Astrostatistics for luminosity calibration in the Gaia era

A dissertation presented
by
Max Palmer
to
The Department of Astronomy and Meteorology

For the degree of
Doctor of Physics
in the subject of
Astrophysics

University of Barcelona
Barcelona, Spain
November 2014

A dissertation presented
by
Max Palmer
to
The Department of Astronomy and Meteorology
November 2014

Thesis directors:
Dr. Xavier Luri Carrascoso
Dr. Frédéric Arenou

Thesis tutor:
Dr. Francesca Figueras Siñol

To trials and travels with E.A.G.

Acknowledgments

I am sincerely grateful for the help and advice which I have received over the last three years from my three excellent supervisors, Xavier Luri, Frédéric Arenou, and Eduard Masana. Our weekly meetings and telecons were invaluable, and this thesis would not exist were it not for the expertise of the three of them. Xavier, as my tutor, brought a concise, organised and laid back style of working which suited me well, and undoubtedly made this period much less stressful than it could otherwise have been.

I am indebted to Frédéric Arenou, who warmly received me in tranquil Paris-Meudon. Under his supervision, and consistent good humour, the goals of the thesis really came into focus. Additionally I am indebted to Gisella Clementini, who received me in Bologna at the other end of the timeline of thesis. It was a pleasure to build new collaborations, and experience life in historic and vibrant Italy.

I would like to thank my colleagues and members of the Gaia and DAM teams, both scientists and engineers, for being welcoming and supportive during my entire stay. Particularly thanks goes to Cesca Figueras for overcoming the countless bureaucratic problems I encountered, and without whom I may never have been able to complete. Additionally I would like to express my appreciation for Lola Balaguer, who organised all of my paperwork, Dani Molina for keeping Linx running with the myriad of Python libraries, and everybody else who took time to explain the complexities of Gaia, GOG, and astrophysics in general.

Thanks also goes to Hoda Abedi, with whom we have completed the various stages of the PhD at the same time, and the other GREAT students around Europe with whom we have spent time with in all of the conferences and secondments. Also to all of the people from around the world who I have lived with in the past three years (particularly *las seis llamas*), who have contributed to making this PhD a profound cultural learning experience as well as a academic one.

Finally I would like to express my gratitude to my parents, who's sacrifices gave me opportunities which made all of this possible.

Chapter specific acknowledgements

For Ch. 3: GOG is the product of many years of work from a number of people involved in DPAC and specifically CU2. The authors would like the thank the various CUs for contributing predicted error models for Gaia. The processing of the GOG data made significant use of the Barcelona Supercomputing Center (MareNostrum), and the authors would specifically like to thank Javi Castañeda, Marcial Clotet, and Aidan Fries for handling our computation and data handling needs. Additionally, thanks go to Sergi Blanco-Cuaresma for his help with Matplotlib. I thankfully acknowledge the computer resources, technical expertise and assistance provided by the Red Española de Supercomputación.

For Ch. 6: I would like to thank Carme Jordi for her useful comments, and Anthony Brown and Jos de Bruijne for their input on recreating the results of Narayanan & Gould (1999). We likewise thank CU2 for the use of GaiaSimu and the GOG simulator, and Erika Antiche for running GOG simulations.

Thesis advisor: X. Luri and F. Arenou Max Palmer

Astrostatistics for luminosity calibration in the Gaia era

Abstract

The European Space Agency astrometric satellite *Gaia* was launched in December 2013. Having completed its commissioning phase, Gaia is transmitting data to Earth that will lead to the creation of an astrometric catalogue of more than 10^9 celestial objects. This catalogue will contain astrometric measurements, including parallaxes, with an expected precision in the 10-300 μas range. Due to the very large quantity of data which will be produced, it is essential to begin preparing for the eventual use of the Gaia catalogue. In this thesis we summarise the expected contents of the Gaia end-of-mission catalogue using state of the art simulations from the Gaia Data Processing and Analysis Consortium.

To prepare for the use of the Gaia catalogue, we define a set of statistical methods for the correct use of trigonometric parallax data. Direct use of the distance obtained from the inverse of the parallax is problematic, due to the possible effect of numerous known statistical biases and selection effects. We provide a set of methods which utilise parallax information correctly, in order to calibrate the absolute luminosity of several objects of interest: normal stars, variable stars, open clusters, and the Magellanic Clouds. The methods make extensive use of statistical modelling and Bayesian statistics in order to treat the data correctly and without bias. The methods have been tested using simulated Gaia data, and are ready to be applied to the real Gaia data after release.

In the case of Cepheid and RR-Lyrae variable stars, we present a method for fitting the period-luminosity relation either with parallaxes, or without. For the parallax free case, the method has been applied to real data in order to calibrate the period-luminosity-metallicity relation for RR-Lyraes in the LMC. In the case of open clusters, the method has been applied to Hipparcos data in order to estimate the distance to the Pleiades and Hyades. We contribute to the long-standing Pleiades distance debate through an analysis of the effect of correlated errors in the original (1997) and new (2007) hipparcos reductions. With the possibility of the correlated errors affecting the maximum precision obtainable with Gaia data, we study the expected effect of error correlations in Gaia.

Director de tesis: X. Luri and F. Arenou Max Palmer

Astrostatistics for luminosity calibration in the Gaia era

Resumen

El satélite astrométrico *Gaia* de la Agencia Espacial Europea fue lanzado en Diciembre 2013. Ahora que ha completado su fase de comisionado, Gaia está transmitiendo a la Tierra datos que permitirán la construcción de un catálogo astrométrico de más que 10^9 objetos celestes. Este catálogo contendrá medidas astronómicas, incluyendo paralaje, con una precisión esperada en el nivel de μas. Dada la extrema cantidad de datos que serán creados, es necesario comenzar a prepararse para el eventual uso del catálogo. En esta tesis resumiremos el contenido esperado del catálogo de final de misión, usando simulaciones de última generación del Consorcio de Análisis y Procesado de los Datos de Gaia.

A fin de investigar las capacidades de Gaia en la calibración de luminosidades, y los problemas potenciales en este campo que deben ser considerados, es necesario estudiar varios detalles específicos de la misión. En primer lugar, es importante determinar y cuantificar el retorno científico que Gaia proporcionará y el contenido esperado del catálogo. Trabajando con los datos del simulador de Gaia para la obtención del catálogo final de la misión, ha sido posible obtener un profundo conocimiento del contenido del mismo y la precisión que tendrá. Este análisis no sólo va servir como una referencia útil para la comunidad científica involucrada en la preparación de la explotación de los datos reales, sino también ha sido fundamental para decidir que aplicaciones de la calibración de luminosidades deben abordarse en los siguientes pasos.

Con el objeto de preparase para el uso del catálogo final de Gaia, hemos definido un conjunto de modelos para el uso correcto de las paralajes trigonométricas. El uso directo de la distancia obtenido por inversión de la paralaje es problemático, debido a los varios sesgos estadísticos y efectos de selección. Proporcionamos un conjunto de métodos que usan correctamente la información contenida en las paralajes, con la finalidad de calibrar la luminosidad absoluta de varios objetos de interés: estrellas normales, estrellas variables, cúmulos abiertos, y las nubes Magellanes. Los métodos hacen uso de estadísticos Bayesiana y de modelos estadísticos para tratar los datos correctamente y sin sesgos. Estos métodos han sido probados con datos simulados y reales de varias fuentes, y están disponibles para aplicarlos a datos reales de Gaia

cuando sean publicados.

Para estrellas variables, presentamos un método para inferir el relación entre periodo y luminosidad, con o sin paralajes. El método que no usa paralajes ha sido aplicado a un conjunto de datos de OGLE-III y VMC para calibrar la relación entre periodo, luminosidad y metalicidad (PLM) para estrellas tipo RR-Lyrae en la Nube Grande de Magellanes. Este método Bayesiano utiliza *Markov Chain Monte Carlo* para estimar la inclinación y punto de cero de la relación PLM y cuantificar la posterior función de densidad de probabilidad.

Para cúmulos abiertos, presentamos un método y su aplicación a datos de Hipparcos para estimar la distancia de las Pleiades y las Hyades. El método combina información de paralaje, magnitud aparente, movimiento propio, color, y velocidad radial para determinar simultáneamente: distancia, velocidad espacial 3D, y una isócrona observada. A través de su aplicación para determinar la distancia hasta las Pleiades, hemos detectado evidencia directa de segregación de masa. Las estimaciones para la distancia de las Pleiades y las Hyades concuerdan con estudios anteriores, además son altamente precisas dado que hemos incluido más información. Hemos simulado observaciones de una cúmulo como las Pleiades para probar el uso del método con datos de Gaia. Encontramos que el método sera precisa hasta unos kpc. La figura 1 muestra la comparación de resultados del metodo Máxima Verosimilitud con el método básico de inversión de paralaje.

Hemos contribuido al debate acerca del problema de la distancia a las Pleiades, a través de un análisis de los efectos de la correlación en los errores en el catálogo original de Hipparcos (1997) y el nuevo (2007), revisando los resultados de autores anteriores, los cuales han argumentado diversos problemas en el catálogo de Hipparcos. Dada la posibilidad de que dichas correlaciones en los errores puedan afectar la precisión máxima obtenible con Gaia, hemos estudiado el efecto potencial en el catálogo de Gaia. Los errores correlacionados son errores que tienen un componente en común en una cantidad de estrellas, normalmente con baja separación en el cielo, por razones diversas, como una pequeña inexactitud en la determinación de la actitud del satélite. Tales efectos pueden limitar la precisión del cálculo de valores medios para el conjunto de las estrellas afectadas. Hemos construido una aproximación de la solución global astrométrica de Gaia de un área pequeña para dos casos: con errores con un componente correlacionado, o con errores totalmente independientes. Comparando los dos casos hemos comprobado que hay un límite en la precisión máxima obtenible de alrededor de 1μas.

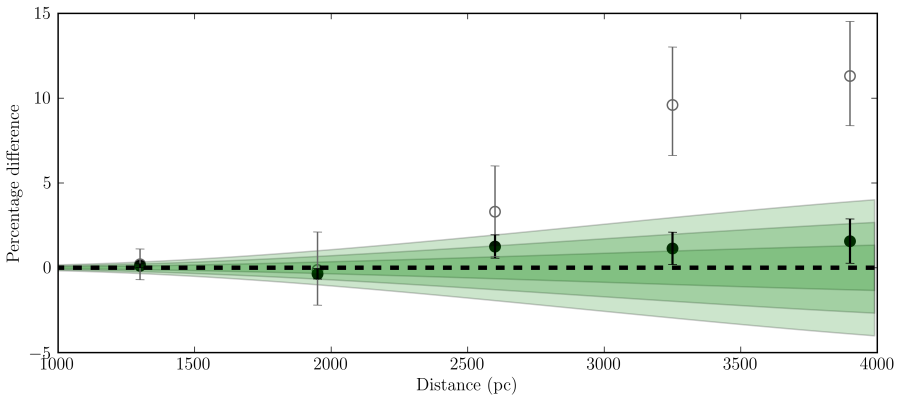


Figure 1: Resultados del estimación de la desestanca a una conjunto de simulados cúmulos como las Pleiades, con distancias desde 10 a 30 veces la distancia original. Los círculos llenas muestran la porcentaje diferencia dentro la distancia estimado por Máxima Verosimilitud y la distancia verdadera. Los círculos vacías muestran la porcentaje diferencia dentro el inverso del mediana paralaje y la distancia verdadero. La área verde pone de manifiesto los errores de 1, 2, and 3σ extrapolado por el rango completa.

Publications

The work in this thesis has formed the basis of all or part of the following publications in peer-reviewed scientific journals:

- Content of Sec. 3: Overview and stellar statistics of the expected Gaia Catalogue using the Gaia Object Generator. Luri, X.; Palmer, M.; Arenou, F.; Masana, E.; et. al. **Astronomy & Astrophysics, Volume 566, 2014, id: A119, 15 pp.**

- Content of Sec. 5:

  A new Period-Luminosity-Metallicity relation for RR Lyrae stars and the impact of Gaia. Muraveva, T.; Palmer, M.; Clementini, G.; Luri, X.; et. al. **Article in preparation.**

- Content of Sec. 6

  An updated maximum likelihood approach to open cluster distance determination. Palmer, M.; Arenou, F.; Luri, X.; Masana, E. **Astronomy & Astrophysics, Volume 564, 2014, id: A49, 14 pp.**

- Content of Sec. 7

  Error correlations in Gaia over small angular scales. Palmer, M.; Arenou, F.; Luri, X.; Masana, E. **Article in preparation.**

Contents

1 Introduction

The understanding of the distance scale of the universe is one of the most fundamental aspects of astronomy. Knowledge of the distance to an object and its absolute luminosity is inextricably linked, and calibration of these properties has implications in studies at all scales of the universe. From resolving the position of neighbouring open clusters to calibrate stellar evolution models, to estimating the distances of ultra-high redshift galaxies in order to determine the corresponding age of the universe, precise distance and luminosity calibration is crucial. Yet, the task is far from easy.

Measuring the distance to a star in the sky by observing its apparent brightness is of course impossible, as the observer has no way of knowing if the bright star they are observing in the night sky is actually a very bright distant star, or a star of modest brightness which happens to be located relatively close to Earth. This severe lack of any measurable distances in space throughout the

majority of human history has lead to many strange and interesting concepts of the structure of the universe.

In early times (pre-1500AD), the study of the cosmos was undertaken solely by theologists and philosophers. Many early civilisations believed that the Sun, Moon, planets, and the stars were equidistant, existing in a single plane, or existed in a set of spherical shells around the earth. It wasn't until the start of the 16^{th} century that measurements of the positions of the planets were combined with robust mathematics by Copernicus and later Kepler in order to correctly order the planets in their orbit around the sun, and in doing so calculate the relative sizes of their orbits. Still, the limiting optical technology of the time prevented any possible determination of the distances of stars, which were still widely seen as existing in a single thin sheet around the earth.

The first and only direct way to measure the distance to stars is through the use of trigonometric *parallax*, the measurement of the apparent movement of stars from the perspective effect caused by the translation of the observer (on earth) around the sun. Attempts were made to measure stellar parallax by Galileo as early as the 17^{th} Century, but again, limitations in the precision of position measurements at that time made any detection of parallax motion impossible. The lack of measured parallaxes was even used in that era as an argument against heliocentricity (Graney, 2006). The first successful measurement of a stellar parallax was made more than 200 years later, by the German astronomer and mathematician Friedrich Bessel.

Due to the tiny angles involved, measurement of parallaxes remained difficult and time consuming until very recently. While some catalogues of trigonometric parallax have been produced (e.g.: Jenkins (1963), van Altena et al. (1995)), the first large and homogeneous all sky sample of parallaxes was produced by the European Space Agency (ESA) mission, *Hipparcos*, in 1989 (Perryman & ESA, 1997). By moving to space based observation, Hipparcos benefited from a very stable environment, and due to having two telescopes separated by a large angle could measure absolute parallaxes, which are not possible from the ground. The release of the 118 thousand star catalogue in 1997 (Perryman &

ESA, 1997), which included a measurement of the absolute trigonometric parallax for every star, represented a huge breakthrough in attempts to measure stellar distances, and it is still in widespread use to this day.

However, the angles involved in measuring distances using parallax are incredibly small. Proxima Centauri, the nearest star to the sun, has a distance of only 1.3 parsecs, leading to a parallax of 2×10^{-4} degrees (769 mas). This means that even for ground breaking space missions such as Hipparcos with astrometric precision of a few miliarcseconds or below, useful distance measurement is confined to stars in the very local neighbourhood. With parallaxes quickly becoming unmeasurable for more distant objects, astronomers have developed a series of innovative and complimentary distance measures which can be used to step out to greater and greater distances. The first step on the so called *distance ladder* is parallax measurements by Hipparcos and other missions such as the Hubble Space Telescope (HST) (Benedict et al., 1995). These form the base calibration of distances with direct distance measurements to objects from the closest stars out to a few hundred parsec. These distances are then used to calibrate a second tier of distance indicators, including nearby standard candles.

Several types of variable stars are known to exhibit a strong correlation between their absolute magnitude and directly observable properties such as the period of pulsation. The most famous of these are classical Cepheid variable stars, which have been known to follow a tight relation between period and luminosity since the discovery of the relation by Leavitt (1908). Other variable star types such as RR-lyrae and Mira variables have also been found to follow similar relations in luminosity, period, and metallicity (see e.g. Longmore et al. (1986), Liu & Janes (1990), Nemec et al. (1994), Freedman & Madore (1990), and work continues to quantify and calibrate these relationships.

Variable stars, and particularly Cepheids as they are massive and bright, have been used to estimate distances across the Milky Way, and have been useful historically in characterising the size and shape of our Galaxy and our position within it (e.g. Shapley, 1939). Such variables are also easily observ-

able in the Magellanic Clouds, where direct parallax measurements have been impossible until now, and many attempts have been made to quantify the distance to both the Large and Small Magellanic clouds using variable stars for which absolute luminosity relations have been calibrated within the galaxy (Caldwell & Coulson, 1986; Udalski et al., 1999a; Haschke et al., 2012).

Another type of standard candle is that of some types of supernovae, particularly type Ia, which, due to their formation mechanism are believed to exhibit a consistent peak luminosity and light curve (Baade, 1938; Riess et al., 1996). Due to their extreme brightness, they are observable in galaxies up to megaparsec distances and are therefore useful at calibrating the distances to their host galaxies in the extragalactic distance scale. The combination of distance and redshift information for distant galaxies allows the study of the structure and evolution of the universe (Riess et al., 2004).

Use of the Hertzsprung-Russell (HR) diagram can also be useful in determining the distances to stars. The concept of main sequence fitting (e.g. Pinsonneault et al. (2003) and following series) can be applied to clusters of stars, and involves comparing the position of the main sequence on the HR diagram with that of a cluster of known distance. As all stars in each cluster are known to be at a similar distance, the shape of the main sequence will be common to both clusters, yet the offset in magnitude will be due to the difference in the distance of the two clusters (with the added complicating factor of interstellar extinction). This method can be used for galactic open and globular clusters clusters.

Alternatively, the observation of red giant branch (RGB) stars is possible out to nearby external galaxies due to their extreme luminosity, and can indicate distance through positioning of the Tip of the RGB (TRGB, Lee et al. (1993), Madore et al. (1997)). The core burning and helium flash which occurs after the transition to the RGB leads to a fairly specific maximum absolute luminosity, independent of the mass of the star. This means that the brightest Population II giant stars in a galaxy can be treated as a sort of standard candle. This method has been used to calibrate the distance to the Magellanic

Clouds (e.g. Sakai et al., 2000), and other galaxies up to several Mpc with the use of the HST (e.g. Sakai & Madore (1999), Sakai et al. (1997)).

The recurring theme with all of the above methods is the need for precise *luminosity calibration*. For variables, the zero point of their absolute luminosity relations must be calibrated. For MS fitting, the absolute luminosity of the zero age main sequence must be determined, and for the TRGB, its position on the HR diagram must be determined. Where methods overlap, each distance estimator (or a combination of several) is used to calibrate the one suitable for determination of yet greater distances. Uncertainty and systematic errors are carried through, leading to progressively worse precision as distances increase. The penalty for imprecision at the smallest scales, which are then used to calibrate other measures, can be significant. Recent controversy about the distance to the LMC, the very nearest significant neighbour of the Milky Way, has been highlighted by the famous plot from Gibson (2000) and later expanded by Benedict et al. (2002), who show the difference of more than 17 kpc, or 45%, in the distance determination of the LMC obtained by using different methods. While this example is extreme, and has been improved upon in recent years, it highlights the requirement for precise luminosity calibration at galactic scales. The European Space Agency (ESA) is taking the lead in this, with the flagship space mission *Gaia*.

1.2 Gaia mission / instruments

Gaia is a cornerstone ESA mission, and was launched on the 9^{th} of December 2013. Gaia is a successor to the Hipparcos astrometric space mission, and is based on the similar observation principles of simultaneous observation using two telescopes separated by a large angle. Gaia combines high precision optics and stable environment at Lagrangian point L2 with the largest CCD camera ever put into space in order to produce a highly precise astrometric catalogue of more than 10^9 stars, extragalactic objects, and solar system objects. Through Gaia's photometric instruments, object detection up to $G = 20$ mag will be

possible (see Jordi et al. (2010) for a definition of G magnitude), including measurements of positions, proper motions, and parallaxes up to micro arcsecond accuracy. The on-board radial velocity spectrometer will provide radial velocity measurements for stars down to a limit of around $G_{RVS} = 16$ mag. With low-resolution spectra providing information on effective temperature, line of sight extinction, surface gravity, and chemical composition, Gaia will yield a detailed catalogue that contains roughly 1% of the entire galactic stellar population.

Gaia will represent a huge advance on its predecessor, Hipparcos (Perryman & ESA, 1997), both in terms of massive increases in precision and in the numbers of objects observed. Thanks to accurate observations of large numbers of stars of all kinds, including rare objects, large numbers of Cepheids and other variable stars, and direct parallax measurements for stars in all galactic populations (thin disk, thick disk, halo, and bulge), Gaia data is expected to have a strong impact on luminosity calibration and improvement of the distance scale. This, along with applications to studies of galactic dynamics and evolution, and of fields ranging from exoplanets to general relativity and cosmology, Gaia's impact is expected to be significant and far reaching.

The Gaia space astrometry mission will map the entire sky in the visible G band over the course of its five-year mission. Located at Lagrangian point L2, Gaia will be constantly and smoothly spinning. It has two telescopes separated by a basic angle of 106.5°. Light from stars that are observed in either telescope is collected and reflected to transit across the Gaia focal plane.

The Gaia focal plane can be split into several main components. The majority of the area is taken up by CCDs for the broad band G magnitude measurements in white light, used in taking the astrometry measurements. Second, there is a pair of low-resolution spectral photometers, one red and one blue, producing low-resolution spectra with integrated magnitude G_{RP} and G_{BP}, respectively. Finally, there is a radial velocity spectrometer observing at near-infrared, with integrated magnitude G_{RVS}. The magnitudes G, G_{RP}, and G_{BP} will be measured for all Gaia sources ($G \leq 20$), whereas G_{RVS} will

be measured for objects up to $G_{RVS} \leq 16$ magnitude. For an exact definition of the Gaia focal plane and the four Gaia bands, see Figs. 1 and 3 of Jordi et al. (2010).

The motion of Gaia is complex, with rotations on its own spin axis occurring every six hours. This spin axis is itself precessing, and is held at a constant 45° degrees from the Sun. With the Earth, Gaia will orbit the Sun over the course of a year. Thanks to the combination of these rotations, the entire sky will be observed repeatedly. The Gaia scanning law gives the number of times a region will be re-observed by Gaia over its five-year mission, and comes from this spinning motion of the satellite and its orbit around the Sun. Objects in regions with more observations have greater precision, while regions with fewer repeat observations have lower precision. The average number of observations per object is 70, although it can be as low as a few tens or as high as 200.

1.3 Rationale

With the successful launch of Gaia in December 2013 and subsequent completion of the commissioning phase, Gaia is in nominal mission mode and is transferring data to ground. There are several planned data releases, with the first expected in summer 2016, with the preliminary release of parallax data expected in early 2017. It is therefore essential to prepare for the use of what will be at the time of its release one of the largest and most complex astronomical datasets ever created. As indicated above, there is a strong need for precise parallax data which can be used to calibrate stellar luminosities and different aspects of the distance ladder. This thesis focuses on methods for luminosity calibration, and takes a statistical approach to facilitate efficient use of the catalogue. However, as has been known since long before Hipparcos, the correct use of trigonometric parallax data is non trivial. In this thesis several tools are developed for luminosity calibration with Gaia data. Chapter 2 contains a summary of the numerous biases and other problems which arise when using parallax data for luminosity calibration, providing a justification

for the mathematical and statistical methods applied during this thesis.

Of course, the Gaia data is not available at the time of writing. This is not problematic, because the methodology can be tested with simulated data, and the resulting algorithms can be applied to currently existing data before Gaia release. In this way, the tools required for luminosity calibration with Gaia will be prepared for use with the Gaia data as soon as it is available. In order to summarise the expected contents of the Gaia catalogue, and to have a set of Gaia-like data in order to test the methods produced during this thesis, it has been necessary to perform a statistical overview of a simulated Gaia mock catalogue. The results of which can be found in Chapter 3.

The core product of this thesis is a set of working methods for luminosity calibration with Gaia data. These tools have been applied to fundamental aspects of luminosity calibration. The tools are given in Chapters 4 to 7 in the form of statistical methods, with full derivations of the mathematics used. They also have been implemented in functioning code, tested with simulated data, and in some cases applied to real data, such as Hipparcos. This has produced several results which are given in the text, and have featured in several academic publications. Additionally the codes and expertise have been developed, and are available, ready for exploitation of the Gaia catalogue when it is made available.

1.4 Main contributions of this thesis

The main contributions of this thesis are listed here.

- An overview of the problems which can arise during the use of parallax data, including statistical biases and selection effects.
- An overview and statistical analysis of the expected Gaia catalogue using the Gaia Object Generator simulated catalogue. We present a selection of the statistics to provide an overview of this one billion star catalogue.

- A Maximum Likelihood (ML) tool for luminosity calibration of stars of a specific luminosity class and spectral type, ready for use with Gaia data.
- A ML tool for calibration of the period-absolute luminosity relation of Cepheids, ready for use with Gaia data.
- A fitting package for linear regression with 2 or 3 dimensional data, used for the calibration of the RR-Lyrae period-luminosity-metallicity relation.
- A tool for simultaneous fitting of an open cluster's mean distance, position, 3D kinematics, and an observational isochrone, ready for use with Gaia data. Tested using Hipparcos data.
- A tool for fitting the mean distance to the LMC, ready for use with Gaia data.
- A study into the effects of error correlations in the Gaia catalogue.

2

Background

2.1 Biases and sample selection effects

It has long been known that the correct use of parallax information is non-trivial. When dealing with observational data, there are many complicating effects such as observational errors and sample incompleteness which can effect results. Use of the Hipparcos astrometric catalogue brought a renewed focus to the debate, and attempts were made to stress the importance of the choice of methods for the use of such data and the understanding of any results. Papers such as Arenou & Luri (1999) and Arenou & Luri (2002) have publicised several of the major effects, and provide several examples in the literature of misuse of such data.

This section contains an overview of the problems which can arise during the use of astronomical data, including statistical biases and selection effects,

with particular focus given to the use of parallax data. A brief description of the main effects are given along with some examples from the literature in order to highlight potential difficulties in the use of the Gaia catalogue and to justify the choice of methods given in Chapter 4.

2.1.1 Malmquist

In 1922 the Sweedish astronomer G. Malmquist highlighted the effects of selection bias in observational astronomy, caused by preferentially selecting brighter objects when constructing a magnitude limited sample.

Given a population of stars, each apparent magnitude is dependent on the star's distance and its absolute magnitude (excluding other effects such as reddening). By making a cut in apparent magnitude at some faintness limit, brighter stars are included from a greater volume of space, because stars which are intrinsically brighter can be seen at greater distances. As such, the underlying population of stars is sampled non-uniformly, resulting in a mean absolute magnitude greater than that of the underlying population.

This effect is present in any magnitude limited sample, and is therefore widespread in astronomy as almost all telescopes are limited primarily by the flux of light arriving at the telescope.

Fig 2.1 shows a simple model of the bias in action. The population of stars is simulated using a Gaussian distribution for the absolute magnitudes and generates distances assuming that stars are uniformly distributed in space. The apparent magnitude of each star is calculated, ignoring any complicating effects such as reddening, mixed population or non-uniform spatial distributions. By assuming some limiting magnitude and selecting a sample from only the stars which are brighter than this limit, a cut is performed in the apparent magnitude distribution. The distribution of the absolute magnitudes of this sample is not the same as the original population, having a non-Gaussian shape and a brighter mean absolute magnitude.

Properties such as the mean absolute magnitude, which have been derived

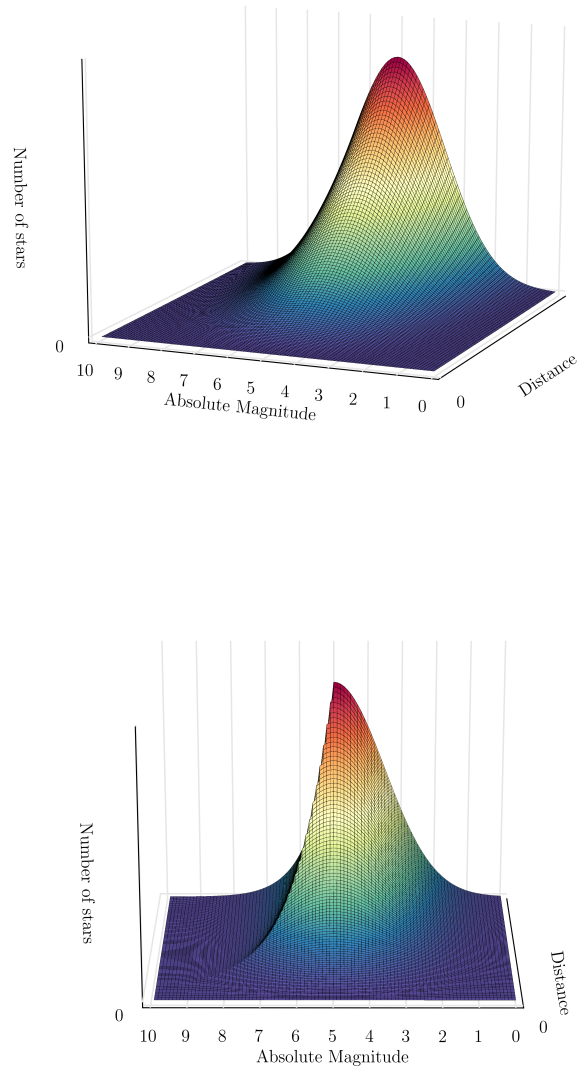


Figure 2.1: A Monte Carlo simulation of a population of stars with uniform spatial distribution and a Normal distribution of absolute magnitudes. Top: the simulated population; Bottom: the population after the application of a cut in apparent magnitude.

from the observed population, will be strongly biased with respect to the true underlying population. In more realistic cases including mixed populations or realistic spatial distribution (e.g. galactic disk), the problem becomes more complex and difficult to correct for.

2.1.2 Lutz-Kelker

The Lutz-Kelker (Lutz & Kelker, 1973) bias is another statistical effect known to bias mean parallaxes in observational data. The theoretical basis of the Lutz-Kelker bias is that the spatial distribution of stars results in a non-uniform distribution of parallaxes. When this non-uniform distribution is affected by observational errors, the mean of the distribution can be shifted significantly.

In the simple case of a uniform spatial distribution of stars, we can divide the space into thin heliocentric spherical shells with equal thickness and radii r to $r + \delta_r$. As the radius increases, there are more stars per shell due to the fact that more distant shells contain a greater volume of space. In error affected observation, stars observed in one shell could have been scattered from shells either closer or further away. However, if there are more stars at greater distances, there is a greater probability that stars observed to be in some specific shell come from greater distances.

This causes measured distances to be on average too small. The situation is further complicated when attempting to account for the true spatial distribution, which is affected by the shape of the Galaxy and our position within it. Lutz and Kelker calculated analytically a bias of -0.43 mag for a relative parallax error $\sigma_\varpi/\varpi = 0.175$. Koen (1992) provide tables of the extent of the bias, along with confidence intervals, showing a large dependence on the choice of spatial distribution and the effect of uncertainty in the distribution of the formal errors.

The bias has been found at similar levels empirically by Oudmaijer et al. (1998), though use of the Hipparcos catalogue. They propose a correction

which can be applied to individual stars or a sample, based on assumptions about the luminosity function, spatial distribution and error in parallax. However, a correction is not necessary if the spatial distribution of the population is included explicitly the calculation of the mean parallax (see Sect. 2.2). This avoids the Lutz-Kelker bias completely, negating the need for any posterior corrections with the numerous assumptions and large confidence intervals associated with such.

2.1.3 Non-linear transformations

Non-linear transformations are present throughout astronomy. In the case of parallaxes, the very simple relation, $D = 1/\varpi$ relates an object's parallax with its distance, when the distance is in parsec and the parallax is in arcseconds. However, if the error distribution in the measured parallax can be described by a symmetrical Gaussian distribution, then the error of its inverse (the distance) can no longer be described with a symmetrical error distribution due to the fact that taking an inverse is a non-linear transformation.

$$\frac{1}{\varpi \pm \sigma} \neq D \pm \sigma'$$

A similar problem exists in famous equation relating absolute and apparent magnitude:

$$M = m + 5\log(\varpi) + 5$$

where, although the parallax is used directly, the logarithm acting on the parallax results in an asymmetric uncertainty and a bias if this is not correctly taken into account.

The problem is further complicated by the possibility of negative parallaxes, which, although geometrically impossible, can occur in real data for distant objects where large observational errors can result in a negative value. As the inverse of a negative parallax results in a negative distance, and the logarithm of a negative number is mathematically non-defined, it could be tempting to

discard such results.

Such measurements do however contain information, and can be used to improve statistics of a sample. By discarding negative values, a bias is introduced due to the more distant, and therefore smaller parallax, stars being more likely to be have negative parallaxes, and that fainter stars are usually more likely to have larger errors.

The above mentioned effects are particularly relevant in the construction of colour-absolute magnitude diagrams, which often require the calculation of each object's absolute magnitude from apparent magnitude and distance or parallax information.

2.2 Methods

In this thesis, several different statistical methods are used. These methods are utilised for applications such as linear and non-linear regression, parameter fitting, and Bayesian inference, and have different strengths and weaknesses making them attractive for some situations above others. Here, an introduction and brief description of the main methods are given, to familiarise the reader before giving the detailed application of such methods in Sect. 4.

In science in general it is often useful to *model* a system or a set of results. This allows one to quantify a process or outcome, make predictions about future outcomes, or to summarise results in a concise manner. Modelling of a system or set of results generally requires a simplification or an assumption about its nature. The development of a model can be as simple as modelling a set of x, y data as following a straight line. In this case, one would assume that a model for the data is

$$y = Ax + B \tag{2.1}$$

where the parameters A and B are the model parameters which are to be found. In the other extreme, models can involve multiple highly complex

distributions and many tens of unknown parameters.

In either case, if the model does not accurately describe the system being modelled, the results can be biased, or in the worst case meaningless. Fitting a straight line to data which follows a quadratic power law will provide a result, albeit one which may not be particularly useful. For this reason models which are chosen must always be justifiable, and checked by eye or by using one or several statistical tests. In general, models which better describe the data, or have more parameters, will reduce the variance between the data and the model.

With one or several models defined, it is necessary to obtain the free fit parameters. There are several methods to do so, and those which are utilised extensively in this thesis are outlined below.

2.2.1 Least Squares / Frequentist statistics

Probably the most basic and widely used method for linear regression and model fitting is Minimum Least Squares (LS). LS has been in use since around 1800, and involves minimising the sum of the square of the residuals between a set of observations and a set of corresponding model predictions. In the simple case of a set of independent x and dependent y data, a model of the data takes the form $y_{\mathrm{model}} = f(x, \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is a vector of the parameters of the model. The least squares solution is therefore found by minimising

$$S = \sum_{i=0}^{N} r_i^2 \tag{2.2}$$

where r is the residual $y_{\mathrm{obs}} - y_{\,model}$.

In 1810 it was shown by Laplace that, in the case of unbiased, uncorrelated errors of equal size, the best possible estimation of model parameters is that given by the LS approach. Due to this, and the ease of learning and applying the LS technique, LS is one of the most commonly used fitting techniques currently in use. It is also very fast to compute, and is therefore easily appli-

cable on large datasets. However, in cases where the assumption of perfectly known, unbiased, uncorrelated and equal errors are not fulfilled, more complex methods such as that given in Sect. 5.2 should be preferred.

Some variations of LS exist, such as weighting the residuals of each observation by the observational error.

2.2.2 Bayesian statistics

Work containing the basis for what is now known as Bayes theorem (Bayes, 1763) was published by the Royal Society London, in 1763, two years after the death of its writer, Thomas Bayes. The work defines probability (page 376, definition 5), as:

> The probability of any event is the ratio between the value at which an expectation depending on the happening of the event ought to be computed, and the value of the thing expected upon its happening

Bayesian methodology uses this definition of probability which, importantly, allows the inclusion of prior information in assigning a probability. The core concept of Bayes theorem can be concisely expressed as:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{2.3}$$

where A is an event and B is the background. Here, $P(A|B)$ is the conditional probability of A occurring, given that B is true. $P(A)$ is the prior on A, which contains any initial belief about its outcome. The above is a very powerful statement, as it combines knowledge about a test (B) with observational data and any prior belief in its result. This inclusion of prior belief can lead to stronger constraints on our posterior probability $P(A|B)$, provided that the prior belief is not wrong, and therefore biasing the result.

In astronomy, there is often a strong degree of previous knowledge about an object or event, which in many cases is ignored when new observational data

is acquired. This previous knowledge can be put to good use, in defining a model which accurately describes the data, and in providing a constraint on the outcome of the fitting though the application of priors.

Maximum Likelihood Estimation

Maximum Likelihood Estimation (MLE) is a method for estimating the parameters of a statistical model, given a set of observations. The ML estimate of a set of model parameters are those that maximise the *Likelihood* of the observations. For a set of n observations, ($\boldsymbol{x} = x_1, x_2, ..., x_n$), and set of model parameters, $\boldsymbol{\theta}$, the Likelihood is defined as the probability

$$\mathcal{L}(\boldsymbol{\theta}; \boldsymbol{x}) = P(\boldsymbol{x}|\boldsymbol{\theta}) \tag{2.4}$$

Maximising $\mathcal{L}$, or equivalently minimising $-\mathcal{L}$, gives the parameters which provide the maximum probability of having taken the observations in $\boldsymbol{x}$. Comparing the above with Bayes theorem in Eqn. 2.3 we see the term $P(\boldsymbol{\theta})/P(\boldsymbol{x})$ missing. Of course, $P(\boldsymbol{x})$ is independent of the parameters $\boldsymbol{\theta}$ and so in the maximisation can be ignored. Secondly, MLE assumes no prior on any of the parameters, leaving $P(\boldsymbol{\theta})$ as a uniform probability.

In many cases of parameter estimation the priors on the parameters should be uninformative, i.e. very wide or flat, in order to take the weight of the results from the data being used, and to avoid biasing the result by using an incorrect prior. In these cases, MLE provides an excellent method for parameter estimation, as has been shown through its numerous applications to diverse problems throughout the literature.

As the maximisation of equation 2.4 is simply a multi-dimensional optimisation problem, common packages are widely available for numerical optimisation without differentiation. Nelder & Mead (1965), Wright (1996) and Powell (1964) have been used extensively in the minimisation problems in the following chapters, and can be quick to compute. Problems can arise due to local minima in highly complex minimisation problems, which more advanced

methods such as that given in Sect. 2.2.2 can solve.

MLE has a long history in astronomy, but specific improvements by Luri et al. (1996) resolved several shortcomings in its use specifically in the field of luminosity calibration. The improvements, described in full in Luri et al. (1996), have been used extensively in this work and allow the following improvements:

- The use of numerical methods avoids the necessity of approximating or simplifying complex equations during the formulation of the MLE, thus avoiding any loss of precision.
- Explicitly taking into account sample selection effects caused by observational constraints is required to correctly model the joint probability density function (PDF) of the sample. These selection effects introduce various biases into the sample, e.g. Malmquist, and their correct treatment within the formulation of the density laws avoids the need for a posteriori corrections, which are often poorly understood.
- Through the construction of the MLE joint PDF, where all of a star's available information is used simultaneously, posterior estimates of a star's properties can be obtained with increased precision compared with the original data.

Maximum A Posteriori estimation

In cases where prior information about the model parameters is available, Maximum A Posteriori (MAP) estimation can be applied. Starting again from Bayes theorem in Eqn. 2.3, the Likelihood is defined as

$$\mathcal{L}(\boldsymbol{\theta}; \boldsymbol{x}) = P(\boldsymbol{x}|\boldsymbol{\theta})P(\boldsymbol{\theta}) \tag{2.5}$$

The prior on the parameters, $P(\boldsymbol{\theta})$, is the PDF of the parameters which, if known, provides an additional constraint on their estimation. Having priors on

the parameters, even if they are wide, has the additional benefit of limiting the search space when attempting to maximise the Likelihood $\mathcal{L}$. If a parameter is known to be, for example, positive, then priors defining a zero probability for negative values limits the search space, preventing wasted computation time. Of course, in the case of uniform priors, the MAP estimate of a set of model parameters is equivalent to that from MLE.

As with MLE, only the parameters maximising the Likelihood are found. This makes the maximisation straightforward, and the same methods as those in 2.4 can be applied for optimisation. As only the global maximum is required, the optimisation can be performed in relatively few steps, depending on the number of dimensions. However, finding only the peak of the posterior PDF of the parameters does not provide any information about the spread of the distribution (uncertainty), or any other characteristics such as asymmetry or irregularity.

Full Bayesian Inference

Evaluating the full posterior PDF of the parameters provides more information. The peak of the PDF is the MAP estimate (of MLE if no priors are used). The width of the PDF gives confidence intervals, through taking upper and lower quartiles, or by more specific methods such as batch means (Flegal & Jones, 2008). Other important properties such as asymmetric or bi-modal distributions are also easy to detect and include in any analysis.

However, evaluating the full posterior PDF can be prohibitively expensive in terms of computation time, especially in cases with a large number of dimensions. The problem of efficiently sampling a large space has been made much simpler through the combination of Markov Chain Monte Carlo (MCMC) methods with modern day computing power.

The introduction of MCMC methods in the 1940's (Robert & Casella, 2008) by the same scientists who developed regular Monte Carlo, was not very widely followed up due to the lack of computing resources at the time. Work by

Metropolis et al. (1953) and later Hastings (1970) on algorithms which overcame several of the original shortcomings of the method allowed much greater utility of MCMC methods, which have become widespread since the 1990's.

MCMC effectively builds a chain of successive iterations of a random walk through the parameter space. When evaluating a Likelihood function $\mathcal{L}$, the direction of the random walk performed by MCMC is weighted in the direction of increasing Likelihood, meaning that the sampler spends more time sampling regions of greater importance (i.e. regions of non-negligible probability in the posterior PDF of the parameters) and little to no time in regions of less importance. The key difference from regular optimisers is the fact that the sampler contains a random element which allows the chain to move away from the peak of the distribution, sampling the entire useful range (but no more). The method is much more efficient than brute force when sampling irregular distributions and in problems with a large number of dimensions.

Many modern implementations of MCMC methods exist, making its application straightforward. Here, the EMCEE algorithm (Foreman-Mackey et al., 2013) is used. EMCEE offers an efficient parallelised solution whereby a large number of initial guesses are used to create a number of parallel chains, which can interact to concentrate the chains towards regions of higher importance. This reduces the importance of good initial guesses and goes some way towards avoiding the problem of an optimiser getting stuck in local minima.

3

The Gaia catalogue

This section contains a complete overview of the expected contents of the Gaia catalogue. This has been achieved through analysis of the mock Gaia catalogue which has been simulated in order to enable perspective users of the Gaia data to get an idea of the contents and quality of the expected end-of-mission catalogue. This mock catalogue has additionally served as a tool for familiarisation with the format of the Gaia data and the use of massive datasets. The simulated data has been used extensively in Chapters 4 to 7 to test the various statistical methods with Gaia-like data. The contents of this chapter has been published as **Astronomy & Astrophysics, Volume 566, 2014, id: A119, 15 pp.**

3.1 Simulations

During its five years of data collection, Gaia is expected to transmit some 150 terabytes of raw data to Earth, leading to production of a catalogue of 10^9 individual objects. After on-ground processing, the full database is expected to be in the range of one to two petabytes of data. Preparation for acquiring this huge amount of data is essential. Work has begun to model the expected output of Gaia in order to predict the content of the Gaia Catalogue, to facilitate the production of tools required to effectively validate the real data before publication, and to analyse the real data set at the end of the mission.

To this end, the Gaia Data Processing and Analysis Consortium (DPAC) has been preparing a set of simulators, including a simulator called the Gaia Object Generator (GOG), which simulates the end-of-mission catalogue, including observational errors. Here a full description of GOG is provided, including the models assumed for the performance of the Gaia satellite and an overview of its simulated end-of-mission catalogue. A selection of statistics from this catalogue is provided to give an idea of the performance and output of Gaia.

Simulation of many aspects of the Gaia mission has been carried out in order to test and improve instrument design, data reduction algorithms, and tools for using the final Gaia Catalogue data. The Gaia Simulator is a collection of three data generators designed for this task: the Gaia Instrument and Basic Image Simulator (GIBIS, Babusiaux, 2005), the Gaia System Simulator (GASS, Masana et al., 2010), and the Gaia Object Generator (GOG, Luri et al., 2014). Through these three packages, the production of the simulated observation images down to pixel level, Gaia telemetry stream, and intermediate or final catalogue data is possible.

3.1.1 Gaia Universe Model Snapshot and the Besançon galaxy model

One basic component of the Gaia Simulator is its Universe Model (UM), which is used to create object catalogues down to a particular limiting magnitude (in our case $G = 20$ mag for Gaia). For stellar sources, the UM is based on

the Besançon galaxy model (Robin et al., 2003). This model simulates the stellar content of the Galaxy, including stellar distribution and a number of object properties. It produces stellar objects based on the four main stellar populations (thin disk, thick disk, halo, and bulge), each population with its own star formation history and stellar evolutionary models. Additionally, a number of object-specific properties are also assigned to each object, dependent on its type. Possible objects are stars (single and multiple), stellar clusters, diffuse light, planets, asteroids, comets, resolved galaxies, unresolved extended galaxies, quasars, AGN, and supernovae. Therefore, the UM is capable of simulating almost every object type that Gaia can potentially observe. It can therefore construct simulated object catalogues down to Gaia's limiting magnitude.

Building on this, the UM creates for any time, over any section of the sky (or the whole sky), a set of objects with positions and assigns each a set of observational properties (Robin et al., 2012). These properties include distances, apparent magnitudes, spectral characteristics, and kinematics.

Clearly the models and probability distributions used in order to create the objects with their positions and properties are highly important in producing a realistic catalogue. The UM has been designed so that the objects it creates fit as closely as possible to observed statistics and to the latest theoretical formation and evolution models (Robin et al., 2003). For a statistical analysis of the underlying potentially observable population (with $G \leq 20$ mag) using the Gaia UM without satellite instrument specifications and error models, see Robin et al. (2012).

3.1.2 The Gaia Object Generator

The GOG is capable of transforming this UM catalogue into Gaia's simulated intermediate and final catalogue data. This is achieved through the use of analytical and numerical error models to create realistic observational errors in astrometric, photometric, and spectroscopic parameters (Isasi et al., 2010). In

this way, GOG transforms 'true' object properties from the UM into 'observed' quantities that have an associated error that depends on the object's properties, Gaia's instrument capabilities, and the type and number of observations made.

3.1.3 Error models

DPAC is divided into a number of coordination units (CUs), each of which specialises in a specific area. In GOG we have taken the recommendations from the various CUs in order to include the most complete picture of Gaia performance as possible. The CUs are divided into the following areas: CU1, system architecture; CU2, simulations; CU3, core processing (astrometry); CU4, object processing (multiple stars, exoplanets, solar system objects, extended objects); CU5, photometric processing; CU6, spectroscopic processing; CU7, variability processing; CU8, astrophysical parameters; and CU9, catalogue access.

Models for specific parameters have been provided by the various CUs, and only an outline is given here. In the following description, *true* refers to UM data (without errors), *transit* refers to simulated individual observations (including errors), and *observed* to the simulated observed data for the end of the Gaia mission (including standard errors). Transit is defined as a single complete pass of the Gaia focal plane, crossing 9 CCDs. Throughout, *error* refers to the formal standard error on a measurement.

Astrometric error model

The formal error on the parallax, σ_ϖ, is calculated following the expression:

$$\sigma_\varpi = m \cdot g_\varpi \cdot \sqrt{\frac{\sigma_\eta^2}{N_{\text{eff}}} + \frac{\sigma_{\text{cal}}^2}{N_{\text{transit}}}} \tag{3.1}$$

- σ_η is the CCD centroid positioning error. It uses the Cramer Rao (CR) lower bound in its discrete form, which defines the best possible preci-

sion of the Maximum Likelihood centroid location estimator. The CR lower bound requires the line spread function (LSF) derivative for each sample[1], the background, the readout noise, and the source integrated signal.

- m is the contingency margin that is used to take environmental effects into account, for example: uncertainties in the on-ground processing, such as uncertainties in relativistic corrections and solar system ephemeris; effects such as having an imperfect calibrating LSF; errors in estimating the sky background; and other effects when dealing with real stars. The default value assumed for the Gaia mission has been set by ESA as 1.2 and is used in GOG.

- $g_{\varpi} = 1.47/\sin\xi$ is a geometrical factor where ξ is known as the solar aspect angle, with a value of 45°.

- N_{eff} is the number of elementary CCD transits ($N_{\text{strip}} \times N_{\text{transit}}$).

- N_{transit} is the number of field of view transits.

- N_{strip} is the number of CCDs in a row on the Gaia focal plane. It has a value of 9, except for the row that includes the wavefront sensor, which has 8 CCDs.

- σ_{cal} is the calibration noise. A constant value of 5.7 μas has been applied. This takes into account that the end-of-mission precision on the astrometric parameters not only depends on the error due to the location estimation with each CCD. There are calibration errors from the CCD calibrations, the uncertainty of the attitude of the satellite and the uncertainty on the basic angle.

We are enabling the activation of gates, as described in the Gaia Parameter Database. The Gaia satellite will be smoothly rotating and will constantly

[1]In Gaia, a sample is defined as a set of individual pixels.

image the sky by collecting the photons from each source as they pass along the focal plane. The total time for a source to pass along the focal plane will be 107 seconds, and the electrons accumulated in the a CCD pixel will be passed along the CCD at the same rate as the source (Time Delay Integration mode). To avoid saturation for brighter sources, and the resulting loss of astrometric precision, gates can be activated that limit the exposure time. Here we are using the default gating system, which could change during the mission.

Following de Bruijne (2012) (see also `http://www.cosmos.esa.int/web/gaia/science-performance`), we have assumed that the errors on the positional and proper motion coordinates can be given respectively by

- $\sigma_\alpha = 0.787\sigma_\varpi$
- $\sigma_\delta = 0.699\sigma_\varpi$
- $\sigma_{\mu_\alpha} = 0.556\sigma_\varpi$
- $\sigma_{\mu_\delta} = 0.496\sigma_\varpi$

Photometric error model

GOG uses the single CCD transit photometry error $\sigma_{\mathrm{p,j}}$ (Jordi et al., 2010) defined as

$$\sigma_{\mathrm{p,j}}[\mathrm{mag}] = 2.5 \cdot \log_{10}(e) \cdot \frac{\sqrt{f_{\mathrm{aperture}} \cdot s_{\mathrm{j}} + (b_{\mathrm{j}} + r^2) \cdot n_s \cdot (1 + \frac{n_{\mathrm{s}}}{n_{\mathrm{b}}})}}{f_{\mathrm{aperture}} \cdot s_{\mathrm{j}}} \tag{3.2}$$

to compute either the single transit errors or the end-of-mission errors.

We assume, following an 'aperture photometry' approach, that the object flux s_{j} is measured in a rectangular 'aperture' (window) of n_{s} along-scan object samples. The sky background b_{j} is assumed to be measured in n_{b} background samples, and r is CCD readout noise. The $f_{\mathrm{aperture}} \cdot s_{\mathrm{j}}$ is expressed in units of photo-electrons (e^-), and denotes the object flux in photometric band j

contained in the 'aperture' (window) of $n_{\rm s}$ samples, after a CCD crossing. The number $f_{\rm aperture}$ thus represents the fraction of the object flux measured in the aperture window.

For the single transit and end-of-mission data, the same expression is used for the standard deviation calculation

$$\sigma_{\rm G,j} = m \cdot \sqrt{\frac{\sigma_{\rm p,j}^2 + \sigma_{\rm cal}^2}{N_{\rm eff}}} \tag{3.3}$$

where $N_{\rm eff}$ is the number of elementary CCD transits ($N_{\rm strip} \times N_{\rm transit}$), with ($N_{\rm strip} = 1$ and $N_{\rm transit}$) $= 1$ for a single transit and equal to the number of transits for the end-of-mission photometry. The calibration noise $\sigma_{\rm cal}$ has a fixed value of $\sigma_{\rm cal} = 30$ mmag. Use of a fixed value of the centroiding error is possible because this error is only relevant for brighter stars, due to their centroiding errors being smaller than the calibration error.

Radial velocity spectrometer errors

CU6 tables (see Table 3.1) using the stars' physical parameters and apparent magnitude are used to obtain $\sigma_{V_{\rm r}}$. They were computed following the prescriptions of Katz et al. (2004) and later updates. Those tables have been provided for one and 40 field-of-view transits, therefore the value for 40 transits is used here to calculate the average end-of-mission errors in RVS.

Given the information on the apparent Johnson V magnitude and the atmospheric parameters of each star (from the UM), we select from Table 3.1 the closest spectral type and return the corresponding radial velocity error. Since Table 3.1 is given for [Fe/H]=0 alone, a correction is made on the apparent magnitude in order to take different metallicities into account: for each variation in metallicity of Δ[Fe/H] $= -1.5$ dex, the magnitude is increased by $V = 0.5$ mag.

We have set a lower limit on the wavelength calibration error, giving a lower limit on the radial velocity error of 1 km·s^{-1}. For the faintest stars

the spectra will be of poor quality and will not contain enough information to enable accurate estimation of the radial velocity. Owing to the limited bandwidth in the downlink of Gaia data to Earth, poor quality spectra will not be transmitted. We therefore set an upper limit on the radial velocity error of 20 km·s^{-1}, beyond which we assume that there will be no data. The exact point at which the data will be assumed to have too low a quality is still unknown[2].

Physical parameters

GOG uses the stellar parametrisation performance given by CU8 to calculate error estimations for effective temperature, line-of-sight extinction, metallicity, and surface gravity. The colour-independent extinction parameter A_0 is used in preference to the band specific extinctions A_V or A_G, because A_0 is a property of the interstellar medium alone (Bailer-Jones, 2011). CU8 use three different algorithms to calculate physical parameters using spectrophotometry (see Liu et al. (2012)).

It should be noted that the errors calculated here are calculated only as a function of apparent magnitude. However, as described in Liu et al. (2012), there are clear dependencies on the spectral type of the star, because some star types may or may not exhibit spectral features required for parameter determination. Additionally, Liu et al. (2012) report a strong correlation between the estimation of effective temperature and extinction. This correlation is not simulated in GOG. Following the recommendation of CU8, calculating errors of physical parameters depends on apparent magnitude and is split into two cases, objects with $A_0 < 1$ mag and $A_0 \geq 1$ mag.

In GOG, $\sigma_{T_{\rm eff}}$, σ_{A_0}, $\sigma_{Fe/H}$ and $\sigma_{\log g}$ are calculated from a Gamma distribu-

[2]And is expected to change due to mitigation efforts of the Gaia stray light issue outlined in Sect. 3.1.4

Type \ V	8.5	9.0	9.5	10	10.5	11.5	12	12.5	13	13.5	14	14.5	15	15.5	16	16.5	17	17.5
B0V	1.2	1.6	2	2.7	3.8	6.8	9.7	14.5	24.8	n	n	n	n	n	n	n	n	n
B5V	1	1.1	1.4	1.9	2.5	5.1	6.9	10	15.3	24.1	n	n	n	n	n	n	n	n
A0V	1	1	1	1	1	1.3	1.8	2.6	3.9	5.7	8.6	14.6	32.5	n	n	n	n	n
A5V	1	1	1	1	1	1	1	1.3	2	4.2	6.9	11.1	20.1	n	n	n	n	n
F0V	1	1	1	1	1	1	1	1	1.5	2.1	3.2	5.3	7.8	12.7	23.4	n	n	n
G0V	1	1	1	1	1	1	1	1	1	1.4	2.1	3	4.8	7.9	12.4	19.6	n	n
G5V	1	1	1	1	1	1	1	1	1	1.2	1.9	2.8	4.4	6.3	10.1	17.6	n	n
K0V	1	1	1	1	1	1	1	1	1	1.1	1.4	2.1	3.3	5.1	8.1	12.6	24.9	n
K4V	1	1	1	1	1	1	1	1	1	1	1.1	1.6	2.7	3.6	5.2	8.4	14.5	30
K1III	1	1	1	1	1	1	1	1	1	1	1	1.2	1.8	2.7	4.2	6.8	10.3	18

Table 3.1: The average end-of-mission formal error in radial velocity (based on Katz et al. (2004)) with an assumed average of 40 field-of-view transits, in km·s^{-1}, for each spectral type. The numbers in the top row are Johnson apparent V magnitudes. Fields marked by "n" are assumed to be too faint to produce spectra with sufficient quality for radial velocity determination. Stars with these magnitudes will have no radial velocity information.

tion, with shape parameter α and scale parameter θ :

$$f(\sigma;\alpha,\theta) = \frac{1}{\Gamma(\sigma)\theta^{\alpha}}\sigma^{\alpha-1}e^{-\frac{x}{\theta}} \tag{3.4}$$

where α and θ are obtained from the following expressions, which have been calculated to give each σ a close approximation to the CU8 algorithm results. A gamma distribution was selected for ease of implementation and for its ability to statistically recreate the CU8 results to a reasonable approximation. A gamma distribution is also only non-zero for positive values of sigma. This is essential when modelling errors because, of course, it is impossible to have a negative error. The following Gamma distribution parameters are from an E. Antiche internal communication:

- For stars with $A_0 < 1$ mag:

$$\alpha_{A_0} = 0.204 - 0.032G + 0.001G^2$$

$$\alpha_{\log g} = 0.151 - 0.019G + 0.001G^2$$

$$\alpha_{Fe/H} = 0.295 - 0.047G + 0.002G^2$$

$$\alpha_{T_{\rm eff}} = 78.2 - 10.3G + 0.46G^2$$

$$\theta_{A_0} = 0.084$$

$$\theta_{\log g} = 0.160$$

$$\theta_{Fe/H} = 0.121$$

$$\theta_{T_{\rm eff}} = 28.2.$$

- For stars with $A_0 \geq 1$ mag:

$$\alpha_{A_0} = 0.178 - 0.026G + 0.001G^2$$

$$\alpha_{\log g} = 0.319 - 0.044G + 0.002G^2$$

$$\alpha_{Fe/H} = 0.717 - 0.115G + 0.005G^2$$

$$\alpha_{T_{\rm eff}} = 67.3 - 7.85G + 0.35G^2$$

$$\theta_{A_0} = 0.096$$

$$\theta_{\log g} = 0.179$$

$$\theta_{Fe/H} = 0.353$$

$$\theta_{T_{\rm eff}} = 33.5.$$

The Gamma distributions thus obtained for each parameter are used to generate a formal error for each parameter for each individual star, aiming to statistically (but not individually) reproduce the results that will be obtained by the application of the CU8 algorithms and then included in the Gaia Catalogue.

It should be noted that in the stellar parametrisation algorithms used in Liu et al. (2012), a degeneracy is reported between extinction and effective temperature owing to the lack of resolved spectral lines only sensitive to effective temperature. In GOG, this degeneracy has not been taken into account, and the precisions of each of the four stellar parameters is simulated independently.

Additionally, the results of Liu et al. (2012) have recently been updated, and Bailer-Jones et al. (2013) gives the latest results regarding the capabilities of physical parameter determination. This latest paper has not been included in the current version of GOG.

3.1.4 Limitations

Since the publication of the work in this chapter, several anomalies have been detected in the functioning of the Gaia satellite. There is a stray light issue effecting the Gaia focal plane, which will impact the precision of Gaia's measurements, particularly for fainter stars. Additionally, problems of con-

tamination of the optics are occurring. The contamination is suspected to be caused to the build up of a fine layer of ice due to the presence of water outgassing from some part of the satellite. Finally, there are some unexpected variations in the basic-angle-monitoring system, which measures the angle of separation between the two telescopes.

These three issues were unexpected, and their effect and mitigation strategies are still not fully known. The stray light is expected to decrease astrometric performance for the faintest stars by a factor of around 50%, while brighter stars will be largely unaffected. The RVS spectrometer will be badly effected, potentially losing around one magnitude from its faint limit. The contamination problem is being treated by heating up the on board optics, and the extent and duration of any effects are still being calculated, along with future plans for further decontamination. A cause for the basic angle variation is still under investigation by teams at ESA.

Due to the fact that these problems were unexpected and their causes and effects are still largely unknown, the GOG simulator only includes error models for the nominal Gaia mission. The reader should therefore be aware that there will potentially be a strong degradation in the precision of measurements for fainter stars in the real Gaia catalogue, compared with what has been reported here. When the effects have been studied and quantified, it is likely that new error models will be included in GOG in order to provide updated information about the expected contents of the catalogue. This would allow a comparison between the nominal mission and the real satellite performance, and enable a quantitative view of any degradation in performance.

In the present work, only the results for single stars are given in detail. All of the figures and tables represent the numbers and statistics of only individual single stars, excluding all binary and multiple systems. Since the performance of the CU4 processing of the data is largely unknown for binary and multiple systems, the implementation into GOG of realistic error models has not yet been possible. While the results presented in Sect. 3.2 are expected to be reliable under current assumptions for the performance of Gaia, the real Gaia

Catalogue will differ from these results due to the presence of binary and multiple systems. By removing binaries from the latter, direct comparison of the results presented here with the real Gaia Catalogue will not be possible because of the presence of unresolved binaries, which are difficult to detect. As a simulator, GOG relies heavily on all inputs and assumptions supplied both from the UM or from the Gaia predicted performance and error models.

In our simulations we used an exact cut at $G = 20$ mag, beyond which no stars are observed. In reality, in regions of low density observations of stars up to 20.5 mag could be possible. Inversely, very crowded regions may not be complete up to 20 mag, or the numbers of observations per star over the five-year mission may be reduced in these regions.

There is no simulation of the impact of crowding on object detection or the detection of components in binary and multiple systems. This can lead to unrealistic quality in all observed data in the most crowded regions of the plane of the Galaxy, to overestimates for star counts in the bulge, and to a lack of features related to the disk and bulge in Figs. 3.4 and 3.21.

Additionally, GOG uses the nominal Gaia scanning law to calculate the number of field of view transits per object over the five years of the mission while Gaia is operating in normal mode. There will be an additional one-month period at the start of the mission using an ecliptic pole scanning law, and this has not been taken into account. It may lead to a slight underestimation of the number of transits, and therefore a slight overestimation of errors, for some stars near the ecliptic poles.

There is the possibility that the Gaia mission will be extended above the nominal five-year mission. Since this idea is under discussion and has not yet been confirmed or discarded, we only present results for the Gaia mission as originally planned.

If the length of the mission is increased, the number of field-of-view transits will increase, and the precision per object will improve. If the proposal is accepted, the GOG simulator could be used to provide updated statistics for the expected catalogue without extensive modification.

3.1.5 Methods and statistics

Considering current computing capabilities, it is not straightforward to make statistics and visualisations when dealing with catalogues of such a large size. A specific tool has been created by DPAC which is capable of extracting information and visualising results, with excellent scalability allowing its use for huge datasets and distributed computing systems.

The Gaia Analysis Tool (GAT) is a data analysis package that allows, through three distinct frameworks: the generation of statistics, validation of data, and generation of catalogues. It currently handles both UM and GOG generated data, and could be adapted to handle other data types.

Every statistical analysis is performed by a Statistical Analysis Module (SAM), with several grouped into a single XML file as an input to GAT. Each SAM can contain a set of filters, enabling analysis of specific subsets of the data. This allows the production of a wide range of statistics for objects satisfying any number of specific user-defined criteria or for the catalogue as a whole.

GAT creates a number of different statistics outputs including histograms, sky density maps and HR diagrams. After the GAT execution, statistics output are stored to either generate a report or to be analysed using the GAT Displaying tool.

Because we have information from not only observations of a population but also of the observed population itself, comparison is possible between the simulated Gaia Catalogue and the simulated 'true' population, allowing large scope for investigating the precision[3] of the observations and differences between the two catalogues. Clearly this is only possible with simulated data and cannot be attempted with the true Gaia Catalogue, so it is an effective way to investigate the possible extent and effect of observational errors and

[3]Here we assume the standard definition of accuracy and precision: accuracy is the closeness of a result (or set) to the actual value, i.e. it is a measure of systematics or bias. Precision is the extent of the random variability of the measurement, i.e. what is called observational errors above.

selection bias on the real Gaia Catalogue, where this kind of comparison is not possible.

GOG can be used in the preparations for validating the true Gaia Catalogue, by testing the GOG catalogue for accuracy[3] and precision. In special cases, observational biases could even be implemented into the code to allow thorough testing of validation methods.

3.2 A statistical analysis

The GOG simulator has been used to generate the simulated final mission catalogue for Gaia, down to magnitude $G = 20$. The simulation was performed on the MareNostrum super computer[4] at the Barcelona Supercomputing Centre (Centre Nacional de Supercomputació), and it took 400 thousand CPU hours. An extensive set of validations and statistics have been produced using GAT to validate performance of the simulator. Below we include a subset of these statistics for the most interesting cases to give an overview of the expected Gaia Catalogue.

3.2.1 General

In total, GOG has produced a catalogue of about one billion objects, consisting of 523 million individual single stars and 484 million binary or multiple systems. The total number of stars, including the components of binary and multiple systems is 1.6 billion. The skymap of the total flux detected over the entire five-year mission is given in Fig. 3.1. Although GOG can produce extragalactic sources, none have been simulated here.

The following discussion is split into sections for different types of objects of interest. Section 3.2.2 covers all galactic stellar sources. Section 3.2.3 covers all variable objects. Section 3.2.4 is a discussion of physical parameters estimated

[4]The authors thankfully acknowledge the computer resources, technical expertise and assistance provided by the Red Española de Supercomputación.

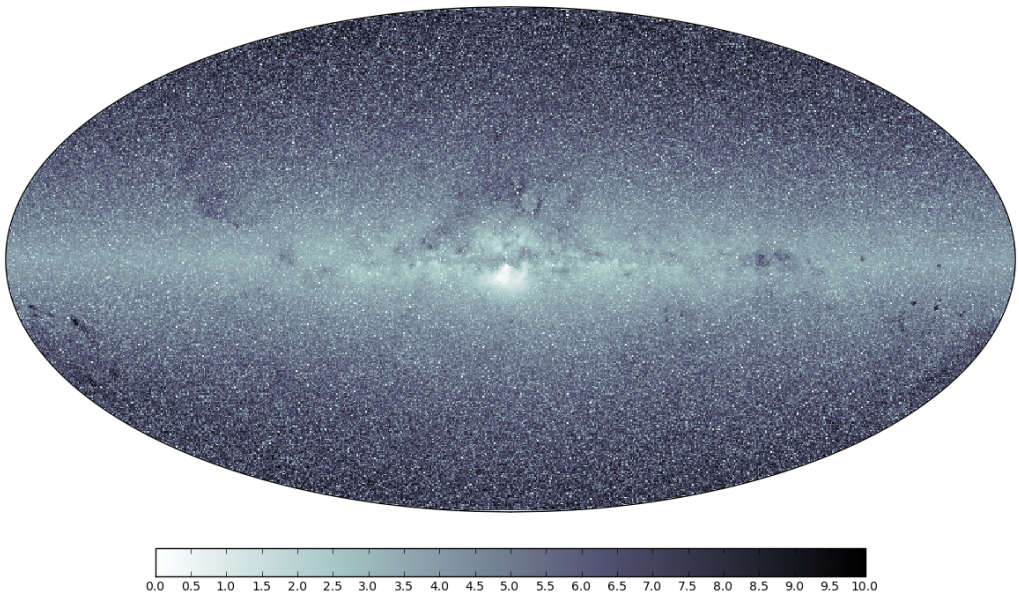


Figure 3.1: Skymap of total integrated flux over the Milky Way, in the G band. The colour bar represents a relative scale, from maximum flux in white to minimum flux in black. The figure is plotted in galactic coordinates with the galactic-longitude orientation swapped left to right.

by Gaia. All objects in these sections are within the Milky Way.

To make the presentation of performance as clear as possible, all binary and multiple systems and their components have been removed from the following statistics. This is due to complicating effects that arise when predicting the performance of the data processing of orbital parameters of binary systems, some of which do not yet have a quantitative description given by the relevant CUs. For example, GOG does not yet contain an orbital solution in its astrometric error models, and the effects of unresolved systems on photometry and astrophysical parameter determination have not yet been well determined. Therefore, the numbers presented are only for individual single stars and therefore around half of the full one billion objects simulated. Of the single stars

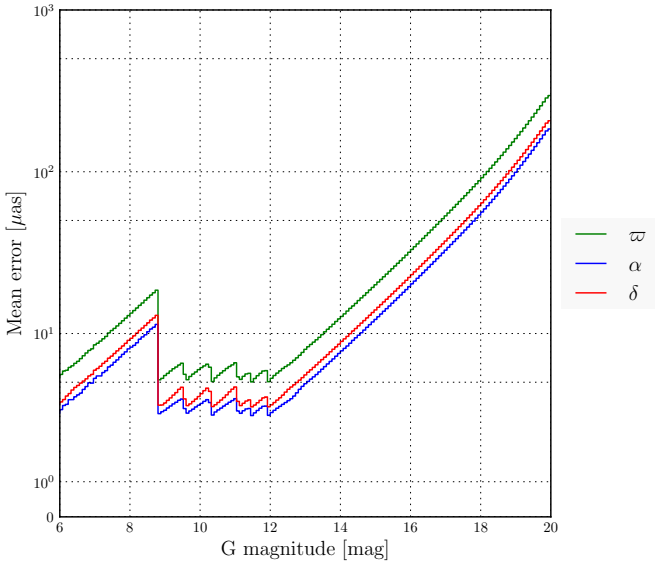
Standard error	LQ	Median	Mean	UQ
Parallax (μas)	80	140	147	210
α^* (μas)	40	80	91	130
δ (μas)	50	100	103	150
μ_α (μas·yr^{-1})	40	80	82	120
μ_δ (μas·yr^{-1})	40	70	73	110
G (mmag)	2	3	3.0	4
G_{BP} (mmag)	6	11	14.6	19
G_{RP} (mmag)	5	7	7.7	10
G_{RVS} (mmag)	6	11	13.2	18
Radial velocity (km·s^{-1})	3	7	8.0	13
Extinction (mag)	0.16	0.21	0.21	0.26
Metallicity (Fe/H)	0.46	0.57	0.57	0.73
Surface gravity ($\log g$)	0.34	0.35	0.45	0.58
Effective temperature (K)	280	350	388	530

Table 3.2: Mean and median value of the end-of-mission error in each observable, along with the upper (UQ) and lower (LQ) 25% quartile. Since the error distributions are not symmetrical, the mean value should not be used directly, and is given only to give an idea of the approximate level of Gaia's precision. The median G magnitude of all single stars is 18.9 mag.

presented in this work, 74 million are within the radial velocity spectrometer magnitude range.

Table 3.2 gives the mean and median error for each of the observed parameters discussed in this work, along with the upper and lower 25% quartile.

In Figs. 3.2 and 3.3, the mean error for parallax, position, proper motion, and photometry in the four Gaia bands are given as a function of G magnitude. Also, the mean error in radial velocity is given as a function of G_{RVS} magnitude. The sharp jumps in the mean error in astrometric parameters between 8 and 12 mag are due to the activation of gates for the brighter sources in an attempt to prevent CCD saturation (see Sect. 3.1.3).



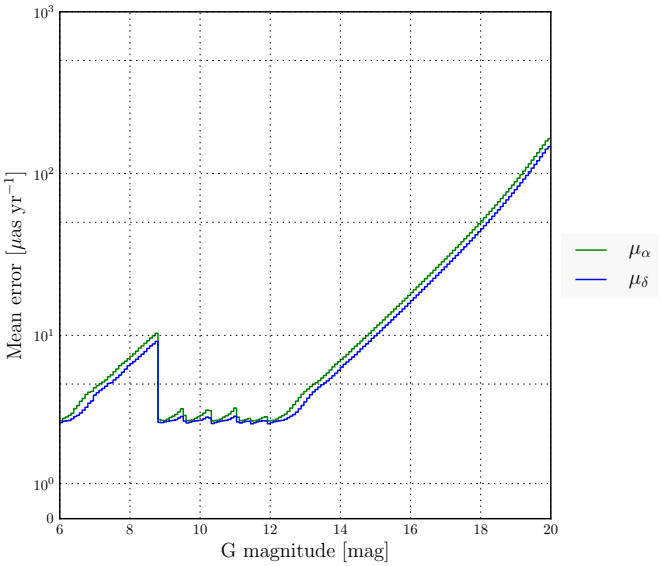


*Figure 3.2: Mean end-of-mission error as a function of G magnitude for parallax and position (*top*), and proper motion (*bottom*).*

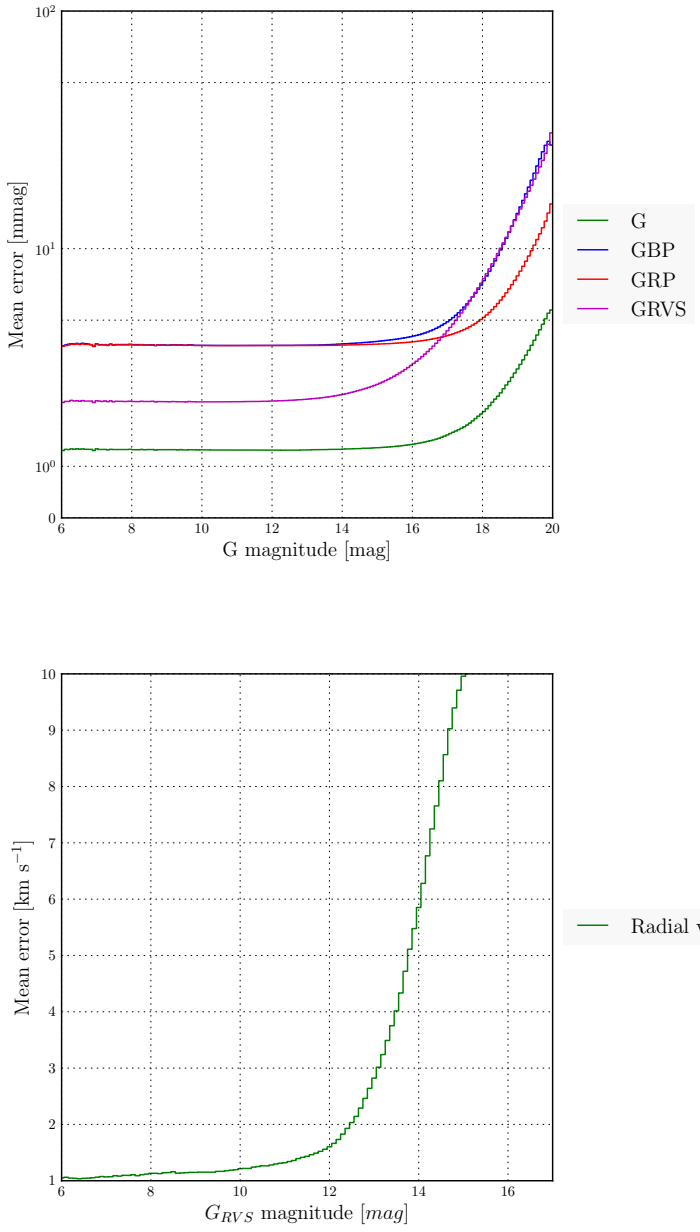


*Figure 3.3: Mean end-of-mission error as a function of G magnitude for photometry in the four Gaia passbands (*top*), and the mean end-of-mission error in radial velocity as a function of G_{RVS} magnitude (*bottom*).*

3.2.2 Stars

Parallax

The error in parallax measurements for Gaia depends on the magnitude of the source, the number of observations made, and the true value of the parallax. Figure 3.4 shows the mean parallax error over the sky. Its shape clearly follows that of the Gaia scanning law. The red area corresponding to the region of worst precision is due to the bulge population, which suffers from high levels of reddening. The faint ring around the centre of the figure corresponds to the disk of the Galaxy, remembering that the plot is given in equatorial coordinates. The blue areas corresponding to regions of improved mean precision are areas with a higher number of observations. The characteristic shape of this plot is due to the Gaia scanning law.

The distribution of parallax measurements for all single stars is given in Fig. 3.5. Note that there is a significant number of stars with negative parallax measurements. Negative parallax measurements arise due to the scatter caused by observational errors being larger than the parallax itself. While non-physical, negative parallaxes do contain information and methods utilising parallax information should include these stars to avoid biasing the result.

The relative parallax error σ_{ϖ}/ϖ is given in Fig. 3.6 for stars split by spectral type. This cumulative histogram shows the number of stars at or below each relative parallax error, and allows comparison of the relative number of each spectral type expected in the catalogue. The distribution of parallax errors is given for each stellar population in Fig. 3.7 and for each spectral type in Fig. 3.8. The error in parallax as a function of measured G magnitude is given in Fig. 3.9 and as a function of the real parallax in Fig. 3.10.

The mean parallax error for all single stars is 147.0 μas. The number of single stars falling below three relative parallax error limits is given in Table 3.3 for each spectral type and in Table 3.4 for each luminosity class. For those interested in a specific type of star, Table 3.5 gives the full breakdown of the number of single stars falling below three relative parallax error limits for every

spectral type and luminosity class.

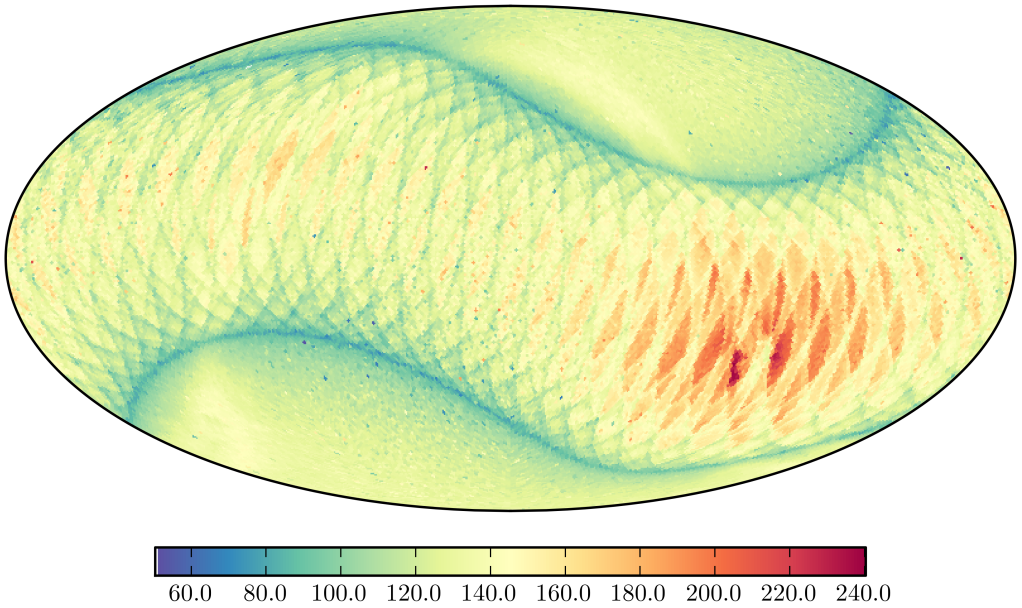


Figure 3.4: Sky map (healpix) of mean parallax error for all single stars in equatorial coordinates. Colour scale is mean parallax error in μas. The red area is the location of the bulge.

Spec. type	Total	$\sigma_\varpi/\varpi < 1$	$\sigma_\varpi/\varpi < 0.2$	$\sigma_\varpi/\varpi < 0.05$
O	3.3×10^2	87.5	57.5	29.2
B	3.4×10^5	74.0	33.0	12.2
A	5.3×10^6	79.7	38.0	14.7
F	1.2×10^8	66.2	20.1	6.1
G	2.0×10^8	67.4	20.0	5.7
K	1.5×10^8	82.4	30.9	8.6
M	4.5×10^7	98.1	68.0	18.6

Table 3.3: Total number of single star for each spectral type, along with the percentage of those that fall below each relative parallax error limit; e.g., 68% of M-type stars have a relative parallax error better than 20%.

Lum. class	Total	$\sigma_\varpi/\varpi < 1$	$\sigma_\varpi/\varpi < 0.2$	$\sigma_\varpi/\varpi < 0.05$
Supergiant	5.6×10^3	91.5	65.9	36.8
Bright giant	6.9×10^5	87.1	57.9	25.1
Giant	6.6×10^7	67.5	21.4	7.0
Sub-giant	7.5×10^7	60.2	16.8	5.3
Main sequence	3.8×10^8	78.1	30.7	8.5
White dwarf	2.1×10^5	100.0	94.3	41.9

Table 3.4: Total number of single stars for each luminosity class, along with the percentage that fall below each relative parallax error limit.

Type	Total	$\sigma_\varpi/\varpi < 1$	$\sigma_\varpi/\varpi < 0.2$	$\sigma_\varpi/\varpi < 0.05$
OII	4	100	100	75
OIII	17	100	65	35
OIV	26	96	50	31
OV	203	89	57	27
BII	80	91	50	21
BIII	1.5×10^5	58	19	8
BIV	1.1×10^5	81	42	17
BV	2.1×10^5	81	38	13
AII	6.7×10^3	79	42	19
AIII	9.5×10^5	76	37	16
AIV	1.4×10^6	80	40	16
AV	2.7×10^6	81	37	14
FII	2.0×10^3	81	46	22
FIII	1.7×10^6	76	33	12
FIV	3.6×10^7	67	22	7
FV	7.9×10^7	66	19	5
GII	1.6×10^5	81	43	20
GIII	2.0×10^7	61	17	5
GIV	3.7×10^7	53	11	3
GV	1.5×10^8	72	23	6
KII	2.5×10^5	81	43	19
KIII	4.0×10^7	69	21	7
KV	1.1×10^8	87	34	9
MII	1.1×10^4	89	59	32
MIII	2.2×10^6	85	46	16
MV	4.2×10^7	99	70	19
WD	2.1×10^5	100	94	42

Table 3.5: Total number of single stars for each stellar classification, along with the percentage that fall below each relative parallax error limit.

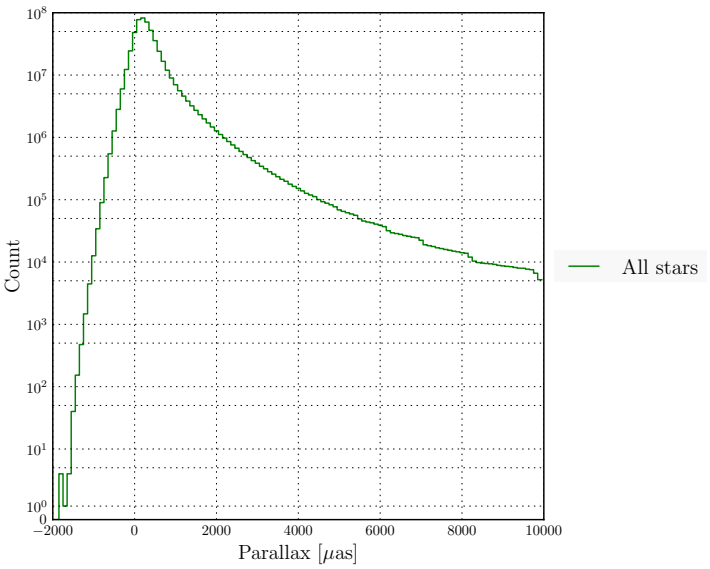


Figure 3.5: Histogram of parallax for all single stars. The histogram contains 99.5% of all data.

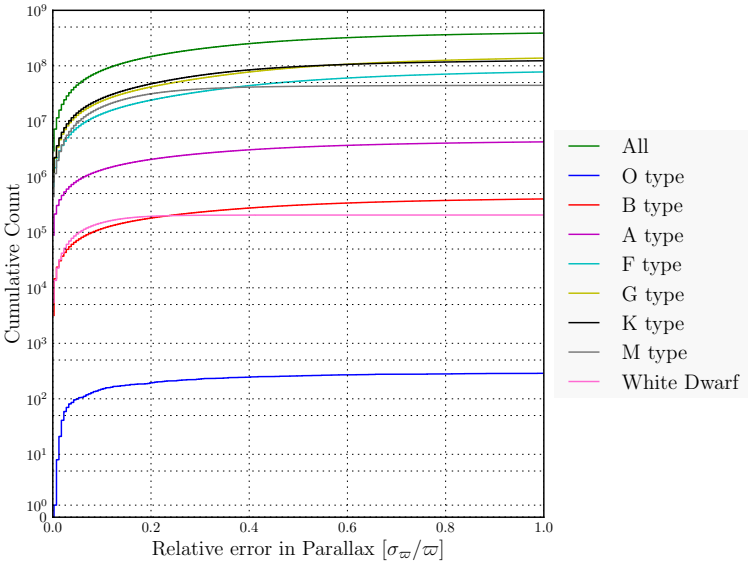


Figure 3.6: Cumulative histogram of relative parallax error for all single stars, split by spectral type. The histogram range displays 74% of all data.

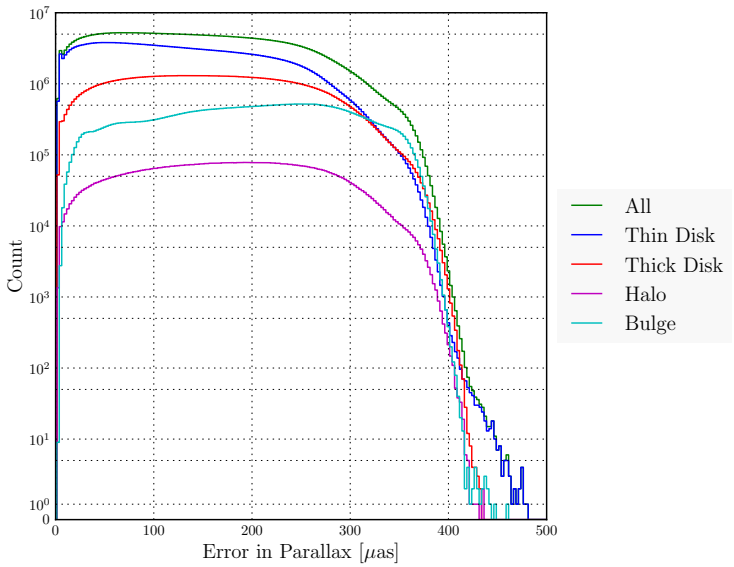


Figure 3.7: Histogram of end-of-mission parallax error for all single stars, split by stellar population.

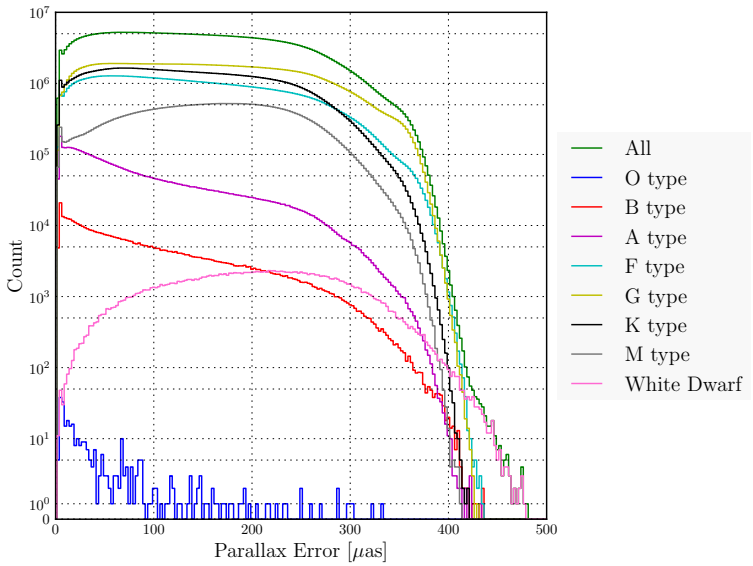


Figure 3.8: Histogram of end-of-mission parallax error for all single stars, split by spectral type.

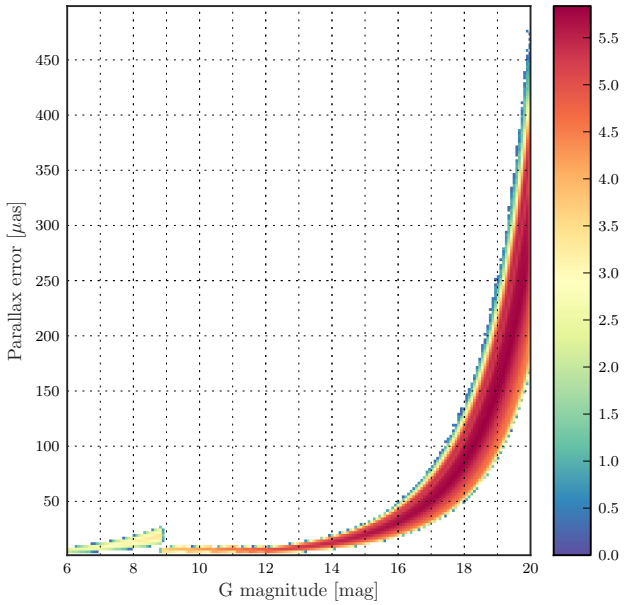


Figure 3.9: End-of-mission parallax error against G magnitude for all single stars. The colour scale represents the log of density of objects in a bin size of 80 mmag by 2.5 μas. White area represents zero stars.

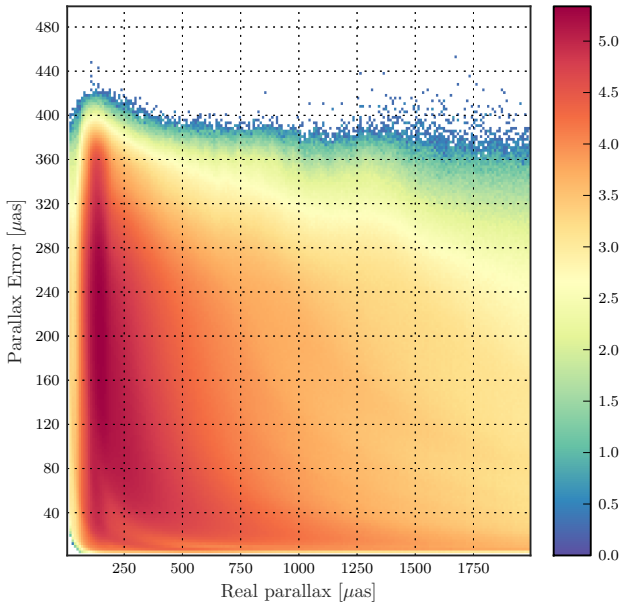


Figure 3.10: End-of-mission parallax error against parallax for all single stars. The colour scale represents the log of density of objects in a bin size of 10 by 2.5 μas. White area represents zero stars.

Position

Gaia will be capable of measuring the position of each observed star at an unprecedented accuracy, producing the most precise full sky position catalogue to date. Position information will also be amongst the data that will form the first preliminary data release, expected in the summer of 2016 (http://www.cosmos.esa.int/web/gaia/release).

The mean error is 90 μas for right ascension and 103 μas for declination. The distribution of error in right ascension and declination as a function of the true value, along with a histogram of the error, are given in Figs. 3.11 and 3.12. The overdensities are due to the bulge of the Galaxy.

Proper motion and radial velocity

In addition to parallax measurements, Gaia will also measure proper motions for all stars it detects. The proper motion in right ascension and declination is labelled as μ_α and μ_δ, respectively. The mean error in μ_α is 81.7 μas·yr^{-1}, and in μ_δ is 72.9 μas·yr^{-1}. The distribution of errors in both components of proper motion is given in Fig. 3.13. The first preliminary data release is expected to contain improved proper motion data for the stars of the Hipparcos catalogue. This preliminary release will combine the hipparcos data with Gaia data collected and processed by the time of release in order to benefit from the long baseline between the two sets of observations.

The radial velocity is measured by the on-board radial velocity spectrometer. This instrument is only sensitive to stars down to $G_{RVS} = 16$ magnitude. We assume an upper limit on the error in radial velocity of 20 km·s^{-1}, and assume that stars with a precision worse than this will not be given any radial velocity information.

Of the 523 million measured individual Milky Way stars, 74 million have a radial velocity measurement[5]. The mean error in the radial velocity mea-

[5]This is expected to change due to the effect of stray light issues onboard Gaia.

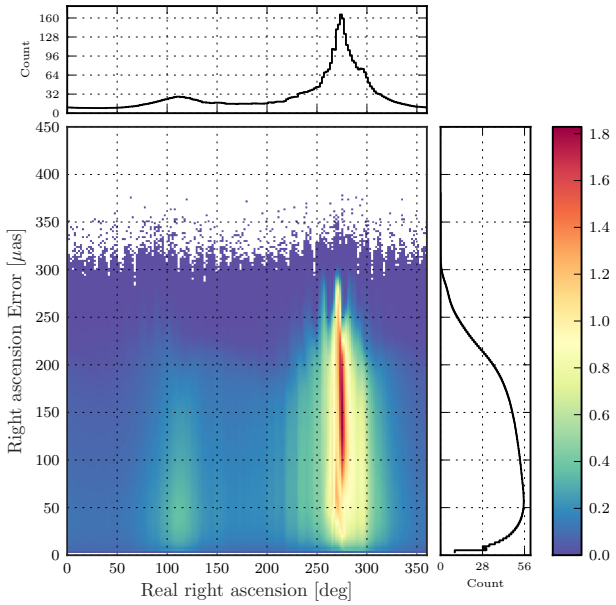


Figure 3.11: Right ascension error against real right ascension. The colour scale is linear, with a factor of 10^5. Histograms are computed for both right ascension and right ascension error. The colour scale represents log density of objects in a bin size of 2 degrees by 7.5 μas. White area represents zero stars.

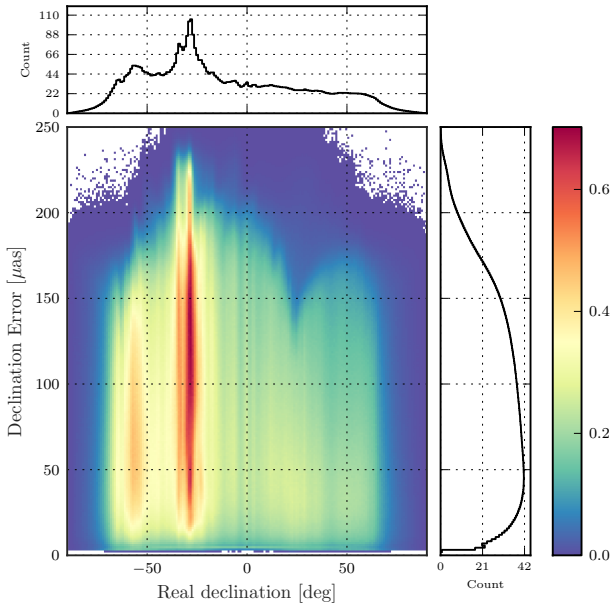


Figure 3.12: Declination error against real declination. The colour scale is linear, with a factor of 10^5. Histograms are computed for both declination and declination error. The colour scale represents log density of objects in a bin size of 1 degrees by 5 μas. White area represents zero stars.

surement is 8.0 km·s^{-1}. The distribution of radial velocity error is given for each G magnitude in Fig. 3.14, and in Fig. 3.15 split by spectral type. The radial velocity error is given as a function of G_{RVS} magnitude in Fig. 3.16. Note that due to the stray light issues highlighted in Sect. 3.1.4, it is expected that there will be a strong degradation in the RVS spectrometer results. Due to the effects of the stray light, the spectra obtained for the faintest stars will be badly effected and it expected that the limiting magnitude will be reduced by around one magnitude. The spectrometer was designed to run in both high resolution and low resolution mode, and was planned to switch between modes depending on the strength of the source. It is likely that the instrument will be used only in HR mode in order to mitigate some of the issues. The exact extent of the effect is still under investigation. Check http://www.cosmos.esa.int/web/gaia/science-performance for the latest information.

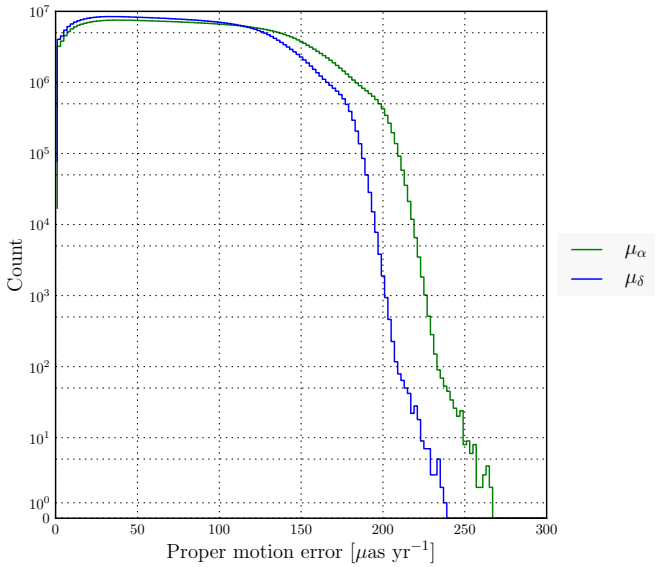


Figure 3.13: Error in proper motion for alpha and delta for all single stars.

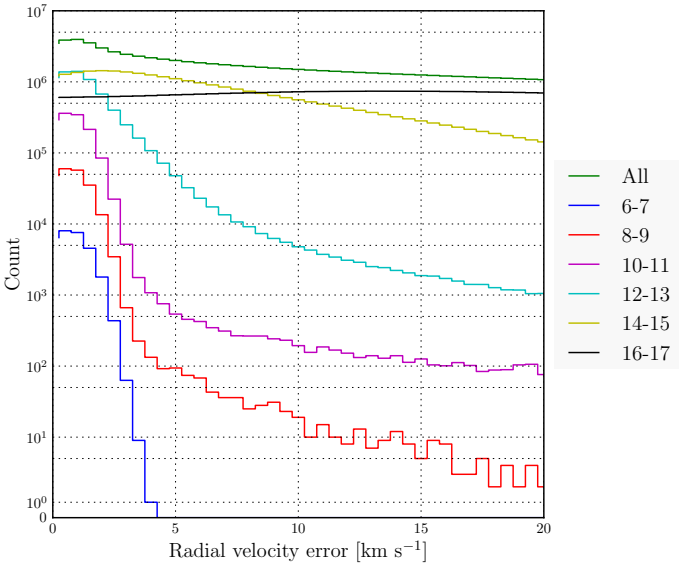


Figure 3.14: Histogram of radial velocity error split by G magnitude range. The histogram contains 100% of all data that have radial velocity information.

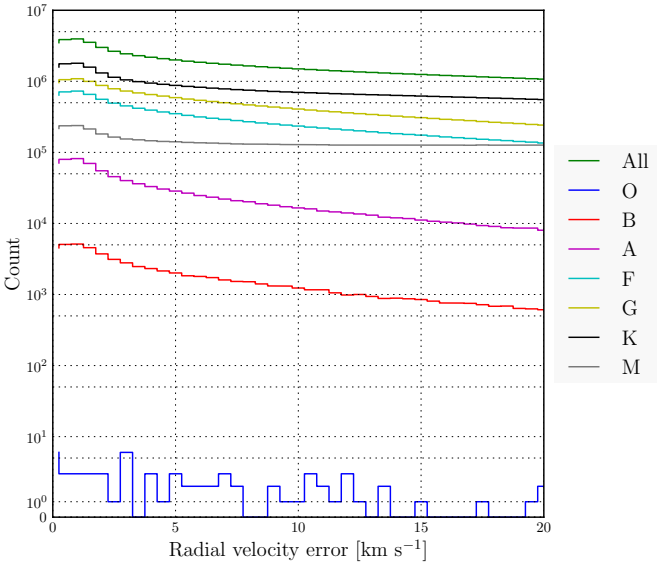


Figure 3.15: Histogram of radial velocity error split by spectral type. The histogram contains 100% of all data that have radial velocity information.

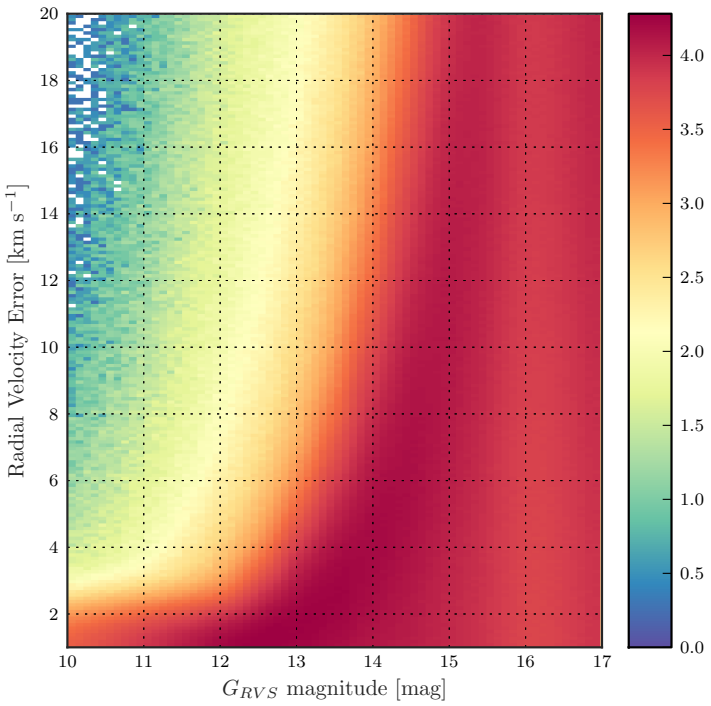


Figure 3.16: End-of-mission error in radial velocity against G_{RVS} magnitude. The colour scale represents log density in a bin size of 50 mmag by 1 km·s^{-1}. White area represents zero stars.

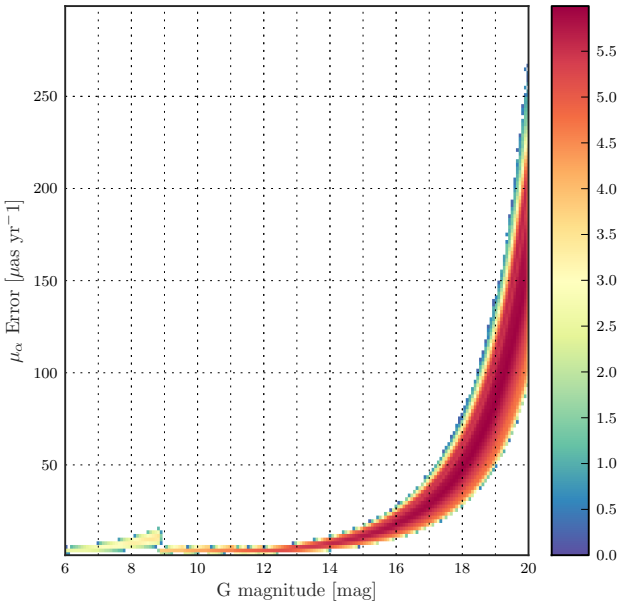


Figure 3.17: 2D histograms showing error in proper motion in alpha against G magnitude. The colour scale represents log density of objects in a bin size of 80 mmag by 2 $\mu as{\cdot}yr^{-1}$. White area represents zero stars.

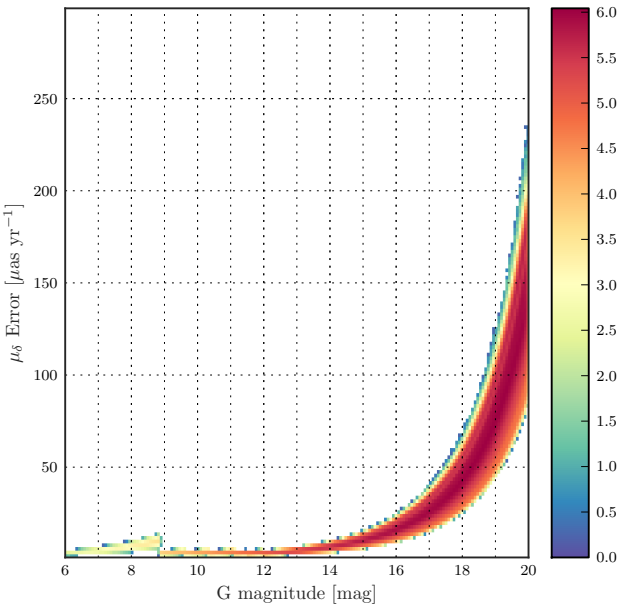


Figure 3.18: 2D histograms showing error in proper motion in delta against G magnitude. The colour scale represents log density of objects in a bin size of 80 mmag by 2 $\mu as{\cdot}yr^{-1}$. White area represents zero stars.

Photometry

The end-of-mission error in each measurement as a function of G magnitude is given in Fig. 3.19.

Gaia will produce low-resolution spectra, in addition to measuring the magnitude of each source in the Gaia bands G, G_{BP}, G_{RP}, and G_{RVS}. Whilst GOG is capable of simulating these spectra, they have not been included in the present simulations owing to the long computation time and the large storage space requirement of a catalogue of spectra for one billion sources.

Figure 3.20 shows the distribution in the error of each photometric measurement. As can be seen in this figure, the error in G is much lower than for the other instruments, and for all stars it is less than 8 mmag. The mean error in G is 3.0 mmag. The mean error in G_{BP} and G_{RP} is 14.6 mmag and 7.7 mmag, respectively. The mean error in G_{RVS} is 13.2 mmag, although it must be remembered that the radial velocity spectroscopy instrument is limited to brighter than $G_{RVS} = 16$.

Figure 3.21 shows the mean photometric error as a function of position on the sky for the four Gaia photometric passbands. The structure seen in all four maps is derived from the Gaia scanning law.

It is interesting to point out the ring in the four plots of Fig. 3.21 caused by the disk of the Galaxy. Owing to significant levels of interstellar dust in the disk of the Galaxy, visible objects are generally much redder. This reddening causes objects to lose flux at the bluer end of the spectrum, making them appear fainter to the G_{BP} photometer. Therefore the plane of the Galaxy can be seen as an *increase* in the mean photometric error in the G_{BP} error map.

Conversely, the disk of the Galaxy shows as a ring of *decreased* mean photometric error in the G_{RP} and G_{RVS} maps, since the sensitivity of their spectra is skewed more towards the redder end of the spectrum. It is important to note, however, that the effect of crowding on photometry is not accounted for in GOG.

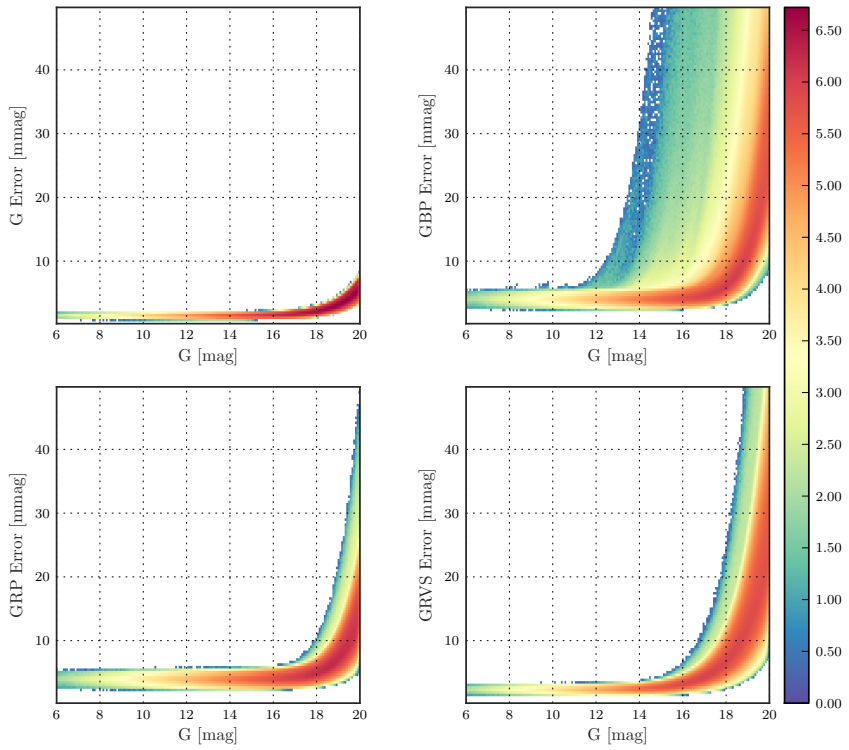


Figure 3.19: End-of-mission errors in photometry as a function of G magnitude. The colour scale represents log density of objects in a bin size of 80 mmag by 0.4 mmag. Top left, G magnitude; top right, G_{BP}; bottom left, G_{RP}; bottom right, G_{RVS}. White area represents zero stars.

3.2.3 Variables

Gaia will be continuously imaging the sky over its full five-year mission, and each individual object will be observed 70 times on average. The scanning law means that the time between repeated observations varies, and Gaia will be incredibly useful for detecting many types of variable stars. GOG produces a total of 10.8 million single variable objects. This number comes from the UM (Robin et al., 2012) and assumes 100% variability detection. The exact

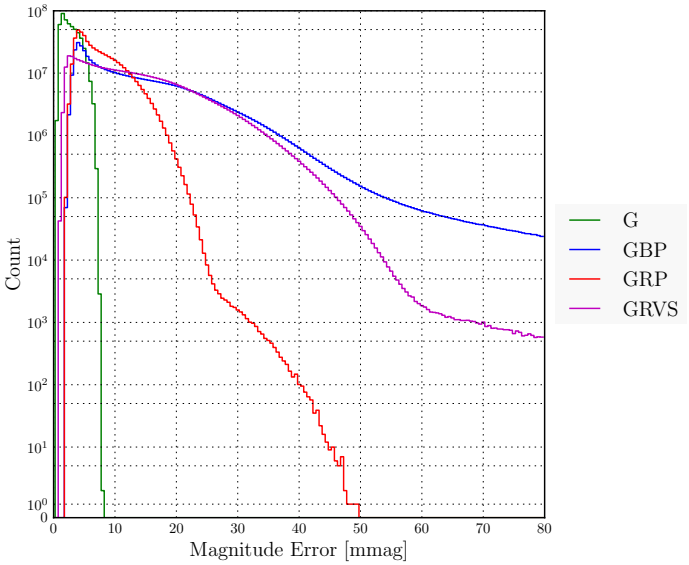


Figure 3.20: Histogram of error in G, G_{RVS}, G_{RP}, and G_{BP} for all single stars.

detection rates and the classification accuracy for each variability type are still unknown. In fact, the numbers of variable objects in the catalogue is expected to be higher than 10.8 million because some variable star types have not yet been implemented (see Robin et al. 2012 for a more detailed description).

The distribution of relative parallax error is given for each type of variable star in Fig. 3.22. The numbers of each type of variable produced by GOG are given in Table 3.6, along with the number of each type that falls below each relative parallax error limit.

In general, the numbers of variables presented in this work are lower than in Robin et al. (2012) by a factor of two or three. This is expected, because in the present work we are excluding all variables that are part of binary or multiple systems, and presenting the number of single variable stars alone.

However, the number of emission variables is higher in the present work. This is due to implementation of new types of emission stars: Oe, Ae, dMe,

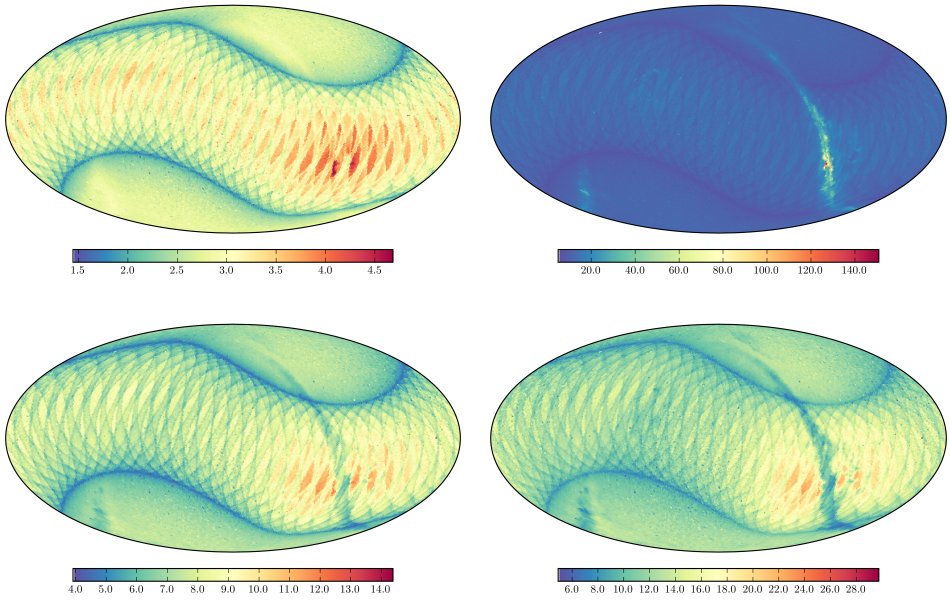


Figure 3.21: HealPixMap in equatorial coordinates of the mean error in: Top left: G*;* top right: G_{BP}*;* lower left: G_{RP}*;* lower right: G_{RVS}*. The colour scale gives the mean photometric error in mmag. The colour scales are different due to differences in the maximum mean magnitude.*

and WR stars. These are now included as emission variables but were not simulated in Robin et al. (2012). Additionally, the number of Mira variable stars is higher in the present work. This is from an implementation error in the version of the UM used in Robin et al. (2012), which has been fixed in the version used in the present work.

Cepheids and RR-Lyrae

Cepheids and RR-Lyrae are types of pulsating variable stars. Their regular pulsation and a tight period-luminosity relation make them excellent standard candles, and therefore of particular interest in studies of galactic structure and the distance scale. Figure 3.23 shows the histogram of error in parallax specifically for Cepheid and RR-Lyrae variable stars, while Fig. 3.24 shows

the errors in proper motions for Cepheids and RR-Lyrae.

It is clear that the vast majority of observed Cepheids will have very good precision, around the lower limit for Gaia. This is because they are very luminous, and will therefore in most cases have their mesaurement precision dominated by calibration effects rather than by photon noise. This is good news for those using Cepheids as distance indicators. We introduce a method for calibration of the Cepheid period-luminosity relation in Ch. 5.

For RR-Lyrae stars, which are older populations stars with lower liminosity, the expected precision is generally worse, ranging from around 50 to 300 μas. Whilst they are fainter, and therefore generally of lower precision than Cepheids, they are much more numerous. Current trigonometric parallax of RR-Lyrae stars is limited to only a handful of stars (e.g. those from Benedict et al., 2002), and the availability of Gaia data for a significant number of these stars will therefore be revolutionary.

Whilst the current results are presented only for the Milky Way, it is also expected that Gaia will directly observe trigonometric parallax for stars within the Magellanic Clouds. See Sect. 5.5 for a discussion of the impact of Gaia in the observation and calibration of LMC Cepheids and RR-Lyraes.

3.2.4 Physical parameters

The addition of low-resolution spectral photometers on-board Gaia has made it capable of providing information on several object parameters including an estimate of line-of-sight extinction, effective temperature, metallicity, and surface gravity. Discussion of each individual physical parameter is given below.

Provided here are results for an approximation of the results of Liu et al. (2012), which reproduces CU8 results statistically but not individually for each star. Therefore for detailed analysis of specific object types, care should be taken. Again, due to the very long tails of the error distributions caused by large numbers of extremely faint stars, the mean values given below should be taken with caution.

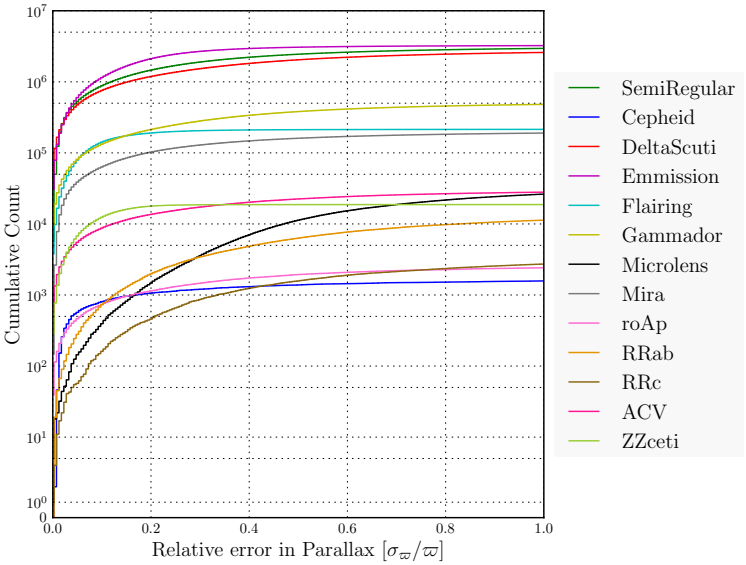


Figure 3.22: Cumulative histogram of the relative parallax error for all single stars, split by variability type. The histogram range displays 85% of all data.

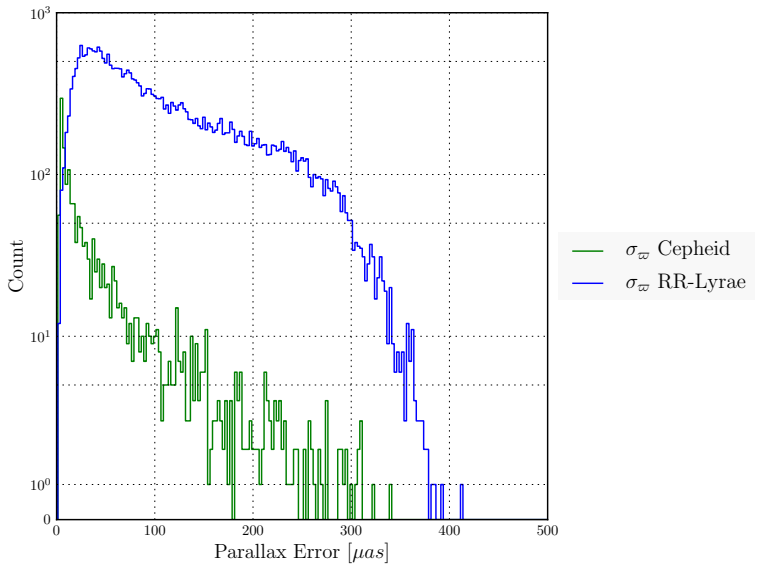


Figure 3.23: Histogram of parallax error for Cepheid and RR-Lyrae variable stars. RR-Lyrae is a combination of the two sub-populations RR-ab and RR-c.

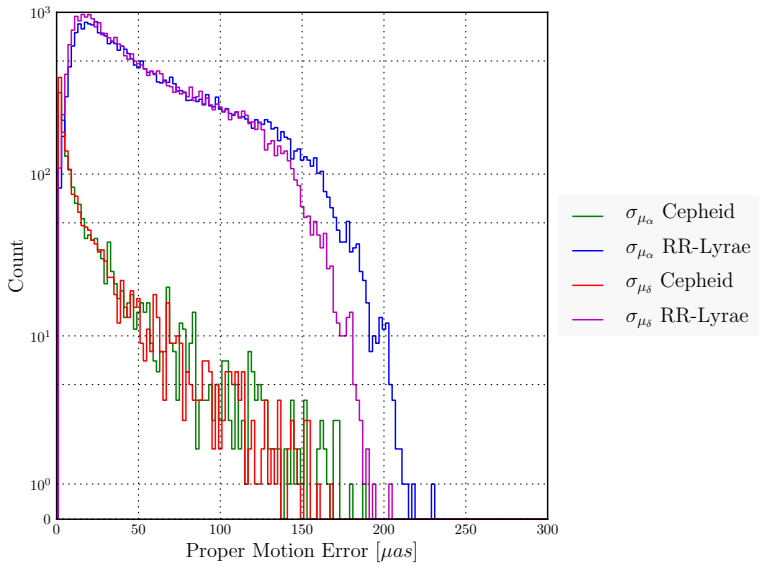


Figure 3.24: Histogram of proper motion error μ_α and μ_δ for Cepheid and RR-Lyrae variable stars. RR-Lyrae is a combination of the two sub-populations RR-ab and RR-c.

Variability type	Total	$\sigma_\varpi/\varpi < 5$	$\sigma_\varpi/\varpi < 1$	$\sigma_\varpi/\varpi < 0.5$	$\sigma_\varpi/\varpi < 0.2$	$\sigma_\varpi/\varpi < 0.05$	$\sigma_\varpi/\varpi < 0.01$
Non-variables	5.1×10^8	88	74	55	27	7.7	1.4
Emission	3.3×10^6	99	97	92	62	16	2.4
Flaring	2.1×10^5	99	99	98	88	33	4.6
δ Scuti	3.3×10^6	90	78	61	35	13	3.5
Semiregular	3.6×10^6	92	82	68	40	14	1.4
γ Dor	6.0×10^5	91	80	63	35	13	3.2
RR Lyrae AB-type	2.4×10^4	67	45	25	7.9	1.0	0.1
Mira	2.3×10^5	92	83	70	44	16	1.2
ZZ Ceti	1.9×10^4	100	100	99	94	33	4.1
ACV	3.5×10^4	91	80	64	39	15	3.8
RR Lyrae C-type	5.6×10^3	68	45	25	7.8	0.9	0.1
ρ Ap	3.0×10^3	92	82	65	38	14	3.8
Cepheid	1.8×10^3	95	88	78	59	30	0.1

Table 3.6: Total number of single stars of each variability type, and the percentage of each that falls below each relative parallax error limit: 500%, 100%, 50%, 20%, 5%, and 1%.

Extinction

Across many fields of astronomy, the effects of extinction on the apparent magnitude and colour of stars can play a major role in contributing to uncertainty. An accurate estimation of extinction will prove highly useful for many applications of the Gaia Catalogue.

Figure 3.25 shows the comparison between true extinction and the simulated Gaia estimate. For the vast majority of stars, the Gaia estimated extinction lies very close to the true value. This could prove very useful when, for example, using parallax and apparent magnitude data from Gaia because accurate extinction estimates are required to constrain the absolute magnitude of an object.

Additionally, these results show that the Gaia data will be highly useful in terms of mapping galactic extinction in three dimensions, thanks to the combination of a large number of accurate parallax and extinction measurements. The negative extinction values in Fig. 3.25 are of course non-physical and are simply the result of applying a Gaussian random error to stars with near zero extinction.

The discontinuity at $A_0 = 1$ in the top left-hand panel of Fig. 3.25 comes from the distinction made between high and low extinction stars in the presentation of the results in Liu et al. (2012). Our algorithm is based on results given in that paper, where the dependence on the extinction has been simplified to two cases, stars with $A_0 < 1$ and those with $A_0 > 1$. This distinction was made only for presentation of the results, and the real results from the DPAC algorithms will not show this discontinuity. Liu et al. (2012) report a degeneracy between extinction and effective temperature due to the lack of resolved spectral lines sensitive only to effective temperature.

Effective temperature

For all objects in the GOG catalogue, the measured effective temperature ranges between 850 and 102000 K. The error in effective temperature is less

than 640 K for all stars, with a mean value of 388 K. Figure 3.25 shows the comparison between true object effective temperature and the Gaia estimation. The thin lines visible in Fig. 3.25 are an artefact from the UM, which uses a Hess diagram to produce stars, leading to some quantisation in the effective temperature of simulated stars.

Metallicity

Metallicity can be estimated by Gaia in the form of $[Fe/H]$. Measured values range from -6.5 to +4.6. The mean error in metallicity estimate is 0.57 dex. The relatively high error in metallicity estimate can lead a large difference between real and observed values, as seen in Fig. 3.25.

Surface gravity

The mean error in surface gravity is 0.45 dex. The comparison between real and observed surface gravity can be seen in Fig. 3.25. As with metallicity, the lines at regular intervals at high gravity in this plot are due to the UM (Robin et al., 2012).

3.3 Conclusions

The Gaia Object Generator provides the most complete picture to date of what can be expected from the Gaia astrometric mission. Its simulated catalogue provides useful insight into how various types of objects will be observed and how each of their observables will appear after including observational errors and instrument effects. The simulated catalogue includes directly observed quantities, such as sky position and parallax, as well as derived quantities, such as interstellar extinction and metallicity.

Additionally, the full sky simulation described here is useful for gaining an idea of the size and format of the eventual Gaia Catalogue, for preparing

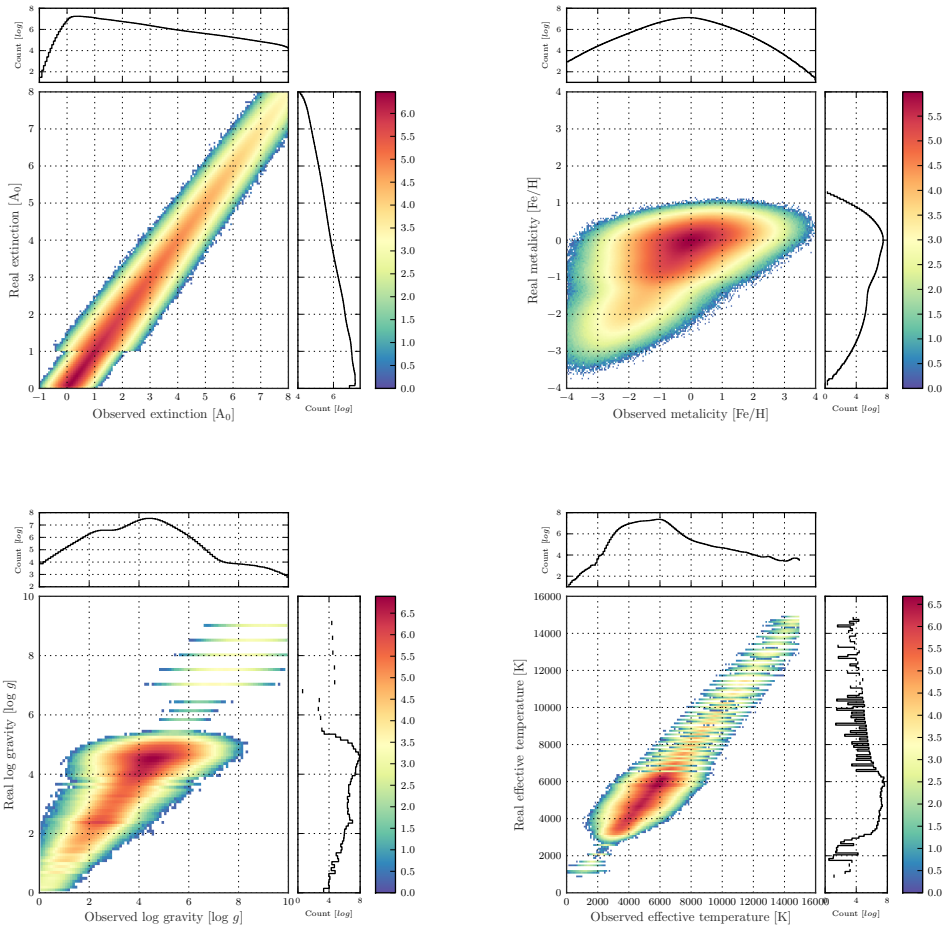


Figure 3.25: Comparison of the true values of physical parameters with the GOG 'observed' values for: Top left*, extinction;* top right*, metallicity;* bottom left*, surface gravity; and* bottom right*, effective temperature. The colour scales represent log density of objects in a bin size of:* top left*, 50 by 50 mmag;* top right*, 0.4 by 0.4 dex;* bottom left*, 0.5 by 0.5 dex; and* bottom right*, 100 by 100 K. White area represents zero stars.*

tools and hardware for hosting and distribution of the data, and for becoming familiar with working with such a large and rich dataset.

In addition to the stellar simulation described in this work, there are plans to generate other simulated catalogues of interest, so that a more complete version of the simulated Gaia Catalogue can be compiled. Already available

in GOG are other objects such as more than 500 open clusters, the Magellanic Clouds, supernovae, and other types of extragalactic objects.

Here we have focussed on the simulated catalogue from the inbuilt Gaia Universe Model, based on the Besançon Galaxy model. However, GOG can alternatively be supplied with an input catalogue generated by the user. This way, simulated data from any other model can be processed with GOG to obtain simulated Gaia observations of specific interest to the individual user. The input can be either synthetic data on a specific star or catalogue, or an entire simulated survey such as those generated using Galaxia (Sharma et al., 2011), provided a minimum of input information is supplied (e.g. position, distance, apparent magnitude, and colour).

With GOG, the capabilities of the instrument can be explored, and it is possible to gain insight into the expected performance for specific types of objects. While only a subset of the available statistics have been reproduced here, it is possible to obtain the full set of available statistics at request.

We are working to make the full simulated catalogue publicly available, so that interested individuals can begin working with data similar to the forthcoming Gaia Catalogue. The full mock catalogue has been ingested into the European Space Astronomy Centre (ESAC) test archive.

4

Basic luminosity calibration

Calculating the absolute magnitude of a star from a its distance and apparent magnitude is possible through the well known Pogson law:

$$M = m - 5\log(d) + 5 - A \tag{4.1}$$

where M is the absolute magnitude, m is the apparent magnitude, d the distance, and A the interstellar extinction. However, as highlighted in the introductory Ch. 2, the result of the direct application of this relation by substituting the parallax $\varpi = 1/d$ can lead to a biased result. The question of how best to obtain the absolute magnitude from parallax data has been tackled various times in the literature. Methods of Maximum Likelihood Estimation (MLE) were introduced for the use of statistical parallaxes by Rigal (1958). Statistical parallaxes are those inferred from proper motion and radial velocity

information for groups of stars on the sky. Although the initial method by Rigal made several very strong assumptions (e.g. that all of the stars had the same absolute magnitude, and that there were no observational errors), it included an explicit definition of the spatial distribution of the population and allowed simultaneous estimation of both the mean absolute magnitude of the population and the mean motion of the sun.

Jung (1970) improved upon the work by Rigal to overcome several of its limitations, for example by including a distribution of absolute magnitudes and a way of estimating the effects of observational errors. The algorithm of Jung was further improved by Heck (1975) by including more parameters and through the use of numerical methods to avoid of some of the simplifications in the equations needed due to the very limited computing resources of the time. The further improvements were made by Heck (1978), which proved a useful tool for luminosity calibration with statistical parallaxes from proper motion and radial velocity information, and led to several applications for determination of the absolute magnitude of specific star types, e.g. Jaschek et al. (1980) for Hg-Mg stars and Grenier et al. (1985) for main sequence stars.

However, the method of MLE at that time still had several major limitations: it used approximations of complex likelihood functions to avoid heavy computation requirements; the lack of a selection function lead to a Malmquist bias on the resulting apparent magnitudes; there was no inclusion of interstellar extinction; and only simple assumed distributions were used, such as a uniform spatial distribution. Additionally, all prior methods had been focusing on the use of statistical parallax, which was more readily obtainable at the time due to the availability of proper motion measurements and a lack of trigonometric parallax measurements.

The MLE method by Ratnatunga & Casertano (1991) included for the first time terms to take account for selection bias due to incomplete samples. The next major evolution of the method was not until Luri (1995), when many the aforementioned limitations of the MLE method were overcome. Additionally, the timing of these improvements coinciding with the Hipparcos mission gave

access to a significant sample of absolute trigonometric parallax measurements, making the method much more widely applicable.

Here, we continue the development following the prescription of Luri (1995). We extend the method from simple calibration of absolute magnitudes to simultaneous estimation of multiple parameters, and derive applications to several key aspects of luminosity calibration and the distance scale. This comes at a time when the Gaia mission will be making the next large advance in astrometric measurements, allowing the application of the derived methods to many more wide ranging applications and to stars at distances up to the Magellanic Clouds.

In the following four chapters we develop several methods for luminosity calibration. The methods make extensive use of the statistical techniques introduced in Ch. 2, particularly those of Maximum Likelihood Estimation (MLE), and Maximum A Posteriori estimation (MAP). As a test case, the basics of the mathematical formulation of MLE are introduced in this chapter. Here a simple case of a galactic population of stars is used, which serves well to introduce the topic and mathematics required.

In Ch. 5, the same MLE method is applied to the more realistic case of fitting a period-luminosity relation for Cepheid variable stars. In Sect. 5.2, we define a MAP model fitting approach which is useful for fitting the period-luminosity or period-luminosity-metallicity relations in the absence of parallax information. This method has wider applications in that it is generally applicable to fitting a straight line to any x, y data, or a plane to any x, y, z data, and is available as a stand-alone algorithm. The method has been applied to real data for RR-Lyrae stars.

In Sect. 5.4, a method for fitting the distance to the Large Magellanic Cloud is presented. This work tests the capability of Gaia in directly measuring its distance, utilising absolute trigonometric parallax alone. As no parallax measurements for stars in the LMC or SMC currently exist, the method has been tested using GOG simulated data.

In Ch. 6, an approach to fitting the distance, kinematics, and isochrone for

an open cluster is presented. The mathematical derivation of the method is shown, along with the results of extensive testing using both real and simulated data.

4.2 Basic definition

In the most basic sense, luminosity calibration of a population of stars will derive the absolute magnitude of the population. Usually, the absolute magnitude will follow some distribution, which depends on the population at hand. Stars of a specific spectral type and luminosity class will exhibit a tight range of absolute magnitudes, whereas when taking a general population of Main Sequence stars one would expect the absolute magnitude to be distributed around the main sequence, which leads to a dependence in colour (or temperature). In this initial example, it is assumed that stars exhibit a tight range of absolute magnitudes, and are distributed in space uniformly. This only holds true for stars of known spectral type and luminosity class, in the local neighbourhood.

Assume that a catalogue contains a list of stars with values of apparent magnitude (m), position (l, b) and parallax (ϖ).

Here a model must be chosen which describes the system at hand. In this case, the distribution in absolute magnitude can be defined as being given by a Normal distribution:

$$\varphi_M = e^{-0.5\left(\frac{M-M_{\text{mean}}}{\sigma_M}\right)^2} \tag{4.2}$$

where M is the observed absolute magnitude, M_{mean} is the mean magnitude of the population, and σ_M is the intrinsic dispersion of the distribution. The absolute magnitude is obtained from the observations through the common relation (neglecting extinction for now):

$$M = m + 5\log(\varpi) + 5 \tag{4.3}$$

A uniform spatial distribution is given by:

$$\varphi_{\varpi} = r^2 \cos b \tag{4.4}$$

Where the distance is obtained from the inverse of the parallax $r = 1/\varpi$. As will be the case for almost all observations (including Gaia), we assume a magnitude limited sample. In order to correctly account for this, a selection function ($\mathcal{S}$) is introduced in the form of a Heaviside function at the limiting apparent magnitude m_{lim}, which is magnitude 20 in the case of Gaia:

$$\mathcal{S} = \Theta(m_g - m_{lim}) \tag{4.5}$$

All of the above distributions are unnormalised. It is convenient to work with the unnormalised functions and then normalise the full joint PDF with a global normalisation factor later. In order to normalise the PDF of the distance distribution, a limit must be introduced to φ_{ϖ}. This upper limit in distance is arbitrary, provided that it is chosen to be high enough that it does not effect the spatial or magnitude distribution of the simulated sample, because very distant stars will be omitted from the sample due to the apparent magnitude limit.

4.3 Formulation of the ML equation

Starting with the most basic form of the definition of ML estimation, the *Likelihood function*, $\mathcal{L}$ is defined such that:

$$\mathcal{L}(\boldsymbol{\theta}) = \prod_{i=1}^{N} \mathcal{P}(\boldsymbol{x}_i|\boldsymbol{\theta}) \tag{4.6}$$

where $\mathcal{P}$ is the combined probability density function (PDF), $\boldsymbol{x}$ is vector containing a set of observables, and $\boldsymbol{\theta}$ is a vector describing the set of parameters which exist in the model. The set of observables refers to a list of 'observa-

tions', but here we assume that the real values are known and there are no observational errors. This likelihood function is the joint probability of having found the data, given the parameters of the model. By finding the values for $\boldsymbol{\theta}$ which maximise the value of the likelihood function, $\mathcal{L}$, we find the ML estimation of the true values of $\boldsymbol{\theta}$ for the given dataset. This is mathematically equivalent to taking the sum of the log of $\mathcal{P}$:

$$\ln\mathcal{L}(\boldsymbol{\theta}) = \sum_{i=1}^{N} \ln\mathcal{P}(\boldsymbol{x}_i|\boldsymbol{\theta}) \tag{4.7}$$

In our case, the set of N observations $\boldsymbol{x} = (\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_N)$ are the stars for which the apparent magnitude, position and parallax are (perfectly) known: $\boldsymbol{x}_i = (m, \varpi, l, b)$. The model parameters are simply the mean and variance of the distribution of absolute magnitudes $\boldsymbol{\theta} = (M_{\text{mean}}, \sigma_M)$.

The combined PDF $\mathcal{P}(\boldsymbol{x}_i|\boldsymbol{\theta})$ is a combination of the PDFs for magnitude and position, the selection function, and a normalisation constant $\mathcal{C}^{-1}$, such that:

$$\mathcal{P}(\boldsymbol{x}|\boldsymbol{\theta}) = \mathcal{C}^{-1}\varphi_M(m)\varphi_r(\varpi, l, b)\mathcal{S}(m) \tag{4.8}$$

4.3.1 Derivation of the normalisation constant

Normalisation of a PDF is achieved through integration the over the full range of all parameters:

$$\mathcal{C} = \int_{\forall \boldsymbol{x}} \varphi_M(m)\varphi_\varpi(\varpi, l, b)\mathcal{S}(m)\, d\boldsymbol{x} \tag{4.9}$$

The selection function acts to provide a finite limit to the integral over M_g:

$$\mathcal{C} = \int_r \int_{-\infty}^{M_{lim}} \varphi_M(m) dM \int_l \int_b \varphi_\varpi(\varpi, l, b)\, db\, dl\, dr \tag{4.10}$$

Here we use the definition of the error function (erf):

$$\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2}\, dt \tag{4.11}$$

and the complementary error function (cerf):

$$\mathrm{erfc}(x) = 1 - \mathrm{erf}(x) = \frac{2}{\sqrt{\pi}} \int_x^\infty e^{-t^2}\, dt \tag{4.12}$$

for the integral over magnitude, giving:

$$\mathcal{C} = \int_0^{r_{lim}} \int_{-\infty}^{M_{lim}} e^{-0.5\left(\frac{M-M_{\mathrm{mean}}}{\sigma_{M_g}}\right)^2} dM \int_{-\pi/2}^{\pi/2} \int_0^{2\pi} r^2 \mathrm{cos}b\, db\, dl\, dr \tag{4.13}$$

$$\mathcal{C} = \int_0^{r_{lim}} \sqrt{\frac{\pi}{2}}\sigma_M \mathrm{erfc}\left(-\frac{m_{lim} + 5\log(\varpi) + 5 - M_{\mathrm{mean}}}{\sqrt{2}\sigma_M}\right) 4\pi r^2\, dr \tag{4.14}$$

Define χ so that:

$$\chi = \frac{m_{lim} + 5log(\varpi) + 5 - M_{\mathrm{mean}}}{\sqrt{2}\sigma_M} \tag{4.15}$$

Finally, the normalisation coefficient is:

$$\mathcal{C} = 2\sqrt{2}\pi^{3/2}\sigma_M \int_0^{r_{lim}} \mathrm{erfc}\,(\chi)\, r^2\, dr \tag{4.16}$$

Therefore substituting this into the original Likelihood function in Eq. 4.7, one obtains:

$$\ln\mathcal{L}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \ln\left(e^{-0.5\left(\frac{M_i-M_{\mathrm{mean}}}{\sigma_{M_g}}\right)^2} r_i^2 \mathrm{cos}b_i \mathcal{C}^{-1} \mathcal{S}(\boldsymbol{x})\right) \tag{4.17}$$

The selection function $\mathcal{S}$ is by definition 1 for all $\boldsymbol{x}_i$, because for the stars exist

within the catalogue they must be brighter than the limiting magnitude.

$$\ln\mathcal{L}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \ln\left(e^{-0.5\left(\frac{M_i - M_{\rm mean}}{\sigma_{Mg}}\right)^2} r_i^2 \cos b_i \mathcal{C}^{-1} \right) \tag{4.18}$$

This is the function which is maximised to find the ML estimation of the parameters $M_{\rm mean}$ and σ_M, the mean and variance of the magnitude of the population.

4.4 Exponential disk population

Clearly a uniform spatial distribution an over simplification and is only a reasonable assumption for nearby stars. The next step is to introduce a more realistic pdf for the spatial distribution. In this case an exponential disk is chosen:

$$\varphi_\varpi = e^{\frac{-|r\sin b|}{Z_0}} r^2 \cos b \tag{4.19}$$

where b is the galactic latitude, and Z is the scale height of the Galaxy. This function is implemented in place of the previous spatial distribution for the uniform case. With the same reasoning as in the previous case, upper limits are introduced to the distance, which provides a limit to the distribution but is large enough not to effect the distribution of the parameters of the sample (after the truncation in absolute magnitude).

Normalisation constant

Following the same method for the derivation of the normilisation constant as in Sect. 4.3.1:

$$\mathcal{C} = \int_{\forall x} \varphi_M(m)\varphi_\varpi(r,l,b)\mathcal{S}(x) \tag{4.20}$$

$$\mathcal{C} = \int_0^{r_{lim}} \int_{-\pi/2}^{\pi/2} \int_0^{2\pi} \int_{-\infty}^{M_{lim}} e^{-0.5\left(\frac{M-M_{\rm mean}}{\sigma_M}\right)^2} e^{\frac{-|r{\rm sin}b|}{Z_0}} r^2 {\rm cos}b \, dM \, dl \, db \, dr \quad (4.21)$$

$$\mathcal{C} = \int_0^{r_{lim}} \int_{-\pi/2}^{\pi/2} \int_0^{2\pi} \sqrt{\frac{\pi}{2}} \sigma_{M_g} {\rm erfc}\left(-\chi\right) e^{\frac{-|r{\rm sin}b|}{Z_0}} r^2 {\rm cos}b \, dl \, db \, dr \quad (4.22)$$

$$\mathcal{C} = 2\sqrt{2}\pi^{3/2} \sigma_{M_g} Z_0 \int_0^{r_{lim}} {\rm erfc}\left(-\chi\right) \left(1 - e^{\frac{-r}{Z_0}}\right) r \, dr \quad (4.23)$$

The error function with χ remains inside the integral because M is dependent on r. This integral will be solved numerically.

Putting everything together we have:

$$\ln\mathcal{L}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \ln\left(e^{-0.5\left(\frac{M-M_{\rm mean}}{\sigma_M}\right)^2} e^{\frac{-|r{\rm sin}b|}{Z_0}} r^2 {\rm cos}b \, \mathcal{C}^{-1}\right) \quad (4.24)$$

This highlights the fact that the user is free to choose the distribution which best suits the problem at hand. The distribution must accurately reflect the physical nature of the system being modelled, and the model can increase significantly in complexity in an attempt to model very detailed aspects of a system. In this example we have swiftly moved from a uniform distribution for the solar neighbourhood to a exponential disk, but it would be possible to extend the model to include the different galactic populations such as the thin and thick disk, the halo, and the bulge. Further detail such as spiral arms can be included if needed.

4.5 Inclusion of observational errors

Up to this point the Maximum Likelihood equations have been formulated for a set of observables, $\boldsymbol{x} = (\varpi, l, b, m)$. However, in reality these values are not known. Instead we have measurements of these parameters, which are

susceptible to various observational errors. To define $\boldsymbol{x}_0 = (\varpi_0, l_0, b_0, m_0)$ as the 'real', exact values of the parameters, we then have $\boldsymbol{x} = (\varpi, l, b, m)$ as the measured value of each quantity (including measurement error).

In many cases, errors in l, b and m_g can be treated as negligible, as their measurement error is generally small in comparison to that of the parallax, and will not have a significant effect on the results. Therefore the distribution of these parameters can effectively be described by a Dirac delta function.

The measured value of the parallax ϖ can be well described by a Gaussian distribution around the true value ϖ_0, with some dispersion ϵ_ϖ given by the formal error on the measurement. Therefore the distribution in errors can be written as:

$$\mathcal{E}(\boldsymbol{x}|\boldsymbol{x}_0) = e^{-0.5\left(\frac{\varpi - 1/r_0}{\epsilon_\varpi}\right)^2} \delta(l_0, b_0, m_0) \tag{4.25}$$

remembering that ϖ_0 and $1/r_0$ are exactly equivalent. To formulate the Likelihood function:

$$\mathcal{L}(\boldsymbol{\theta}) = \prod_{i=1}^{n} \mathcal{P}(\boldsymbol{x}_i|\boldsymbol{\theta}) \tag{4.26}$$

where:

$$\mathcal{P}(\boldsymbol{x}_i|\boldsymbol{\theta}) = \mathcal{C}^{-1}\mathcal{S}(\boldsymbol{x}) \int_{\forall \boldsymbol{x}_0} \varphi_{m_0}(m_0)\varphi_\varpi(\varpi_0, l_0, b_0)\mathcal{E}(\boldsymbol{x}|\boldsymbol{x}_0)\, d\boldsymbol{x}_0 \tag{4.27}$$

This step provides a link between a set of observations and the true values of the observables, which are not known. The observed parallax (and potentially any other error effected quantity) is used directly, with no non-linear transformations. Non-linear transformations can be applied on the true values without issue, as in the case of calculating M_0. Of course, true values of the observables are not known, so they must be marginalised through integration.

Substituting in the various models gives:

$$\mathcal{P}(\boldsymbol{x}_i|\boldsymbol{\theta}) = \mathcal{C}^{-1}\mathcal{S}(\boldsymbol{x}) \int_{\forall \boldsymbol{x}_0} e^{-0.5\left(\frac{M_0 - M_{mean}}{\sigma_{Mg}}\right)^2} e^{\frac{-|\sin b_0|}{Z_0 \varpi_0}} \frac{1}{\varpi_0^2} \cos b_0 \, e^{-0.5\left(\frac{\varpi - \varpi_0}{\epsilon_\varpi}\right)^2} \delta(l_0, b_0, m_0) \, d\boldsymbol{x}_0 \tag{4.28}$$

As l_0, b_0 and m_0 are all described by the delta function their integral is known to be exactly 1 by definition, leaving:

$$\mathcal{P}(\boldsymbol{x}_i|\boldsymbol{\theta}) = \mathcal{C}^{-1}\mathcal{S}(\boldsymbol{x}) \int_{0}^{\infty} e^{-0.5\left(\frac{M_0 - M_{mean}}{\sigma_{M}}\right)^2} e^{\frac{-|\sin b|}{Z_0 \varpi_0}} \frac{1}{\varpi_0^2} \cos b \, e^{-0.5\left(\frac{\varpi - \varpi_0}{\epsilon_\varpi}\right)^2} \, d\varpi_0 \tag{4.29}$$

This integration will be performed numerically.

4.5.1 Normalisation coefficient

The normalisation coefficient is calculated using the same method as that in Sect. 4.3:

$$\mathcal{C} = \int_{\forall \boldsymbol{x}_0} \int_{\forall \boldsymbol{x}} \varphi_{m_0}(m_0) \varphi_{\varpi_0}(\varpi_0, l_0, b_0) \mathcal{S}(\boldsymbol{x}) \mathcal{E}(\boldsymbol{x}|\boldsymbol{x}_0) \, d\boldsymbol{x} \, d\boldsymbol{x}_0 \tag{4.30}$$

Here the integration is simplified by defining the temporary variable I such that:

$$\mathcal{C} = \int_{\forall \boldsymbol{x}_0} \varphi_{m_0}(m_0) \varphi_{\varpi_0}(\pi_0, l_0, b_0) \underbrace{\int_{\forall \boldsymbol{x}} \mathcal{S}(\boldsymbol{x}) \mathcal{E}(\boldsymbol{x}|\boldsymbol{x}_0) \, d\boldsymbol{x}}_{\mathrm{I}} \, d\boldsymbol{x}_0 \tag{4.31}$$

$$I = \int_{\forall \boldsymbol{x}} \Theta(m - m_{lim}) e^{-0.5\left(\frac{\varpi - \varpi_0}{\epsilon_\varpi}\right)^2} \delta(l_0, b_0, m_{g0}) \, d\boldsymbol{x} \tag{4.32}$$

The delta function of (l, b, m) integrates to 1 by definition. Additionally, because the parallax ϖ is a measured quantity which is ϖ_0 plus or minus some Gaussian error, the possible values of ϖ are theoretically positive as well as

negative. Therefore the limits of the integration are $-\infty$ to $+\infty$:

$$I = \int_{-\infty}^{\infty} \Theta(m - m_{lim}) e^{-0.5\left(\frac{\varpi - \varpi_0}{\epsilon_\varpi}\right)^2} d\varpi \tag{4.33}$$

For any fixed absolute magnitude, the apparent magnitude limit is equivalent to a limit in the observable distance, or parallax, such that $\Theta(m - m_{lim}) = \Theta(\varpi_{lim} - \varpi)$ with: $\varpi_{lim} = 10^{-\frac{m_{lim} - M + 5}{5}}$ Therefore:

$$I = \int_{-\infty}^{\infty} \Theta(\varpi_{lim} - \varpi) e^{-0.5\left(\frac{\varpi - \varpi_0}{\epsilon_\varpi}\right)^2} d\varpi \tag{4.34}$$

$$I = \int_{\varpi_{lim}}^{\infty} e^{-0.5\left(\frac{\varpi - \varpi_0}{\epsilon_\varpi}\right)^2} d\varpi \tag{4.35}$$

$$I = \sqrt{\frac{\pi}{2}} \epsilon_\varpi \text{erfc}\left(\chi_1\right) \tag{4.36}$$

where:

$$\chi_1 = \frac{10^{-\frac{m_{lim} - M + 5}{5}} - \varpi_0}{\sqrt{2}\epsilon_\varpi} \tag{4.37}$$

Substituting this back into $\mathcal{C}$:

$$\mathcal{C} = \int_{\forall \boldsymbol{x}_0} \varphi_{m_0}(m_0) \varphi_{\varpi_0}(\varpi_0, l_0, b_0) \sqrt{\frac{\pi}{2}} \epsilon_\varpi \text{erfc}\left(\chi_1\right)\, d\boldsymbol{x}_0 \tag{4.38}$$

$$\mathcal{C} = \sqrt{\frac{\pi}{2}} \epsilon_\varpi \int_{1/r_{lim}}^{\infty} \int_{-\pi/2}^{\pi/2} \int_{0}^{2\pi} \int_{-\infty}^{\infty} e^{-0.5\left(\frac{M_0 - M_{mean}}{\sigma_M}\right)^2} e^{\frac{-|\sin b_0|}{\varpi_0 Z}} \frac{1}{\varpi_0^2} \cos b_0 \text{erfc}\left(\chi_1\right)\, dm_0\, dl_0\, db_0\, d\varpi_0 \tag{4.39}$$

$$\mathcal{C} = 2\pi \sqrt{\frac{\pi}{2}} \epsilon_\varpi \int_{1/r_{lim}}^{\infty} \frac{1}{\varpi_0^2} \text{erfc}\left(\chi_1\right) \int_{-\pi/2}^{\pi/2} \cos b_0 e^{\frac{-|\sin b_0|}{\varpi_0 Z_0}}\, db_0 \int_{-\infty}^{\infty} e^{-0.5\left(\frac{M_0 - M_{mean}}{\sigma_M}\right)^2} dm_0\, d\varpi_0 \tag{4.40}$$

$$\mathcal{C} = 2\pi\sqrt{\frac{\pi}{2}}\epsilon_{\varpi}\int_{1/r_{lim}}^{\infty}\frac{1}{\varpi_0^2}\mathrm{erfc}\left(\chi_1\right)2Z\left(1-e^{-\frac{1}{\varpi_0 Z_0}}\right)\frac{1}{\varpi_0}\sqrt{2\pi}\sigma_M\, d\varpi_0 \qquad (4.41)$$

$$\mathcal{C} = 4\pi^2 Z\epsilon_{\varpi}\sigma_M\int_{1/r_{lim}}^{\infty}\frac{1}{\varpi_0^3}\mathrm{erfc}\left(\chi_1\right)2Z\left(1-e^{-\frac{1}{\varpi_0 Z_0}}\right)\, d\varpi_0 \qquad (4.42)$$

With the normalisation coefficient derived, every component of the Likelihood function (Eq. 4.26) has been defined. This formulation is the full prescription necessary for Maximum Likelihood Estimation of the absolute magnitude of a population of stars. When combined with a set of observational data and the numerical methods required for the calculation and optimisation of the Likelihood function, the Maximum Likelihood estimate of the model parameters can be obtained. The explicit inclusion of a spatial distribution takes account of Lutz-Kelker bias, and the inclusion of a selection function avoids the Malmquist bias. By defining the models in terms of 'true' values of the parameters, and linking these unknown 'true' values with the observations via the error function, we simultaneously take into account the effect of observational errors and avoid any non-linear transformations on error effected values (e.g. on the parallax).

This introductory example has made several simplifying assumptions. The spatial distribution of the population of stars is assumed to be an infinite disk (Eq. 4.19). It is possible to increase the complexity of this distribution, to include the various galactic populations (bulge, halo, thin disk, thick disk), or the spiral arms. The effects of galactic rotation and interstellar extinction can also be taken into account. This has not been calculated here as it has been described in detail in Luri (1995).

4.6 Conclusions

Here, the basic mathematical formulation for luminosity calibration has been introduced. This is the core concept which will be utilised in the following chapters. The rational for the method which has been introduced is that:

- Estimation of derived quantities such as distance and absolute magnitude is best achieved through statistical *inference*.
- The properties of the system being analysed should be parametrised, and modelled by explicitly including the expected distributions of all parameters.
- The observed data are linked to their true values of via the error function, which takes into account the precision of the data being used.
- Explicitly including a selection function takes into account the fact that the data is taken from a magnitude limited sample.
- Integration of the functions should be performed as far as possible, and numerical methods employed where no analytical solution can be found. This avoids simplifying assumptions in the mathematics.

Using the mathematical description given here, the estimation of the parameters is achieved by using numerical optimisation routines. Classical optimisation algorithms such as 'Nelder-Mead' or 'Powell' efficiently allow estimation of the peak of the posterior PDF, providing the Maximum Likelihood estimate of the parameters. More advanced methods such as Markov Chain Monte Carlo provide more information by evaluating the entire PDF. Both methods have been employed in the following chapters.

5

Variables, and the Large Magellanic Cloud

One of the strengths of the MLE method is that it can be applied in a wide range of situations, simply by adapting the models for the situation at hand. While the method from Ch. 4 could be useful for calibrating the absolute magnitude of different star populations in the Gaia catalogue, there are many other applications which can prove useful. In this chapter we look at two fundamental parts of the distance ladder: variable stars as standard candles, and the distance to the Large Magellanic Cloud (LMC).

In Sect. 5.1 we revisit the MLE method introduced in the last chapter and extend it to the calibration of a period-luminosity (PL) or period-luminosity-metallicity (PLZ) relation. The method uses position, parallax, period and apparent magnitude in order to fit the slope and zero point of a PL relation,

and will therefore be useful when applied to Gaia data for Milky Way variable stars such as Cepheids. As Gaia parallax data is not yet available, we derive a method which is applicable in the absence of parallax information in Sect. 5.2. In Sect. 5.3 we apply this method to real data for calibration of RR-Lyrae stars in the LMC.

The Magellanic Clouds are a key step in the distance ladder, and are the first point in the calibration of the extragalactic distance scale. In Sect. 5.4 we derive a method for the estimation of the distance to the LMC using trigonometric parallax data. The distance to the LMC is often used to calibrate the the zero points of absolute PL and PLZ relations, therefore we test the capabilities of Gaia in calibrating these relations for LMC variables in Sect. 5.5.

5.1 The Galactic PL relation

Extending the formulation of Ch. 4 to the calibration of variable stars has been performed for the example case of Cepheid variable stars. However the method is applicable to any type of variable star which follows a quantifiable relationship in its absolute magnitude. The assumed linear period-luminosity (PL) relation is easily interchangeable with a period-luminosity-metallicity (PLZ) relation, or a non-linear relation.

Due to the fact that Gaia is constantly rotating and observing the sky repeatedly over its five year mission, Gaia will provide on average 80 observations per star. This makes the Gaia transit data useful for characterising stellar variability, and the end-of-mission catalogue will contain flags for variability with information on period. As shown in Sect. 3.2.3 for the GOG catalogue, Gaia is expected to observe several million variable stars, including standard candles such as thousands of Cepheid variable stars and tens of thousands of RR-Lyrae. Additionally, the Gaia catalogue will contain an estimate of the interstellar extinction. Combining the variability period with the apparent magnitude, position, parallax, and extinction, it will be possible to calibrate

the absolute PL relation, including the zero point, with an unprecedented precision by using the Gaia data.

Figure 3.23 shows the parallax error for Cepheid and RR-Lyrae stars. For Cepheids, it is clear that the vast majority of stars will have very high precision due to their extreme brightness. RR-Lyraes are generally older and fainter, and for this reason the majority of galactic RR-Lyraes are expected to have precision ranging from around 50 to 300 μas. Although with lower precision, RR-Lyrae stars are expected to be around a factor of 10 more numerous than Cepehids, making both variable types very promising candidates for precise calibration with Gaia data.

5.1.1 Likelihood function

As in previous sections, the Likelihood function can be defined as:

$$\mathcal{L}(\boldsymbol{\theta}) = \prod_{i=1}^{n} \mathcal{P}(\boldsymbol{x}_i|\boldsymbol{\theta}) \tag{5.1}$$

where the joint density function is made up of the unnormalised density function and a normalisation constant such that:

$$\mathcal{P}(\boldsymbol{x}_i|\boldsymbol{\theta}) = \mathcal{C}^{-1}\mathcal{D}(\boldsymbol{x}|\boldsymbol{\theta}) \tag{5.2}$$

The vector ($\boldsymbol{x} = m, l, b, P, \varpi$) describes the observed properties of each object (where P is the period), and ($\boldsymbol{x}_0 = m_0, l_0, b_0, P_0, r_0$) describes the 'true' underlying object properties.

Defining a model for the distribution in absolute magnitude, φ_M, position, $\varphi_{r,l,b}$, and variability period, φ_P, the unnormalised density function becomes:

$$\mathcal{D}(\boldsymbol{x}|\boldsymbol{\theta}) = \mathcal{S}(\boldsymbol{x}) \int_{\forall \boldsymbol{x}_0} \varphi_{M_0}\varphi_{\varpi_0,l_0,b_0}\varphi_P \mathcal{E}(\boldsymbol{x}|\boldsymbol{x}_0)\, d\boldsymbol{x}_0 \tag{5.3}$$

We assume that the absolute magnitude of the star is Normally distributed around some mean magnitude. Assuming that the star follows a period lu-

minosity relation of the form $M = \rho \log P + \delta$, the expected distribution of absolute magnitude is given by:

$$\varphi_M = e^{-0.5\left(\frac{M - M_{\text{mean}}}{\sigma_{PL}}\right)^2} = e^{-0.5\left(\frac{M - (\rho \log P + \delta)}{\sigma_{PL}}\right)^2} \tag{5.4}$$

where σ_{PL} is the intrinsic dispersion of the PL relation. The parameters ρ and δ are the slope and zero point of the PL relation, which are to be found. Similarly to Sect. 4.4, the spatial distribution $\varphi_{r,l,b}$, is assumed to be an exponential disk:

$$\varphi_{r,l,b} = e^{\frac{-|r \sin b|}{Z_0}} r^2 \cos b \tag{5.5}$$

Additionally, we approximate the period distribution, φ_P, for the population of Cepheids as being Normally distributed around some mean period:

$$\varphi_P = e^{-0.5\left(\frac{P - P_{\text{mean}}}{\sigma_P}\right)^2} \tag{5.6}$$

The above choices are reasonable approximations but can easily be improved by simply substituting in a different distribution. Finally, the error distribution for observed quantities is given by:

$$\mathcal{E}(\boldsymbol{x}|\boldsymbol{x}_0) = e^{-0.5\left(\frac{\varpi - \varpi_0}{\epsilon_\varpi}\right)^2} \delta(l_0, b_0, m_{g0}, P_0) \tag{5.7}$$

As l_0, b_0, m_0 and P_0 are all described by the delta function their integral is known to be exactly 1 by definition, the only remaining integral is in r_0:

$$\mathcal{D}(\boldsymbol{x}|\boldsymbol{\theta}) = \int_0^\infty \varphi_{M_0} \varphi_{\varpi_0, l_0, b_0} \varphi_P \mathcal{E}(\boldsymbol{x}|\boldsymbol{x}_0)\, dr_0 \tag{5.8}$$

This integration will be performed numerically. The final free-fit parameters are therefore the slope and zero point of the PL relation, and the scale height of the Galaxy: $\boldsymbol{\theta} = (\alpha, \rho, Z)$.

5.1.2 Normalisation constant

The normalisation constant is found by integrating the joint probability density function over all ($\boldsymbol{x} = m, l, b, P, \varpi$):

$$\mathcal{C} = \int_{\forall \boldsymbol{x}_0} \int_{\forall \boldsymbol{x}} \varphi_{M_{g0}} \varphi_{\varpi_0, l_0, b_0} \varphi_{P_0} \mathcal{S}(\boldsymbol{x}) \mathcal{E}(\boldsymbol{x}|\boldsymbol{x}_0)\, d\boldsymbol{x}\, d\boldsymbol{x}_0 \tag{5.9}$$

Here the integration is simplified by defining the temporary variable I such that:

$$\mathcal{C} = \int_{\forall \boldsymbol{x}_0} \varphi_{M_{g0}} \varphi_{\varpi_0, l_0, b_0} \varphi_{P_0} \underbrace{\int_{\forall \boldsymbol{x}} \mathcal{S}(\boldsymbol{x}) \mathcal{E}(\boldsymbol{x}|\boldsymbol{x}_0)\, d\boldsymbol{x}}_{\mathrm{I}}\, d\boldsymbol{x}_0 \tag{5.10}$$

$$I = \int_{\forall \boldsymbol{x}} \Theta(m - m_{lim}) e^{-0.5\left(\frac{\varpi - \varpi_0}{\epsilon_\varpi}\right)^2} \delta(l_0, b_0, m_{g0}, P_0)\, d\boldsymbol{x} \tag{5.11}$$

As the delta function of (l, b, m, P) integrates to 1 by definition, this term can be removed. Because the observed parallax ϖ is a measured quantity which is the true parallax ϖ_0 plus or minus some Gaussian error, its possible values are theoretically positive as well as negative. Therefore the limits of the integration are $-\infty$ to $+\infty$:

$$I = \int_{-\infty}^{\infty} \Theta(m - m_{lim}) e^{-0.5\left(\frac{\varpi - \varpi_0}{\epsilon_\varpi}\right)^2} d\varpi \tag{5.12}$$

$$I = \Theta(m - m_{lim}) \sqrt{\frac{\pi}{2}} \epsilon_\varpi \tag{5.13}$$

Therefore substituting I back into Eqn. 5.10:

$$\mathcal{C} = \sqrt{\frac{\pi}{2}} \int_{\forall \boldsymbol{x}_0} \Theta(m - m_{lim}) \varphi_{M_{g0}} \varphi_{\varpi_0, l_0, b_0} \varphi_{P_0} \epsilon_\varpi\, d\boldsymbol{x}_0 \tag{5.14}$$

The remaining integral over all $\boldsymbol{x}$ can be performed for each variable in steps.

Integrating with respect to l_0 and b_0 gives:

$$\mathcal{C} = \sqrt{\frac{\pi}{2}} 4\pi Z \int_{\forall m_0, P_0, r_0} \Theta(m - m_{lim}) \varphi_{M_{g0}} \varphi_{P_0} \left(1 - e^{-\frac{r_0}{Z}}\right) r_0 \epsilon_{\varpi} \, dm_0 \, dP_0 \, dr_0 \tag{5.15}$$

Integrating with respect to m_0 gives:

$$\mathcal{C} = 2\sqrt{2}\pi^{3/2} \epsilon_{\varpi} Z \int_{\forall P_0, r_0} \varphi_{P_0} \left(1 - e^{-\frac{r_0}{Z}}\right) r_0 \int_{\forall m_0} \Theta(m - m_{lim}) \varphi_{M_{g0}} \, dm_0 \, dP_0 \, dr_0 \tag{5.16}$$

$$\mathcal{C} = 2\sqrt{2}\pi^{3/2} \epsilon_{\varpi} Z \int_0^{\infty} \int_0^{\infty} \varphi_{P_0} \left(1 - e^{-\frac{r_0}{Z}}\right) r_0 \int_{-\infty}^{m_{lim}} \varphi_{M_{g0}} \, dm_0 \, dP_0 \, dr_0 \tag{5.17}$$

$$\chi = \frac{m_{lim} - 5log(r_0) + 5 - (\rho \log P_0 + \delta)}{\sqrt{2}\sigma_{PL}} \tag{5.18}$$

$$\mathcal{C} = 2\sqrt{2}\pi^{3/2} \sigma_{PL} \epsilon_{\varpi} Z \int_0^{\infty} \int_0^{\infty} \varphi_{P_0} \left(1 - e^{-\frac{r_0}{Z}}\right) r_0 \mathrm{erfc}\left(\chi\right) \, dP_0 \, dr_0 \tag{5.19}$$

The remaining double integral will be performed numerically. The model parameters to be fit are therefore: the slope, zero point, and intrinsic dispersion of the PL relation (ρ, δ, σ_{PL}); the scale height of the Galaxy (Z); and the mean and dispersion of the period distribution (P_{mean}, σ_P).

5.2 Fitting when no parallax data is available

In the near future, before the full release of the final Gaia catalogue, it is still useful to attempt to improve on the calibration of the PL and PLZ relations for variable stars. Even after the Gaia parallax data is made available, in distant objects such as the LMC the relative parallax precision will be low, often in the order of $\sigma_{\varpi}/\varpi = 1 - 10$ (see Fig. 5.1). In such cases, it is useful to calibrate the apparent PL or PLZ relations, without the use of parallax data.

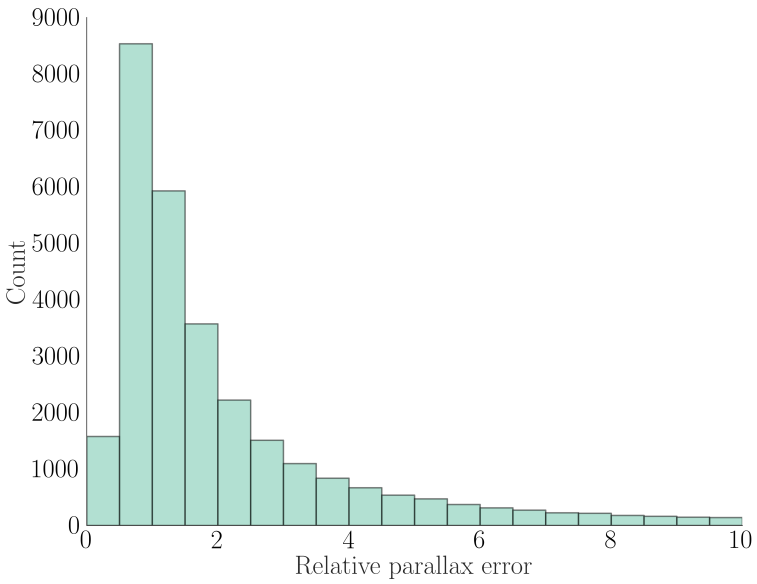


Figure 5.1: Histogram of relative parallax error σ_ϖ/ϖ for GOG simulated RR-Lyrae stars in the LMC. Synthetic catalogue generated from OGLE-III and EROS-II survey data.

This is in theory much more straightforward, as it is essentially just fitting a line to (x, y), or (x, y, z) data.

Fitting a line to data is a common exercise in science. Most common approaches use Minimum-Least-Squares methods, however these are often based on assumptions which do not always hold for real observational data. The most basic methods assume that data are drawn from a perfect straight line, and that errors are Gaussian, perfectly known, and exist in one axis only.

In the example case of fitting the RR-Lyrae PLZ relation, it is clear that we have significant errors in at least two axis, and that there can be substantial intrinsic dispersion around the relation due to other, un-modelled effects. Therefore an attempt has been made to model the data, using a Bayesian approach to fitting a line in three dimensions which will take into account these effects correctly. The method was introduced by Hogg. et. al. 2010, but has been extended for use in three dimensions here.

5.2.1 Basic approach

Initially, we model the data as being drawn from a plane defined by:

$$z = f(x, y) = A\,x + B\,y + C \tag{5.20}$$

where A is the slope in the x axis, B the slope in the y axis, and C is the intercept. In this initial model we assume that we have data in three axis, x, y and z, with errors only in the z axis.

In this model, given an independent position (x_i,y_i), an uncertainty σ_{zi}, a slope A, B, and an intercept C, the frequency distribution $p(z_i|x_i, y_i, \sigma_{zi}, m, b)$ for z_i is

$$p(z_i|x_i, y_i, \sigma_{zi}, m, b) = \frac{1}{\sqrt{2\,\pi\,\sigma_{zi}^2}}\,\exp\left(-\frac{[z_i - A\,x_i - B\,y_i - C]^2}{2\,\sigma_{zi}^2}\right) \tag{5.21}$$

We therefore define the Likelihood as:

$$\mathscr{L} = \prod_{i=1}^{N}\,p(z_i|x_i, y_i, \sigma_{zi}, m, b) \tag{5.22}$$

Taking the logarithm,

$$\begin{aligned}\ln\mathscr{L} &= K - \sum_{i=1}^{N}\frac{[z_i - A\,x_i - B\,y_i - C]^2}{2\,\sigma_{zi}^2} \\ &= K - \frac{1}{2}\,\chi^2\end{aligned} \tag{5.23}$$

which is effectively the least-squares solution. K is a constant. Returning to Bayes rule it is possible to define:

$$p(A, B, C|\{z_i\}_{i=1}^N, I) = \frac{p(\{z_i\}_{i=1}^N|A, B, C, I)\,p(A, B, C|I)}{p(\{z_i\}_{i=1}^N|I)} \tag{5.24}$$

$\{z_i\}_{i=1}^N$ is all the data z_i, I is a all of the knowledge of x_i, y_i and the σ_{zi}, plus

any other prior information we may have.

5.2.2 Multiple errors, no dispersion

However, in our case we have errors in more than one axis, so these can be put together into a covariance matrix $\boldsymbol{S}_i$

$$\boldsymbol{S}_i \equiv \begin{bmatrix} \sigma_{xi}^2 & \sigma_{xyi} & \sigma_{xzi} \\ \sigma_{xyi} & \sigma_{yi}^2 & \sigma_{yzi} \\ \sigma_{xzi} & \sigma_{yzi} & \sigma_{zi}^2 \end{bmatrix} \tag{5.25}$$

With errors in several dimensions, our observed data point (x_i,y_i,z_i) could have been drawn from any true point along the plane (x,y,z). Making the probability of the data, given the model and the true position:

$$p(x_i, y_i, z_i | \boldsymbol{S}_i, x, y, z) = \frac{1}{2\,\pi\,\sqrt{\det(\boldsymbol{S}_i)}}\,\exp\left(-\frac{1}{2}\,[\boldsymbol{Z}_i - \boldsymbol{Z}]^\top\,{\boldsymbol{S}_i}^{-1}\,[\boldsymbol{Z}_i - \boldsymbol{Z}]\right) \tag{5.26}$$

where we have implicitly made column vectors

$$\boldsymbol{Z} = \begin{bmatrix} x \\ y \\ z \end{bmatrix} \quad ; \quad \boldsymbol{Z}_i = \begin{bmatrix} x_i \\ y_i \\ z_i \end{bmatrix} \tag{5.27}$$

In only two dimensions (e.g. y and z), the slope (e.g. B) can be described by a unit vector $\hat{\boldsymbol{v}}$ *orthogonal* to the line or linear relation (at any x):

$$\hat{\boldsymbol{v}} = \frac{1}{\sqrt{1+B^2}} \begin{bmatrix} -B \\ 1 \end{bmatrix} = \begin{bmatrix} -\sin\theta \\ \cos\theta \end{bmatrix} \tag{5.28}$$

where we have defined the angle $\theta = \arctan B$ made between the line and the y axis. The orthogonal displacement Δ_i of each data point (y_i, z_i) from the

line is given by

$$\Delta_i = \hat{\boldsymbol{v}}^\top \begin{bmatrix} y_i \\ z_i \end{bmatrix} - C\,\cos\theta \tag{5.29}$$

Instead of extending fully into three dimensions, we will assume a negligible error in x (which will be the period, so has justifiably higher precision). x can be directly input into Δ_i without worrying about the interplay between the other parameters.

$$\Delta_i = \hat{\boldsymbol{v}}^\top \begin{bmatrix} y_i \\ z_i \end{bmatrix} - (C\,\cos\theta + A\,x_i) \tag{5.30}$$

Assuming negligable errors in x also redefines the covariance matrix of the errors as:

$$\boldsymbol{S}_i \equiv \begin{bmatrix} \sigma_{yi}^2 & \sigma_{yzi} \\ \sigma_{yzi} & \sigma_{zi}^2 \end{bmatrix} \tag{5.31}$$

Similarly, each data point's covariance matrix $\boldsymbol{S}_i$ projects down to an orthogonal variance Σ_i^2 given by

$$\Sigma_i^2 = \hat{\boldsymbol{v}}^\top\,\boldsymbol{S}_i\,\hat{\boldsymbol{v}} \tag{5.32}$$

and then the log Likelihood for (A, B, C) or $(A, \theta, C\,\cos\theta)$ can be written as

$$\ln\mathscr{L} = K - \sum_{i=1}^{N} \frac{\Delta_i^2}{2\,\Sigma_i^2} \tag{5.33}$$

where K is some constant. This Likelihood can be maximized to find A, B and C.

5.2.3 Intrinsic dispersion

The final step is to introduce an intrinsic variance in the line, V, orthogonal to the line.

According to Hogg 2010, each data point can be treated as being drawn from a projected distribution function that is a convolution of the projected uncertainty Gaussian, of variance Σ_i^2 defined above, with the intrinsic scatter Gaussian of variance V. Therefore the Likelihood becomes:

$$\ln \mathscr{L} = K - \sum_{i=1}^{N} \frac{1}{2} \ln(\Sigma_i^2 + V) - \sum_{i=1}^{N} \frac{\Delta_i^2}{2\,[\Sigma_i^2 + V]} \tag{5.34}$$

where again K is a constant, everything else is defined as above.

5.3 The RR-Lyrae PLZ relation

The method from Sect. 5.2 has been implemented as a line-fitting package, and accepts a set of input data (x, y, z, σ_y, σ_z), using MCMC (Foreman-Mackey et al., 2013) to sample the posterior PDF of the four parameters (A, B, C, V). The method provides the MAP estimates for the model parameters, the formal error on the estimates, and the full posterior PDF. The method has been tested extensively using simulations. The method has been used in collaboration with the INAF-University of Bologna team (G. Clementini, T. Muraveva) to calibrate the RR-Lyrae PLZ relation for a catalogue of 71 RR-Lyrae stars in the LMC, with spectroscopically determined metallicity, precise periods from the third Optical Gravitational Lensing Experiment (OGLE III) catalogue (Udalski et al., 2008) and multi-epoch K_s photometry from the Vista for Magelanic Clouds (VMC) survey (Cioni et al., 2011). K_s photometry has been used because it is in the infrared that RR-Lyrae stars exhibit a tight PLZ relation. Additionally, the effect of interstellar extinction is weaker in the infrared passbands, and therefore the effects of differential reddening along the depth of the LMC are reduced. The LMC is an excellent location to study variable stars because they can often be assumed to be at the same distance. For study of variable stars in the Galaxy, the distance to each individual star must be known in order to meaningfully compare the magnitudes of the stars, and for this reason calibration of PL relations can be more difficult to achieve.

The results of the application of the method are detailed in an article which is under preparation: **A new Period-Luminosity-Metallicity relation for RR Lyrae stars and the impact of Gaia. Muraveva, T.; Palmer, M.; Clementini, G.; Luri, X.; et. al. Article in preparation.**

In order to calibrate the $PL_{K_s}Z$ relation a large sample of RR Lyraes is required, which span a wide range of metallicities and for which accurate K_s and [Fe/H] measurements are available. We selected the 71 RR Lyraes in the Large Magellanic Cloud (LMC) which have spectroscopically determined metallicities in the range [-2.06; -0.63] dex (Gratton et al., 2004). All these RR Lyraes have counterparts in the OGLE III catalogue (Soszyński et al., 2009), therefore very precise periods are available. In order to increase the accuracy of the determination of the mean K_s magnitudes, multi-epoch photometry is needed. For this reason we are using the data of the near-infrared *Vista for Magellanic Clouds (VMC)* survey, which performs K_s observations in 12 epochs. Many previous studies use single-epoch photometry from the Two Micron All-Sky Survey (2MASS, Cutri et al. (2003)). To fit the $PL_{K_s}Z$ relation we apply the Bayesian model fitting approach from Sect. 5.2.

By applying this method we found the following relation between periods of pulsation, metallicities and mean apparent K_s magnitudes determined with templates[1]:

$$\begin{aligned} K_{s,0} &= (-2.70 \pm 0.22)\log P + (0.03 \pm 0.06)[Fe/H] \\ &+ (17.44 \pm 0.05) \end{aligned} \tag{5.35}$$

Note that the actual fitting was for a function of the form: $z = f(x, y) = A\,(x+0.25) + B\,(y+1.5) + C$ in order to avoid a strong correlation between the gradients and zero point. The data was shifted by 0.25 in period and 1.5 in

[1]Note that this work is still in progress and the selection of stars may be modified in the near future. For the full list of selected RR-Lyraes and the final PLZ relation, please refer to Muraveva et. al (in prep).

metallicity to give a zero point more or less central to the data, and after the fitting procedure both the data and fit are returned to the initial configuration.

The intrinsic dispersion of the relation is found to be 0.09 mag. The RMS deviation of the data around the relation, neglecting the intrinsic dispersion, is 0.09 mag. The resulting fit and posterior PDF of the parameters are shown in Figs. 5.2 and 5.3. The projections of the $PL_{K_s}Z$ relation (Eqn. 5.35) on the $\log(P) - K_s$ and $K_s - [Fe/H]$ planes is shown in Figure 5.2. The grey lines in the figure are lines of equal metallicity (top) or equal period (bottom). For a clear representation of the scatter of the points around the relation, fig. 5.4 shows the relation in three dimensions.

It is worth noting that we find a very small (effectively zero) dependence of the K_s magnitude on the metallicity.

5.4 The distance to the Large Magellanic Cloud

As mentioned in Sect. 5.2, the LMC is an excellent object for study of variable stars. In fact, observation of the population of variable stars in the LMC was responsible for the initial discovery of a PL relation by Leavitt (1908). The relatively short depth of the LMC allows the approximation that all of the stars are at the same distance, meaning that calibration of the apparent PL or PLZ relation is possible with no distance information whatsoever (as seen in Sect. 5.3). An absolute relation would give information on the luminosity of the stars, aiding work on pulsation mechanisms and stellar evolution theory, and would additionally allow direct comparison between LMC variables and those within the Milky Way. To calibrate an absolute relation, distance measurements are required for the individual variable stars involved, or globally for the entire system.

Gaia will for the first time provide a statistically significant sample of measurements of absolute trigonometric parallax for stars within the Magellanic Clouds, and for this reason could provide an excellent tool for constraint of their distances. This would allow direct calibration of absolute relations for

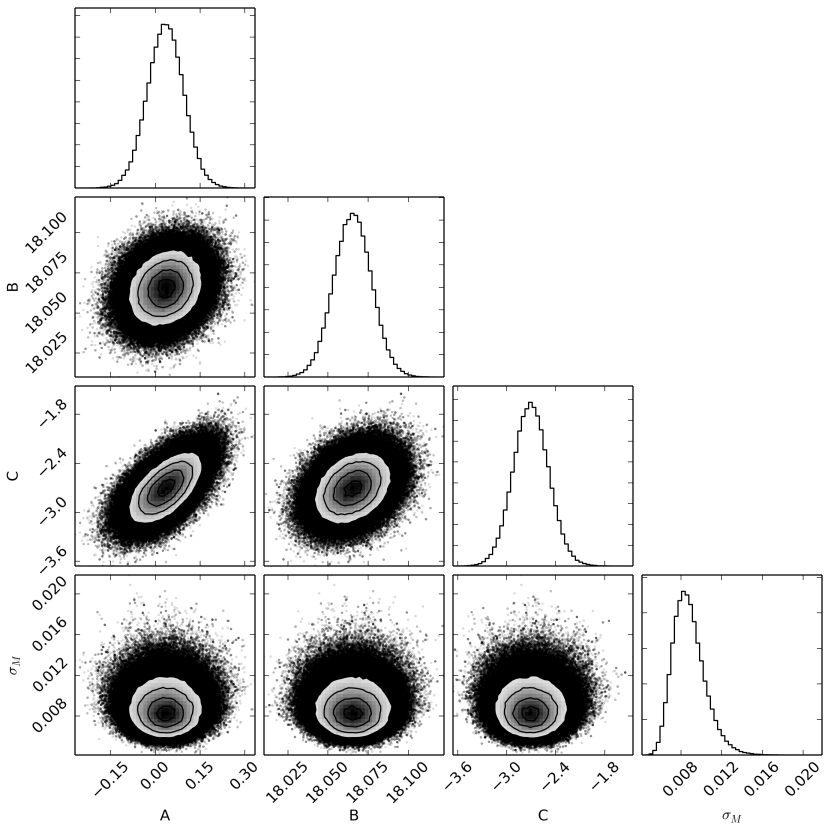


Figure 5.2: Triangle plot showing the posterior PDF of the parameters, along with correlations between each pair of parameters.

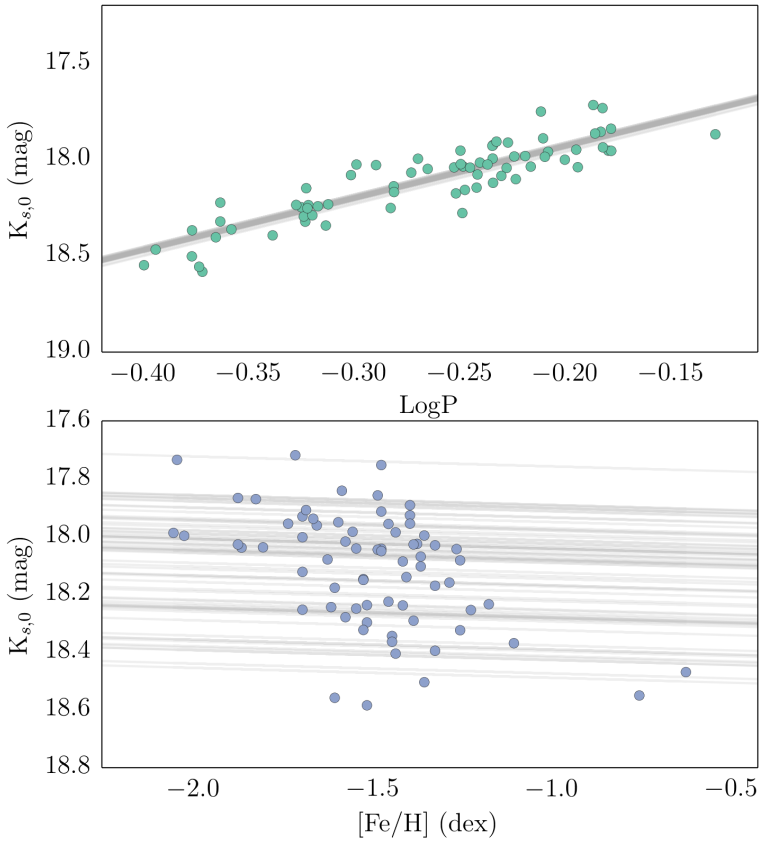


Figure 5.3: Projections of the PLK_s relation on the log(P)-K_s (top panel) and [Fe/H]-K_s (bottom panel) planes. Grey lines represent lines of equal metallicities (top panel) and periods (bottom panel) intersecting each star.

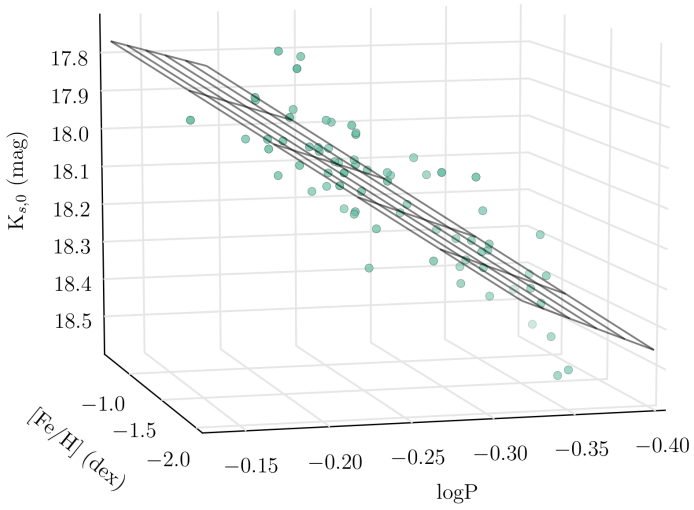


Figure 5.4: The PLZ relation in three dimensions, showing the plane of the fitted relation and the dispersion of the stars around it.

variable stars within the LMC. Additionally, a precise estimate for the distance to the LMC would allow calibration of several other distance indicators, and an improvement in the calibration of the extragalactic distance scale.

In this section we define a method for estimation of the distance to the LMC, and test it using simulated Gaia data. In the following section we look at the calibration of the distances to individual variable stars for the cases of Cepheids and RR-Lyraes and the potential impact for the calibration of their respective PL relations.

5.4.1 Distance estimation with Gaia

Given the very large distance to stars in the Magellanic Clouds, determination of their distances is one of the most extreme use cases for Gaia parallaxes. Gaia is expected to observe several million stars in the LMC, however the distances involved are such that the error in parallax will be in most cases many times larger than the parallax itself.

To test Gaia's capability in obtaining a distance estimate to the LMC, simulated Gaia data has been generated using GOG. Synthetic catalogue data for the LMC is obtained from a photometric catalogue (Belcheva private communication, as indicated in GAIA-C2-TN-LAOB-AR-004-10), and so is a realistic catalogue of stars which Gaia will be capable of observing. The synthetic catalogue is processed using GOG, which contains a full mathematical description of the nominal performance of the Gaia mission, in order to produce simulated Gaia end-of-mission catalogue data including realistic observational errors. The sky density of the simulated stars in the region of the LMC can be seen in Fig. 5.5.

GOG uses an input catalogue of roughly 7.5 million LMC stars which are brighter than $G = 20$ mag. Position, apparent magnitude and colour information is obtained from the photometric catalogue on which the simulation is based. A mean LMC distance of 48 kpc (or a mean parallax of 20.83 μas) from Macri et al. (2006) is assumed, and stars are simulated with a Gaussian

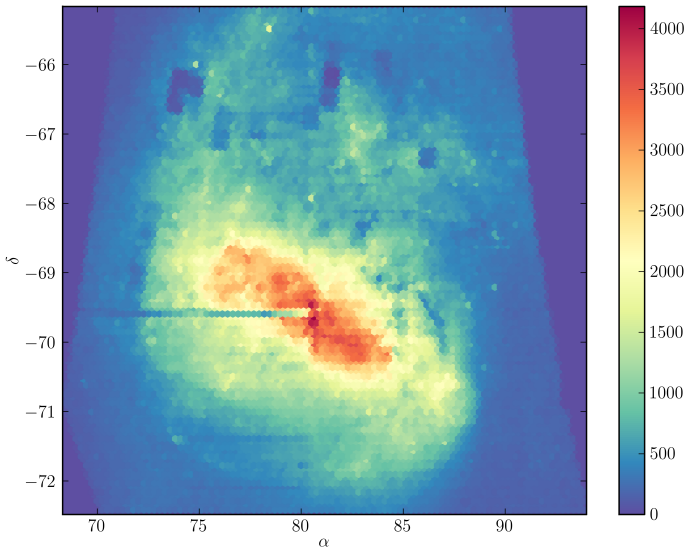


Figure 5.5: Sky density map of the GOG simulated stars in the LMC.

distribution around the mean with a standard deviation of 750 pc, following Sakai et al. (2000). After simulation of Gaia observations of LMC stars using GOG, the distribution of relative parallax error in the GOG simulated Gaia data is shown in Fig. 5.6. With the precision of Gaia's astrometric measurements around 10 μas for the brightest stars and dropping to a few hundred μas at the faint limit, it is clear that the majority of stars will be individually very poorly determined.

5.4.2 Method

Taking the mean of all parallax measurements could theoretically provide a mean parallax with a precision of around 0.25 μas, taking into account the individual parallax errors and the scaling of the precision with $1/\sqrt{N}$, where N is the number of stars. However, the direct calculation of the mean parallax is susceptible to several of the biases highlighted in Sect. 2.

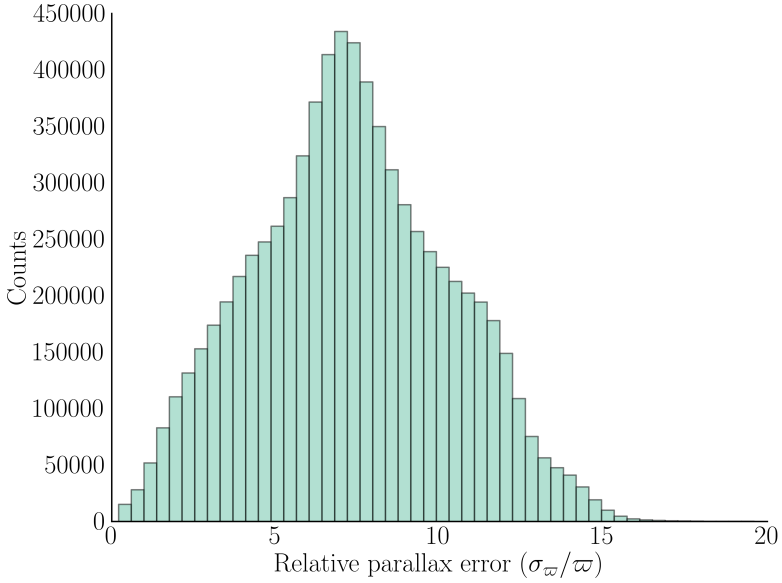


Figure 5.6: Histogram of the relative error in parallax (σ_ϖ/ϖ) for all of the 7.5 million GOG simulated stars in the LMC.

The situation is improved by explicitly modelling the true parallax distribution and fitting the mean distance and dispersion as parameters. We assume a Gaussian distribution around some mean distance, $d_{\rm mean}$ with dispersion, σ_d. We label the parameters as the vector $\boldsymbol{\theta} = (d_{\rm mean}, \sigma_d)$, and the set of observations as the vector $\boldsymbol{x} = [(\varpi, \epsilon_\varpi)_1, (\varpi, \epsilon_\varpi)_2, ..., (\varpi, \epsilon_\varpi)_N]$.

We can therefore define the Likelihood of our parameters $\boldsymbol{\theta}$, given a set of observations $\boldsymbol{x}$:

$$\mathcal{L}(\boldsymbol{\theta}|\boldsymbol{x}) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma_d} e^{0.5\left(\frac{d_{r,i}-d_{\rm mean}}{\sigma_d}\right)^2} \frac{1}{\sqrt{2\pi}\epsilon_i} e^{0.5\left(\frac{\varpi_i - 1/d_{r,i}}{\epsilon_i}\right)^2} \tag{5.36}$$

Or equivalently taking the log Likelihood as:

$$\ln\mathcal{L}(\boldsymbol{\theta}|\boldsymbol{x}) = \sum_{i=1}^{N} \ln\left(\frac{1}{2\pi\sigma_d\epsilon_i}\right) - 0.5\left(\left(\frac{d_{r,i}-d_{\rm mean}}{\sigma_d}\right)^2 + \left(\frac{\varpi_i - 1/d_{r,i}}{\epsilon_i}\right)^2\right) \tag{5.37}$$

Here, $d_{r,i}$ is the real distance to star i. This can be fit as a parameter, to obtain the a posteriori estimate of the distance to an individual star, or marginalised through integration.

The log Likelihood function as defined above has several advantages in terms of estimating a mean distance: observational errors are taken into account; the explicit inclusion of a distance distribution and non-truncation of the observed parallax avoids the Lutz-Kelker bias; the observed parallax is used directly so there are no non-linear transformations in the equations which can bias the result. As the model relies only on trigonometric parallax, there is no need for external assumptions or corrections due to effects such as reddening.

Using Bayes theorem to define the log Likelihood function, we find the Maximum A Posteriori Probability (MAP) estimate of the parameters of our model by maximising:

$$\theta_{\rm MAP}(\boldsymbol{x}) = \ln\mathcal{L}(\boldsymbol{x}|\boldsymbol{\theta})g(\boldsymbol{\theta}) \tag{5.38}$$

where the function $g(\boldsymbol{\theta})$ is our prior PDF. We have assumed wide priors.

This method has been implemented using an adaptive Markov Chain Monte Carlo (MCMC) sampler (Foreman-Mackey et al., 2013) to construct the posterior probability density function (PDF) of the parameters. The method returns the Maximum A Posteriori Probability (MAP) estimate of the parameters, the formal error, and the complete posterior PDF.

The MCMC sampler performs a semi-random walk in the parameter space in order to evaluate the PDF in regions of non-negligible probability in an efficient way. The type of MCMC algorithm used here is the Goodman and Weare Affine Invariant MCMC Ensemble sampler, which allows numerous simultaneous Markov chains, called walkers, which start at some range of initial guesses and move with a semi-random walk with preference towards regions of higher probability. The sum of all of the positions of all of the walkers over this time gives the sampling of the PDF of the parameters. The initial few steps, before the system has reached a state of equilibrium, is called the burn in period. These steps will not be included in the final sample as they are affected by the initial guess.

The method has been tested using the simulated GOG data for the brightest 700 stars, without correlated errors (see Ch. 7 for a discussion of error correlations in Gaia), and recovers the mean distance of the LMC to within 0.2 kpc, or 0.4%, of the true value. The mean distance is estimated to be $R = 48.19^{+0.45}_{-0.44}$ kpc ($R_{\rm true} = 48.0$), with a variance around the mean position of $\sigma_R = 2.12^{+0.93}_{-0.70}$ kpc ($\sigma_{R_{\rm true}} = 0.750$). The posterior PDF of the parameters obtained using MCMC is shown in Fig. 5.7.

One of the key strengths of the use of MCMC is the possibility of reconstructing the entire posterior PDF. In Fig. 5.7 we can see that the PDF of the intrinsic dispersion σ_R is highly asymmetric. Additionally there is a visible correlation between the distance and the dispersion. This slight correlation is expected, as the expected size of the LMC should increase if the distance parameter is overestimated.

This case study highlights the possibility of fitting mean parallaxes to a much

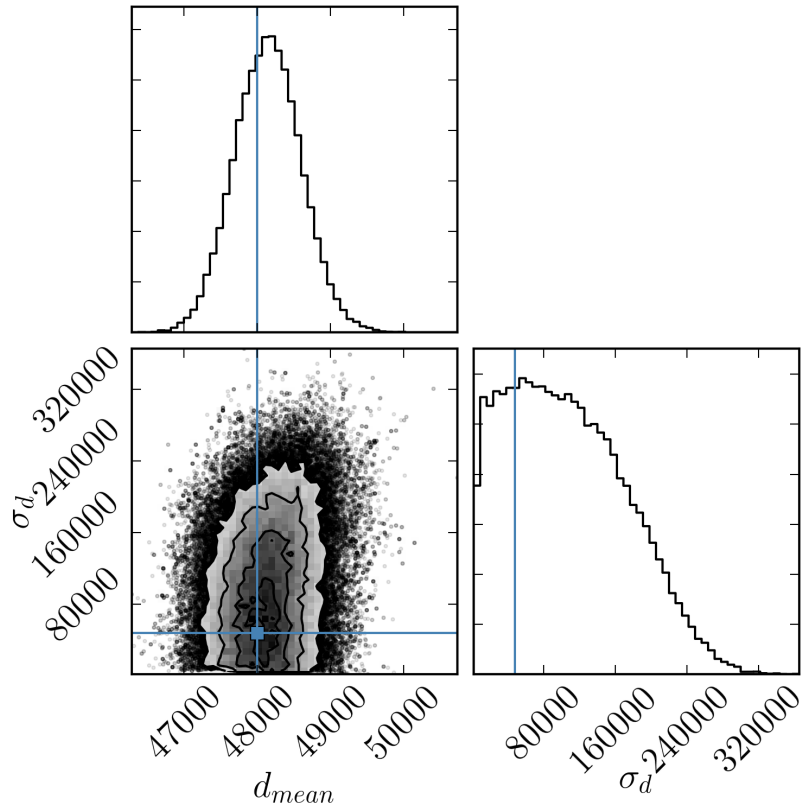
greater precision than that of the individual stars in the sample. However, error correlations (see Ch. 7) may play a role in limiting the maximum achievable accuracy if the stars used are concentrated in a small area on the sky.

5.4.3 The orientation of the LMC

As a brief aside, we will look at the application of this method to determine the orientation of the LMC disk. The LMC is widely known to be slightly tilted with respect to the viewing angle on the plane of the sky. Inclination angles of $35.8° \pm 2.4°$ (Olsen & Salyk, 2002) and $32° \pm 4°$ (Haschke et al., 2012) have been reported, although a new analysis of LMC RR-Lyrae stars finds an inclination angle of only $11.84° \pm 0.8°$ (Klein et al., 2014), in strong disagreement with previous studies. Methods for determining the tilt angle of the LMC rely on distance indicators such as Cepheids, RR-Lyraes or red clump stars, and are therefore dependent on the calibration of these distance indicators, and other important effects such as interstellar extinction.

The method from Sect. 5.4.2 estimates the mean distance to the LMC directly using trigonometric parallax. While it would be possible to model the 3D structure of the LMC disk in order to fit its angle and orientation, a more straight forward approach is possible. By simply splitting up the LMC stars into a grid on the sky, it is possible to obtain an estimate of the mean distance to each grid cell.

To test this approach the distances of the GOG LMC input catalogue were modified to include a tilt angle of 35° (for simplicity the line of nodes was chosen parallel with the galactic longitude, l). The results of the estimation of the distance to a grid imposed over the LMC area are shown in Fig. 5.8. The colour scale shows the relative difference between the global mean distance and the distance estimated for each grid cell. From the figure it is clear that the line of nodes (parallel with galactic longitude for convinence) has been accurately obtained. Taking a 3° angular scale at a distance of 48 kpc provides a size perpendicular to the plane of the sky of 2.5 kpc. Over 3° in either direction



[ht!]

Figure 5.7: Triangle plot of the results of the MCMC sampling of the Likelihood function. Histograms are the posterior PDF of: the mean LMC distance (R), the depth of the LMC ($_R$, $\times 64$ to give same magnitude to each parameter in the minimisation). Also shown is the correlation between each pair of parameters. The blue lines represent the true value, included in the initial simulation.

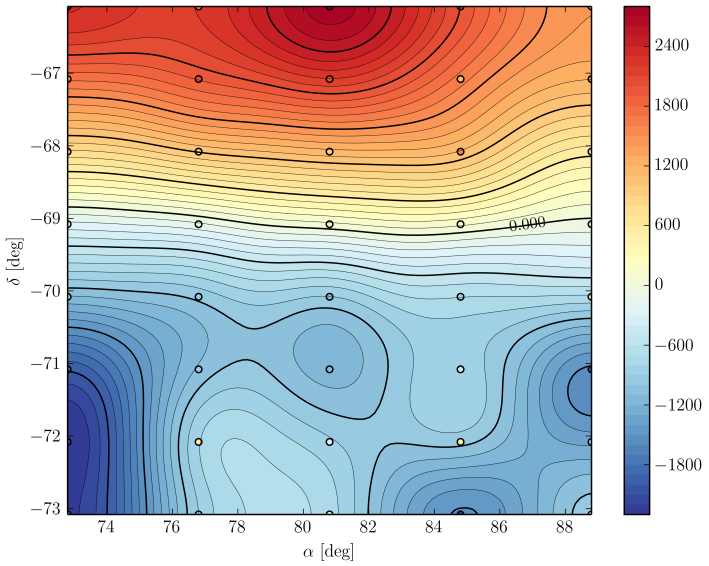


Figure 5.8: Points show the centre of each cell of a grid covering the LMC area. The size and orientation of the grid is arbitrary and chosen to allow a reasonable number of stars per cell. The colour scale represents the residual of the global mean distance and the mean distance of stars within a cell $d_{mean} - d_{cell}$. Using the angular size and the distance to the LMC, the residual difference corresponds to an observed tilt angle of 37°. The contour plot and colour fill is the smoothed residual extrapolated over the entire area.

from the line of nodes there is a distance residual of around 2 kpc (see Fig. 5.8), which translates to a tilt angle of 37°. Before performing a more rigorous analysis, we direct the reader to Sect. 7.4 on error correlation effects on Gaia LMC data.

5.5 Variables in the Large Magellanic Cloud

Included within the large number of stars which Gaia will observe in the Magellanic Clouds will be a significant population of variable objects. Variable stars have a long history of use in luminosity calibration and the distance scale, due to their potential use in calibration as standard candles and though rela-

tions between the absolute magnitude of a star to observables such as period, metallicity or colour.

Study of variable stars in the LMC is ongoing. Because observables such as period are measurable at kpc distances, stars in the Magellanic Clouds are easily accessible, and well calibrated relations can serve as a excellent tool for distance determination to the Milky Way's satellite galaxies and in the extra-galactic distance scale (Kanbur et al., 2003). Calibration of the relations is still being undertaken, either for absolute relations such as those by Gieren et al. (1998), or by extrapolating Galactic relations out to LMC variables which can then be used to determine the LMC distance (e.g. Clement et al., 2008). One major problem in this extrapolation of relations is the possibility of there being a dependence on age or metallicity which would result in different relations applying for different galaxies (Storm et al., 2011). This is of particular concern for the LMC which is much more metal poor than the Milky Way. It can therefore be useful to provide a calibration for the absolute luminosity of LMC variable stars and their relations by direct means, rather than relying on relations calibrated within the Milky Way.

In order to test the potential of Gaia in LMC variable star calibration it is first important to get an idea of the numbers and types of variables which will be observed. Two recent surveys of the LMC, the Optical Gravitational Lensing Experement (OGLE) (Udalski et al., 2008) and The Expérience pourla Recherched 'Objets Sombres (EROS) (Tisserand et al., 2007), have produced catalogues of both Cepheid variable stars and RR-Lyrae stars in the LMC.

EROS is a microlensing experiment, and due to its sampling period was very useful for detecting periodic variable stars and sampling their light curves. Work has been performed in order to characterise the different variable types contained within the EROS data, finding 6607 RR Lyraes, 638 Classical Cepheids, and 178 Type II Cepheids (Kim et al., 2014).

OGLE is another gravitational lensing experiment, which in the OGLE-III data release provided a catalogue of variable stars in the LMC with very good coverage over the LMC area. The catalogue contains classification of

3361 classical Cepheids (Soszyński et al., 2008) and 24906 RR Lyrae stars (Soszyński et al., 2009). The OGLE-III catalogue contains more variable stars due to covering the outer regions of the LMC, well beyond the main disk.

Work has been undertaken to use these stars to calibrate the PL, PLZ, or PLC relations (e.g. Udalski et al., 1999b), yet for the zero point it is always necessary to rely on external distance estimates for the mean distance to the LMC.

The recent VISTA near-infrared Magellanic Cloud survey has combined the OGLE-III and EROS variabe star data with multi-epoch photometry in Y, J, and K_S bands with sampling specifically selected for the lightcurves of RR Lyrae and Cepheids with periods up to 30 days. To test Gaia's capability in the calibration of the Cepheid and RR Lyrae PL relation in the LMC, synthetic catalogues of both types of variables have been created from this combined catalogue.

The photometric catalogue provides a complete list of Cepheids in the LMC, including positions, apparent magnitude in V and I, pulsation period, and pulsation mode (fundamental, 1^{st} overtone, etc.). Lacking from the input information is the distance, which is required in order to simulate the parallax observations.

The OGLE-III catalogue is used here due to its use of the Johnson V and I bands, and the internal classification of pulsation mode. As can be seen in Fig. 5.9, the survey area covers the majority of the LMC. All 1831 of the fundamental pulsating Cepheids in OGLE-III are included in the following discussion.

The production of the synthetic catalogue is achieved though the following steps:

- The catalogue provides apparent V and I magnitudes, period and position. These are taken to be known 'perfectly' with negligible errors.

- The distance is generated randomly using a Gaussian distribution with mean 48 kpc and variance of 750 pc. This distance is an arbitrary choice,

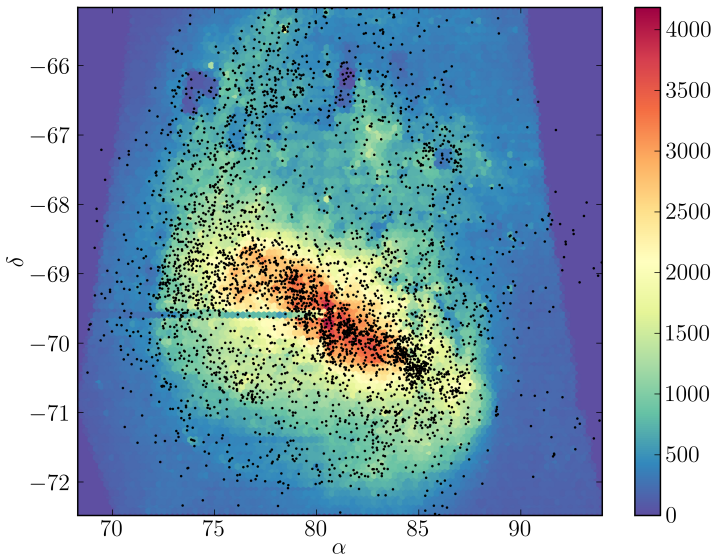


Figure 5.9: Density map of the 7.5 million stars simulated by GOG, with the OGLE-III and EROS Cepheids overlaid as black dots.

with a reasonable approximation used for the distance and taken from now to be the 'true' distance.

- The absolute V and I magnitudes are obtained from the period using PL relations for the LMC from Udalski et al. (1999c):

$$M_V = -2.760 \log P - 1.18 \qquad \sigma = 0.159 \tag{5.39}$$

$$M_I = -2.962 \log P - 1.66 \qquad \sigma = 0.109 \tag{5.40}$$

- The extinction is calculated from the above parameters, using:

$$A_x = m_x + 5 - (M_x + 5 \log d) \tag{5.41}$$

where x is either V or I, and assuming that the extinction is the only cause of loss of flux. Given that for a given star the distance and the dispersion around the PL relation are drawn as random variables, the extinction is also a random quantity, but statistically will be representative of the true extinction.

- We therefore have position, period, absolute magnitude in V and I, distance and extinction in the V band. These quantities are taken to be the 'real' parameters of our catalogue and are processed using the Gaia Object Generator (GOG) to obtain simulated Gaia observations of each star, including observational errors. GOG handles the conversions between Johnson and Gaia bands internally.

5.5.1 Method

GOG provides a catalogue of 'observed' stars, including astrometric and photometric information. As can be seen by the distribution in apparent magnitudes in Fig. 5.10, the magnitudes of the Cepehids in this sample are brighter than

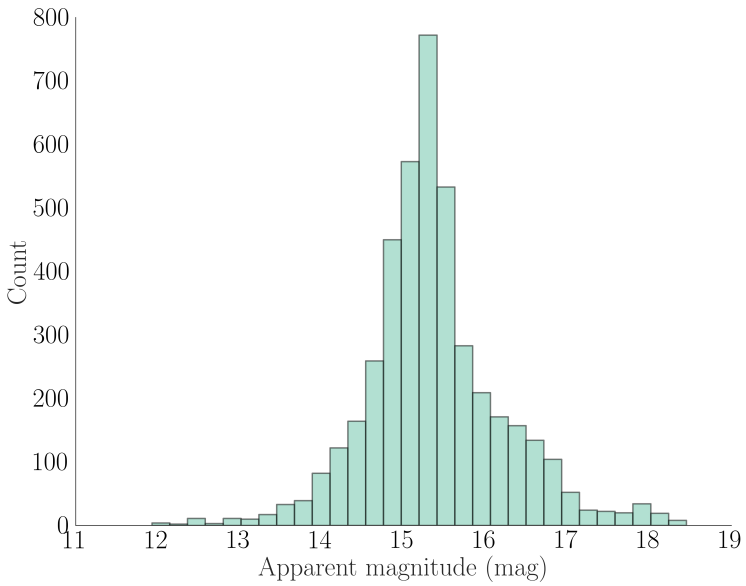


Figure 5.10: Distribution of apparent magnitues.

$G = 17$ mag. We can therefore assume that, to a reasonable approximation, all LMC Cepheids will be observed and there is no limiting magnitude. For the estimation of distances we apply the same formulation as that of Sect. 5.4.

5.5.2 Application

The MCMC method is used to obtain the mean and variance of the distribution in distances. These values are the median of the posterior distribution, with the errors obtained from the 15.87 and 84.13 quartiles respectively (representing 1σ assuming that the posterior distributions are Gaussian).

Using all 1831 fundamental mode classical Cepheids, the mean distance to the LMC is found to be $R = 50.96^{+1.54}_{-1.30}$ kpc ($R_{\text{true}} = 48.0$), with a variance around the mean position of $\sigma_R = 8.1^{+2.0}_{-2.6}$ kpc ($\sigma_{R_{\text{true}}} = 0.750$). Clearly the determination of the variance of the distribution, which is a measure of the depth of the LMC, is not accurate, and the very large error bars confirm that

there is insufficient data to accurately constrain the depth[2] of the LMC using only Cepheid parallax information. The mean distance obtained directly from the mean of the inverse of the parallax is 111.4 kpc, which serves to reiterate that the directly calculated mean distance is unreliable due to being susceptible to various biases. The above estimate of the distance to the LMC is accurate to roughly 6%.

As there are a total of 7.5 million stars in the entire simulated LMC, there are many very bright stars with better parallax precision than the stars in the Cepheid sample. By using the global distance estimate to the LMC given in Sect. 5.4.1 it is possible to improve on the distance estimate for Cepheid variable stars alone. However, it is necessary to assume that the distance distribution of the Cepheid population is the same as that of the general stellar population.

5.5.3 Recovering the Cepheid P-L relation

In the construction of the synthetic catalogue, the period luminosity relation assumed from Udalski et al. (1999c) was used, as shown in Fig. 5.11.

GOG is capable of transforming from Johnson bands into Gaia bands, and so the GOG output is given for Gaia band G as will be in the final Gaia catalogue. In Fig. 5.11 the fit to the PL relation to the true values (not affected by observational errors) is shown, for the Gaia band G. It is clear that the slope and dispersion are largely the same in both bands. This will be assumed to be the 'real' Gaia PL relation and an attempt will be made to recover this from the observations.

Given the distance estimate of $R = 48.19^{+0.45}_{-0.44}$, the absolute magnitude of the star can be calculated using the method derived for estimating variable star PL relations in Sect. 5. We will assume that all stars are at the same distance. This assumption will cause a slight overestimation of the intrinsic

[2]Note that we approximate the depth of the LMC as a Gaussian distribution in distance. This could easily be changed to a more complex distribution if required.

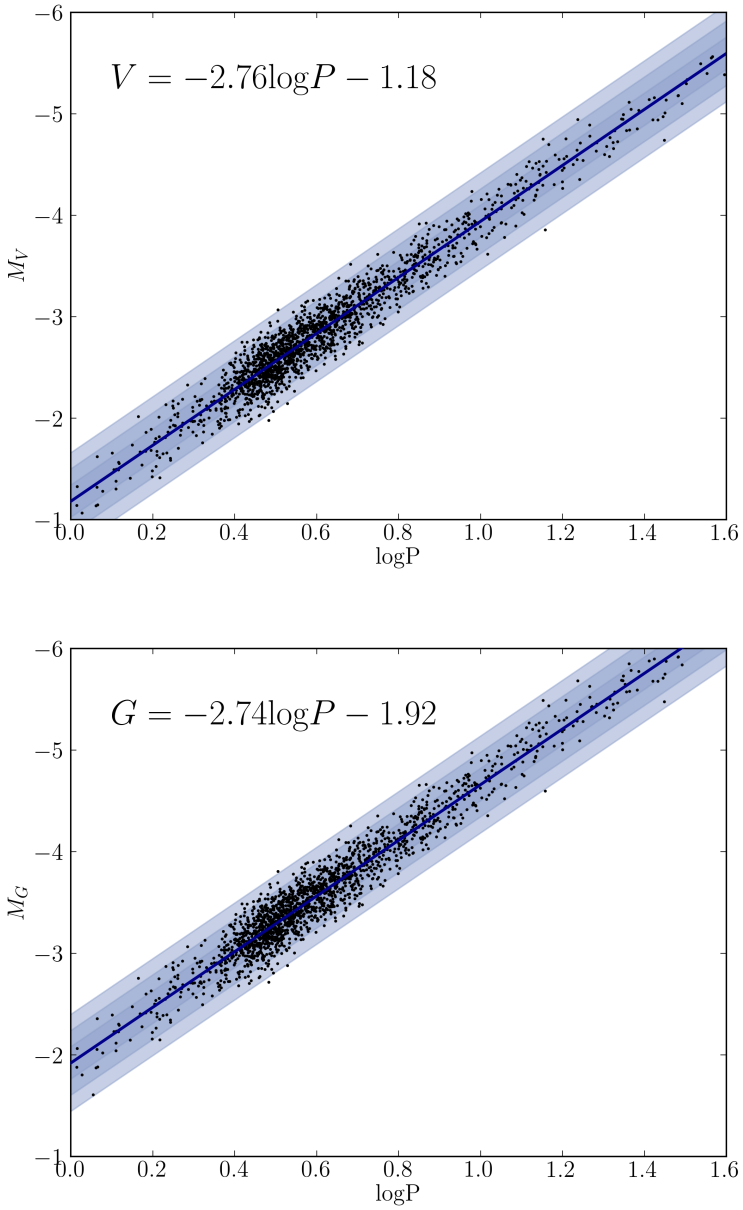


Figure 5.11: Johnson V band (top) and Gaia G band (bottom) period luminosity relation assumed for LMC fundimental mode classical Cepheids. The blue bands are the 1, 2 and 3 σ dispersion around the PL relation. No observational errors have been applied.

dispersion, due to the differences in apparent magnitude caused by differences in distance not being taken into account. Using the LMC distance and depth of 48 kpc and 750 pc from Macri et al. (2006) and Sakai et al. (2000), the depth of the LMC causes relative differences of only 1.5%. Recent results by Haschke et al. (2012) specifically for the Cepheid population find a LMC depth of 1.7 ± 0.2 pc, with a distance of 53.9 ± 1.7 pc for the same population. This corresponds to effects of the depth of the LMC causing relative differences of only 3%.

Defining a model for the distribution in absolute magnitude

$$\varphi_M = e^{-0.5\left(\frac{M-M_{mean}}{\sigma_{PL}}\right)^2} = e^{-0.5\left(\frac{M-(\rho \log P+\delta)}{\sigma_{PL}}\right)^2} \tag{5.42}$$

where σ_{PL} is the intrinsic dispersion of the PL relation, and the parameters ρ and δ are the slope and zero point. Assuming that the uncertainty in the period is negligible compared with the other parameters, the error distribution for observed quantities is given by:

$$\mathcal{E}(\boldsymbol{x}|\boldsymbol{x}_0) = e^{-0.5\left(\frac{m_g-m_{g0}}{\epsilon_{m_g}}\right)^2} e^{-0.5\left(\frac{R-R_0}{\epsilon_R}\right)^2} e^{-0.5\left(\frac{A_g-A_{g0}}{\epsilon_{A-g}}\right)^2 \delta(\ logP)} \tag{5.43}$$

Therefore the joint Likelihood function is:

$$\mathcal{P}(\boldsymbol{x}|\boldsymbol{\theta}) = \frac{\mathcal{S}(\boldsymbol{x})}{\mathcal{C}} \int_{\forall \boldsymbol{x}_0} \varphi_M \mathcal{E}(\boldsymbol{x}|\boldsymbol{x}_0)\, d\boldsymbol{x}_0 \tag{5.44}$$

where the normalisation coefficient is:

$$\mathcal{C} = 4\pi^2 \sigma_{PL} \epsilon_m \epsilon_R \epsilon_{A_g} \tag{5.45}$$

This integration will be performed numerically. The final free-fit parameters are therefore the slope, zero point and intrinsic dispersion of the PL relation. While it would be more complete to fully model the spatial distribution along with the distribution in absolute magnitude, the precision on the individual

parallaxes is too low to allow precise fitting of the distance using only Cepheid parallax measurements, as seen in Sect. 5.5. For this reason it is beneficial to include the global distance estimate calculated from the inclusion of all of the stars in the catalogue.

The result of the application of the above method is shown in Fig. 5.12, for fitting the PL relation for the GOG simulated LMC Cepheid population. In the simulated data there are 316 stars with negative prallax, making up around 20% of the total, which can not be used to calculate the absolute magnitude due to having to take the logarithm of a negative number. These data points do however contain information, and removing them would bias the sample by including stars for which random measurement errors have scattered the result in the positive direction rather than negative. This is one benefit of the method used here, as the parallax can be used directly even if negative. When plotting the observed PL relation, the stars with negative parallax can not be included. Additionally the errors bars are asymmetric. It is therefore useful to define the Astrometric Based Luminosity (ABL, Arenou & Luri (1999)) as:

$$\mathrm{ABL} = 10^{0.2M_V} = 10^{\frac{m_V+5}{5}} \tag{5.46}$$

where m_V has been corrected for extinction and the parallax is in arcsec. Plotting the PL relation using the ABL results in symmetric error bars and allows the inclusion of negative parallax data, leading to a clearer visualisation of the resulting fit.

5.6 Conclusions

Here the ML method introduced in Ch. 4 has been extended to calibration of the Cepheid PL relation. The method has been implemented and is ready to use when Gaia data becomes available. Currently, there is a severe lack of parallax measurements for Cepheids and RR-Lyraes. This has made application of the method to real data impossible, but highlights the potential use

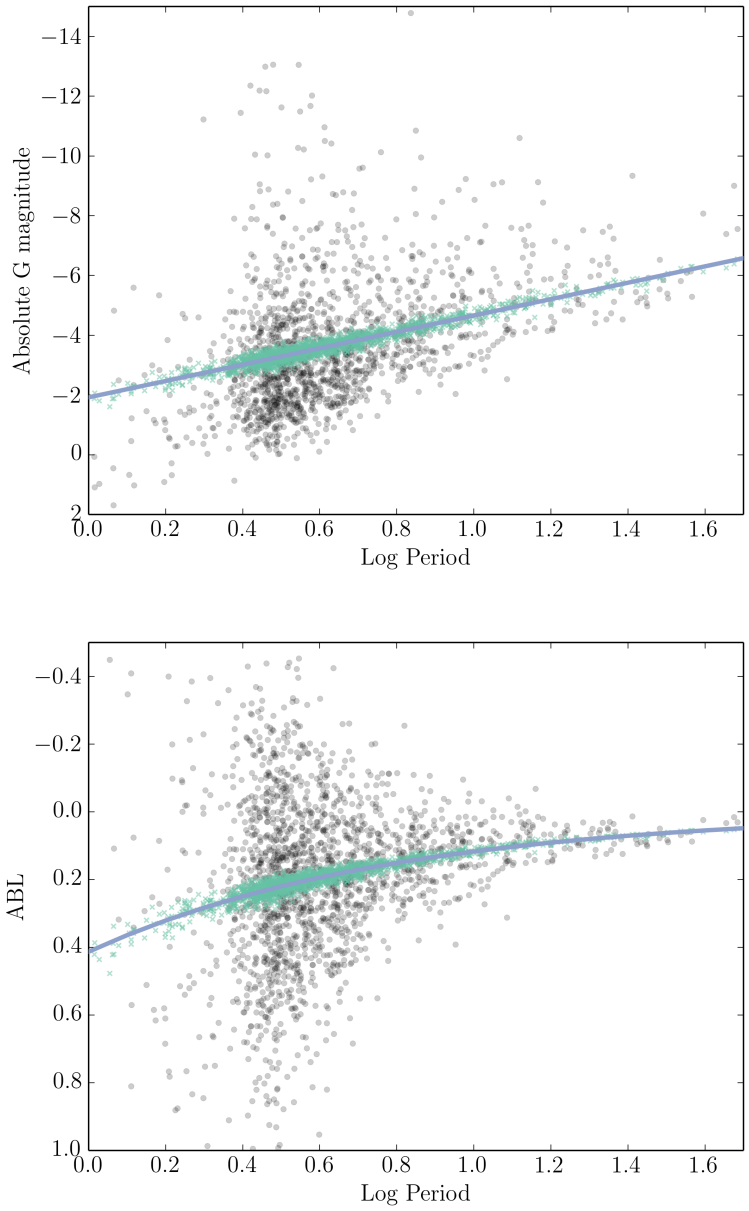


Figure 5.12: Top*: Observational PL relation with the absolute magnitude calculated from the observed parallax and apparent magnitude. The 316 stars with negative parallax have not been included.* Bottom*: The relationship between the period and the Astrometric Based Luminosity calculated for all of the observed Cepheids. In both plots, the black circles are the observed data and the blue crosses are the 'true' values. The fit line is that obtained using the method in Sect. 5.5.3.*

for precise calibration of the PL relation when Gaia data becomes available. With an expected number of Cepheids in the thousands, the method provided here has the potential to provide a precise and unbiased calibration of both the slope and zero point of the absolute PL relation. With no statistically significant sample of parallax measurements available, a method was developed for calibration for use without parallax data. This was applied to real data for RR-Lyraes in the LMC, leading to a new PLZ relation.

With the potential for Gaia to obtain absolute trigonometric parallax measurements for stars in the Magellanic Clouds, the ML method introduced in Ch. 4 has been adapted for obtaining the distance to the LMC. Testing with simulated data shows that Gaia will be capable of obtaining the global mean distance to the LMC to within 0.4%, though there is insufficient information to constrain specific details about the depth or structure. While information could theoretically be obtained on the orientation and angle of the tilt of the disk of the LMC, error correlation effects will likely make detection impossible (see Ch. 7).

The LMC contains many variables and the sky position of the LMC away from the disk makes them an ideal candedate for calibration of relations which can be used out to the extragalactic distance scale. To test the capability of Gaia in this calibration, the ML method is applied to synthetic catalogues of Cepheid and RR-Lyrae stars. Cepheids, being bright, will allow a reasonable distance determination (5%), however we find that it is beneficial to estimate the mean distance to the LMC using the full quantity of available stars. This allows an excellent precision on the mean distance to the LMC which can then be used to calibrate the absolute PL relation.

6

Open clusters

6.1 Introduction

Open clusters have long been used as a testing ground for a large number of astronomical theories. Determining the distances to nearby open clusters is critical, since they have historically formed the first step in the calibration of the distance scale. Because all stars within an open cluster are expected to have the same age and metallicity, accurate distance estimates are highly useful in calibrating the main sequence and checking stellar evolutionary theory through comparison with theoretical isochrones.

Until recently, accurate distances to even the most nearby clusters have not been possible. The Hipparcos astrometric mission of 1989 (Perryman & ESA, 1997) for the first time gave accurate parallax measurements for over one hundred thousand stars and has been used extensively to give direct distance

measurements to more than 30 open clusters.

Still, many questions remain. While revolutionary in its time, the milliarcsecond astrometry and limiting magnitude (H_p<12.5) of Hipparcos allow calculating distances to only the nearest open clusters, and even then do so with a precision no better than a few percent. Owing to the relatively small size of most open clusters compared with their distances and the precision of measurements, the Hipparcos data has been unable to give definitive answers about the internal structure and physical size of such clusters.

Additionally, the release of the Hipparcos catalogue in 1997 has led to some controversy. The most famous is the case of the Pleiades, where methods based on Hipparcos data (van Leeuwen & Hansen Ruiz, 1997; Robichon et al., 1999a; Mermilliod et al., 1997) gave a distance estimate that was some 10% shorter than works based on photometric methods (Pinsonneault et al., 1998; Robichon et al., 1999b; Stello & Nissen, 2001) (see Sect. 6.8). Even more precise results from Melis et al. (2014) are contra Hipparcos based distance estimates.

With the above in mind, it is apparent that a new method is required that is capable of squeezing the maximum precision from the currently available data and is capable of utilising information from the full range of observed quantities (astrometric, photometric, and kinematic information). This will be particularly true after the launch of Gaia, which will produce a rich dataset that will include not only accurate parallax measurements, but also photometry at millimag precision and a full set of kinematics obtained from proper motions combined with radial velocity measurements from the on-board radial velocity spectrometer (for stars with G_{RVS}<16).

Maximum Likelihood Estimation (MLE) has been used in relation to open clusters since 1958, where Vasilevskis et al. (1958) used MLE to perform cluster membership selection from proper motions.

For the Hyades and Pleiades, Chen and Zhao used a combination of the convergent point method with MLE in order to simultaneously determine mean parallax and kinematics for the Hyades (Zhao & Chen, 1994) and later the

Pleiades (Li & Junliang, 1999). Here, we continue the formulation introduced in Sect. 4.3.

In Sect. 6.2 the exact mathematical formulation is given. A description of the observational data used is given in Sect. 6.6, and the results of application of the method to the Pleiades and Hyades given in Sects. 6.7 and 6.9. The possible effects of correlated errors in the Hipparcos catalogue are discussed in Sect. 6.8. The use of the method with Gaia data is tested using simulations in Sect. 6.10. The contents of this chapter has been published as **Palmer et al. (2014)**, **Astronomy & Astrophysics, Volume 564, 2014, id: A49, 14 pp.**

6.2 Mathematical formulation

Definition

The Likelihood function can be defined as

$$\mathcal{L}(\boldsymbol{\theta}) = \prod_{i=1}^{n} \mathcal{P}(\boldsymbol{x}_i|\boldsymbol{\theta}) \tag{6.1}$$

where the joint PDF $\mathcal{P}(\boldsymbol{x}_i|\boldsymbol{\theta})$ is made up of the un-normalised PDF $\mathcal{D}(\boldsymbol{x}_i|\boldsymbol{\theta})$ and a normalisation constant $\mathcal{C}_i$, such that

$$\mathcal{P}(\boldsymbol{x}_i|\boldsymbol{\theta}) = \mathcal{C}_i^{-1}\mathcal{D}(\boldsymbol{x}_i|\boldsymbol{\theta}), \tag{6.2}$$

and $\boldsymbol{\theta}$ is the vector giving the parameters of the model.

To model every major aspect of an open cluster, we have chosen that $\boldsymbol{\theta} = (R, \sigma_R, M(V-I), \sigma_M, U, V, W, \sigma_{UVW})$, where R is the distance to the centre of the cluster assuming a spherical Gaussian distribution, σ_R the intrinsic dispersion around the mean distance, $M(V-I)$ the mean absolute magnitude as a function of colour, σ_M the intrinsic dispersion around the mean absolute magnitude, U, V, and W are the three components of the clusters velocity ellipsoid in galactic Cartesian coordinates, and σ_{UVW} is the intrinsic dispersion

in the clusters velocity.

The vector ($\boldsymbol{x} = m, l, b, \varpi, \mu_l, \mu_\delta, v_r$) describes the observed properties of each star, and ($\boldsymbol{x}_0 = m_0, l_0, b_0, r_0, \mu_{\alpha^*,0}, \mu_{\delta 0}, v_{r0}$) is the vector describing the 'true' underlying stellar properties unaffected by observational errors.

We can then define the un-normalised[1] PDF such that

$$\mathcal{D}(\boldsymbol{x}_i|\boldsymbol{\theta}) = \mathcal{S}(\boldsymbol{x}_i) \int_{\forall \boldsymbol{x}_0} \varphi_{M_0} \varphi_{rlb_0} \varphi_{v_0} \mathcal{E}(\boldsymbol{x}_i|\boldsymbol{x}_0)\, d\boldsymbol{x}_0, \tag{6.3}$$

where φ_{M_0}, φ_{rlb}, φ_v are the PDFs describing the nature of the open cluster, and $\mathcal{S}(\boldsymbol{x})$ is the selection function, which takes the probability of observing a star into account, given the properties of the star and the instruments' observational capabilities. To take the fact that Hipparcos is a magnitude limited sample into account, a step function is used with

$$\mathcal{S}(\boldsymbol{x}) = \begin{cases} 1, & \text{if } H_p{<}12.5. \\ 0, & \text{otherwise.} \end{cases} \tag{6.4}$$

The case is more complicated in reality , with Hipparcos only being complete up to magnitude $H_p < 7$. At fainter magnitudes, stars were selected using a number of criteria and used as an input catalogue (Turon et al., 1992). Hipparcos had a physical limit in the number of stars observable in a single field of view, and it suffered from glare effects on the telescope when observing stars very close together on the sky. Therefore, decisions were made on a case-by-case basis as to which stars to observe, depending on the number and position of stars in each field of view. The Hipparcos Mission Pre Launch Status Report states that: "*Once the list of likely cluster members had been established and the worst veiling glare cases excluded, a somewhat arbitrary compromise had to be found*". The hand selection of stars for observation makes it impossible

[1]Although the PDF must be normalised, it is convenient to define the un-normalised PDF first, and then come back to the normalisation constant when all of the components of the PDF have been defined.

to accurately model a selection function based on apparent magnitude that can describe the selection probability for the underlying cluster population, which does not strictly follow the definition given in Eq. 6.4. However, as the Hipparcos star selection mostly depended on the proximity of target stars to each other on the sky and not on magnitude, the effects on the results given in Sects. 6.7 and 6.9 are minimal.

6.2.1 Models

The distribution of absolute magnitudes in a given photometric band is assumed to be a Gaussian distribution around a mean colour-absolute magnitude relation, $M_{\rm mean}$, with some dispersion, σ_M:

$$\varphi_M = e^{-0.5\left(\frac{M-M_{\rm mean}(V-I)}{\sigma_M}\right)^2}. \tag{6.5}$$

Here, the absolute magnitude is given by

$$M = m + 5\log_{10}(\varpi) + 5 - A, \tag{6.6}$$

where m is the apparent magnitude of the star, ϖ its parallax, and A is the interstellar extinction of the star in the same photometric band.

In the following work, the interstellar extinction is assumed to be known, although it can be left as a free fit parameter. For the Pleiades, a single extinction value of A_{H_p}=0.0975 magnitudes in the Hipparcos band H is used for all members. This is derived from a reddening of $E(B-V) = 0.025 \pm 0.003$ found by Groenewegen et al. (2007), which is converted to an extinction estimate in the visual band through $A_V = 3.1E(B-V)$ and then into the Hipparcos band H_p. For the Hyades, the level of extinction is assumed to be negligible, and an extinction of A_{H_p}=0 is assumed. If available, individual extinction estimates for each star could be used, in order to correctly take effects of differential reddening into account.

The term $M_{\rm mean}(V-I)$ gives the mean absolute magnitude of stars as a

function of colour, while σ_M is the intrinsic dispersion. For simplicity, known binaries are removed. Additionally, the given distribution does not support the giant branch, since this would require the magnitude function to turn back on itself, giving a non-unique solution. Therefore, giants are also removed, enabling the application of this method to all clusters, irrespective of age.

The fitting procedure has two options:

1. It fits the position of points on the colour-absolute magnitude diagram, to which a spline function is fitted in order to determine a colour-dependent absolute magnitude distribution approximating the isochrone of the cluster; or

2. If a theoretical isochrone is supplied as input, the shape of the magnitude distribution is taken from the isochrone. While the shape of the isochrone is conserved, the isochrone is free to be shifted in absolute magnitude (equivalent to a shift in distance).

If using the second option, the age and metallicity of the cluster must be assumed. This makes the first option preferable for clusters with little information on age and metallicity, because parameters can be determined without having to be concerned with models that depend on this information.

6.3 Cartesian to galactic coordinate transformation

The spatial distribution of the cluster is given in Cartesian coordinates by a Gaussian distribution in each axis:

$$\varphi_x = e^{-0.5\left(\frac{x-X_0}{\sigma_S}\right)^2} \tag{6.7}$$

$$\varphi_x = e^{-0.5\left(\frac{y-Y_0}{\sigma_S}\right)^2} \tag{6.8}$$

$$\varphi_z = e^{-0.5\left(\frac{z-Z_0}{\sigma_S}\right)^2}, \tag{6.9}$$

where X_0, Y_0, and Z_0 define the centre of the cluster in each axis, and σ_S gives the variance of the distribution.

To transform this Cartesian PDF into polar coordinates as required for our observables r, l, and b, we start with the relationship between our two sets of variables and find the inverse:

$$r = \sqrt{x^2 + y^2 + z^2} \Longrightarrow x = \sqrt{r^2 - y^2 - z^2} \tag{6.10}$$

$$l = \tan^{-1}\left(\frac{y}{x}\right) \Longrightarrow y = \tan(l)x \tag{6.11}$$

$$b = \sin^{-1}\left(\frac{z}{r}\right) \Longrightarrow \boxed{z = r\sin(b)}. \tag{6.12}$$

Then we have two equations and two unknowns:

$$x^2 = r^2 - y^2 - z^2 = r^2 - \tan^2(l)x^2 - r^2\sin^2(b) \tag{6.13}$$

$$x^2(1 + \tan^2(l)) = r^2 - r^2\sin^2(b) \tag{6.14}$$

$$x^2 = \frac{r^2 - r^2\sin^2(b)}{1 + \tan^2(l)} \tag{6.15}$$

$$\boxed{x = r\cos(b)\cos(l)} \tag{6.16}$$

$$\boxed{y = r\cos(b)\sin(l)}. \tag{6.17}$$

Then we require the Jacobian: $r^2\cos(b)$.

By substituting the x, y, and z found above into the original PDF and multiplying by the Jacobian of the transformation, we find the PDF in the

new coordinate system:

$$\varphi_{rlb} = r^2\cos(b)\, e^{-\frac{0.5}{\sigma_S^2}\left((r\cos(b)\cos(l)-X_0)^2+(r\cos(b)\sin(l)-Y_0)^2+(r\sin(b)-Z_0)^2\right)}. \tag{6.18}$$

By rotating our coordinate system $l, b \rightarrow l', b'$ to align the cluster centre with the X axis, we have new coordinates l' and b' for all the stars. In this rotated coordinate system, the cluster has a position $Y_0' = Z_0' = 0$, and X' is equivalent to the distance to the clusters centre. The above spatial probability distribution function can then be simplified as

$$\varphi_{r'l'b'} = r^2\cos(b')e^{-\frac{0.5}{\sigma_S^2}\left(R^2+r^2-2rR\cos(b')\cos(l')\right)} \tag{6.19}$$

where the term $r^2\cos(b')$ is the Jacobian of the coordinate transformation, R the mean distance to the cluster, r the distance to the individual star, and l' and b' are the rotated coordinates of the star.

It should be noted that the two coordinate systems are used simultaneously. The rotated coordinates (l', b') are used in the integration over position to simplify the integrals as explained above. However, in the analytic solution to the integrals over $\mu_{\alpha^*}\mu_\delta v_r$, the unrotated coordinates l and b are used.

The Gaussian spatial distribution was chosen for ease of implementation, and as a first approximation of the cluster's spatial structure. It is possible to use a distribution specifically suited to open clusters, such as King's profile (King, 1962), which could make the distribution more realistic. This is a possible improvement for later additions to this work.

Finally, the velocity distribution in Cartesian coordinates is defined as the velocity ellipsoid:

$$\varphi_v = e^{-0.5\left(\frac{U-U_{\text{mean}}}{\sigma_U}\right)^2-0.5\left(\frac{V-V_{\text{mean}}}{\sigma_V}\right)^2-0.5\left(\frac{W-W_{\text{mean}}}{\sigma_W}\right)^2}, \tag{6.20}$$

where $U_{\rm mean}$, $V_{\rm mean}$, and $W_{\rm mean}$ are the components of the cluster's mean velocity along each Cartesian axis.

The distribution in observational errors is given by

$$\mathcal{E}(\boldsymbol{x}|\boldsymbol{x}_0) = \mathcal{E}(\varpi|\varpi_o)\mathcal{E}(\mu_{\alpha^*}|\mu_{\alpha^*,0})\mathcal{E}(\mu_\delta|\mu_{\delta,0})\mathcal{E}(v_r|v_{r_0})\delta(m, l', b'). \tag{6.21}$$

All observational errors are assumed to follow a Gaussian distribution with a variance given by the formal error $\epsilon_{\boldsymbol{x}}$. Here, ϖ is the parallax, μ_{α^*} and μ_δ the proper motions, and m the apparent magnitude. The delta function $\delta(m, l', b')$ describes the case for which a parameter's observational error is small enough to be deemed negligible.

An additional benefit of using this formulation with 'true' object parameters in the models and then linking these true parameters with the observed quantities is that observational data including negative parallaxes can be used directly without problems attempting to calculate the logarithm of a negative number (e.g. in Eq. 6.6). The inclusion of stars with negative parallaxes is essential for avoiding biasing the sample by preferentially removing more distant stars and biasing the average distance by selecting the stars with only positive errors.

6.4 Integration of the Likelihood function

To evaluate $\mathcal{D}(\boldsymbol{x}|\boldsymbol{\theta})$ in Eq. 6.1 we must integrate over all $\boldsymbol{x}_0$, giving a multiple integral that can be split into three parts. First is the integral over variables with assumed zero error, second kinematics, and finally distance.

6.4.1 Integration over m_0, l'_0 and b'_0

As these variables have errors given by the delta function, $(m, l', b') = (m_0, l'_0, b'_0)$ so we can use (m, l', b'). This avoids integrating over these three parameters.

6.4.2 Integration over $\mu_{\alpha^*,0}$, $\mu_{\delta,0}$ and v_{r0}

The triple integral over $\mu_{\alpha^*,0}$, $\mu_{\delta,0}$ and v_{r0} is

$$\int_{\forall \mu_{\alpha^*,0}\mu_{\delta,0}v_{r0}} \varphi_v(U,V,W)\mathcal{E}(z|z_0)\, d\mu_{\alpha^*,0}\, d\mu_{\delta,0}\, dv_{r0} \tag{6.22}$$

where

$$\mathcal{E}(z|z_0) = e^{-0.5\left(\frac{\mu_{\alpha^*}-\mu_{\alpha^*,0}}{\epsilon_{\mu_{\alpha^*}}}\right)^2} e^{-0.5\left(\frac{\mu_{\delta}-\mu_{\delta,0}}{\epsilon_{\mu_{\delta}}}\right)^2} e^{-0.5\left(\frac{v_r-v_{r0}}{\epsilon_{v_r}}\right)^2}. \tag{6.23}$$

In order to perform the integral, the function $\varphi_v(U,V,W)$ must be expressed in terms of $\mu_{\alpha^*,0}$, $\mu_{\delta,0}$, and v_{r0}. This is achieved through the following expressions:

$$\begin{aligned} U &= a_1\mu_{\alpha^*}r + b_1\mu_{\delta}r + c_1 v_r \\ a_1 &= -k\cos(b)\sin(l) \\ b_1 &= -k\sin(b)\cos(l) \\ c_1 &= \cos(b)\cos(l) \end{aligned} \tag{6.24}$$

$$\begin{aligned} V &= a_2\mu_{\alpha^*}r + b_2\mu_{\delta}r + c_2 v_r \\ a_2 &= k\cos(b)\cos(l) \\ b_2 &= -k\sin(b)\sin(l) \\ c_2 &= \cos(b)\sin(l) \end{aligned} \tag{6.25}$$

$$
\begin{aligned}
W = a_3\mu_{\alpha^*}r + b_3\mu_\delta r + c_3 v_r \\
a_3 = 0 \\
b_3 = k\cos(b) \\
c_3 = \sin(b)
\end{aligned}
\tag{6.26}
$$

where $k = 4.74\frac{Km\,year}{"\,s\,pc}$.

Therefore $\varphi_v(U,V,W)$ can be written in terms of $\mu_{\alpha^*,0}$, $\mu_{\delta,0}$ and v_{r0} as

$$
\varphi_v(U,V,W) = e^{p(\mu_{\alpha^*,0},\mu_{\delta,0},v_{r0}|r)}. \tag{6.27}
$$

This can be integrated using the definite integral,

$$
\int_{-\infty}^{\infty} e^{-(\alpha x^2+\beta x+\gamma)}dx = \sqrt{\frac{\pi}{\alpha}}\, e^{\left(\frac{\beta^2}{4\alpha}-\gamma\right)}, \tag{6.28}
$$

giving the solution

$$
\int_{\forall\mu_{\alpha^*,0}\mu_{\delta,0}v_{r0}} \varphi_v(U,V,W)\mathcal{E}(\boldsymbol{x}|\boldsymbol{x}_0)\, d\mu_{\alpha^*,0}\, d\mu_{\delta,0}\, dv_{r0} = K\, e^{\left(\frac{Y^2}{4X}-Z\right)} \tag{6.29}
$$

where K, X, Y, and Z are defined as:

$$
K = \sqrt{-\frac{\pi^3}{J_2 E_1 X}}
$$

$$
\begin{aligned}
X &= \frac{D_1^2}{4E_1} - F_1 \\
Y &= \frac{B_1 D_1}{2E_1} - C_1 \\
Z &= \frac{B_1^2}{4E_1} - A_1
\end{aligned}
$$

$$
\begin{aligned}
A_1 &= \frac{A_2^2}{4J_2} - D_2 \\
B_1 &= \frac{A_2 B_2}{2J_2} - E_2 \\
C_1 &= \frac{A_2 C_2}{2J_2} - F_2 \\
D_1 &= \frac{B_2 C_2}{2J_2} - G_2 \\
E_1 &= \frac{B_2^2}{4J_2} - H_2 \\
F_1 &= \frac{C_2^2}{4J_2} - I_2
\end{aligned}
$$

$$
\begin{aligned}
A_2 &= -\frac{\mu_{li}}{\epsilon^2_{\mu_{\alpha^*}}} - \left(\frac{a_1 U_0}{\sigma_U^2} + \frac{a_2 V_0}{\sigma_V^2} + \frac{a_3 W_0}{\sigma_W^2} \right) r \\
B_2 &= \left(\frac{a_1 b_1}{\sigma_U^2} + \frac{a_2 b_2}{\sigma_V^2} + \frac{a_3 b_3}{\sigma_W^2} \right) r^2 \\
C_2 &= \left(\frac{a_1 c_1}{\sigma_U^2} + \frac{a_2 c_2}{\sigma_V^2} + \frac{a_3 c_3}{\sigma_W^2} \right) r
\end{aligned}
$$

$$
\begin{aligned}
D_2 &= \frac{1}{2} \left(\frac{\mu_{li}^2}{\epsilon^2_{\mu_{\alpha^*}}} + \frac{\mu_{bi}^2}{\epsilon^2_{\mu_\delta}} + \frac{v_{ri}^2}{\epsilon^2_{v_r}} + \frac{U_0^2}{\sigma_U^2} + \frac{V_0^2}{\sigma_V^2} + \frac{W_0^2}{\sigma_W^2} \right) \\
E_2 &= -\frac{\mu_{bi}}{\epsilon^2_{\mu_\delta}} - \left(\frac{b_1 U_0}{\sigma_U^2} + \frac{b_2 V_0}{\sigma_V^2} + \frac{b_3 W_0}{\sigma_W^2} \right) r \\
F_2 &= -\frac{v_{ri}}{\epsilon^2_{v_r}} - \left(\frac{c_1 U_0}{\sigma_U^2} + \frac{c_2 V_0}{\sigma_V^2} + \frac{c_3 W_0}{\sigma_W^2} \right)
\end{aligned}
$$

$$
\begin{aligned}
G_2 &= \left(\frac{b_1 c_1}{\sigma_U^2} + \frac{b_2 c_2}{\sigma_V^2} + \frac{b_3 c_3}{\sigma_W^2}\right) r \\
H_2 &= \frac{1}{2}\frac{1}{\epsilon_{\mu_\delta}^2} + \frac{1}{2}\left(\frac{b_1^2}{\sigma_U^2} + \frac{b_2^2}{\sigma_V^2} + \frac{b_3^2}{\sigma_W^2}\right) r^2 \\
I_2 &= \frac{1}{2}\frac{1}{\epsilon_{v_r}^2} + \frac{1}{2}\left(\frac{c_1^2}{\sigma_U^2} + \frac{c_2^2}{\sigma_V^2} + \frac{c_3^2}{\sigma_W^2}\right) \\
J_2 &= \frac{1}{2}\frac{1}{\epsilon_{\mu_{\alpha*}}^2} + \frac{1}{2}\left(\frac{a_1^2}{\sigma_U^2} + \frac{a_2^2}{\sigma_V^2} + \frac{a_3^2}{\sigma_W^2}\right) r^2
\end{aligned} \tag{6.30}
$$

6.4.3 Integration over R

The remaining integral has no analytical solution and will therefore be performed numerically:

$$
\mathcal{D}(\boldsymbol{x}|\boldsymbol{\theta}) = \int_r \varphi_{M_0}\varphi_{\varpi 0} l_0' b_0' K \; e^{\left(\frac{Y^2}{4X} - Z\right)} e^{-0.5\left(\frac{\varpi - \varpi_0}{\epsilon_\varpi}\right)^2} \, dr_0 \tag{6.31}
$$

6.5 Normalisation coefficient

Until now we have been using the un-normalised joint probability distribution. Normalisation is achieved by dividing by a normalisation constant, $\mathcal{C}$. The normalisation constant is found by integrating the un-normalised joint probability distribution $\mathcal{D}(\boldsymbol{x}|\boldsymbol{\theta})$ over all $\boldsymbol{x}$:

$$
\mathcal{C} = \int_{\forall \boldsymbol{x}_0} \int_{\forall \boldsymbol{x}} \varphi_{M_0}\varphi_{\varpi 0} l_0' b_0' \varphi_{v0}(U,V,W) \mathcal{S}(\boldsymbol{x}) \mathcal{E}(\boldsymbol{x}|\boldsymbol{x}_0) \, d\boldsymbol{x} \, d\boldsymbol{x}_0. \tag{6.32}
$$

This integral can be performed in two parts, where I is defined such that

$$
\mathcal{C} = \int_{\forall \boldsymbol{x}_0} \varphi_{M_0}\varphi_{\varpi 0} l_0' b_0' \varphi_{v0}(U,V,W) \underbrace{\int_{\forall \boldsymbol{x}} \mathcal{S}(\boldsymbol{x}) \mathcal{E}(\boldsymbol{x}|\boldsymbol{x}_0) \, d\boldsymbol{x}}_{\mathrm{I}} \, d\boldsymbol{x}_0. \tag{6.33}
$$

Substituting in the selection function $\mathcal{S}(\boldsymbol{x})$ and the PDF of the observational

errors $\mathcal{E}(\boldsymbol{x}|\boldsymbol{x}_0)$ gives the following seven-dimensional integral:

$$I = \int_{\forall \boldsymbol{x}} \boldsymbol{\theta}(m - m_{lim}) \mathcal{E}(\boldsymbol{x}|\boldsymbol{x}_0)\, d\boldsymbol{x}. \tag{6.34}$$

This integral can be split into two parts. The integral over the delta function in $\mathcal{E}(\boldsymbol{x}|\boldsymbol{x}_0)$ that, by definition, gives one; and the integral over each Gaussian error,

$$\begin{aligned} I = & \\ & \boldsymbol{\theta}(m - m_{lim}) \\ & \int_{\forall \varpi \mu_{\alpha^*} \mu_\delta v_r} e^{-0.5\left(\frac{\varpi - \varpi_0}{\epsilon_\varpi}\right)^2} e^{-0.5\left(\frac{\mu_{\alpha^*} - \mu_{\alpha^*,0}}{\epsilon_{\mu_{\alpha^*}}}\right)^2} e^{-0.5\left(\frac{\mu_\delta - \mu_{\delta,0}}{\epsilon_{\mu_\delta}}\right)^2} \\ & e^{-0.5\left(\frac{v_r - v_{r0}}{\epsilon_{v_r}}\right)^2} d\varpi\, d\mu_{\alpha^*}\, d\mu_\delta\, dv_r \int_{-\infty}^{\infty} \delta(m, l', b') dm\, dl'\, db' \end{aligned} \tag{6.35}$$

$$I = \boldsymbol{\theta}(m - m_{lim}) (2\pi)^2 \epsilon_\varpi \epsilon_{\mu_{\alpha^*}} \epsilon_{\mu_\delta} \epsilon_{v_r}. \tag{6.36}$$

Here, $\boldsymbol{\theta}(m - m_{lim})$ acts to provide an upper limit to the integral over all M. Substituting I back into C we have

$$\begin{aligned} \mathcal{C} = & \\ & (2\pi)^2 \epsilon_\varpi \epsilon_{\mu_{\alpha^*}} \epsilon_{\mu_\delta} \epsilon_{v_r} \\ & \int_{-\infty}^{m_{\rm lim}} \int_0^{\infty} \int_{-\pi/2}^{\pi/2} \int_0^{2\pi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \varphi_{m0} \varphi_{r_0 l'_0 b'_0} \\ & \varphi_v(U, V, W)\, dU\, dV\, dW\, dl'_0\, db'_0\, dr_0\, dm_0 \end{aligned}. \tag{6.37}$$

As with in the previous section, the integral can be split up into a number of parts.

6.5.1 Integration over m

Evaluating first the integral over apparent magnitude gives

$$\int_{-\infty}^{m_{\rm lim}} \varphi_{m0}\, dm_0 = \frac{\sqrt{2\pi}}{2}\sigma_M \mathrm{erfc}\left(\frac{A - m_{\rm lim}}{\sqrt{2}\sigma_M}\right) \tag{6.38}$$

where erfc is the complementary error function, and

$$A = 5\log(r_0) - 5 + M_{\rm mean}. \tag{6.39}$$

6.5.2 Integration over (U, V, W)

Integrating over (U,V,W) gives

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \varphi_v(U,V,W)\, dU\, dV\, dW = (2\pi)^{3/2}\sigma_U\sigma_V\sigma_W. \tag{6.40}$$

6.5.3 Integration over l'_0, b'_0, and r_0

The remaining triple integral has no analytical solution and will be performed numerically:

$$\mathcal{C} = B\int_0^{\infty}\int_{-\pi/2}^{\pi/2}\int_0^{2\pi} \mathrm{erfc}\left(\frac{A - m_{\rm lim}}{\sqrt{2}\sigma_M}\right)\varphi_{r_0 l'_0 b'_0}\, dl'_0\, db'_0\, dr_0 \tag{6.41}$$

with: $B = \frac{(2\pi)^4}{2}\sigma_U\sigma_V\sigma_W\sigma_M\epsilon_{\varpi}\epsilon_{\mu_{\alpha^*}}\epsilon_{\mu_\delta}\epsilon_{v_r}$

6.5.4 Formal errors

The Hessian matrix is constructed through numerical differentiation of the Likelihood function at its global maximum. The inverse of the Hessian matrix is the covariance matrix, and the formal errors are calculated from the square root of the diagonal of the covariance matrix. After calculating the covariance

matrix and formal errors, correlations between each of the parameters can easily be obtained.

6.5.5 Data binning

While some parameters, such as mean distance, are 'global' parameters describing some general property of the open cluster, a number of the parameters show a dependence on colour. For example, the physical spatial distribution of some clusters is believed to change as a function of mass, and therefore colour, through the process of mass segregation.

With a sufficiently precise data set containing enough stars, it is possible to fit a smooth function describing a parameter's colour dependence, if applicable, by estimating the parameters of, for example, a polynomial approximation of the dependence. Owing to the limited available data in the Hipparcos catalogue for the Pleiades and Hyades, there is insufficient information to constrain such distributions in all cases, so where necessary an approximation has been made by binning the data.

A star's normalised PDF can be thought of as the sum of several other PDFs, such that

$$\frac{\mathcal{D}(\boldsymbol{x}_i|\boldsymbol{\theta})}{\mathcal{C}} = \frac{\mathcal{W}(V-I)_1\mathcal{D}(\boldsymbol{x}_i|\boldsymbol{\theta}) + \mathcal{W}(V-I)_2\mathcal{D}(\boldsymbol{x}_i|\boldsymbol{\theta}) + ...}{\mathcal{C}} \tag{6.42}$$

where $\mathcal{C}$ is the normalisation constant and $\mathcal{W}$(V-I) a selection function dependent on colour:

$$\mathcal{W} = \begin{cases} w, & \text{if } (V-I)_{min} < (V-I)_{star} < (V-I)_{max}. \\ 0, & \text{otherwise,} \end{cases} \tag{6.43}$$

where w is a normalised coefficient that depends on the relative number of stars per colour bin.

The normalisation constant C is found by integrating the PDF over all $\boldsymbol{x}$:

$$\mathcal{C} = \int_{\forall \boldsymbol{x}} \mathcal{D}(\boldsymbol{x}|\boldsymbol{\theta})\ d\boldsymbol{x} \tag{6.44}$$

$$\equiv \int_{\forall \boldsymbol{x}} \mathcal{W}(V-I)_1 \mathcal{D}(\boldsymbol{x}|\boldsymbol{\theta}) d\boldsymbol{x} + \int_{\forall \boldsymbol{x}} \mathcal{W}(V-I)_2 \mathcal{D}(\boldsymbol{x}|\boldsymbol{\theta}) d\boldsymbol{x} + ... \tag{6.45}$$

The strength of this approach is that parameters with a single value of interest (with no colour dependence) are described by only one parameter in $\boldsymbol{\theta}$. Where more information can be gained by producing an estimate of a parameter in each colour bin, it is possible to define a separate parameter for each colour bin.

The parameters to be estimated therefore become

$$\boldsymbol{\theta} = (R, \sigma_R^{(n)}, M^{(n+1)}, \sigma_M^{(n)}, U, V, W, \sigma_{UVW}), \tag{6.46}$$

where n is the number of bins, R the mean distance to the cluster (assuming a spherical Gaussian distribution), σ_R^n the intrinsic dispersion around the mean distance in each colour bin, M^{n+1} are the absolute magnitude values used for fitting the spline function of the isochrone, U, V, and W are the three components of the velocity ellipsoid, and σ_{UVW} is the velocity dispersion.

6.5.6 Testing with simulations

The method described in Sect. 6.2 was implemented and tested extensively using simulations. During development and testing of the MLE implementation, simulated catalogues were constructed using Monte Carlo techniques. Each star in the simulated population is given a position in Cartesian coordinates, which is then converted into a sky position and distance, and given a velocity in Cartesian coordinates, which is converted into proper motions and a radial velocity. The values are randomly chosen from a spherical Gaussian spatial distribution and a velocity ellipsoid, with the mean and variance of each distribution chosen by hand. The relationship between colour and

absolute magnitude is assumed to be linear for simplicity.

Basic simulation of observational errors was achieved by adding an error value derived from a Gaussian random number generator to each parameter in the simulated catalogue, with the variance chosen to be a value similar to the errors found in the Hipparcos catalogue. Additionally, the more realistic AMUSE open cluster simulator (Pelupessy et al., 2013) was used to simulate open clusters with more realistic distributions, including a realistic isochone.

The final stage of testing with simulations uses a set of 500 simulated open clusters found in the Gaia Simulator (Masana et al., 2010; Robin et al., 2012)) in order to test the suitability of the method in use with Gaia data (see Sect. 6.10). The Gaia Object Generator (Isasi et al., 2010) was used to simulate realistic errors as expected for Gaia catalogue data. Using the ML method on this simulated data showed the successful extraction of an isochrone-like sequence from error-effected data (Fig. 6.1) and could reproduce the input distance of the cluster, within formal error bounds and, after repeated testing, with no significant bias.

In the following sections (6.6, 6.7, 6.8, and 6.9), we apply the method to the best currently available data on the Pleiades and the Hyades. In Sect. 6.10 we use simulations to extrapolate the use of the method out to greater distances, in expectation of the Gaia data.

6.6 Data

The method described here has been applied to the new Hipparcos reduction (van Leeuwen, 2007) data for 54 well known Pleiades members. The 54 cluster members are believed to be a clean sample and have been identified and used in numerous previous studies (e.g. van Leeuwen & Hansen Ruiz (1997); Robichon et al. (1999a); Makarov (2002)).

The new Hipparcos reduction has several advantages over the original Hipparcos catalogue, including a reduction in the formal errors by up to a factor of 4 for the brightest stars and a claimed reduction, by up to a factor of 10, in

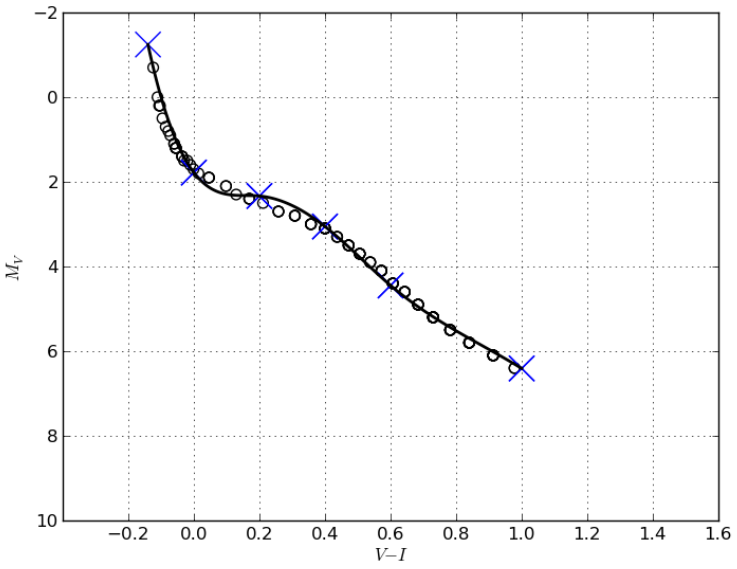


Figure 6.1: CMD for a simulated Pleiades like cluster found in the GaiaSimu library, with a distance of 2.6 kpc. Open circles show the 'true' simulated absolute magnitudes without errors. Blue crosses are the points fitted with the ML method applied to the data after the simulation of observational errors. The solid line is the resulting isochrone-like sequence, showing accurate reconstruction of the isochrone-like sequence from error effected data.

the correlations between stars observed over small angles. This was achieved through a completely new method for formulating the Hipparcos satellite's attitude model, replacing the previous 'great-circles' reduction process with a fully iterative global solution. This new method of catalogue reconstruction supersedes an earlier attempt by Makarov (2002) to reduce correlation in the Hipparcos catalogue specifically for the Pleiades open cluster.

This reduction in correlations is particularly important in the study of open clusters. This is because correlated errors in the original Hipparcos catalogue were blamed for the cases of discrepancy between distances calculated with Hipparcos parallaxes, compared with photometric data or other methods.

Where radial velocity information has been used, it was obtained from Mermilliod et al. (2009); Raboud & Mermilliod (1998); Morse et al. (1991); Liu et al. (1991) through the WEBDA database.

6.7 Results - Pleiades

The results of the application of the method described in Sect. 6.2 can be seen in Tables 6.1 and 6.3. Parameters describing the general properties of the cluster, such as mean distance and mean proper motion are given in Table 6.1, whereas colour-dependent parameters are given in Table 6.3. For the latter, stars were divided into four color bins. The choice of bins is arbitrary and have been selected to give roughly the same number of stars per bin.

6.7.1 Distance

The mean distance to the Pleiades has been estimated to be 120.3 ± 1.5 pc. This agrees with van Leeuwen (2009), who finds 120.2 ± 2.0 pc using the same dataset. The formal errors assigned to each parameter are calculated from the square root of the covariance matrix of the Likelihood function. The formal error is approximately 25% smaller than from van Leeuwen (2009), with the increased precision from the use of this method attributed to the

inclusion of parallax, position, proper motion, colour, and apparent magnitude information.

The distance is given as a general property of the cluster alone (Table 6.1), because there is no physical reason to expect differing distances for different colour bins, unlike, for example, the size parameter, σ_R.

The mean distance to the cluster has been included as a colour-dependent parameter during testing, in order to check the ML method functions as expected and to check that there are no biases present in the Hipparcos catalogue. The mean distance to the cluster was found to be consistent across all colour bins, within the error bounds. This is a good test that there is no colour-dependent bias in the Hipparcos Pleiades data.

6.7.2 Kinematics

Testing was carried out for two cases: first where only Hipparcos data has been used, and second including the use of radial velocity data where available. Fifteen Hipparcos Pleiades members have radial velocity data (less than one third of the sample). Working in a mixed case, where some stars have radial velocity information and some do not, is possible through marginalisation of terms with missing data from the joint PDFs.

With the inclusion of radial velocity data, the variance in the space velocity has been found to be nearly 6 kms^{-1}. We believe that a lack of homogeneity in the data is responsible for the large variance, because the data is compiled from several sources. While there is no change in the estimation of the mean distance to the cluster between the two cases, the formal errors are in fact larger with the inclusion of radial velocity data. Therefore, the lack of an accurate and homogeneous catalogue of radial velocities for a significant number of Hipparcos Pleiades stars means that radial velocity information has been ignored, and the results shown are for a fitting to a Gaussian distribution in proper motion rather than a three-dimensional space velocity.

The variance in the distribution in proper motions is found to be 1.7 ± 0.3

and 1.5 ± 0.2 mas year^{-1} for μ_{α^*} and μ_δ, respectively. Taking the distance to be 120.3 pc, this variance is equivalent to a variance in the velocity distribution of the cluster of 1.3 ± 0.4 kms^{-1}.

6.7.3 Size

The spatial distribution of the open cluster is modelled using a spherical Gaussian distribution, with its center at some distance R and a variance around the mean distance σ_R. The variance has been included as a colour-dependent term and is estimated for each colour bin. The results show an increase in σ_R with an increase in colour $(V - I)$, which corresponds directly to a decrease in stellar mass.

As the variance in the spatial distribution increases from 3.3 pc to 13.2 pc over the full colour range for the observed stars, we find strong evidence for mass segregation. This relationship between stellar mass and the spatial distribution of stars in a cluster has been reported for the Pleiades with a similar degree of segregation, using a star counting technique for the two-dimensional case of the cluster projected onto the plane of the sky (Converse & Stahler, 2008).

6.7.4 Absolute magnitude distribution

An approximation to the isochrone of the cluster is found by fitting points in the colour-absolute magnitude diagram to find the absolute magnitude distribution of the star as a function of colour. A spline function converts these points into a smooth line, which can be thought of as the observed isochrone of the cluster. The points used for the spline function are found by the fitting method and labelled A and B for each colour bin in Table 6.3. The resolution of the resulting fit only depends on the number of stars, which limits the possible number of bins, and the precision of the data.

The intrinsic dispersion, σ_M, is the dispersion in absolute magnitude around the mean magnitude given by the magnitude distribution, over any narrow

Parameter	Estimated	Error
Distance (pc)	120.3	1.5
μ_{α^*} (mas year^{-1})	19.9	0.3
μ_δ (mas year^{-1})	-45.3	0.3
$\sigma_{\mu_{\alpha^*}}$ (mas year^{-1})	1.7	0.3
σ_{μ_δ} (mas year^{-1})	1.5	0.2

Table 6.1: Colour-independent results obtained for the Pleiades from the method applied to the new Hipparcos reduction.

colour range $(V-I)$ to $(V-I)+\delta_{(V-I)}$. Since the dispersion is not constant over the whole colour range, the value of σ_M is given for each colour bin.

The results of the fitting can be seen in Table 6.3, where they have been plotted onto the colour magnitude diagram of the Pleiades in Fig. 6.2. The theoretical isochrone taken from the PARSEC library (Bressan et al., 2012) can be seen in green. The isochrone was generated using PARSEC v1.1, with an age of 100 Myr, and $Z = 0.03$.

Excluding the turn off ($V-I < 0.1$), the shape of the theoretical isochrone is accurately obtained by the fitting procedure (see Fig. 6.2). The ~0.3 mag difference in absolute magnitude between the theoretical isochrone and the sequence found by the fitting procedure is the discrepancy historically reported between Hipparcos and photometry-based methods.

Above the main sequence turn off, neither the theoretical isochrone nor the results from this work accurately fit the data. This is due to the presence of stars migrating to the giant branch.

The intrinsic dispersion of absolute magnitude around the isochrone is found for each colour bin. The large dispersion in the first bin is due to the presence of the turn off, and subsequently the dispersion around the main sequence decreases with increased $V-I$.

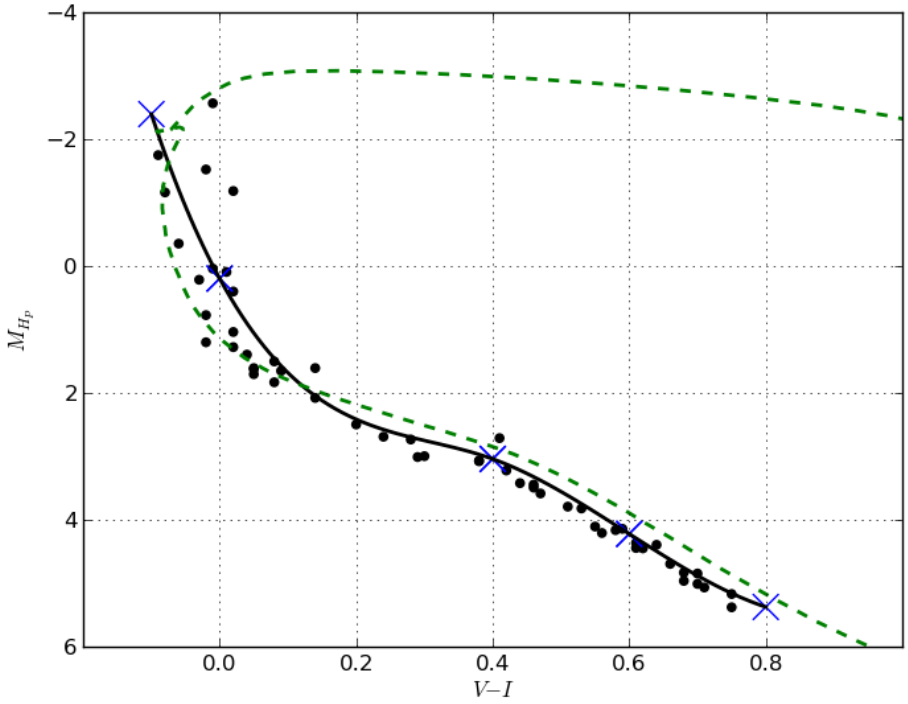


Figure 6.2: Colour-absolute magnitude diagram for the 54 Hipparcos Pleiades members. M_H is the absolute magnitude in the Hipparcos photometric band calculated using the posterior parallax. The blue crosses are the results of the fitting, with the spline function giving the magnitude dependence as the thick line. The green dashed line is the theoretical isochrone generated using PARSEC v1.1, with an age of 100 Myr, and $Z = 0.03$ (Bressan et al., 2012).

Parameter	Estimated	Error
Distance (pc)	122.1	3.7
μ_{α^*} (mas year^{-1})	20.0	0.5
μ_δ (mas year^{-1})	-45.9	0.3
$\sigma_{\mu_{\alpha^*}}$ (mas year^{-1})	2.0	0.4
σ_{μ_δ} (mas year^{-1})	1.3	0.3

Table 6.2: Results obtained for the Pleiades after cutting all stars at or above the apparent location of the main sequence turn off: $(V - I) > 0.1$

	$(-0.1 < V-I < 0.0)$		$(0.0 < V-I < 0.4)$		$(0.4 < V-I < 0.6)$		$(0.6 < V-I < 0.8)$	
Parameter	Estimated	Error	Estimated	Error	Estimated	Error	Estimated	Error
σ_R (pc)	3.4	1.0	4.9	0.9	10.3	2.9	13.1	3.8
σ_M (mag)	1.6	0.5	0.45	0.06	0.22	0.06	0.17	0.06
A (start point)	-2.4	0.7	0.2	0.2	3.0	0.1	4.2	0.1
B (end point)	0.2	0.2	3.0	0.1	4.2	0.1	5.4	0.3

Table 6.3: Colour-dependent results obtained for the Pleiades from the method applied to the new Hipparcos reduction. In the four bins there are 9, 21, 12, and 12 stars (low $(V-I)$ to high). A and B are the points found for the spline function fitting to the colour-absolute magnitude relationship, with A at the start of the bin and B at the end.

	$(0.1 < V-I < 0.4)$		$(0.4 < V-I < 0.6)$		$(0.6 < V-I < 0.8)$	
Parameter	Estimated	Error	Estimated	Error	Estimated	Error
σ_R (pc)	9.3	2.5	8.2	2.3	12.5	3.6
σ_M (mag)	0.14	0.06	0.22	0.06	0.13	0.06
A (start point)	1.6	0.2	3.1	0.1	4.1	0.1
B (end point)	3.1	0.1	4.1	0.1	5.8	0.2

Table 6.4: Results obtained for the Pleiades after cutting all stars at or above the apparent location of the main sequence turn off. The results remain unchanged within the error range. In the three bins there are 9, 12, and 12 stars (low $(V-I)$ to high).

6.8 Correlations

The main argument against Hipparcos-based distance estimates of open clusters has been that Hipparcos trigonometric parallaxes have correlated errors on small angular scales. Indeed, this has been cited as the cause of the large (0.3 mag) discrepancy in the Pleiades distance modulus between Hipparcos-based and photometry-based methods. Narayanan & Gould (1999) argue that correlated parallax errors cause a stark difference between Hipparcos trigonometric parallaxes and the Hipparcos kinematic parallaxes derived from proper motions.

According to Narayanan & Gould (1999), the proper-motion-based parallax can be determined from Hipparcos data using

$$\varpi_{\mathrm{pm},i} = \frac{\left\langle (\boldsymbol{V}_t)_i \,\middle|\, \boldsymbol{C}_i^{-1} \,\middle|\, \boldsymbol{\mu}_{\mathrm{HIP},i} \right\rangle}{\left\langle (\boldsymbol{V}_t)_i \,\middle|\, \boldsymbol{C}_i^{-1} \,\middle|\, (\boldsymbol{V}_t)_i \right\rangle}, \tag{6.47}$$

where $(\boldsymbol{V}_t)_i$ is the transverse velocity of the cluster in the plane of the sky at the position of the star i, $\boldsymbol{C}_i$ is the sum of the velocity dispersion tensor divided by the square of the mean distance to the cluster and the covariance matrix of the Hipparcos proper motion, and $\boldsymbol{\mu}_{\mathrm{HIP},i}$ is the vector describing the proper motion of the star.

Narayanan and Gould used a plot showing the contours of the difference between Hipparcos parallaxes and the parallaxes derived from proper motions (Eq. 6.47) to argue that the Hipparcos parallaxes are systematically larger by up to two mas throughout the inner 6° of the Pleiades. This figure has been recreated in Fig. 6.3 with the 54 member stars used in this work.

Figure 6.4 has been produced using the method described by Narayanan & Gould (1999), but using parallax data from the new Hipparcos reduction. While in the original plot from Narayanan and Gould there is clearly a region where the trigonometric parallaxes are larger than those derived from proper motions, this feature has been reduced in severity by a factor of two by using the new Hipparcos reduction.

This new figure confirms, for the case of the Pleiades, claims made by van Leeuwen (2007) that correlations in the new reduction have been reduced significantly. Additionally, this disagrees with claims that the shorter distance for the Pleiades derived from Hipparcos is due to the correlated errors in parallaxes for Pleiades stars. If this was the case, one would expect the distance estimate from the new Hipparcos reduction to be greater now that the correlations have been reduced.

In fact, the distance derived from the new reduction in this work and by van Leeuwen (2009) put the Hipparcos distance to the Pleiades at 2.5 pc longer than those derived from the original Hipparcos catalogue, a much smaller difference than the roughly 10% historic discrepancy. The method presented here has also been applied to the original Hipparcos catalogue. The difference between the estimated distance from the old and new Hipparcos reductions is only 2%. The small change in distance estimate despite the reduction in correlations by a factor of 10 implies that correlations cannot be responsible for the long-standing discrepancy.

Additionally, in the calculation of the proper-motion-based parallax, a mean distance to the cluster must be assumed. Narayanan & Gould (1999) derived a distance to the Pleiades of 131 pc using Hipparcos proper motions, which was then used in calculating the proper-motion-based parallaxes that form the basis of Fig. 6.3 and their argument against Hipparcos. Using the distance of 120 pc as found in this work and implied by studies using Hipparcos parallaxes, the baseline for the correlation plots is shifted, as can be seen in Fig. 6.6. In this final figure no significant residual biasing the results is apparent, contrary to the original claims.

6.9 Results - Hyades

This method has also been applied to the new Hipparcos reduction for the Hyades open cluster. Of the 282 potential Hyades members used by Perryman et al. (1998), a detailed study by de Bruijne et al. (2001) finds 218 probable

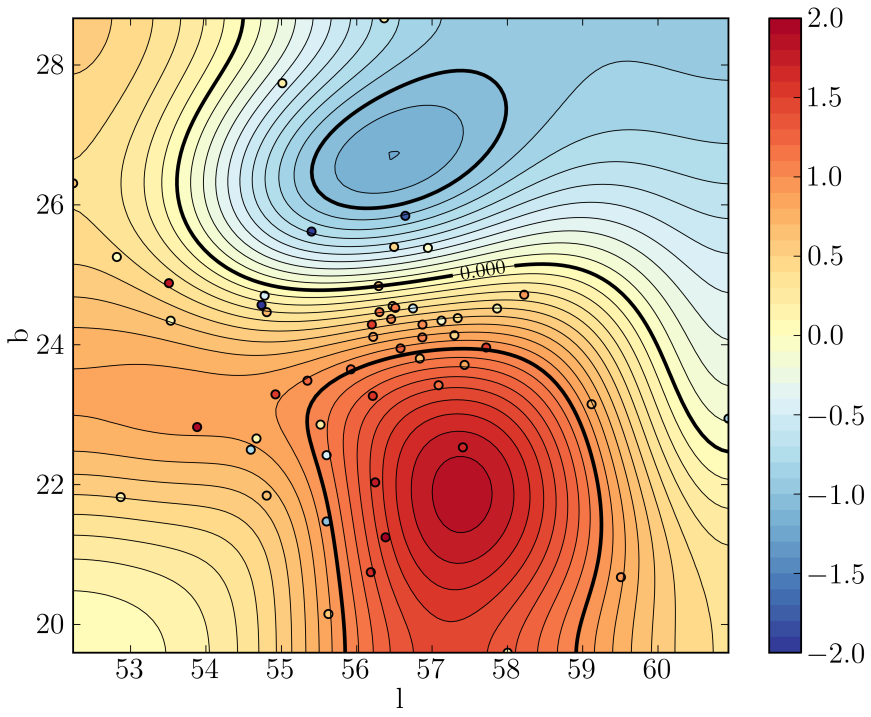


Figure 6.3: Smoothed contours of the difference between the Hipparcos parallaxes and the proper-motion-based parallax ($\varpi_{Hip} - \varpi_{pm}$) of the 54 Pleiades cluster members, with data taken from the original Hipparcos catalogue (1997). Thin contours are at intervals of 0.1 mas, thick contours at intervals of 1 mas.

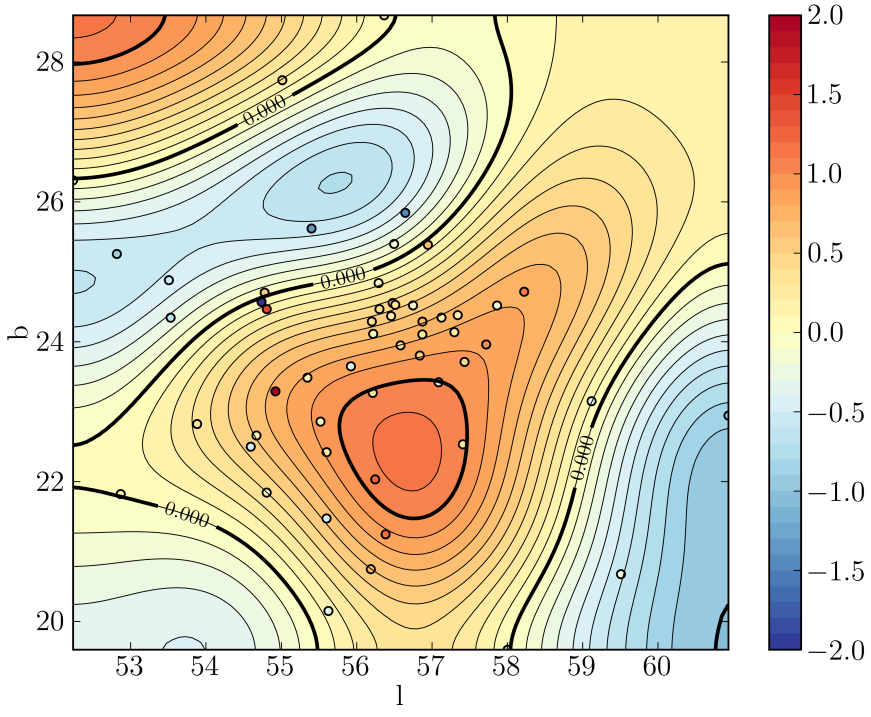


Figure 6.4: Same as Fig. 6.3, but with parallax data taken from the new Hipparcos reduction (2007).

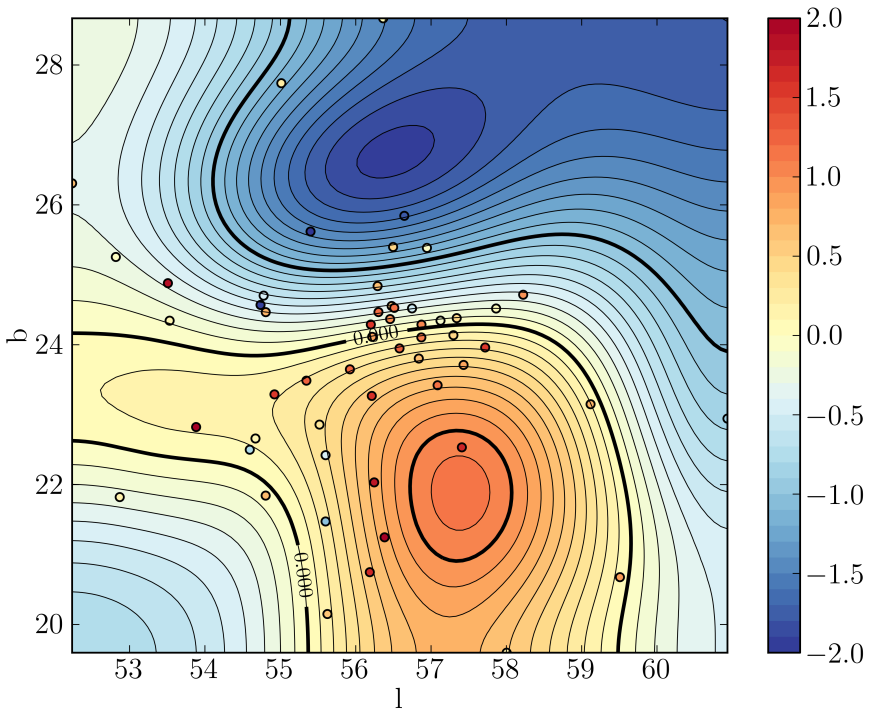


Figure 6.5: Same as Fig. 6.3, but with an assumed mean cluster distance of 120 pc as implied by the Hipparcos data for the computation of the proper motion based parallax.

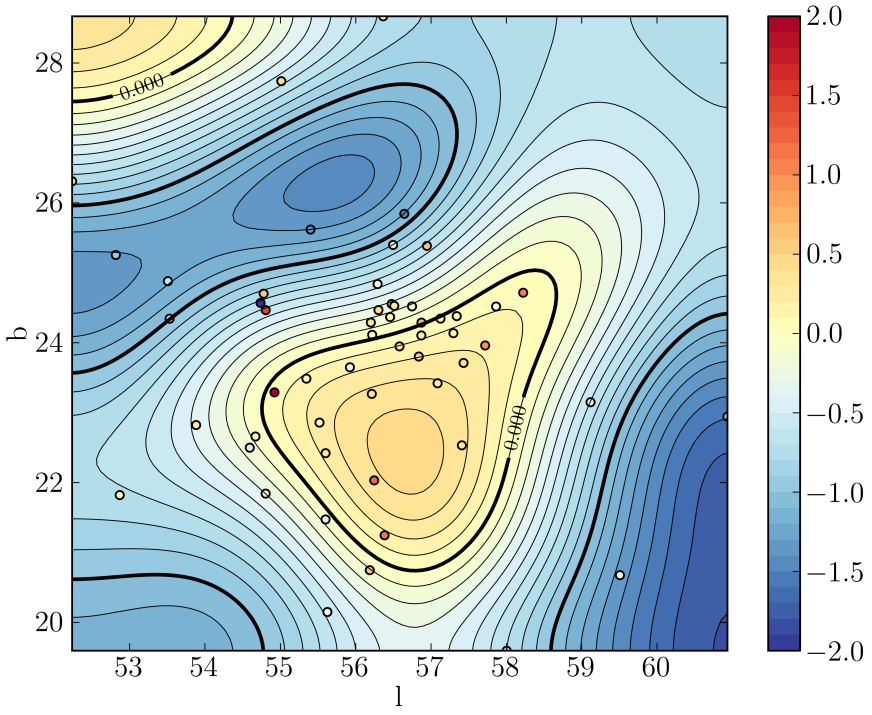


Figure 6.6: Same as Fig. 6.4, but with an assumed mean cluster distance of 120 pc as implied by the Hipparcos data for the computation of the proper motion based parallax.

members, which are used as the basis of this study. Radial velocities from ground-based observations have been collected in Perryman et al. (1998) and used extensively here.

The star HIP20205 was rejected as a giant, because the isochrone fitting does not currently support fitting of the giant branch.

Galli et al. (2012) reject the star HIP28774 in their analysis of the Hyades cluster due to conflicting data for this star in the old and new Hipparcos reductions. In the following sections, this star is not present in our membership list after selecting only the stars in the inner 10 pc. When using the full 218 probable members, the star is included, however its removal has no real effect on the results. This is due to the very low precision of the data on this star in the Hipparcos catalogue, and therefore its very low statistical weight within the method.

6.9.1 Distance

As in Perryman et al. (1998), we select only the stars within the inner 10 pc for the distance estimation, since the large spatial dispersion and presence of numerous halo stars is not modelled well by a spherical Gaussian distribution. The distance to the Hyades has been estimated as 46.35 ± 0.35 pc. This is identical to Perryman et al. (1998), who finds 46.34 ± 0.27 pc, and slightly smaller than van Leeuwen (2009), who finds 46.45 ± 0.50 pc. Perryman et al. (1998) used the original Hipparcos catalogue, whereas van Leeuwen (2009) used the new reduction from 2007. Both authors used differing membership selection and determination techniques.

The three-dimensional dispersion in the spatial distribution of the cluster's core is estimated to be 3.4 ± 0.2 pc, which is consistent with the observed distribution in sky position. Assuming this dispersion in the spatial distribution, the selection of stars with cluster radius of less than 10 pc corresponds to a 3 σ limit, and so the spatial distribution of the core of the cluster should not be significantly affected.

Using the ML method on the full 218 probable member stars, including halo stars, the mean distance is found to be 42.6 ± 0.5 pc. The decrease in precision with the increase in the number of stars is attributed to the clearly non-Gaussian distribution of a dense core and a large number of disperse halo stars. In this case the spatial distribution of the cluster's core is estimated to be 15.8 ± 3.9 pc. The modelling of the spatial distribution could be improved through the use of a King's profile (King, 1962) or an exponential distribution. This is being considered for future improvements.

That member stars are found with distances from the center greater than its tidal radius of ~10 pc (Madsen et al., 2001) is reasonable, and is expected due to the large number of so-called halo stars. These stars have been found to exist in the Hyades at a radius of 10 to 20 pc (Brown & Perryman, 1997), and although they are effected by the galactic gravitational field, they remain bound to the cluster for some significant time.

6.9.2 Kinematics

The space velocity of the Hyades has been found to be 46.5 ± 0.2 kms^{-1}, with an internal dispersion on the velocity of 1.11 ± 0.05 kms^{-1}. The internal dispersion is slightly larger than those in the literature reviewed by de Bruijne et al. (2001), who find a dispersion of ~0.3 kms^{-1}.

As for the case of the Pleiades, the high velocity dispersion is attributed to an inhomogeneous radial velocity data. As highlighted by Brown & Perryman (1997), the radial velocity data for the Hyades comes from a combination of sources with greatly varying precision and zero points.

6.9.3 Absolute magnitude distribution

As described in Sect. 6.7.4, an estimate of the isochrone of the cluster is produced through fitting a smoothed line to the mean absolute magnitude M_{mean} in Eq. 6.5. The results of the fitting can be seen in Fig. 6.7, with the theoretical isochrone overlaid in green.

Parameter	Estimated	Error
Distance (pc)	46.35	0.35
U (kms^{-1})	-42.24	0.11
V (kms^{-1})	-19.27	0.12
W (kms^{-1})	-1.55	0.11
σ_{UVW} (kms^{-1})	1.10	0.05

Table 6.5: Colour-independent results obtained from the method applied to Hyades stars in the new Hipparcos reduction.

In contrast to the results for the Pleiades, the isochrone from the ML method and the theoretical isochrone from the PARSEC library are in strong agreement in both shape and position over most of the main sequence, with some divergence at the extreme ends of the colour range, where there are few stars to constrain the model fit.

The offset in the case of the Pleiades was caused by the long-standing discrepancy between Hipparcos and photometric methods. This is not present in the case of the Hyades, where Hipparcos-based distances generally agree with other methods.

Dispersion around the main sequence has been greatly reduced compared with computing the absolute magnitude from the data directly, by computing posterior distances for each star from the results of the fitting and the individual Hipparcos observations.

6.10 Outlook for Gaia

The method described in Sect. 6.2 will be particularly useful after the release of the Gaia astrometric catalogue. That the Gaia catalogue will include all of the information required for applying this method, including radial velocities for G_{RVS} <16, in one self-consistent catalogue makes Gaia ideal for studying open clusters to greater precision and at greater distances than was possible previously. Indeed, Gaia is expected to observe some one billion stars,

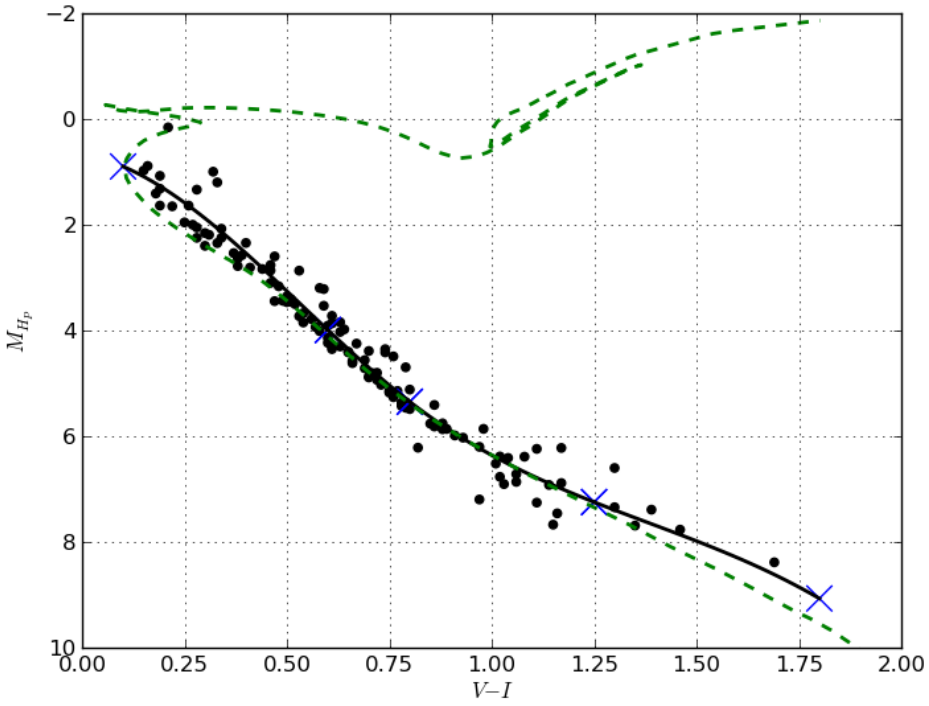


Figure 6.7: Colour-absolute magnitude diagram for the 128 Hipparcos Hyades members found in the inner 10 pc of the clusters core. M_{H_P} is the absolute magnitude in the Hipparcos band, and is calculated using the posterior parallax. The blue crosses are the results of the fitting, with the spline function giving the magnitude dependence as the thick line. The dashed line is the theoretical isochrone from the PARSEC library, with an age of 630Myr and $Z = 0.024$.

	$(0.1 < V-I < 0.6)$		$(0.6 < V-I < 0.8)$		$(0.8 < V-I < 1.25)$		$(1.25 < V-I < 1.8)$	
Parameter	Estimated	Error	Estimated	Error	Estimated	Error	Estimated	Error
σ_R (pc)	3.42	0.27	5.50	0.67	2.30	0.15	2.55	0.45
σ_M (mag)	0.39	0.04	0.32	0.04	1.63	0.19	0.34	0.12
A (start point)	0.77	0.24	3.98	0.05	5.25	0.07	7.25	0.15
B (end point)	3.98	0.05	5.25	0.07	7.25	0.15	9.26	0.57

Table 6.6: Colour-dependent results obtained from the method applied to the new Hipparcos reduction for the Hyades open cluster. In the four bins there are 76, 55, 53, and 21 stars (low $(V-I)$ to high).

.

including stars in numerous open clusters. This will allow application of this method to many more clusters, including those at much greater distances than was previously possible.

To test the performance of the ML method with Gaia data, a simulated open cluster was created, and simulated Gaia observational errors applied. To continue with the case study of the Pleiades, a simulated star catalogue for a Pleiades-like cluster was obtained from GaiaSimu (Masana et al., 2010; Robin et al., 2012). GaiaSimu is a set of libraries containing the Gaia Universe Model and instrument models used by the Gaia Data Processing and Analysis Consortium (DPAC). It contains a database of 500 simulated open clusters, including the Pleiades, constructed from Padova isochrones and a Chabrier/Salpeter IMF. To apply simulated Gaia observational errors, the data was processed using GOG (see Ch. 3).

In the Gaia case, the selection function in Eq. 6.4 is modelled as a step function at $G = 20$ mag. Because Gaia is expected to be complete up to this magnitude, the step function should be a good approximation to the real case.

6.10.1 Pleiades with Gaia

The GaiaSimu and GOG simulated Pleiades contains some 1000 stars, placed at a distance of 130 pc and occupying the same region of the sky as the real Pleiades. With simulated parallax errors of between 10 and 100 μas, the vast majority of the star's distances are very accurately measured.

With such a precise data set, both the estimated distance from MLE and the distance obtained directly from the mean of the parallax are both within 0.01 pc of the true value. This highlights that, for the nearest open clusters, it will be possible with Gaia data to go further beyond the current goal of determining the distance and kinematic and structural parameters, to having highly detailed information on many aspects of open clusters.

In such cases, the mean distance of a cluster determined through the ML method will not in itself be useful, although individual stars' posterior dis-

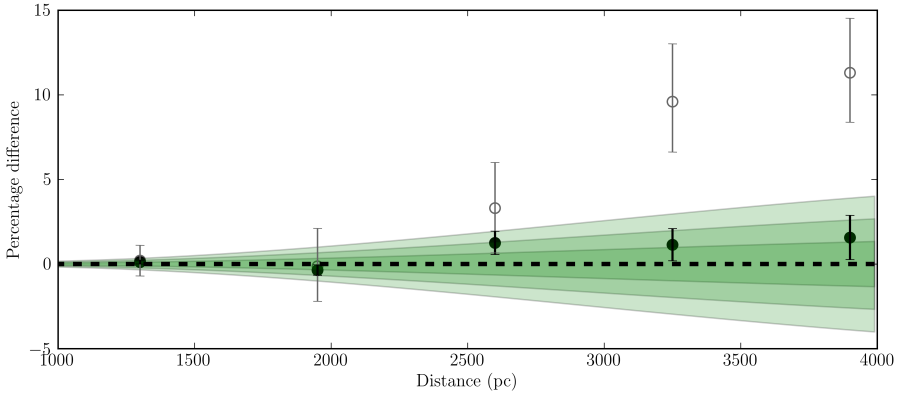


Figure 6.8: Results of the distance estimation to a simulated Pleiades-like cluster placed at a factor of 10 to 30 times the original distance. Filled circles show the percentage difference between the MLE-estimated distances and the true distance, open circles are the percentage difference between the inverse of the mean of the parallax and the true distance, and green shading highlights 1, 2, and 3σ errors extrapolated over the entire range.

tance estimates from the method will be unbiased and therefore preferable to distances found by inverting individual parallaxes. In terms of determining a cluster's spatial distribution, the direct use of parallax information results in a bias in the results not present during the application of the ML method.

6.10.2 Distant clusters with Gaia

To test the performance of the ML method with open clusters at greater distances, the GaiaSimu simulated Pleiades was modified, increasing the distance while conserving all other properties. The open cluster was moved to a range of distances between 10 and 30 times the originally assumed distance of 130 pc (i.e. up to a distance of 4 kpc). Then the Pleiades-like clusters were processed with GOG to simulate Gaia observational errors.

Using the same simulated open cluster moved to different distances allows a direct comparison of the ML method performance at different distances. Figure 6.8 shows the results of the distance estimation using the ML method, showing the percentage difference between the 'true' distance and the distance

estimated from MLE, $\Delta(d_{real} - d_{MLE})/d_{real}$, and comparing this with the percentage difference between the 'true' distance with the inverse of the mean parallax, $\Delta(d_{real} - d_{1/\bar{\varpi}})/d_{real}$. As the distance to the cluster increases, the stars become fainter and the observational errors larger. As can be clearly seen in Fig. 6.8, the mean of the parallax is susceptible to large random error, in addition to Lutz-Kelker effects and other statistical biases, and is unsuitable for accurate distance determination in magnitude-limited data sets and those with significant observational errors.

As with the Hipparcos Pleiades and Hyades data, the CMD is plotted with the isochrone-like sequence obtained from the observational data and shown in Fig. 6.9. These plots have been created using the simulated clusters at 1300, 2600, and 3900 pc, showing that it is possible to obtain a reliable observational isochrone from Gaia observations even when individual parallaxes are strongly affected by observational errors.

With this simulated dataset, the ML method is confirmed as not suffering from significant statistical biases, and it is expected to perform well with real Gaia data.

6.10.3 Membership selection

When studying open clusters, especially those at greater distances, membership selection has always been important and problematic. When applying the ML method to distant open clusters in the Gaia catalogue, the density of stars on the sky could cause problems with misclassification and source confusion.

However, with an expected source density at $G < 20$ mag in the galactic plane of around $3{\times}10^5$ stars per square degree, Gaia's windowing system for object detection is small enough to give a low probability of source confusion even when observing distant open clusters.

In terms of membership selection, the ML method's estimation of an open cluster's parameters can be used directly to perform membership probability tests. If the ML method is primed using a sample of probable members, a

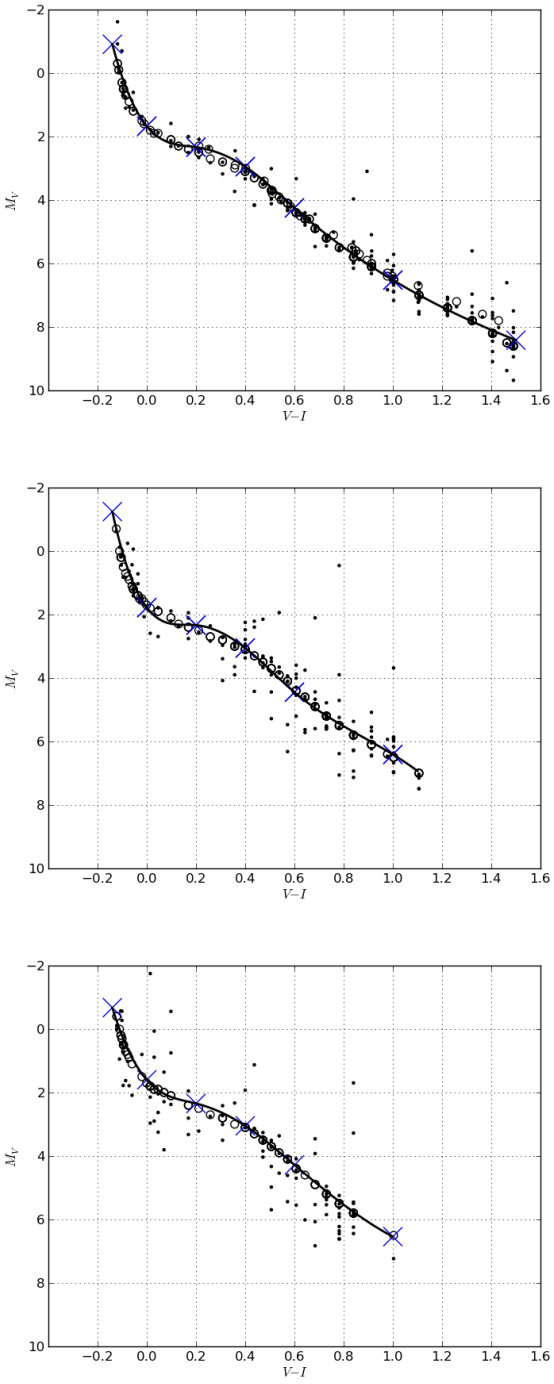


Figure 6.9: CMD for the simulated Pleiades like cluster at distances of 1300 pc (top), 2600 pc (middle) and 3900 pc (bottom). The three clusters have 216, 143 and 114 observed members respectively. Open circles show the 'true' simulated absolute magnitudes without errors, and filled circles show the absolute magnitudes calculated directly from the simulated observations including simulated gaia errors. Stars with negative parallaxes are omitted from the figure but included in the ML estimation. Reddening in $V-I$ is assumed known. The blue crosses are the points fitted with the ML method and the solid line is the resulting isochrone-like sequence.

χ^2 test can be applied to calculate membership probability for each star in the sample using the parallax, proper motion, and radial velocity information simultaneously:

$$K^2 = \mathbf{A}C^{-1}\mathbf{A}^T, \tag{6.48}$$

where $\mathbf{A}$ is the vector $(\varpi_i - \varpi_{ML},\ \mu_{\alpha,i} - \mu_{\alpha,ML},\ \mu_{\delta,i} - \mu_{\delta,ML},\ v_{ri} - v_{r,ML})$, and C^{-1} is the sum of the catalogue's covariance matrix and the variance on each parameter due to the clusters intrinsic dispersion in distance, proper motion, and radial velocity. Here, ϖ_{ML}, $\mu_{\alpha,ML}$, $\mu_{\delta,ML}$, and $v_{r,ML}$ are the mean parallax, mean cluster proper motion and mean radial velocity of the cluster, determined from the ML method.

Stars with $K^2 > 16.25$ are rejected, in correspondence with a χ^2 test at three-sigma level with four degrees of freedom. This process should be performed using an iterative process, rejecting the worst outliers a few at a time and recalculating all fitting parameters, until no further outliers remain.

Here, the estimated cluster distance and space velocity, including intrinsic dispersion, are combined with the individual observations and their associated errors in order to distinguish between members and non-members in a single step.

Testing using a 1000-star sample of GOG simulated field stars added to the five simulated Pleiades samples used in Fig. 6.8, the χ^2 test excluded field stars with a misclassification rate between 0.8 and 0.3%, assuming a worst case of zero radial velocity information.

6.11 Conclusions

An improved method for estimating the properties of open clusters has been presented, and tested using real data on two nearby and well studied open clusters. In addition to distance estimation, internal kinematics and spatial structure were probed, with mass segregation detected in the case of the Pleiades.

These results confirm that the method performs as expected and highlight the potential future uses of such a method when high quality parallax information is available from the Gaia mission.

After revisiting the 'Pleiades problem', an explanation was not found in error correlation problems in Hipparcos[2]. Through the use of simulations we find that Gaia will measure the distance to Pleiades stars with precision of a fraction of a percent, enabling a conclusion to this long running discrepancy.

The ML method can be extended further to give more detailed information, such as including a model for cluster ellipticity and orientation. It is possible to include and compare different spatial and kinematic distributions, allowing one to test predictions on spatial structure, mass segregation, and peculiar motions, and to test for other properties such as cluster rotation. In the case of the absolute magnitude distribution, it would be possible to give age and metallicity estimates by fitting and comparing sequences of different theoretical isochrones.

Unresolved binaries, which complicate studies of open clusters, can be detected using the posterior distances calculated using the ML method and the resulting colour-magnitude diagram. It is possible to extend the method to use a distribution in absolute magnitude that is asymmetrical around the main sequence, in order to consider undetected unresolved binaries within the method.

As mentioned in Sect. 6.7, a lack of quality radial velocity data for Hipparcos Pleiades stars limits the application of the method in fitting the full three-dimensional kinematics of the open cluster. This is expected to change when Gaia comes to fruition, because all stars with $G_{RVS} < 16$ will have radial velocity information from the on-board radial velocity spectrometer. Additionally, very high quality radial velocities for stars in more than 100 open clusters will become available through the Gaia ESO Survey (Gilmore et al., 2012), expanding the scope of the method's application to clusters at much greater

[2]The discrepancy is however still there, and it is not clear if this is due to a problem in the Hipparcos data or in the stellar models. The problem is still under debate.

distances.

7

Error correlations in Gaia

Due to the very large number of stars which will be observed, there will be many cases where large numbers of Gaia parallaxes can be averaged in order to find very precise estimates of the mean distance to certain groups of objects, as will be the case, for example, in studies of open clusters.

However, the precision of these mean distances can be reduced by the error correlations in the Gaia data. The idea of error correlations has been suggested as having a real and measurable effect in Gaia's predecessor, Hipparcos (Lindegren, 1988). In the most extreme case, work such as Narayanan & Gould (1999) assert that error correlations are responsible for the so-called Pleiades problem, where the derived distance to the Pleiades open cluster is around 10% closer from methods based on the Hipparcos data (e.g. van Leeuwen & Hansen Ruiz, 1997) than the distance obtained by other methods (e.g. Pinsonneault et al. (1998), Robichon et al. (1999b), Stello & Nissen (2001)).

A complete re-reduction of the Hipparcos raw data by van Leeuwen (2007), lead to a revised catalogue with error correlations reduced by up to a factor of 10. Distance estimation to the Pleiades using the new data in Sect. 6.8 and by van Leeuwen (2009) show a slight change in the mean distance, but not enough to reconcile the two groups. The method used to construct the new Hipparcos reduction from the raw data was similar to that used by Gaia's Astrometric Global Iterative Solution (AGIS).

In Sect. 7.1, we summarise the concept of error correlations, and provide a method to test the effect of error correlations in Gaia. Then, we apply this method to two important case studies, a simulated open cluster in Sect. 7.3, and the simulated Large Magellanic Cloud (LMC) in Sect. 7.4. Conclusions are given in Sect. 7.5.

7.1 Basic description

Gaia will repeatedly measure the position of numerous stars, and will use this position information to fit a five parameter astrometric solution, including the parallax, for each star observed. The observed position of a star in each transit (observation) will be effected by observational errors such that:

$$\alpha_{\text{observed}} = \alpha_{\text{real}} + \epsilon_{\alpha} \tag{7.1}$$

where α is the position and ϵ_{α} the random error caused by various real-world observational effects.

For two stars observed close together on the sky, the error will be made up of two components, a random component and a component which will be common to both stars, here called the correlated or common error. For a measurement of the positions of these two stars, the observed position of each is a combination of these random and common errors:

$$\alpha_{\text{observed}} = \alpha_{\text{real}} + \epsilon_{\text{random}} + \epsilon_{\text{common}} \tag{7.2}$$

The common error can be caused for example by a slight inaccuracy in the Gaia satellite attitude determination, or the measuring of the basic angle separating Gaia's two telescopes, which will equally effect all stars observed within the same field of view.

Looking at any one star alone, the effect is present yet undetectable, and the formal error on the observations is the standard combination of both the random and common error:

$$\sigma_{\text{total}} = \sqrt{\sigma^2_{\text{random}} + \sigma^2_{\text{common}}} \tag{7.3}$$

However when looking at many stars, as would be the case of an open cluster, the common component of the error in the averaging of many individual measures can not be reduced by the square root of the number of stars, as would be the usual case for a random error. Because the common error is an effect present at an individual transit level, we simulate correlated errors for individual Gaia transits of an open cluster, and then from the obtained position information derive an estimate for the parallax of each star. We take the mean of the parallax measurements of many stars in an attempt to obtain the mean parallax of the open cluster, and observe how the presence of a common error between stars effects the convergence on the real value.

7.2 Method

Gaia transit data contains along and across scan positions in the Gaia focal plane (see Jordi et al. (2010) for details of the Gaia focal plane), transit angle, observation time, and parallax factor. From this transit data, the five astrometric parameters are obtained using a Minimum Least Squares fit. The method is similar to that employed by the Gaia Data Processing and Analysis Consortium (DPAC), the group tasked with the reduction and processing of the Gaia data. The five astrometric parameters, $(\alpha, \delta, \mu_\alpha, \mu_\delta, \varpi)$, are the position, proper motion and parallax for each star. The exact method for fitting

the astrometric parameters to the transit data (as an approximation of AGIS) is given below:

1. For each transit of every star, the along scan CCD position, AL, and the formal error on the along scan position, σ_{AL}, are used to calculate the weighted parameter:

$$\mathrm{AL}_{\mathrm{weighted}} = \mathrm{AL}/\sigma_{\mathrm{AL}} \tag{7.4}$$

2. For each transit of every star the vector P is constructed from the position angle θ, the parallax factor F_{ϖ} and the time t:

$$P = \begin{pmatrix} \sin\theta \\ \cos\theta \\ F_{\varpi} \\ t\sin\theta \\ t\cos\theta \end{pmatrix} \tag{7.5}$$

3. The vector P is weighted by the along scan error:

$$P_{\mathrm{weighted}} = P/\sigma_{\mathrm{AL}} \tag{7.6}$$

4. A least squares fit is obtained using the QR decomposition of P_{weighted}, solving for the five parameters.

5. The solution is a five parameter matrix, representing the five astrometric parameters $(\alpha, \delta, \mu_{\alpha}, \mu_{\delta}, \varpi)$.

For this study, Gaia transit data is simulated using the Gaia Object Generator (GOG) (Luri et al., 2014). The GOG transit data already contains observational errors due to CCD centroiding errors (detailed in Sect. 7.2.1). An additional error is included on top of this, aiming to simulate the so-called Calibration Noise (detailed in Sect. 7.2.2).

7.2.1 Single transit astrometric error model

As mentioned previously, Gaia transit data contains along and across scan positions in the Gaia focal plane, transit angle, observation time, and parallax factor. Below we summarise how GOG simulates each component, including the formal error on the position:

Local plane coordinates are a localised coordinate system defined to simplify computations of position which would require satellite attitude, calibration and orbital data. These local coordinates defined tangentially on the plane of the sky are further simplified by transforming from the celestial plane coordinates to CCD scan coordinates w and z.

The error for the average of a CCD row transit follows the expression:

$$\sigma_w = \frac{\sigma_\eta}{\sqrt{n}} \tag{7.7}$$

$$\sigma_z = p_r \frac{\sigma_\eta}{\sqrt{n}} \tag{7.8}$$

where n is the along scan number of CCDs (typically 9) and p_r is the relation between the across scan and along scan pixel sizes ($p_r = 3$). For both expressions we have used the centroid positioning error, σ_η, whose calculation uses the Cramer Rao lower bound in its discrete form, which defines the best possible precision of the Maximum Likelihood centroid location estimator.

When the position of a star is obtained from a CCD image, its apparent position is shifted with respect to its mean position due to the parallax, e.g..:

$$\alpha = \alpha_{mean} + \varpi f_\alpha \tag{7.9}$$

$$\delta = \delta_{mean} + \varpi f_\delta \tag{7.10}$$

where f is the parallax factor, and ϖ is the parallax. For any given observation, it is possible to calculate the parallax factors, as they do not depend

on the parallax itself.

The parallax factors for the equatorial coordinates are computed following expressions from Green (1985) (with a correction on a sign in the f_δ expression due to a book error):

$$f_\alpha = \frac{1}{\cos\delta}\left(X\sin\alpha - Y\cos\alpha\right) \tag{7.11}$$

$$f_\delta = X\sin\delta\cos\alpha + Y\sin\alpha\sin\delta - Z\cos\delta \tag{7.12}$$

where (X, Y, Z) are the coordinates of the barycentric position of the observer, Gaia in our case (obtained from the satellite ephemeris).

Using the Gaia Model simulation time-scale and the attitude the transit scan angle θ is computed. The respective f_w and f_z parallax factors for the local plane coordinates are:

$$f_w = f_\alpha\sin\theta + f_\delta\cos\theta \tag{7.13}$$

$$f_z = -f_\alpha\cos\theta + f_\delta\sin\theta \tag{7.14}$$

7.2.2 Correlation noise error models

Correlation noise has optionally not been included in the GOG simulation, and so is added manually here. To find the calibration noise per transit, the calibration noise per CCD is taken, as a function of magnitude, from the Gaia Parameter Database, and is divided by $\sqrt{9}$ to take into account that there are 9 CCDs in a row of the Gaia focal plane.

This correlation noise is assumed to be made up of four main components[1]: the basic-angle measurement error, itself composed of 0.5 μas basic-angle mea-

[1]Values are taken from: 'Science Performance Budget Report', EADS-Astrium, 25 February 2011, GAIA.ASF.RP.SAT.00005, issue 5, revision 0. The values are also published in the Gaia Parameter Database.

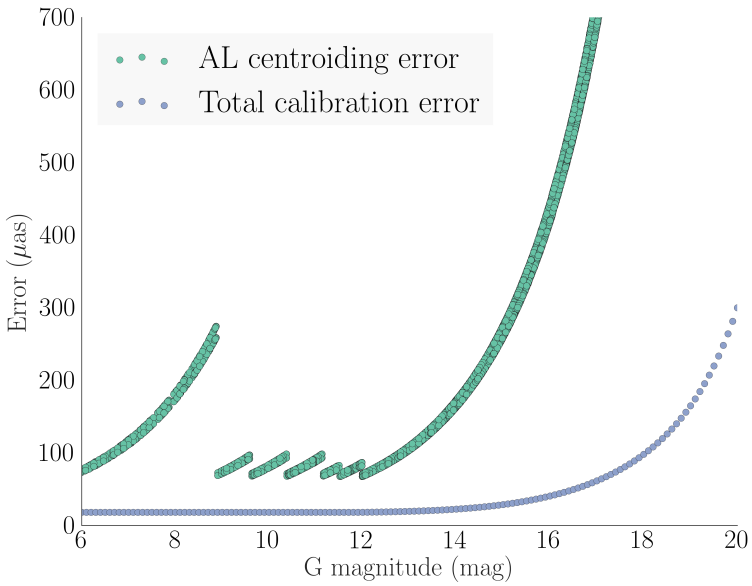


Figure 7.1: The error in a single transit along scan position from photon noise (green) and from calibration noise (blue). The y axis is cut at 700 for clarity, but the AL centroiding error continues much higher.

surement error and 0.3 μas basic-angle measurement efficiency error; the attitude error induced by high-frequency attitude disturbances at 3.4 μas and the residual attitude-reconstruction error at 3.1 μas; the residual geometric-calibration error, which is assumed to be 0.3 μas; and a residual chromaticity-calibration error, assumed to be 0.3 times the end-of-mission random parallax error. A breakdown of the error contributions for some example apparent magnitudes is given in Table 7.1. The comparison between the photon noise and the total calibration error is shown in Fig. 7.1.

7.2.3 Application

Two cases are considered, one with correlation and one without:

1. **With correlation**, all stars are assumed to be within the correlated area. There is no spatial dependence on the effect. All stars are assumed

Component \ G	6	8	10	12	14	16	18	20
Photon[1]	75.91	181.14	81.66	67.92	164.26	421.11	1146.57	3848.92
Attitude[2]	4.59	4.59	4.59	4.59	4.59	4.59	4.59	4.59
Basic angle[2]	0.58	0.58	0.58	0.58	0.58	0.58	0.58	0.58
Geometry[3]	3.09	3.09	3.09	3.09	3.09	3.09	3.09	3.09
Residual chromaticity-calibration[3]	17.20	17.07	17.39	17.43	21.55	40.56	97.44	299.43

Table 7.1: The error from the main error sources evaluated at different apparent G magnitudes. The different contributions to the error can be either unique or common to many stars. Each row is labelled with a superscript: 1: unique to each star, 2: common to all stars in FOV, 3: common to all stars in a single CCD. As stars pass down a row of CCDs in the focal plane, the geometry error (3) is equivalently common to all stars passing down a CCD row in the focal plane. These simulated results are for the nominal Gaia mission only, and can change due to real-world effects causing changes away from Gaia's predicted performance. The errors are per transit. Values are taken from: 'Science Performance Budget Report', EADS-Astrium, 25 February 2011, GAIA.ASF.RP.SAT.00005, issue 5, revision 0. The values are also published in the Gaia Parameter Database.

to be observed in each transit. Holl & Lindegren (2012) and Holl et al. (2012) show that correlation is at a maximum with angular separation of less than 0.1 deg, and reduces with increased separation until reduced by around a factor of five beyond 0.7 deg.

The effect of attitude and basic angle errors are common for all stars within the field of view, therefore for each transit a random change in observed position is added to all of the stars. The change in position is the same for all stars (direction, size), and for each transit is drawn from a random Gaussian distribution with mean zero and standard deviation of the sum of the basic angle error and the attitude errors above (added in quadrature).

The effect of the geometric and chromaticity calibration errors is common to all stars in a single CCD. As stars always transit all CCDs in a row, the effect of the geometric calibration error can be thought of as being common to all stars which are in the same CCD row (after a division by $\sqrt{9}$ to take into account the number of CCDs per row). To simulate this effect, for each transit each star is randomly assigned a CCD row (i.e. one of 7 groups) and all stars within each row are subject to a change in observed position drawn from a random Gaussian distribution with mean zero and a standard deviation of the sum of the geometric calibration error and the chromaticity calibration error given above (added in quadrature). Therefore for each transit stars have a common error with other stars in their CCD row, but are independent from other rows. Row selection is random for every transit.

The correlated error is therefore made up of three components, a random error independent to each star (from GOG, due to CCD centroiding error), a random error common to all stars in a CCD row (geometry+chromaticity calibration), and a random error common to all stars (BAM+Attitude).

2. **Without correlation**, a random change in the observed position is

applied to every star in each transit. The random change in observed position is drawn from the a random Gaussian distribution with standard deviation equivalent to the combination of the various effects above, but is a unique draw for every star and for every transit. Therefore the error in each transit is different for each star.

Given both sets of transit data (one with correlated errors and one without), the minimum least squares method given above is used to obtain the observed parallax for each star. Then the mean parallax, $\overline{\varpi}$, to the whole cluster is calculated using this observed data, once for each data set.

The difference between the two measurements $\overline{\varpi_{\mathrm{withCor}}} - \overline{\varpi_{\mathrm{noCor}}}$ is calculated using the mean cluster parallax with and without correlations. This difference is the effect of the correlated errors. The residual between the with correlation and without correlation cases does not decrease with the inclusion of more stars, as can be seen in Fig. 7.2. In this figure, it is clear that the random error is large with few stars in the sample, and decreases with the square root of the number of stars through the inclusion of more stars in the sample. The mean tends towards the true value with the inclusion of more stars, however the residual due to the common error can not be averaged away because it is common to all stars. Including even thousands of stars does not decrease this effect. Therefore the effect of correlated errors can be thought of as an additional random error on the mean calculated from correlated observations.

By taking the GOG transit data, and repeating the process many times (adding random and correlated errors, finding the mean parallax for each case and taking the difference between the two), the distribution of the correlation values is obtained. Assuming that the effect follows a Gaussian distribution, the mean and standard deviation tell us about the size and type of the effect of our correlated errors.

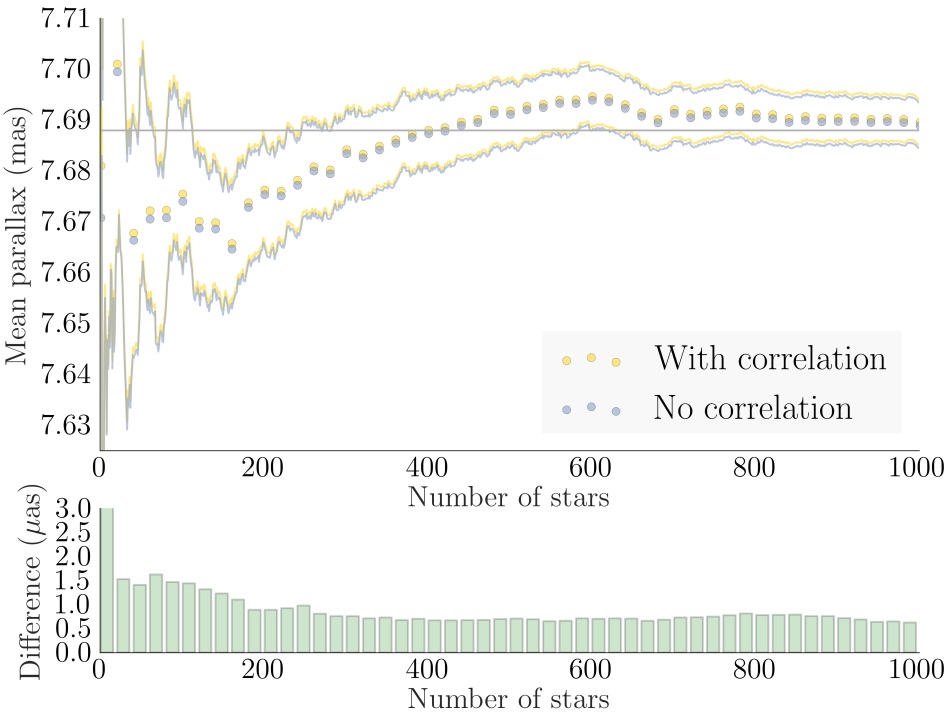


Figure 7.2: Filled circles are the mean cluster distance with increasing number of stars included in the sample. For clarity, a point is drawn for only one in ten stars. The coloured lines around the mean are the one sigma error on the mean. The parallax and residual are given in mas. The grey line at 7.69 is the true mean parallax.

7.3 Open clusters

To test the impact of correlated errors in open cluster distance estimation, transit data for the a Pleiades-like open cluster has been simulated with GOG. This data contains along and across scan positions, transit angle, observation time, and parallax factor. The simulated cluster contains roughly 3000 members and covers an area of around four square degrees, although all stars are assumed to be in the correlated area for this exercise, corresponding to a worst case scenario. The true mean distance to the open cluster is taken as 130 pc.

By applying the method outlined in Sect. 7.2, the effect of correlated errors has been evaluated. As the effect is a random variable, its distribution is found by repeating the inclusion of correlated errors many times. The mean is found to be close to zero, and the standard deviation is 0.7 μas, as seen in Fig. 7.3.

The effect can be thought of as being an additional error on the mean, on top of the one derived from the individual error of each star. This error can not be reduced by including a greater number of stars unless they come from a larger, uncorrelated area.

7.4 LMC

The effect of correlated errors is limited to stars which are observed simultaneously, and is therefore only present over angular separation of less than 1 degree. From the results for fitting the LMC distance given in Sect. 5.4, it is clear that the precision of less than 0.1 μas as achieved above will not be possible in a compact region with strongly correlated errors. The error due to correlation (found to be at 0.7 μas in Sect. 7.3) constrains the determination of the assumed 20.8 μas mean distance to the LMC to being determined with a accuracy no better than 3%.

However, as can be seen in Fig. 5.5, the LMC is extended over several degrees in each direction. The effect of correlated errors, being derived from a random process, is itself a random variable, with mean zero. By taking several

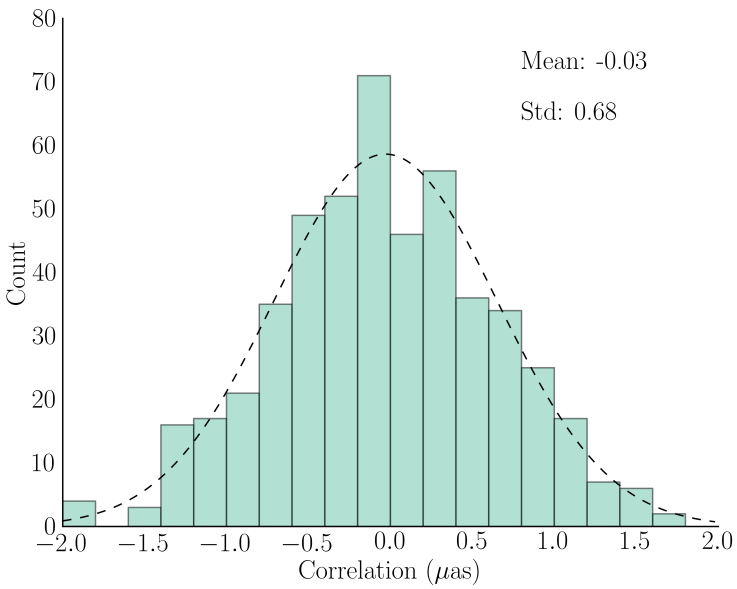


Figure 7.3: Normalised histogram of the residual $\overline{\varpi_{withCor}} - \overline{\varpi_{noCor}}$ found for 500 tests, with a fitted Gaussian distribution.

uncorrelated regions, the effect is reduced. As the LMC covers approximately 90 square degrees on the sky, the effect of correlated errors will be reduced by a factor of around 9.5, reducing the size of the correlated error to be comparable with the precision of the no-correlation case above.

While a global mean distance estimate is therefore possible using Gaia data, it is known that the LMC is highly inhomogeneous, with different variable star types and stellar populations occupying different regions. As the effect of error correlations will be consistent, the relative distances of the different populations can be explored using methods such as that presented in Sect. 5.4. However the absolute distance to sub-populations in the LMC, such as Cepheids which are constrained to the central bar region of the LMC, will be limited in the possible accuracy due to correlation effects.

In Sect. 5.4.3, tests have been made in applying the distance determination method from Sect. 5.4 to the LMC area divided into a grid on the sky. Doing so allowed the estimate of mean distances to the LMC area in different parts of the sky, and the precision was found to be sufficient to allow reconstruction of the angle and orientation of the LMC disk. However, after the inclusion of the effect of error correlations, the small area of the required grid means that the correlated errors play a significant role and detail of the LMC disk orientation is lost.

7.5 Conclusions

In Gaia there will be abundant cases where many parallax measurements will be used to obtain average properties. By using a statistical approach, such as the one outlined in Sect. 5.4, it is possible to obtain an improvement in precision over the direct calculation of the mean, and to avoid many statistical biases. Correct statistical treatment is essential, else the very high precision of Gaia can be lost through untreated observational biases or poorly known posterior corrections.

The number of stars which Gaia will observe provides the opportunity to

obtain very precise distances. For example, observed members of open clusters could number as high as thousands of stars per cluster. In these cases, problems of error correlations can limit the possible accuracy of mean properties. Taking stars from a wide area will negate the effect, however this is not always possible. In compact cases (less than a few square degree), the additional component in the error due to correlations should be included in the formal error propagation.

If μ is the mean parallax, ρ is the correlation between two stars, and σ_ϖ is the formal error on the parallax, the formal variance on the mean, σ_μ, is:

$$\sigma_\mu^2 = E\left[\mu^2\right] \tag{7.15}$$

$$\sigma_\mu^2 = E\left[\left(\sum \varpi_i\right)^2\right]/n^2 \tag{7.16}$$

$$\sigma_\mu^2 = \sum E\left[\varpi_i.\varpi_j\right]/n^2 \tag{7.17}$$

$$\sigma_\mu^2 = (n\sigma_\varpi^2 + (n^2-n)\rho\sigma_\varpi^2)/n^2 \tag{7.18}$$

So, for large n, the precision on the average parallax is

$$\sigma_\mu \sim \sigma_\varpi\sqrt{(1/n+\rho)} \sim \sigma_\varpi\sqrt{\rho} \quad \text{asymptotically} \tag{7.19}$$

Assuming an end-of-mission 10 μas parallax precision for a bright star in Gaia, and taking the expected per transit correlation error given in Sect. 7.2 and dividing by $\sqrt{80}$ for the mean number of transits, we obtain an upper limit on the formal error of 2 μas, in accordance with Fig. 7.3. In most cases of course the systematics will be much smaller, and are within the expected range for Gaia performance.

This does however shows that, as soon as $n > 1/\rho$, having more stars is not beneficial. With $\rho = 0.05$, having more than only some tens of bright stars in a compact area does not help. This number of stars is very small for Gaia,

which is capable of observing 10^6 stars per square degree in compact regions.

This puts a strict limit on the accuracy which can be obtained using Gaia over small regions of the sky. The formal error on a mean distance estimate, obtained from the variance in the measurements though the standard methods, will not contain the correlated part and may therefore state high precision regardless of the accuracy limit.

Taking stars from distinct regions, such as taking parallax measurements for Cepheid variable stars from all over the sky in order to calibrate the Cepheid period-luminosity relation, poses no theoretical limit in the achievable accuracy. In compact cases such as open and globular clusters, and external resolved galaxies, the effects of correlated errors must be taken into account.

The concepts covered here will be subject to further investigation.

8

Conclusions

The aim of this thesis was to develop and prepare a set of tools for luminosity calibration with the Gaia data. To the uninitiated this would appear easy: invert the parallaxes of one thousand million stars to obtain the largest ever compilation of distance measurements, and combine these distances with apparent magnitudes to obtain absolute luminosities. As we have seen in the introductory chapters, it is possible to do much better through statistical *inference* of derived quantities, whereby biases are avoided and the precision of the result is improved.

Within this thesis we have seen several examples of how the straightforward approach leads to inaccurate or misleading results. For open clusters we have seen that the inverse of the mean of the parallax ($\frac{1}{<\varpi>}$) progressively diverges from the true value with increasing observational errors. For the LMC we have seen that the mean of the inverse of the parallax of the Cepheid population

gives 111 kpc while the true value was 48 kpc. Such examples validate the initial argument that simply inverting the parallax to obtain a distance is the incorrect solution. Of course, when using the real Gaia catalogue to obtain distances and absolute magnitudes there will be no possible comparison with the 'true' values or in many cases even with results from other missions. A consumer of the Gaia data must be sure that they are not committing such mistakes, otherwise the benefit of the μas precision of Gaia (its main selling point and mission justification) will be lost.

This thesis therefore presents a method which can be used to successfully obtain distances, absolute magnitudes, and other information of interest (space velocity, isochrones, etc.) in a way which utilises the full range of available data to optimise the result. The method was introduced for a simplified case and then progressively amended and reformed for several key problems in the field of the distance scale and luminosity calibration. The resulting statistical methods have been implemented and tested so as to be available for use with the release of the Gaia data. However, the additional benefit of the formulation used here is its adaptability, and by simply extracting a model from the presented case studies and replacing it with another, the method can be applied in a range of cases which have not been covered here. Calibration of the main sequence, the tip of the Giant branch, or the distance to some specific structure in the Milky Way can all be possible with the Gaia data, and the tools developed during the course of this thesis can be applied to several problems beyond the presented case studies.

The future legacy of this thesis is therefore not only in applying the derived methods to real Gaia data, but (similarly to work undertaken with Hipparcos) keeping the justification and rationale of the thesis in the collective consciousness of potential users of parallax data to ensure that the Gaia data can be utilised to its full potential.

In order to fully investigate the capabilities of Gaia for luminosity calibration and the potential caveats which would need to be taken into consideration, it has been necessary to explore various aspects of the Gaia mission. The clear

starting point was to determine and quantify the expected scientific output and the contents of the catalogue. By working with data from the Gaia simulator for the mock end-of-mission catalogue, it has been possible to get an in-depth view of both the contents of the catalogue and the expected precision of its contents. Not only will this analysis of the mock catalogue serve as a useful reference for the scientific community involved in preparing for scientific exploitation of the real data, but the analysis has been essential within the thesis for deciding which applications for luminosity calibration to focus on throughout the remainder of the thesis. For example, the expected number and precision of parallax measurements for Cepheid and RR-Lyrae variable stars has highlighted the potential for calibration of the period luminosity and period-luminosity-metallicity (PLZ) relations, and the need for a method to do so. Furthermore, the initial analysis of the mock catalogue has also made apparent the potential limitations of the Gaia data. Following from the previous example, it was clear that the precision on the metallicity may not be sufficient for a RR-Lyrae PLZ relation calibration, and therefore combination with other datasets would be required.

From a technical point of view, this initial analysis of the mock Gaia catalogue provided much needed familiarisation with the format and the quantity of the expected Gaia data. Extracting subsets of interesting information from the vast quantities of data in the several terabyte catalogue proved an initial challenge, and the incentive to learn crucial computing and data mining skills. The contents of the catalogue itself then became the ideal testing ground for the various applications of luminosity calibration which form the main body of this thesis.

During this thesis it has been necessary to perform detailed study of specific aspects of the inner workings of the Gaia satellite, beyond the global analysis of the expected Gaia catalogue. After returning to the Hipparcos 'distance problem' during the application of the method for open cluster distance determination, it was clear that a discussion of error correlations was necessary. As highlighted by the recent VLBI observations from Melis et al. (2014) being

again contra the Hipparcos distance estimate, it is clear that there is still an open question about cause of the discrepancy (either a problem in Hipparcos or in the stellar evolution models). As the basic principle of observation is the same for Hipparcos and Gaia, it was clear that a more in-depth study of error correlation effects was required, specifically related to the calibration of distances. This study drew from several other topics from this thesis, using GOG to simulate transit data and using simulated open clusters and the LMC to test the possible extent of resulting correlation effects. The results quantify a minimum possible precision for deriving mean quantities from Gaia catalogue data, and provide a good estimate of error correlation effects so that they can be clearly considered when working on small angular scales. While beyond the scope of this thesis, it would be possible to take a similar approach to the Hipparcos data, by using the raw transit data for the Pleiades stars to reconstruct error correlation effects in a realistic manner. Such an approach can be taken as part of future work, and if combined with the robust method for open cluster distance determination for the distance to the Pleiades from this thesis applied to real Gaia data, the question of the distance to the Pleiades can be settled conclusively.

In light of the results of Ch. 7, it should be noted that it will be necessary to modify the method for open cluster distance determination given in Ch. 6, in cases where correlations may be important. For the given case studies of the Hyades an Plieades, it is unlikely that correlated errors will be significant in Gaia data, due to the fact that these nearby clusters cover large areas of the sky. The same can not be said for more distant clusters, which can appear much more compact. In such cases, the covariance matrix of the measurements between stars should be taken into account explicitly within the method in order to correctly account for correlation effects.

One major potential criticism of this thesis is that while Gaia is the core focus of the work, Gaia data is not yet available. This has lead to extensive use of the aforementioned mock catalogue and other simulations. Of course, no new discoveries can derive from luminosity calibration using simulated data,

but due to the complexity of the mathematical methods and their implementation in functioning algorithms, it is essential to undertake rigorous testing using simulated data with known inputs in order to ensure that mistakes have not been made. Where possible, this has been followed up with applications to real data (e.g. Hipparcos, OLGE, EROS, VMC), which has lead to interesting discoveries (e.g. direct evidence of mass segregation in the Pleiades; a new RR-Lyrae PLZ relation in the LMC). For all of the methods presented in this thesis, the completed mathematical derivations and their respective implementations are ready to be applied when the Gaia data is released.

The future legacy of this thesis is threefold. First, is the personal knowledge and experience which I have gained in astrophysics, statistics and computer science which can be utilised in future projects related to Gaia science. Specific skills such as working with massive datasets, extracting and sorting complex information and visualising results in a clear and concise manner, will all compliment the specific statistical experience gained and will surely be of great use when working on luminosity calibration with the real Gaia data upon release. The second aspect is in the potential extension and application of the specific tools outlined in the previous chapters. As seen, the models defined for Bayesian methods can increase in complexity indefinitely. For example the spatial distribution of Galactic stars can be defined as an exponential disk, but can be extended to include thick and thin disk components, the halo, the bulge, the spiral arms and other major features. Variable star PL relations, defined in Ch. 5 as a linear PL relation, can include dependences on metallicity, colour, or others, where there is a physical justification for doing so. The relations could be non-linear, or contain a dog-leg. In fact, several alternative models could be implemented and compared using quantitative statistical methods to accept or reject conflicting models and choose the optimal description of the observational data. In future applications, and with the very high precision of the Gaia data, some of the methods in this thesis may need extending, or tailoring to specific cases and situations.

The third aspect, which incorporates the previous two, is the potential for

the construction of a general tool for extraction of derived quantities such as distance and absolute magnitude from the Gaia catalogue. As the Gaia catalogue is being eagerly awaited for by scientists working in diverse fields of astrophysics, many will be tempted to invert the parallax of their object of interest. The correct extraction of distance information from the Gaia catalogue can be time consuming, and the correct methodology is not always followed. By utilising the experience gained in this thesis to make a general tool which would be useful to a large user base, the methodology developed in this thesis can be made easily accessible. This would have strong implications to the eventual ease-of-use of the Gaia catalogue, and has the potential to result in a tangible improvement on the core output of the Gaia mission.

Listing of figures

References

Arenou, F. & Luri, X. 1999, in Astronomical Society of the Pacific Conference Series, Vol. 167, Harmonizing Cosmic Distance Scales in a Post-HIPPARCOS Era, ed. D. Egret & A. Heck, 13–32

Arenou, F. & Luri, X. 2002, Highlights of Astronomy, 12, 661

Baade, W. 1938, ApJ, 88, 285

Babusiaux, C. 2005, in ESA Special Publication, Vol. 576, The Three-Dimensional Universe with Gaia, ed. C. Turon, K. S. O'Flaherty, & M. A. C. Perryman, 417

Bailer-Jones, C. A. L. 2011, MNRAS, 411, 435

Bailer-Jones, C. A. L., Andrae, R., Arcay, B., et al. 2013, A&A, 559, A74

Bayes, T. P. 1763, Philosophical Transactions of the Royal Society of London, 48

Benedict, G. F., Jefferys, W. H., McArthur, B., et al. 1995, in IAU Symposium, Vol. 166, Astronomical and Astrophysical Objectives of Sub-Milliarcsecond Optical Astrometry, ed. E. Hog & P. K. Seidelmann, 89

Benedict, G. F., McArthur, B. E., Fredrick, L. W., et al. 2002, AJ, 123, 473

Bressan, A., Marigo, P., Girardi, L., et al. 2012, MNRAS, 427, 127

Brown, A. G. A. & Perryman, M. A. C. 1997, in IAU Joint Discussion, Vol. 14, IAU Joint Discussion

Caldwell, J. A. R. & Coulson, I. M. 1986, MNRAS, 218, 223

Cioni, M.-R., Clementini, G., Girardi, L., et al. 2011, The Messenger, 144, 25

Clement, C. M., Xu, X., & Muzzin, A. V. 2008, AJ, 135, 83

Converse, J. M. & Stahler, S. W. 2008, ApJ, 678, 431

Cutri, R. M., Skrutskie, M. F., van Dyk, S., et al. 2003, 2MASS All Sky Catalog of point sources.

de Bruijne, J. H. J. 2012, Ap&SS, 341, 31

de Bruijne, J. H. J., Hoogerwerf, R., & de Zeeuw, P. T. 2001, A&A, 367, 111

Flegal, J. M. & Jones, G. L. 2008, ArXiv e-prints

Foreman-Mackey, D., Hogg, D. W., Lang, D., & Goodman, J. 2013, PASP, 125, 306

Freedman, W. L. & Madore, B. F. 1990, ApJ, 365, 186

Galli, P. A. B., Teixeira, R., Ducourant, C., Bertout, C., & Benevides-Soares, P. 2012, A&A, 538, A23

Gibson, B. K. 2000, Mem. Soc. Astron. Italiana, 71, 693

Gieren, W. P., Fouqué, P., & Gómez, M. 1998, ApJ, 496, 17

Gilmore, G., Randich, S., Asplund, M., et al. 2012, The Messenger, 147, 25

Graney, C. M. 2006, ArXiv e-prints

Gratton, R. G., Bragaglia, A., Clementini, G., et al. 2004, A&A, 421, 937

Green, R. M. 1985, Book, Cambridge ed., 339

Grenier, S., Gomez, A. E., Jaschek, C., Jaschek, M., & Heck, A. 1985, A&A, 145, 331

Groenewegen, M. A. T., Decin, L., Salaris, M., & De Cat, P. 2007, A&A, 463, 579

Haschke, R., Grebel, E. K., & Duffau, S. 2012, AJ, 144, 106

Hastings, W. 1970, Biometrika, 57, 97

Heck, A. 1975, Applications du principe du maximum de vraisemblance a la calibration de criteres de luminosite stellaire

Heck, A. 1978, Vistas in Astronomy, 22, 221

Holl, B. & Lindegren, L. 2012, A&A, 543, A14

Holl, B., Lindegren, L., & Hobbs, D. 2012, A&A, 543, A15

Isasi, Y., Figueras, F., Luri, X., & Robin, A. C. 2010, in Highlights of Spanish Astrophysics V, ed. J. M. Diego, L. J. Goicoechea, J. I. González-Serrano, & J. Gorgas, 415

Jaschek, M., Jaschek, C., Grenier, S., Gomez, A. E., & Heck, A. 1980, A&A, 81, 142

Jenkins, L. F. 1963, General catalogue of trigonometric stellar paralaxes

Jordi, C., Gebran, M., Carrasco, J. M., et al. 2010, A&A, 523, A48

Jung, J. 1970, A&A, 4, 53

Kanbur, S. M., Ngeow, C., Nikolaev, S., Tanvir, N. R., & Hendry, M. A. 2003, A&A, 411, 361

Katz, D., Munari, U., Cropper, M., et al. 2004, MNRAS, 354, 1223

Kim, D.-W., Protopapas, P., Bailer-Jones, C. A. L., et al. 2014, A&A, 566, A43

King, I. 1962, AJ, 67, 274

Klein, C. R., Cenko, S. B., Miller, A. A., Norman, D. J., & Bloom, J. S. 2014, ArXiv e-prints

Koen, C. 1992, MNRAS, 256, 65

Leavitt, H. S. 1908, Annals of Harvard College Observatory, 60, 87

Lee, M. G., Freedman, W. L., & Madore, B. F. 1993, ApJ, 417, 553

Li, C. & Junliang, Z. 1999, in Astronomical Society of the Pacific Conference Series, Vol. 167, Harmonizing Cosmic Distance Scales in a Post-HIPPARCOS Era, ed. D. Egret & A. Heck, 259–262

Lindegren, L. 1988, A correlation study of simulated Hipparcos astrometry., Tech. rep.

Liu, C., Bailer-Jones, C. A. L., Sordo, R., et al. 2012, MNRAS, 426, 2463

Liu, T. & Janes, K. A. 1990, ApJ, 354, 273

Liu, T., Janes, K. A., & Bania, T. M. 1991, ApJ, 377, 141

Longmore, A. J., Fernley, J. A., & Jameson, R. F. 1986, MNRAS, 220, 279

Luri, X. 1995, Servei de publicacions de la Universitat de Barcelona, ISBN 84-475-1072-7

Luri, X., Mennessier, M. O., Torra, J., & Figueras, F. 1996, A&AS, 117, 405

Luri, X., Palmer, M., Arenou, F., et al. 2014, A&A, 566, A119

Lutz, T. E. & Kelker, D. H. 1973, PASP, 85, 573

Macri, L. M., Stanek, K. Z., Bersier, D., Greenhill, L. J., & Reid, M. J. 2006, ApJ, 652, 1133

Madore, B. F., Freedman, W. L., & Sakai, S. 1997, in The Extragalactic Distance Scale, ed. M. Livio, M. Donahue, & N. Panagia, 239–253

Madsen, S., Lindegren, L., & Dravins, D. 2001, in Astronomical Society of the Pacific Conference Series, Vol. 228, Dynamics of Star Clusters and the Milky Way, ed. S. Deiters, B. Fuchs, A. Just, R. Spurzem, & R. Wielen, 506

Makarov, V. V. 2002, AJ, 124, 3299

Masana, E., Isasi, Y., Luri, X., & Peralta, J. 2010, in Highlights of Spanish Astrophysics V, ed. J. M. Diego, L. J. Goicoechea, J. I. González-Serrano, & J. Gorgas, 515

Melis, C., Reid, M. J., Mioduszewski, A. J., Stauffer, J. R., & Bower, G. C. 2014, Science, 345, 1029

Mermilliod, J.-C., Mayor, M., & Udry, S. 2009, A&A, 498, 949

Mermilliod, J.-C., Turon, C., Robichon, N., Arenou, F., & Lebreton, Y. 1997, in ESA Special Publication, Vol. 402, Hipparcos - Venice '97, ed. R. M. Bonnet, E. Høg, P. L. Bernacca, L. Emiliani, A. Blaauw, C. Turon, J. Kovalevsky, L. Lindegren, H. Hassan, M. Bouffard, B. Strim, D. Heger, M. A. C. Perryman, & L. Woltjer, 643–650

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. 1953, J. Chem. Phys., 21, 1087

Morse, J. A., Mathieu, R. D., & Levine, S. E. 1991, AJ, 101, 1495

Narayanan, V. K. & Gould, A. 1999, ApJ, 523, 328

Nelder, J. A. & Mead, R. 1965, The Computer Journal, 7, 308

Nemec, J. M., Nemec, A. F. L., & Lutz, T. E. 1994, AJ, 108, 222

Olsen, K. A. G. & Salyk, C. 2002, AJ, 124, 2045

Oudmaijer, R. D., Groenewegen, M. A. T., & Schrijver, H. 1998, MNRAS, 294, L41

Palmer, M., Arenou, F., Luri, X., & Masana, E. 2014, A&A, 564, A49

Pelupessy, F. I., van Elteren, A., de Vries, N., et al. 2013, A&A, 557, A84

Perryman, M. A. C., Brown, A. G. A., Lebreton, Y., et al. 1998, A&A, 331, 81

Perryman, M. A. C. & ESA, eds. 1997, ESA Special Publication, Vol. 1200, The HIPPARCOS and TYCHO catalogues. Astrometric and photometric star catalogues derived from the ESA HIPPARCOS Space Astrometry Mission

Pinsonneault, M. H., Stauffer, J., Soderblom, D. R., King, J. R., & Hanson, R. B. 1998, ApJ, 504, 170

Pinsonneault, M. H., Terndrup, D. M., Hanson, R. B., & Stauffer, J. R. 2003, ApJ, 598, 588

Powell, M. J. 1964, The Computer Journal, 7, 155

Raboud, D. & Mermilliod, J.-C. 1998, A&A, 329, 101

Ratnatunga, K. U. & Casertano, S. 1991, AJ, 101, 1075

Riess, A. G., Press, W. H., & Kirshner, R. P. 1996, ApJ, 473, 88

Riess, A. G., Strolger, L.-G., Tonry, J., et al. 2004, ApJ, 607, 665

Rigal, J.-L. 1958, Bulletin Astronomique, 22, 171

Robert, C. & Casella, G. 2008, ArXiv e-prints

Robichon, N., Arenou, F., Mermilliod, J.-C., & Turon, C. 1999a, A&A, 345, 471

Robichon, N., Lebreton, Y., & Arenou, F. 1999b, in Galaxy Evolution: Connecting the Distant Universe with the Local Fossil Record, ed. M. Spite, 279

Robin, A. C., Luri, X., Reylé, C., et al. 2012, A&A, 543, A100

Robin, A. C., Reylé, C., Derrière, S., & Picaud, S. 2003, A&A, 409, 523

Sakai, S. & Madore, B. F. 1999, ApJ, 526, 599

Sakai, S., Madore, B. F., Freedman, W. L., et al. 1997, ApJ, 478, 49

Sakai, S., Zaritsky, D., & Kennicutt, Jr., R. C. 2000, AJ, 119, 1197

Shapley, H. 1939, Proceedings of the National Academy of Science, 25, 113

Sharma, S., Bland-Hawthorn, J., Johnston, K. V., & Binney, J. 2011, ApJ, 730, 3

Soszyński, I., Poleski, R., Udalski, A., et al. 2008, Acta Astron., 58, 163

Soszyński, I., Udalski, A., Szymański, M. K., et al. 2009, Acta Astron., 59, 1

Stello, D. & Nissen, P. E. 2001, A&A, 374, 105

Storm, J., Gieren, W., Fouqué, P., et al. 2011, A&A, 534, A95

Tisserand, P., Le Guillou, L., Afonso, C., et al. 2007, A&A, 469, 387

Turon, C., Crézé, M., Egret, D., et al., eds. 1992, ESA Special Publication, Vol. 1136, The HIPPARCOS input catalogue

Udalski, A., Soszynski, I., Szymanski, M., et al. 1999a, Acta Astron., 49, 223

Udalski, A., Szymanski, M., Kubiak, M., et al. 1999b, Acta Astron., 49, 201

Udalski, A., Szymanski, M., Kubiak, M., et al. 1999c, ACTA ASTRONOMICA, 49, 201

Udalski, A., Szymanski, M. K., Soszynski, I., & Poleski, R. 2008, Acta Astron., 58, 69

van Altena, W. F., Lee, J. T., & Hoffleit, E. D. 1995, The general catalogue of trigonometric [stellar] parallaxes

van Leeuwen, F. 2007, A&A, 474, 653

van Leeuwen, F. 2009, A&A, 497, 209

van Leeuwen, F. & Hansen Ruiz, C. S. 1997, in ESA Special Publication, Vol. 402, Hipparcos - Venice '97, ed. R. M. Bonnet, E. Høg, P. L. Bernacca, L. Emiliani, A. Blaauw, C. Turon, J. Kovalevsky, L. Lindegren, H. Hassan, M. Bouffard, B. Strim, D. Heger, M. A. C. Perryman, & L. Woltjer, 689–692

Vasilevskis, S., Klemola, A., & Preston, G. 1958, AJ, 63, 387

Wright, M. H. 1996, Proceedings of the 1995 Dundee Biennial Conference in Numerical Analysis (Eds. D F Griffiths and G A Watson), 191

Zhao, J. L. & Chen, L. 1994, A&A, 287, 68