(t_i)$ change between the magnitude bins in which *Gaia* downloads data however, within any download bin, it is independent of magnitude.

5.3.5 Astrometry Spread Function

In the previous sections I have derived the expected precision, $\mathbb{E}\left[\Sigma^{-1}\right]$ for simple point sources as observed by *Gaia*. The DR2 data has been used to estimate $\sigma_{\mathrm{AL}}(G), R(G)$ and $f_{\mathrm{good}}(G)$ as running medians as a function of magnitude. $\boldsymbol{\Phi}(l,b) = \sum_{i=1}^{N_{\mathrm{AL}}} f(t_i)\mathbf{A}_i$ is a function of the scanning law only and has no dependence on the *Gaia* astrometry data. For the remainder of this chapter, I simplify the notation taking $\Sigma = \mathbb{E}\left[\Sigma^{-1}\right]^{-1}$ as the expected 5D astrometry covariance for a simple point source in *Gaia*.

For a point source moving without acceleration with true astrometric coordinates $\mathbf{r}$ observed in *Gaia* DR2, the expected measured astrometric coordinates are drawn from a multivariate normal distribution with covariance Σ,

$$\mathbf{r}' \sim \mathcal{N}(\mathbf{r}, \Sigma(G,l,b)). \tag{5.24}$$

This normal distribution is the Astrometry Spread Function where G, l, b are the apparent magnitude and position of the source on the sky.

To demonstrate how the astrometry is fit in practice, I show the expected observations and astrometric uncertainty for a hypothetical source at $l = 30$ degrees, $b = 10$ degrees with apparent magnitude $G = 16$ in Fig. 5.5. The source is given proper motion $\mu_\alpha^* = 20\,\mathrm{mas/y}$, $\mu_\delta = 20\,\mathrm{mas/y}$ which produces a trajectory from South East to North West. Adding the parallax ellipse for $\varpi = 12$ mas generates a spiralling apparent position observed by *Gaia* throughout DR2 given by the black-dashed line in the top panel of Fig. 5.5.

Gaia scans this region of the sky 15 times in DR2 given by the blue and red arrows for scans from FoV1 and FoV2 respectively. Each scan improves the constraint on each of the five astrometry parameters the uncertainties for which are given in the bottom panel of Fig. 5.5. *Gaia* selects sources for the 5D astrometry catalogue which have at least 6 VISIBILITY_PERIODS_USED where a visibility period is a group of observations separated by less than four days. Where fewer than 6 visibility periods have been observed the AGIS pipeline places priors on the astrometry derived in Michalik et al. 2015b and only

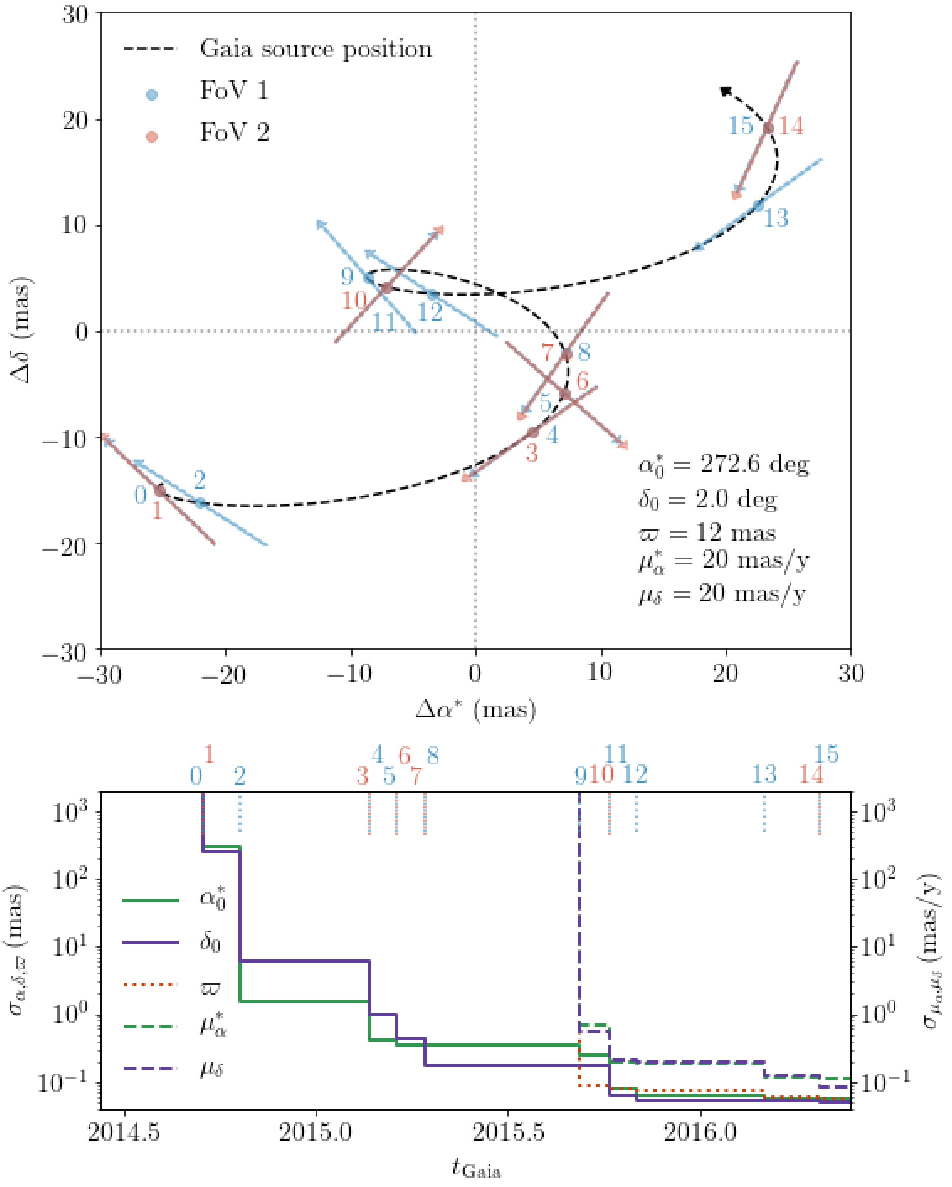


Fig. 5.5 **Top**: Observed from L2, any source on the sky follows a curved track given by the combination of source proper motion and apparent parallax ellipse as represented here by the black-dashed line for a source with $\mu_\alpha^* = 20\,\mathrm{mas/y}$, $\mu_\delta = 20\,\mathrm{mas/y}$ and $\varpi = 10\,\mathrm{mas}$ located near the galactic bulge with $l = 30$, $b = 10$. *Gaia* scans this position on the sky 15 times in the DR2 time frame shown by blue and red arrows for FoV1 and FoV2 respectively. **Bottom**: Each scan, marked by the vertical red and blue dashed lines, contributes to the 5D astrometry constraints. The expected uncertainty on each astrometry parameter is shown for a source with $G = 16$ and therefore $\sigma_{\mathrm{AL}} = 0.37$ mas and reduce with each subsequent scan. When fewer than 6 visibility periods are observed, only α_0^* (green solid) and δ_0 (purple solid) are shown with priors placed on all parameters. With at least 6 visibility periods, uncertainties are also given for μ_α^* (green dashed), μ_δ (purple dashed) and ϖ (red dotted).

the 2D position constraints are published. I replicate this using the same priors and only providing uncertainties for the α_0^* (green solid) and δ_0 (purple solid) parameters before the sixth visibility period (9th scan).

After the sixth visibility period, the priors were dropped and the uncertainties on μ_α^* (green dashed), μ_δ (purple dashed) and ϖ (red dotted) parameters are also shown. For simplicity, this demonstration assumes all observations were successful and equally likely to contribute to the astrometry however as discussed in Section 5.3.3, this is not always the case and this is corrected for by weighting observations.

5.4 Results

To test that my method is producing reasonable covariance matrices, I compare my predictions with the published 5D astrometry covariances. From the *Gaia* DR2 astrometry sample I determine the median published covariance on a level 7 HEALPix grid (Górski et al., 2005) in the magnitude range $G \in [18.1, 19.0]$ which represents a single Star Packet bin in which the scan fractions, f are unchanged.

I estimate the predicted covariance using the formula in Eq. 5.22 where G is taken as the median apparent magnitude of stars in the given magnitude bin and HEALPix pixel. The scan angles and times are inferred at the central coordinates of the HEALPix pixel.

The diagonal elements of both the median observed and predicted covariance matrices are shown in Fig. 5.6 demonstrating excellent agreement down to degree scales in all components. In all coordinates the variance is significantly enhanced in regions which have been scanned less in DR2, most notably around the Galactic bulge. Thin streaks of boosted variance on the sky correspond to time periods in *Gaia* DR2 where data was not taken due to mirror decontamination or other disruptive processes.

In Fig. 5.7 I compare the correlation coefficients by dividing the off-diagonal covariance elements by the square root of the products of their respective variances. Correlation coefficients are less dependent on the number of observations, which has largely been divided out, and more on the scan directions and time variance leading to a more complex and varied structure on the sky. The observed correlation (upper right triangle) and predicted correlation (lower left) show excellent agreement down to small scale variations.

In Fig. 5.8, diagonal elements give the ratio of predicted to observed variance. Across the vast majority of the sky, there is strong agreement with noise dominating in underscanned regions. Two features stand out. A streak of scans in the South East and North West show underestimated uncertainties from the model. The scans in *Gaia* responsible for this are constrained and discussed in Section 5.6. The Galactic bulge also shows a significant systematic underestimate against the observed variance. This is not unexpected as high source crowding can cause single windows to be allocated to multiple sources generating spurious centroid positions. The third panel in L18 shows the same issue but manifested in the ASTROMETRIC_EXCESS_NOISE of the source fits.

I demonstrate this issue in Fig. 5.9 where I show the median $\sigma_{\mathrm{AL}}(G, l, b)$ evaluated using Eq. 5.21 with the median taken in every 0.1 mag magnitude bin and HEALPix level

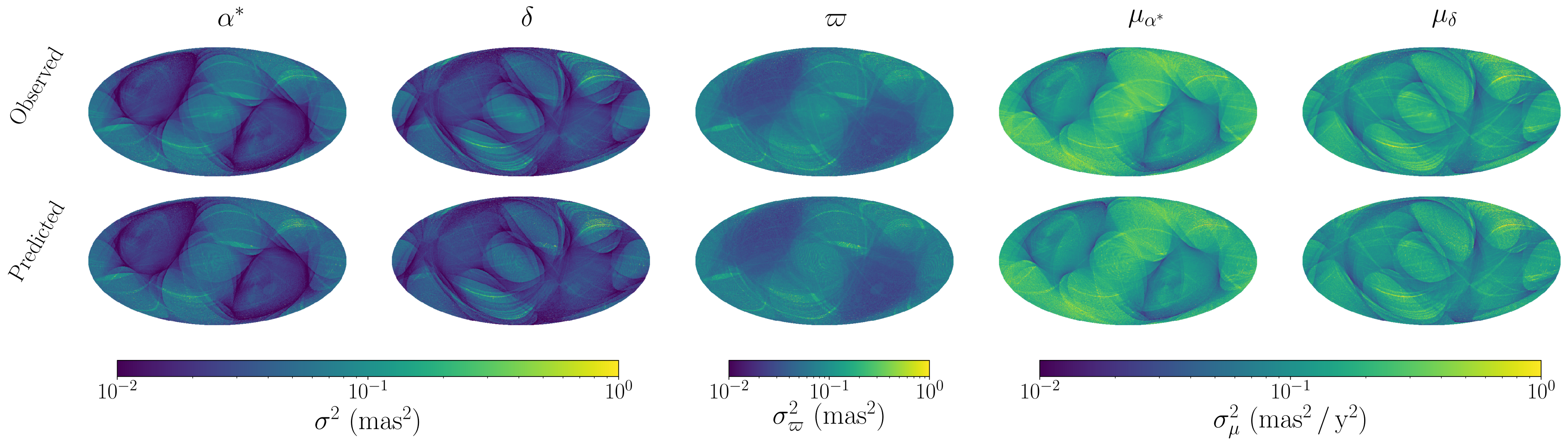


Fig. 5.6 The predicted variance in each component of the astrometry for the $\boldsymbol{G} \in [18.1, 19.0]$ magnitude bin (bottom row) matches the median variances for observed *Gaia* astrometry sources in the same magnitude range (top row) in HEALPix level 7 bins shown in Galactic coordinates. In all plots, the lighter shades near the ecliptic plane show increased variance due to lack of scans in *Gaia* DR2. The top row is similar to L18, Fig. B3 where instead of showing individual components, they instead show the position and proper motion semi-major axes.

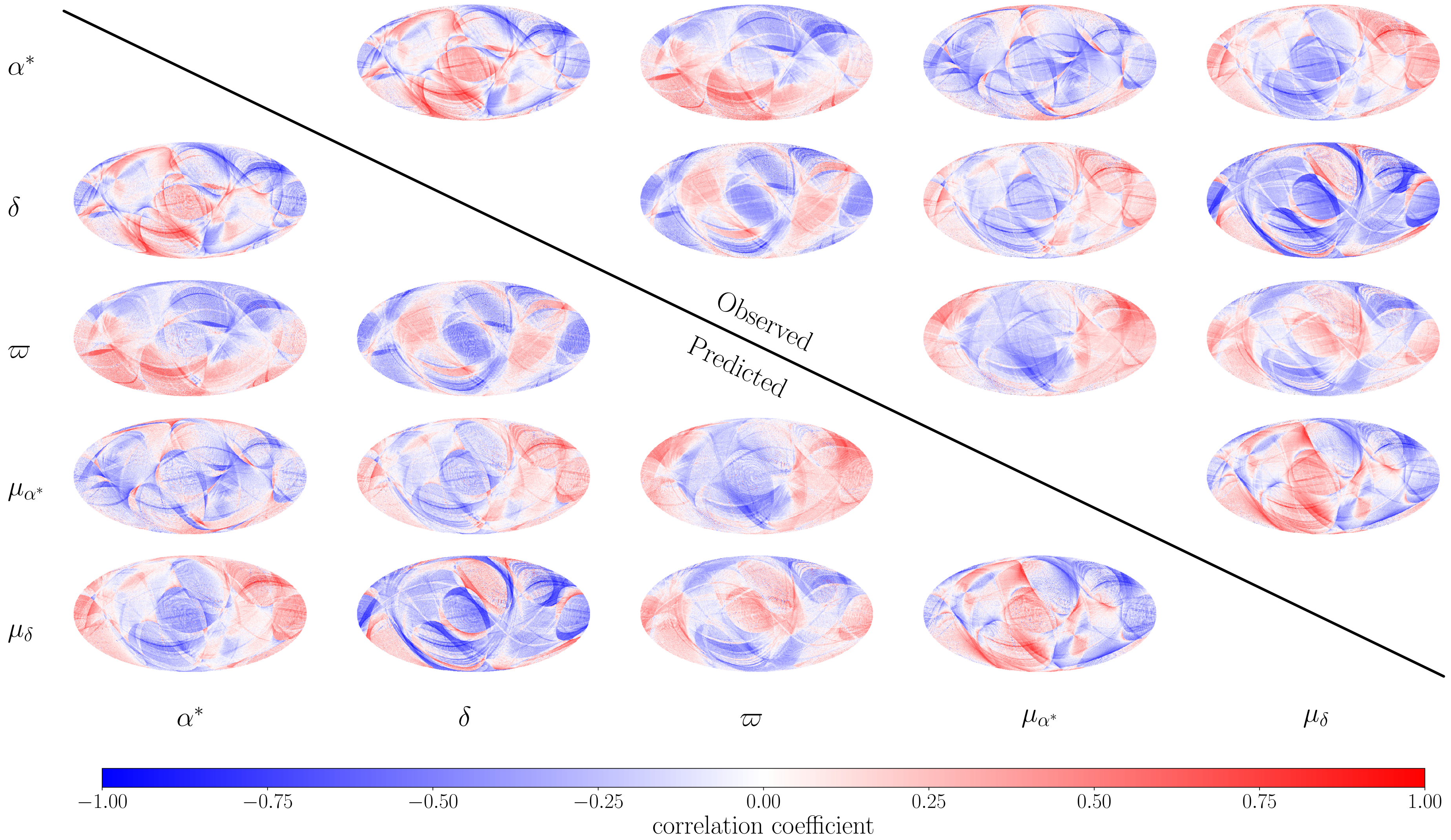


Fig. 5.7 The predicted correlation coefficients for $G \in [18.1, 19.0]$ (lower triangle) match up well to the correlation coefficients of the median covariances from the *Gaia* astrometry (upper triangle) in HEALPix level 7 bins shown in Galactic coordinates. The detailed and complex structure in the correlation coefficients is driven by the directions and time separations between subsequent scans of the same position on the sky. The $\varpi - \mu_\alpha^*$, $\varpi - \mu_\delta$ and $\mu_\alpha^* - \mu_\delta$ in the upper triangle correspond to L18, Fig. B5 although at a different magnitude.

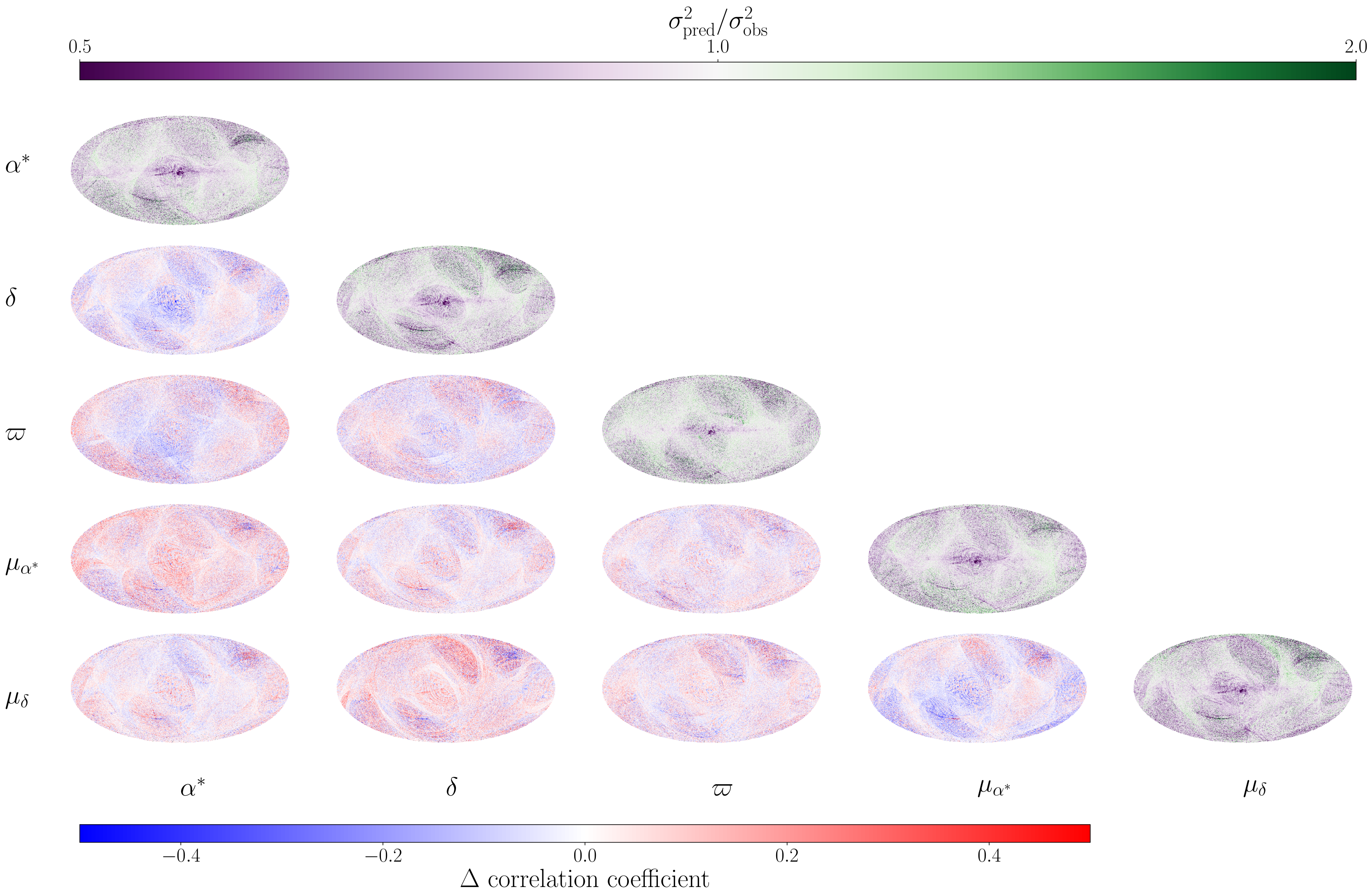


Fig. 5.8 The ratio of predicted to median observed variance for $\boldsymbol{G} \in [18.1, 19.0]$ (diagonal) in HEALPix level 7 bins shows structure in the bulge where the prediction has underestimated the observed variance driven by crowding of sources and lack of observations in *Gaia* DR2. The difference between predicted and median observed correlation coefficients (lower triangle) shows no strong structural bias in correlation.

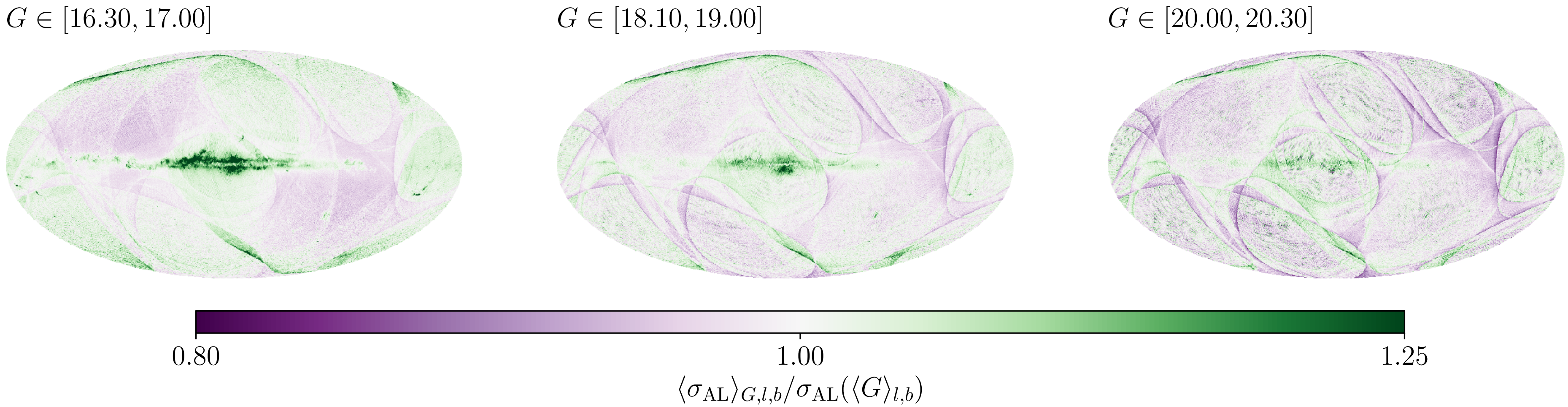

$G \in [16.30, 17.00]$
$G \in [18.10, 19.00]$
$G \in [20.00, 20.30]$
0.80
1.00
1.25
$\langle\sigma_{\mathrm{AL}}\rangle_{G,l,b}/\sigma_{\mathrm{AL}}(\langle G\rangle_{l,b})$

Fig. 5.9 The ratio of the median σ_{AL} in HEALPix level 7 bins to the value of σ_{AL} evaluated at the median magnitude of stars in the HEALPix bins highlights any dependence of σ_{AL} on sky position in Galactic coordinates. Particularly in the highest source density regions of the Galactic plane and bulge at brighter magnitudes, sources have significantly higher astrometric measurement uncertainty than the average across the sky.

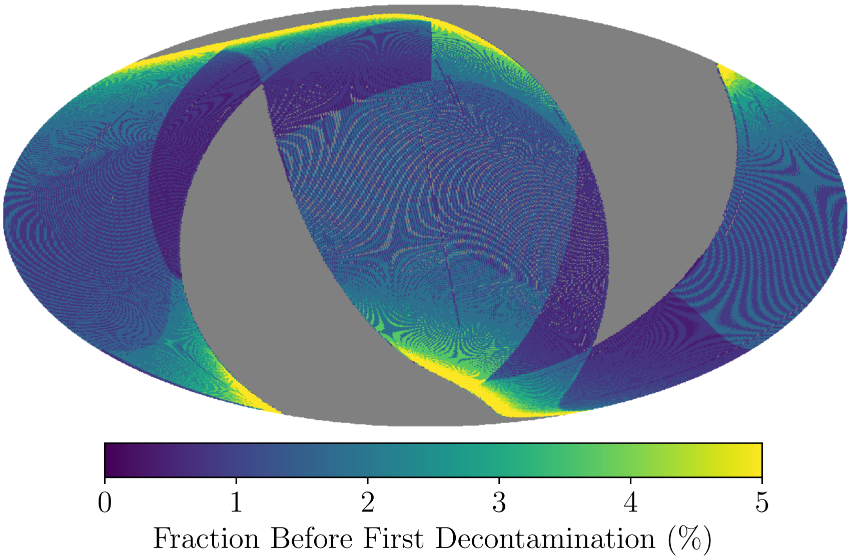


Fig. 5.10 The fraction of scans of any position on the sky which occurred before the first decontamination event, shown here as a percentage in HEALPix 7 bins in Galactic coordinates, corresponds to regions of the sky with enhanced measurement uncertainty.

7 pixel divided by $\sigma_{\mathrm{AL}}(G)$ evaluated at the median magnitude of stars in the HEALPix pixel. From Section 5.3.4, I expect σ_{AL} to be independent of position on the sky which is a key assumption in my model. Across the sky σ_{AL} shows only weak dependence on the scanning law at less than 10%. However, particularly for brighter magnitude bins, σ_{AL} is significantly higher in regions of the disc and bulge with the highest source density. This issue is further exacerbated for the bulge as it happens to reside in a region of the sky which has been scanned very few times in *Gaia* DR2 whereas the LMC and SMC, which have been scanned more heavily, show no clear signal. In future *Gaia* data releases, the Galactic bulge will receive significantly more scans reducing this issue.

Fig. 5.9 also shows residual scanning law structure which is likely caused by the 20% variation in the instrument precision discussed in Section 5.3.1. For example, the green strips in the North East and South where σ_{AL} is systematically higher correspond to areas which received many observations before the first decontamination when the satellite measurement precision was at its worst as shown in Fig. 5.3. Fig. 5.10 shows the percentage of observations which took place before the first decontamination event in DR2 for which the highest regions match exactly with regions of the sky in Fig. 5.9 with enhanced σ_{AL}. The diagonal elements of Fig. 5.8 show that these features are comparable to the background noise level and so are not of significant concern.

The off-diagonal elements of Fig. 5.8 show the difference between predicted and observed correlation coefficients. The structure of the scanning law can be seen in white as the regions which are most heavily scanned have the lowest uncertainty. There is some marginal bias in the α^* and δ components but this is small compared with the overall signal seen in Fig. 5.7.

From these results, I demonstrate that the ASF is accurate across the majority of the sky across all magnitudes at the 10% level. However for bright sources ($G \lesssim 18$) close in crowded regions ($|b| \lesssim 5$ degrees) un-corrected calibration effects become significant inflating the systematic uncertainties. When using *Gaia* DR2 astrometry to search for excess noise from genuine source characteristics, these systematic uncertainties should be taken into account.

5.5 Unit Weight Error

Unit Weight Error (UWE) is the reduced chi-squared statistic of the astrometric fit to observations.

$$\mathrm{UWE} = \sqrt{\frac{1}{\nu}(\mathbf{x}' - \mathbf{x})^{\mathrm{T}}\mathbf{K}^{-1}(\mathbf{x}' - \mathbf{x})} \tag{5.25}$$

where $\mathbf{x}'$ and $\mathbf{x}$ are the measured and expected position measurements of a source, $\mathbf{K} = \mathrm{diag}[\sigma_1^2, \sigma_2^2 ... \sigma_N^2]$ is the measurement covariance and $\nu = N - 5$ is the number of degrees of freedom.

For simple point sources UWE is drawn from a Gamma distribution, UWE $\sim \Gamma\left[\nu/2, \nu/2\right]$ such that the expected value is 1 and the variance is inversely proportional to the degrees of freedom. However any excess stellar motion or an extended flux distribution introduces an excess UWE above 1 as happens for binary systems (Penoyre et al., 2020) or astrometric microlensing events (McGill et al., 2020). *Gaia* publishes χ^2 and the degrees of freedom $\nu = N - 5$ for all stars with 5D astrometry in DR2 from which UWE can be calculated. However, the published χ^2 is plagued by the DoF bug which makes values unreliable to use for estimating the excess noise (L18, provides a full description of the DoF bug which I have summarised in Appendix B.1).

This can be remedied by renormalising the published UWE as a function of colour and apparent magnitude to produce a new statistic, RUWE[4]. RUWE is normalised such that the 41st percentile is 1 as this was found to represent well behaved sources where the median showed significant contamination from sources with excess error. This works well at face value and produces a usable statistic however there are two limitations. Firstly RUWE does not follow a well defined χ^2 distribution as would be expected from UWE, therefore estimating the significance of excess noise is challenging. Secondly, in cases where excess noise is not equally likely in all colours and apparent magnitudes, the renormalisation can hide some of the expected excess. This would be problematic when establishing the binary fraction as a function of colour and absolute magnitude which is expected to vary considerably between stellar populations (Belokurov et al., 2020b; Price-Whelan et al., 2020).

An alternative of UWE for a source with measured 5D astrometry is given by

$$\mathrm{UWE} = \sqrt{\frac{1}{n}\mathbb{E}[\boldsymbol{\delta}^{\mathrm{T}}\Sigma^{-1}\boldsymbol{\delta}]} \tag{5.26}$$

[4] http://www.rssd.esa.int/doc_fetch.php?id=3757412

where $n = 5$ is the dimensionality of the astrometry. Given $\boldsymbol{\delta} = (\mathbf{r}' - \mathbf{r}) \sim \mathcal{N}(0, \mathbf{C})$

$$\begin{aligned}\mathbb{E}[\boldsymbol{\delta}^{\mathrm{T}}\Sigma^{-1}\boldsymbol{\delta}] &= \int \frac{1}{(2\pi)^{n/2}\sqrt{||\mathbf{C}||}}(\boldsymbol{\delta}^{\mathrm{T}}\Sigma^{-1}\boldsymbol{\delta}) \exp\left(-\frac{1}{2}(\boldsymbol{\delta}^{\mathrm{T}}\mathbf{C}^{-1}\boldsymbol{\delta})\right) \mathrm{d}\boldsymbol{\delta} \\ &= \int \frac{1}{(2\pi)^{n/2}}(\mathbf{y}^{\mathrm{T}}\sqrt{\mathbf{C}}^{\mathrm{T}}\Sigma^{-1}\sqrt{\mathbf{C}}\mathbf{y}) \exp\left(-\frac{1}{2}(\mathbf{y}^{\mathrm{T}}\mathbf{y})\right) \mathrm{d}\mathbf{y} \qquad (5.27)\end{aligned}$$

where $\mathbf{y} = \sqrt{\mathbf{C}^{-1}}\boldsymbol{\delta}$ and $\frac{\mathrm{d}\mathbf{y}}{\mathrm{d}\boldsymbol{\delta}} = \sqrt{||\mathbf{C}||}$. Letting $\mathbf{W} = \sqrt{\mathbf{C}}^{\mathrm{T}}\Sigma^{-1}\sqrt{\mathbf{C}}$

$$\mathbb{E}[\boldsymbol{\delta}^{\mathrm{T}}\Sigma^{-1}\boldsymbol{\delta}] = \int \frac{1}{(2\pi)^{n/2}}(\mathbf{y}^{\mathrm{T}}\mathbf{W}\mathbf{y}) \exp\left(-\frac{1}{2}(\mathbf{y}^{\mathrm{T}}\mathbf{y})\right) \mathrm{d}\mathbf{y}. \qquad (5.28)$$

All off-diagonal elements of $\mathbf{W}$ produce antisymmetric integrands in $\mathbf{y}$ leaving only the diagonal elements

$$\begin{aligned}\mathbb{E}[\boldsymbol{\delta}^{\mathrm{T}}\Sigma^{-1}\boldsymbol{\delta}] &= \sum_{i=1}^{n} \int \frac{1}{(2\pi)^{n/2}}\mathbf{W}_{i,i}\mathbf{y}_i^2 \exp\left(-\frac{1}{2}\mathbf{y}_i^2\right) \mathrm{d}\mathbf{y} \\ &= \sum_{i=1}^{n} \mathbf{W}_{i,i}. \qquad (5.29)\end{aligned}$$

Substituting Σ back into this I have UWE in terms of the published covariance $\mathbf{C}$

$$\begin{aligned}\mathrm{UWE} &= \sqrt{\frac{1}{5}\mathrm{Tr}(\sqrt{\mathbf{C}}\Sigma^{-1}\sqrt{\mathbf{C}})} \\ &= \sqrt{\frac{1}{5}\mathrm{Tr}(\mathbf{C}\Sigma^{-1})}. \qquad (5.30)\end{aligned}$$

Using this formula, I estimate UWE for all stars with 5D astrometry in *Gaia* DR2. The distribution of UWE as a function of magnitude, shown in the Fig. 5.11, is uniform with the median $\langle\mathrm{UWE}\rangle \gtrsim 1$. The fact that the median UWE sits slightly higher than 1 is due to the contribution from sources with excess noise. The spread of UWE which is greatest at $G \sim 13$ and narrows to fainter magnitude is a clear signature of excess error which is resolvable at brighter magnitude but becomes increasingly dominated by photon count noise for fainter sources.

Our estimate is compared with the published UWE for sources with $G \in [18.1, 19.0]$ in Fig. 5.12. At these dim magnitudes, the impact of the DoF bug is small. Across the sky, my estimate of UWE is in excellent agreement with the published value producing no systematic residual signal in the right hand panel down to 10% uncertainty.

In *Gaia* EDR3, the DoF bug is fixed and my estimate of UWE is superseded by the published value. However, the fact that my measurement is in good agreement with the published UWE is indicative that the published covariance alongside my prediction of the ASF contains all of the information contained in UWE and more. Whilst UWE can be used to determine the probability and amplitude of any excess variance, the ASF has the potential to decode the orientation and time variation of excess noise.

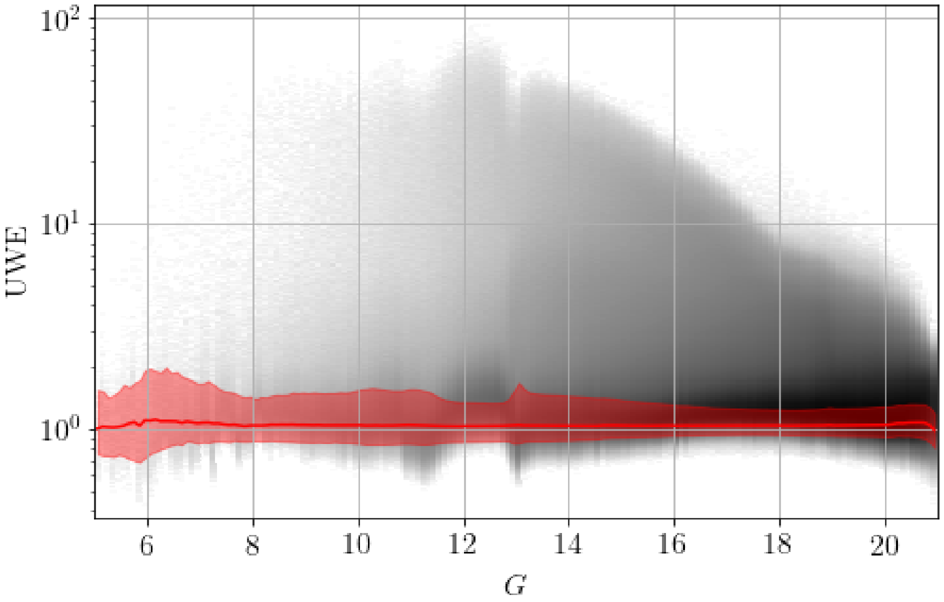


Fig. 5.11 The reduced χ^2 of the astrometric solution, UWE, is estimated from the published covariances using the predicted covariance for simple point sources producing a distribution with median ~ 1 (red solid). The distribution of source (black histogram, log normalised) extends out to high values of UWE due to sources with high excess noise.

5.6 Astrometric Selection

In order to construct unbiased dynamical models of the Milky Way, it is critically important that we have a strong understanding of the completeness of the sample. In Chapter 4 I introduced the selection functions for the *Gaia* source catalogue and science subsets. In this section I investigate one aspect of the astrometry selection function in more detail to understand the impact of source variance and demonstrate another application of the ASF. As a brief reminder, the *Gaia* DR2 5D astrometry sample is the subset of the full sample that satisfies the cuts (L18, Section 4.3):

- $G < 21$
- VISIBILITY_PERIODS_USED > 5
- ASTROMETRIC_SIGMA5D_MAX $> 1.2 \times \gamma(G)$

where $\gamma(G) = \max\left[1, 10^{0.2(G-18)}\right]$.

To construct the selection function for the 5D astrometry sample I can combine the effect of these cuts with the full sample selection function

$$\mathrm{P}(\mathcal{S}_{\mathrm{5Dast}}) = \mathrm{P}(\mathcal{S}_{\mathrm{5Dast}}|\mathcal{S}_{\mathrm{DR2}})\mathrm{P}(\mathcal{S}_{\mathrm{DR2}}) \tag{5.31}$$

where $\mathcal{S}_{\mathrm{5Dast}}$ is the event that a source is published with 5D astrometry and $\mathcal{S}_{\mathrm{DR2}}$ is the event that a source is included in DR2 with or without 5D astrometry. $\mathrm{P}(\mathcal{S}_{\mathrm{DR2}})$ is the full *Gaia* DR2 selection function estimated in Boubert & Everall 2020. The probability of a star in DR2 receiving 5D astrometry, $\mathrm{P}(\mathcal{S}_{\mathrm{5Dast}}|\mathcal{S}_{\mathrm{DR2}})$, is governed by the three cuts outlined above.

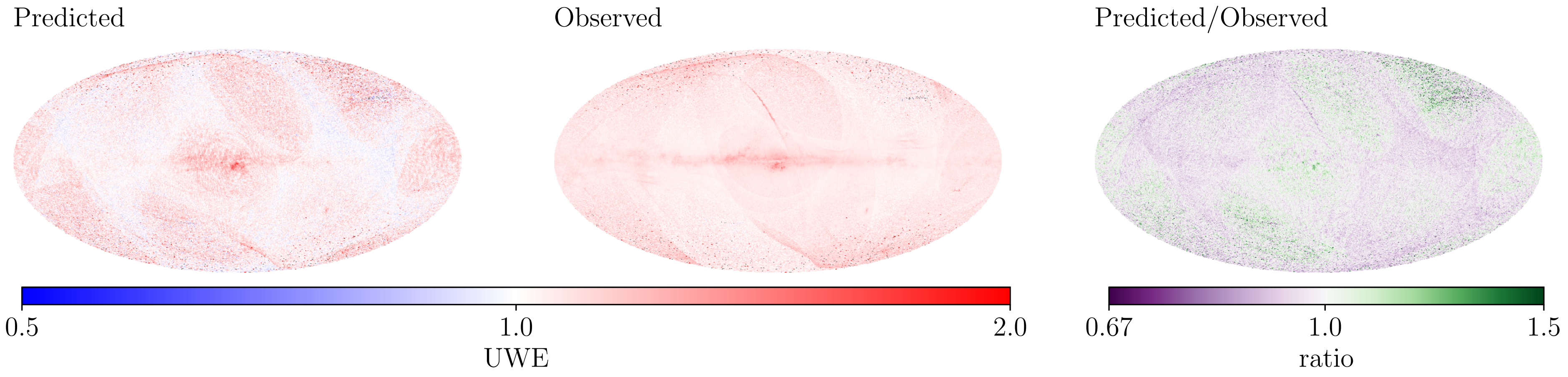


Fig. 5.12 Predicted UWE for sources with $G \in [18.1, 19.0]$ in Galactic coordinates (left) show little structure on the sky as expected. The median observed UWE in HEALPix level 7 bins (middle) has some limited structure relating to problematic individual scans in *Gaia* DR2. The ratio between predicted and observed UWE shows strong agreement down to the 10% level at which point the DoF bug introduces a bias to the published values for $G \sim 18$.

The second cut on VISIBILITY_PERIODS_USED (k_VP) is a complex function of the scanning law.Here I focus on the ASTROMETRIC_SIGMA5D_MAX (σ_5Dmax) cut.

σ^2_5Dmax is the maximum eigenvalue of the scaled astrometric covariance matrix

$$\sigma^2_\mathrm{5Dmax} = \lambda_\mathrm{max}\left[S\,C\,S\right] \tag{5.32}$$

where $C \in \mathbb{R}^{5\times5}$ is the published 5D covariance matrix and $S = \mathrm{diag}[1, 1, \sin(\xi), T/2, T/2]$ where $\xi = 45$ deg is the solar aspect angle of the *Gaia* satellite and $T = 1.75115$ yr is the time window of observations used in *Gaia* DR2 (see Section 4.3 L18).

My aim is to estimate the contribution to the selection function solely from the cut on σ_5Dmax,

$$\mathrm{P}(\sigma_\mathrm{5Dmax} < 1.2\gamma \,|\, k_\mathrm{VP} > 5, G, l, b). \tag{5.33}$$

σ_5Dmax and k_VP are published for all sources in *Gaia* DR2 so this could be approximated by taking the ratio of number of sources with $\sigma_\mathrm{5Dmax} < 1.2\gamma(G)$ and $k_\mathrm{VP} > 5$ to only those with $k_\mathrm{VP} > 5$ as a function of apparent magnitude and position on the sky

$$\mathrm{P}(\sigma_\mathrm{5Dmax} < 1.2\gamma \,|\, k_\mathrm{VP} > 5, G, l, b) = \frac{N(\sigma_\mathrm{5Dmax} < 1.2\gamma,\ k_\mathrm{VP} > 5, G, l, b)}{N(k_\mathrm{VP} > 5, G, l, b)}.$$

This approach is limited by Poisson count noise. To resolve scanning law variations, one would need to resolve the sky to at least HEALPix level 7. Using 200 magnitude bins, this results in an average of ~ 30 stars with astrometry per bin which is dominated by the Milky Way disc. At high latitudes the inference is entirely dominated by Poisson noise.

I could apply the method introduced in Chapter 4 which provides very impressive results for the astrometry and RVS selection function. But given the ASF, can I do better?

Instead, I can use *Gaia*'s predicted covariance as a function of position on the sky given in Section 5.3. This enables me to reach unlimited resolution on the sky without HEALPix binning the data. I can predict σ_5Dmax for any source in *Gaia* as a function of magnitude and position on the sky

$$\sigma_\mathrm{5Dmax} = \sqrt{\lambda_\mathrm{max}\left[S\Sigma S\right]} = \frac{1}{\sqrt{\rho(G)}}\sqrt{\lambda_\mathrm{max}\left[S\Phi^{-1}S\right]}. \tag{5.34}$$

where I have used the substitution $\Sigma^{-1} = \rho(G)\Phi$ from Eq. 5.22 and $\rho(G)$ is defined in Eq. 5.23. A comparison of the running median of the predicted σ_5Dmax (red dashed) and observed ASTROMETRIC_SIGMA5D_MAX (blue solid) in Fig. 5.13 shows that the prediction overestimates for $G < 13$ and underestimates for $13 < G < 16$. The cause of this is the 'DoF' bug detailed in Appendix A of L18. My predicted σ_5Dmax has been corrected for the DoF bug whilst the published values, on which the astrometry was selected, had not been corrected. The DoF bug is de-corrected from my prediction dividing through by a factor F from Eq. B.1 to produce the red solid line, in good agreement with the published σ_5Dmax as a function of magnitude. The predicted value marginally systematically underestimates σ_5Dmax across all magnitudes by ~ 10% which I conjecture

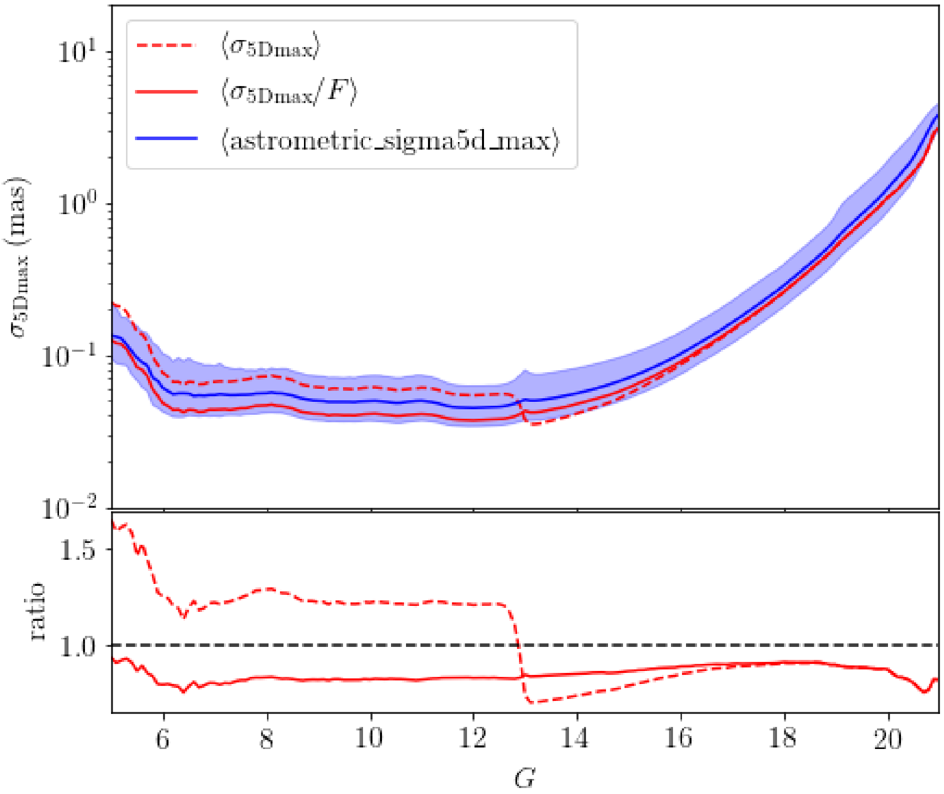


Fig. 5.13 Predicted $\sigma_{5\mathrm{Dmax}}$ after correcting for the DoF bug (red solid) as a function of magnitude for all sources in the *Gaia* DR2 astrometry shows strong agreement with the published values (median - blue solid, $16^{\mathrm{th}} - 84^{\mathrm{th}}$ percentiles - blue shaded). The model before correcting for the DoF bug (red dashed) shifts at $G \sim 13$ the magnitude at *Gaia* switches from 2D to 1D observations. The systematic underestimate of the prediction against the median published astrometry is expected to be due to remaining calibration uncertainties which I have not fully accounted for.

may be linked to time dependence of σ_{AL} which produces systematic uncertainties at the same level however the exact cause of this discrepancy for $\sigma_{5\mathrm{Dmax}}$ is unclear.

The predicted and observed distribution of $\sigma_{5\mathrm{Dmax}}$ on the sky are shown in Fig. 5.14 with the right panel showing strong agreement across the majority of the sky. Some residual streaks still persist in the South East and North West regions of the sky which match those seen in Section 5.4 when comparing the predicted and observed astrometry variances. These correspond to broken scans in *Gaia* DR2 which have not previously been diagnosed. I use the HEALPix time extractor tool (Holl, 2021) to constrain the times at which these scans happened in DR2. The clearest time ranges are given in Table 5.2 where the time range OBMT= 1556 – 1560rev is the direct cause of the residual streaks discussed above.

$\sigma_{5\mathrm{Dmax}}$ is published for all sources in *Gaia* DR2 whether or not they have published 5D astrometry. I can therefore use the published $\sigma_{5\mathrm{Dmax}}$ to estimate ρ for all stars in DR2

$$\rho = \frac{\lambda_{\max}\left[S\Phi^{-1}S\right]}{\sigma^2_{5\mathrm{Dmax}}}. \tag{5.35}$$

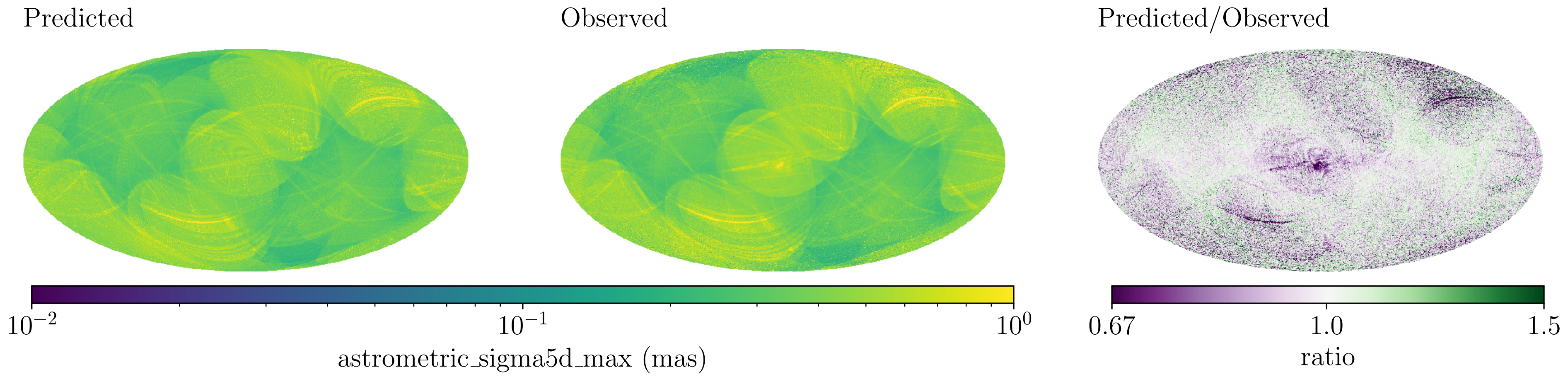


Fig. 5.14 The predicted $\sigma_{5\mathrm{Dmax}}$ (left) for $G \in [18.1, 19.0]$ agrees well with the HEALPix level 7 binned median published values (middle) across the sky in Galactic coordinates. The ratio between the predicted and observed shows some weak residuals in the Galactic bulge and a bad scan which impacts areas of sky which have not been significantly observed in *Gaia* DR2.

Start	End	Magnitudes
1447	1449	19.05 - 19.95
1453	1457	20.00 - 21.00
1556	1560	18.10 - 21.00
1730	1732	20.00 - 21.00

Table 5.2 Time periods producing un-modelled scan features in ASTROMETRIC_SIGMA5D_MAX. All times are given in OBMT (rev).

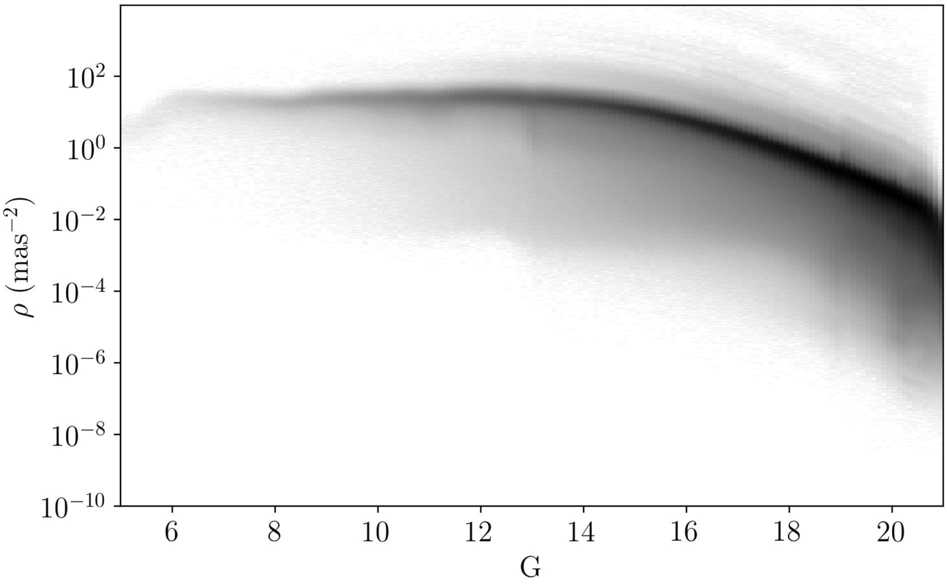


Fig. 5.15 ρ encodes the magnitude dependence of the predicted astrometric precision of *Gaia* DR2. 5D astrometric covariance is only published for the subset of DR2 with 5D astrometry however σ_{5Dmax} is published for all sources in DR2. I estimate ρ for all sources in DR2 with $k_{\mathrm{VP}} > 5$ using Eq. 5.35 shown here as a function of magnitude.

The distribution of ρ as a function of magnitude is shown in Fig. 5.15 where the distribution is largely flat at brighter magnitudes whilst declining for $G > 13$ due to low photon count noise. The spread to lower values is driven by excess noise due to binaries and other accelerating or extended sources.

In every 0.1 mag bin I fit a two component Gamma mixture model (ΓMM) to model the distribution of ρ,

$$\mathrm{P}(\rho) = \pi_1\,\Gamma(\rho;\alpha_1,\beta_1) + \pi_2\,\Gamma(\rho;\alpha_2,\beta_2). \tag{5.36}$$

One component of the mixture model fits the peak of the distribution which is dominated by well behaved simple point sources whilst the second component has an extended tail to low ρ which accounts for sources with significant excess noise. Examples of these fits in four magnitude bins are shown in Fig. 5.16 demonstrating reasonable agreement at dim magnitudes whilst somewhat cutting through the low ρ tail at bright magnitudes. At dim magnitudes, there is also a small excess of sources at large ρ. The precise cause of this

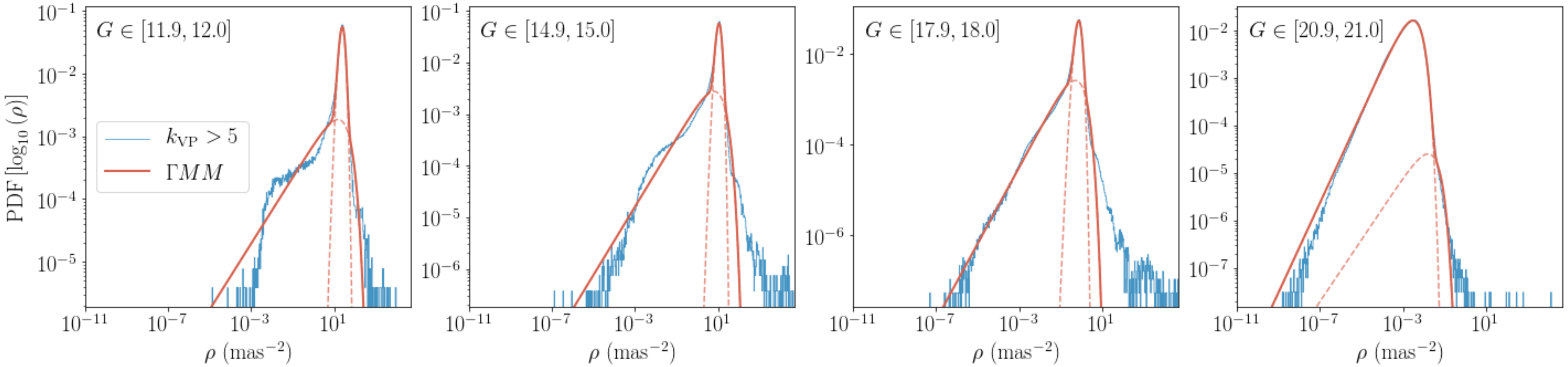


Fig. 5.16 The distribution of ρ in 0.1mag bins for *Gaia* DR2 sources with $k_{\mathrm{VP}} > 5$ (blue histograms) consists of a sharp peak of well behaved inertial point sources with a long wing to low ρ from sources with high excess error. This is fit with a two component Gamma Mixture Model (ΓMM) with one component fitting the peak and the second accounting for the low-ρ wing (red solid line). Red-dashed lines show the individual Γ components.

$\log(\alpha)$	$\mathrm{U}[-\infty, \infty]$
$\log(\beta)$	$\mathrm{U}[-\infty, \infty]$
π	Dirichlet$(a = [2, 2])$

Table 5.3 Priors used for ΓMM fit to ρ distribution.

tail is unclear but since any cuts on $\sigma_{5\mathrm{Dmax}}$ are on the low ρ end, the fact that I have not correctly modelled the high ρ tail would only generate a $< 1\%$ systematic uncertainty in the inferred selection function. Priors used for each of the parameters in the ΓMM are given in Table 5.3. The parameters are fit using expectation maximisation and posterior distributions produced using emcee (Foreman-Mackey et al., 2013).

The behaviour of the ΓMM parameters as a function of magnitude is modelled with a single Gaussian Process. For values of the same parameter at different magnitudes, the $\mathcal{GP}$ uses a square exponential kernel with variance s and scale length l. For different parameters, I assume no intrinsic correlation, however, correlations are introduced between different parameters of the same magnitude bin through the covariance of MCMC samples. Applying k-fold cross validation with $k = 5$ I infer hyperparameter values of $l = 0.224$, $s = 2.578$. The posterior $\mathcal{GP}$ is shown in Fig. 5.17 where the blue solid and red dashed lines are the two components for each parameter. Due to a lack of bright sources in *Gaia* DR2 astrometry, the $\mathcal{GP}$ at the bright end is dominated by the prior from the kernel. Since a negligible proportion of stars are influenced by the $\sigma_{5\mathrm{Dmax}}$ cut at these magnitudes, this is not a significant issue for the model.

Using the ΓMM as a function of magnitude, the selection function probability is given by

$$\mathrm{P}(\sigma_{5\mathrm{Dmax}} < 1.2\gamma(G)|k_{\mathrm{VP}} > 5) = \int_{\rho_{\min}}^{\infty} \sum_{j=1}^{2} \pi_j(G) \Gamma\left[\rho; \alpha_j(G), \beta_j(G)\right] \mathrm{d}\rho$$

where $\rho_{\min} = \frac{\lambda_{\max}[S\Phi S]}{(1.2\gamma(G))^2}$ from substituting $\sigma_{5\mathrm{Dmax}} = 1.2\gamma(G)$ into Eq. 5.35.

The selection probability is given at three magnitudes in Fig. 5.18 demonstrating that the cut only has a significant effect for $G > 20$. At the faintest magnitudes, regions of the sky which have been only sparsely scanned in *Gaia* DR2 are most likely to be removed due to the cut on $\sigma_{5\mathrm{Dmax}}$. In the most extreme cases such as in the Milky Way bulge, this can result in $< 1\%$ completeness in the *Gaia* DR2 astrometry sample.

Due to the simplicity of my 2 component ΓMM, the fits to the distribution of stars can produce significant offsets from the true distribution of data at the low ρ tail as is seen in the fourth panel of Fig. 5.16. The overestimate of the number of sources at low ρ in this case leads to a significant overestimate of the number of stars with high $\sigma_{5\mathrm{Dmax}}$ which subsequently get cut from the 5D astrometry sample. For now I consider the method a proof of principle for applying the ASF in order to derive the selection function.

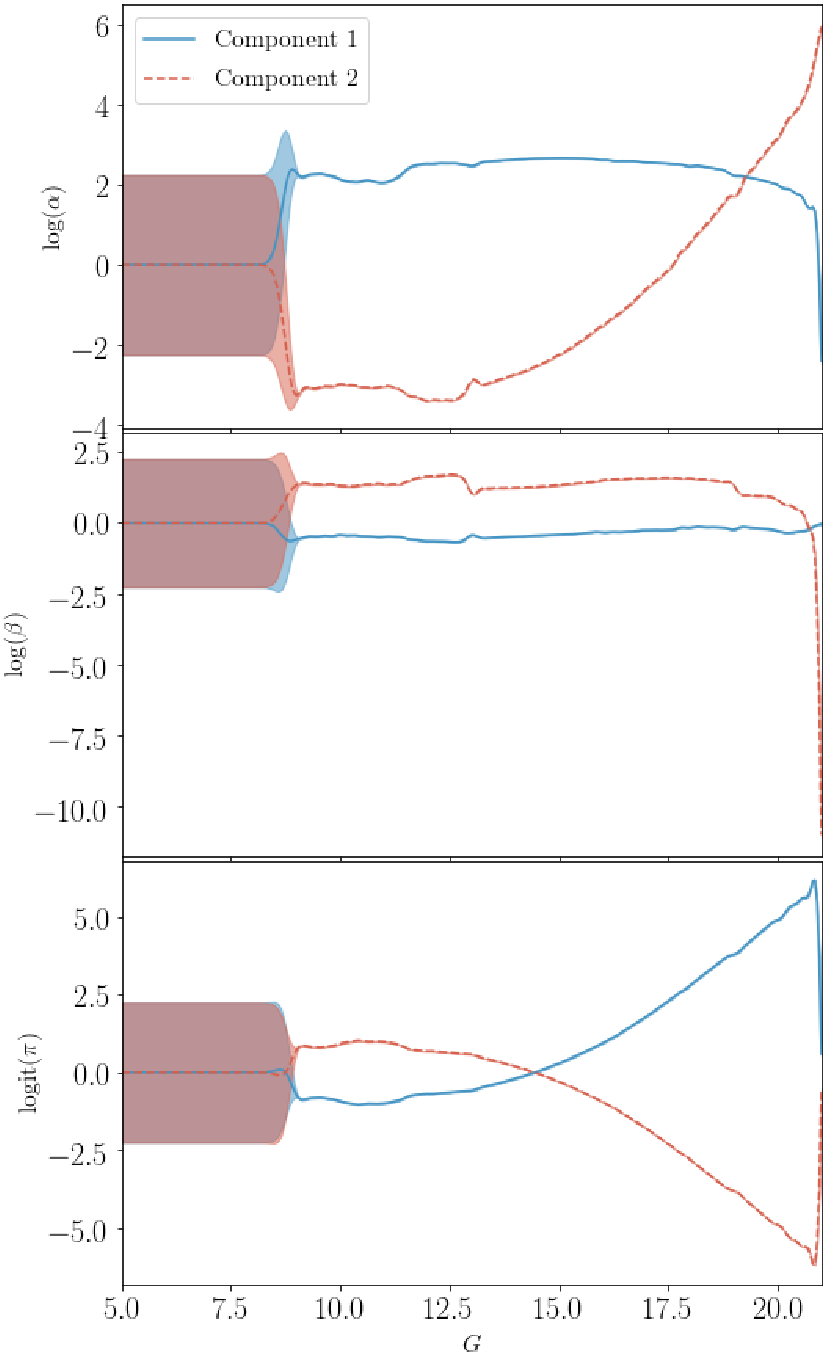


Fig. 5.17 The five parameters of the two component ΓMM are fit with with a single $\mathcal{GP}$ as a function of magnitude with a square exponential covariance kernel for matching parameters using the posterior MCMC samples from each magnitude bin. Small volumes of data at the bright end mean that the $\mathcal{GP}$ is dominated by the prior with mean 0 and variance $s = 2.578$.

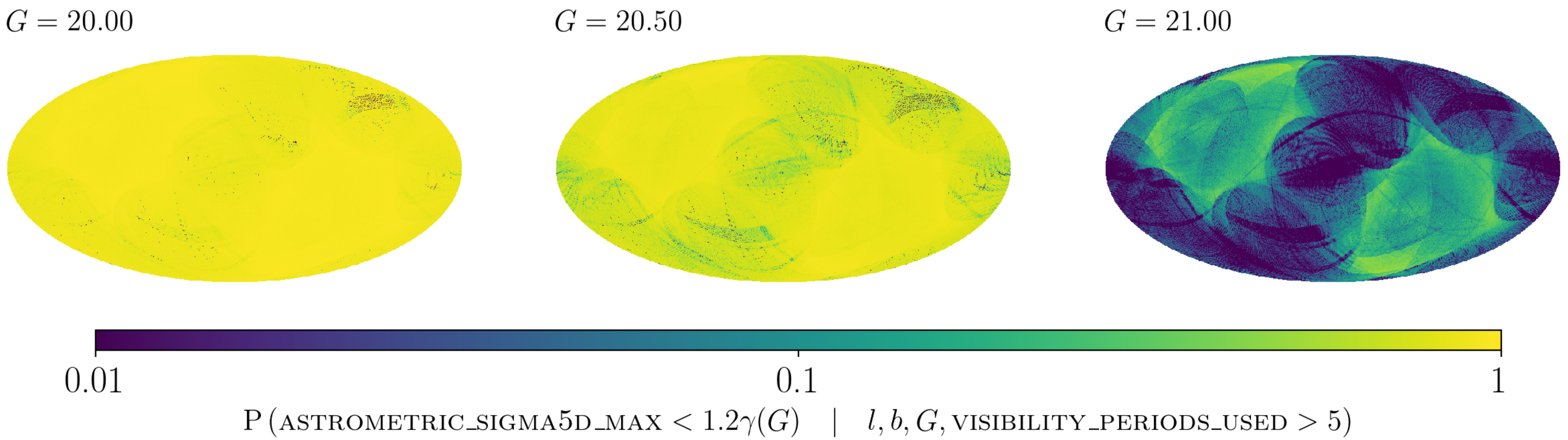

$G = 20.00$
$G = 20.50$
$G = 21.00$
0.01
0.1
1
P (ASTROMETRIC_SIGMA5D_MAX $< 1.2\gamma(G)$ | l, b, G, VISIBILITY_PERIODS_USED > 5)

Fig. 5.18 The selection probability of passing the $\sigma_{5\mathrm{Dmax}}$ cut is estimated from the ΓMM fits as a function of magnitude and position on the sky in HEALPix level 7 bins presented in Galactic coordinates. For $\boldsymbol{G} \sim 20$ (left), a negligible portion of stars are removed by the cut however at $\boldsymbol{G} \sim 21$ (right), the faintest magnitude for *Gaia* astrometry, the vast majority of sources in low scanned regions of the sky are removed from the sample.

5.7 Discussion

5.7.1 Excess Covariance

In this chapter I derived and discussed the importance of the ASF for analysing simple point sources in *Gaia* DR2. However I have not established how to use the ASF to estimate the excess covariance or precisely how this can be interpreted.

Consider a source with true 5D astrometry, r. However the source is not a simple point source such that the apparent position as a function of time is not well modelled by the 5D astrometric solution. If the excess noise may be parameterised by a 5D covariance, E, the probability of measuring the apparent 5D astrometry as $\mathbf{r}_E$ is given by

$$\mathrm{P}(\mathbf{r}_E) = \mathcal{N}(\mathbf{r}_E\,;\,\mathbf{r},\mathsf{E}). \tag{5.37}$$

If one attempts to measure this source, the uncertainty with which the 5D astrometry is measured is given by the ASF

$$\mathrm{P}(\mathbf{r}') = \mathcal{N}(\mathbf{r}'\,;\,\mathbf{r}_E,\Sigma). \tag{5.38}$$

By multiplying the two distributions together and marginalising over $\mathbf{r}_E$, I can determine the probability distribution of the measured 5D astrometry

$$\begin{aligned}\mathrm{P}(\mathbf{r}') &= \int \mathrm{d}^5\mathbf{r}_E\, \mathcal{N}(\mathbf{r}'\,;\,\mathbf{r}_E,\Sigma)\, \mathcal{N}(\mathbf{r}_E\,;\,\mathbf{r},\mathsf{E}) \\ &= \mathcal{N}\left(\mathbf{r}'\,;\,\mathbf{r},\left(\Sigma^{-1}+\mathsf{E}^{-1}\right)^{-1}\right) \\ &= \mathcal{N}(\mathbf{r}'\,;\,\mathbf{r},\mathsf{C}).\end{aligned} \tag{5.39}$$

Therefore, in this vastly oversimplified situation, the final measurement uncertainty for the 5D astrometry is given by the convolution of the excess noise and the ASF (providing the contribution from the observation measurement uncertainty).

There are two significant issues with this interpretation when considering the astrometry published by *Gaia*. Firstly, the AGIS pipeline does not formally infer the measurement uncertainty induced by excess noise. Residuals beyond simple point source astrometry are absorbed into a 1D excess noise parameter for each source as well as impacting the weights used for the given observations. The second problem is that source excess noise can disguise itself as a shift in the simple point source astrometry. As shown in Penoyre et al. 2020, excess binary motion can have complex effects on the posterior astrometry from *Gaia* including a phenomenon called the proper motion anomaly (Kervella et al., 2019). Interpretation of the excess covariance requires simulating stellar populations and emulating the AGIS pipeline in order to forward model how the intrinsic properties of the source relate to the posterior excess.

5.7.2 Mock Observations

Whilst I have entirely focused on the implications of the ASF for constraining excess source noise, it is also directly applicable to simulations in order to generate mock *Gaia* catalogues for Milky Way analogues.

Recent simulations such as Auriga (Grand et al., 2017) and VINTERGATAN (Agertz et al., 2020) have demonstrated the ability of the latest generation of cosmological simulations to produce Milky Way analogues which are excellent tools for studying the physical processes which govern the evolution of our galaxy. Performing a direct comparison with *Gaia* observations requires the *Gaia* selection functions and measurement uncertainty. The ASF provides the expected uncertainty of 5D astrometry for a simple point source. Given a simulated star with astrometry $\mathbf{r}$ as observed from the sun, the astrometry that would be measured by *Gaia*, $\mathbf{r}'$ can be inferred by sampling from the ASF

$$\mathbf{r}' \sim \mathcal{N}(\mathbf{r}, \Sigma(G, l, b)). \tag{5.40}$$

5.8 Summary of Uncertainty

The Astrometry Spread Function is the astrometric uncertainty distribution which would be expected for a point source with linear motion relative to the solar system barycentre (simple point source) given the source apparent magnitude and position on the sky. *Gaia*'s DPAC estimate the astrometric solution using an iterative linear regression algorithm. Given the uncertainty of individual observations and the scanning law, I have been able to reconstruct the astrometric covariance that would be expected for a simple point source observed by *Gaia* DR2. The ASF is a 5D multivariate Gaussian distribution with mean $\mathbf{0}$ and covariance $\Sigma \in \mathbb{R}^{5\times5}$ where I have formally derived $\Sigma(G, l, b)$.

Assuming the bulk of stars in the *Gaia* DR2 5D astrometry sample are simple point sources down to *Gaia*'s detection limit, I compare my result with the published covariances and find extremely good agreement down to sub-degree scales on the sky. The only regions with marginal disagreement are the highest source density regions of the bulge where the combination of source crowding and few scans in *Gaia* DR2 invalidate my assumptions. Therefore I caution the use of the ASF in highly crowded regions with low scan counts.

I have outlined three core applications of the ASF relevant to studies of the content and kinematics of the Milky Way.

(i) The ASF can be used to find sources with excess astrometric error indicitave of binary systems and extended sources. I used the ASF in combination with the published covariance to infer unit weight error for *Gaia* DR2 sources. The strong agreement with the published UWE demonstrates that the ASF can be used to find the excess error in *Gaia* observations due to physical source characteristics. The ASF is a valuable tool for exploiting *Gaia* data to model binary stars, astrometric microlens events and extended sources.

(ii) I applied the ASF to predict the selection function contribution from the cut on ASTROMETRIC_SIGMA5D_MAX used to generate the *Gaia* DR2 5D astrometry sample.

This is a key component of the full astrometry selection function which is a vital tool for unbiased modelling of Milky Way kinematics from *Gaia*'s 5D astrometry. Whilst my application here produced less reliable results than I achieved in Chapter 4, I have demonstrated that I can reach much greater spatial resolution using the scanning law so this is an avenue worth pursuing further.

(iii) The ASF can be used to generate *Gaia*-like astrometric solutions for mock catalogues produced from simulations such as *Auriga* (Grand et al., 2017) and synthetic population generators such as *Galaxia* (Sharma et al., 2011).

The final point comes in handy in the following chapter where I demonstrate the efficacy of Milky Way structure inference on a mock *Gaia* sample.

5.9 Accessing the ASF

The ASF is accessible through the Python package scanninglaw (https://github.com/gaiaverse/scanninglaw) (Boubert et al., 2021b). The user can ask the question 'What astrometric covariance would *Gaia* have published if my star was a simple point source?'.

I demonstrate this by determining the ASF covariance of the fastest main-sequence star in the Galaxy (S5-HVS1, Koposov et al., 2020) for *Gaia* DR2. The diagonal elements of the output covariance give the variance in α_0^*, δ_0, ϖ (mas^2), μ_α^*, μ_δ^* (mas^2/y^2).

```
import scanninglaw.asf as asf
from scanninglaw.source import Source

dr2_sl = asf.dr2_asf(version='cog')
s5_hvs1 = Source('22h54m51.68s',
                 '-51d11m44.19s',
           photometry={'gaia_g':16.02},
           frame='icrs')
Sigma = dr2_sl(s5_hvs1)

print('ASF Covariance: \n', Sigma)

>> ASF Covariance Position:
[[ 0.0005,  0.0004, -0.0005,  0.0003,  0.0006],
 [ 0.0004,  0.0029, -0.0023,  0.0016,  0.0013],
 [-0.0005, -0.0023,  0.0057, -0.0026, -0.0034],
 [ 0.0003,  0.0016, -0.0026,  0.0038,  0.0017],
 [ 0.0006,  0.0013, -0.0034,  0.0017,  0.0096]]
```

6

Photo-Astrometric Tracer Density of the Milky Way

Obi-Wan Kenobi   Jocasta Nu

"It should appear in this quadrant here, just south of the Rishi maze."

"It looks like the system you're searching for doesn't exist."

"Impossible, perhaps the archives are incomplete."

"If an item does not appear in our records, it does not exist."

Star Wars, Attack of the Clones

I'm not aware of how many systems are supposedly stored in the Jedi archive but the *Gaia* archive holds data on 1 811 709 771 sources which we believe is a small fraction of the number of stars in the Milky Way. If there are no stars in a given region of space, such as just south of Obi-Wan's "Rishi maze", is that because no stars exist in that region or because they couldn't be observed? And if stars in that region of parameter space couldn't be observed, what is the likelihood that there's anything there?

I've produced selection functions for the *Gaia* source catalogue and subset of 1.5 billion objects with measured parallax. For any region of parameter space, this tells me "If there were a source here, what is the probability it would be in the *Gaia* catalogue with measured parallax?". By defining and fitting a model to the observed data, I can interpolate and extrapolate to regions of parameter space which *Gaia* was unable to observe and therefore infer the expected number density of sources in that region. Constructing and fitting such a model to the vertical distribution of Milky Way stars is the subject of this chapter.

6.1 Complications when Fitting 1.5 Billion Sources

The 3D distribution of stars throughout the Milky Way is vital for understanding the formation history of our Galaxy. This 'tracer density' is also a key ingredient in methods attempting to estimate the distribution of dark matter in the Milky Way with important implications for both cosmological models and direct detection experiments (Read, 2014).

The primary aim of the *Gaia* mission is to measure the three-dimensional spatial and three-dimensional velocity distribution of stars (Gaia Collaboration et al., 2016) of which I focus on the spatial part. To achieve this, *Gaia* DPAC[1] have published parallax measurements for 1,467,744,818 sources (Gaia Collaboration et al., 2021a) with precisions down to 10^{-2} mas (Lindegren et al., 2021a) providing geometric distance estimates with no assumptions about source intrinsic brightness. Given this quality of data, one would be forgiven for thinking a detailed 3D map of the Milky Way would be a trivial task. Inferring spatial distributions of sources from *Gaia* data is a complex statistical problem for two key reasons.

(i) Until recently, the completeness limits of the *Gaia* catalogues were largely unknown. The observation strategy of the mission results in a completeness which varies significantly across the sky on sub-degree scales. Without a selection function, it is impossible to generate an unbiased map of the Milky Way using the full power of the *Gaia* data.

(ii) The second reason is that parallax-based distances are statistically awkward to work with. Much of our statistical methodology is constructed around the assumption of Gaussian measurement uncertainties motivated by the central limit theorem. Parallax uncertainties are Gaussian distributed which means that distances are reciprocal Gaussian distributed. This is a highly asymmetric distribution which, under an improper uniform prior, cannot be normalised. As such the distribution has no finite mean. Detailed discussions on how to use *Gaia* parallaxes for distance inference on individual stars are given in Bailer-Jones 2015b and Luri et al. 2018.

In spite of these hurdles, the structure of the Milky Way has been studied in detail, more often than not without *Gaia* data. A work-around to the challenges of parallax uncertainties is to focus on particular stellar populations for which the intrinsic brightness can be modelled. The distance can then be inferred from the measured apparent brightness. In some cases simple stellar colour-absolute magnitude relations are used for either a large population of sources across the CMD (e.g. Bilir et al., 2006a; Dobbie & Warren, 2020) or a small subset (e.g. horizontal branch stars, Fukushima et al., 2019). This approach has been taken further by using full stellar evolution models to infer intrinsic source brightness (de Jong et al., 2010). Period-luminosity relations for certain variable sources are also incredibly valuable distance indicators. Ak et al. 2008 used cataclysmic variables to estimate the vertical profile of the Milky Way disc whilst Mateu & Vivas 2018 used RRLyrae to determine the structure of the old thick disc and radial profile of the halo.

Some of these approaches apply uncertain colour-magnitude relations to large populations across the CMD leaving the results susceptible to systematic biases. Other approaches use more carefully chosen sub-samples of specific stellar types such that only a small fraction of the data are used. Many choose not use *Gaia* data at all (Dobbie & Warren, 2020; Fukushima et al., 2019; Mateu & Vivas, 2018).

Thanks to the results of Chapter 4, I have selection functions for the subsample of *Gaia* with measured parallax. In this chapter, I develop a method to overcome the challenges of

[1] DPAC is the *Gaia* Data Processing and Analysis Consortium who we have to thank for producing the exquisite quality of data.

directly using *Gaia* parallaxes and demonstrate its veracity on a *Gaia*-like mock sample with a known ground-truth. I limit the scope of this work to a high latitude region of the sky for statistical and computational reasons and due to the challenge of dust extinction which I do not attempt to solve here. I leverage this method to estimate the scale height of the Milky Way thin and thick discs, the radial profile of the halo and the local number density of stars for each component.

In Section 6.2 I show the likelihood optimization method used for this work followed by a full description of the model in Section 6.3. The *Gaia*-like mock sample is explained in Section 6.4 where I demonstrate the application of the method in Section 6.5. In Section 6.6 I introduce the *Gaia* EDR3 sample and describe the cuts used to remove a small number of contaminants. The model is fit to the *Gaia* data and I publish the results in Section 6.7. There are various simplifications and approximations used in the method and model which could, in principle, bias the parameter fits. These are discussed and tested in Section 6.8 and I explain how my tests are used to quantify statistical and systematic uncertainties in Section 6.9. Finally, I discuss the results in comparison to the literature along with additional considerations for this method such as source astrometric excess noise and the importance of kinematic information (which is not included in this work) in Section 6.10.

6.2 Method

The probability of drawing a population of objects $\{\mathbf{x}_i\}$ from a density profile $\lambda(\mathbf{x})$ is given by the Poisson likelihood function (for which the derivation is given in Appendix A.2),

$$\log\mathcal{L} = \sum_{i=1}^{N} \log\left(\lambda(\mathbf{x}_i)\right) - \int \mathrm{d}\mathbf{x}\lambda(\mathbf{x}). \tag{6.1}$$

The observed population of objects is drawn from the true underlying distribution of sources multiplied by a selection function which gives the probability of a source being included in the survey. Therefore I can substitute $\lambda(\mathbf{x}) = f(\mathbf{x},\boldsymbol{\psi})\mathcal{S}(\mathbf{x})$ where $\mathcal{S}$ is the selection function and f is the true underlying source density with model parameters $\boldsymbol{\psi}$,

$$\log\mathcal{L} = \sum_{i=1}^{N} \log\left(f(\mathbf{x}_i,\boldsymbol{\psi})\mathcal{S}(\mathbf{x}_i)\right) - \int \mathrm{d}\mathbf{x} f(\mathbf{x},\boldsymbol{\psi})\mathcal{S}(\mathbf{x}). \tag{6.2}$$

The aim of density estimation is to fit the parameters of the true underlying distribution, $\boldsymbol{\psi}$. Since the selection function is independent of the model parameters, it can be dropped out of the first term in the likelihood function,

$$\log\mathcal{L} \sim \sum_{i=1}^{N} \log\left(f(\mathbf{x}_i,\boldsymbol{\psi})\right) - \int \mathrm{d}\mathbf{x} f(\mathbf{x},\boldsymbol{\psi})\mathcal{S}(\mathbf{x}). \tag{6.3}$$

The source properties, $\mathbf{x}$, need to be chosen according to the dependencies of the model and selection function.

Mateu & Vivas 2018 use this method on a sample of RR Lyrae to constrain the structure of the thick disc and halo considering only spatial dimensions, while Bovy et al. 2012b apply a more complex model to a population of G-dwarfs to fit the Milky Way disc using measured apparent magnitude, colour and metallicity. The aim of this work is to model the purely spatial distribution of sources, however the selection function, which I discussed in Chapters 3 and 4, is a function of position on the sky and apparent magnitude. Therefore, I must also consider the intrinsic brightness of a source, so my source properties are $\mathbf{x} = (l, b, s, M_G)$.

An additional complexity I introduce beyond previous works is accounting for parallax measurement uncertainties, which is vital when working with *Gaia* astrometry. Suppose instead that $\mathbf{x}$ are the *measured* source properties and $f(\mathbf{x}, \boldsymbol{\psi})$ is the expected distribution of measured source properties given the model. Source measurements are drawn from an uncertainty distribution, $\mathrm{P}(\mathbf{x}\,|\,\mathbf{x}_\mathrm{T})$ where $\mathbf{x}_\mathrm{T}$ are the underlying *true* source properties. The measured model (f) is given by a convolution between the true underlying model (f_T) and the measurement error distribution,

$$f(\mathbf{x}, \boldsymbol{\psi}) = \int \mathrm{d}\mathbf{x}_\mathrm{T}\, \mathrm{P}(\mathbf{x}\,|\,\mathbf{x}_\mathrm{T})\, f_\mathrm{T}(\mathbf{x}_\mathrm{T}, \boldsymbol{\psi}). \tag{6.4}$$

Substituting this into the likelihood, I get

$$\begin{aligned}\log \mathcal{L} \sim \sum_{i=1}^{N} \log \left(\int \mathrm{d}\mathbf{x}_\mathrm{T}\, \mathrm{P}(\mathbf{x}_i\,|\,\mathbf{x}_\mathrm{T})\, f_\mathrm{T}(\mathbf{x}_\mathrm{T}, \boldsymbol{\psi}) \right) \\ - \int \mathrm{d}\mathbf{x}_\mathrm{T}\, f_\mathrm{T}(\mathbf{x}_\mathrm{T}, \boldsymbol{\psi}) \int \mathrm{d}\mathbf{x}\, \mathrm{P}(\mathbf{x}\,|\,\mathbf{x}_\mathrm{T})\, \mathcal{S}(\mathbf{x})\end{aligned} \tag{6.5}$$

where I have reversed the order of integration in the second term and brought f_T outside the integral over measured parameters.

My measured source properties are $\mathbf{x} = (l, b, G, \varpi)$, or Galactic longitude and latitude, apparent magnitude and parallax. In this work, I consider parallax error as the only significant measurement uncertainty. Positional uncertainties in (l, b) are extremely small and I test the impact of neglecting error in G in Section 6.8. Therefore, the error term becomes

$$\mathrm{P}(\mathbf{x}\,|\,\mathbf{x}_\mathrm{T}) = \delta(l - l_T)\, \delta(b - b_T)\, \delta\left(G - G_T(s, M_G)\right)\, \mathrm{P}(\varpi\,|\,s). \tag{6.6}$$

I integrate over all delta functions in the first term of the likelihood function

$$\int \mathrm{d}\mathbf{x}_\mathrm{T}\, \mathrm{P}(\mathbf{x}_i\,|\,\mathbf{x}_\mathrm{T})\, f_\mathrm{T}(\mathbf{x}_\mathrm{T}, \boldsymbol{\psi}) = \int \mathrm{d}s\, \mathrm{P}(\varpi_i\,|\,s)\, f_\mathrm{T}(l_i, b_i, G_i, s, \boldsymbol{\psi}). \tag{6.7}$$

The selection function is a function of l, b and G only; there is no dependence on measured parallax (see Chapter 4). This makes it easy to integrate over

$$\int \mathrm{d}\mathbf{x}\, \mathrm{P}(\mathbf{x}\,|\,\mathbf{x}_\mathrm{T})\, \mathcal{S}(l, b, G) = \mathcal{S}\left(l_T, b_T, G_T(s, M_G)\right). \tag{6.8}$$

Finally, I can substitute this into the likelihood function,

$$\log \mathcal{L} \sim \sum_{i=1}^{N} \log\left(\int \mathrm{d}s\, \mathrm{P}(\varpi_i \mid s)\, f_{\mathrm{T}}(l_i, b_i, G_i, s, \boldsymbol{\psi})\right) \\ - \int \mathrm{d}\mathbf{x}_{\mathrm{T}}\, f_{\mathrm{T}}(l_T, b_T, M_G, s, \boldsymbol{\psi})\, \mathcal{S}\left(l_T, b_T, G_T(s, M_G)\right). \tag{6.9}$$

This is the likelihood function which I use to fit the model parameters, $\boldsymbol{\psi}$, to the observed data. For the remainder of the chapter, I drop the subscript T with f always referring to the true underlying source distribution.

6.2.1 Parallax Error Integration

The biggest numerical challenge for my method is the parallax error convolution. I need to integrate over parallax for every source at every proposed set of model parameters. In this section I use slightly different notation where $\varpi = 1/s$ is the true parallax distance which I am marginalising over, and ϖ_i is the measured parallax for source i. The integral I need to evaluate is

$$\int_0^\infty \mathrm{d}s\, \mathrm{P}(\varpi_i \mid s)\, f(l_i, b_i, G_i, s, \boldsymbol{\psi}) = \int_0^\infty \mathrm{d}\varpi\, \varpi^{-2}\, \mathcal{N}(\varpi; \varpi_i, \sigma_{\varpi,i})\, f(l_i, b_i, G_i, s, \boldsymbol{\psi}) \\ \equiv \int_0^\infty \mathrm{d}\varpi\, I(\varpi) \tag{6.10}$$

where $\sigma_{\varpi,i}$ is the parallax error of source i. In Section 6.3 I introduce the absolute magnitude model which is broken into sections with an upper absolute magnitude limit (minimum brightness) for the model. I can then write the integral as a sum of definite integrals

$$\int_0^\infty \mathrm{d}\varpi\, I(\varpi) = \sum_j \int_{\varpi_j}^{\varpi_{j+1}} \mathrm{d}\varpi\, I(\varpi) \tag{6.11}$$

where

$$\varpi_j = 10^{(M_j + 10 - G_i)/5} \tag{6.12}$$

and M_j are the magnitude boundaries of the sections. For an unconstrained lower absolute magnitude limit, $\varpi_0 = 0$.

I numerically evaluate the integral of each section using the following five step recipe.

(i) Transform into logit-parallax space using the substitution

$$x' = \log\left(\frac{\varpi - \varpi_j}{\varpi_{j+1} - \varpi}\right). \tag{6.13}$$

This gives

$$\int_{\varpi_j}^{\varpi_{j+1}} \mathrm{d}\varpi I(\varpi) = \int_{-\infty}^{\infty} \mathrm{d}x' \frac{I}{J} \tag{6.14}$$

where the Jacobian is

$$J = \left|\frac{\partial x'}{\partial \varpi}\right| = \frac{\varpi_{j+1} - \varpi_j}{(\varpi - \varpi_j)(\varpi_{j+1} - \varpi)}. \tag{6.15}$$

(ii) Find the peak of the logit-transformed integrand by solving

$$\frac{\partial}{\partial x'}\left(\frac{I}{J}\right) = 0 \tag{6.16}$$

using the bisection algorithm with respect to ϖ initialising at the integration boundaries, ϖ_j, ϖ_{j+1}. Transform the parallax of the peak into logit space giving us the mode, x'_0.

(iii) Estimate the width of the peak from the curvature around x'_0,

$$\sigma_{x'} = \left(\frac{\partial^2 I/J}{\partial x'^2}\right)^{-\frac{1}{2}}\Bigg|_{x'=x'_0}. \tag{6.17}$$

(iv) Recentre and rescale via

$$x = \frac{x' - x'_0}{\sqrt{2}\sigma_{x'}}, \tag{6.18}$$

such that the integrand is approximately $I \sim \exp\left(-x^2\right)$ around the peak.

(v) Apply Gauss-Hermite quadrature in x-space which gives

$$\int_{\varpi_j}^{\varpi_{j+1}} \mathrm{d}\varpi\, I = \sum_k w_k \frac{\sqrt{2}\sigma_{x'}\, I(\varpi(x_k))}{J(\varpi(x_k))} \exp\left(x_k^2\right). \tag{6.19}$$

In my application of the method, I use Gauss-Hermite quadrature with 11 sample points. Increasing the number of sampling points has no appreciable effect on my inferred likelihood.

A major limitation of this method is that it cannot accurately integrate multimodal integrands. The integrand must be unimodal such that I can integrate around the single peak. I discuss the implications of this in Section 6.3 when introducing my model. However, since this is purely a numerical rather than conceptual challenge, I hope future work can improve on my method to allow for more general models to be evaluated and with greater computational efficiency.

6.3 Model

For this work, I only consider high latitudes, $|b| > 80°$. There are several reasons for this:

- Dust extinction is negligible at high latitudes. Modelling the 3D distribution of dust throughout the Milky Way is a complicated problem on its own (Green et al., 2014; Marshall et al., 2006). I quantify the impact of dust extinction on my results in Section 6.8.
- The in-plane structure of the Milky Way disc is complex with waves, spiral arms and the bar which add vast numbers of free parameters to any spatial model.

- Parallax integration is computationally expensive and scales linearly with the number of sources. By focusing on a subset of *Gaia* data, I am left with a computationally tractable problem.

The aim of this work is to demonstrate how *Gaia* parallax information can be used to obtain an unbiased model of the Milky Way's stellar content. The vertical distribution at the Solar neighbourhood is a tractable first step in this direction.

The vertical distribution of sources is assumed to be a mixture of three distinct components: thin disc, thick disc and halo. This canonical model has been used for decades since the addition of the second disc component by Gilmore & Reid 1983. More recent work has shown that – rather than a dichotomy into thin and thick discs – there may be a continuous evolution of disc height with stellar metallicity (Bovy et al., 2012a, 2016b). However, since metallicity is not an observable in my sample, I keep to the canonical distinct thin and thick disc model.

Within each component, I assume the spatial and absolute magnitude distributions are separable such that

$$f(l, b, \varpi, M_G) = \sum_{c=\{\mathrm{Tn},\mathrm{Tk},\mathrm{H}\}} w_c\, \nu_c(l, b, \varpi, \boldsymbol{\psi}_\nu)\, \phi_c(M_G, \boldsymbol{\psi}_\phi). \tag{6.20}$$

This is a significant assumption. The thin disc has undergone star formation over long periods and will have correlations between age and metallicity and the vertical and radial dispersion of orbits (e.g. Ivezić et al., 2008; Martig et al., 2016; Recio-Blanco et al., 2014; Snaith et al., 2015). Likewise, the halo is made of multiple stellar populations from in situ star formation and historical merger events (e.g. Belokurov et al., 2018; Belokurov et al., 2020a; Helmi et al., 2018). Nonetheless, I maintain this assumption here in the interests of keeping a simple and tractable model. Note that w_c is a free parameter of the model for each component and gives the total number of stars for that component within the given region of the sky and absolute magnitude range.

6.3.1 Spatial Distributions

I consider the thin and thick discs to have exponential profiles vertically $\nu_c \propto \exp\left(-\frac{|z|}{h_c}\right)$ similar to previous work (e.g. Bovy et al., 2012b, 2016b; Jurić et al., 2008). Other possibilities include sech or sech^2 profiles, but there is a moderate preference in the data for an exponential profile (Dobbie & Warren, 2020).

Since I am only considering high latitudes, I neglect any radial dependence of the vertical density profile. This makes the numerical integral described in Section 6.2.1 significantly more tractable. The complexity introduced by adding radial dependence is explained in more detail in Section 6.3.3. The impact of this simplification on the results is tested and quantified in Section 6.8.

Transforming into heliocentric coordinates $z = s\sin(b)$ and normalising I get the density distribution

$$\nu_c(l, b, s)\mathrm{d}V = \frac{\tan^2(|b|_{\min})}{2\pi\, h_c^3} s^2 \exp\left(-\frac{|s\sin b|}{h_c}\right) \mathrm{d}l\, \mathrm{d}\sin(b)\, \mathrm{d}s, \tag{6.21}$$

where $|b|_{\min} = 80^\circ$ is the on-sky latitude limit of my sample. In Section 6.7 I consider the northern and southern high latitude samples independently; however here, I assume a Milky Way symmetric about the Galactic plane, in which the Sun lies. This introduces a ~ 20.8pc systematic offset into my results (Bennett & Bovy, 2019), whose effect on the posterior distributions is quantified in Section 6.8.

For the spatial distribution of the halo, I use a spherically symmetric single power law profile centred on the Galactic centre, $\nu_\mathrm{H}(r)\mathrm{d}V \propto r^{-n_\mathrm{H}}$. Many other works also include a free parameter for the halo axis ratio (Jurić et al., 2008; Mateu & Vivas, 2018), however, as I am only using a narrow window on the sky, there is limited information to independently constrain the profile and axis ratio of the halo. Furthermore, previous works have either implicitly or explicitly truncated the halo or included a broken power-law profile. The halo profile used in this work is assumed to extend infinitely so we constrain $n_\mathrm{H} > 3$ to maintain a finite halo normalisation. This is in tension with Deason et al. 2014 and Fukushima et al. 2019 who find a steeper halo profile beyond $r \sim 50$ kpc and 160 kpc respectively. This corresponds to a parallax $\varpi < 0.02$ mas which is pushing the precision limit of *Gaia* parallaxes even for bright sources (see Fig. 7 of Lindegren et al., 2021a). Therefore, my model should not be significantly sensitive to this shift.

As I did for the disc profile, I neglect cylindrical radius dependence for the halo by placing all sources at the Solar radius such that

$$r^2 = s^2\sin^2(b) + R_\odot^2. \tag{6.22}$$

Again this is only valid at high latitudes. The dependence of the source distribution on Galactic longitude can then be neglected. This may lead to systematic biases which are tested in Section 6.8. The spatial model of the halo is given by

$$\nu_\mathrm{H}(l, b, s)\mathrm{d}V = \mathcal{N}_{\nu_\mathrm{H}} s^2 \left(s^2\sin^2 b + R_\odot^2\right)^{-\frac{n_\mathrm{H}}{2}}\ \mathrm{d}l\, \mathrm{d}\sin b\, \mathrm{d}s \tag{6.23}$$

where

$$\mathcal{N}_{\nu_\mathrm{H}} = \frac{1}{2\pi}\frac{8\tan^2(b_{\min})}{\sqrt{\pi}R_\odot^{3-n}}\frac{\Gamma\left(n/2\right)}{\Gamma\left(n/2 - 3/2\right)}. \tag{6.24}$$

This spatial distribution adds three parameters to the model: the exponential scale height of the thin and thick discs (h_Tn and h_Tk) and the power-law index of the halo (n_H).

6.3.2 Luminosity Functions

The luminosity distribution function of stars in the Milky Way is an intricate function of the star formation history, accretion history and dynamical evolution of the Galaxy. The

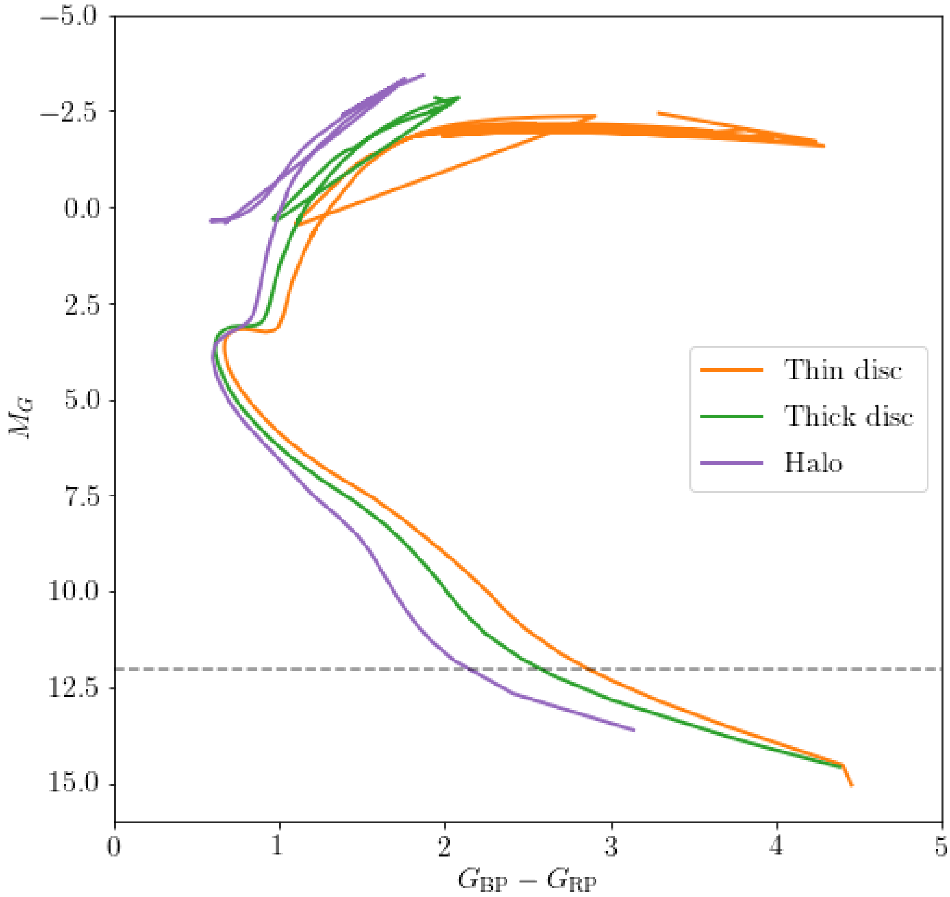


Fig. 6.1 HR diagram showing the isochrones used for my mock model of Milky Way sources, with ages $\tau = 6.9, 7.8, 12.5$ Gyr and metallicities [Fe/H] $= -0.3, -0.7, -1.5$ for the thin disc, thick disc and halo respectively (orange, green and purple). The grey dashed line shows the minimum absolute magnitude of my model – $M_G = 12$.

aim of this work is to derive the spatial distribution of sources in the Galaxy - independent of stellar populations - and so the magnitude distribution is only included in order to formally account for the survey selection function. In this section, I explain how to derive an adequate parameterisation for the luminosity function for each Milky Way component.

Each of the three Milky Way components is assumed to be a single mono-age, mono-abundance stellar population. Using the results of Kilic et al. 2017 from white dwarf populations, the ages used for the thin disc, thick disc and halo are 6.9 Gyr, 7.8 Gyr and 12.5 Gyr respectively. Using SDSS spectroscopy, Ivezić et al. 2008 derived halo and thick disc metallicities of [Fe/H] $= -1.5, -0.7$ respectively, whilst Recio-Blanco et al. 2014 used the *Gaia*-ESO survey (Gilmore et al., 2012b) to find the thin disc metallicity fell in the range $[-0.8, 0.2]$ and the thick disc between $[-1.0, -0.25]$. Combining these results, I assume the thin disc, thick disc and halo have metallicities of $-0.3, -0.7$ and -1.5. The HR diagram in Fig. 6.1 shows the three isochrones which are taken from PARSEC v1.2s (Bressan et al., 2012; Chen et al., 2014, 2015; Tang et al., 2014).

I then draw a random sample from the broken power law initial mass function (IMF) of Kroupa 2001 for initial masses greater than $0.09\,\mathrm{M}_\odot$ with $\mathcal{M}_{\mathrm{ini}} \sim \mathcal{M}_{\mathrm{ini}}^{-1.3}$ for $\mathcal{M}_{\mathrm{ini}} < 0.5\,\mathrm{M}_\odot$ and $\mathcal{M}_{\mathrm{ini}} \sim \mathcal{M}_{\mathrm{ini}}^{-2.3}$ otherwise. This is shown in the top panel of Fig. 6.2. The individual

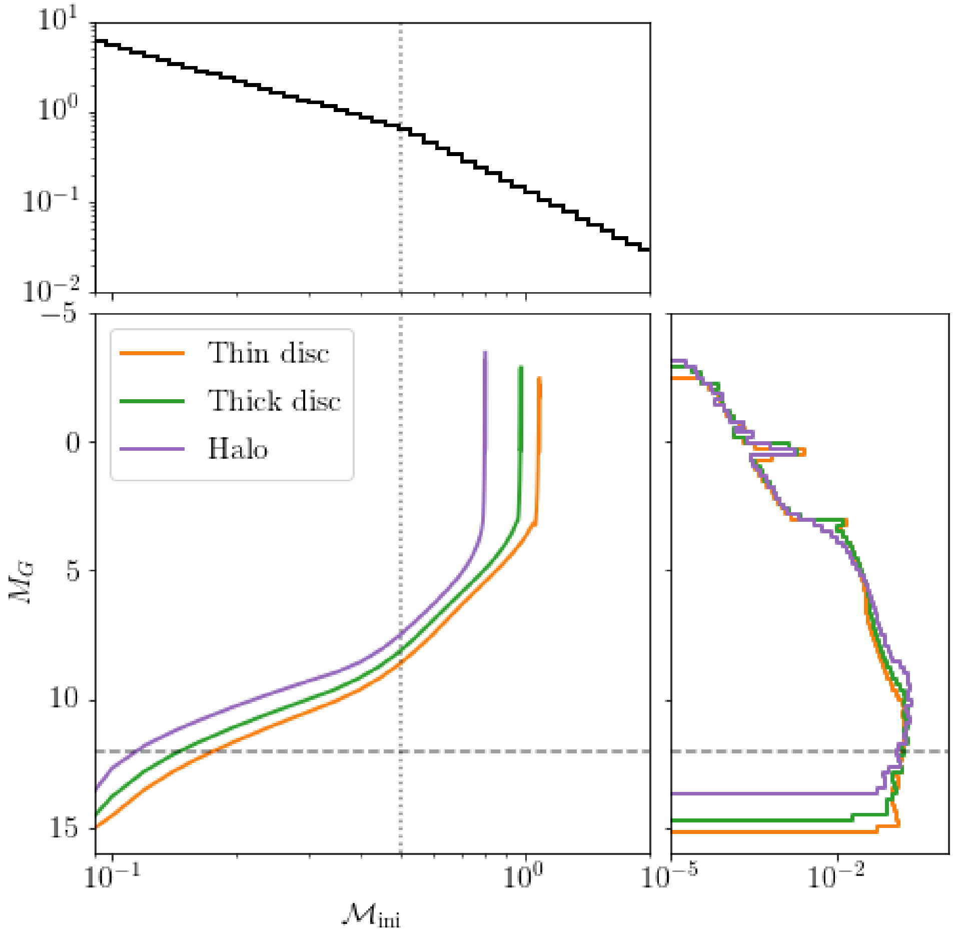


Fig. 6.2 The thin disc, thick disc and halo isochrones (orange, green and purple) are used to transform a mock sample of stars from initial mass ($\mathcal{M}_{\rm ini}$) to absolute magnitude (M_G). The initial mass (top panel) is drawn from a Kroupa IMF (Kroupa, 2001) with $\mathcal{M}_{\rm ini} > 0.09\mathrm{M}_\odot$ where the vertical grey dotted line is the break mass $0.5\,\mathrm{M}_\odot$. This produces the absolute magnitude distribution shown in the right hand panel. The horizontal grey-dashed line shows the maximum absolute magnitude as my model only includes sources with $M_G < 12$.

component isochrones, shown in the middle panel, are then used to transform the IMF into an absolute magnitude distribution which is shown in the right hand panel. This sample is not used as my mock catalogue, it is only for deriving my model absolute magnitude distribution.

The absolute magnitude distributions of the three components from the right hand panel of Fig. 6.2 are shown as shaded histograms in Fig. 6.3. They are made up of four regimes. At the bright end ($M_G \lesssim 3$), sources evolve much faster along the giant branch than the main sequence (MS), generating a sharp drop at the turn-off above which the number density of sources falls quickly aside from a spike at the red clump ($M_G \sim 0$). The MS has three components, a relatively shallow upper sequence for $M_G \sim [3, 7]$, a steeper section for $M_G \sim [7, 9]$ where the slope of the main sequence in Fig. 6.1 shifts which is also around the power-law break of the IMF (I refer to this section as the 'gap'), and a very flat lower MS for $M_G \gtrsim 9$. Sources continue fainter to the brown dwarf regime; however, stellar models in these regions of parameter space are poorly constrained by observations

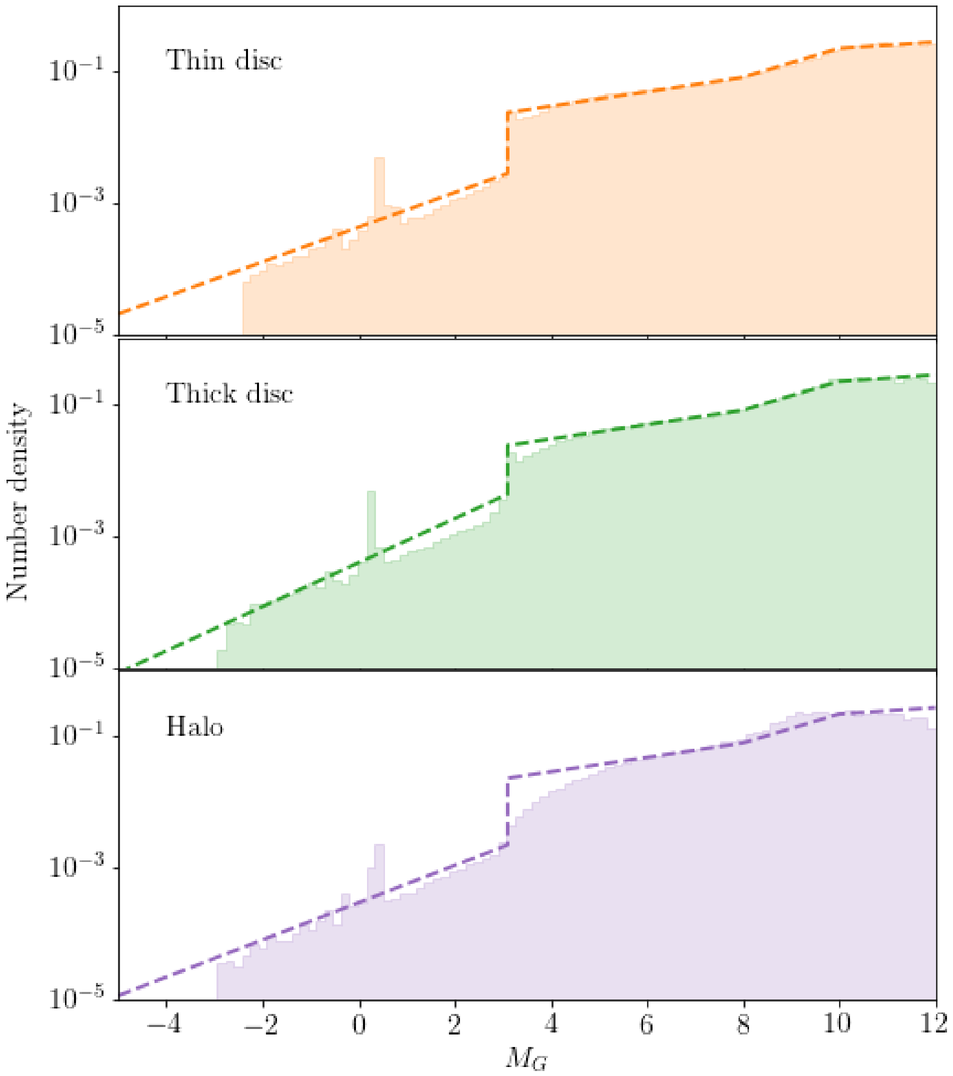


Fig. 6.3 The mock distributions produced by transforming the Kroupa IMF through isochrones from Fig. 6.1 (shaded histograms) are fit with the approximate absolute magnitude distribution used for the model (dashed lines).

as there are few stars this dim yet bright enough for current observatories. For this reason, I only consider sources with $M_G < 12$ in this work. This is especially beneficial when I model the *Gaia* data in Section 6.7 as the majority of sources with spurious astrometric solutions as classified by Rybizki et al. 2021a and Gaia Collaboration et al. 2021b have absolute magnitudes fainter than $M_G = 12$.

If one takes a population of sources with a power-law mass distribution and power-law mass-luminosity relation, the absolute magnitude distribution of the population is exponentially distributed. I assume each component of the absolute magnitude distribution is modelled by an exponential distribution. The absolute magnitude is drawn from a broken exponential distribution,

$$M_G \sim \exp(-\alpha M_G), \tag{6.25}$$

with four components

$$\alpha = \begin{cases} \alpha_1 & 9 < M_G < 12 \qquad \text{(Lower Main Sequence)} \\ \alpha_g & 7 < M_G < 9 \qquad \text{(Main Sequence `gap')} \\ \alpha_2 & M_{\mathrm{TO}} < M_G < 7 \qquad \text{(Upper Main Sequence)} \\ \alpha_G & M_G < M_{\mathrm{TO}} \qquad \text{(Giants)} \end{cases} \tag{6.26}$$

where M_{TO} is the turn-off magnitude.

The distribution is continuous everywhere other than the turnoff where the discontinuous change in the gradient of the magnitude-initial mass relation leads to a discontinuity in the magnitude distribution. Continuity conditions at $M_G = 7, 9$ constrain the exponential profile α_g and the normalisation A_g of the gap profile

$$\alpha_g = \frac{\log\left(\frac{\alpha_1\,(\epsilon_2-1)}{\alpha_2\,(\epsilon_1-1)}\right) - \alpha_1(M_{\mathrm{MS}}-9) + \alpha_2(M_{\mathrm{MS}}-7)}{9-7} \tag{6.27}$$

$$A_g = -\frac{2.5\alpha_1}{\log(10)\,(\epsilon_1-1)} \exp\left((\alpha_g-\alpha_1)(M_{\mathrm{MS}}-9)\right) \tag{6.28}$$

where $M_{\mathrm{MS}} = 8$ is the magnitude of the transition from the lower to upper main sequence and $\epsilon_1 = 1.3, \epsilon_2 = 2.3$ are the power law profiles of the Kroupa 2001 IMF.

The full magnitude distribution is given by

$$f(M)\mathrm{d}M = \begin{cases} (1-f_G)\mathcal{N}_D \frac{1}{\alpha_1} \exp\left(-\alpha_1(M-M_{\mathrm{MS}})\right) \mathrm{d}M & 9 < M < 12 \\ (1-f_G)\mathcal{N}_D A_g \exp\left(-\alpha_g(M-M_{\mathrm{MS}})\right) \mathrm{d}M & 7 < M < 9 \\ (1-f_G)\mathcal{N}_D \frac{1}{\alpha_2} \exp\left(-\alpha_2(M-M_{\mathrm{MS}})\right) \mathrm{d}M & M_{\mathrm{TO}} < M < 7 \\ f_G\mathcal{N}_G \exp(-\alpha_G(M-M_{\mathrm{TO}}))\mathrm{d}M & M < M_{\mathrm{TO}} \end{cases} \tag{6.29}$$

where $\mathcal{N}_D$ and $\mathcal{N}_G$ are the normalisations of the dwarf and giant magnitude distributions respectively.

The magnitude distribution introduces five parameters: α_1, α_2, M_{TO}, α_G and f_G, the fraction of the population which are giants, which constrains the size of the discontinuity at the turn-off. I could fix all parameters using the IMF-isochrone sample just constructed, however, this is only an approximate representation of the magnitude distribution which may introduce large systematics. To avoid this problem, I free up α_1, α_2 and f_G to be constrained by the real data. α_1 and α_2 are assumed to be the same for all populations as the MS is dominated by older stars which show a similar distribution independent of population parameters.

The position of the turn-off, M_{TO}, defines a discontinuity for the model. Depending on the location of individual sources in relation to the turnoff, this can generate sample-dependent local optima in the likelihood space which is challenging for optimization. For this reason, I fix $M_{\mathrm{TO}} = 3.1$ for all models and address the implications of this in Section 6.8. α_G has a strong degeneracy with f_G as both control the number of sources at bright magnitudes. I fix α_G to values which are discussed in Section 6.4 to avoid this degeneracy. All free parameters are listed in Table 6.2 with their respective components.

This fully defines the model which I fit to the *Gaia* data. In total, there are 11 free parameters of the model.

6.3.3 Integrand Limitations

In Section 6.2.1 I stated that the integral over parallax uncertainty becomes intractable for more complex models, I briefly justify that statement here where I use the example of the exponential disc model to demonstrate.

The integrand including R-dependence is

$$I\,\mathrm{d}\varpi \propto \varpi^{-4}\exp\left(-\frac{z}{h}-\frac{R}{L}\right)\exp\left(\frac{(\varpi-\varpi_i)}{2\sigma_{\varpi i}^2}\right)\mathrm{d}\varpi \tag{6.30}$$

where h and L are the scale height and length of the disc being considered. The Jacobian for the logit transformation I applied is

$$J \propto \frac{1}{(\varpi_{j+1}-\varpi)(\varpi-\varpi_j)}. \tag{6.31}$$

Taking the gradient of I/J, setting to zero (as in Eq. 6.16) and simplifying down, I am left with

$$-\frac{4}{\varpi}-\frac{1}{h}\frac{\partial z}{\partial\varpi}-\frac{1}{L}\frac{\partial R}{\partial\varpi}-\frac{(\varpi-\varpi_i)}{\sigma_{\varpi i}^2}+\frac{1}{(\varpi_{j+1}-\varpi)}-\frac{1}{(\varpi-\varpi_j)}=0 \tag{6.32}$$

where

$$z=\frac{\sin b}{\varpi}\quad\text{and}\quad R^2=R_\odot^2+\left(\frac{\cos b}{\varpi}\right)^2-\frac{2R_\odot\cos b\,\cos l}{\varpi}. \tag{6.33}$$

In my application, I have assumed no R-dependence, i.e. setting $L=\infty$. I have

$$\frac{\partial z}{\partial\varpi}=-\frac{\sin b}{\varpi^2} \tag{6.34}$$

and Eq. 6.32 simplifies to a quintic polynomial in terms of ϖ. I know that at least two solutions of the quintic are outside $[\varpi_j,\varpi_{j+1}]$ since

$$\frac{I}{J}\begin{cases}=0 & \text{for } \varpi=\varpi_j,\varpi_{j+1}\\ <0 & \text{for } \varpi\lesssim\varpi_j,\ \varpi\gtrsim\varpi_{j+1}\\ =0 & \text{for } \varpi=0,\infty\end{cases} \tag{6.35}$$

so there must be a stationary point above and below the boundaries. This leaves three stationary points in the integration range corresponding to two peaks or modes. My model is equivalent to the exponentially-decreasing square distance prior used by Section 7 of Bailer-Jones 2015b and they also find the same two modes. However two of the roots are often either complex, or, for $\varpi_i<0$, there is a mode with negative parallax which is outside the integration limits. Whilst I cannot guarentee that the integrand is always

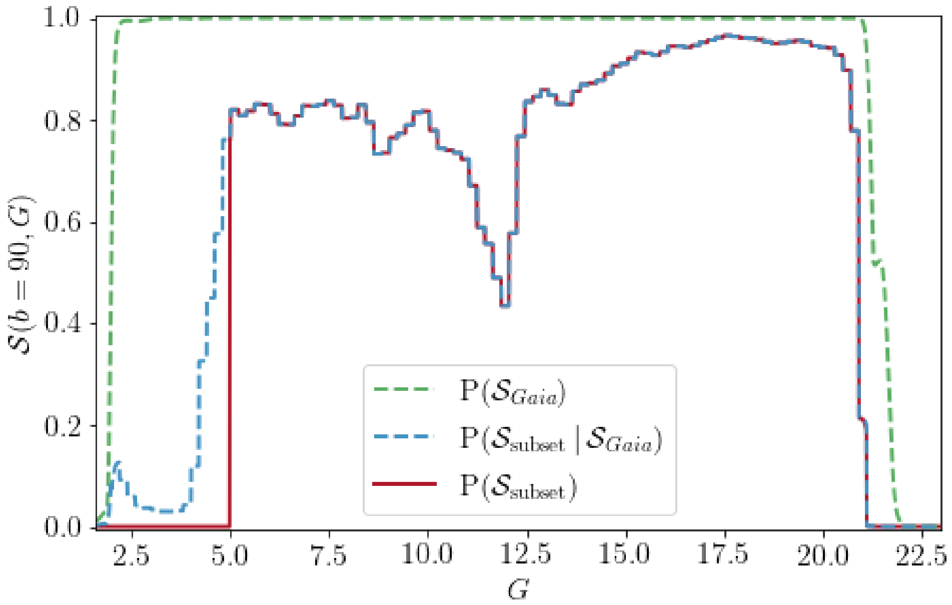


Fig. 6.4 The selection function probability at $b = 90°$ for the *Gaia* EDR3 source catalogue (green dashed) drops off at bright magnitudes ($G < 2$, due to CCD over-saturation) and faint magnitudes ($G \gtrsim 21$) however remains high across the rest of apparent magnitude space. The *Gaia* EDR3 astrometry with RUWE < 1.4 relative selection function (blue dashed) is more restrictive over the entire magnitude range and dominates the total selection function (red solid). The cut-off at $G < 5$ is deliberately imposed to remove regions of apparent magnitude with poor astrometry calibration.

unimodal, Section 6.5 demonstrates that this does not have a measurable affect on my results.

If, however, I include R-dependence and have L of order unity (kpc), then the integrand significantly changes. Eq. 6.32 now includes

$$\frac{\partial R}{\partial \varpi} = \frac{1}{R}\left(-\frac{\cos^2 b}{\varpi^3} + \frac{R_\odot \cos b \cos l}{\varpi^2}\right) \tag{6.36}$$

where R is given in Eq. 6.33. Expanding this out, Eq. 6.32 is now an 11th order polynomial in ϖ. Again, two of the stationary points are outside the integration bounds due to the logit transformation but that leaves 9 stationary points meaning up to 5 modes in the integrand.

One simplification I could take which would avoid adding any more modes to the integrand is

$$R \approx R_\odot - X = R_\odot - s \cos l \cos b. \tag{6.37}$$

This would provide a slight improvement on my previous models however also makes the model normalisation non-analytic which adds another layer of complexity. This may be an avenue worth pursuing however I consider it beyond the scope of this work.

6.3.4 Selection Function

A major obstacle to using a catalogue of sources to fit a distribution is the selection function. Many surveys have complex and unknown observation limitations which are a strong function of observatory properties and observing conditions. *Gaia* is no exception due in part to the complexity of the scanning law (Boubert et al., 2020, 2021b).

In most previous works, the sample is either assumed to be magnitude complete to some limit (e.g. Ak et al., 2008; Bilir et al., 2006a; Jurić et al., 2008), or the sample is bright and nearby for which there are larger, complete catalogues against which the selection function has been estimated (e.g. Bennett & Bovy, 2019; Bovy, 2017; Mateu & Vivas, 2018). *Gaia* is neither complete in position on the sky or apparent magnitude, nor is there a larger, more complete sample against which to compare the *Gaia* source catalogue.

Fortunately, a solution for the *Gaia* source catalogue selection function has been developed and applied to *Gaia* DR2 (Boubert & Everall, 2020). Section 4.1 provides a simple extension to model the selection function of the *Gaia* EDR3 source catalogue using the nominal EDR3 scanning law. This may have some limitations in crowded regions due to changes in *Gaia*'s data processing pipeline. However, since I am only considering high latitude fields, it should be sufficient for my purposes. The selection probability as a function of apparent magnitude for $b = 90°$ is given by the green dashed line in Fig. 6.4 showing that the source catalogue is nearly complete for $3 < G < 21$.

Given the source catalogue selection function, the selection functions of subsets can be estimated by comparison (Boubert & Everall, 2021). In Section 6.6, I introduce the *Gaia* astrometry catalogue with RUWE < 1.4 where apparent G-band magnitude is available. The selection function for this dataset is given by the product of the source catalogue and subset selection functions

$$\mathcal{S}_{\mathrm{subset}}(l, b, G) = \mathrm{P}(\mathcal{S}_{\mathrm{subset}} \,|\, \mathcal{S}_{\mathrm{source}}, l, b, G)\, \mathrm{P}(\mathcal{S}_{\mathrm{source}} \,|\, l, b, G) \tag{6.38}$$

where $\mathrm{P}(\mathcal{S}_{\mathrm{source}} \,|\, l, b, G)$ is the probability of selection in the *Gaia* source catalogue with published G and $\mathrm{P}(\mathcal{S}_{\mathrm{subset}} \,|\, \mathcal{S}_{\mathrm{source}}, l, b, G)$ is the probability of an object in the source catalogue having published parallax with RUWE < 1.4, modelled in Chapter 4, as a function of G and position on the sky only. When fitting the model parameters to data, Eq. 6.38 is substituted into Eq. 6.9.

The results are applied in 0.2 mag bins in G in NSIDE = 64 HEALPix pixels (Górski et al., 2005) across the sky. The selection probability for $b = 90°$ is given by the red line in Fig. 6.4. Due to the challenges of modelling sources which saturate the *Gaia* CCDs at the bright end of the magnitude distribution I use a selection function which truncates at $G = 5$. My sample also only includes those sources with $G > 5$.

6.4 Mock

To test and demonstrate the efficacy of the method, I generate a mock catalogue from my model with realistic parameters. Information on the true parameters is then removed, the

Component	Parameter	Input	Full	SF	SF & σ_ϖ
Thin disc	$\log_{10}(w)$	4.0792	$4.0700^{+0.0194}_{-0.0191}$	$4.0637^{+0.0331}_{-0.0359}$	$3.9816^{+0.0586}_{-0.0657}$
	$h_{\rm Tn}$	0.300	$0.301^{+0.007}_{-0.006}$	$0.301^{+0.010}_{-0.010}$	$0.281^{+0.015}_{-0.015}$
	f_G	4.50×10^{-3}	$3.73^{+1.00}_{-0.99}\times10^{-3}$	$3.91^{+1.23}_{-1.22}\times10^{-3}$	$3.76^{+1.43}_{-1.30}\times10^{-3}$
	$M_{\rm TO}$	3.1			
	α_3	-0.6			
Thick disc	$\log_{10}(w)$	4.6335	$4.6249^{+0.0051}_{-0.0050}$	$4.6253^{+0.0092}_{-0.0093}$	$4.6221^{+0.0200}_{-0.0198}$
	$h_{\rm Tk}$	0.900	$0.891^{+0.012}_{-0.011}$	$0.884^{+0.029}_{-0.028}$	$0.812^{+0.052}_{-0.045}$
	f_G	5.40×10^{-3}	$5.76^{+0.54}_{-0.52}\times10^{-3}$	$5.80^{+0.64}_{-0.60}\times10^{-3}$	$5.83^{+0.69}_{-0.66}\times10^{-3}$
	$M_{\rm TO}$	3.1			
	α_3	-0.77			
Halo	$\log_{10}(w)$	5.9754	$5.9759^{+0.0005}_{-0.0005}$	$5.9662^{+0.0106}_{-0.0105}$	$5.9450^{+0.0247}_{-0.0229}$
	$n_{\rm H}$	3.740	$3.745^{+0.001}_{-0.001}$	$3.753^{+0.020}_{-0.020}$	$3.812^{+0.068}_{-0.066}$
	f_G	3.50×10^{-3}	$3.47^{+0.06}_{-0.06}\times10^{-3}$	$3.49^{+0.10}_{-0.09}\times10^{-3}$	$3.48^{+0.15}_{-0.15}\times10^{-3}$
	$M_{\rm TO}$	3.1			
	α_3	-0.64			
Shared	α_1	-0.1100	$-0.1109^{+0.0003}_{-0.0004}$	$-0.1094^{+0.0014}_{-0.0015}$	$-0.1098^{+0.0020}_{-0.0020}$
	α_2	-0.2500	$-0.2524^{+0.0020}_{-0.0019}$	$-0.2534^{+0.0045}_{-0.0046}$	$-0.2521^{+0.0089}_{-0.0084}$

Table 6.1 The input parameters for the mock sample catalogue generation and the results of the fit to the data are shown when using the full sample with no observational errors (”Full”), the selection function with no observational errors (”SF”) and the sample with both the selection function and the added parallax errors (”SF & σ_ϖ”). For all parameters I provide the median and 16th and 84th percentile uncertainties.

Gaia selection function and *Gaia*-like parallax uncertainties are applied, and I attempt to infer the input parameters from the mock sample. I note that this only tests the method. Because the data is drawn from the same model which is being refit, any inconsistencies between the model and true Milky Way distribution of stars do not show up here. These inconsistencies are discussed, tested and quantified in Section 6.8.

6.4.1 Input Parameters

Parameters for the scale heights and power law indices of the discs and halo respectively are taken from the literature. For the thin disc $h_{\mathrm{Tn}} = 300\mathrm{pc}$ and thick disc $h_{\mathrm{Tk}} = 900\mathrm{pc}$ (Jurić et al., 2008). The power law index used is $n_{\mathrm{H}} = 3.74$ from Fukushima et al. 2019.

The relative stellar mass density of the discs is $\rho_{\mathrm{Tk}}/\rho_{\mathrm{Tn}} = 0.12$ and $\rho_{\mathrm{H}}/\rho_{\mathrm{Tn}} = 0.005$ (Jurić et al., 2008). Instead of local mass density, my model fits the total number of sources in each component with $|b| > 80°$. To convert mass density to number density in the Solar neighbourhood, I divide by the mean mass of a star. The mean mass is estimated using the IMF-isochrone sample in Section 6.3.2 as $\mathcal{M} \sim 0.413, 0.369, 0.308\,\mathrm{M}_{\odot}$ for the thin disc, thick disc and halo respectively. I then divide the number density by the value of the normalised component at $s = 0$ to get the total number of sources in each component. The result is that $w_{\mathrm{Tn}}/w_{\mathrm{Tk}} = 0.275$ and $w_{\mathrm{Tn}}/w_{\mathrm{H}} = 0.0127$. The halo dominates the total counts because my observing volume is a cone with $|b| > 80°$. This significantly reduces the relative contribution from the disc to the sample.

The absolute magnitude distributions for each Milky Way component are shown by the shaded histograms in Fig 6.3. To estimate magnitude parameters for the luminosity function described in Section 6.3.2, I directly fit the parameters to the magnitude distributions. For each component, the turn-off magnitude is at $M_G \sim 3.1$. f_G is approximated from the ratio of sources with $M_G < 3.1$ to those with $M_G > 3.1$. For $M_G < 3.1$ I fit a power law profile to each component independently using the Poisson likelihood function from Eq. 6.1. This gives $\alpha_3 = -0.60, -0.77, -0.64$ and $f_G = 0.0045, 0.0054, 0.0035$ for the thin disc, thick disc and halo respectively.

The lower main sequence is dominated by old, long-lived stars which evolve slowly on the HR diagram. Therefore, I assume that the main sequence profiles are similar between different Milky Way components such that the values of α_1, α_2 are shared between profiles. I draw a sample of sources from each of the components according to the component's respective weight and fit the main sequence profiles to the sources with $M_G > 3.1$, which gives $\alpha_1 = -0.12$, $\alpha_2 = -0.26$. The dashed lines in Fig. 6.3 give the absolute magnitude distributions implied by the parameters I have just derived.

All selected and evaluated parameter values are listed as 'Input' in Table 6.1.

6.4.2 Parallax Error

To generate a realistic mock, I also need to sample measurement uncertainties. Since the *Gaia* astrometry was fit using an iterative linear regression process, the covariance may be estimated from information theory (neglecting excess noise) using only the scanning law and individual observation centroid uncertainties. This process is performed in Chapter 5 for *Gaia* DR2 and I use the *Gaia* EDR3 nominal scanning law to extend this to the EDR3 baseline.

The covariance estimates break down for sources with significant excess noise, such as in heavily crowded regions and for sources with intrinsic astrometric variability like binaries. Since I only consider sources with $|b| > 80°$, crowding is negligible. By focusing on the sample with RUWE < 1.4, I expect to have removed sources with observable binary motion.

6.4.3 Mock Samples

A sample of one million sources with distance, latitude and absolute magnitude is drawn from the model using MCMC sampling (Foreman-Mackey et al., 2013). Since all sources are assumed to be at the Solar Galactocentric cylindrical radius, the full model is Galactic longitude-independent so the longitude is drawn from a uniform distribution $l \sim \mathrm{U}[0, 2\pi]$. The distribution of drawn sources as a function of distance from the Galactic disc and absolute magnitude is given by the blue histograms in the top panels of Fig. 6.5.

The selection function probability is evaluated for all sources given their position on the sky and apparent magnitude as described in Section 6.3.4. To generate the mock *Gaia* astrometry with RUWE < 1.4 sample, the event of a source being included is drawn from a Bernoulli distribution with the given selection probability $\mathrm{S}_i \sim \mathrm{Bernoulli}(\mathcal{S}(l_i, b_i, G_i))$ where $\mathrm{S}_i = 0, 1$. Of the 1 000 000 source in the full sample, 73 132 survive the selection cuts, shown by the red histograms in the middle and bottom panels of Fig. 6.5.

Parallax error is evaluated from the Astrometric Spread Function described in Section 5. The observed parallax is drawn from a Gaussian distribution with the given error for each source $\varpi \sim \mathcal{N}(1/s, \sigma_\varpi)$. The red histograms in the bottom panels of Fig. 6.5 show the distribution of measured $z = \sin(b)/\varpi$, $M_G = G - 10 + 5\log_{10}(\varpi/\mathrm{mas})$ after sampling ϖ from the parallax error. This significantly affects the distributions, demonstrating the importance of properly accounting for parallax uncertainty when modelling the structure of the Milky Way from *Gaia* data.

This produces three samples which can each be used to independently fit the model parameters demonstrating each stage of the method:

(i) Full sample fit with Eq. 6.1: $l^i, b^i, si, G^i \,\forall\, i$,

(ii) SF sample fit with Eq. 6.3: $l^i, b^i, s^i, G^i \,\forall\, i$ where $S_i = 1$,

(iii) SF & σ_ϖ fit with Eq. 6.9: $l^i, b^i, \varpi^i, G^i \,\forall\, i$ where $S_i = 1$.

To be clear, in sample (iii) the selection function is not dependent on measured parallax or parallax error as discussed in Section 6.2. I simply mean that the selection function is applied and parallax error on sources is also included. Samples (ii) and (iii) contain the exact same subset of sources from the mock catalogue. Sample (ii) has no parallax error, whilst measured parallaxes in (iii) have been drawn from the parallax uncertainties.

6.5 Parameter Inference

In this section I use the method introduced in Section 6.2 to fit the model parameters to the three mock samples described in Section 6.4.

6.5.1 Priors

Priors for all free parameters of the fits are given in Table 6.2. As is common with mixture model fits to density distributions, the likelihood space is strongly multi-modal. For the thin and thick discs there is of course a complete degeneracy where the components can be switched, but there are also problematic modes where, for example, a single component is expanded to fit the full data-set whilst remaining components are suppressed.

Priors are chosen specifically to avoid local optima in the model. All weights are assumed to be drawn from a Dirichlet distribution with $a = 2$ to remove modes where any component is completely suppressed relative to the others. To avoid the disc degeneracy, the possible disc scale heights are limited to non-overlapping ranges with $h_{\mathrm{Tn}} \sim \mathrm{U}[0.1\mathrm{kpc}, 0.6\mathrm{kpc}]$ and $h_{\mathrm{Tk}} \sim \mathrm{U}[0.6\mathrm{kpc}, 3.0\mathrm{kpc}]$. The power-law index of the halo is also limited to $n_{\mathrm{H}} \sim \mathrm{U}[3.0, 7.3]$ as $n_H < 3.0$ would produce an unnormalised halo and $n_H > 7.3$ produces an incredibly steep halo profile which can mimic the exponential discs (for $n_H = 7.3$ the mean halo source distance is the same as an exponential profile with $h = 3.0$ kpc).

For numerical stability, the fits are made on the transformed parameters where transformations are given in Table 6.2. The transformations scale parameters to the range $[-\infty, \infty]$ in all cases. For logit transformed parameters, I include a logistic prior in logit space which is equivalent to a uniform prior in untransformed space. Therefore the logit transformation has no effect on the prior.

The L-BFGS-B algorithm requires boundaries on all parameters which are given in the final column of Table 6.2. The boundaries are chosen to avoid regions of parameter space which suffer from numerical precision issues. None of the parameter posterior distributions push up against the boundaries.

6.5.2 Optimization

The likelihood optimization is performed in three stages. All MCMC processes used EMCEE (Foreman-Mackey et al., 2013). First, a set of samples is drawn from the parameter priors using MCMC with 44 walkers (this is 4× the number of free parameters in my model), with 100 step burn-in and 100 steps of sampling. Secondly, ten samples are randomly selected from the prior samples as initialisation for gradient descent using L-BFGS-B (Zhu et al., 1997) as implemented in SCIPY. Finally, the maximum likelihood estimate with the highest likelihood is taken as the best fit solution. A secondary MCMC process is initialised with 44 walkers drawn from a Gaussian ball around the maximum likelihood estimate with variance of 10^{-10} times the boundary width. These walkers were run with the likelihood × prior for 5000 steps. The latter 2 500 steps are used at 5 step intervals as the posterior samples. This process is used for fitting all mock samples and the real *Gaia* data in the rest of thus chapter.

Component	Parameter	Prior	Transformation	Bounds
Thin disc	w	Dirichlet$(a = 2)$	$\log(w)$	[-10,50]
	h_{Tn}	$\mathrm{U}[0.1, 0.6]$	$\mathrm{logit}\left(\frac{h-0.1}{0.6-0.1}\right)$	[-10,10]
	f_D	$\mathrm{U}[0, 1]$	$\mathrm{logit}(f_D)$	[-10,10]
Thick disc	w	Dirichlet$(a = 2)$	$\log(w)$	[-10,50]
	h_{Tk}	$\mathrm{U}[0.6, 3.0]$	$\mathrm{logit}\left(\frac{h-0.6}{3.0-0.6}\right)$	[-10,10]
	f_D	$\mathrm{U}[0, 1]$	$\mathrm{logit}(f_D)$	[-10,10]
Halo	w	Dirichlet$(a = 2)$	$\log(w)$	[-10,50]
	n_{H}	$\mathrm{U}[3, 7.3]$	$\mathrm{logit}\left(\frac{h-3}{7.3-3}\right)$	[-10,10]
	f_D	$\mathrm{U}[0, 1]$	$\mathrm{logit}(f_D)$	[-10,10]
Shared	α_1	$-\alpha_1 \sim \log \mathrm{U}[\mathrm{e}^{-5}, \mathrm{e}^{3}]$	$\log(-\alpha_1)$	[-5,3]
	α_2	$-\alpha_2 \sim \log \mathrm{U}[\mathrm{e}^{-5}, \mathrm{e}^{3}]$	$\log(-\alpha_2)$	[-5,3]

Table 6.2 The 11 free parameters used to model the spatial and absolute magnitude distributions of sources along with their priors. The method fits directly to the parameters under the given transformations where logistic priors are also included to correct for the logit transform. The bounds are applied to the transformed parameters for numerical stability of the optimization.

6.5.3 Results

The 'Full' sample posteriors, given by the blue contours in Fig. 6.6, provide tight solutions around the input parameter values which are shown by the black dot. A more quantitative comparison can be made from Table 6.1 which shows that the majority of input parameters fall within the $16-84^{\mathrm{th}}$ percentile range of the posterior distribution. The top panels of Fig. 6.5 compare the ground truth input model, shown with dotted lines, to the refit model, shown by the narrow shaded regions. To produce the shaded posteriors in Fig. 6.5 I draw 100 samples from the posterior parameter distributions and plot the $16-84^{\mathrm{th}}$ percentile range as a function of z, M_G and G. The posteriors are so tight in most cases that the shaded regions appear as lines perfectly tracking the input model and the total of the components in black sits exactly on top of the blue histograms which show the distribution of the data in the sample.

The 'SF' sample, fit to only 73 132 of the initial one million mock sources, has a significantly less tight constraint around the true parameters, shown by the red contours in Fig 6.6, but the parameters show no significant bias. The fits to the halo parameters are slightly shifted from the true values but all parameters are well within 2σ of the input so this can be well explained by correlated noise, particularly considering the negative correlation between the halo weight and power-law index, n_{H}. The red histograms in the middle panels of Fig. 6.5 show the selection-limited sample which drops significantly at

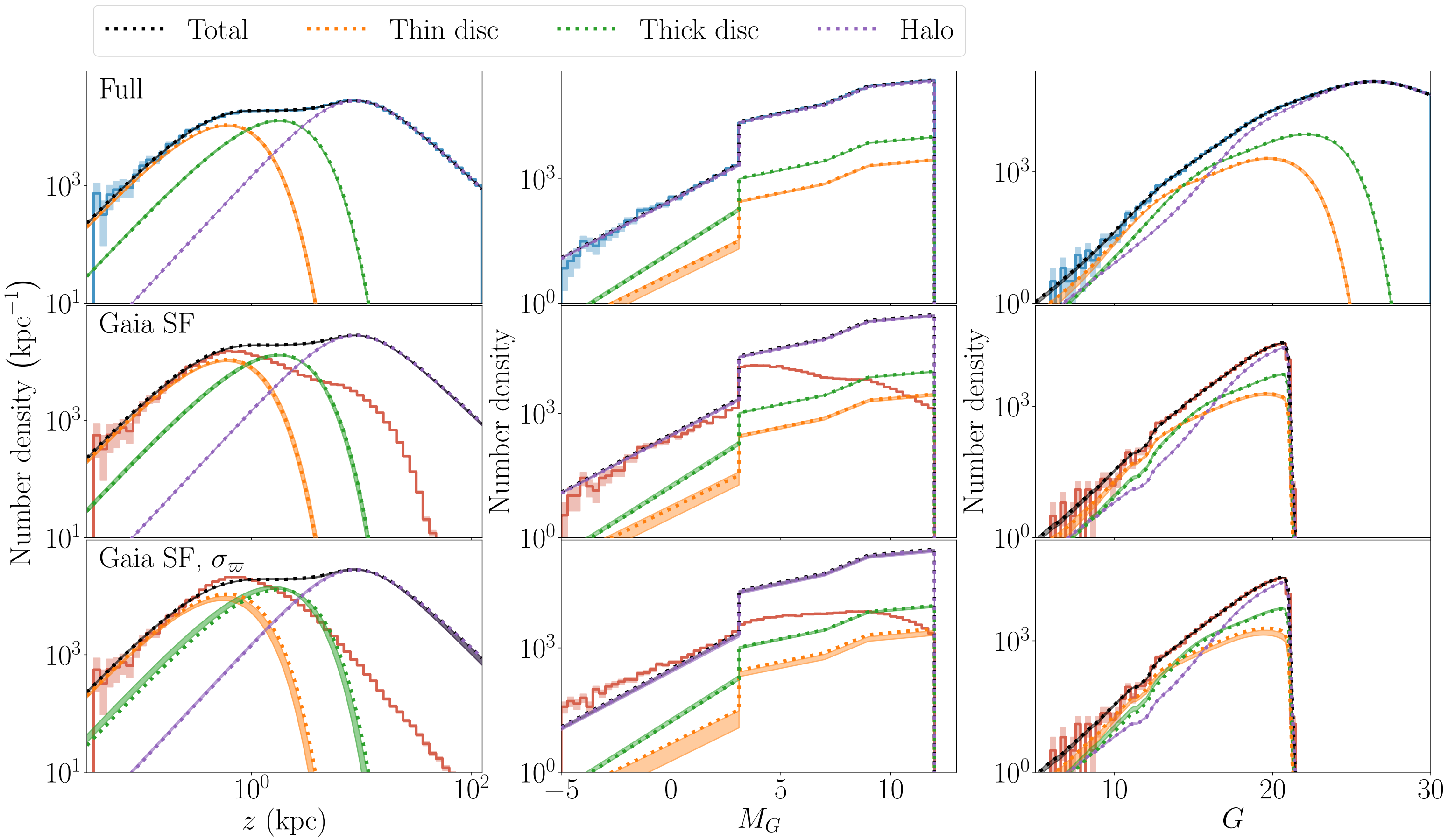


Fig. 6.5 The posterior distribution of fits to the mock sample are shown by the shaded regions for the thin disc (orange), thick disc (green), halo (purple) and the sum total (black) as a function of vertical height (z, left), absolute magnitude (M_G, middle) and apparent magnitude (G, right). Blue histograms in the top row show the full sample which the model fit perfectly cut through. The red histograms in the middle and bottom rows show the distribution of selection function selected samples with the bottom row showing the distribution of $z = \sin(b)/\varpi$ and $M_G = G + 5\log_{10}(\varpi) - 10$, demonstrating the impact of parallax uncertainties on measured quantities. The posteriors agree extremely well with the ground truth shown by the dotted lines in all panels. This is true when fitting to the full sample (top), the selection function-limited sample (middle) and the selection function-limited sample with measured parallaxes sampled from their error distributions (bottom). The posterior distributions are evaluated by randomly selecting 100 samples from the MCMC posteriors and taking the 16th – 84th percentile range. In several cases, particularly for the “Full” fits in the top row, the posterior is so tight that the distribution appears as a line in the figure.

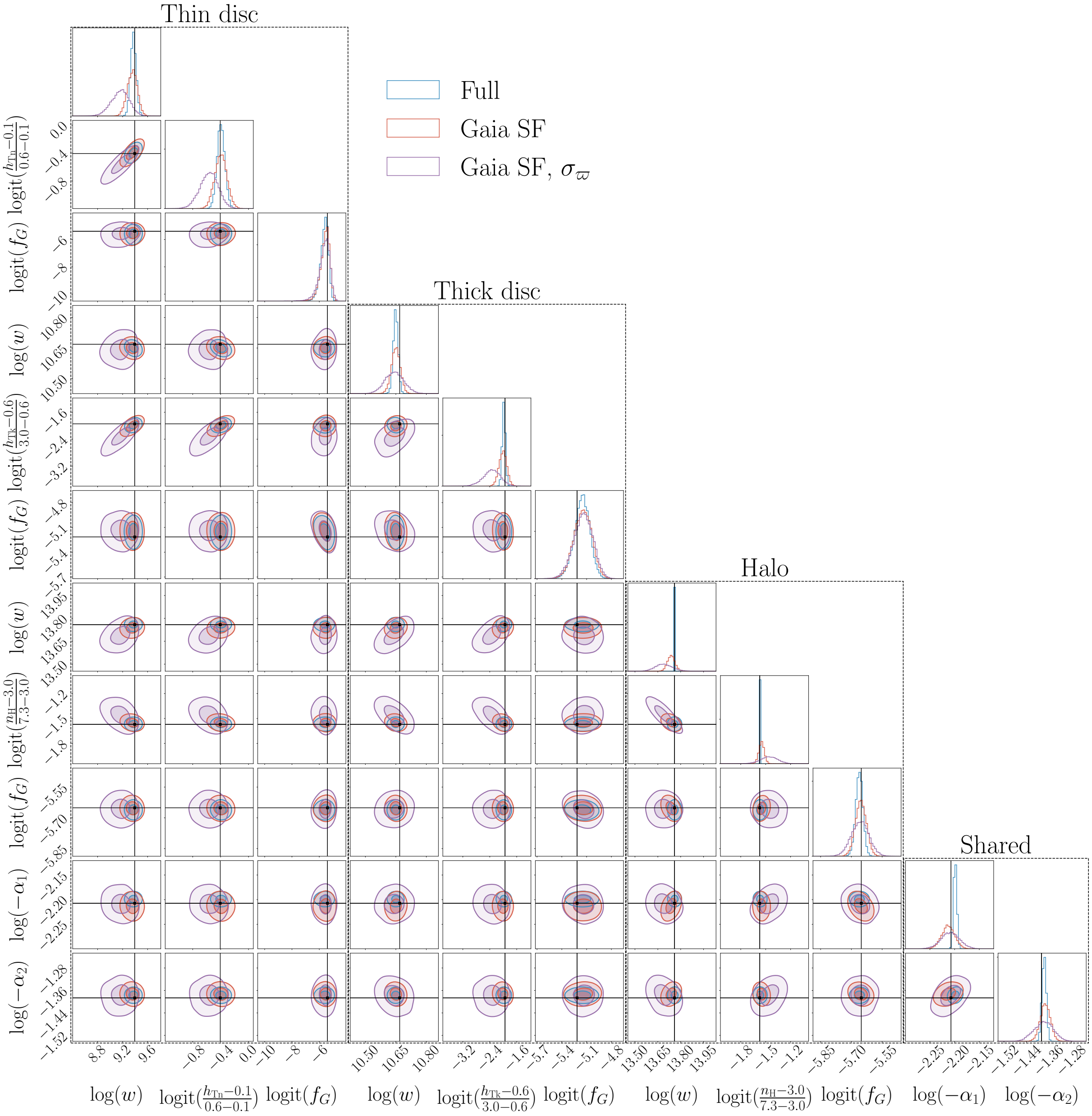


Fig. 6.6 The posterior distributions for all mock samples are shown as a function of transformed parameters which are fit to the data. The Full sample fits (blue), selection function sample (red) and selection function with parallax error (purple) all show strong agreement with one another and the input parameters (black lines). The enhancement of the statistical uncertainty by introducing parallax error can clearly be seen by the increased spread of the posterior for the purple contours.

large vertical heights and faint apparent magnitudes demonstrating how much the model has to extrapolate using the selection function. Again, the model posteriors sit perfectly on the input model shown by the dotted lines.

For the apparent magnitude distribution in the middle right panel of Fig. 6.5 I show the model multiplied by the selection function probability. The total model (black) sits perfectly on top of the red sample histograms demonstrating how successfully the model is fit to the data. This distribution will be especially important when analysing fits to the real *Gaia* data when I cannot directly infer the distance of stars from the Galactic plane or their absolute magnitudes due to significant parallax uncertainties.

The 'SF & σ_ϖ' posterior, given by the purple contours in Fig. 6.6, has significantly enhanced uncertainty compared with the solely selection function limited data. This demonstrates how much information is held in the parallax and how information is lost when realistic *Gaia* parallax uncertainties are included. In spite of this, the input parameters are still recovered with reasonable precision and good accuracy. In the bottom panels of Fig. 6.5, the posterior samples produce a clearer spread around the input distribution. This time the thin and thick discs have not been perfectly fit within the posteriors however the difference is still small enough to be well explained by statistical noise.

These results demonstrate that the Poisson-likelihood method accounting for the *Gaia* selection function and parallax error is a powerful tool for recovering the spatial distribution of sources in the Milky Way. However this only tests the self-consistency of the method; the results may still be susceptible to systematic uncertainties if the model does not represent the real Milky Way. These systematics are tested and quantified in Section 6.8.

6.6 Data

My initial sample of *Gaia* sources consists of all objects in EDR3 with $|b| > 80°$, published parallax with RUWE < 1.4 and published G-band apparent magnitude with $G > 5$. Brighter sources saturate the *Gaia* CCDs which significantly affects the reliability of astrometric solutions. My sample is extracted with the following query which returns 673 926 sources in the Galactic north and 702 599 in the south.

```
select ra, dec, parallax, parallax_error, phot_g_mean_mag
from gaiaedr3.gaia_source
where (b<-80 or b>80)
and parallax is not NULL
and phot_g_mean_mag>5
and ruwe<1.4
```

A recurring challenge with *Gaia* astrometry is the zero-point parallax offset, which leads to a small bias for any individual source but can significantly bias models fit to an entire population (e.g. see Everall et al., 2019). I apply the zero-point correction recommended in Lindegren et al. 2021b for sources with 5 and 6 parameter astrometric solutions. Many other groups have attempted to measure the zero point parallax offset from Cepheid variables (Riess et al., 2021), Red Clump stars (Huang et al., 2021), eclipsing binaries

(Ren et al., 2021; Stassun & Torres, 2021) and quasars (Groenewegen, 2021) (although the Lindegren et al., 2021b model was constructed using quasars so it is unsurprising that these results match well). The conclusions are that for the majority of sources, the parallax offset is reduced to under $10\,\mu$as. Zinn 2021 and Riess et al. 2021 find the parallaxes of sources brighter than $G = 10.8$ are overestimated by $\sim 15\,\mu$as after the correction, so I adjust the offset for the small portion of my sample with $G < 10.8$. I test and discuss the effect of any residual offset in Section 6.8.

Parallax errors in *Gaia* are found to be typically underestimated when considering globular clusters (Vasiliev & Baumgardt, 2021) and wide binaries (El-Badry et al., 2021). I use the model from Equation (16) of El-Badry et al. 2021 to revise the parallax errors of my *Gaia* sample, as this is appropriate for uncrowded fields which broadly applies to my sample.

The *Gaia* G-band apparent magnitude also has some small systematic bias for sources with 6-parameter astrometric solutions. I apply the apparent magnitude correction recommended in Riello et al. 2021 to the sources where $G_{\rm BP} - G_{\rm RP}$ colour is available. One issue this raises is that the G-band apparent magnitude used for the data is subtly different from the measurements used to derive the *Gaia* selection function. However, the magnitude correction is at most -0.025 mag which is much smaller than my 0.2 mag resolution of the selection function. Therefore this inconsistency will have a negligible effect on the results.

As I am only using objects at high Galactic latitude, there is likely to be a sizable contamination from extragalactic sources (both quasars and distant galaxies). If left in the sample, these would bias the inferred distribution of stars towards larger distances.

Classifiers have been constructed to determine the probability of a source being extragalactic based on *Gaia* astrometry and photometry complemented with other surveys (Bailer-Jones et al., 2019; Shu et al., 2019). The issue is that these classifications are not 100% pure and will likely remove dim stars with low parallaxes which are misclassified as extragalactic. This is particularly clear in Fig. 10 of Bailer-Jones et al. 2019 where the 'quasar' population is dominated by the LMC, SMC and particular scans. The most prominent scans are the same as those found in Section 4.1.2 which were caused by missing calibration data in the *Gaia* photometric processing pipeline. To avoid introducing a bias to my data when removing extragalactic sources I avoid selecting on apparent magnitude and astrometry.

Galaxies have an extended flux distribution on the sky. Due to the larger window size used to measure BP and RP on-board *Gaia*, galaxies will typically produce an excess flux in these bands over the G-band (see Fig. 21 Riello et al., 2021). The flux ratio between the combined BP and RP measurements and the G-band is published as PHOT_BP_RP_EXCESS_FACTOR in the *Gaia* archive (Evans et al., 2018). The published excess flux has some residual colour-dependence which needs correcting. I use the formula provided in Section 6 of Riello et al. 2021 to estimate the corrected flux excess C^*. Galaxies are selected as sources with $C^* > 1.8$. The distribution of sources in excess flux vs

$G_{\mathrm{BP}} - G_{\mathrm{RP}}$ is shown in the left panel of Fig. 6.7 with the red dashed line showing the Galaxy cut.

Quasars are well distinguished using the WISE photometry's $W_1 - W_2$ colour (e.g., Shu et al., 2019). I crossmatch my sample with the unWISE sample which has improved resolution over the original WISE catalogue (Lang, 2014). Taking the nearest object within 2 arcseconds correcting for proper motions with the *Gaia* epoch set to 2016 and unWISE to 2010 produces a successful match for 88% of sources in my sample. Quasars are removed from my sample using the colour-colour cut

$$W_1 - W_2 > 0.5 \quad \& \quad G_{\mathrm{BP}} - G_{\mathrm{RP}} < 0.7(W_1 - W_2) \tag{6.39}$$

which is shown by the blue dashed line in the middle panel of Fig. 6.7.

These cuts select 2 933 galaxies and 50 726 quasars with 553 sources classified as both a galaxy and quasar. However, this does not tell us how successful my selection has been. For this, I crossmatch with spectroscopically classified sources in SDSS-IV (Blanton et al., 2017). I again use a proper motion corrected crossmatch for sources within 2 arcseconds with the SDSS epoch set at 2000. In this case, only 1.8% of my sample receive SDSS spectra, the vast majority of which are in the northern field. The objects classified as galaxies and quasars by SDSS are shown as the red and blue points respectively in the left and middle panels of Fig. 6.7.

Of those with successful crossmatches, 8 900 are classified as galaxies or quasars by SDSS whilst my cuts select 8 275 sources, of which 8 114 are classified as extragalactic by both. This implies that my selection criteria correctly classifies 91.2% of extragalactic sources with only 1.7% of Milky Way sources incorrectly classified as extragalactic. The remaining 8.8% of missing sources account for $\sim 0.3\%$ of my final sample, so I consider this completeness to be sufficient.

Extragalactic sources are far too distant for *Gaia* parallax measurements therefore the measured parallax signal to noise will be distributed as $\varpi/\sigma_{\varpi} \sim \mathcal{N}(0, 1)$. I show this distribution in the right hand panel of Fig. 6.7 for galaxies (red), quasars (blue) and the remainder of the sample (purple). The extragalactic sources are close to normally distributed. The galaxy sample has a small amount of stellar contamination which marginally enhances the $+\varpi$ wing, but overall this shows that my classification has performed well.

The number density of sources in pixels around the north and south Galactic poles is shown in the top panels of Fig. 6.8. For the most part, the distribution is reasonably smooth and noise dominated which is good when fitting a smooth model. However, the south field has two significant overdensities. The overdensity close to the south Galactic pole is the globular cluster NGC 288 which sits at a distance of approximately 9 kpc from the Sun with a scale radius of ~ 3 arcminutes (Vasiliev & Baumgardt, 2021). The other overdensity at slightly higher latitudes east of the Galactic Centre direction is the Sculptor dwarf spheroidal at $l = 288°, b = -83°$ with a half-light radius of ~ 11.3 arcminutes (McConnachie, 2012).

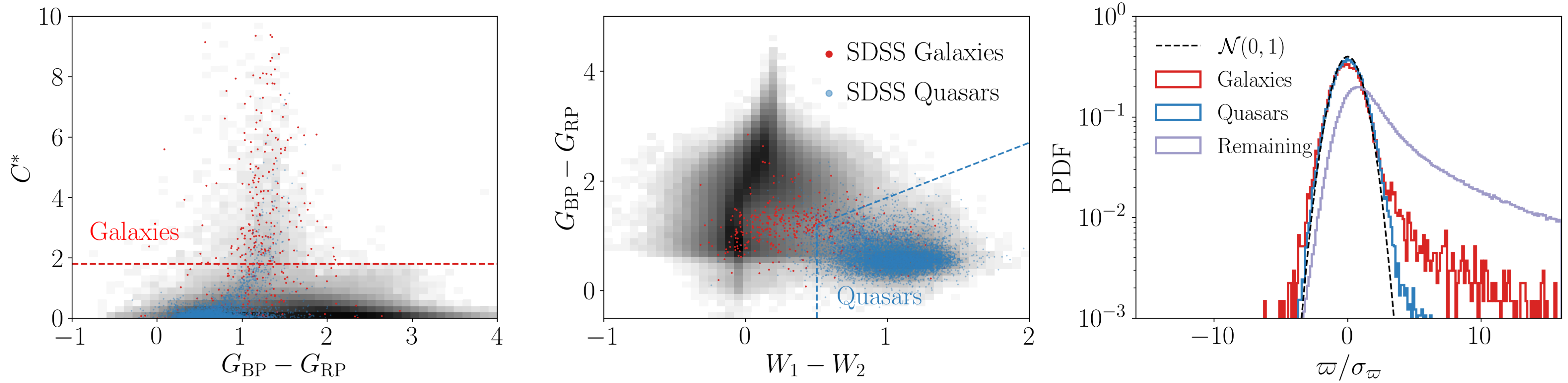


Fig. 6.7 Cuts on *Gaia* and unWISE colour photometry are used to remove extragalactic sources from my sample. ***Left***: Galaxies are removed using a cut on PHOT_BP_RP_EXCESS_FACTOR after correcting for colour dependence, $C^* < 1.8$. Red points show the SDSS spectroscopically classified galaxies which clearly extend to high excess flux levels. ***Middle***: Quasars are removed with colour-colour cuts on *Gaia* $G_{\rm BP} - G_{\rm RP}$ and unWISE $W_1 - W_2$ shown by the blue dashed lines. The SDSS quasars (blue points) are clearly clustered in the region of colour-colour space beyond these cuts. ***Right***: The parallax SNR distribution of the galaxy and quasar samples are nearly Gaussian unit-variance distributed with a small enhancement at high ϖ due to a small number of stars which are incorrectly removed from the sample.

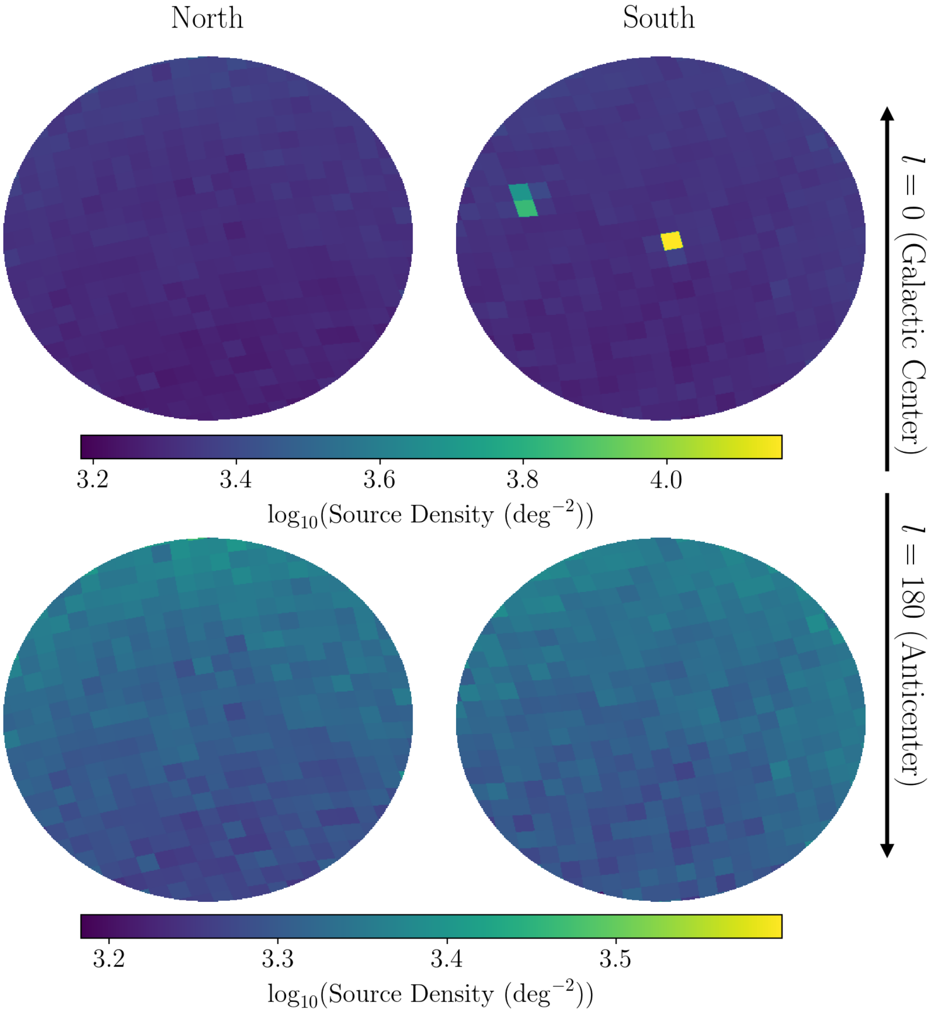


Fig. 6.8 ***Top***: The number density of sources in HEALPix pixels across the north (left) and south (right) regions of the sky with $|b| > 80°$ is mostly uniform. The two clear exceptions are NGC 288 at the south Galactic pole and the Sculptor dwarf spheroidal at $(l, b) = (288°, -83°)$ both appearing in the upper right panel. ***Bottom***: After masking these contributions I am left with the bottom panels which are almost uniform with a slight number density gradient from towards the Galactic centre at the top to the anticentre at the bottom.

To prevent these objects from contaminating my smooth models, I mask the regions of the sky occupied by the structure out to four scale radii. I then renormalise the pixels by the fraction of the area which remains unmasked. This is the same treatment that I apply to pixels sitting on the edge of the 10° radius fields. The resulting source density after masking NGC 288 and Sculptor is given by the bottom panels in Fig. 6.8, showing no further significant residual substructure. The gradient of the source density from the Galactic Centre (top of the figure) to the outer galaxy can now be seen. This shows the cylindrical radius dependence of the Milky Way distribution of stars which is not factored into my model, but I discuss its impact in Section 6.8.

The absolute magnitude model defined in Section 6.3 is limited by $M_G < 12$ in order to avoid use of uncertain stellar evolutionary models. The *Gaia* sample may still contain sources dimmer than this limit, which are nonetheless near enough that *Gaia* is able to detect them. The issue is that I cannot directly measure absolute magnitude and parallax error is large enough for many sources that they will be scattered to that region of absolute magnitude space independent of their true brightness. My compromise is to cut out sources which are likely to be fainter than $M_G = 12$ by one sigma uncertainty in parallax. In other words, removing all sources with greater than 84% likelihood of $M_G > 12$. This means only keeping sources with

$$\varpi - \sigma_\varpi < 10^{\frac{22-G}{5}}. \tag{6.40}$$

The effect of this cut is shown in Fig. 6.9. The left panel shows the naive absolute magnitude distribution calculated with $s = 1/\varpi$. The cut removes a large fraction of objects which fall outside the boundary. Importantly, from the middle and right panels, all of the sources removed from the sample are measured with $1/\varpi$ within 400pc of the Sun with a parallax signal-to-noise ratio (SNR) greater than 1.4. Any error in this cut will introduce a dependence of the selection function on measured parallax and parallax error. However, given the high parallax SNR of the removed sources, I expect that this dependence should be negligibly small. An added benefit of the cut I have placed here is that it will likely remove sources with poor astrometric solutions as classified by Rybizki et al. 2021a and Gaia Collaboration et al. 2021b which are typically fainter than $M_G = 12$. This cut removes a further 13 792 and 13 731 sources from the north and south fields respectively.

After all of the cleaning, I am left with 633 289 north and 640 072 south sources in my sample. I emphasise that, through all of these cuts, I remove less than 11% of the sample with published parallax, G apparent magnitude and RUWE < 1.4. By comparison, a cut on $\varpi > 0$ *alone* (which is a serious crime, according to Luri et al. 2018) removes over 15% and a signal-to-noise cut of $\varpi/\sigma_\varpi > 4$ removes over 61%. I am modelling the vast majority of *Gaia* sources using the reliable astrometric and photometric data that is available.

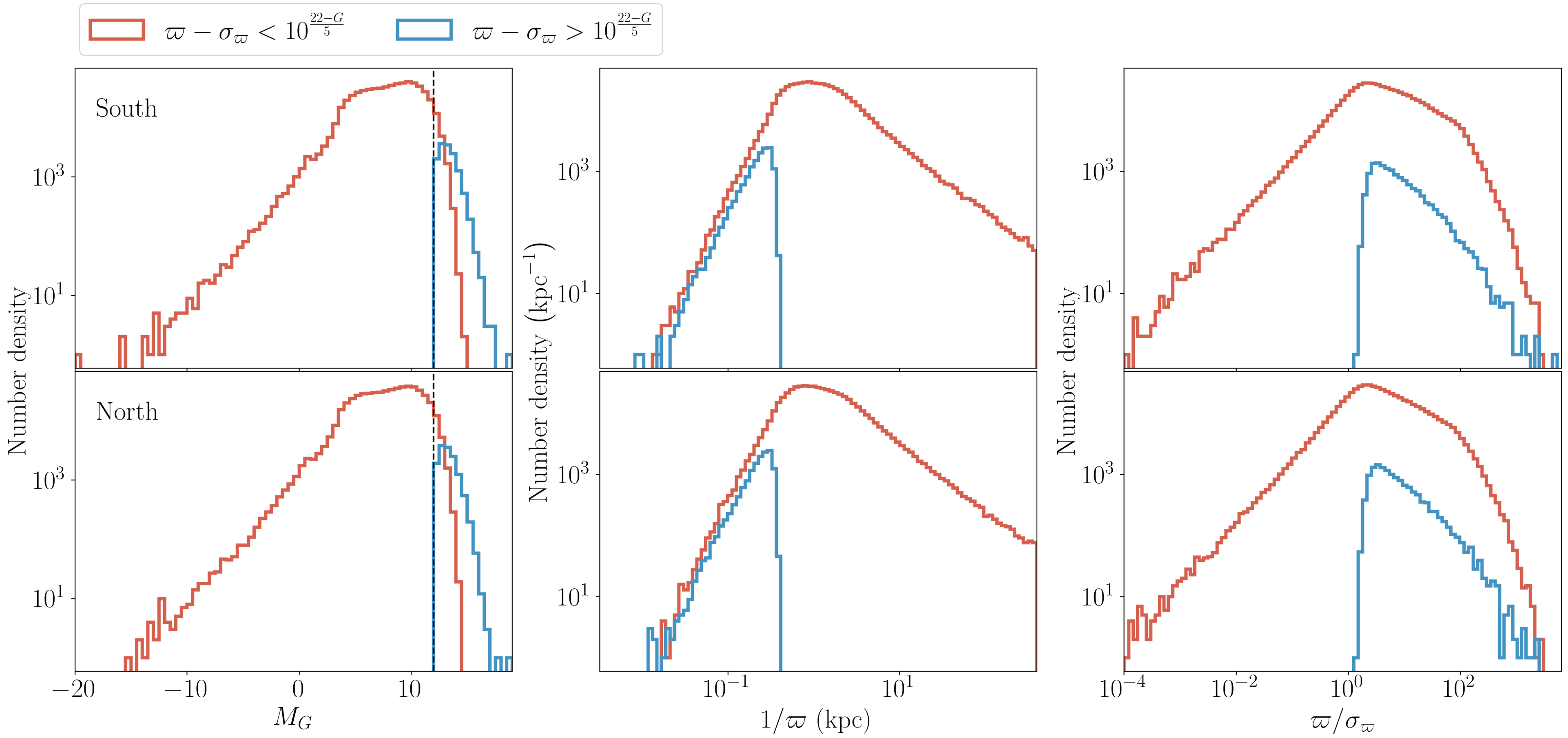
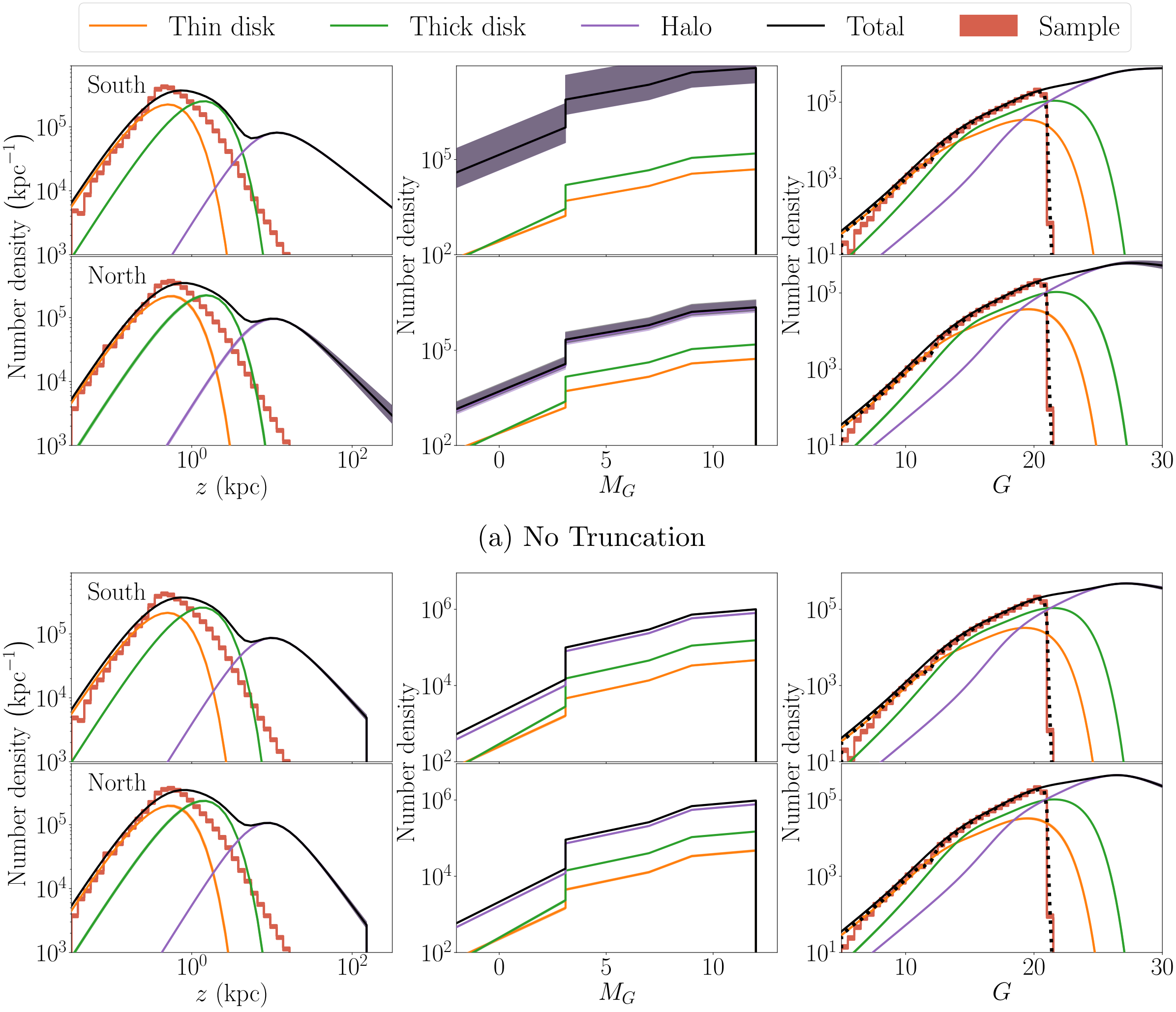


Fig. 6.9 The effect of removing sources with $\varpi - \sigma_\varpi > 10^{\frac{22-G}{5}}$ is shown as a function of absolute magnitude ($M_G = G + 5\log_{10}(\varpi) - 10$, left), inverse parallax (middle) and parallax signal-to-noise (right) for samples in the south (top) and north (bottom) regions. The cut conservatively removes sources which are likely to be intrinsically dimmer than the maximum absolute magnitude of the model ($M_G = 12$). The removed sources (blue histograms) don't extend beyond $1/\varpi = 400$ pc and all have parallax SNR greater than unity so I can be confident in their high absolute magnitudes. Some sources with $M_G > 12$ will pass this very cautious cut but I expect these will be dominated by the number of bright sources with well measured parallax in the Solar neighbourhood. My remaining samples after applying this cut are shown by the red histograms.



(a) No Truncation

(b) s < 160 kpc

Fig. 6.10 Fitted models and sample number densities per unit z (kpc, left), M_G (middle) and G (right) for the thin disc (orange) and thick disc (green) and halo (purple) and their sum total (black). Lines show the median model fits with shaded regions providing the 1$^{\mathrm{st}}$ – 99$^{\mathrm{th}}$ percentile range of the posterior fits to the *Gaia* data. In most cases the posterior is so tightly constrained that the uncertainties cannot be picked out in these plots. **a**: For the infinite halo model there is qualitative agreement between south (top) and north (bottom) disc samples with a steeper northern halo profile. Due to the large total normalisation of the infinite halo within $b > 80°$, the halo dominates the absolute magnitude profile and it sits directly under the total profile. **b**: The model with halo truncated such that $s < 160$ kpc also has similar north and south profiles with a marginally steeper south halo. For both models the red histograms in the left panels show the distribution $\sin(|b|)/\varpi$ which is significantly different to the fit model due to a combination of the selection function and parallax error which I have demonstrated need to be treated properly ($\sim 14\%$ of the sample has negative parallax and cannot even be plotted). Red histograms in the right column show the G distribution of the data which agrees very well with the product of my model with the selection function (black dotted line). At the bright end the model slightly overestimates the data which is likely because my model does not truncate at the tip of the red giant branch (see Fig. 6.3).

6.7 Results

The model is independently fit to the northern and southern *Gaia* samples. This halves the sample size in either fit but means I can draw a comparison between the Milky Way structure above and below the disc. The method is described in detail in Section 6.2.

The resultant model is shown in Fig. 6.10a. Solid lines and shaded regions show the median and $1^{\mathrm{st}} - 99^{\mathrm{th}}$ percentile ranges for the fits to the individual components and sum total. I evaluate this by drawing 1000 samples from the MCMC posterior, evaluating the model and taking the percentiles as a function of z, M_G and G. Each of the three Galaxy components are well constrained with the thin disc dominating the model for $z < 0.5$ kpc, the thick disc being the main contribution for $0.5 < z < 5$ kpc and the halo taking over at large distances. The thin and thick disc profiles are qualitatively very similar between the north and south samples however the halo profile in the south fit declines much more slowly with distance.

In the left panels of Fig. 6.10a red histograms show the number density of stars as a function of $z = \sin(b)/\varpi$, which provides a biased estimate of height above the Milky Way disc. The distribution is significantly lower than my model at large scale heights both due to the selection function and because parallax uncertainty scatters measurements to either larger positive or negative observed parallax. Faint sources in *Gaia* have typical parallax uncertainties $\sigma_\varpi > 0.1$ mas and so measuring $0 < \varpi < 0.1$ mas (which corresponds to $z > 10$ kpc) is unlikely due purely to measurement noise. Some of these sources are scattered up in parallax and down in distance generating the excess of sources with measured $z \sim 0.4$ kpc. This can also be seen in Fig. 6.5 where the imposed *Gaia*-like selection function and parallax uncertainties have the same effect on the naive distribution of $z = \sin(b)/\varpi$. The point I am making here is that one must account for both parallax uncertainty and the *Gaia* selection function to obtain an unbiased model of the Milky Way distribution of stars.

Unlike several previous works such as Jurić et al. 2008 and Mateu & Vivas 2018, my model extends to infinity so I require $n_\mathrm{H} > 3$ to keep the model normalisable. However, this is not a physical constraint and other studies have shown that the halo drops off much steeper beyond $r \gtrsim 50$ kpc (Deason et al., 2014) or $r \sim 160$ kpc (Fukushima et al., 2019). My model is dominated by information from the inner, shallower component of their profiles. This leads to an overestimate of the overall halo normalisation, which consequently are untrustworthy.

To obtain a more realistic halo normalisation, I rerun the fits truncating the parallax integral and halo normalisation with $s < 160$ kpc (i.e. $1/s > 6.25\,\mu$as) and changing the halo exponent prior to $n_\mathrm{H} \sim \mathrm{U}[2, 7.3]$. The spatial and absolute magnitude profiles are shown in Fig. 6.10b. In this case, the north and south halo profiles are both significantly steeper.

The right hand panels of Figs. 6.10a and 6.10b show the apparent magnitude distribution marginalised over position on the sky and distance. I weight the total distribution by the selection function which produces the black dotted line. This sits directly on the red histograms which give the apparent magnitude distribution of the *Gaia* data. The model slightly overestimates the apparent magnitude distribution at the bright end ($G \lesssim 7$) which I expect is due to the truncation of the absolute magnitude distribution at the tip of the red giant branch which can be seen at $M_G \sim -3$ in Fig. 6.5 but which I do not account for in my model.

The posteriors on each parameter are shown in Fig. 6.11 for the north and south samples (blue and red respectively) with distance truncated fits shown with dashed contours. Across all parameters there are systematic differences between the results from the north and south samples. For the thin and thick disc parameters these differences are small. However, in the case of the halo, the effect is far more substantial. Transitioning from an infinite to a truncated halo also significantly modifies the halo parameters with small knock-on effects to the disc. Given previous work (Deason et al., 2014; Fukushima et al., 2019), I consider the truncated model to be the more appropriate and will use those fits for my final results.

The posterior median, 16th and 84th percentiles for all components and parameters in each of the runs are given in Table 6.3.

6.7.1 Stellar Mass Density

My parameterisation, in particular the component normalisation (w_c), is specific to this sample as it is the total number of source with $M_G < 12$ within the cone $|b| > 80°$. The local stellar mass density (ρ^*_{local}), local surface density ($\Sigma^*_{\mathrm{local}}$) and halo total stellar mass (M^*_{Halo}) are more generally interesting to the Galactic dynamics community and can be estimated from my results as I explain here.

The number density of sources in the Solar neighbourhood with $M_G < 12$ is given by $w_c \cdot \nu_c(s = 0)$ where subscript c refers to each of the three Milky Way components. I can inflate this to include main sequence sources with $M_G > 12$ using the isochrones from Section 6.3 and the IMF. The isochrones translate $M_G = 12$ to a minimum initial mass of sources in my sample for each component, giving $\mathcal{M}_{\mathrm{ini,min}} = 0.177, 0.147, 0.115\,\mathrm{M}_{\odot}$ for the thin, thick disc and halo respectively. The maximum initial mass of stars before they reach the post-AGB evolution phase – eventually leading to a compact object remnant and thus disappearing from my sample – is $\mathcal{M}_{\mathrm{ini,max}} = 1.083, 0.980, 0.801\,\mathrm{M}_{\odot}$. To get the total pre-compact object local number density of sources, I inflate my local number density by a factor

$$X_c = \frac{\int_0^{\mathcal{M}_{\mathrm{ini,max}}} \xi(\mathcal{M}_{\mathrm{ini}}) \mathrm{d}\mathcal{M}_{\mathrm{ini}}}{\int_{\mathcal{M}_{\mathrm{ini,min}}}^{\mathcal{M}_{\mathrm{ini,max}}} \xi(\mathcal{M}_{\mathrm{ini}}) \mathrm{d}\mathcal{M}_{\mathrm{ini}}} \tag{6.41}$$

where $\xi(\mathcal{M}_{\mathrm{ini}})$ is the IMF (I use Kroupa 2001). This gives $X_c = 3.167, 2.785, 2.398$ for the three components.

To estimate the local stellar mass density, I need the mean mass of sources in the population. I can use the IMF again for this however I need to account for stellar mass loss. I use the three component isochrones to transform from initial mass to current stellar mass, $\mathcal{M}(\mathcal{M}_{\mathrm{ini}})$. The stellar evolution models do not extend all the way to zero mass so I

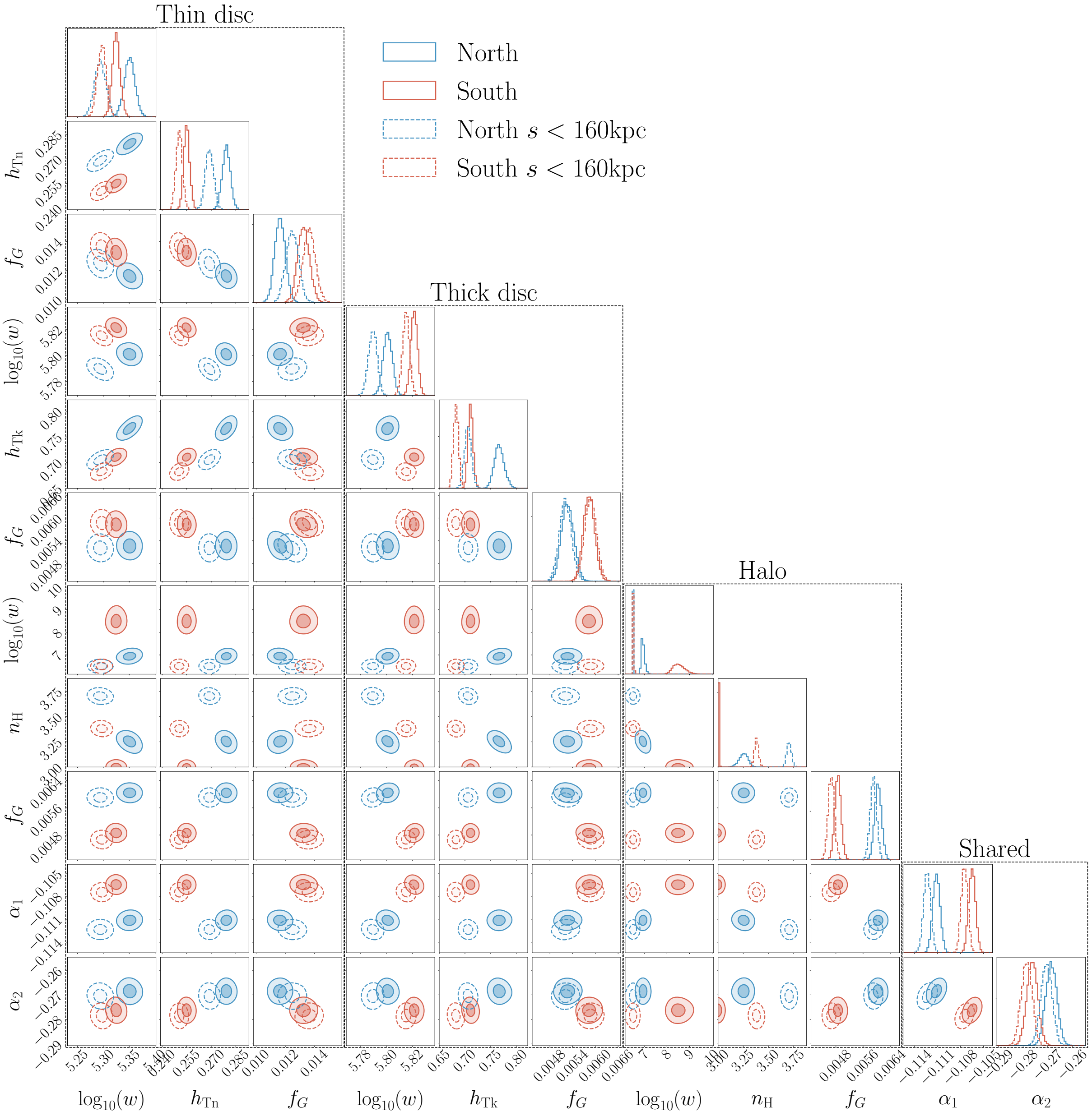


Fig. 6.11 The posterior distributions for the north (red solid contours) and south (blue solid contours) sample fits show a small but significant disagreement across most parameters suggesting a weak asymmetry. Constraining the model to s < 160 kpc (dashed contours) has a small impact on disc parameters however the halo model is much more significantly affected. Notably, the halo power-law index, which pushes close to the lower bound for an un-truncated model, is fit with a significantly steeper profile when the truncation is applied. The truncated model is better suited to the Milky Way for which the halo will not extend indefinitely.

Component	Parameter	North	South	North (s < 160kpc)	South (s < 160kpc)
Thin disc	w	$2.24^{+0.05}_{-0.05} \times 10^5$	$2.11^{+0.04}_{-0.03} \times 10^5$	$1.97^{+0.05}_{-0.05} \times 10^5$	$1.98^{+0.04}_{-0.04} \times 10^5$
	h_{Tn}	$0.279^{+0.002}_{-0.002}$	$0.255^{+0.002}_{-0.002}$	$0.269^{+0.003}_{-0.002}$	$0.250^{+0.002}_{-0.002}$
	f_G	$1.16^{+0.04}_{-0.04} \times 10^{-2}$	$1.32^{+0.04}_{-0.04} \times 10^{-2}$	$1.25^{+0.04}_{-0.04} \times 10^{-2}$	$1.36^{+0.04}_{-0.04} \times 10^{-2}$
	M_{TO}	3.1			
	α_3	-0.6			
Thick disc	w	$6.32^{+0.05}_{-0.05} \times 10^5$	$6.63^{+0.04}_{-0.04} \times 10^5$	$6.16^{+0.05}_{-0.05} \times 10^5$	$6.53^{+0.04}_{-0.04} \times 10^5$
	h_{Tk}	$0.766^{+0.010}_{-0.009}$	$0.711^{+0.005}_{-0.005}$	$0.706^{+0.007}_{-0.007}$	$0.683^{+0.005}_{-0.005}$
	f_G	$5.26^{+0.17}_{-0.16} \times 10^{-3}$	$5.83^{+0.15}_{-0.14} \times 10^{-3}$	$5.21^{+0.16}_{-0.16} \times 10^{-3}$	$5.87^{+0.15}_{-0.15} \times 10^{-3}$
	M_{TO}	3.1			
	α_3	-0.73			
Halo	w	$8.64^{+1.69}_{-1.07} \times 10^6$	$3.26^{+3.11}_{-1.38} \times 10^8$	$3.19^{+0.05}_{-0.05} \times 10^6$	$3.39^{+0.06}_{-0.05} \times 10^6$
	n_{H}	$3.254^{+0.041}_{-0.046}$	$3.005^{+0.004}_{-0.003}$	$3.705^{+0.023}_{-0.023}$	$3.382^{+0.020}_{-0.020}$
	f_G	$6.04^{+0.10}_{-0.10} \times 10^{-3}$	$4.86^{+0.09}_{-0.09} \times 10^{-3}$	$5.91^{+0.09}_{-0.09} \times 10^{-3}$	$4.66^{+0.09}_{-0.09} \times 10^{-3}$
	M_{TO}	3.1			
	α_3	-0.64			
Shared	α_1	$-0.1111^{+0.0005}_{-0.0005}$	$-0.1065^{+0.0004}_{-0.0004}$	$-0.1123^{+0.0005}_{-0.0005}$	$-0.1075^{+0.0004}_{-0.0004}$
	α_2	$-0.2685^{+0.0023}_{-0.0023}$	$-0.2763^{+0.0022}_{-0.0022}$	$-0.2705^{+0.0022}_{-0.0023}$	$-0.2785^{+0.0022}_{-0.0023}$

Table 6.3 The median and 16th – 84th percentile ranges of the posterior parameter distributions from fits to the *Gaia* samples are shown for the north and south samples for an infinitely extending model and for a halo truncated at $s = 160$ kpc. Across all parameters there is a significant asymmetry between the results of fitting to the north and south samples. The south disc profiles are steeper than the north with a smaller scale height, however, the southern halo is significantly shallower than the northern halo even pushing up against the prior boundary for the un-truncated model.

assume any stars with $\mathcal{M}_{\mathrm{ini}} < 0.1$ experience negligible mass loss in their lifetimes such that $\mathcal{M}(\mathcal{M}_{\mathrm{ini}}) = \mathcal{M}_{\mathrm{ini}}$. The mean mass of all non-compact object stars is

$$\langle \mathcal{M} \rangle_c = \frac{\int_0^{\mathcal{M}_{\mathrm{ini,max}}} \mathcal{M}(\mathcal{M}_{\mathrm{ini}})\,\xi(\mathcal{M}_{\mathrm{ini}})\,\mathrm{d}\mathcal{M}_{\mathrm{ini}}}{\int_0^{\mathcal{M}_{\mathrm{ini,max}}} \xi(\mathcal{M}_{\mathrm{ini}})\,\mathrm{d}\mathcal{M}_{\mathrm{ini}}}$$

which gives $\langle \mathcal{M}_c \rangle = 0.174, 0.168, 0.155$ for the three components. Finally, the local mass density of non-compact object stars is

$$\rho^*_{\mathrm{local},c} = w_c\, \nu_c(s = 0) \cdot X_c \cdot \langle \mathcal{M} \rangle_c. \tag{6.42}$$

A critical assumption I have made is that any stars born with an initial mass larger than $\mathcal{M}_{\mathrm{ini,max}}$ will not appear in my sample. In reality, the White Dwarf sequence extends up to $M_G \sim 8$ (see Rix et al., 2021) and so there may be many White Dwarfs in my sample. However, these will be dominated by the main sequence dwarfs of the same absolute magnitude and will only provide a severely sub-dominant contribution to the number density (see Fig. 2 Gaia Collaboration et al., 2021b).

I estimate the surface densities by integrating my components with respect to z. Since my power-law halo has $n > 3$, the total halo stellar mass is not well normalised at $r = 0$. I estimate the total halo mass, M^*_{Halo}, by integrating my halo profile for $r > 1$ kpc and taking $n_H = 2$ (uniform density) inside. As a result, the halo mass is largely dominated by stars inside the Solar radius and is an extrapolation of the local halo stellar mass density so this should be taken with caution.

The means and standard deviations of these parameters using the $s < 160$ kpc fits are μ^{North} and μ^{South} in Table 6.4.

6.8 Systematic Errors

There are various aspects of the model which may lead to systematic errors in the posterior parameter fits. In most cases, these originate from simplifications to make the optimization computationally tractable. Here, I address some of the aspects which are capable of biasing the results and test the significance of their impact on the inferred parameters.

The tests are all performed using mock catalogues. The first four systematic tests (Sections 6.8.1-6.8.4) use the exact same sample as Section 6.4, but resampling from the selection function where apparent magnitudes are altered. In Fig. 6.12 and Table 6.5 I have provided the posteriors of the “SF & σ_ϖ” fit from Section 6.5 for comparison and labeled it “Good” as this was fit under ideal circumstances where the data correctly represents the model. The tests in Sections 6.8.5 and 6.8.6 use re-sampled catalogues applying the same method as Section 6.2 including parallax uncertainties from the Astrometric Spread Function (see Chapter 5). There is a level of statistical error in population sampling which affects the posteriors for tests when a new catalogue is generated.

	μ^{North}	μ^{South}	μ	$\sigma_{\mathrm{sys}}^{\mathrm{N/S}}$	$\sigma_{\mathrm{sys}}^{\mathrm{Tests}}$	
ρ^*_{local} $(\mathrm{M_\odot/pc^3})$	$3.28 \pm 0.03 \times 10^{-2}$	$4.05 \pm 0.04 \times 10^{-2}$	$3.66 \pm 0.03 \times 10^{-2}$	$\pm 0.39 \times 10^{-2}$	$\pm 0.34 \times 10^{-2}$	$(z_\odot)$
$\Sigma^*_{\mathrm{local}}$ $(\mathrm{M_\odot/pc^2})$	22.02 ± 0.09	24.83 ± 0.10	23.42 ± 0.09	± 1.40	± 2.90	(R)
$\log_{10}\left(M^*_{\mathrm{Halo}}\ (\mathrm{M_\odot})\right)$	8.97 ± 0.01	8.74 ± 0.01	8.86 ± 0.01	± 0.12	± 0.15	(R)
ρ_{Tn} $(\mathrm{M_\odot/pc^3})$	$2.86 \pm 0.03 \times 10^{-2}$	$3.56 \pm 0.03 \times 10^{-2}$	$3.21 \pm 0.03 \times 10^{-2}$	$\pm 0.35 \times 10^{-2}$	$\pm 0.38 \times 10^{-2}$	$(z_\odot)$
ρ_{Tk} $(\mathrm{M_\odot/pc^3})$	$4.19 \pm 0.13 \times 10^{-3}$	$4.92 \pm 0.11 \times 10^{-3}$	$4.57 \pm 0.13 \times 10^{-3}$	$\pm 0.34 \times 10^{-3}$	$\pm 2.17 \times 10^{-3}$	(A_V,σ_G)
ρ_{H} $(\mathrm{M_\odot/pc^3})$	$2.13 \pm 0.03 \times 10^{-5}$	$1.51 \pm 0.02 \times 10^{-5}$	$1.82 \pm 0.03 \times 10^{-5}$	$\pm 0.31 \times 10^{-5}$	$\pm 0.22 \times 10^{-5}$	(R)
$\rho_{\mathrm{Tk}}/\rho_{\mathrm{Tn}}$	0.147 ± 0.005	0.138 ± 0.003	0.141 ± 0.005	± 0.000	± 0.075	(A_V,σ_G)
$\rho_{\mathrm{H}}/\rho_{\mathrm{Tn}}$	$7.46 \pm 0.14 \times 10^{-4}$	$4.25 \pm 0.07 \times 10^{-4}$	$5.85 \pm 0.14 \times 10^{-4}$	$\pm 1.60 \times 10^{-4}$	$\pm 1.10 \times 10^{-4}$	(R)
Σ_{Tn} $(\mathrm{M_\odot/pc^2})$	15.36 ± 0.15	17.82 ± 0.13	16.59 ± 0.15	± 1.22	± 2.22	$(z_\odot)$
Σ_{Tk} $(\mathrm{M_\odot/pc^2})$	5.92 ± 0.13	6.72 ± 0.11	6.33 ± 0.13	± 0.38	± 1.67	(σ_G,A_V)
Σ_{H} $(\mathrm{M_\odot/pc^2})$	0.74 ± 0.04	0.29 ± 0.01	0.51 ± 0.04	± 0.22	± 0.37	(R)
h_{Tn} (kpc)	0.269 ± 0.003	0.250 ± 0.002	0.260 ± 0.003	± 0.009	± 0.024	(σ_G,A_V)
h_{Tk} (kpc)	0.706 ± 0.007	0.682 ± 0.005	0.693 ± 0.007	± 0.010	± 0.121	(R)
n_{H}	3.705 ± 0.023	3.382 ± 0.020	3.543 ± 0.023	± 0.160	± 0.204	(R)
$f^{\mathrm{G}}_{\mathrm{Tn}}$	$1.25 \pm 0.04 \times 10^{-2}$	$1.36 \pm 0.04 \times 10^{-2}$	$1.31 \pm 0.04 \times 10^{-2}$	$\pm 0.04 \times 10^{-2}$	$\pm 0.84 \times 10^{-2}$	$(M^{\mathrm{Tn}}_{\mathrm{TO}},\sigma_G)$
$f^{\mathrm{G}}_{\mathrm{Tk}}$	$5.21 \pm 0.16 \times 10^{-3}$	$5.87 \pm 0.15 \times 10^{-3}$	$5.55 \pm 0.16 \times 10^{-3}$	$\pm 0.29 \times 10^{-3}$	$\pm 1.05 \times 10^{-3}$	$(\Delta\varpi,A_V)$
$f^{\mathrm{G}}_{\mathrm{H}}$	$5.91 \pm 0.10 \times 10^{-3}$	$4.67 \pm 0.09 \times 10^{-3}$	$5.28 \pm 0.10 \times 10^{-3}$	$\pm 0.61 \times 10^{-3}$	$\pm 0.64 \times 10^{-3}$	$(\Delta\varpi)$
α_1	-0.1123 ± 0.0005	-0.1075 ± 0.0004	-0.1099 ± 0.0005	± 0.0024	± 0.0050	$(\Delta\varpi)$
$\alpha 2$	-0.2705 ± 0.0023	-0.2785 ± 0.0022	-0.2745 ± 0.0023	± 0.0034	± 0.0244	$(\Delta\varpi)$

Table 6.4 Transformed results from the model fits to *Gaia* data are given for the north, south and combined samples along with one standard deviation uncertainties. While the statistical uncertainties for each sample are incredibly tight for most parameters, the systematic uncertainties due to north-south asymmetry and model oversimplifications are much larger. It is important to consider these additional systematics when using my results. The systematic uncertainties should be added in quadrature to the statistical uncertainty.

6.8.1 Solar Vertical Offset

In my model, I assume that the Sun sits directly on the mid-plane of the Milky Way and as such I have a symmetric view of the Galaxy towards the north and south. In fact, the Sun is slightly vertically offset from the Galactic plane to the north by $\sim 14 - 21$pc (Bennett & Bovy, 2019; Binney et al., 1997; Joshi, 2007; Widmark & Monari, 2019). As a result, my model assumes the distribution of stars in the south is closer than it actually is and in the north, too far away. This may impact the inferred scale height of the discs.

To test the significance of this assumption, I use my mock sample and introduce a vertical shift to the effective Solar position. This is done for stars assuming the sample is entirely in the north. The vertical position is changed for all stars such that the new coordinate, z' is given by

$$z' = z - z_{\odot}, \tag{6.43}$$

with $z_{\odot} = 21$ pc – towards the upper end of estimates of the Solar position offset from the Galactic plane. This reduces both the latitude and distance of sources and therefore also reduces the apparent magnitudes

$$\tan b' = \frac{s \tan b - z_{\odot} \sec b}{s} \tag{6.44}$$

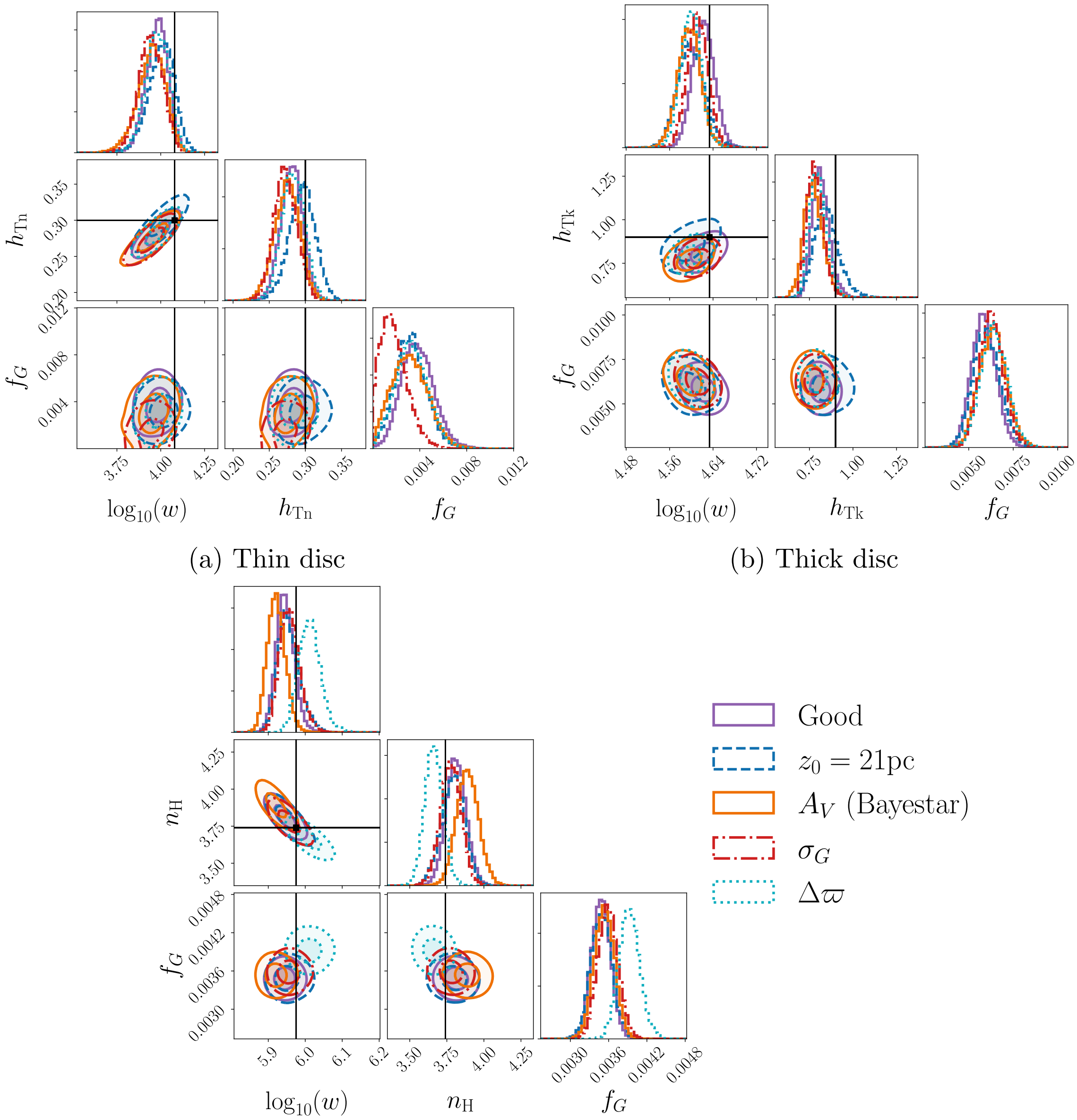
$$s' = s \left[1 + \frac{z_{\odot}}{s} \left(\frac{z_{\odot}}{s} - 2 \sin b \right) \right]^{\frac{1}{2}} . \tag{6.45}$$

My latitude cut is applied on the updated latitudes, $|b'| > 80$. I do not use the southern population, as this requires re-sampling the mock outside the original selection bounds which would be more complicated to interpret. For the north sample, this cut simply removes some sources from the original data set. The source apparent magnitudes are then recomputed from their original absolute magnitudes and the new distances, after which the selection function is applied to the sample and finally observed parallaxes are re-sampled from the expected uncertainties. The sample size is reduced by $\sim 1.8\%$ over the original.

The results of the parameter fits to the new sample are given in Table 6.5 and shown by the blue dashed contours in Fig. 6.12. The shifts of parameters from the true values are marginally significant in some cases. Specifically, the scale height of the thin disc is slightly increased which may be considered counter intuitive given that I am effectively pushing the Sun closer to sources. However, pairing this with the increased weight of the thin disc and reduced weight of the thick disc it suggests the thin disc is taking on some thick disc sources. Overall, these results suggest that the simplification to the model of setting $z_{\odot} = 0$ is only likely to have a marginal effect on parameter estimates.

6.8.2 Dust Extinction

Extinction due to inter-stellar dust causes stars to appear dimmer than they would otherwise be at a given distance and absolute magnitude. This is one of the motivations behind narrowing the sample to high-latitude regions. In these areas, the effects of dust



(a) Thin disc

(b) Thick disc

(c) Halo

Fig. 6.12 Posterior parameter fits to the mock sample from Section 6.4 for the thin disc parameters (a, top left), thick dick (b, top right) and halo (c, bottom) under alterations to the data which could introduce systematic errors. The purple solid "Good" contours in all panels show the posteriors from Section 6.5, fit to the sample without any imposed systematics and black dot and lines show the input parameters used to generate the sample. Adding a $z_\odot$ offset (blue dotted) has only a marginal impact on most parameter estimates. *Gaia*-like magnitude error (σ_G, red dot-dashed) leads us to underestimate the thin disc giant fraction. Extinction from Bayestar (A_V, orange solid) biases the model towards an overly steep halo whilst a $-10\,\mu$as parallax offset ($\Delta\varpi$, cyan dotted) has the opposite effect.

Component	Parameter	Input	Good fit	z_0 = 21pc	A_V	σ_G	$\Delta\varpi$
Thin disc	w	1.20×10^4	$9.59^{+1.38}_{-1.35} \times 10^3$	$1.01^{+0.17}_{-0.17} \times 10^4$	$8.93^{+1.65}_{-1.55} \times 10^3$	$8.76^{+1.52}_{-1.27} \times 10^3$	$9.56^{+1.52}_{-1.37} \times 10^3$
	h_{Tn}	0.300	$0.281^{+0.015}_{-0.015}$	$0.296^{+0.017}_{-0.018}$	$0.275^{+0.017}_{-0.018}$	$0.272^{+0.017}_{-0.015}$	$0.283^{+0.016}_{-0.015}$
	f_G	4.50×10^{-3}	$3.76^{+1.43}_{-1.30} \times 10^{-3}$	$3.20^{+1.37}_{-1.27} \times 10^{-3}$	$3.11^{+1.60}_{-1.49} \times 10^{-3}$	$1.67^{+1.28}_{-0.98} \times 10^{-3}$	$3.28^{+1.43}_{-1.36} \times 10^{-3}$
	M_{TO}	3.1					
	α_3	-0.6					
Thick disc	w	4.30×10^4	$4.19^{+0.20}_{-0.19} \times 10^4$	$3.97^{+0.24}_{-0.21} \times 10^4$	$3.95^{+0.19}_{-0.18} \times 10^4$	$4.09^{+0.19}_{-0.17} \times 10^4$	$3.98^{+0.17}_{-0.17} \times 10^4$
	h_{Tk}	0.900	$0.812^{+0.052}_{-0.045}$	$0.846^{+0.072}_{-0.064}$	$0.772^{+0.058}_{-0.050}$	$0.780^{+0.050}_{-0.042}$	$0.798^{+0.052}_{-0.045}$
	f_G	5.40×10^{-3}	$5.83^{+0.69}_{-0.66} \times 10^{-3}$	$5.99^{+0.77}_{-0.72} \times 10^{-3}$	$6.35^{+0.74}_{-0.76} \times 10^{-3}$	$6.28^{+0.69}_{-0.65} \times 10^{-3}$	$6.40^{+0.76}_{-0.70} \times 10^{-3}$
	M_{TO}	3.1					
	α_3	-0.77					
Halo	w	9.45×10^5	$8.81^{+0.52}_{-0.45} \times 10^5$	$8.97^{+0.69}_{-0.51} \times 10^5$	$8.34^{+0.49}_{-0.43} \times 10^5$	$9.05^{+0.59}_{-0.51} \times 10^5$	$1.03^{+0.07}_{-0.07} \times 10^6$
	n_{H}	3.740	$3.812^{+0.068}_{-0.066}$	$3.795^{+0.076}_{-0.081}$	$3.889^{+0.075}_{-0.073}$	$3.778^{+0.068}_{-0.068}$	$3.661^{+0.063}_{-0.060}$
	f_G	3.50×10^{-3}	$3.48^{+0.15}_{-0.15} \times 10^{-3}$	$3.49^{+0.16}_{-0.17} \times 10^{-3}$	$3.53^{+0.16}_{-0.16} \times 10^{-3}$	$3.59^{+0.16}_{-0.15} \times 10^{-3}$	$3.92^{+0.16}_{-0.16} \times 10^{-3}$
	M_{TO}	3.1					
	α_3	-0.64					
Shared	α_1	-0.110	$-0.110^{+0.002}_{-0.002}$	$-0.110^{+0.002}_{-0.002}$	$-0.111^{+0.002}_{-0.002}$	$-0.108^{+0.002}_{-0.002}$	$-0.105^{+0.002}_{-0.002}$
	α_2	-0.250	$-0.252^{+0.009}_{-0.008}$	$-0.252^{+0.009}_{-0.009}$	$-0.257^{+0.009}_{-0.009}$	$-0.246^{+0.009}_{-0.009}$	$-0.228^{+0.009}_{-0.009}$

Table 6.5 The results of systematics due to shifting the Sun from the Galactic plane ($z_\odot$ = 21pc), introducing dust extinction (A_V), adding apparent magnitude error (σ_G) and introducing a $-10\,\mu$as zero point parallax offset ($\Delta\varpi$) are tested for the mock sample. "Good fit" provides the results from Section 6.5 where the parameters are fit to data drawn from the same model with no systematics imposed. Most tests have only a marginal impact on parameter estimates with $\lesssim 2$ standard deviation offsets from the input parameters. Magnitude error significantly affects thin disc giant fraction whilst A_V and $\Delta\varpi$ have a stronger impact on the halo parameters.

extinction are small for any individual star. However, since this systematically affects all sources in the same direction, there can still be a sizable affect on the model parameter estimates.

Why don't I use the published dust maps to de-redden the sources in the *Gaia* samples in the first place? At first glance, this is an appealing suggestion, but there is a subtle issue here which would lead to an underestimate in uncertainties. Because the Green et al. 2019 extinction map is evaluated using *Gaia* parallax information, as is my model, there would be a double counting of information. The formally correct way to handle this problem is to simultaneously fit the structure of the Milky Way and the extinction map. One immediate challenge is a strong degeneracy between extinction, distance and absolute magnitude of sources. This is a significantly more complex problem and well beyond the scope of my thesis.

Nonetheless, I can gauge the impact of extinction by applying a Milky Way extinction map to the mock catalogue, re-sample the selection function from the new observed apparent magnitudes and fit the model parameters to this.

The most detailed 3D extinction map to date for the Milky Way is that of Green et al. 2019. This uses apparent magnitudes from a wide range of pass-bands throughout the optical and infra-red to estimate stellar reddening, whilst *Gaia* parallaxes are used to provide distance information to the model. Using the DUSTMAPS Python module (Green, 2018), I take a single sample of the extinction parameter for each source. As a proxy for the reddening vector component in the *Gaia* G band, I use the Pan-STARRS' g-filter value of 3.518 (Table 1 of Green et al., 2019). The mean extinction for sources in the selected sample is $\langle \delta G \rangle \sim 0.01$.

The addition of stellar extinction from Green et al. 2019 has a marginally significant effect on parameter estimates shown by the orange contours in the posteriors in Fig 6.12 and more quantitatively in the A_V column of Table 6.5. The scale height and normalisation of the discs and halo are pushed down and the halo is too steep. Stars further away will be more obscured by dust and less likely to be included in my sample. My method doesn't account for this so I instead fit a marginally steeper model than is actually the case. At high latitudes, this is a small effect but if I widened my on-sky sample, this could dramatically impact the results.

6.8.3 Magnitude Uncertainty

The method used to fit the model assumes no apparent magnitude measurement uncertainty. Of course this is not the case and *Gaia* has uncertainties on all apparent magnitude measurements. To estimate the systematic effect of this uncertainty, I re-sample the mock apparent magnitudes from *Gaia*-like uncertainties and apply the method to the new sample.

For any source in the *Gaia* catalogue, the apparent magnitude in the G-band is estimated up to nine times whenever it is scanned by the nine columns of CCDs in the field of view. The set of all apparent magnitude measurements, which can number in the hundreds, is used to estimate the magnitude uncertainty. I reverse engineer this process to

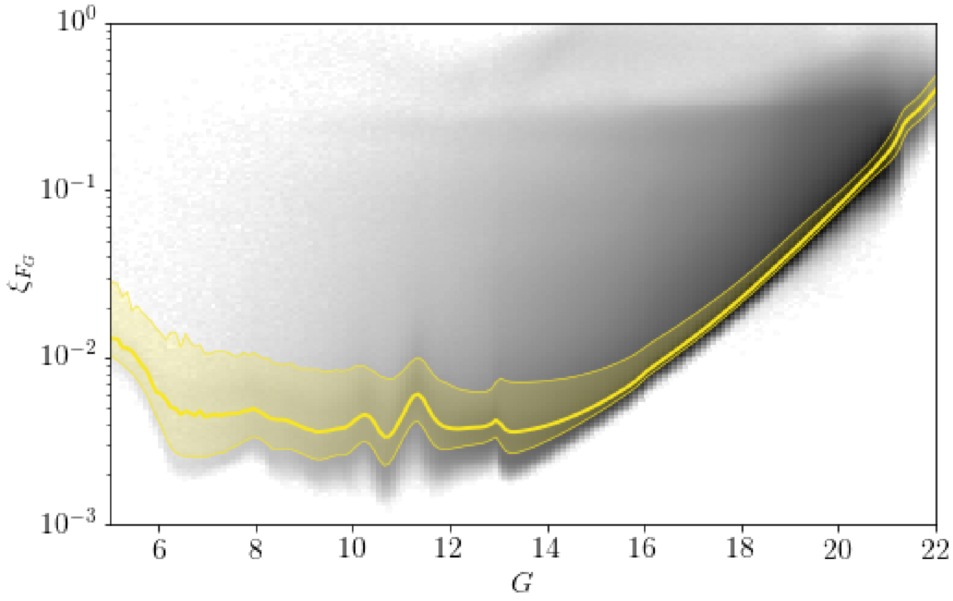


Fig. 6.13 The distribution of G-band flux error amplitude (uncertainty per observation) of all sources in *Gaia* EDR3 is shown by the log-normalised grey-scale histograms. The yellow line and shaded regions provide the median and $16^{\mathrm{th}} - 84^{\mathrm{th}}$ percentile range in 0.1mag bins. The median is used with the scanning law to estimate the expected apparent magnitude error for sources in *Gaia* as a function of position on the sky and apparent magnitude.

estimate the apparent magnitude uncertainty per observation. The amplitude of apparent magnitude measurement uncertainty is given by

$$\xi_{F_G} = \sqrt{N}\frac{\sigma_{F_G}}{F_G} \tag{6.46}$$

where F_G is the measured source flux and ξ_{F_G} is the flux error per observation (Eq. 2 Belokurov et al., 2017). I estimate this for all sources in *Gaia* EDR3 and take the median as a function of apparent magnitude shown in Fig. 6.13.

To estimate apparent magnitude errors for sources in the mock sample, I then replicate *Gaia*'s observations for those sources again. The error per observation is taken from the median in Fig. 6.13 and the number of scans of the source is given by the number of scans of that position on the sky in the EDR3 nominal scanning law[2]. Ideally, on average, 62/7 observations are taken with each scan as there are nine CCD columns but in one of the seven rows, a CCD is replaced by a wave-guide sensor (Gaia Collaboration et al., 2016). However, *Gaia* is not 100% efficient and not all observations are successfully recorded or make it through the data processing pipeline. To account for this, I also multiply by the efficiency at a given magnitude taken from Boubert & Everall 2020. Whilst this efficiency is estimated from Data Release 2 (DR2), it should give a rough approximation of the

[2] Gaia EDR3 nominal scanning law: http://cdn.gea.esac.esa.int/Gaia/gedr3/auxiliary/commanded_scan_law/

behaviour in EDR3. Eq. 6.46 is now reversed to estimate the flux error

$$\sigma_{F_G} = \sqrt{N(l,b,G)}\, \xi_{F_G}(G)\, F_G(G). \tag{6.47}$$

where $N(l,b,G)$ is the product of the number of scans at the given position on the sky and 62/7 times the observation efficiency. This is used to sample an observed flux $F'_G \sim \mathcal{N}(F_G, \sigma_{F_G})$. The mean magnitude change for all sources in my sample is $\langle|\delta G|\rangle \sim 0.003$. Finally, I apply the selection function to the newly estimated magnitudes which replicates the fact that the selection function was estimated as a function of measured apparent magnitude.

The introduction of apparent magnitude error has marginal effects on some parameter posteriors, shown by the red dot-dashed contours in Fig. 6.12 and included in Table 6.5 as σ_G, with the most significant being the thin disc dwarf fraction. This may be explained by a blurring of the sharp dwarf-giant absolute magnitude boundary which, given the low giant fraction in the thin disc, will lead to more dwarfs being estimated as giants than vice versa and reduce the dwarf fraction.

6.8.4 Parallax Offset

As discussed in Section 6.6, a significant amount of work has been devoted to constraining the zero-point parallax offset of the *Gaia* astrometry sample. However, most tests are applied on sources at the bright end of the *Gaia* magnitude range. At the faint end, the correction from Lindegren et al. 2021b reduces the parallax bias to ∼ a few micro-arcseconds as a function of apparent magnitude. However, as can be seen in the third panel of Fig. 2 in Lindegren et al. 2021b, there are variations over the sky of $\sim 10\,\mu$as towards the north and south Galactic poles.

To test the impact of a residual parallax offset, I subtract $10\,\mu$as from the parallax measurements in my mock sample and rerun the fits without correcting for this. The posteriors are shown in the final column of Table 6.5 and with cyan dotted contours in Fig. 6.12. The parallax bias has no significant impact on the thin disc and thick disc parameters however the impact on the halo is considerable. Because this is tested using the same sample as the “Good” fits, I am comparing with those posteriors rather than the input parameters. In the Fig. 6.12c I can see how much the parallax offset shifts the halo parameter posteriors away from the “Good” results towards a shallower, more extended halo profile. This is not too surprising – a constant parallax offset is small compared to the true source parallax for nearby sources, but becomes much more significant with increasing distance.

A similar effect was found in Everall et al. 2019 when measuring the local velocity ellipsoid with *Gaia* DR2. As I stray further from the Solar neighbourhood, the impact of a negative parallax offset becomes more significant, systematically overestimating the distances to sources. In my case, this causes an overestimate of the radial extent of the halo.

6.8.5 Turn-off Magnitude

When setting up the model, I fixed some parameter values. The most notable is the main-sequence turn-off which determines the absolute magnitude at which the population transitions from dwarf-dominated to giant-dominated. For all populations, I set this to $M_G = 3.1$ motivated by the theoretical isochrones, however, the turn-off magnitude is a function of stellar age. I used a thin disc model with $\tau = 6.9$ Gyr, but this is only approximately the mean age of the thin disc. The disc has formed over a long period of time and so is made up of sources with a wide range of ages (Fantin et al., 2021; Katz et al., 2021; Snaith et al., 2015). This leads to a main sequence turn-off which is extended over a range of magnitudes.

Ruiz-Lara et al. 2020 demonstrated that there was a large burst of star formation in the thin disc ~ 5.9 Gyr ago which, by inspection of isochrones, corresponds to a turn-off magnitude of $M_{\mathrm{TO}} = 2.9$. To test this, I generate a new mock *Gaia*-like catalogue with the same input parameter values as described in Section 6.4 except that $M_{\mathrm{TO}} = 2.9$ for the thin disc population. I then refit this incorrectly assuming a fixed $M_{\mathrm{TO}} = 3.1$ for all components.

The results are provided in Table 6.6 and shown by the blue solid contours in Fig. 6.14. This systematic error increases the giant fraction of the thin disc, whilst not significantly affecting any other parameter. By moving M_{TO} lower and fitting with a higher value, I am classifying many dwarfs as giants in the model. Therefore, the effect on f_{G} is unsurprising but it is reassuring to see that the remaining parameters are not sensitive to small changes in the absolute magnitude distribution.

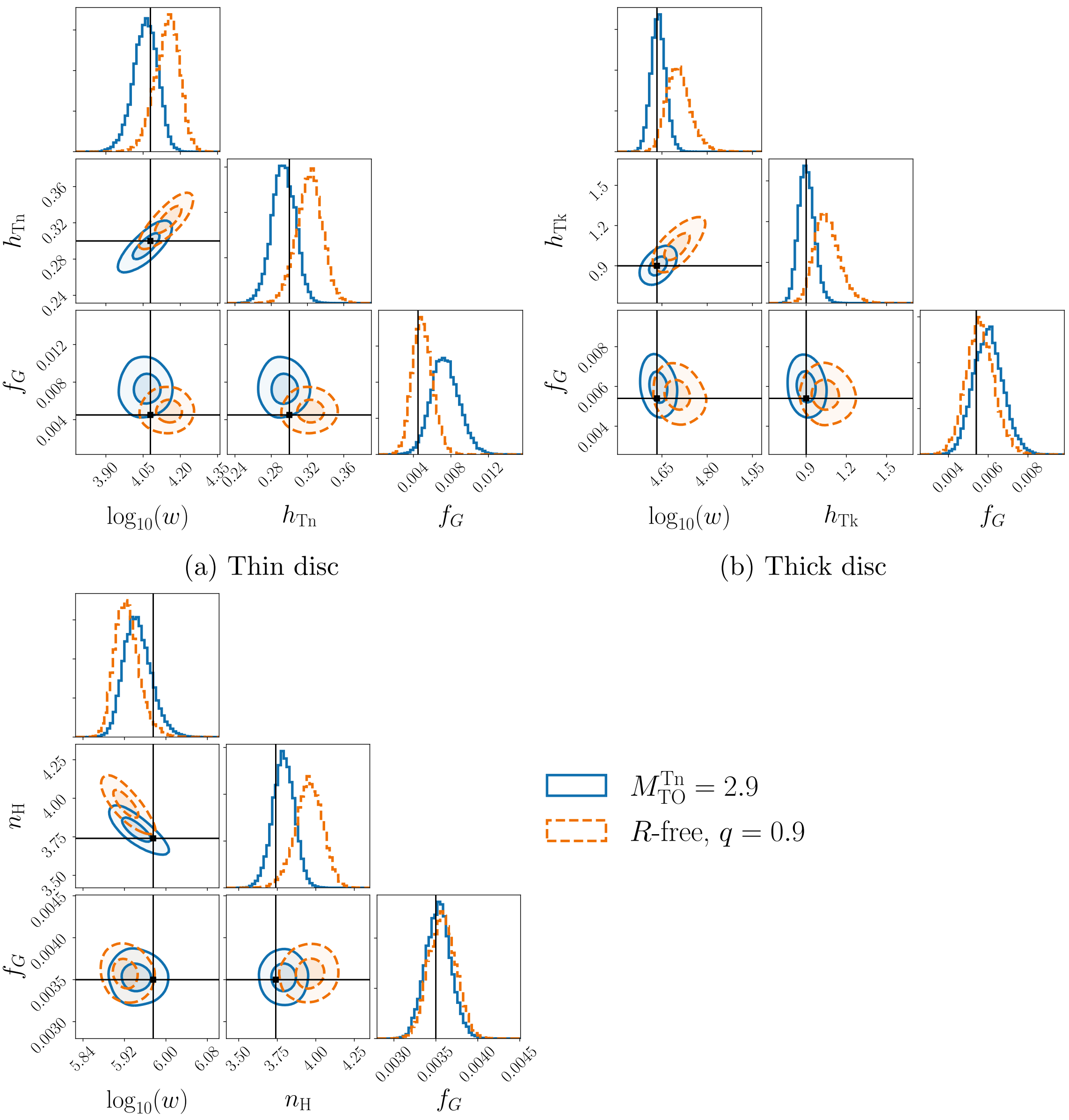
6.8.6 Galactocentric Radius and an Oblate Halo

The model used in this work has no dependence on Galactocentric cylindrical radius or azimuth. Due to the complexity of the model integration, I made the simplifying assumption that all sources have the same cylindrical radius as the Sun, $R_{\odot}$. For a detailed discussion of this, see Section 6.3.3.

For the thin and thick disc profiles, this means approximating the Milky Way disc as a uniform sheet. In the most extreme case, where sources are against the edge of the cone with $l = 0°$ or $180°$, the cylindrical radius is incorrect by

$$\delta R = z/\tan 80° = 0.176\, z. \tag{6.48}$$

Whilst this maximum offset is significant, it does not provide much information on how the disc profile will affect the results. For that, I examine the mean cylindrical radius offset integrating over the disc profile. Using radial scalelengths of $L_{\mathrm{Tn}} = 2.6$ kpc and $L_{\mathrm{Tk}} = 3.6$ kpc for the thin and thick disc respectively from Jurić et al. 2008, I draw a sample within my $|b| > 80°$ cone at fixed z for a thin and thick disc weighted by $\exp(-R/L)$ and estimate the mean R. For the thin disc, I get $\delta R \sim 0.2$ pc at $z = 0.3$ kpc, whilst the thick disc produces $\delta R \sim 1.7$ pc at $z = 0.9$ kpc. Therefore, the average offset of sources from their true position is small.



(a) Thin disc

(b) Thick disc

(c) Halo

Fig. 6.14 Posterior distributions for parameters fit to mock samples with systematic differences to the type of model assumed in the fitting procedure. Ground truth input parameters for the mock sample are shown by the black lines and points for the thin disc (a, top left sub-figure), thick disc (b, top right sub-figure) and halo parameters (c, bottom sub-figure). The fits to the mock sample with a shifted thin disc turn-off ($M_{\rm TO}^{\rm Tn}$, blue contours) overestimate the population of thin disc giants which is unsurprising as the fits assume a giant is any source with $M_G < 3.1$. Orange dashed contours are the fits to a mock sample with cylindrical radius dependence and an oblate halo ($q = 0.9$, orange dashed). This significantly impacts all parameters producing extended thin and thick discs with overestimated scale heights and an overly-steep halo.

Component	Parameter	Input	$M_{\rm TO}^{\rm Tn} = 2.9$	R-free, $q = 0.9$
Thin disc	w	1.20	$1.15^{+0.13}_{-0.13} \times 10^4$	$1.41^{+0.15}_{-0.17} \times 10^4$
	$h_{\rm Tn}$	0.300	$0.294^{+0.013}_{-0.013}$	$0.323^{+0.014}_{-0.015}$
	f_G	4.50	$7.34^{+1.56}_{-1.39} \times 10^{-3}$	$4.93^{+1.12}_{-1.01} \times 10^{-3}$
Thick disc	w	4.30	$4.34^{+0.25}_{-0.23} \times 10^4$	$5.05^{+0.49}_{-0.41} \times 10^4$
	$h_{\rm Tk}$	0.900	$0.901^{+0.061}_{-0.055}$	$1.049^{+0.099}_{-0.085}$
	f_G	5.40	$5.99^{+0.75}_{-0.73} \times 10^{-3}$	$5.58^{+0.71}_{-0.69} \times 10^{-3}$
Halo	w	9.45	$8.78^{+0.59}_{-0.48} \times 10^5$	$8.39^{+0.48}_{-0.40} \times 10^5$
	$n_{\rm H}$	3.740	$3.791^{+0.070}_{-0.070}$	$3.957^{+0.086}_{-0.089}$
	f_G	3.50	$3.53^{+0.15}_{-0.15} \times 10^{-3}$	$3.58^{+0.16}_{-0.16} \times 10^{-3}$
Shared	α_1	−0.110	$-0.108^{+0.002}_{-0.002}$	$-0.111^{+0.002}_{-0.002}$
	α_2	−0.250	$-0.249^{+0.009}_{-0.008}$	$-0.238^{+0.008}_{-0.008}$

Table 6.6 I provide the median parameter estimate with 16th – 84th percentiles for the mock sample fits with thin disc $M_{\rm TO} = 2.9$ and the cylindrical radius dependent sample with halo oblateness $q = 0.9$.

The halo spatial distribution is defined as a power-law profile of Galactocentric spherical radius. An incorrect cylindrical radius leads to an incorrect spherical radius. The spherical radius will be incorrect by

$$\delta r = \sqrt{z^2 + R_\odot^2} - \sqrt{z^2 + y^2 + (R_\odot - x)^2}, \tag{6.49}$$

where x, y, z are the standard Galactic Cartesian coordinates with x positive towards the Galactic centre. At the edges of my cone with $l = 0°$ and $180°$ with $z = 1$ kpc, this corresponds to $\delta r \sim +0.175$ and -0.175 kpc respectively. This increases to $\delta r \sim +1.05$ and -1.19 kpc for $z \sim 10$ kpc. To test the impact of this on the model fits, I sample a halo profile with $n_{\rm H} = 3.724$ within the $b > 80°$ vertical cone and use the sample to estimate the mean spherical radius error. I find that the spherical radius of sources is underestimated by $\sim 0.3\%$ on average.

I also assumed my halo was spherically symmetric and ignored any flattening. Given the small high latitude region I have used, it is unlikely that I would have been able to fit a halo oblateness parameter separately. However, this assumption may still impact the inferred steepness of the halo. I test the impact of placing all sources at the Solar radius and assuming a spherical halo by regenerating my mock sample with the correct cylindrical radius with an oblate halo $q = 0.9$ (Mateu & Vivas, 2018). Using the same parameters as discussed in Section 6.4, I generate a mock catalogue with thin and thick disc scale lengths of $L_{\rm Tn} = 2.6$ and $L_{\rm Tk} = 3.6$ kpc (Jurić et al., 2008). I also include the cylindrical radius dependence of the spherical radius for the halo model.

The results are shown in Table 6.6 and orange dashed contours of Fig. 6.14. This has had a more significant impact on the fits with the posteriors $\sim 2-3$ sigma off the input parameter values in several cases. The disc scale heights and halo power law profile all have significant offsets from the true input parameters. It appears that the dominant effect is a level of source confusion between the components. The sample is no longer exactly representative of my assumed vertical exponential profile and power law halo but instead a marginalisation over this with radius. This leads to overestimated disc normalisation and scale heights and an overly steep halo.

6.9 Statistical and Systematic Uncertainties

I have produced fits to the observed data around the northern and southern Galactic poles and to mock samples to test the effects of limitations in the model. To provide results which are informative and usable, I quantify what my results mean for the true model parameters and their statistical and systematic uncertainties.

I do this by assuming the posterior distributions for all parameters in Table 6.4 are drawn from independent normal distributions. This enables us to parameterise all posteriors with a means (μ) and uncertainties (σ).

The first two columns of Table 6.4 provide the means and standard deviation uncertainties of the MCMC posteriors for the north and south fits to the *Gaia* data. I model the combined north/south posteriors as being drawn from a normal distribution with mean μ, uncertainty $\sigma^2_{\mathrm{sys,N/S}}$ convolved with an additional normal distribution, $\mathcal{N}(0,\sigma^2)$ providing the standard deviation uncertainty for each sample. Therefore the likelihood of the posteriors is

$$\begin{aligned}\log\mathcal{L} = &\sum_{i,\mathrm{North}} \log\mathcal{N}\left(x_i\,|\,\mu,\sigma^2_{\mathrm{North}}+\sigma^2_{\mathrm{sys,N/S}}\right)\\ &+\sum_{i,\mathrm{South}} \log\mathcal{N}\left(x_i\,|\,\mu,\sigma^2_{\mathrm{South}}+\sigma^2_{\mathrm{sys,N/S}}\right)\end{aligned} \tag{6.50}$$

where x_i are the posterior samples provided by the MCMC chains. I then maximise the log-likelihood with respect to μ, $\sigma_{\mathrm{sys,N/S}}$. The results are given in the third and forth columns of Table 6.4. The method I am using here is similar to inflating systematics until the chi-squared reaches a ‘reasonable’ value, however, I avoid defining an arbitrary chi-squared target by instead maximising the Gaussian log-likelihood.

Given a sample drawn from an equally weighted sum of Gaussian distributions with the same mean but different variance, the sample variance will be the mean of the individual component variances. Therefore, the statistical error for the *Gaia* data fits is the root-mean-square of the north and south fits

$$\sigma_{\mathrm{stat}} = \sqrt{\frac{\sigma^2_{\mathrm{North}}+\sigma^2_{\mathrm{South}}}{2}}. \tag{6.51}$$

This is given as the statistical error on μ in Table 6.4.

I play a similar game with the results of my systematic test runs from Section 6.8, however, in this case I know the true parameters because I provided the input parameters. The likelihood is given by

$$\log \mathcal{L} = \sum_{i,\mathrm{test}} \log \mathcal{N}\left(x_i \,|\, \mu_{\mathrm{true}}, \sigma^2_{\mathrm{test}} + \sigma^2_{\mathrm{sys,test}}\right) \tag{6.52}$$

where σ_{test} is the statistical uncertainty of the fit given by the standard deviation of the posterior and μ_{true} is the input parameter. I maximise this with respect to $\sigma_{\mathrm{sys,test}}$ to estimate the systematic uncertainty contribution from the given test. I then rescale the systematic errors by the measured Milky Way parameters, μ/μ_{true}, to estimate the systematic error on my fits to the *Gaia* data. In Table 6.4, I provide $\sigma_{\mathrm{sys,tests}}$ which is the maximum systematic error for the given parameter from my tests. I also state the test(s) which dominate the systematic uncertainty contribution. Where more than one test is listed, it is because they provided a similar systematic uncertainty to within 10%.

I recommend that anyone using my results should take the root mean square sum of all quoted uncertainties to obtain the total uncertainty on each parameter.

6.10 Discussion

Here, I interpret the results of the structural parameters of the Milky Way thin disc, thick disc and halo given in Table 6.4, comparing them with previous work as well as considering future developments.

6.10.1 Results

The most striking thing to notice when examining my results is the comparison between statistical and systematic uncertainties. In general, the total systematic uncertainty is more than an order of magnitude greater than statistical uncertainty. In some cases, it is over two orders of magnitude larger. This demonstrates two things. First, *Gaia* has ushered in an era where it is necessary to model systematic errors once considered insignificant. Rigorous systematic analysis of the kind I have performed is essential to provide accurate and reliable results. Secondly, the precision which can be achieved with *Gaia* data is impressive. I used a deliberately constrained sample of objects on the sky consisting of less than 0.1% of the entire *Gaia* catalogue and yet the precision on most parameters is more than an order of magnitude better than anything in the literature.

I infer a local stellar mass density for pre-compact object stars with of $\rho^*_{\mathrm{local}} = 3.66 \pm 0.03\,(\mathrm{stat}) \pm 0.52 \times 10^{-2}\,\mathrm{M}_\odot/\mathrm{pc}^3\,(\mathrm{sys})$. This is smaller than the value of $\rho^*_{\mathrm{local}} \approx 4.2 \times 10^{-2}\,\mathrm{M}_\odot/\mathrm{pc}^3$ as derived (without errors bars) in Flynn et al. 2006, using the *Hipparcos* and *Tycho* surveys, together with the Catalogue of Nearby Stars.

I compute a surface density of $\Sigma^*_{\mathrm{local}} = 23.42 \pm 0.09\,(\mathrm{stat}) \pm 3.22\,\mathrm{M}_\odot/\mathrm{pc}^2\,(\mathrm{sys})$. This is significantly smaller than Bovy et al. 2012a who estimate $30 \pm 1\,M_\odot/\mathrm{pc}^2$, as well as Flynn et al. 2006 who estimate $35.5 M_\odot/\mathrm{pc}^2$.

I expect that the most significant difference is that their works include compact objects in the stellar mass estimates. In particular, Bovy et al. 2012a use the initial mass function to infer the contribution from all sources similar to my work in Section 6.7.1. However, I account for stellar mass loss and only include stars with mass low enough that they would not have evolved into a compact object or gone supernova. Bovy et al. 2012a extrapolate to higher mass stars which will have evolved to a compact object without accounting for mass loss. This means their results will significantly overestimate the total stellar mass density for evolved stars which have undergone significant mass loss. I do not extrapolate my results to include compact objects as there is significant uncertainty over how much of the initial mass is kept in the final compact object remnant.

My relative thick-to-thin disc local density ratio sits between the values of Mackereth et al. 2017 and Jurić et al. 2008, although the systematic uncertainties on this due to extinction and magnitude error are quite large.

Ample past research has been dedicated to estimating the scale heights of the thin and thick disc. There is some discrepancy between studies, with thin disc estimates in the range $h_{\mathrm{Tn}} \sim 120 - 300$pc and thick disc in the range $h_{\mathrm{Tk}} \sim 500 - 1900$pc (Ak et al., 2008; Bilir et al., 2006a; Dobbie & Warren, 2020; Jurić et al., 2008; Kuijken & Gilmore, 1989; Mateu & Vivas, 2018; de Jong et al., 2010). I constrain the thin disc scale height as $h_{\mathrm{Tn}} = 260 \pm 3\,(\mathrm{stat}) \pm 26\,\mathrm{pc}\,(\mathrm{sys})$ and thick disc $h_{\mathrm{Tk}} = 693 \pm 7\,(\mathrm{stat}) \pm 121\,\mathrm{pc}\,(\mathrm{sys})$. My estimates are broadly in agreement with Jurić et al. 2008, de Jong et al. 2010 and Mateu & Vivas 2018 with reasonably strong constraints on the thin disc scale height (±26 pc), but the thick disc scale height is dominated by systematic uncertainty (±121 pc) due to the cylindrical radius dependence.

The power law profile of the halo has received substantial attention with typical estimates in the range $n_{\mathrm{H}} \sim 2.5 - 4.4$ (Cohen et al., 2017; Gould et al., 1996; Hernitschek et al., 2018; Iorio et al., 2018; Jurić et al., 2008; Mateu & Vivas, 2018; Newberg & Yanny, 2006; Saha, 1985; Smith et al., 2009a; Yanny et al., 2000; de Jong et al., 2010). My model sits in the middle of these estimates with $n_{\mathrm{H}} = 3.542 \pm 0.023\,(\mathrm{stat}) \pm 0.259\,(\mathrm{sys})$. Recent works have suggested the halo is better represented by a broken power law distribution (Deason et al., 2011; Fukushima et al., 2019; Thomas et al., 2018). Whilst I have focused in the inner halo by the definition of Fukushima et al. 2019 and truncated at $s = 160$ kpc, there is a wide range of distances inferred for the truncation, e.g. 25 kpc (Watkins et al., 2009) and 42 kpc for (Cohen et al., 2017). In reality I expect I am covering both sides of the break especially considering many stars from other author's samples will likely have made it into the *Gaia* astrometry sample. Another issue, extensively discussed in the literature, which can impact halo fits is accreted substructure (Bell et al., 2008). The current standard halo model is composed of stars from from GES (a major merger event ~ 8 Gyr ago Belokurov et al., 2018; Helmi et al., 2018), the "Splash" (in-situ stars kicked up by the merger event Belokurov et al., 2020a) and other accreted substrucuture such as Sagittarius and smaller streams. I masked problematic regions of the southern field in Section 6.6, however, there are likely to be more diffuse substructures which are not so easy to mask.

Another notable feature of my results is the north-south asymmetry across several parameters. I find the northern thin disc scale height is larger than the south at $\lesssim 10\%$. Dobbie & Warren 2020 found a similar asymmetry, although they claimed a much larger 25% difference. I also find that the southern halo is significantly shallower with a smaller power law exponent than the north which was also seen by Hernitschek et al. 2018. These effects may be caused by dynamical instabilities which asymmetrically excite the disc (Antoja et al., 2018; Widrow et al., 2012) and diffuse halo substructure such as Sagittarius which may contribute many more stars to the southern high latitude field (Vasiliev et al., 2021).

I used my halo local mass density and profile to estimate the total halo stellar mass, using a flat uniform density for $r < 1$ kpc in order to prevent the integral from diverging, obtaining $M^*_{\mathrm{Halo}} \sim 7.2 \times 10^8\,\mathrm{M}_\odot$. This is quite a rough estimate of the total halo stellar mass however my results do agree reasonably well with the broad range of literature results $M^*_{\mathrm{Halo}} \sim 2 - 14 \times 10^8\,M_\odot$ (Bell et al., 2008; Deason et al., 2011, 2019).

The dwarf fractions and absolute magnitude profiles were defined specifically for this work and were mainly fit as nuisance parameters in order to get at the spatial distribution of stars so I do not discuss these in detail here.

6.10.2 More General Models

My results have small statistical uncertainty compared with the dominant systematic uncertainty. This implies that the model I have chosen to fit is over-constrained by the data. The solution to this is to significantly increase the amount of freedom in the model until my systematic and statistical uncertainties are comparable.

Some generalizations of my model are obvious: inclusion of radial dependence; provision of free parameters on the radial disc profile; allowing a free Solar vertical position $z_\odot$; introduction of halo oblateness as a free parameter. I have not provided these freedoms due to numerical complications in the parallax error integral discussed in Sections 6.3.2 and 6.3.3.

I could also move away from the simplistic two-component disc model towards a continuous distribution of discs with age, as proposed by Bovy et al. 2012a. Alternatively, I could go for a much more data-driven approach and fit the source density at nodes with a smooth model such as a Gaussian Process to enable correlations between neighbouring points.

More ambitiously still, I could leverage the BP and RP photometry provided in *Gaia* EDR3 for 1.5 billion sources. Rather than using my simple magnitude model, I could directly infer the population of the HR diagram as a function of position in the Milky Way, from which the star formation histories and metallicity distributions could also be inferred. This will require selection functions for the BP and RP photometry samples which have not yet been produced.

Ultimately the model I have fit to the data is grossly oversimplified such that the results do not provide significant insight into the physical processes governing our Galaxy's formation and evolution. The suggestions provided here improve on this to some extent but it is still not clear what question these models would answer. What we would really like to do is combine this with kinematic information to infer a dynamical model of the Milky Way from which we could learn about the mass and distribution of Dark Matter. Or we would work with metal abundance data to learn something about how the elements came to be where we see them in the Galaxy today.

6.10.3 Extragalactic Component

In this work, I filtered extragalactic sources from my sample using cuts on colour and excess flux. However, another option is to add an additional component to the model for sources at infinite distance.

The spatial distribution of the extragalactic sources would simply be

$$\nu_{\mathrm{EG}}(l, b, s)\mathrm{d}V = \frac{1}{2\pi(1 - \sin(b_{\mathrm{min}}))}\, \delta(1/s)\, s^2\, \mathrm{d}l\, \mathrm{d}\sin(b)\, \mathrm{d}s \tag{6.53}$$

where extragalactic sources have zero parallax and are uniformly distributed across the sky for $b > b_{\mathrm{min}}$ or $b < -b_{\mathrm{min}}$ with $b_{\mathrm{min}} = 80°$ in this work.

However, this spatial model needs an apparent magnitude distribution for all extragalactic sources to which the selection function can be applied. This makes the model significantly more complicated as the apparent magnitude distribution is dependent on the distance and luminosity distribution which are different for quasars and galaxies. For this reason, I have not chosen to model the extragalactic population in this work. However, adding this additional component would be an interesting and worthwhile route forward, measuring the population of galaxies and quasars as a function of apparent magnitude with *Gaia* data.

6.11 My Milky Way Model

I have developed a method to fit the distribution of stars in the Milky Way using the Poisson likelihood function. My method correctly accounts for the sample selection function and parallax measurement uncertainty.

The method is used to fit the vertical distribution of stars with $|b| > 80°$. For the model I use two exponential disc components and a power-law halo. The data are also simultaneously fit with a four-piece exponential absolute-magnitude distribution. The efficacy of my method is demonstrated against a mock sample. By refitting the model parameters I demonstrate that the method produces results which are accurate to within the statistical uncertainties of the parameter posteriors.

I used the *Gaia* Early Data Release 3 (EDR3) photometry and astrometry to model the vertical distribution of stars in the Milky Way at the Solar radius. My sample includes the majority of stars with measured parallax in *Gaia* within 10° of the Galactic north and south Poles. My method formally accounts for parallax measurement uncertainty and the *Gaia* selection function.

I represent the vertical density of the thin and thick discs by exponentials with scale heights h_{Tn} and h_{Tk} respectively. The stellar halo density is a power-law of spherical radius, i.e. $\rho \propto r^{-n_{\mathrm{H}}}$. I thoroughly test possible sources of systematic uncertainty in my approach, in particular from oversimplifications of the model. This enables me to quantify the systematic uncertainty associated with all parameter estimates.

I find the scale height of the thin disc is $h_{\mathrm{Tn}} = 260 \pm 3\,(\mathrm{stat}) \pm 9 \pm 24\,\mathrm{pc}\,(\mathrm{sys})$, Here, the two levels of systematic error correspond to north-south asymmetry about the Galactic plane and simplifying model assumption (particularly the treatment of extinction and the assumption of halo spherical symmetry). The scale height of the thick disc is $h_{\mathrm{Tk}} = 693 \pm 7\,(\mathrm{stat}) \pm 10 \pm 121\,\mathrm{pc}\,(\mathrm{sys})$ where the larger systematic error contribution is introduced by my assumption that all sources have the same cylindrical polar radius as the Sun. For the stellar halo, I am able to constrain a power law profile of $n_{\mathrm{H}} = 3.542 \pm 0.023\,(\mathrm{stat}) \pm 0.160 \pm 0.204\,(\mathrm{sys})$.

I infer a local stellar mass density for non-compact object stars $\rho^*_{\mathrm{local}} = 3.66 \pm 0.03\,(\mathrm{stat}) \pm 0.39 \pm 0.34 \times 10^{-2}\,\mathrm{M}_\odot/\mathrm{pc}^3\,(\mathrm{sys})$ and surface density $\Sigma^*_{\mathrm{local}} = 23.42 \pm 0.09\,(\mathrm{stat}) \pm 1.4 \pm 2.9\,\mathrm{M}_\odot/\mathrm{pc}^2\,(\mathrm{sys})$. Whilst these values are lower than previous estimates (Bovy et al., 2012a; Flynn et al., 2006), this discrepancy may be explained by the absence of any contribution from compact object remnants to the total stellar mass. I have not included this due to the uncertain correction for stellar mass loss, itself not well accounted for in previous works.

I also find a north-south asymmetry with respect to the Galactic plane. The thin and thick disc scale heights are larger in the north, and the halo profile is shallower in the south. However, this asymmetry is only at the $\lesssim 10$ percent level, much less than the 25 percent claimed by Dobbie & Warren 2020.

The impressive information content of the *Gaia* data produces parameter estimates with significantly improved precision over previous studies, even for my sample using only a small region of the sky. Note, though, systematics now completely dominate the error budgets, meaning that I need better models to fully realise the potential of the *Gaia* data. As I discussed in Secion 6.3, the model I have applied is not an accurate representation of the Milky Way. The work I have performed does not provide dramatic new insight into the physics governing the formation and evolution of the Galaxy but it does provide an avenue through which that can be achieved with far greater accuracy and precision than ever before.

The approach taken here demonstrates the power of information available from *Gaia* which has yet to be unlocked. There is a substantial prize available for controlling the systematic uncertainties involved with modelling the *Gaia* data.

7

Conclusions

My PhD aimed to model the spatial and velocity distribution of stars in the Milky Way. This 6D distribution of stars is hugely important for our understanding of the Milky Way. It facilitates Galactic archaeology, studying how the Galaxy formed and evolved to its current state. It is hugely valuable for stellar physics, matching populations to their kinematic and inferring the intrinsic luminosity distribution of stars. It can provide detailed tests of cosmological models such as comparing the numbers of Milky Way satellite galaxies against predictions. Perhaps most excitingly, it allows us to model the gravitational potential of the Milky Way and therefore estimate the distribution of missing mass, dark matter, which is important for direct detection experiments.

The *Gaia* mission has published positions, parallaxes and proper motions for 1.5 billion sources and radial velocities for a much smaller but no less impressive seven million stars. This makes the problem appear incredibly tractable, one simply needs to fit a density model to the number density of observed sources. However, there are two dominant obstacles which significantly complicate the issue.

If no stars are observed in a region of parameter space, is this because there are no stars there or because the telescope was unable to observe them? We need the answer to this question to reliably fit the distribution of stars to the data. The answer is provided by the selection function.

The second problem is that parallax uncertainties are significant for most stars observed by *Gaia* and modelling spatial distributions from parallax distances is a statistically complicated task. I demonstrated this in Chapter 2 where I showed that a small systematic bias in the parallax measurements can significantly change the inferred tilt of the velocity ellipsoid. This systematic is important – parallax biases can completely change the conclusions inferred about the velocity ellipsoid alignment and the Milky Way potential.

7.1 My solutions

I first solved the local velocity ellipsoid problem by using the work of Schönrich et al. 2019 who provide posterior distance estimates to all stars in the *Gaia* radial velocity sample. They fit a geometric parallax zero-point offset to the *Gaia* RVS sample and, using a prior Milky Way density model, evaluate posterior distance distributions to all stars. In Chapter 2 I used this data to measure the velocity ellipsoid in the solar neighbourhood which is close to spherical alignment with deviations consistent with the contribution from baryonic matter in the Milky Way disk. This is consistent with the canonical picture of a close to spherical Milky Way mass distribution with an additional disk component.

Whilst Schönrich et al. 2019 distance estimates were hugely valuable for this study, they had some significant limitations. The parallax zero-point offset could only be modelled for sources with measured radial velocity which limited the sample to the seven million sources in *Gaia* RVS. Their method included fitting a distance-dependent selection function for the catalogue, however, this didn't factor in the dramatic variations of the selection function across the sky. Furthermore, they include an *a priori* density model such that any spatial model with Milky Way data could be influenced by the prior.

To model the spatial distribution of stars in the Milky Way I would need selection functions for *Gaia* science samples and complementary spectrographs and a method to formally account for parallax uncertainties.

I presented my method for selection functions of multi-fibre spectrographs in Chapter 3. I used a Poisson likelihood density modelling method to fit the distribution of sources in the spectrograph sample against a colour-magnitude complete catalogue. Using the field-by-field nature of spectrograph observations I incorporated spatial dependence on the sky. I applied isochrones to transform the selection function into intrinsic source parameters. This selection function reproduced the data extremely well, passing Kolmogorov-Smirnov sample tests.

However, my method has its limitations. Firstly it could not be applied the the full *Gaia* source catalogue as it required a more complete catalogue to compare against. *Gaia* subsets are all-sky samples which means the field-by-field method was not applicable. The Poisson likelihood method also didn't leverage our knowledge that the spectrograph isn't just less complete than the photometric catalogue, it is a direct subset.

In Chapter 4 I introduced new methods for modelling the *Gaia* source catalogue and science subsets. The observing strategy of the *Gaia* satellite presents an opportunity to evaluate the source catalogue selection function leveraging information from the scanning law. For science subsets of *Gaia* I applied the method developed in Boubert & Everall 2021. The model uses spherical needlets to encode on-sky spatial dependence and a Gaussian process prior across colour-apparent magnitude bins. Improving on Chapter 3, I used a Binomial likelihood function which includes the information that the subset is drawn from the source catalogue. Using Binomial p-value tests I demonstrated that the new selection function models are consistent with the data down to 2 degree scales on the sky.

Using *Gaia* astrometry to model the spatial distribution of stars requires an in depth understanding of the astrometric fitting procedure. I developed the Astrometric Spread Function (ASF), the expected *Gaia* astrometry covariance of a point source moving linearly with respect to the Solar system barycenter, which I introduced in Chapter 5. The ASF predicts the *Gaia* DR2 published covariances beautifully well except in crowded regions where the satellite struggles to measure centroid positions for individual sources.

Using my comprehensive toolkit for *Gaia* data, I have fitted the vertical structure of stars in the Milky Way at the solar radius. I developed a method to fit the tracer density of stars formally accounting for parallax uncertainty and survey incompleteness through the selection function. I demonstrated that this performs extremely well on *Gaia*-like mock samples. Applying this to a small subset of *Gaia*, the tight statistical uncertainties in my posterior fits demonstrate the as-of-yet unexplored power of the *Gaia* data with orders of magnitude improvement in precision over previous studies. However, better models are needed to reduce the systematic uncertainties to obtain comparable accuracy.

I measured the scale height of the thin disk, $h_{\mathrm{Tn}} = 260 \pm 3\,(\mathrm{stat}) \pm 26\,\mathrm{pc}\,(\mathrm{sys})$, and thick disk, $h_{\mathrm{Tk}} = 693 \pm 7\,(\mathrm{stat}) \pm 121\,\mathrm{pc}\,(\mathrm{sys})$, and the radial power law exponent of the spherical halo, $n_{\mathrm{H}} = 3.542 \pm 0.023\,(\mathrm{stat}) \pm 0.259\,(\mathrm{sys})$. These are consistent with various previous works although the halo profile varies heavily depending on the choice of truncation radius.

I also used my results to infer the local stellar mass density, $\rho^*_{\mathrm{local}} = 3.66 \pm 0.03\,(\mathrm{stat}) \pm 0.52\,(\mathrm{sys}) \times 10^{-2}\,\mathrm{M}_\odot/\mathrm{pc}^3$, and surface density, $\Sigma^*_{\mathrm{local}} = 23.42 \pm 0.09\,(\mathrm{stat}) \pm 3.22\,(\mathrm{sys})\,\mathrm{M}_\odot/\mathrm{pc}^2$. My result are consistent with the bulk of previous work to within total uncertainties and demonstrate the first pure-*Gaia* photo-astrometric tracer density model of the Milky Way. These values are lower than previous estimates as I don't include mass contributions from sources which have experienced supernovae and evolved into compact objects.

I found an asymmetry between the north and south of the Milky Way disk and halo consistent with previous works but less extreme at the $\lesssim 10\%$ level. This is likely the result of disk disequilibrium and halo substructure.

7.2 Value for the Community

I have introduced an armory of tools for analysing *Gaia* data. Through my PhD I have produced and contributed to the open source GitHub repositories:

- SEESTAR: For estimating selection functions for multi-fibre spectrographs using my method from Chapter 3.
- SELECTIONFUNCTIONS: For retrieving selection probabilities for stars as a function of position on the sky, colour and apparent magnitude using the selection functions from Chapter 4.
- SCANNINGLAW: For retrieving the ASF covariance of any source as explained in Chapter 5 and learning the *Gaia* scan times, directions and probabilities expected for a star.
- MWTRACE: My code used in Chapter 6 to evaluate the vertical tracer density of stars in the Milky Way.

These tools enable the user to produce generative models of *Gaia* and multi-fibre spectrograph data for predictions of the phase-space structure of the Milky Way. This is a vital aspect of forward modelling. I applied this technique to fit the Galactic tracer density in Chapter 6 and produced *Gaia*-like mock catalogues akin to Aurigaia (Grand et al., 2018).

7.3 The Ones that Got Away

Whilst I have achieved many scientific goals in my PhD, there are some problems which remain unsolved. I hope these problems will keep future PhD students and GaiaUnlimited[1] busy.

7.3.1 General Spatial Models

My tracer density model demonstrated the power of the *Gaia* data but was not able to fully leverage it due to an over-simplified model. There are several ways this could be improved which I discussed in Section 6.10. A key challenge is integrating over parallax error for every source at every iteration of the optimization process. Any model used needs to be tractable to efficiently integrate over.

7.3.2 Crowded Regions

Gaia doesn't behave well in crowded regions of the sky. The astrometric solution relies on centroid fits to the images of stars on the *Gaia* focal plane. If multiple sources are contributing to the flux distribution this can become significantly more challenging. Both the ASF and selection functions are significantly dependent on the number density of sources in a region of space. In the case of the selection function I have managed to model this for individual stars, however, this introduces inter-source correlations which means a source's selection function can no longer be fully defined by it's observable properties alone. This is the key un-solved challenge of *Gaia* selection functions.

7.3.3 Stellar Type Selection Functions

Particular stellar types are often extremely valuable for understanding the properties of the Milky Way and indeed nearby galaxies. Henrietta Leavitt's modelling of the Cepheid variable period luminosity relation which facilitated the resolution of "The Great Debate" is a beautiful example. These types of objects enable us to extrapolate geometric distance information from parallax measurements to much greater distances. However, in order to use them to fit the structure of the Milky Way we need their selection function.

As I discussed in Section 4.7, this cannot be modelled simply as a subset of a parent sample as most sources in the parent will not be stars of the given type. We need the probability of selection given that a source is, for example, a Cepheid variable. As of yet we do not know how to achieve this other than comparing with surveys such as OGLE (Udalski et al., 1992) which are expected to be significantly more complete for variable stars in the Milky Way bulge, LMC and SMC.

[1] GaiaUnlimited is a new collaboration seeking to model the completeness limits of catalogues published by *Gaia*, building on the work of the *Completeness of the Gaia-verse* collaboration.

7.3.4 Binary Systems and Exoplanets

The *Gaia* mission will introduce a detection method for finding binary systems and exoplanets which has the capacity to outshine all previous techniques, astrometric wobble. The orbit of a star around the system barycentre due to an unseen companion can generate a detectable deviation from expected single object astrometry (Penoyre et al., 2020), a phenomenon previously seen in the *Hipparcos* data (Lindegren et al., 1997).

To date ~ 20 000 close binary systems (Price-Whelan et al., 2020) and 4 331 exoplanets[2] have been discovered, most from variations in stellar radial velocity and binary eclipses or exoplanet transits. Using astrometric wobble, *Gaia* is capable of finding 60 million binary systems including many Brown Dwarf and Black Hole companions (Andrews et al., 2019). Similarly, it is expected that *Gaia* will be able to find over 21 000 exoplanets (Perryman et al., 2014).

The potential of this has already been demonstrated by Belokurov et al. 2020b using *Gaia*'s published goodness-of-fit statistic, RUWE. Estimating the sensitivity of *Gaia* to binaries is extremely hard due to the complex Image Parameter Determination (IPD Fabricius et al., 2021) and astrometry pipeline (Lindegren et al., 2012). Given a binary system at a known distance with known orbital parameters, what is the probability that *Gaia*'s astrometry would be significantly affected by the binary motion? This questions pertains to the binary selection function and will be important when using *Gaia* astrometry to estimate the distribution of binary systems and exoplanets.

7.4 Back Down to Earth

My journey into the *Gaia*-verse has been a fascinating learning experience about the Milky Way's past, present and future. But the most exciting thing for me has been discovering how to distill the full information content from the vast wealth of data we have available. Galactic dynamics is in a data-dominated regime where we have not perfected the statistical and computational techniques required to fully exploit the huge volume and incredible precision of data available.

There are many avenues to take building on the work I have done, some of which I have suggested. I hope that others in the field will pick up on them. As for me, my next steps will be in biostatistics for genomics which has undergone an arguably even more significant data revolution in the last decade. I hope the techniques I have learned and developed to model the building blocks of the Galaxy will prove similarly useful for modelling the building blocks of life.

[2] https://exoplanetarchive.ipac.caltech.edu/index.html

References

Abolfathi B., et al., 2018, ApJS, 235, 42

Adelmam-McCarthy J. K., et al., 2008, ApJS, 175, 297

Agertz O., et al., 2020, arXiv e-prints, p. arXiv:2006.06008

Ak T., Bilir S., Ak S., Eker Z., 2008, New Astron., 13, 133

An J., Evans N. W., 2016, ApJ, 816, 35

Anders F., et al., 2019, arXiv e-prints, p. arXiv:1904.11302

Andrae R., et al., 2018, A&A, 616, A8

Andrews J. J., Breivik K., Chatterjee S., 2019, ApJ, 886, 68

Antoja T., et al., 2018, Nature, 561, 360

Arenou F., et al., 2018, A&A, 616, A17

Astropy Collaboration et al., 2013, A&A, 558, A33

Astropy Collaboration et al., 2018, AJ, 156, 123

Bailer-Jones C. A. L., 2015a, PASP, 127, 994

Bailer-Jones C. A. L., 2015b, PASP, 127, 994

Bailer-Jones C. A. L., Rybizki J., Fouesneau M., Mantelet G., Andrae R., 2018, AJ, 156, 58

Bailer-Jones C. A. L., Fouesneau M., Andrae R., 2019, MNRAS, 490, 5615

Baldi P., Kerkyacharian G., Marinucci D., Picard D., 2006, arXiv Mathematics e-prints, p. math/0606154

Bell E. F., et al., 2008, ApJ, 680, 295

Belokurov V., et al., 2006, ApJ, 642, L137

Belokurov V., Erkal D., Deason A. J., Koposov S. E., De Angeli F., Evans D. W., Fraternali F., Mackey D., 2017, MNRAS, 466, 4711

Belokurov V., Erkal D., Evans N. W., Koposov S. E., Deason A. J., 2018, MNRAS, 478, 611

Belokurov V., Sanders J. L., Fattahi A., Smith M. C., Deason A. J., Evans N. W., Grand R. J. J., 2020a, MNRAS, 494, 3880

Belokurov V., et al., 2020b, MNRAS, 496, 1922

Bennett M., Bovy J., 2019, MNRAS, 482, 1417

Bessel F. W., 1838, MNRAS, 4, 152

Bilir S., Karaali S., Ak S., Yaz E., Hamzaoğlu E., 2006a, New Astron., 12, 234

Bilir S., Karaali S., Güver T., Karataş Y., Ak S. G., 2006b, Astronomische Nachrichten, 327, 72

Binney J., McMillan P., 2011, MNRAS, 413, 1889

Binney J., Tremaine S., 2008, Galactic Dynamics: Second Edition. Princeton University Press

Binney J., Gerhard O., Spergel D., 1997, MNRAS, 288, 365

Binney J., et al., 2014, MNRAS, 439, 1231

Blanton M. R., et al., 2017, AJ, 154, 28

Bond N. A., et al., 2010, ApJ, 716, 1

Boubert D., Everall A., 2020, MNRAS, 497, 4246

Boubert D., Everall A., 2021, *submitted*

Boubert D., et al., 2019, MNRAS, 486, 2618

Boubert D., Everall A., Holl B., 2020, MNRAS, 497, 1826

Boubert D., Fraser J., Everall A., 2021a, *in prep*

Boubert D., Everall A., Fraser J., Gration A., Holl B., 2021b, MNRAS, 501, 2954

Bovy J., 2017, MNRAS, 468, L63

Bovy J., Rix H.-W., Hogg D. W., 2012a, ApJ, 751, 131

Bovy J., Rix H.-W., Liu C., Hogg D. W., Beers T. C., Lee Y. S., 2012b, ApJ, 753, 148

Bovy J., et al., 2014, ApJ, 790, 127

Bovy J., Rix H.-W., Green G. M., Schlafly E. F., Finkbeiner D. P., 2016a, The Astrophysical Journal, 818, 130

Bovy J., Rix H.-W., Schlafly E. F., Nidever D. L., Holtzman J. A., Shetrone M., Beers T. C., 2016b, ApJ, 823, 30

Boyle W. S., Smith G. E., 1970, The Bell System Technical Journal, 49, 587

Bradley J., 1727, Philosophical Transactions of the Royal Society of London Series I, 35, 637

Bradley J., 1748, Philosophical Transactions of the Royal Society of London Series I, 45, 1

Bramich D. M., 2018, A&A, 618, A44

Bressan A., Marigo P., Girardi L., Salasnich B., Dal Cero C., Rubele S., Nanni A., 2012, MNRAS, 427, 127

Brown A. G. A., et al., 2016, A&A, 595, A2

Büdenbender A., van de Ven G., Watkins L. L., 2015, MNRAS, 452, 956

Bullock J. S., Johnston K. V., 2005, ApJ, 635, 931

Burrows C. J., Holtzman J. A., Faber S. M., Bely P. Y., Hasan H., Lynds C. R., Schroeder D., 1991, ApJ, 369, L21

Carrasco J. M., et al., 2016, A&A, 595, A7

Casertano S., et al., 2008, A&A, 482, 699

Chambers K. C., et al., 2016, arXiv e-prints, p. arXiv:1612.05560

Chen Y., Girardi L., Bressan A., Marigo P., Barbieri M., Kong X., 2014, MNRAS, 444, 2525

Chen Y., Bressan A., Girardi L., Marigo P., Kong X., Lanza A., 2015, MNRAS, 452, 1068

Chen B.-Q., Liu X.-W., Yuan H.-B., Xiang M.-S., Huang Y., Wang C., Zhang H.-W., Tian Z.-J., 2018, Monthly Notices of the Royal Astronomical Society, 476, 3278

Cirasuolo M., et al., 2014, in Ramsay S. K., McLean I. S., Takami H., eds, Vol. 9147, Ground-based and Airborne Instrumentation for Astronomy V. p. 91470N, doi:10.1117/12.2056012, http://proceedings.spiedigitallibrary.org/proceeding.aspx?doi=10.1117/12.2056012

Cohen J. G., Sesar B., Bahnolzer S., He K., Kulkarni S. R., Prince T. A., Bellm E., Laher R. R., 2017, ApJ, 849, 150

Cropper M., et al., 2018, A&A, 616, A5

Crowley C., et al., 2016, A&A, 595, A6

Cui X.-Q., et al., 2012, Research in Astronomy and Astrophysics, 12, 1197

Dalton G., et al., 2012, in McLean I. S., Ramsay S. K., Takami H., eds, Vol. 8446, Ground-based and Airborne Instrumentation for Astronomy IV. p. 84460P, doi:10.1117/12.925950, http://proceedings.spiedigitallibrary.org/proceeding.aspx?doi=10.1117/12.925950

Das P., Binney J., 2016, MNRAS, 460, 1725

Das P., Williams A., Binney J., 2016, MNRAS, 463, 3169

Davis M., Efstathiou G., Frenk C. S., White S. D. M., 1985, ApJ, 292, 371

De Silva G., et al., 2015, MNRAS, 449, 2604

Deason A. J., Belokurov V., Evans N. W., 2011, MNRAS, 416, 2903

Deason A. J., Belokurov V., Koposov S. E., Rockosi C. M., 2014, ApJ, 787, 30

Deason A. J., Belokurov V., Sanders J. L., 2019, MNRAS, 490, 3426

Dehnen W., 1998, AJ, 115, 2384

Dehnen W., Binney J., 1998, MNRAS, 294, 429

Dobbie P. S., Warren S. J., 2020, The Open Journal of Astrophysics, 3, 5

Eddington A. S., 1915, MNRAS, 76, 37

El-Badry K., Rix H.-W., Heintz T. M., 2021, MNRAS,

Evans N. W., Sanders J. L., Williams A. A., An J., Lynden-Bell D., Dehnen W., 2016, MNRAS, 456, 4506

Evans D. W., et al., 2018, A&A, 616, A4

Everall A., Boubert D., 2021, *submitted*

Everall A., Das P., 2020, MNRAS, 493, 2042

Everall A., Evans N. W., Belokurov V., Schönrich R., 2019, MNRAS, 489, 910

Everall A., Evans N. W., Belokurov V., Boubert D., Grand R., 2021a, *submitted*

Everall A., Belokurov V., Evans N. W., Boubert D., Grand R., 2021b, *submitted*

Everall A., Boubert D., Koposov S. E., Smith L., Holl B., 2021c, MNRAS, 502, 1908

Eyer L., et al., 2017, arXiv e-prints, p. arXiv:1702.03295

Fabricius C., et al., 2021, A&A, 649, A5

Fantin N. J., et al., 2021, arXiv e-prints, p. arXiv:2103.14721

Flynn C., Holmberg J., Portinari L., Fuchs B., Jahreiß H., 2006, MNRAS, 372, 1149

Foreman-Mackey D., Hogg D. W., Lang D., Goodman J., 2013, Publications of the Astronomical Society of the Pacific, 125, 306

Fraser J., Boubert D., Everall A., 2021, *in prep*

Fukushima T., et al., 2019, PASJ, 71, 72

Gaia Collaboration et al., 2016, A&A, 595, A1

Gaia Collaboration et al., 2018a, A&A, 616, A1

Gaia Collaboration et al., 2018b, A&A, 616, A11

Gaia Collaboration et al., 2021a, A&A, 649, A1

Gaia Collaboration et al., 2021b, A&A, 649, A6

Geller D., Mayeli A., 2010, Theory Signal Image Process, 9, 1

Gilmore G., Reid N., 1983, MNRAS, 202, 1025

Gilmore G., et al., 2012a, The Messenger, 147, 25

Gilmore G., et al., 2012b, The Messenger, 147, 25

Gómez F. A., Minchev I., O'Shea B. W., Beers T. C., Bullock J. S., Purcell C. W., 2013, MNRAS, 429, 159

Górski K. M., Hivon E., Banday A. J., Wandelt B. D., Hansen F. K., Reinecke M., Bartelmann M., 2005, ApJ, 622, 759

Gould A., Bahcall J. N., Flynn C., 1996, ApJ, 465, 759

Grand R. J. J., et al., 2017, MNRAS, 467, 179

Grand R. J. J., et al., 2018, MNRAS, 481, 1726

Gravity Collaboration et al., 2018, A&A, 615, L15

Green G., 2018, The Journal of Open Source Software, 3, 695

Green G. M., et al., 2014, ApJ, 783, 114

Green G., Ford Schlafly E., Finkbeiner D. P., 2015, American Astronomical Society Meeting Abstracts #225, 225, 256.19

Green G. M., Schlafly E., Zucker C., Speagle J. S., Finkbeiner D., 2019, ApJ, 887, 93

Griffin S., 2021, BMJ, 372

Groenewegen M., 2021, arXiv e-prints, p. arXiv:2106.08128

Gunn J. E., et al., 1998, AJ, 116, 3040

Hagen J. H. J., Helmi A., de Zeeuw P. T., Posti L., 2019, A&A, 629, A70

Halley E., 1717, Philosophical Transactions of the Royal Society of London Series I, 30, 736

Harrison D. L., 2011, Experimental Astronomy, 31, 157

Helmi A., Babusiaux C., Koppelman H. H., Massari D., Veljanoski J., Brown A. G. A., 2018, Nature, 563, 85

Henderson T., 1840, Mem. RAS, 11, 61

Hernandez X., Valls-Gabaud D., Gilmore G., 2000, MNRAS, 316, 605

Hernitschek N., et al., 2018, ApJ, 859, 31

Høg E., 2011, Baltic Astronomy, 20, 221

Hogg D. W., Bovy J., Lang D., 2010, arXiv e-prints, p. arXiv:1008.4686

Holl B., 2021, *in prep*

Holl B., Hobbs D., Lindegren L., 2010, in Klioner S. A., Seidelmann P. K., Soffel M. H., eds, Vol. 261, Relativity in Fundamental Astronomy: Dynamics, Reference Frames, and Data Analysis. pp 320–324, doi:10.1017/S1743921309990573

Holl B., et al., 2018, A&A, 618, A30

Holl B., Perryman M., Lindegren L., Segransan D., Raimbault M., 2021, *submitted*

Huang Y., Yuan H., Beers T. C., Zhang H., 2021, ApJ, 910, L5

Hubble E. P., 1925, The Observatory, 48, 139

Iorio G., Belokurov V., 2019, MNRAS, 482, 3868

Iorio G., Belokurov V., Erkal D., Koposov S. E., Nipoti C., Fraternali F., 2018, MNRAS, 474, 2142

Ivezić Ž., et al., 2008, ApJ, 684, 287

Jeans J. H., 1922, MNRAS, 82, 122

Jenkins L. F., 1952, General catalogue of trigonometric stellar parallaxes. Yale University Observatory

Joshi Y. C., 2007, MNRAS, 378, 768

Jurić M., et al., 2008, ApJ, 673, 864

Kains N., et al., 2017, ApJ, 843, 145

Kapteyn J. C., 1922a, Bull. Astron. Inst. Netherlands, 1, 69

Kapteyn J. C., 1922b, ApJ, 55, 302

Katz D., et al., 2019, A&A, 622, A205

Katz D., Gomez A., Haywood M., Snaith O., Di Matteo P., 2021, arXiv e-prints, p. arXiv:2102.02082

Kawata D., Baba J., Ciucă I., Cropper M., Grand R. J. J., Hunt J. A. S., Seabroke G., 2018, MNRAS, 479, L108

Kazantzidis S., Bullock J. S., Zentner A. R., Kravtsov A. V., Moustakas L. A., 2008, ApJ, 688, 254

Kepler J., Brahe T., Eckebrecht P., 1627, Tabvlae Rudolphinae, qvibvs astronomicae scientiae, temporum longinquitate collapsae restauratio continetur, doi:10.3931/e-rara-8742.

Kervella P., Arenou F., Mignard F., Thévenin F., 2019, A&A, 623, A72

Khan S., et al., 2019, arXiv:1904.05676

Kilic M., Munn J. A., Harris H. C., von Hippel T., Liebert J. W., Williams K. A., Jeffery E., DeGennaro S., 2017, ApJ, 837, 162

Klüter J., Bastian U., Demleitner M., Wambsganss J., 2018, A&A, 620, A175

Klüter J., Bastian U., Wambsganss J., 2020, A&A, 640, A83

Koposov S. E., et al., 2020, MNRAS, 491, 2465

Kordopatis G., et al., 2013, AJ, 146, 134

Krone-Martins A., et al., 2013, A&A, 556, A102

Kroupa P., 2001, MNRAS, 322, 231

Kuijken K., Gilmore G., 1989, MNRAS, 239, 605

Kunder A., et al., 2017, AJ, 153, 75

Lang D., 2014, AJ, 147, 108

Laplace P. S., 1921, Essai philosophique sur les probabilités.. Maîtres de la pensée scientifique, Gauthier-Villar, Paris

Laporte C. F. P., Minchev I., Johnston K. V., Gómez F. A., 2019, MNRAS, 485, 3134

Leavitt H. S., Pickering E. C., 1912, Harvard College Observatory Circular, 173, 1

Lindegren L., Perryman M. A. C., 1996, A&AS, 116, 579

Lindegren L., et al., 1997, A&A, 323, L53

Lindegren L., Lammers U., Hobbs D., O'Mullane W., Bastian U., Hernández J., 2012, A&A, 538, A78

Lindegren L., et al., 2016, A&A, 595, A4

Lindegren L., et al., 2018, A&A, 616, A2

Lindegren L., et al., 2021a, A&A, 649, A2

Lindegren L., et al., 2021b, A&A, 649, A4

Lorimer D. R., Bailes M., McLaughlin M. A., Narkevic D. J., Crawford F., 2007, Science, 318, 777

Luri X., et al., 2018, A&A, 616, A9

Lutz T. E., Kelker D. H., 1973, PASP, 85, 573

Mackereth J. T., et al., 2017, MNRAS, 471, 3057

Majewski S. R., et al., 2017, The Astronomical Journal, 154, 94

Malmquist K. G., 1922, Meddelanden fran Lunds Astronomiska Observatorium Serie I, 100, 1

Marigo P., Girardi L., Bressan A., Groenewegen M., Silva L., Granato G., 2008, A&A, 482, 883

Marinucci D., et al., 2008, MNRAS, 383, 539

Marshall D. J., Robin A. C., Reylé C., Schultheis M., Picaud S., 2006, A&A, 453, 635

Martig M., Minchev I., Ness M., Fouesneau M., Rix H.-W., 2016, ApJ, 831, 139

Mateu C., Vivas A. K., 2018, MNRAS, 479, 211

McConnachie A. W., 2012, AJ, 144, 4

McGill P., Smith L. C., Evans N. W., Belokurov V., Lucas P. W., 2019, MNRAS, 487, L7

McGill P., Everall A., Boubert D., Smith L. C., 2020, MNRAS, 498, L6

Michalik D., Lindegren L., Hobbs D., 2015a, Astronomy & Astrophysics, 574, A115

Michalik D., Lindegren L., Hobbs D., Butkevich A. G., 2015b, A&A, 583, A68

Miles R., 2007, Journal of the British Astronomical Association, 117, 172

Mints A., Hekker S., 2019, Astronomy & Astrophysics, 621, A17

Miralda-Escude J., 1996, ApJ, 470, L113

More J. J., Garbow B. S., Hillstrom K. E., 1980, doi:10.2172/6997568

Muraveva T., Delgado H. E., Clementini G., Sarro L. M., Garofalo A., 2018, MNRAS, 481, 1195

Nandakumar G., Schultheis M., Hayden M., Rojas-Arriagada A., Kordopatis G., Haywood M., 2017, Astronomy & Astrophysics, 606, A97

Newberg H. J., Yanny B., 2006, in Journal of Physics Conference Series. pp 195–204 (arXiv:astro-ph/0507671), doi:10.1088/1742-6596/47/1/024

Nitschai M. S., Cappellari M., Neumayer N., 2020, MNRAS, 494, 6001

Norris R. P., 2016, Publ. Astron. Soc. Australia, 33, e039

Pedregosa F., et al., 2011, Journal of Machine Learning Research, 12, 2825

Penoyre Z., Belokurov V., Wyn Evans N., Everall A., Koposov S. E., 2020, MNRAS, 495, 321

Perryman M. A. C., et al., 1997, A&A, 500, 501

Perryman M. A. C., et al., 2001, A&A, 369, 339

Perryman M., Hartman J., Bakos G. Á., Lindegren L., 2014, ApJ, 797, 14

Pickering E. C., 1908, Annals of Harvard College Observatory, 50, 1

Planck Collaboration et al., 2020, A&A, 641, A6

Price-Whelan A. M., et al., 2020, ApJ, 895, 2

Prieto C. A., et al., 2008, Astronomische Nachrichten, 329, 1018

Queiroz A. B. A., et al., 2018, MNRAS, 476, 2556

Ranalli P., Hobbs D., Lindegren L., 2018, A&A, 614, A30

Read J. I., 2014, Journal of Physics G Nuclear Physics, 41, 063101

Recio-Blanco A., et al., 2014, A&A, 567, A5

Reid N., 1982, MNRAS, 201, 51

Reid N., Gilmore G., 1982, MNRAS, 201, 73

Ren F., Chen X., Zhang H., de Grijs R., Deng L., Huang Y., 2021, ApJ, 911, L20

Riello M., et al., 2018, A&A, 616, A3

Riello M., et al., 2021, A&A, 649, A3

Riess A. G., et al., 2018, ApJ, 855, 136

Riess A. G., Casertano S., Yuan W., Macri L. M., Scolnic D., 2019, ApJ, 876, 85

Riess A. G., Casertano S., Yuan W., Bowers J. B., Macri L., Zinn J. C., Scolnic D., 2021, ApJ, 908, L6

Rix H.-W., et al., 2021, arXiv e-prints, p. arXiv:2106.07653

Robin A., Reylé C., Derrière S., Picaud S., 2003, A&A, 409, 523

Rowell N., et al., 2021, A&A, 649, A11

Ruiz-Lara T., Gallart C., Bernard E. J., Cassisi S., 2020, Nature Astronomy, 4, 965

Rybizki J., Green G., Rix H.-W., Demleitner M., Zari E., Udalski A., Smart R. L., Gould A., 2021a, arXiv e-prints, p. arXiv:2101.11641

Rybizki J., Rix H.-W., Demleitner M., Bailer-Jones C. A. L., Cooper W. J., 2021b, MNRAS, 500, 397

Saha A., 1985, ApJ, 289, 310

Saha K., Pfenniger D., Taam R. E., 2013, ApJ, 764, 123

Sale S. E., Magorrian J., 2014, Monthly Notices of the Royal Astronomical Society, 445, 256

Sanders J., Binney J., 2015, MNRAS, 449, 3479

Sartoretti P., et al., 2018, A&A, 616, A6

Schlesinger F., Jenkins L., Observatory Y. U., 1935, General Catalogue of Stellar Parallaxes: Compiled at Yale University Observatory. Yale University Observatory, https://books.google.co.uk/books?id=3uooAAAAYAAJ

Schönrich R., 2012, MNRAS, 427, 274

Schönrich R., Aumer M., 2017, MNRAS, 472, 3979

Schönrich R., Binney J., 2009, MNRAS, 396, 203

Schönrich R., Binney J., Dehnen W., 2010, MNRAS, 403, 1829

Schönrich R., McMillan P., Eyer L., 2019, MNRAS, 487, 3568

Schwarz G., 1978, Annals of Statistics, 6, 461

Scodeller S., Rudjord Ø., Hansen F. K., Marinucci D., Geller D., Mayeli A., 2011, ApJ, 733, 121

Sellwood J. A., 2014, Rev. Mod. Phys., 86, 1

Shapley H., 1918, PASP, 30, 42

Sharma S., Bland-Hawthorn J., Johnston K., Binney J., 2011, ApJ, 730, 3

Shu F. H., 1969, ApJ, 158, 505

Shu Y., Koposov S. E., Evans N. W., Belokurov V., McMahon R. G., Auger M. W., Lemon C. A., 2019, MNRAS, 489, 4741

Silverwood H., Sivertsson S., Steger P., Read J. I., Bertone G., 2016, MNRAS, 459, 4191

Sivertsson S., Silverwood H., Read J. I., Bertone G., Steger P., 2018, MNRAS, 478, 1677

Skowron D. M., et al., 2019, Acta Astron., 69, 305

Skrutskie M., et al., 2006, AJ, 131, 1163

Smart R. L., Nicastro L., 2014, A&A, 570, A87

Smith M. C., et al., 2009a, MNRAS, 399, 1223

Smith M. C., Evans N. W., An J. H., 2009b, ApJ, 698, 1110

Snaith O., Haywood M., Di Matteo P., Lehnert M. D., Combes F., Katz D., Gómez A., 2015, A&A, 578, A87

Sozzetti A., Giacobbe P., Lattanzi M. G., Micela G., Morbidelli R., Tinetti G., 2014, MNRAS, 437, 497

Springel V., et al., 2005, Nature, 435, 629

Stassun K. G., Torres G., 2021, ApJ, 907, L33

Steinmetz M., et al., 2006, AJ, 132, 1645

Stonkutė E., et al., 2016, MNRAS, 460, 1131

Strömberg G., 1927, ApJ, 65, 238

Tang J., Bressan A., Rosenfield P., Slemer A., Marigo P., Girardi L., Bianchi L., 2014, MNRAS, 445, 4287

Thomas G. F., et al., 2018, MNRAS, 481, 5223

Toomer G. J., 1984, Ptolemy's Almagest. Springer

Torra F., et al., 2021, A&A, 649, A10

Udalski A., Szymanski M., Kaluzny J., Kubiak M., Mateo M., 1992, Acta Astron., 42, 253

Vasiliev E., Baumgardt H., 2021, arXiv e-prints, p. arXiv:2102.09568

Vasiliev E., Belokurov V., Erkal D., 2021, MNRAS, 501, 2279

Vickers J. J., Smith M. C., 2018, The Astrophysical Journal, 860, 91

Watkins L. L., et al., 2009, MNRAS, 398, 1757

Wegg C., Gerhard O., Bieth M., 2019, arXiv:1806.09635

Widmark A., Monari G., 2019, MNRAS, 482, 262

Widrow L. M., Gardner S., Yanny B., Dodelson S., Chen H.-Y., 2012, ApJ, 750, L41

Wielen R., 1997, A&A, 325, 367

Williams M. E. K., et al., 2013, MNRAS, 436, 101

Wojno J., et al., 2017, MNRAS, 468, 3368

Xiang M.-S., et al., 2017, MNRAS, 467, 1890

Xu Y., Newberg H. J., Carlin J. L., Liu C., Deng L., Li J., Schönrich R., Yanny B., 2015, ApJ, 801, 105

Xu S., Zhang B., Reid M. J., Zheng X., Wang G., 2019, ApJ, 875, 114

Xue X.-X., et al., 2011, The Astrophysical Journal, 738, 79

Yanny B., et al., 2000, ApJ, 540, 825

Yanny B., et al., 2009, AJ, 137, 4377

York D. G., et al., 2000, AJ, 120, 1579

Zhao G., Zhao Y.-H., Chu Y.-Q., Jing Y.-P., Deng L.-C., 2012, Research in Astronomy and Astrophysics, 12, 723

Zhu C., Byrd R. H., Lu P., Nocedal J., 1997, ACM Trans. Math. Softw., 23, 550–560

Zinn J. C., 2021, AJ, 161, 214

Zinn J. C., Pinsonneault M. H., Huber D., Stello D., 2019, ApJ, 878, 136

de Bruijne J. H. J., et al., 2018, Gaia DR2 documentation Chapter 1: Introduction, Gaia DR2 documentation

de Jong J. T. A., Yanny B., Rix H.-W., Dolphin A. E., Martin N. F., Beers T. C., 2010, ApJ, 714, 663

de Jong R. S., et al., 2012, in McLean I. S., Ramsay S. K., Takami H., eds, Vol. 8446, Ground-based and Airborne Instrumentation for Astronomy IV. p. 84460T, doi:10.1117/12.926239, http://proceedings.spiedigitallibrary.org/proceeding.aspx?doi=10.1117/12.926239 de Swart J. G., Bertone G., van Dongen J., 2017, Nature Astronomy, 1, 0059

van Altena W. F., Lee J. T., Hoffleit E. D., 1995, The general catalogue of trigonometric [stellar] parallaxes. Yale University Observatory

van de Ven G., van den Bosch R. C. E., Verolme E. K., de Zeeuw P. T., 2006, A&A, 445, 513

von Struve O. W., 1840, Astronomische Nachrichten, 17, 177

A

Statistics

A.1 Union Expansion

Considering the simplest case of two overlapping fields, A and B, Equation (3.5) expands to

$$\mathrm{P}(\mathcal{S} \mid \boldsymbol{\theta}, \mathbf{v}) = \mathrm{P}(\mathcal{S}_{\mathrm{A}} \mid \boldsymbol{\theta}, \mathbf{v}) + \mathrm{P}(\mathcal{S}_{\mathrm{B}} \mid \boldsymbol{\theta}, \mathbf{v}) - \mathrm{P}(\mathcal{S}_{\mathrm{A}}, \mathcal{S}_{\mathrm{B}} \mid \boldsymbol{\theta}, \mathbf{v}). \tag{A.1}$$

In most surveys I am considering, there will be substantially more than two fields. For M fields, I expand the union using inclusion-exclusion principle.

$$\begin{aligned}\mathrm{P}\left(\bigcup_{i=1}^{M} \mathcal{S}_i\right) &= \sum_{i=1}^{M} \mathrm{P}(\mathcal{S}_i) - \sum_{i=2}^{M} \sum_{j=1}^{i-1} \mathrm{P}(\mathcal{S}_i, \mathcal{S}_j) \\ &+ \sum_{i=3}^{M} \sum_{j=2}^{i-1} \sum_{k=1}^{j-1} \mathrm{P}(\mathcal{S}_i, \mathcal{S}_j, \mathcal{S}_k) - ... \\ &... + (-1)^{M+1} \mathrm{P}(\mathcal{S}_1, \mathcal{S}_2 .... \mathcal{S}_M) \\ &= \sum_{k=1}^{M} (-1)^{k+1} \left[\sum_{1 \leq i_1 < ... < i_k \leq M} \mathrm{P}(\mathcal{S}_{i_1}, \mathcal{S}_{i_2} ... \mathcal{S} i_k) \right],\end{aligned} \tag{A.2}$$

where I have dropped the conditionals such that $\mathrm{P}(\mathcal{S}_i) \equiv \mathrm{P}(\mathcal{S}_i \mid \boldsymbol{\theta}, \mathbf{v})$ for ease of notation.

Assuming independence of observations from different fields (i.e. the event that a star is selected on field A is independent of whether it has been observed on field B) I can expand the joint probability as the product of the probability of each event.

$$\mathrm{P}(\mathcal{S}_1, \mathcal{S}_2, ... \mid \boldsymbol{\theta}, \mathbf{v}) = \prod_{i=1,2,...} \mathrm{P}(\mathcal{S}_i \mid \boldsymbol{\theta}, \mathbf{v})$$

Using Equation 3.4 I expand the conditional selection probabilities in terms of the event that the positional coordinates lie on the field

$$\mathrm{P}(\mathcal{S}_1, \mathcal{S}_2, ... \mid \boldsymbol{\theta}, \mathbf{v}) = \prod_{i=1,2,...} \mathrm{P}(\mathcal{S}_i \mid \Theta_i, \mathbf{v})\, \mathrm{P}(\Theta_i \mid \boldsymbol{\theta})$$

where $\mathrm{P}(\Theta_i \mid \boldsymbol{\theta}) = 1$ if $\boldsymbol{\theta}$ is on field i and $\mathrm{P}(\Theta_i \mid \boldsymbol{\theta}) = 0$ otherwise. Therefore any joint probability terms with fields which don't contain $\boldsymbol{\theta}$ will vanish from Equation A.2.

I can simplify Equation (A.2) for some specific circumstances:

(1) $\boldsymbol{\theta}$ is not located on any fields. $\mathrm{P}(\Theta_i \mid \boldsymbol{\theta}) = 0 \quad \forall \quad i$. All terms in the expansion are 0

$$\mathrm{P}\left(\bigcup_{i=1}^{M} \mathcal{S}_i\right) = 0$$

(2) $\boldsymbol{\theta}$ is located on only one patch, A. $\mathrm{P}(\Theta_i \mid \boldsymbol{\theta}) = 0 \quad \forall \quad i \neq A$, $\mathrm{P}(\Theta_A \mid \boldsymbol{\theta}) = 1$

$$\mathrm{P}\left(\bigcup_{i=1}^{M} \mathcal{S}_i\right) = \mathrm{P}(\mathcal{S}_A \mid \Theta_A, \mathbf{v})$$

(3) $\boldsymbol{\theta}$ is on the intersection between two fields denoted by A and B:

$$\begin{aligned}\mathrm{P}\left(\bigcup_{i=1}^{M} \mathcal{S}_i\right) =& \mathrm{P}(\mathcal{S}_A \mid \Theta_A, \mathbf{v}) + \mathrm{P}(\mathcal{S}_B \mid \Theta_B, \mathbf{v}) \\ & - \mathrm{P}(\mathcal{S}_A \mid \Theta_A, \mathbf{v}) \times \mathrm{P}(\mathcal{S}_B \mid \Theta_B, \mathbf{v})\end{aligned} \tag{A.3}$$

A.2 Poisson Likelihood

I start from the likelihood of observing a particular object with coordinates v_i given that I only pick one point from the density function, λ

$$\mathrm{P}(v_i \mid n = 1, \lambda) = \frac{\lambda(v_i)}{\int \mathrm{d}v\, \lambda(v)}. \tag{A.4}$$

Expanding this to observations of a population of N objects

$$\mathrm{P}(v_1, v_2 ... v_N \mid n = N, \lambda) = \prod_{i=1}^{N} \frac{\lambda(v_i)}{\int \mathrm{d}v\, \lambda(v)}. \tag{A.5}$$

The likelihood of the data is then given by

$$\begin{aligned}&\mathrm{P}(v_1, v_2 ... v_N, n = N \mid \lambda) \\ &\qquad =\mathrm{P}(v_1, v_2 ... v_N \mid n = N, \lambda)\mathrm{P}(n = N \mid \lambda) \\ &\qquad =\frac{\prod_{i=1}^{N} \lambda(v_i)}{(\int \mathrm{d}v\, \lambda(v))^N} \frac{(\int \mathrm{d}v\, \lambda(v))^N \exp\left(-\int \mathrm{d}v\, \lambda(v)\right)}{N!} \\ &\qquad =\frac{\prod_{i=1}^{N} \lambda(v_i) \exp\left(-\int \mathrm{d}v\, \lambda(v)\right)}{N!}.\end{aligned} \tag{A.6}$$

The denominator here is independent of the parameters of the density model so the likelihood of my dataset, $\{v\}$ is given by

$$\mathrm{P}(\{v\} \mid \lambda) \propto \prod_{i=1}^{N} \lambda(v_i) \exp\left(-\int \mathrm{d}v\, \lambda(v)\right) \tag{A.7}$$

and taking the log likelihood

$$\ln \mathrm{P}(\{v\} \mid \lambda) \propto -\int \mathrm{d}v\, \lambda(v) + \sum_{i=1}^{N} \ln \lambda(v_i). \tag{A.8}$$

A.3 Binomial One-tailed p-value

The one-tailed p-value test provides the probability that the observations would be smaller (in some sense) than the actual measured data, d given the hypothesised model.

For a hypothesised model with parameters ψ and measured data d, the likelihood of the data given the parameters is $\mathrm{P}(d \,|\, \psi)$. The p-value is given by the integral over this (i.e. the CDF)

$$P = \int_{d'_{\mathrm{min}}}^{d} \mathrm{d}d'\, \mathrm{P}(d' \,|\, \psi) \tag{A.9}$$

where d'_{min} is the minimum value the data can take under the model.

Let's say I have a sample of k marbles drawn from a bag of n. I hypothesise that the probability of any marble being selected is Bernoulli (i.e. randomly) distributed with probability q. Therefore my likelihood is Binomial$(k \,|\, n, q)$. The p-value is given by

$$P = \int_{k'=0}^{k} \mathrm{d}k'\, \mathrm{Binomial}(k' \,|\, n, q). \tag{A.10}$$

However, the CDF at $k' = k$ discontinuously jumps. This is shown by the example Binomial CDF in Fig. A.1 where I have used $n = 5$, $q = 0.5$ to demonstrate. The CDF at $k = 3$ has a discontinuity and uniformly covers a range of p-values in the CDF. Therefore I can write the p-value as

$$P = \mathrm{U}\left[\sum_{k'=0}^{k-1} \mathrm{Binomial}(k' \,|\, n, q), \sum_{k'=0}^{k} \mathrm{Binomial}(k' \,|\, n, q)\right]. \tag{A.11}$$

This is shown by the red shaded region in Fig. A.1 as the p-value for $k = 3$ given $n = 5$, $q = 0.5$ is $P \sim \mathrm{U}[0.50, 0.81]$.

I would also like to point out here that without any data ($n = 0$) the p-value becomes uniformly distributed

$$P = \mathrm{U}\left[\sum_{k'=0}^{-1} \mathrm{Binomial}(0 \,|\, 0, q), \sum_{k'=0}^{0} \mathrm{Binomial}(0 \,|\, 0, q)\right] = \mathrm{U}\left[0, 1\right]. \tag{A.12}$$

As one would hopefully expect, if there's no data to test the hypothesis against, then the p-value is model independent (i.e. there's no dependence on q).

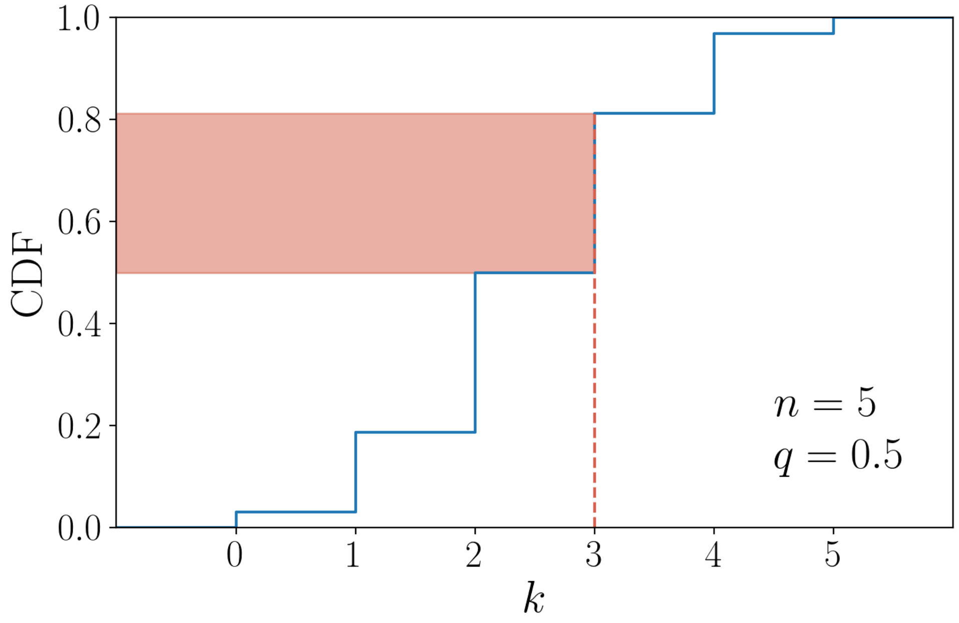


Fig. A.1 The CDF of the Binomial distribution for $q = 0.5$ and $n = 5$ is shown with the blue line as a series of steps. The value of the CDF at $k = 3$ is uniformly distributed between 0.50 and 0.81 shown by the red shaded region.

B

Measurements

B.1 DoF Bug

During the calibration of excess noise in *Gaia* DR2, the degrees of freedom parameter was erroneously used as the total number of AL and AC observations rather than only AL observations as intended. For a full explanation I refer the interested reader to Appendix A in L18.

The result of this bug was that sources brighter than $G = 13$, for which 2D observations were used, received overestimated measurement uncertainties. Through the attitude calibration, this indirectly impacted sources with $G > 13$ although the increased photon count noise at dimmer magnitudes dampens the effect for dim stars.

To correct for this, the astrometric covariance was multiplied by a correction factor

$$F = (1 + 0.8R)\sqrt{\frac{2}{1 + \sqrt{1 + 4(1 + 0.8R)^2\left(\frac{0.025\mathrm{mas}}{\sigma_\varpi}\right)}}} \tag{B.1}$$

taken from Equation A.6 of L18. The published χ^2 and ASTROMETRIC_SIGMA5D_MAX didn't receive this correction, the latter being because the selection of the DR2 astrometry sample was performed before the bug was corrected.

As a result, when estimating ASTROMETRIC_SIGMA5D_MAX in Section 5.6, in order to obtain a good agreement with the data, I must decorrect for the DoF bug by dividing through by the correction factor, F.