UPPSALA
UNIVERSITET

*Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology* 2052

# Statistical Methods in Quantitative Spectroscopy of Solar-Type Stars

ALVIN GAVEL

ACTA
UNIVERSITATIS
UPSALIENSIS
UPPSALA
2021

ISSN 1651-6214
ISBN 978-91-513-1237-8
URN urn:nbn:se:uu:diva-446479

Dissertation presented at Uppsala University to be publicly examined on Zoom, Friday, 10 September 2021 at 13:00 for the degree of Doctor of Philosophy. The examination will be conducted in English. Faculty examiner: Dr. Coryn Bailer-Jones (Max Planck Institute for Astronomy).

**Online defence**: https://uu-se.zoom.us/my/alvinzoomroom

**Abstract**

Gavel, A. 2021. Statistical Methods in Quantitative Spectroscopy of Solar-Type Stars. *Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology* 2052. 115 pp. Uppsala: Acta Universitatis Upsaliensis. ISBN 978-91-513-1237-8.

Galactic archaeology is the research field that attempts to reconstruct the history of the Milky Way, using primarily the tools of astrometric studies and chemical studies. The latter in turn uses stellar spectroscopy. Thanks to technological advances, the field of stellar spectroscopy now has access to much larger amounts of observational data than it used to. At the same time, also thanks to technological advances, the field able to use increasingly more sophisticated modelling. This opens up for the possibility of attacking research problems in Galactic archaeology that were previously intractable. However, it also creates a problem: Access to greater amounts of data means that the random errors in studies will tend to shrink, while the systematic errors tend to stay of the same size. At the same time, improvements in modelling means that studies can look for increasingly subtle effects in their data.

Each article in this thesis attempts solve some specific problem within Galactic archaeology - where possible also developing a general method for handling that type of problem in a way that takes systematic errors into account. In Article I we document a code for estimating stellar parameters from spectra observed with UVES. We use a set of benchmark stars to evaluate the performance of the pipeline, and develop a general method for benchmarking similar codes. In Article II we estimate elemental abundances in spectra in the globular cluster M30 as a means of estimating the Parameter T0 in AddMix models of stellar evolution. At the same time we develop a general method for taking into account systematic errors in derived abundances when estimating parameters in stellar evolution models. In Article III we test whether it is possible to use machine learning to estimate alpha abundances from low-resolved BP/RP spectra from the Gaia satellite.

*Alvin Gavel, Department of Physics and Astronomy, Theoretical Astrophysics, 516, Uppsala University, SE-751 20 Uppsala, Sweden.*

ISSN 1651-6214
ISBN 978-91-513-1237-8
URN urn:nbn:se:uu:diva-446479 (http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-446479)

*Dedicated to whoever is the next PhD student in stellar spectroscopy*
*I've tried to write this thesis so you can use parts of it as a "how-to"*
*and occasionally "how not to"*
*Best of luck!*

# List of papers

This thesis is based on the following papers, which are referred to in the text by their Roman numerals.

I **Gavel, A.**, Gruyters, P., Heiter, Ulrike., Korn, A. J., Lind, K., Nordlander, T. (2019)
*The LUMBA UVES stellar parameter pipeline*
Astronomy & Astrophysics, 629, A74

II **Gavel, A.**, Gruyters, P., Heiter, Ulrike., Korn, A. J., Nordlander, T., Scheutwinkel, K. H., Richard, O. A. (2021)
*Atomic diffusion and mixing in old stars*
*VII. Abundances of Mg, Ti, and Fe in M30*
Astronomy & Astrophysics, accepted 24 May 2021

III **Gavel, A.**, Andrae, R., Fouesneau, M., Korn, A. J., Sordo, R. (2021)
*Estimating $[\alpha/Fe]$ from Gaia low-resolution BP/RP spectra using the ExtraTrees algorithm*
Astronomy & Astrophysics, submitted 18 June 2021

Reprints were made with permission from the publishers.

# List of papers not included in the thesis

The following are publications to which I have contributed but that are not included in this thesis.

1. Rendle, B. M., Miglio, A., Chiappini, C., Valentini, M., Davies, G. R., Mosser, B., Elsworth, Y., García, R. A., Mathur, S., Jofré, P., Worley, C. C., Casagrande, L., Girardi, L., Lund, M. N., Feuillet, D. K., **Gavel, A.**, Magrini, L., Khan, S., Rodrigues, T. S., Johnson, J. A., Cunha, K., Lane, R. L., Nitschelm, C., Chaplin, W. J. (2019)
   *The K2 Galactic Caps Project - going beyond the Kepler field and ageing the Galactic disc*
   Monthly Notices of the Royal Astronomical Society, 490, 4465-4480
2. Worley, C. C., Jofré, P., Rendle, B., Miglio, A., Magrini, L., Feuillet, D., **Gavel, A.**, Smiljanic, R., Lind, K., Korn, A., Gilmore, G., Randich, S., Hourihane, A., Gonneau, A., Francois, P., Lewis, J., Sacco, G., Bragaglia, A., Heiter, U., Feltzing, S., Bensby, T., Irwin, M., Gonzalez Solares, E., Murphy, D., Bayo, A., Sbordone, L., Zwitter, T., Lanzafame, A. C., Walton, N., Zaggia, S., Alfaro, E. J., Morbidelli, L., Sousa, S., Monaco, L., Carraro, G., Lardo, C. (2020)
   *The* Gaia*-ESO Survey: Spectroscopic-asteroseismic analysis of K2 stars in* Gaia*-ESO: The K2 Galactic Caps Project*
   Astronomy & Astrophysics, 643, A83

# Contents

# 1. Introduction

Stellar spectroscopy is the field of astronomy that uses the spectra of stars as a tool for measuring their properties. These properties tend to be divided into on the one hand the *abundances* of different elements, and on the other hand the *stellar parameters*, such as temperature and surface gravity. These can in turn be used to make inferences about processes inside the stars.

The field is relatively old, having appeared thanks to some technological breakthroughs in the late 19th century. However, it has recently undergone several observational and theoretical advances, which have themselves been driven by technological development. One of the major observational improvements is the development of *multifibre spectrographs*. These allow the measurement of spectra from hundreds of stars in the same amount of observation time as was previously necessary to observe just one. This has greatly increased the amount of data available to spectroscopists. Two of the major theoretical improvements consist of increasingly transitioning from one-dimensional to three-dimensional models of stellar atmospheres, and loosening the assumption of Local Thermodynamic Equilibrium (LTE). These have also become possible by improvements in technology. While it has always been possible to write down such models on paper, it is only recently that computers have become fast enough for them to become computationally tractable.

That said, a scientific field does not only consist of observation and theory. An equally important part is the methods used to compare theory to observations. We believe that the development of such methods has somewhat lagged behind the observational and theoretical improvements. This is a problem for two reasons: Access to larger amounts of data means that the impact of statistical errors shrinks, while the systematic errors tend to stay the same. At the same time, improvements in theory means that we will be chasing increasingly subtle effects in our data. At worst, this can lead to purported discoveries that in reality are only measurements of the internal errors of the method used.

This problem is obviously too broad to tackle in its entirety. In this thesis, we will propose improved statistical methods for handling two common problems in stellar spectroscopy: Estimating the performance of a pipeline for estimating stellar parameters by fitting synthetic to observed spectra, and estimating a parameter based on measured stellar abundances. We will also explain how to use the techniques for broader classes of mathematically similar problems.

This line of argument will form one of the main threads running through this thesis. The seconds thread will concern the field of *Galactic archaeology*. This is the field of astronomy that attempts to reconstruct the history of the Milky Way. In this field stellar spectroscopy is one of the most widely used tools, since spectroscopic measurements can be used to distinguish populations of stars that otherwise appear similar. When used together with kinematic studies such as the Gaia survey, this can be used to infer the formation history of those stellar populations.

Since texts are usually easier to follow if they start with the concrete and then go on to the abstract, we will begin with outlining the field of Galactic archaeology in Chap. 2. Chapter 3 describes the spectroscopic techniques used in the articles included in the thesis. We attempt to do this in enough detail for it to be useful to another beginner spectroscopist. Chapter 4 describes the statistical techniques that we developed for the articles included in the thesis. In some cases we outline ways of generalising those techniques to seemingly very different research questions. Chapter 5 summarises the three articles included in this thesis. Chapter 6 summarises our conclusions, and gives a final outlook. Chapter 7 briefly lists my contributions to the articles included in this thesis. Chapter 8 gives a Swedish summary of the thesis. Chapter 9 tries to acknowledge everyone who has in some way contributed to the thesis. Chapter 10 contains two appendices, which discuss issues that have no logical place elsewhere in the thesis, but I found too interesting to cut out entirely. My own scientific work is mostly contained in Chaps. 4, 5, and 10, while Chaps. 2 and 3 are intended to provide background knowledge.

# 2. Galactic archaeology

This chapter attempts to outline the field of Galactic archaeology. Section 2.1 describes the current distribution of directly or indirectly observable matter within the Milky Way. Section 2.2 describes how we arrived at our current picture of the Milky Way, and the gaps in our knowledge. Section 2.3 describes the ongoing work to get a more detailed picture of Galactic history, using astrometric and spectroscopic studies.

## 2.1 The observable geometry of the Milky Way

We have a fairly good picture of the current shape of the Milky Way. Much of the Galaxy can be observed using telescopes, working either in the optical or some other wavelength range. There are some regions which we cannot see, for the simple reason that there is too much material in the way, but in practise this does not matter much: The regions we can see show a very clear symmetry, as do other galaxies, which means that we can assume that the hidden regions look qualitatively similar to what we can see. We conventionally divide the Galaxy into three main structures: the *Disk*, the *Bulge* and the *Halo*, shown schematically in Fig. 2.1 These can in turn be divided into several substructures (Schneider, 2015, Sect. 1.2.1).

The Disk is the approximately circular portion of the Galaxy, containing most of the stars. There are two different ways one could divide the Disk into sub-structures. The most visually obvious distinction is between the *spiral arms* and the *inter-arm region*, shown in Fig. 2.2. The spiral arms are believed to be temporary density waves in the rotating galaxy. That is, as stars orbit the Galaxy they will pass in and out of the spiral arms, but during their passage through an arm they slow down, and this is what creates the arm-shaped overdensity. The reason that the arms are brighter than the inter-arm regions is not only that the stars are more densely packed, but also because they are regions of star formation. Since the brightest stars are also the most short-lived, they typically die before the density wave passes, leaving the inter-arm region populated mostly by long-lived, faint stars (Schneider, 2015, Sect. 3.3.6).

While the spiral arms are visually striking, they are not of great interest to us as stellar spectroscopists, since they are a purely kinematic phenomenon. We are more interested in the distinction between the *thick disk*, *old thin disk* and *young thin disk*. This is not obvious from pictures of the Galaxy, but the distribution of stars can be described well in terms of three distinct groups.

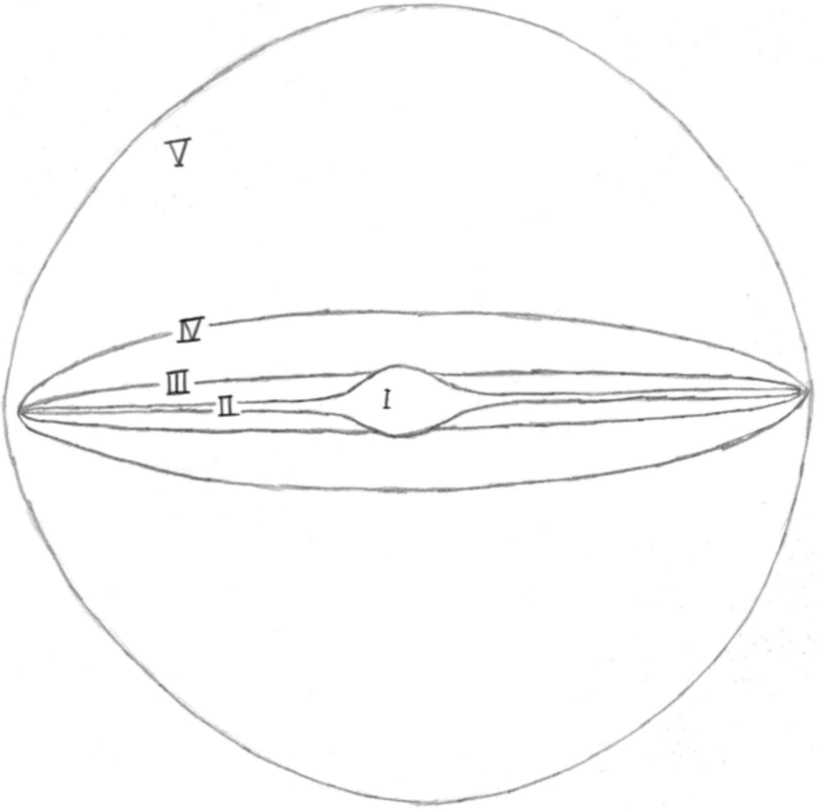

*Figure 2.1.* Schematic structure of the Milky Way galaxy, seen from the side. I: Bulge. II: Young thin disk. III: Old thin disk. IV: Thick disk. V: Halo. Dark matter halo not shown.

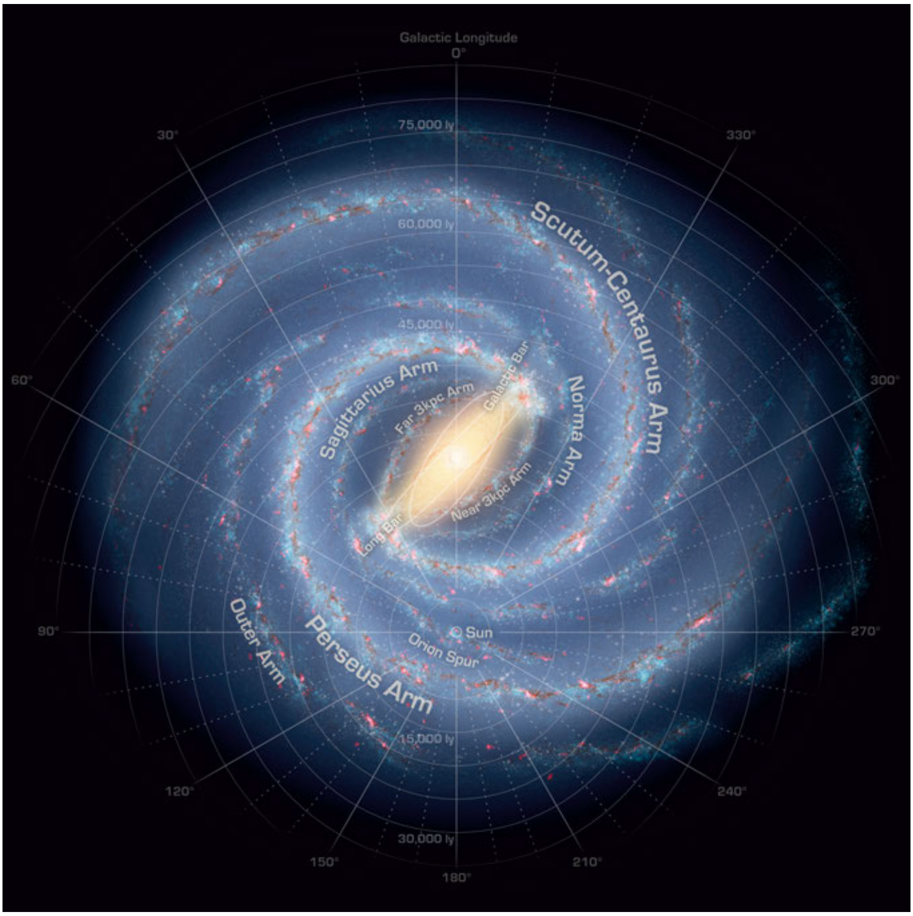

*Figure 2.2.* Structure of the Milky Way galaxy, seen from above. Image credit: NASA/JPL-Caltech/ESO/R. Hurt

Each one of these has a number density that drops off exponentially with the perpendicular distance $z$ from the Galactic plane. That is, they drop off as $\propto \exp(|z|/h)$, where $h$ is the *scale height* of the population. For the thick disk, the scale height is 1500 pc, while for the old thin disk it is 325 pc and for the young thin disk it is 100 pc (Schneider, 2015, Sect. 2.3.1). For reasons we will see, it is believed that the thick disk formed first, followed by the old and then the young thin disk.

The Bulge is the central part of the Galaxy. It is bar-shaped and is a bit thicker than the old thin disk, at a scale height around 400 pc. It is considerably redder than the rest of the Galaxy (Schneider, 2015, Sect. 2.3.5). Near the centre of mass is the supermassive black hole Sagittarius A* (Schneider, 2015, Sect. 2.6.3).

The Halo is an approximately spherical distribution of mass around the Galactic centre. It is fairly dim since it contains very few stars, but it does contain most of the Galactic mass, in the form of dark matter. It is conventionally divided into the *stellar halo*, which is made up of stellar clusters and lone stars, the *Galactic corona*, which is made up of hot gas, and the *dark matter halo* (Schneider, 2015, Sects. 2.3.6, 3.3.7, 7.6). We do not currently know what the latter is made of (Schneider, 2015, Sect. 1.1). ‘Dark matter’ is essentially a placeholder term for “something which we cannot observe, but which based on its indirect effect seems to behave mostly like matter”.

## 2.2 Known and unknown Galactic history

Currently, we understand the history of the Universe fairly well on the scale of stars, on the scale of atoms, and on scales larger than galaxies. On the scale of the Galaxy itself, we still only know the broad brush-strokes.

### 2.2.1 Stars

Stars form when molecular gas clouds undergo *gravitational collapse*: When the outwards force from the gradient in the gas pressure and the inwards force from gravity cancel out in a gas cloud, it is in *hydrostatic equilibrium*. This equilibrium is stable to small perturbations, as the gas pressure gradient and gravitational pull both increase if the cloud contracts, and decrease if it expands. However, there is a tipping point beyond which a cloud contracts so quickly that the increase in the pressure gradient cannot keep up with the increase in gravitational pull, since changes in gas pressure can only propagate at the speed of sound in the gas. Once this happens, a runaway process begins where the cloud keeps contracting until counteracted by some new source of outwards pressure[1]. The tipping point can be alternately be described in terms of the *Jeans mass* – the mass above which a spherical gas cloud with a particular temperature and particle density will collapse – or the *Jeans radius* – the radius below which the spherical gas cloud will collapse. Usually, this process is triggered by some external perturbation that compresses the cloud[2], but it can also happen through a cloud simply gradually cooling down, since the Jeans mass increases with temperature. This process of gravitational collapse only stops when individual clumps of gas in the cloud become so dense that they start undergoing *nuclear fusion*. That is, nuclei start to fuse, forming other nuclei that have more binding energy per nucleon. This releases energy which is emitted as electromagnetic radiation, including visible light. This creates a stronger pressure gradient, which halts the collapse. The radiation pressure from these newly-formed stars also starts to disperse what remains of the cloud (Carroll and Ostlie, 2014, Sect. 12.2).

The mass of a star determines its future life-span. As long as it has hydrogen to burn, it will remain in more-or-less hydrostatic equilibrium, as the outwards force from the gradient in gas pressure counteracts the inwards pull of gravity. But at some point, it will exhaust the supply of hydrogen in the regions hot and dense enough for fusion to happen (Carroll and Ostlie, 2014, Chapter 10). This will tend to be quicker the more massive the star is: For a $60\,\mathrm{M}_\odot$ solar-mass star it takes three million years, for a star like the Sun it takes ten billion years, and for a $0.1\,\mathrm{M}_\odot$ it takes in excess of the current age of the Universe (Bertulani, 2013, Table 1.1).

[1] Assuming it stops at all, that is. Supermassive black holes have formed through *direct collapse* of gas clouds in the early universe (Begelman et al., 2006).

[2] Hence the aforementioned increased star formation in the Galactic arms.

In stars above $0.25\,M_{\odot}$[3], the helium accumulates in the centre of the star, forming an inert core while hydrogen fusion shifts to a shell around the core. What happens near and around the core is necessary to understand the future development of those stars. Initially, the helium itself will not undergo fusion. This means that the core does not itself generate heat to counteract the pull of gravity. In stars above $1.8\,M_{\odot}$, the gas pressure gradient is enough to counteract gravity, but in the stars below the core will contract until the gravity is counteracted by the *electron degeneracy pressure* – the pressure due to Pauli's exclusion principle preventing two fermions from occupying the same quantum state (Carroll and Ostlie, 2014, Chap. 13).

For a star between 0.25-$8\,M_{\odot}$, the shift of hydrogen fusion away from the core causes the star to expand and become brighter but colder, turning it first into a *subgiant* and then into a *red giant* (Carroll and Ostlie, 2014, Chap. 13). If the stellar mass is above a threshold that lies somewhere in the range 0.25-$0.5\,M_{\odot}$, the core will at some point start to undergo helium fusion (Laughlin et al., 1997). In the stars light enough for the core to be degenerate, the material initially cannot expand in response to the rising temperature, which causes much of the core to ignite practically all at once in a *helium flash*. In heavier stars, the onset of helium fusion is a more gradual process. As the star consumes helium, it will form a core of oxygen and carbon, and fusion will shift to a shell around the core. The star will then shift from hydrogen fusion and very rapid helium fusion, becoming an *asymptotic giant branch star* (AGB star). During this variable phase, strong stellar winds will eject much of the material of the star. The outer layers will form a *planetary nebula* around the star, while the remaining core becomes a *white dwarf* – it will be small and initially very hot, but it will no longer sustain nuclear fusion and is held up by electron degeneracy pressure (Carroll and Ostlie, 2014, Chap. 13).

The stars lighter than about $0.25\,M_{\odot}$ have a less dramatic life cycle: Since convective flows move around material throughout the entire star they never form a distinct core. Hence, they keep burning hydrogen until the entire mass of the star turns into helium. During this process they are theoretically expected to first heat up, becoming *blue dwarfs* before cooling into white dwarfs. However, this has not yet been observed, since the process takes longer than the current age of the Universe (Adams et al., 2005, Sect. 3).

For a star above about $8\,M_{\odot}$, the end of the life cycle is considerably more dramatic, ending in some form of *core-collapse supernova* – a release of energy so sudden that it tears the star apart, possible leaving a small remnant behind. The pressure and temperature in the core eventually become high enough for the carbon to begin fusion, forming neon, sodium, magnesium and oxygen. In stars above about $10\,M_{\odot}$, gradual fusion will continue and produce

[3] All stellar mass thresholds in this section are fairly approximate. Different sources give slightly different numbers, depending on the state of the art of stellar evolution modelling when they were written.

heavier elements, forming a structure of nested shells in the core. However, this process cannot form elements heavier than iron in appreciable amounts: Above this atomic number, the binding energy per nucleon decreases[4], meaning that fusion consumes energy rather than releasing it. Instead, a degenerate iron core will form at the centre of the star. Once this passes a threshold mass, two things start to happen: There is widespread *electron capture* as electrons and protons start to merge into neutrons and neutrinos. There is also *photodisintegration*, as highly-energetic photons break apart massive nuclei, being absorbed in the process. Both of these remove much of the pressure that used to resist the gravitational pull, causing the core of the star to collapse. This usually results in the star being torn apart by an explosion, and usually leaves behind some kind of remnant (Carroll and Ostlie, 2014, Chap. 15). In a narrow mass range somewhere around $10\,\mathrm{M}_\odot$, the supernova will instead occur when the degenerate oxygen-neon-magnesium core starts to undergo electron capture, which causes it to collapse while undergoing rapid nuclear fusion (Nomoto and Hashimoto, 1986, Sect. VI).

The type of remnant depends on the mass of the remaining material, which depends on the original mass of the star, and to some extent on chance. Below 2.2-2.9$\,\mathrm{M}_\odot$, depending on the rotation speed, the remnant will form a *neutron star*. That is, the protons will merge with the electrons to form neutrons, causing the atomic nuclei to merge into a single mass of neutrons, supported by *neutron degeneracy pressure*, analogously to the electron degeneracy pressure in the white dwarf (Carroll and Ostlie, 2014, Chapter 16). If the mass is above the threshold, even this will be insufficient to support the star, and it will collapse into a *black hole* (Carroll and Ostlie, 2014, Chapter 17).

A common tool for visualising the lives of stars is the *Hertzsprung-Russell Diagram*[5] (HR diagram) – a scatter plot of stars with axes of luminosity and effective temperature[6]. We show an example in Fig. 2.3. On the HR-diagram, the positions of newborn stars of different mass forms a curve called the *main sequence* (Carroll and Ostlie, 2014, Sect 8.2). When they become red giants they move upwards to the *Red Giant Branch* (RGB) above (Carroll and Ostlie, 2014, Sect 13.2).

[4] It is sometimes stated that iron-56 is *the* element with the most binding energy per nucleon. This is not quite true: Nickel-62 has more, but it so happens that there are no reaction pathways that can produce it in large amounts (Shurtleff and Derringh, 1989).

[5] It should probably be called the *Rosenberg-Hertzsprung-Russell diagram*, considering Hans O. Rosenberg's important work, but even getting Ejnar Hertzsprung's contributions acknowledged took more than a decade of argument within the astronomical community, during which many would insist on only saying *Russell diagram* (Valls-Gabaud, 2014, Sect. 2).

[6] Sometimes, absolute magnitude is used instead of luminosity and $B-V$ colour instead of effective temperature, making a *colour-magnitude diagram* instead. Since there is a monotonous relationship between the quantities, this is essentially a rescaled version of the corresponding HR diagram.

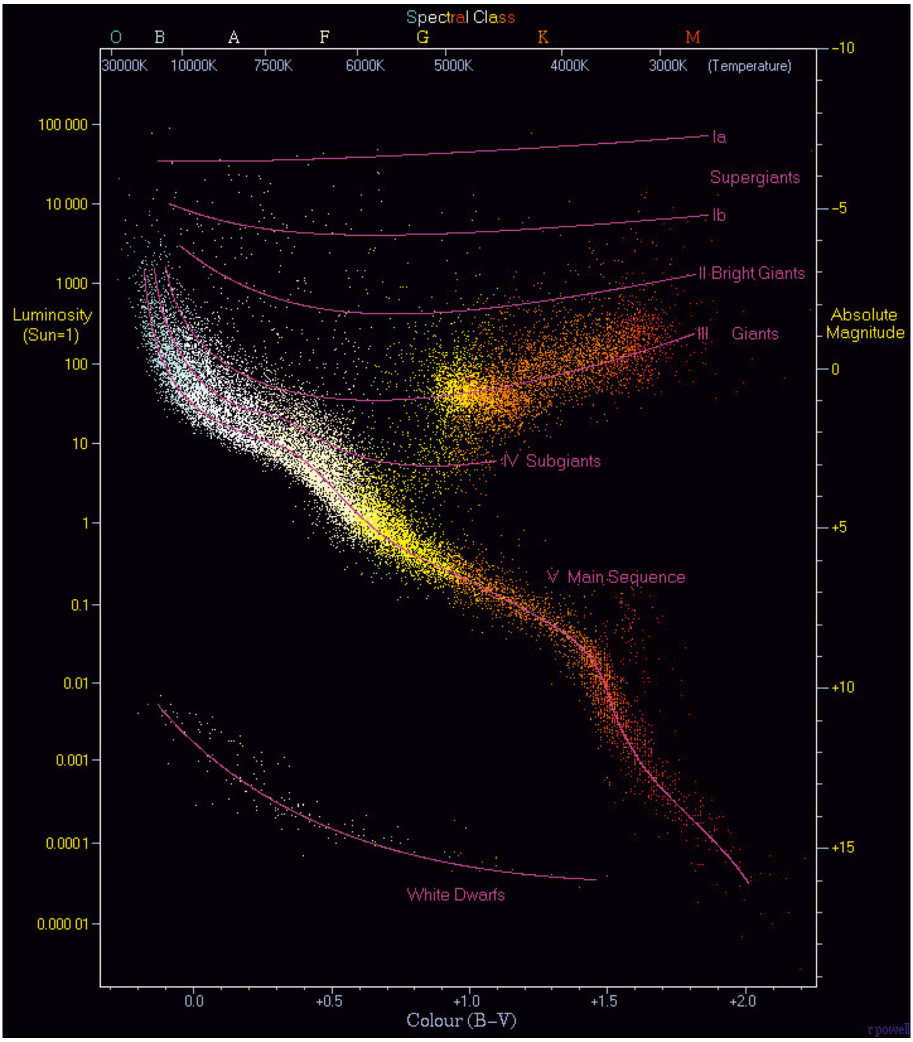

*Figure 2.3.* Example Hertzsprung-Russell diagram. Image credit: Richard Powell. Retrieved June 15, 2021 from Wikimedia Commons, `https://commons.wikimedia.org/w/index.php?curid=1736396`. Used under Creative Commons Attribution-Share Alike 2.5 Generic license.

### 2.2.2 Stellar clusters and stellar atmospheres

The group of stars created when a molecular gas cloud collapses and disperses is called a *stellar cluster*. We typically distinguish between the *globular clusters*, which formed early in galactic history, and the more recent *open clusters*. The globular clusters are usually found in the galactic halo, while the open clusters are more common in the galactic disk (Carroll and Ostlie, 2014, Sect 13.3). Open clusters typically disperse in less than 300 Myr since they are very loosely gravitationally bound (Schneider, 2015, Sect. 2.3.3).

The newborn stars will vary greatly in mass following a statistical distribution called the *initial mass function*. Since the stars have approximately the same age, they will approximately trace out a curve in the HR diagram called an *isochrone*. The isochrone will have a distinct hook somewhere at the hot end, marking the temperature above which all stars have already left the main sequence. This feature is called the *Turn-off Point* (TOP). Isochrones are sometimes used in the technique of *isochrone fitting* for estimating the ages of stellar clusters (Carroll and Ostlie, 2014, Sect 13.3). This technique can also be used to estimate temperature and other parameters describing the stars, as discussed in Sect. 3.5.1.

Aside from being approximately the same age, stars in a cluster typically start out chemically similar, since the gas in the cloud is fairly homogeneous. However, they do not start out chemically identical, and they become more distinct over time. The reason that they are not identical from the start – other than inhomogeneities in the cloud – is that the stars do not form exactly at once. There is enough time for short-lived members of the first generation of stars to die and eject heavier elements into the cloud. In globular clusters this creates the phenomenon of *anticorrelations*, where an excess in one element statistically correlates with a deficiency in another, since nuclear fusion transforms the latter element into the former (Cohen, 1978).

There are two processes that make the stars more distinct over time. The obvious one is the formation of new elements in the stellar core. However, this is uninteresting to us as spectroscopists, since it does not affect the stellar atmosphere, which is what we can actually observe. The only exception to this is the phenomenon of *dredge-up*, when convective flows reach so deep into the star that they start pulling up elements from the core. What mostly changes the stellar atmosphere is the tendency for elements to separate by depth over time: Heavier elements tend to sink – *gravitational settling* – and opaque elements tend to rise – *radiative acceleration*. These tendencies are counteracted by the convective motions in the atmosphere, which mix up the chemical composition again. The relative effectiveness of these processes depends on the mass of the star, which means that in a stellar cluster the observed elemental abundances will vary with the temperature, or some other proxy for mass (Michaud, 1970; Richard et al., 2001).

However, stellar evolution models which take all of these processes into account are not able to reproduce the abundances actually observed in stellar clusters. They invariably predict variations in abundance that are much larger than those actually seen. There appears to be some *Additional Transport or Mixing Process* (AddMix) that evens out the abundances even more than convection can do on its own. This process is not well understood[7] but it is possible to reproduce observed abundances fairly well by postulating that it takes the form of some kind of turbulent diffusion, with a diffusion coefficient $D_T$ such that

$$D_T \equiv \omega D_{\mathrm{He}}\left(T_0\right)\left(\frac{\rho\left(T\right)}{\rho\left(T_0\right)}\right)^{-n} \tag{2.1}$$

where $\omega$ is some real number, $T_0$ is some reference temperature $D_{\mathrm{He}}\left(T_0\right)$ is the atomic diffusion coefficient of helium at the temperature $T_0$, $\rho\left(T\right)$ is the density at some temperature $T$, and $n$ is some integer[8] (Richer et al., 2000; Richard et al., 2005). So far, it has been possible to constrain $n$ to 3 and $\omega$ to 400 (Proffitt and Michaud, 1991; Richard et al., 2002, 2005). In Article II we attempt to constrain $T_0$ for the cluster M30, and find that it is somewhere in the range $\log_{10}\left(T_0/\left[\mathrm{K}\right]\right) = 6.09$-$6.2$.

### 2.2.3 Elements

The matter in the early Universe was almost exclusively hydrogen, helium and lithium. It was formed through *Big Bang nucleosynthesis* (BBN): The collision of elementary particles in the first minutes after the Big Bang. While it might seem difficult to model events very soon after something as dramatic as the Big Bang, the early Universe was actually a fairly uncomplicated place, consisting only of a homogeneous, isotropic gas. It is only afterwards that complicated structures have developed. Because of that, the processes behind BBN are mostly well understood, and current models can reproduce observed abundances relatively closely (Coc and Vangioni, 2017).

The exception to this is lithium. Attempts at measuring primordial lithium abundances consistently end up with values around a factor of 3 below those predicted by BBN models. This is known as the *cosmological lithium problem*. There are many proposed solutions, which fall into four categories that are not mutually exclusive. Some solutions propose that our observations are simply flawed, so that measurements of lithium have systematic errors which make them appear lower than they are (Fields, 2011). For example, Wang et al. (2021) looks at how much lithium abundance estimates are affected by

[7] We describe it with a deliberately vague name to reflect the fact, but some authors use *turbulent mixing* instead.

[8] Note that this formula is more general than it might seem: One could drop the interpretation of $D_{\mathrm{He}}\left(T_0\right)$ as having anything to do with diffusion of helium and simply write Eq. (2.1) as $D_T = k\left(\rho\left(T\right)/\rho\left(T_0\right)\right)^{-n}$, where $k$ is some quantity with dimensions of length$^2$/time.

the approximations of one-dimensional stellar model atmospheres and Local Thermodynamic Equilibrium (LTE)[9]. The next two categories propose that our BBN models are flawed. At the simplest, parameters in the models could be poorly measured. For example, Cyburt and Pospelov (2012) look at a poorly-constrained nuclear energy level in boron-9, which could contribute to enhancement of reactions that destroy beryllium-7, which otherwise produces lithium. At the more intriguing, it could be that accurate modeling requires taking into account physics beyond the standard model. For example, Jedamzik and Pospelov (2009) look at how different models of dark matter could introduce reaction pathways that destroy lithium or prevent it from forming. The fourth category postulates that the measurements of current abundances and BBN model predictions of primordial abundances are both correct, but that some astrophysical process has since changed the abundances from the primordial values. For example, Piau et al. (2006) proposes that Pop III stars destroyed much of the primordial lithium.

This thesis does not attempt to resolve the cosmological lithium problem, but we touch upon it in Article II. This article is part of the article series *Atomic diffusion and mixing in old stars*. The article Gruyters et al. (2016) – the sixth paper in that series – did study lithium and estimated primordial abundances of $2.48 \pm 0.10$ dex in the globular cluster M30, where BBN models would predict 2.72 dex. However, it did so assuming AddMix with $\log_{10}(T_0/[\mathrm{K}]) = 6.0$. Since our Article II also investigated M30 and found a higher value of $T_0$, we had to revise the estimated primordial abundance to 2.42-2.46 dex, making the measured discrepancy even worse.

Practically all elements heavier than lithium have since been formed through *stellar nucleosynthesis*: As described above, the fusion processes in the stellar core can create elements up to and including iron. The heavier elements are created by *neutron capture*, as nuclei absorb neutrons. Depending on whether the neutron flux is strong enough for unstable nuclei to decay between neutron captures, we distinguish between the *rapid neutron-capture process* (r-process) and the *slow neutron-capture process* (s-process) (Carroll and Ostlie, 2014, Sect 16.3). r-process elements are probably created in neutron-star mergers and some supernovæ while s-process elements are mostly created in AGB stars. In both cases, the deaths of the stars cause the elements to be injected into the interstellar medium (Freiburghaus et al., 1999; Heiter, 2017).

### 2.2.4 Galaxies

In general, galaxies are believed to form through the gravitational collapse of clumps of dark matter attracting primordial gas created in the Big Bang. This gas invariably has some angular momentum to begin with – it would be an extreme coincidence for the net moment of any large lump of matter to be

[9] We discuss these approximations in more detail in Sect. 3.8

exactly zero – and as it contracts the moment of inertia decreases, forcing the angular velocity to increase in order to conserve angular momentum. As the material starts to spin it acquires a disk shape, through the same process as flattens spun dough into a pizza. Since the dark matter does not interact by anything other than gravity, it keeps its original shape and forms a dark halo around the galaxy (Schneider, 2015, 10.4.1).

As described above, the primordial gas starts out consisting almost only of hydrogen, helium and lithium, but gradually gets enriched with heavier elements as successive generations of stars form and die.

Several simulations have been made that attempt to describe this process from first principles, such as the Millenium Run and the Illustris project. These simulations have managed to reproduce the large-scale structure of the Universe fairly well. They even lead to the formation of structures that are clearly galaxy-like. Unfortunately, while they often have some kind of spiral structure, the detailed morphology looks very different from any actual galaxy (Lemson and Virgo Consortium, 2006; Vogelsberger et al., 2014; Genel et al., 2014)

### 2.2.5 The Milky Way

The Milky Way is believed to have formed through the same process as other Galaxies. As the primordial gas started undergoing gravitational collapse, it formed the first generation of stars, the *Population III stars* (Pop III) (Schneider, 2015, Sect. 10.3.2). So far, none of these stars have been observed, which is a natural consequence of the early Galaxy being both hot and metal-poor: The Jeans mass increases with temperature, meaning that the only stars that could form were very massive. When that first generation of stars died, they seeded the interstellar medium with elements, which meant that the next generation, the *Population II stars* (Pop II), were formed from gas that was metal-poor but not entirely free of metals (Bromm et al., 2009). Many of these stars still exist, but those that have since died went on to enrich the Galactic medium with more metals. From this gas the *Population I* (Pop I) stars formed, to which the Sun belongs (Schneider, 2015, Sect. 2.3.2).

This happened concurrently with the flattening of the disk, so that Pop II is mostly found in the Halo, while Pop I is mostly found in the disk. In addition to that, the stars in the Thin Disk are more metal-rich than those in the Thick Disk. The reason for this is not completely clear: The straightforward explanation would be that there was a gradual flattening of the gas component of the Galactic disk, so that the Thick Disk stars have their current greater velocity dispersion because they were originally created from gas with a greater velocity dispersion, but this may not be the case. It could be that the Thick Disk stars originally had trajectories similar to the current Thin Disk, but that over time they have been scattered by more close encounters with other stars, which most Thin Disk stars have not had the time for yet. It is also possible that the Thick Disk was formed by several dwarf galaxies being absorbed by the early Galaxy (Schneider, 2015, Sect. 2.3.2). It may even have been formed by the absorption of a single dwarf galaxy dubbed *Gaia-Enceladus* (Helmi et al., 2018).

This gives us a general picture of Galactic history, but note that it is still fairly vague. Based on the information given here, an astronomer could make some very general inferences, such as: The more metal-poor a star is, and the more it moves transversely to the Galactic plane, the more likely it is to have formed early, and vice-versa. To get a more detailed understanding, we need to make use of several complementary observational techniques, which are described in the next section.

## 2.3 Inferring the details of Galactic history

In principle, given only the positions and velocities of the stars in the Milky Way, it should be possible to go quite far in reconstructing the history of the Galaxy. After all, given this dynamic information one can simply describe the Galaxy as a large number of point masses following Newton's laws, meaning that it should be possible to 'rewind' Galactic history at least some distance in time. The main problem would seem to be that far enough back, stars 'collide' in the gas clouds that birthed them, which cannot be described purely as an interaction of point masses. In practise, even within the regime where stars can be described as point masses, this approach would very quickly run into difficulties: It turns out that, for most initial conditions, a system of more than two gravitationally interacting bodies is a *chaotic* system. That is, it is a system such that if the state of the system is known at time $t_0$ with some uncertainty, then an extrapolation forward or backwards to some time $t$ will have an uncertainty that is exponential in $|t - t_0|$. This means that it is practically impossible to make measurements exact enough to allow meaningful extrapolations outside a fairly narrow span of time (Goldstein et al., 2002, Chap. 11). Hence, some additional constraint is needed.

One such constraint is given in the form of chemical information. When stars are formed, they form in clusters, from material that is more-or-less chemically homogeneous. These clusters tend to gradually disperse, but the chemical similarities mostly remain, which can be used to identify their common origin. This is not completely foolproof, though: There are occasional cases of *doppelgänger* stars, which by chance happen to be chemically very similar to one another (Ness et al., 2018).

### 2.3.1 Astrometric surveys

Large-scale surveys measuring both 3D positions and 3D velocities of stars appear surprisingly late in the history of astronomy, given that the scientific question is fundamentally just "Where are the stars and where are they going?" There turn out to be basic technical reasons why this is so hard to do. Purely positional surveys in 2D have been made through all of human history – they began the moment anyone in pre-history noticed that the sky looks similar from night to night, and started memorising the positions of the stars – but it is very difficult to measure either the distance to a star or its velocity: In principle, estimating the distance to a star is just a matter of trigonometry. Given a well-defined reference frame, if one measures the direction to the star at one point in time and waits half a year and repeats the measurement, the angle between the two lines and the perpendicular distance traversed between the measurements is enough to calculate the distance to the star. In practise, this angle is so small that prior to the 20th century it was only possible to do for a handful of very nearby stars (Carroll and Ostlie, 2014, Sect. 3.1). Once the distance is known, measuring the transversal component of the velocity of a star is in principle just a matter of observing how the direction from a particular point in the Earth's orbit to the star changes over time – *the proper motion* (Carroll and Ostlie, 2014, Sect. 1.3). The radial component of the velocity can be estimated by observing the Doppler shift in the stellar spectrum, which is simple as long as the stars are bright enough for spectral lines to be visible (Carroll and Ostlie, 2014, Sect. 5.1).

From the ground, it is almost impossible to measure radial distance and transverse velocity for stars that are not close to the Earth. Irrespective of the technology used, atmospheric effects practically put a hard limit on how small angles can be resolved. Hence, large-scale kinematic surveys of 3D positions and 3D velocities have only become possible through the use of space telescopes (Gaia Collaboration, 2016b, Sect. 1).

The first large-scale space-based survey measuring position in 3D and velocity in 2D ran in the years 1989-93, when the *HIgh Precision PARallax COllecting Satellite* (HIPPARCOS) was launched by ESA. This led to the publication of two star catalogues: The HIPPARCOS Catalogue, which gives precise 3D positions and 2D velocities for 118200 stars; and the Tycho-2 Catalogue, which gives less precise data for 2.5 million stars[10] (Perryman et al., 1997; Høg et al., 2000).

The first, and so far only, space-based survey measuring both position and velocity in 3D began in 2013, with the launch of the Gaia[11] space observatory (Gaia Collaboration, 2016b). At the moment of writing, this mission is still ongoing, and is expected to continue for as long as Gaia is able to keep functioning – it has already exceeded its initial expected lifespan, but in 2024 it will have to shut down for lack of fuel.

The data collection of Gaia is a gradual process. It is not the case that during one month, Gaia makes an authoritative measurement of the positions and velocities of a certain number of stars, and the next month it goes on to measure a different set of stars. Rather, Gaia is constantly rotating, and as its field of view sweeps across the sky it incrementally improves its positional data for the stars covered. With a single observation of a star, it is impossible to say anything more than the 2D position. After two observations, the distance and transversal motion are still degenerate. But as the number of observations increases, the error bars on those parameters shrink and the correlations between them will decrease. Over time, more and more stars move above the quality threshold where the data can be used in scientific applications (Gaia Collaboration, 2016b).

[10] This supercedes the previous Tycho Catalogue containing a subset of 1.0 million out of those stars (Høg et al., 1997).

[11] The name is formally not an acronym: It was originally derived from *Global Astrometric Interferometer for Astrophysics*, during early planning stages when it seemed that an interferometric method might be used. When the decision was made not to use interferometry, the name was kept.

As data has been collected and analysed, there has been a steady output of public data releases. Data Release 1 (DR1) of 2016 contained one sample of Gaia-only data, and a sample which combined information from Gaia and the Tycho-2 catalogue. The Gaia-only data gave 2-dimensional positions for a billion objects, while the Gaia-Tycho data gave 3D-positions and 2D-velocities for 2 million objects (Gaia Collaboration, 2016a). Data Release 2 (DR2) of 2018 is Gaia-only, and contains 3D-positions and 2D-velocities for 1.3 billion objects and fully 3D data for a subset of 7 million of them (Forveille, Thierry et al., 2018; Gaia Collaboration, 2018, a). The Early Data Release 3 (EDR3) of late 2020 contains fully 3D data for 882 million objects and only 3D-positions and 2D-velocities for another 585 million (Gaia Collaboration, 2021, b). At the time of writing, the full Data Release 3 (DR3) is expected to be finished in the middle of 2022.

### 2.3.2 Spectroscopic surveys

So far, the only way of studying the chemical composition of stars other than the Sun is through *spectroscopy*. We describe the technical details in Sect. 3, but the basic principle is fairly uncomplicated: The spectrum of a star is to first approximation a black body curve, since the deeper regions of the atmosphere are dense enough to function like a black body. Overlaid on this black body curve are absorption lines, coming from atoms, ions and molecules in the less dense region of the outer atmosphere. These absorption lines have wavelengths that are unique to the species in question. The overall shape of the line is a function of the number density of that species – as well as the temperature and other physical properties of the star, collectively called the *stellar parameters*. This means that if the stellar parameters are known, it is in principle possible to use one or more absorption lines of a species to estimate the *abundance* of that species. The stellar parameters can in turn be estimated through several methods, including spectroscopy.

Unlike kinematic observations, high-quality spectroscopy can be made from the ground. Hence, there have been multiple large-scale spectroscopic surveys. During 2003-13, the RAdial Velocity Experiment (RAVE) used the Six Degree Field (6dF) multiobject spectrograph at the Anglo-Australian Telescope (AAT) of the Anglo-Australian Observatory/Australian Astronomical Observatory (AAO)[12] to observe more than 400000 stars. This was in some strict sense a kinematic survey, since the main goal was to use the Doppler shift of the spectra to estimate radial velocities, but as a happy side effect it was possible to use the spectra to estimate stellar parameters and abundances (Steinmetz et al., 2006). As of 2021, the survey is on its sixth data release (Steinmetz et al., 2020b,a). During 2011-14, the Apache Point Observatory Galactic Evolution Experiment (APOGEE) used the Sloan 2.5 m Telescope to observe spectra of

[12] The name was changed in 2010, preserving the acronym.

146000 stars. Unlike RAVE, this survey was dedicated specifically to estimating chemical abundances (Majewski et al., 2017). Starting in 2013, the ongoing GALactic Archaeology with HERMES (GALAH) has been using the High Efficiency and Resolution Multi-Element Spectrograph (HERMES) on the AAT to observe more than 500000 stars (Freeman et al., 2013). At the time of writing, the survey is on its third data release (Buder et al., 2021).

During 2011-2018 Gaia-ESO Survey (GES)[13] has been complementing the astrometric Gaia survey with ground-based observations (Gilmore et al., 2012; Randich et al., 2013). It uses the spectrographs GIRAFFE and UVES, which we describe in more detail in Sect. 3.9. In brief, UVES has better resolution than GIRAFFE, but requires more time to observe the same number of spectra.

The GES spectra are not analysed by a single method. Instead, for either spectrograph there are several *nodes*, each tasked with estimating stellar parameters[14] and abundances for the spectra. These estimates are then combined in a way that uses scatter between and internal to nodes to estimate their reliability, and then uses this to estimate the uncertainty in each final abundance estimate. The underlying assumption is that since the major source of errors in spectroscopically determined quantities is not noise, but systematics due to choices in the analysis method, the only way to robustly estimate the uncertainties in these quantities is through consulting multiple spectroscopists and looking at the scatter between their results.

The author participated in the Internal Data Releases 5 and 6 (iDR5, iDR6) of GES as part of the Lund-Uppsala-MPIA-Bordeaux-ANU (LUMBA) node. The LUMBA pipeline for estimating stellar parameters is described in Article I, which is summarised in Sect. 5.1. We use a branch of the pipeline for estimating abundances in Article II, which is summarised in Sect. 5.2. The abundance pipeline is not described in quite as much detail, since it is conceptually similar but simpler in implementation: The stellar parameter pipeline attempts to estimate six free parameters based on several spectral lines, while the abundance pipeline uses a single spectral line at a time to estimate two free parameters, an abundance and a generic broadening parameter.

[13] Note that this abbreviation is mostly used internally within the Gaia-ESO Survey, and rarely appears in publications.

[14] See Sect. 3.2 for a detailed explanation of what stellar parameters are.

# 3. Stellar spectroscopy

This chapter attempts to outline the field of stellar spectroscopy. It is written to be useful to somebody who is just starting out as a spectroscopist. Hence, it covers the basic theory, but occasionally jumps into technical details such as commonly encountered problems in spectroscopy or commonly used tools for spectroscopy.

Section 3.1 describes how stellar spectra form in the first place. Section 3.2 describes the parameters typically used in models of stellar atmospheres and their resulting stellar spectra. Section 3.3 explains what the spectral continuum is, and how spectra are normalised to the continuum. Section 3.4 describes how stellar spectra are used to estimate the stellar abundances of elements. Section 3.5 describes how stellar spectra are used to estimate the stellar parameters described in Sect. 3.2. Section 3.6 describes common practical issues that tend to appear during attempts at making the estimates described in Sects. 3.4 and 3.5, and how we dealt with them in the articles included in the thesis. Section 3.7 describes the basic principles behind models of stellar atmospheres and their resulting stellar spectra, and how they are used in the technique of spectral fitting. Section 3.8 describes some common approximations that are typically made when calculating stellar spectra. Section 3.9 describes the spectrographs UVES and GIRAFFE, which were used in Articles I and II. Section 3.10 describes the most important software tools that were involved in making the articles in this thesis.

## 3.1 The formation of stellar spectra

Stellar spectroscopy studies the *spectra* of the light[1] coming from stars. For our purposes, a spectrum is any measurement of the intensity of the light as a function of wavelength, possibly subject to some normalisation factor that may change slowly with the wavelength. Stellar spectra contain information about the star, since they are shaped by the processes in the stellar atmosphere.

In a single atom[2], the electrons can inhabit certain discrete energy levels. Transitioning between two energy levels requires the atom to either absorb or emit a photon, with an energy corresponding to the difference between the energy levels. This means that in isolation, an atom will be able to emit and absorb light of wavelengths that are specific to that element. A collection of atoms which are relatively widely separated, such as a diffuse gas, will still tend to absorb and emit light at those energy levels. Hence, if light with a continuous spectrum shines through the gas, this will leave *absorption lines* at those wavelengths. At these lines the intensity typically does not drop to zero. Instead, the lines will have a depth and width that correspond to the temperature, density and other properties of the gas. As the abundance of an absorber increases, a line will pass through the regime of being *weak*, *saturated* and *strong*. In the weak regime, the line is narrow, and responds to increases in abundance mostly by becoming deeper. In the strong regime, the line has reached its maximum depth, and responds to increases in abundance mostly by becoming even wider. In the saturated stage in between, the line responds only weakly to changes in abundance (Gray, 2008, Chaps. 1 & 13).

In addition to this *line absorption*, two types of *continuous absorption* appear once the energies involved are high enough. If the light is sufficiently energetic, it will be able to detach electrons entirely. This *bound-free absorption* affects all wavelengths shorter than a threshold given by the minimum amount of energy needed. While a certain amount of energy is needed for this to happen in the first place, once there is an appreciable number of free electrons moving around in the gas, they will start to attach to atoms and form ions. In many of these, such as $H^-$, the extra electron is very loosely bound, meaning that bound-free absorption will start to affect much less energetic light. In addition the free electrons themselves will absorb light, creating *free-free absorption* (Gray, 2008, Chap. 8).

Everything said so far applies as long as the atoms are sufficiently widely separated that they can be treated in isolation. If two atoms are put in close proximity, the energy levels will split, allowing them to emit and absorb light of a larger number of discrete wavelengths. If very many atoms are put in close proximity, the energy levels become so numerous that they are practically impossible to distinguish. Hence, a dense lump of matter will be able to emit or absorb a continuous range of wavelengths (Holgate, 2009, Sect. 6.2.2). Using

[1] For brevity, I will say ‘light’ rather than ‘electromagnetic radiation’ throughout this chapter.

[2] For brevity, I will say ‘atom’ rather than ‘atom or ion’, except when the distinction is important.

statistical mechanics, it is possible to show that the emitted light will follow a *black body spectrum*: a spectrum with a very specific shape which is determined by the temperature of the material. The peak of this spectrum will be pushed towards shorter wavelengths as the temperature increases, at the same time as the intensity increases at every wavelength (Gray, 2008, Chap. 6).

In a stellar atmosphere, both of these occur at different depths in the atmosphere. In the deeper regions of the star, the material is dense enough to emit light following a black body spectrum, which is reshaped by both line and continuous absorption as it passes through the less dense outer layers of the stellar atmosphere. The position and shape of these spectral lines in principle contains information which elements exist in the star, as well as in what amounts. Stellar spectroscopy is the craft of actually extracting that information as accurately as possible.

## 3.2 Stellar parameters

In principle, a stellar spectrum is a function of everything that happens anywhere in the outer layers of the star. In practise, it is possible to replicate stellar spectra with models that only require the elemental abundances and a handful of *stellar parameters* as their input. In our work, we use a set of seven stellar parameters. Note that more complex models do not necessarily require more parameters: As we discuss in more detail in Sect. 3.8, the parameters below describing microturbulence and macroturbulence are necessary in one-dimensional models of the stellar atmosphere, but disappear in three-dimensional models.

### 3.2.1 Effective temperature

The *effective temperature* is defined as the temperature of a perfect black body with the same power output per unit area as the star. Intuitively, this can be thought of as a way of expressing whether a star is hot or cold overall, even though the temperature varies from place to place. As a rough rule of thumb, the spectral lines tend to be deeper and wider in colder stars, and shallower and narrower in hotter stars. For most lines, there is some temperature above which the line becomes so weak that it is effectively impossible to measure. As a shorthand for this, we tend to say that lines ‘disappear’ at some temperature, and that hotter stars have ‘fewer lines’ than colder stars.

Over the range of stars discussed in this thesis, roughly 4000-8000 K, this has the effect that very hot stars and very cold stars are both difficult to study. In the hottest stars there are so few visible lines that there is almost no information there to work with. In the coldest stars the lines are instead so many and so broad that most wavelengths are covered by multiple overlapping lines, which makes it difficult to untangle what is actually seen.

We denote the effective temperature with $T_{\mathrm{eff}}$ and express it in units of Kelvin. Note that some authors prefer to look at the 10-logarithm of the effective temperature divided by Kelvin. Since the temperatures of the stars we study are within a factor of 2 of each other, there is no reason for us to do that.

### 3.2.2 Surface gravity

The *surface gravity* is defined as the acceleration due to gravity somewhere in the outer layers of the star. The exact position is obviously a bit arbitrary – stars have no clearly defined surface – but for the stars we deal with in this thesis the transition region between the stellar atmosphere and the vacuum of space is thin enough that it does not matter.

We denote the effective temperature with

$$\log g \equiv \log_{10}\left(a/\left[\mathrm{cm/s^2}\right]\right) \tag{3.1}$$

where $a$ is the acceleration due to gravity. Strictly speaking $\log g$ is a dimensionless quantity, but it is conventional within spectroscopy to speak of any quantity calculated by taking a dimensionful quantity, dividing it by some unit, and then taking the 10-logarithm, as though it had a unit called *dex*.

### 3.2.3 Metallicity

The *metallicity* is a measure of the fraction of the star made up of elements other than hydrogen and helium. Note that strictly speaking this is not a stellar parameter but an average over abundances, but it is treated as a stellar parameter in most applications.

We denote it as

$$[\mathrm{M}/\mathrm{H}] \equiv \log_{10}\left(\frac{N_\mathrm{M}}{N_\mathrm{H}}\bigg/\frac{N_{\mathrm{M},\odot}}{N_{\mathrm{H},\odot}}\right) \tag{3.2}$$

where $N_\mathrm{M}$ denotes the number density of metals in the stellar atmosphere, and $N_\mathrm{H}$ denotes the number density of hydrogen. $N_{\mathrm{M},\odot}$ and $N_{\mathrm{H},\odot}$ denote the corresponding quantities in the Sun. Like the surface gravity, this is referred to as having units of dex. Note that by this definition $[\mathrm{M}/\mathrm{H}] = 0$ exactly for the Sun.

Note that sometimes [Fe/H] is used instead to denote the metallicity, rather than specifically the abundance of iron. This is done in SME (described in Sect. 3.10.2) as well as in Article I.

### 3.2.4 Macroturbulence

The *macroturbulence* parameter is used in one-dimensional models of stellar atmospheres to capture the effects of *some kind* of large-scale motions in the stellar atmosphere. The exact physical interpretation is still controversial, enough so that we defer discussion of the matter until Sect. 3.8. The observable effect of the macroturbulence is fairly simple: Features of the spectrum are broadened by a *radial-tangential* kernel, which assumes Doppler shift from a Gaussian distribution of velocities in the atmosphere.

We denote the macroturbulence with $v_\mathrm{mac}$, with units of km/s.

### 3.2.5 Microturbulence

Analoguously to the macroturbulence parameter, the *microturbulence* parameter is used in one-dimensional models to describe some kind of small-scale motion. In Sect. 3.8 we discuss different interpretations of what physical phenomenon it actually is that the parameter captures the effects of. Whatever the origin, microturbulence affects the overall line shape in ways that cannot be modelled as a simple convolution.

We denote the microturbulence with $v_{\mathrm{mic}}$, with units of km/s. Note that some other authors prefer $\xi$ instead.

### 3.2.6 Projected rotational velocity

Most stars rotate to some extent. This only affects the radiation coming from a particular point on the stellar surface by Doppler-shifting it. Overall, however, it causes the features of the stellar spectrum to be more smeared out, as it becomes a superposition of spectra from points moving towards and away from the observer. The strength of the effect on the observed spectra depends on the inclination between the rotational axis of the star and the line-of-sight to the observer – the effect is strongest if the equator of the star is closest to us, and non-existent if a pole is closest.

We denote the projected rotational velocity with $v \sin i$, with units of km/s. Note that some authors prefer $v_{\mathrm{rot}}$ instead.

### 3.2.7 Radial velocity

As discussed in Sect. 2.3.1, the radial velocity of stars is usually measured using spectroscopic methods, since it has the effect of uniformly Doppler-shifting the spectrum. In a strict sense this is not a stellar parameter, since it is not actually an intrinsic property of the star but an effect of the relative motion of the Earth and the star, but it is convenient to treat it as a stellar parameter.

We denote the radial velocity with $v_{\mathrm{rad}}$, with units of km/s.

## 3.3 Continuum normalisation

In this thesis we mostly study stellar spectra observed with the echelle spectrographs UVES and GIRAFFE, which are described in Sect. 3.9. The intensity measured with an echelle spectrograph is equal to the true intensity multiplied by some factor, which varies slowly with the wavelength. Before it is possible to use a spectral line for any kind of analysis, it is necessary to compensate for this – to *normalise* the spectrum. Typically this is done by shifting the measured intensity so that the intensity which would have been there in the absence of spectral lines – the *continuum level* – is set to 1. This seemingly trivial technical problem turns out to be surprisingly difficult to solve accurately, and inaccuracies can have a strong impact on all subsequent analysis.

In the articles discussed in this thesis, continuum normalisation is done by first throwing away the parts of the spectrum that are not actively used in the analysis, keeping around each wavelength region of interest a segment which is short enough that the continuum level is approximately linear. Then two regions are selected on either side of each segment, in which the flux is thought to be close to the continuum. A straight line is fitted to those points, and the intensity is shifted so that the line has a constant intensity of 1. In some situations, the assumption of local linearity breaks down, which we discuss in Sect. 3.6.2.

In Sect. 4.6 we describe an algorithm we developed for Article I for selecting the wavelengths to be used in the continuum normalisation. In Sect. 4.2, we describe the analysis we did for Article II of the impact of errors in the continuum normalisation.

## 3.4 Determining abundances

Within the context of Galactic archaeology, the stellar abundances are typically the quantities of interest. Estimating stellar abundances is also relatively simple, but requires that somebody has already estimated the stellar parameters – which we describe how to do in the next section.

If the stellar parameters are known, the shape of a spectral line is usually an invertible function of the corresponding abundance: If the abundance increases, the line will deepen and widen. The simplest, and oldest, method is to simply look at the drop in integrated normalised flux of the line: the *equivalent width*. With the improvement of models and computer hardware it has become possible to attempt to estimate abundances by fitting a model spectrum to the observed spectrum, so-called *spectral synthesis*.

In itself, determining stellar abundances from spectral lines is relatively simple. However, there are many practical problems that can arise, which we describe in Sect. 3.6. There is also the question of how to actually use the abundance estimates afterwards. In Sect. 4.2 we describe the statistical framework that we had to develop to use abundances to answer the science case of Article II, which required us to take into account the systematic errors.

In some situations it is not possible – or simply not necessary – to know the abundances of individual elements. It may be sufficient to determine the overall metallicity [M/H], which we defined in Sect. 3.2.3, or the *alpha abundance*. The latter is denoted $[\alpha/\text{Fe}]$ and describes an average abundance over the elements oxygen, magnesium, silicon, sulphur, calcium, and titanium, which are mainly produced by the *alpha-capture process* – absorption of helium nuclei. In Article III, which we summarise in Sect. 5.3 we describe a method for estimating [M/H] and $[\alpha/\text{Fe}]$ from Gaia spectra, which have too poor resolution for individual spectral lines to be visible.

## 3.5 Determining stellar parameters

Estimating stellar parameters is typically a more complicated process than estimating abundances, since very few spectral lines are sensitive to one parameter alone. It is necessary to either study several lines at once, after selecting them so that the degeneracies are broken, or to avoid spectroscopic methods altogether.

All lines are sensitive to $T_{\rm eff}$, in ways which are not degenerate with the other parameters. Hence, it is usually not necessary to look for lines that are specifically sensitive to that parameter. In fact, it may be inadvisable – in Sect. 3.6.2 we describe a failed attempt at using specifically $T_{\rm eff}$-sensitive lines.

Strong metal lines are typically sensitive to $\log g$. However, they are also sensitive to the abundance of the element in question, meaning that weak lines have to be included in the analysis as well to break the degeneracy. In Paper I we use calcium and magnesium lines for this purpose. In addition, we use singly ionised iron lines, which are also sensitive to $\log g$.

[M/H] can be estimated by looking at iron lines, since iron abundance is usually a good proxy for the abundances of all elements. Neutral iron lines are only sensitive to [M/H] while ionised iron lines are sensitive to $\log g$ as well. That said, in the context of pure abundance estimation it is usually not necessary to determine [M/H]. As long as the lines used in the abundance determination are not blended, the overall metallicity is not very important.

Strong lines are sensitive to $v_{\rm mic}$, but weak lines are not. Hence it is necessary to include both strong and weak lines to break the degeneracy.

$v_{\rm mac}$ and $v \sin i$ can usually not be determined separately, unless the spectrum is very well resolved. In most applications they can be described with a single kernel that describes their combined effect. Alone or combined, these parameters conserve equivalent width, which means they are not degenerate with the abundances, and can be determined simultaneously with abundances. This is what we do in Article II.

$v_{\rm rad}$ can usually be estimated by comparison to some other well-known spectrum at rest wavelength. As long as the spectra have roughly similar parameters and abundances, there will be spectral lines in the same places, albeit with different shapes. This can be used to fit $v_{\rm rad}$ manually, visually comparing the two spectra. In Paper II, we automatically calculate $v_{\rm rad}$ by using a cross-correlation, followed by $\chi^2$-minimisation.

### 3.5.1 Non-spectroscopic determination

For stars in a stellar cluster, it is possible to estimate stellar parameters using photometry instead of spectroscopy. As described in Sects. 2.2.1-2.2.2, in the space of the photometric quantities magnitudes and colour – the *colour-magnitude diagram* – a group of stars with identical ages will lie along a uniquely defined curve – the *isochrone*. For each isochrone there is a corresponding curve in $T_{\rm eff}$-$\log g$ space. By fitting an isochrone to the photometric quantities and then translating the position of each star to the corresponding position in $T_{\rm eff}$-$\log g$ space, it is possible to estimate those stellar parameters. We use this technique in Article II, using $V$ magnitudes and $V-I$ colour to determine those stellar parameters for stars in the cluster M30. However, there are some complications that we discuss in Sect. 4.4.

In situations where the parameter $v_{\rm mic}$ cannot be determined directly, it can sometimes be estimated indirectly by using empirical correlations with other parameters such as $T_{\rm eff}$ and $\log g$. We used this in Article II, since unlike $T_{\rm eff}$ and $\log g$ the parameter cannot be determined through photometry. We strongly advise all spectroscopists to avoid this method if at all possible. In Sect. 4.2.3 we describe how it impacted our systematic errors, and in Appendix 10.1 we discuss how many different empirical relations we had to choose between. We also caution that some studies use empirical relations without taking into account the effect on the systematic errors.

## 3.6 Common issues

The basic reasoning behind both abundance and parameter fitting is simply "The spectrum is a function of these things. By selecting wavelength regions appropriately, you can pick out a part of the segment which is both in principle an invertible function, and in practise is possible to model. That allows us to estimate those things". Unfortunately, the assumptions behind this reasoning occasionally break down, since neither our observations nor our models are perfect. Here we describe some common problems, and how to solve or work around them.

### 3.6.1 Noise

Spectrographs measure photons, which are discrete quantities. Because of this, there is an intrinsic amount of Poissonian noise in any spectrum. In addition, there is some level of internal noise in the spectrograph. The relative error coming from these sources becomes smaller the longer the exposure of the spectrum, but since observing time is expensive, most observations are made according to some trade-off between noise and expense.

In addition, spectroscopic observations are usually made from the ground. This means that the spectrum they observe is in reality the light from the star, overlaid with the light from the night sky. The relative error from this source does not decrease with the observation time. This is typically handled by on each observation night having one or more *sky fibres* in the spectrograph, which simply measure the spectrum of the night sky. This is then subtracted from the stellar spectra, which is called *sky subtraction*.

Both of these sources of error were major issues in Article II, since the spectra studied came from the distant cluster M30. In fact, the article focused on that cluster precisely because the TOP stars are so faint that few people have gone to the trouble of making this kind of measurement. In Article I, on the other hand, we used spectra that by design had very low noise levels.

### 3.6.2 Excessive curvature

As described in Sect. 3.3 continuum normalisation typically assumes that the continuum level of a wavelength segment is approximately linear. This approximation becomes worse the wider a segment is.

This became a problem during our work on Article I, where we initially used hydrogen absorption lines – *Balmer lines* – of stellar spectra to estimate $T_{\mathrm{eff}}$. These lines are in principle ideal for this purpose, since they are insensitive to most other parameters. However, for reasons including the high abundance of hydrogen, these lines are very wide. Because of their width, the Balmer lines are blended (see Sect. 3.6.3 below) with multiple iron lines. We did have an algorithm for ignoring the blended regions in our analysis, but we discovered that the width of the lines meant that the approximation of linear continuum simply broke down. If the problem had been purely physical this might have been resolvable – it might be possible to estimate the true curvature of the continuum level over the Balmer lines – but unfortunately there was an observational part to it as well: At least for UVES spectra, the Balmer lines for multiple spectra of the same star will usually look visibly different. Happily, we were able to solve the problem by simply removing the Balmer lines from our analysis, but we advise other spectroscopists that Balmer lines probably need to be used in manual fitting or not at all – attempts at including them in an automated analysis pipeline like ours are unlikely to be successful[3].

Note that there are spectral synthesis codes which allow fitting quadratic curves to the continuum, but we would caution against attempting this: In the spectra we have looked at, the ratio of continuum regions and spectral lines is such that a segment almost always has to consist of one or two spectral lines, with a little bit of continuum at the far edges of the segment. In this situation, the curvature of the continuum is inherently so poorly constrained that a quadratic fit can easily end up being worse than a linear fit.

### 3.6.3 Blending

Spectral lines are often *blended*, meaning that they overlap with other spectral lines. If one line is used in an analysis, while being blended with another line that is not taken into account by the analysis method, this will throw off the results to some extent. This risk can be diminished by attempting to use lines that are believed to be unblended – or that are blended with lines that are so well characterised that they can be taken into account – but there is usually a risk that some as-yet unknown line will affect the analysis. The impact of this can sometimes be minimised by observing multiple lines of the same element, and averaging together their results.

Note that what would be considered ‘blended’ is context-dependent. As we mentioned in Sect. 3.2.1, most lines are strictly speaking blended, but the lines they are blended with are so weak that they do not affect the analysis.

### 3.6.4 Saturation

As described in Sect. 3.1, there is a range of abundances for which a line will be *saturated*, meaning that it only responds weakly to changes in abundance. While the abundance is still in some strict sense an invertible function of the line shape, noise and instrumental resolution mean that the abundance can only be constrained to somewhere within a wide range. With some methods, even this is only possible in principle, and in practise they will simply return

[3] See also Korn et al. (2007) for a detailed discussion of how the results in Gratton et al. (2001) were partly an artefact of unreliable normalisation of Balmer lines.

a spuriously precise value somewhere within that range. This happened to the titanium line dubbed Ti4571 in Article II, which forced us to drop that line from the analysis.

## 3.7 Spectral synthesis and fitting

Model atmospheres can be made almost arbitrarily complex, but it is possible to reproduce observed spectra fairly well by modelling stellar atmospheres as one-dimensional (1D) systems with some elemental composition and two *thermodynamic parameters* which vary with depth. There are several choices one could make of which thermodynamic parameters to use, but a common pick is pressure and temperature. The stellar atmosphere is 1D not in the sense that the star is literally modelled as a line, but in the sense that the state of the atmosphere is only assumed to change with one spatial coordinate, depth. This can be done with either *spherically symmetric geometry*, where the star is treated as a rotationally symmetric ball, or *plane-parallel geometry* which assumes that the local curvature of the star is negligible. In Sect. 3.8, we discuss 3D models.

Given this model, it is possible to calculate which fraction of the nuclei of each element take the form of atoms, ions, and molecules. In the hotter stars molecules cannot form in any appreciable amount, since chemical bonds are immediately broken up. In this case, the problem simplifies to the *Saha-Boltzmann* equations.

To go on to calculate the spectrum, it is necessary to have a *line list*, which describes at least some of the spectral lines within the wavelength range of interest. This again can be made very complex, but at a minimum it must contain the wavelength, excitation energy and transition probability for each line. All work in this thesis uses the line list compiled for use within GES (Heiter et al., 2021).

Given this, the next step is so solve the *radiative transfer equation*, which gives the intensity of radiation at a specific wavelength, emitted along a line through the star. In principle, the solution is given by the integral over emission and absorption along the line all the way through the star. In practise, it is common to assume that at a particular depth the star is effectively opaque, so that it is enough to start with the black body emission when the line reaches that depth, and then calculate the integral over emission and absorption.

Once this is done, it is necessary to take into account several sources of *line broadening*, which shape the profile of each spectral line. There is *natural broadening*, which stems from the fact that the energy levels in an atom cannot be perfectly precise, due to Heisenberg's uncertainty principle. There is *thermal broadening*, which stems from the Doppler shift due to the random thermal motions of the atoms. There are multiple forms of *pressure broadening*, all of which stem from the energy levels of the atoms being shifted due to perturbations from other particles in the stellar atmosphere. The most important sub-types are *linear Stark broadening*, *quadratic Stark broadening* and *van der Waals broadening*.

In the next-to-final step, the stellar spectrum is calculated by, for each wavelength solving the radiative transfer equation at a range of angles through the stellar atmosphere, and then integrating over the entire stellar disk.

Finally, a realistic *instrumental profile* must be used to describe the spectrum one would actually see in a particular spectrograph. For the UVES and GIRAFFE spectra in Paper I and II respectively, this was done by simply convolving the spectrum with a Gaussian with a width corresponding to the resolution of the instrument. This is accurate enough to allow us to work with wavelength segments short enough that the normalisation can be approximated as a linear function, but as we discovered with the Balmer lines (see Sect. 3.6.2), this leaves out the effect of drift in the normalisation on larger scales. In Paper III, which used Gaia spectra, it was necessary to use a more complicated model of the instrumental profile.

In Articles I and II, we used synthetic spectra for the purpose of *spectral fitting*. That is we adjust the parameters of the model spectra until they are as similar they can be to the observed spectra, and take the corresponding parameters to be our best estimate of the true parameters[4]. The fitting itself is done by minimising a modified $\chi^2$-sum. In Sect. 4.2.7, we discuss the question of how to estimate errors in the resulting estimates.

[4] We avoid the philosophical question of to what extent a star actually has, say, a 'true' microturbulence.

## 3.8 Common approximations

For Articles I and II we used the spectral synthesis code *Spectroscopy Made Easy* (SME), which is described in more detail in Sect. 3.10.2. SME makes several approximations when modelling stellar spectra, which are shared by many other synthesis codes. It assumes that radiation transport in the stellar atmosphere occurs in Local Thermodynamic Equilibrium (LTE), meaning that the temperature does not vary significantly compared to the mean free path of photons. When solving the radiative transfer equation, this means that the absorption and emission at each point in the atmosphere is a black body spectrum given by the temperature at that depth. SME treats the stellar atmosphere in 1D. Both approximations become poorer in certain situations.

The circumstances under which the LTE approximation gives poor results for line formation models are complex, but it is possible to state some very rough rules of thumb: In general the LTE approximation becomes worse when the atmosphere becomes less dense, since this increases the mean free path of photons. This makes NLTE effects especially important in giant stars. They also tend to be more important in the cores of strong lines, since the cores mostly form further out in the atmosphere (Gray, 2008). In stars similar to the Sun, there is a tendency for lines to become shallower in the bluewards portion of the spectrum and deeper in the redwards. There is also a tendency for strong lines to become deeper. NLTE effects do not just change the shape of the individual lines, they also change the ionisation balance for the element in question from what the Saha equation would predict, which affects all lines of that element. If a line is deeper that means that there is more absorption, which means that more particles are being lifted out of that particular atomic level. For lines involving levels for which little additional energy is needed to detach an electron, this can lead to overionisation. For neutral strong lines it can instead lead to underionisation (Kiselman, 1999). In practise, the dominant effect tends to be overionisation. For the temperatures and elements we looked at in Article II – Mg, Ti and Fe – singly ionised particles tend to greatly predominate over neutral particles. This means that that neutral lines become noticeably shallower, while the deepening of singly ionised lines is fairly small. In metal-poor stars there are two partly-counteracting tendencies: On the one hand, the radiation field is stronger, which tends to increase overionisation due to NLTE. On the other hand, lines are weaker, meaning that they form in deeper regions of the atmosphere, where NLTE is less important.

The main drawback of the 1D approximation is that it essentially imposes the assumption of hydrostatic equilibrium, since large-scale mass flows other than pure rotation or pulsation are only possible if material is free to move up in some regions and down in others at the same depth[5]. This means that a naïve 1D model will predict stellar spectra with much narrower lines than are actually observed. It is possible to reproduce fairly well the line broadening

[5] That said, energy transport by convection is taken into account in model atmospheres.

seen in real spectra by introducing turbulent flow, which is taken to consist of isotropic flows of material. This is typically parametrised in terms of *macroturbulence* and *microturbulence*, which describe flows on distance scales respectively longer and shorter than the mean free path of photons. However, this has several drawbacks. Pragmatically, it has the problem that it introduces two free parameters into analyses, which makes calculations longer and conclusions less certain. Theoretically, it has the problem that neither parameter has a clear physical interpretation: Both parameters were originally introduced to describe isotropic turbulent flow, but comparison to 3D simulations indicate that the effects of both parameters can be well replicated even when the models do not actually include such turbulent flow. The effects ascribed to macroturbulence seem to be better explained by large-scale convective flow, while the effects ascribed to microturbulence are better explained by gradients in the convective flow (Asplund et al., 2000, Sect. 8).

## 3.9 UVES and GIRAFFE

The *Very Large Telescope* (VLT) at Paranal, Chile, consists of four 8.2-meter *Unit Telescopes* (UT). The second of these, *UT2*, is connected to an instrument called the *Fibre Large Array Multi Element Spectrograph* (FLAMES). FLAMES in turn feeds the two spectrographs Ultraviolet and Visual Echelle Spectrograph (UVES) and GIRAFFE[6] (Pasquini et al., 2002). The spectrographs are both *echelle spectrographs*, meaning that they separate light into a spectrum using a diffraction grating with relatively widely spaced grooves, and a fairly high angle of incidence for the incoming light.

The spectrographs have been designed with different trade-offs in performance, which makes them to some extent complementary. As a rule of thumb UVES has higher resolution but can only observe 8 stars at a time, while GIRAFFE has lower resolution but can observe 132 stars at a time. Roughly speaking, GIRAFFE has a resolution of $R \equiv \delta\lambda/\lambda$ around 20000-25000 for the settings used within GES, while UVES has $R$ around 47000 (Pasquini et al., 2002; Dekker et al., 2000).

The GES survey almost only uses spectra taken with either GIRAFFE or UVES. However, some of the *benchmark spectra* used to estimate the performance of different analysis methods use spectra observed with other spectrographs. These benchmark stars are described in more detail in Sect. 4.1.1

[6] This is not an acronym, despite the uppercase spelling. The name was chosen by the team on the basis that the instrument looks like a giraffe.

## 3.10 Software tools

Throughout the work on the articles in this thesis, we have used several important pieces of code. We describe them quickly here, with occasional digressions about features, misfeatures or outright bugs that may be useful to other spectroscopists.

### 3.10.1 MARCS

Model Atmospheres with a Radiative and Convective Scheme (MARCS) is a code for modelling the stellar atmospheres. It uses a 1D atmosphere in hydrostatic equilibrium, with either a spherical or plane-parallel geometry. It has existed in various forms since the 1970s, with a standardised version being published in 2008 (Gustafsson et al., 2008).

In Articles I and II we used a grid of MARCS model atmospheres which were in turn used by SME to model spectra. In Article III we instead used spectra calculated directly with MARCS.

### 3.10.2 SME

Spectroscopy Made Easy (SME) is a code for modelling stellar spectra, and fitting synthetic spectra to observed spectra. The first published version is from 1996, and it has been continuously updated and improved since. A description of the first release of SME can be found in Piskunov and Valenti (1996) and a more up-to-date description in Piskunov and Valenti (2017). In Articles I and II we used SME to fit observed spectra. In both cases we used SME with a grid of model spectra calculated using MARCS.

In order to fit to observed spectra, SME requires the user to define a number of *masks*. It is necessary to specify a *line mask*, which states which pixels should actually be used in the fit. To do continuum normalisation, it is also necessary to specify a *continuum mask*. Note that these pixels do not need to be true continuum. SME can do acceptable fitting if the continuum mask covers spectral lines, as long as the lines are sufficiently well characterised that SME can model them.

When fitting synthetic spectra to observed spectra, SME minimises a goodness-of-fit metric defined as

$$\chi^2_{\mathrm{SME}} \equiv \sum_{i=0}^{n} \frac{O_i \left(O_i - E_i\right)^2}{\sigma_i^2} \tag{3.3}$$

where the $n$ pixels covered by the line mask are indexed by $i$, $O_i$ is the observed intensity of pixel $i$, $E_i$ is the model intensity of pixel $i$ and $\sigma_i$ is the uncertainty in $O_i$. This differs from the usual Pearson's $\chi^2$-sum in the additional factor of $O_i$. This is intended to capture a rule-of-thumb that spectroscopists tend to use when fitting spectra manually: There are two things that make observed spectra and model spectra differ from each other. There is noise, which shifts the observed intensity (almost) symmetrically, and there are imperfections in the model. One of the most important imperfections is that the line list is incomplete: There are several lines in reality that are not included in the model. Hence, the best fit is usually one where the model spectrum lies slightly above the observed spectrum (Nikolai Piskunov, priv. comm). To minimise $\chi^2_{\mathrm{SME}}$, SME uses the Levenberg algorithm[7].

Different versions of SME have used different methods for estimating the errors in derived quantities. The oldest method assumes that the only source of error is pixel noise, while the newer method attempts to take systematic errors into account. Neither method can be said to be better than the other in an absolute sense. The newer method is better for estimating errors in quantities estimated using multiple lines, while the older method is probably better for abundances estimated using a single line (Thomas Nordlander, priv. comm). While SME does not have a simple toggle for switching between the methods, it is relatively easy to modify the SME code to restore the older method. This is described in more detail in Sect. 4.5.

### 3.10.3 IDL

A surprising amount of astronomy software, including SME[8], is written in an otherwise-obscure programming language from the 70s: Interactive Data Language (IDL). This is mostly a matter of historical contingency: Up until the 1990's, IDL was more-or-less unique as a combination of programming language and visualization tool, which caused many astronomers at the time to decide to use it (priv. comm. Andreas Korn). Today it is no longer unique, but there is so much code written in IDL that it would require a very large investment of labour to migrate it to some other language.

This is problematic for several reasons. One is that since IDL is very old, it is simply less easy to use than more recent languages, and permits coding practices that are now considered harmful. A more serious problem is that IDL is proprietary, and the licence management system is fairly strict: In addition to having to own an IDL licence to run IDL code, the user must be able to continously access the IDL licence server. It is also only possible to run a limited number of IDL processes in parallel using one licence. This is a severe problem when using IDL for large surveys such as GES, since these require running multiple analyses in parallel. It is possible to run IDL without contacting the licence server by running it in *virtual machine mode*, but this

[7] It is sometimes referred to as the *Levenberg-Marquardt algorithm*, but Levenberg (1944) and Marquardt (1963) proposed slightly different algorithms, and the one used in SME is the version proposed by Levenberg.

[8] However, Ansgar Wehrhahn has recently been working on the development of the Python-based PySME. We encourage all SME users to use this above IDL SME.

requires pressing a button in the IDL Virtual Machine splash screen that appears on start-up. Writing scripts to do this automatically is trivial in principle, but explicitly forbidden by the licence agreement (Harris Geospatial Solutions, 2021, Sect. 9).

When using IDL, we also used two important libraries. One is the IDL Astronomy User's Library (IDLAstro/astrolib), which contains a selection of tools specifically for astronomy. We also made extensive use of the Coyote IDL library, which simply contains a large number of useful functions. We also used the Coyote homepage, which contains many guides explaining otherwise baffling quirks of IDL[9] (Fanning, 2015).

### 3.10.4 Python

Where possible, we have used the programming language Python. This has the twin advantages of being both new and widely used. Being new means that the syntax is very streamlined. Being widely used means that for most problems, somebody somewhere already has a ready solution. The process of using other people's solutions has been simplified through the use of *modules*. These are open-source code packages, containing Python functions that have been assembled to solve some overall type of problem. In the work on this thesis, we have made use of the following modules:

- *SciPy*: A collection of libraries for scientific computing (Virtanen et al., 2020). Among other things, it contains the following modules:
  - *NumPy*: The main purpose of this library is to allow efficient calculations using arrays (Harris et al., 2020).
  - *matplotlib*: This library consists entirely of functions for making plots (Hunter, 2007). All plots in this thesis and the included articles were made with matplotlib.
- *astropy*: This is a library of functions used within astronomy (Astropy Collaboration et al., 2013, 2018). In particular, we have made use of *astropy.io* to handle files in the *FITS* format (described below), and *astropy.coordinates* when handling the Gaia spectra in Article III.
- *scikit-learn*: A library of tools for machine learning (Pedregosa et al., 2011). We used their implementation of the ExtraTrees algorithm in Article III.
- *emcee*: A library implementing Goodman and Weare's *Affine Invariant Markov Chain Monte Carlo Ensemble sampler* (Foreman-Mackey et al., 2013). We used this for estimating the probability distributions in Articles I and II.

In addition to the modules themselves, use of Python is simplified by the fact that the solutions to both common and uncommon problems can be found on

[9] In particular, a user having problems with IDL mysteriously accessing files that are not actually in the IDL path should read *The IDL Path Problem from Hell*.

*stackoverflow*[10]. While we do not cite them individually, this thesis owes a great debt to countless pseudonymous stackoverflow users.

### 3.10.5 The FITS format

The astronomy community often uses the *Flexible Image Transport System* (FITS) format (Pence, W. D. et al., 2010). The spectral files used for Articles I and II use the FITS format. The GES line list used in both articles was distributed internally in the FITS format. GES also uses FITS for communicating the results of analyses.

In IDL, there are several built-in functions for handling FITS files. In Python, the *astropy* module contains functions for handling FITS files inside the sub-module *astropy.io* (Astropy Collaboration et al., 2013, 2018). There is also the open-source program *topcat*, which can be used to inspect files in several formats including FITS (Taylor, 2005). However, we caution that the format is not implemented quite consistently on different platforms. We have occasionally found crucial differences between a FITS file opened with topcat and the same file opened with astropy.

[10] `https://stackoverflow.com/`

# 4. Statistics

This chapter discusses the statistical methods that were used in the articles included in this thesis. Where we developed the methods, it attempts to explain not only what the methods were and how they work, but why they were not defined some other way. This is not described in much detail in any of the articles.

Section 4.1 describes the statistical method that we developed for Article I to estimate the performance of a pipeline for estimating stellar parameters from spectra observed with the UVES spectrograph. Section 4.2 describes the statistical method that we developed for Article II to constrain a parameter in a stellar evolution model, based on abundances derived from spectra observed with the GIRAFFE spectrograph. Section 4.3 describes the Extra-Trees algorithm, which we used in Article III. Section 4.4 discusses some issues with spectroscopic techniques which involve comparisons to stellar isochrones, since we used such techniques in Articles I and II. Section 4.5 discusses the still mostly-unsolved problem of how to estimate uncertainties in abundances or stellar parameters derived through spectroscopy. Section 4.6 describes the algorithm that we developed for Article I to define the continuum mask used for normalising our spectra. Section 4.7 describes the method for coadding spectra that we developed for Article II.

## 4.1 Estimating performance for stellar parameter pipelines

Part of the work that led to Article I consisted of trying to improve the pipeline used by LUMBA to estimate stellar parameters. After some false starts we realised that we did not have a well-defined method for estimating the performance of the pipeline, which meant that it was hard to evaluate whether a particular change had actually made the pipeline better or worse. This prompted us to develop a method that we believe can be used in general to test stellar parameter pipelines.

Within GES, and many similar surveys, the usual tool for estimating the performance of a stellar parameter pipeline is a sample of *benchmark stars*. These are stars whose stellar parameters have been measured using methods other than spectroscopy, such as through their radius and bolometric flux (Blanco-Cuaresma et al., 2014). By letting a pipeline analyse spectra of the benchmark stars, one can then get an independent estimate of how well it performs, which would not be possible by comparing spectroscopic estimates to each other. Ideally, this should reveal approximately what the systematic errors of the pipeline are, for a particular type of star. If, for example, a pipeline consistently gives parameter estimates that are 100 K too high for Solar-type stars, one can then lower its estimates for Solar-type stars by that much.

In practise, there are many different ways one could go about this estimation of the systematic errors. Here, we describe the method used within Article I. We also compare it to a method used within Gaia-ESO, and attempt to show that our method is somewhat more accurate. Unfortunately, as we will see, it is 'accurate' in the sense that it reports larger uncertainties on the estimates of the systematic errors, but these uncertainties more correctly reflect reality.

### 4.1.1 Gaia FGK benchmark stars

The GES benchmark sample is described in detail in Blanco-Cuaresma et al. (2014). The stars are sorted into the four categories *Solar-type stars*, *F dwarfs*, *FGK subgiants* and *red giants*. Table 4.1 shows the full list of benchmark stars, and the number of spectra per star.

In Article I, we investigate the performance for each category separately. This on the assumption that the performance of the pipeline does vary with the stellar parameters, but that it does so slowly enough to be approximately constant within a type of benchmark stars.

The spectra were not all observed with UVES. Many were instead convolved to the resolution of UVES after being observed with the spectrographs ESPaDOnS, HARPS, NARVAL and ATLAS. In some cases this was due to necessity, since not all of the stars are visible from the latitude of Paranal. We were able to use the spectra from ESPaDOnS and NARVAL, but not those from HARPS and ATLAS, since the dichroic gaps of those spectrographs

**Table 4.1.** *Gaia-ESO benchmark stars. A similar but less detailed table is shown in Article I. A number of the benchmark stars were left out of the analysis, for reasons described in detail in the body text of Sect. 4.1.1.*

| Star type | Star | Spectra | Comment |
|---|---|---|---|
| Solar-type stars | The Sun | 9 | |
| | 18 Sco | 3 | |
| | $\mu$ Ara | 3 | |
| | $\mu$ Cas A | 1 | |
| | $\alpha$ Cen A | 2 | |
| | $\alpha$ Cen B | 1 | Not used |
| | $\tau$ Cet | 3 | |
| | $\beta$ Vir | 3 | |
| | HD 22879 | 2 | |
| F dwarfs | Procyon | 6 | |
| | HD 49933 | 2 | |
| | HD 84937 | 5 | |
| FGK subgiants | $\eta$ Boo | 2 | |
| | $\delta$ Eri | 4 | |
| | $\varepsilon$ For | 1 | |
| | $\beta$ Hyi | 3 | |
| | HD 140283 | 5 | |
| Red giants | Arcturus | 5 | |
| | $\beta$ Ara | 1 | |
| | $\alpha$ Cet | 3 | Not used |
| | $\xi$ Hya | 2 | |
| | $\mu$ Leo | 2 | |
| | $\psi$ Phe | 2 | Not used |
| | $\gamma$ Sge | 1 | Not used |
| | $\alpha$ Tau | 2 | Not used |
| | $\varepsilon$ Vir | 3 | |
| | HD 107328 | 2 | |
| | HD 122563 | 5 | Not used |
| | HD 220009 | 2 | |

overlapped with lines that were used by our pipeline. This forced us to drop $\alpha$ Cen B from the analysis, since the one spectrum of that star was observed with HARPS.

Note that it is possible to criticise the assumption of spectrographs being interchangeable up to resolution. During the work on internal Data Release 6 (iDR6) of GES, we had access to 13 UVES spectra of HD 22879. We found that the parameter estimates for those spectra tended to cluster with time, so that spectra that were observed during the same night tended to have similar errors in the estimated parameters. In most cases it was not clear what caused the difference, but we did find that at some point between 2013 and 2015 the appearance of the spectra changes so strongly that it is visible to the eye, which correlated with a large shift in the errors. This could be taken to imply that UVES is not equivalent to itself at different points in time, in which case it cannot be equivalent to any other spectrograph at all points in time. However, it is also possible that the clustering is not due to the UVES spectrograph itself, and instead comes from spectral reduction issues internal to GES. During iDR6 we did find and report some spectra being affected by line doubling, which led to the sample being sent through a second round of quality control, and part of the sample being updated. HD 22879 was not found to be among the affected stars, but it is possible that the spectra had more subtle issues that were not detected at the time. In any case, we believe that the spectrographs are close enough to equivalent to allow us to get meaningful results.

### 4.1.2 Model of pipeline error

Let us assume that we are trying to estimate some stellar parameter $p$, such as $T_{\text{eff}}$. Let index $i$ denote different benchmark stars and index $k$ denote individual observations of a particular star. Let $p_{i,k}$ denote the estimate of stellar parameter $p$ given by running spectrum $k$ of star $i$ through the pipeline. Let $p_i^{\text{true}}$ denote the true value of $p$ for star $i$.

We model the difference between the estimate and the true value as having three components

$$p_{i,k} = p_i^{\text{true}} + e^{\text{pipe}} + e_i + e_{i,k} \tag{4.1}$$

The error term $e^{\text{pipe}}$ is assumed to depend on $p_i^{\text{true}}$, but with a sufficiently slow variation that it can be treated as constant for stars of the same type as star $i$. The error term $e_i$ is assumed to vary from star to star according to some statistical distribution with mean 0, while remaining constant for all spectra of a particular star. The error term $e_{i,k}$ is assumed to vary from spectrum to spectrum of a star, according to some statistical distribution with mean 0. We will make the assumption that the distributions of $e_i$ and $e_{i,k}$ are normal distributions[1] with standard deviations $\sigma^{\text{star}}$ and $\sigma^{\text{spec}}$, respectively. In this

[1] We will assume that most random variables follow normal distributions. This is common practise, but we want to point out that we are not entirely comfortable with it. See MacKay

model, the performance of any pipeline for a particular type of star can be fully characterised by the parameters $e^{\text{pipe}}$, $\sigma^{\text{star}}$ and $\sigma^{\text{spec}}$.

Each one of the error terms has a relatively straightforward interpretation: The term $e^{\text{pipe}}$ is intended to represent the errors inherent to the pipeline, and which have a fairly consistent effect within a volume of parameter space. For example, if a line used by the pipeline is modelled with an incorrect line strength, this will affect all similar stars by a similar amount. The term $e_i$ is intended to represent errors coming from how the idiosyncracies of a particular star violates the assumptions built into the pipeline. For example, if a line used by the pipeline is blended with another line, then this will affect that star to an extent determined by the abundance of the blended element within that particular star. The term $e_{i,k}$ is intended to represent errors coming from the idiosyncrasies of the individual spectra. For example, the Poissonian noise will be uncorrelated from spectrum to spectrum.

We believe that each of the terms in Eq. (4.1) is necessary to model the error in a way that is appropriate for testing the pipeline using benchmark stars. In Sect. 4.1.7 we will look at a similar model which leaves out the term $e_i$, and explain why we believe this is a mistake. That said, once the errors have been estimated for a particular pipeline using this method, it is likely to be unnecessarily complex for most actual science cases. In most studies there will only be one spectrum for each star, meaning that $e_i$ and $e_{i,k}$ cannot be distinguished. Only the combined scatter with width $\sqrt{(\sigma^{\text{star}})^2 + (\sigma^{\text{spec}})^2}$ will appear. In some studies the number of stars studied may be large enough that only $e^{\text{pipe}}$ matters in practise. Even so, Eq. (4.1) is necessary to use benchmark stars to make a reasonable estimate of $e^{\text{pipe}}$.

It would be possible to make this model more realistic by including more terms: one could also have a term $e_{\text{spectrograph}}$, representing the different characteristics of the different spectrographs used for observing the benchmark spectra; one could have a term $e_{\text{time}}$, representing the effect of when an observation was made; one could have a term $e_{\text{S/N},i,k}$, representing asymmetry in the error caused by Poissonian noise. These suggestions are not as frivolous as they might seem: as discussed in Sect. 4.1.1, we have found indications that different spectrographs have slightly different characteristics; as also discussed in Sect. 4.1.1, we have also found indications that the characteristics of the UVES may change slightly over time; as discussed in Sect. 4.6, the effect of Poissonian noise on the derived parameters does not necessarily have mean zero. However, we do not believe we could constrain a model of the error with more parameters than those included in Eq. (4.1), with the number of benchmark stars we have.

(2003, 23.2) for an argument that this practise is at least partly based on a misunderstanding of the central limit theorem.

### 4.1.3 Model of benchmark error

What makes the benchmark stars qualify as benchmarks is not that their parameters are known exactly, but that they are known with methods independent of spectroscopy. For all benchmark stars except for the Sun, there are appreciable uncertainties on the stellar parameters. (The Sun in turn has other problems, which we discuss in Sect. 4.1.9). This uncertainty also needs to be taken into account, but the model does not need to be as complex as that for the pipeline error.

We assume that each fundamental estimate $p_i^{\rm fund}$ can simply be described as

$$p_i^{\rm fund} = p_i^{\rm true} + e_i^{\rm fund} \tag{4.2}$$

where $e_i^{\rm fund}$ follows a normal distribution with standard deviation $\sigma_i^{\rm fund}$. We take the literature values of the fundamental values and their stated uncertainties as our estimates of $p_i^{\rm fund}$ and $\sigma_i^{\rm fund}$.

Note that we do not take into account the possibility of systematic errors in the fundamental estimates. While such errors certainly do exist, there is no way of disentangling them from $e^{\rm pipe}$. This means that if we wanted to include them, we could do so by keeping Eq. (4.1) as it is, but reinterpreting $e^{\rm pipe}$ as the difference in systematic errors between fundamental and spectroscopic parameter estimates.

### 4.1.4 Likelihood function

We want to estimate the error parameters $e^{\rm pipe}$, $\sigma^{\rm star}$ and $\sigma^{\rm spec}$. The information that we in principle have available is a sequence of stellar parameter estimates $\{p_{j,k}\}$, a sequence of fundamental parameter estimates $\left\{p_j^{\rm fund}\right\}$, and the estimated standard deviations in those estimates $\left\{\sigma_j^{\rm fund}\right\}$. Whichever statistical framework we want to use for this[2], we will need to start with the likelihood of getting those estimates, given the known $\left\{p_j^{\rm fund}\right\}$ and $\left\{\sigma_j^{\rm fund}\right\}$, and the unknown $e^{\rm pipe}$, $\sigma^{\rm star}$ and $\sigma^{\rm spec}$.

**Definitions and notation**

We denote[3] the likelihood function as:

$$P^{\rm full} \equiv P\left(\{p_{j,k}\} \middle| \left\{p_j^{\rm fund}\right\}, \left\{\sigma_j^{\rm fund}\right\}, e^{\rm pipe}, \sigma^{\rm star}, \sigma^{\rm spec}\right) \tag{4.3}$$

[2] As already mentioned, we will use a Bayesian framework, but it would be trivial to reframe our approach in Frequentist terms.

[3] In everything that follows, we will use the notation that $P(x|y)$ refers to the probability of $x$ given $y$.

This probability depends in some way on the likelihoods of getting the estimates $\left\{p_{i,k}\right\}$ for each individual star $i$, given some values of $p_i^{\text{fund}}$, $\sigma_i^{\text{fund}}$, $e^{\text{pipe}}$, $\sigma^{\text{star}}$ and $\sigma^{\text{spec}}$. We denote these as:

$$P_i \equiv P\left(\left\{p_{i,k}\right\}\middle| p_i^{\text{fund}}, \sigma_i^{\text{fund}}, e^{\text{pipe}}, \sigma^{\text{star}}, \sigma^{\text{spec}}\right) \tag{4.4}$$

Each $P_i$ in turn depends on both the probability of getting that sequence $\left\{p_{i,k}\right\}$, given some specific value of $p_i^{\text{true}}$, as well as the probability of a star actually having that $p_i^{\text{true}}$, given the values $p_i^{\text{fund}}$ and $\sigma_i^{\text{fund}}$ of the fundamental parameters. We denote these as:

$$P_i^{\text{true}} \equiv P\left(\left\{p_{i,k}\right\}\middle| p_i^{\text{true}}, e^{\text{pipe}}, \sigma^{\text{star}}, \sigma^{\text{spec}}\right) \tag{4.5}$$

$$P_i^{\text{fund}} \equiv P\left(p_i^{\text{true}}\middle| p_i^{\text{fund}}, \sigma_i^{\text{fund}}\right) \tag{4.6}$$

The probability $P_i^{\text{fund}}$ is, as we will see below, given directly by our model of the benchmark error. The probability $P_i^{\text{true}}$ in turn depends on the probability $P_i^{e_i \Rightarrow p_i}$ of getting a sequence $\left\{p_{i,k}\right\}$ – given some value of the error term $e_i$, together with $p_i^{\text{true}}$, $e^{\text{pipe}}$ and $\sigma^{\text{spec}}$ – and on the probability $P_i^{e_i}$ of getting that value $e_i$. We denote these as:

$$P_i^{e_i \Rightarrow p_i} \equiv P\left(\left\{p_{i,k}\right\}\middle| p_i^{\text{true}}, e_i, e^{\text{pipe}}, \sigma^{\text{spec}}\right) \tag{4.7}$$

$$P_i^{e_i} \equiv P\left(e_i | \sigma^{\text{star}}\right) \tag{4.8}$$

The probability $P_i^{e_i}$ is, as we will see below, given directly by our model of the error. The probability $P_i^{e_i \Rightarrow p_i}$ in turn depends on the probability of getting a particular estimate $p_{i,k}$, given the error parameters and the true parameter value $p_i^{\text{true}}$. We denote this as:

$$P_{i,k} \equiv P\left(p_{i,k} | p_i^{\text{true}}, e_i, e^{\text{pipe}}, \sigma^{\text{spec}}\right) \tag{4.9}$$

**Formalisation of model**

Since we assume that the star-to-star error is uncorrelated outside of that taken account in the term $e^{\text{pipe}}$, it follows that

$$P^{\text{full}} = \prod_{i=0}^{N} P_i \tag{4.10}$$

The probability for a particular sequence $\left\{p_{i,k}\right\}$ is given by the integral over the probability of getting that sequence given a particular value of $p_i^{\text{true}}$, multiplied by the probability of that $p_i^{\text{true}}$, given the estimated $p_i^{\text{fund}}$:

$$P_i = \int_{-\infty}^{\infty} P_i^{\text{fund}} P_i^{\text{true}} dp_i^{\text{true}} \tag{4.11}$$

Given the assumptions in Sect. 4.1.3, the probability of a particular value of $p_i^{\text{true}}$ given a particular $p_i^{\text{fund}}$ is given by:

$$P_i^{\text{fund}} = \frac{1}{\sqrt{2\pi}\sigma_i^{\text{fund}}} \exp\left(-\frac{\left(p_i^{\text{true}} - p_i^{\text{fund}}\right)^2}{2\left(\sigma_i^{\text{fund}}\right)^2}\right) \tag{4.12}$$

The probability of getting a particular sequence of values $\{p_{i,k}\}$ is the integral over all possible errors $e_i$ over the probability of getting that sequence given a particular value of the error term $e_i$, multiplied by the probability of that error term

$$P_i^{\text{true}} = \int_{-\infty}^{\infty} P_i^{e_i \Rightarrow p_i} P_i^{e_i} de_i \tag{4.13}$$

The probability of getting a sequence of values $\{p_{i,k}\}$ given some value of the error term $e_i$ is the product of the probability for all the individual values $p_{i,k}$.

$$P_i^{e_i \Rightarrow p_i} = \prod_{k=0}^{n_i} P_{i,k} \tag{4.14}$$

Given the assumptions in Sect. 4.1.2, the probability of a particular value of $p_{i,k}$, given some value of $e_i$, is given by:

$$P_{i,k} = \frac{1}{\sqrt{2\pi}\sigma^{\text{spec}}} \exp\left(-\frac{\left(p_i^{\text{true}} + e^{\text{pipe}} + e_i - p_{i,k}\right)^2}{2\left(\sigma^{\text{spec}}\right)^2}\right) \tag{4.15}$$

The assumptions in Sect. 4.1.2 also mean that the probability of a particular value of $e_i$ is given by

$$P_i^{e_i} = \frac{1}{\sqrt{2\pi}\sigma^{\text{star}}} \exp\left(-\frac{e_i^2}{2\left(\sigma^{\text{star}}\right)^2}\right) \tag{4.16}$$

Given this, we can see that $P^{\text{full}}$ is in principle a function of $e^{\text{pipe}}$, $\sigma^{\text{star}}$ and $\sigma^{\text{spec}}$, with no parameters that are not already known.

**Derivation**

We insert Eq. (4.15) into Eq. (4.14), which gives us

$$P_i^{e_i \Rightarrow p_i} = \frac{1}{(2\pi)^{n_i/2}\left(\sigma^{\text{spec}}\right)^{n_i}} \exp\left(-\sum_{k=0}^{n_i} \frac{\left(p_i^{\text{true}} + e^{\text{pipe}} + e_i - p_{i,k}\right)^2}{2\left(\sigma^{\text{spec}}\right)^2}\right) \tag{4.17}$$

We insert Eq. (4.17) together with Eq. (4.16) into Eq. (4.13), and move everything independent of $e_i$ outside the integral.

$$P_i^{\text{true}} = \frac{\exp\left(\sum_{k=0}^{n_i} \frac{\left(p_i^{\text{true}}+e^{\text{pipe}}-p_{i,k}\right)^2}{2(\sigma^{\text{spec}})^2}\right)}{(2\pi)^{\frac{n_i+1}{2}}\left(\sigma^{\text{spec}}\right)^{n_i}\sigma^{\text{star}}} \times$$
$$\times \int_{-\infty}^{\infty} \exp\left(-\frac{n_i\left(\sigma^{\text{star}}\right)^2+\left(\sigma^{\text{spec}}\right)^2}{2\left(\sigma^{\text{spec}}\sigma^{\text{star}}\right)^2}e_i^2 - 2\sum_{k=0}^{n_i}\frac{e_i\left(p_i^{\text{true}}+e^{\text{pipe}}-p_{i,k}\right)}{2\left(\sigma^{\text{spec}}\right)^2}\right)de_i \tag{4.18}$$

We use the standard integral

$$\int_{-\infty}^{\infty} \exp\left(-ax^2-2bx\right)dx = \frac{\pi}{a}\exp\left(\frac{b^2}{a}\right) \tag{4.19}$$

For brevity, we introduce the shorthand

$$A_i \equiv \frac{n_i}{2\left(n_i\left(\sigma^{\text{star}}\right)^2+\left(\sigma^{\text{spec}}\right)^2\right)} \tag{4.20}$$

$$B_i \equiv \frac{1}{2\left(n_i\left(\sigma^{\text{star}}\right)^2+\left(\sigma^{\text{spec}}\right)^2\right)}\left(n_ie^{\text{pipe}}-\sum_{k=0}^{n_i}p_{i,k}\right) \tag{4.21}$$

$$C_i \equiv \frac{1}{2\left(\sigma^{\text{spec}}\right)^2}\left(\sum_{k=0}^{n_i}\left(e^{\text{pipe}}-p_{i,k}\right)^2-\frac{\left(\sigma^{\text{star}}\right)^2\left(\sum_{k=0}^{n_i}\left(e^{\text{pipe}}-p_{i,k}\right)\right)^2}{n_i\left(\sigma^{\text{star}}\right)^2+\left(\sigma^{\text{spec}}\right)^2}\right) \tag{4.22}$$

$$D_i \equiv (2\pi)^{n_i/2}\left(\sigma^{\text{spec}}\right)^{n_i-1}\left(n_i\left(\sigma^{\text{star}}\right)^2+\left(\sigma^{\text{spec}}\right)^2\right) \tag{4.23}$$

This turns Eq. (4.18) into

$$P_i^{\text{true}} = \frac{\exp\left(-A_i\left(p_i^{\text{true}}\right)^2-2B_ip_i^{\text{true}}-C_i\right)}{D_i} \tag{4.24}$$

For brevity and consistency, we introduce an analogous shorthand for Eq. (4.12):

$$E_i \equiv \frac{1}{2\left(\sigma_i^{\text{fund}}\right)^2} \tag{4.25}$$

$$F_i \equiv -\frac{p_i^{\text{fund}}}{2\left(\sigma_i^{\text{fund}}\right)^2} \tag{4.26}$$

$$G_i \equiv \frac{\left(p_i^{\text{fund}}\right)^2}{2\left(\sigma_i^{\text{fund}}\right)^2} \tag{4.27}$$

$$H_i \equiv \sqrt{2\pi}\sigma_i^{\text{fund}} \tag{4.28}$$

This shortens Eq. (4.12) into

$$P_i^{\text{fund}} = \frac{\exp\left(-E_i\left(p_i^{\text{true}}\right)^2 - 2F_i p_i^{\text{true}} - G_i\right)}{H_i} \tag{4.29}$$

We insert Eq. (4.18) and Eq. (4.29) into Eq. (4.11)

$$P_i = \frac{1}{D_i H_i}\int_{-\infty}^{\infty} \exp\left(-\left(A_i + E_i\right)\left(p_i^{\text{true}}\right)^2 - 2\left(B_i + F_i\right)p_i^{\text{true}} - \left(C_i + G_i\right)\right) dp_i^{\text{true}} \tag{4.30}$$

We use the standard integral

$$\int_{-\infty}^{\infty} \exp\left(-ax^2 - 2bx - c\right) dx = \sqrt{\frac{\pi}{a}} \exp\left(\frac{b^2}{a} - c\right) \tag{4.31}$$

This turns Eq. (4.30) into

$$P_i = \frac{1}{D_i H_i}\sqrt{\frac{\pi}{A_i + E_i}} \exp\left(\frac{\left(B_i + F_i\right)^2}{A_i + E_i} - \left(C_i + G_i\right)\right) \tag{4.32}$$

We insert Eq. (4.32) into Eq. (4.10)

$$P^{\text{full}} = \pi^{N/2}\prod_{i=0}^{N} \frac{1}{D_i H_i \sqrt{A_i + E_i}} \exp\left(\frac{\left(B_i + F_i\right)^2}{A_i + E_i} - \left(C_i + G_i\right)\right) \tag{4.33}$$

This gives us the likelihood function, expressed only in variables that are either known, or that are the three that we wish to estimate.

Note that for numerical reasons, all actual calculations are done using the log-probability corresponding to Eq. (4.33):

$$\begin{aligned} \log P^{\text{full}} &= \frac{N}{2}\log \pi + \\ &+ \sum_{i=0}^{N}\left(-\log\left(D_i H_i\right) - \frac{1}{2}\log\left(A_i + E_i\right) + \frac{\left(B_i + F_i\right)^2}{A_i + E_i} - \left(C_i + G_i\right)\right) \end{aligned} \tag{4.34}$$

### 4.1.5 Priors and resulting posterior

To calculate the posterior on the parameters, we need to assume some priors. In Article I, we choose to use flat priors, so that any value of $e^{\text{pipe}}$, $\sigma^{\text{spec}}$ and $\sigma^{\text{star}}$ is initially taken to be equally likely. This means that the posterior is simply equal to the normalised likelihood.

In retrospect, this was not an optimal choice. A Bayesian could object to it on the following grounds: While it might seem that the choice of a flat prior makes the least assumptions about the parameters, this is not actually true. There are flat priors and there are uninformative priors, and it is only in special cases that they are the same thing. In our case, a flat prior happens to be an uninformative prior for a location parameter such as $e^{\text{pipe}}$, but it is not for scale parameters such as $\sigma^{\text{star}}$ and $\sigma^{\text{spec}}$ (von Toussaint, 2011, Sect. II.D.1). Meanwhile, a Frequentist could object to it on the grounds that by picking a flat prior we have made our method completely equivalent to a maximum-likelihood method, and that this makes it unnecessary to bring in the Bayesian framework to begin with.

While this part of analysis is aesthetically unpleasant, we do not believe that it affects the results much. In general, while there is much controversy about the correct choice of Bayesian priors – and whether there even is a 'correct' choice – it usually has a surprisingly small effect on the results (Trotta, 2017, Sect. 3.1).

### 4.1.6 Implementation

Given a set of parameter estimates $\{p_{i,k}\}$ for a set of benchmarks, together with fundamental parameter estimates $\{p_i^{\text{fund}}\}$ and corresponding uncertainties $\{\sigma_i^{\text{fund}}\}$, Eq. (4.34) in principle gives the (unnormalised) posterior probability of any particular combination of error parameters $e^{\text{pipe}}$, $\sigma^{\text{star}}$ and $\sigma^{\text{spec}}$.

We use an affine invariant ensemble sampler to estimate the normalised posterior. We refer the reader to Foreman-Mackey et al. (2013) for a detailed description of the sampler, and to MacKay (2003, Chap. 29) for the underlying theory of sampling probability distributions.

### 4.1.7 Comparison to a Gaia-ESO method

Gaia-ESO has tried and used a number of different methods for using the benchmark stars. This includes one that is similar to ours, with one important difference: Notation aside, it uses a formula analoguous to Eq. (4.1) that is essentially

$$p_{i,k} = p_i^{\text{true}} + e^{\text{pipe}} + e_{i,k}. \tag{4.35}$$

That is, the term $e_i$ is missing. This makes a very large difference for how much information about $e^{\text{pipe}}$ one could hope to get out of a single benchmark star. By simply having enough spectra of a star, the impact of $e_{i,k}$ can be made negligible. Hence, according to Eq. (4.35), the only real limitation on how well benchmark star $i$ allows us to constrain $e^{\text{pipe}}$ is set by the error $e_i^{\text{fund}}$. In principle, a single benchmark star with very many spectra and very well-known fundamental parameters will be sufficient to cancel out the effect of systematic errors on all similar stars. On the other hand, the error model in Eq. (4.1) implies that even if we had all those things for that benchmark star, the quirks inherent to that star itself would put a bound on how well we can know $e^{\mathrm{pipe}}$, even within that region of parameter space.

This might not be a big problem, if one could expect $e_i$ to be comparatively small. Then one could in practise still constrain $e^{\mathrm{pipe}}$ very well with enough spectra and sufficiently well-known fundamental parameters. Unfortunately, in Article I, we found that $\sigma^{\mathrm{star}}$ is generally of comparable size to $\sigma^{\mathrm{spec}}$. This means that at least for our pipeline, multiple benchmark stars are necessary to constrain $e^{\mathrm{pipe}}$ even within a small volume of parameter space.

Note that it could in principle turn out that most pipelines do not behave like this, and our pipeline has some design flaw which makes it uniquely sensitive to the quirks of the individual benchmark stars. If so, that would be a happy discovery for everyone except us. Since we are obviously a biased source, we will not spend time on apologia for the LUMBA pipeline, but we note that during iDR5 and iDR6 the pipeline did not stick out as performing either unusually badly or unusually well compared to the other pipelines.

If this is representative of pipelines in general, all pipelines in surveys using a similar approach to Gaia-ESO would need to use an approach similar to what we have described here, using an error model resembling Eq. (4.1). This has an immediate implication for anyone trying to define a set of benchmark stars for that survey. Within Gaia-ESO there has been a widespread assumption that relatively few benchmark stars are sufficient to constrain performance within the most-populated regions of parameter space, and that since the survey already has those, the natural next step is to find benchmark stars in the less-populated regions of parameter space. Our findings imply that this would not necessarily be a good investment of resources, and that the survey might benefit more from finding more benchmark stars in the most-populated regions. This is a bit unfortunate, since finding benchmark stars in less-populated regions would mean finding very rare and unusual stars, which would be interesting to do, while our recommendation is just to find large numbers of very ordinary stars instead.

### 4.1.8 Generality of method

We want to emphasise that our method is very general. In principle, the underlying assumptions can be summarised as:

“We have made noisy measurements of a quantity $x$ for a large number of doodads. However, we do not actually care about $x$. We care about a different quantity $y$. We have a black box which when given measurements of $x$ outputs estimates of $y$. We have some special doodads for which we have managed to make noisy but uncorrelated measurements of $y$. For some of those, we may have multiple measurements of $x$. Based on this, how can we figure out how good our black box is?”

It so happens that in our case the doodads are stars; each $x$ is a vector of wavelengths, a vector of pixel fluxes and a vector of pixel flux uncertainties; each $y$ is the stellar parameters $T_{\text{eff}}$, $\log g$ and [Fe/H]; the special doodads are stars with fundamental parameters; and the black box is the LUMBA UVES stellar parameter pipeline. However, none of the mathematics discussed throughout Sect. 4.1 crucially depends on this. Anyone facing a situation similar enough that it can be summarised as in the paragraph above should be able to use an approach very similar to ours.

### 4.1.9 Using only Sun as benchmark

Early in our work on Article I, we only used the Sun as a benchmark, and not the full set of benchmarks. Our assumption was that the Sun is the ideal benchmark, so that once we got correct results for the Sun, the pipeline would be almost functioning. At that point, only a little tweaking would be needed to get good results for the entire benchmark sample. This turned out not to work, and looking in detail at what happened and why might help clarify why we ended up choosing the approach that we did.

By tweaking many details in the pipeline – in our minds optimising it – we did manage to get almost exactly correct stellar parameters for the Sun. Once we started testing other benchmark stars, we discovered something puzzling. The derived parameters for the benchmarks were in many cases very far off, even for stars very similar to the Sun. Most dramatically, we found that the derived $\log g$ for the Sun and the Solar twin 18 Sco consistently differed by 0.4 dex, four times as much as the true discrepancy of 0.1 dex. This even though both the observed and fitted model spectra for the stars are almost indistinguishable by eye. This meant that at least one of the implicit assumptions behind our work was wrong.

The mistake was not our assumption that the Sun is the best benchmark star: The stellar parameters genuinely are known with great precision[4], and there are Solar spectra with such high S/N that they are practically noise-free. The mistake was in our assumption that the performance for one high-quality benchmark star is necessarily representative for similar stars, which as described in Sect. 4.1.2 is not necessarily true: Sometimes a benchmark star will violate the assumptions built into a pipeline in a way that causes a measurement error that has nothing to do with noise, yet differs from otherwise similar stars.

This is what prompted us to develop an explicit model of the error instead of relying on our gut feelings for how errors should behave, and to include an error term that varies from star to star but is independent of the S/N and other qualities of individual spectra. This is the term $e_i$ in Eq. (4.1). Once we had this explicit model, it became clear that our attempts at 'optimising'

[4] In the case of [Fe/H], it is known exactly by definition.

the pipeline had in fact introduced a systematic error such that $e^{\mathrm{syst}} + e^{\mathrm{Sun}} \approx 0$, where $e^{\mathrm{syst}}$ is the systematic error for Sun-like stars while $e^{\mathrm{Sun}}$ is the error term $e_i$ for the Sun specifically. This created a pipeline that gave good results for the Sun, but not necessarily for any other star in existence. Improving the pipeline then required us to accept that the results would end up being worse for the Sun, even as they became better for most stars. Later, we learned that this is a known issue with SME: Analyses of the Sun often give strange results, and it is not entirely clear why. It probably has something to do with the observation conditions: Since it is not actually possible to point a spectrograph directly at the Sun, all 'Solar' spectra are actually spectra of sunlight reflecting off other celestial bodies, or Rayleigh-scattered day- or twilight (Thomas Nordlander, priv. comm.)

**Cause of apparent intuitive correctness of method**

I personally have a strong intuition telling me that the argument for using the Sun alone as a benchmark should be correct. It seems that a noise-free spectrum *should* give an estimate free of random error, and that it should be impossible for a deterministic process applied to data without random error to create an estimate that itself has a random error. Even so, we have empirically shown that this is not correct. That makes it interesting for me to try to understand where that intuition comes from and why it fails. (To readers who do not share this intuition, this section is likely to be completely uninteresting).

I believe that the answer is that the decomposition of an error into a random and a systematic part depends on what larger set the measurement is supposed to be sampled from. However, this is not sufficiently emphasised in statistics courses, which causes us – or at least me – to develop strong, incorrect intuitions about this.

As a concrete example, imagine that we use SME to estimate $T_{\mathrm{eff}}$ for some spectrum. This estimate will differ from the true value by some amount. Now we ask whether the pixel noise in the spectrum will contribute to the random or the systematic part of the error. My gut response is that since pixel noise is random, it should contribute to random error. In reality, it depends. If we imagine that this estimate is done as part of a larger study, where we estimate $T_{\mathrm{eff}}$ for several different spectra of the star, then the pixel noise will tend to contribute to the random error in the estimate. On the other hand, say that this estimate is done as part of a study to test the performance of SME by analysing that specific spectrum repeatedly with different starting parameters. In that case, the pixel noise will be constant from one estimate to the next, and therefore contributes to the systematic error. In short, the decomposition into random and systematic is not well-defined for a single measurement, but requires us to refer to some larger set of measurements.

In my experience, this is not emphasised in statistics courses. For example Taylor (1997) makes no mention of anything like this. That makes it easy to think of 'randomness' and 'systematicness' as being inherent properties of the noise, which one could then expect to propagate separately through an analysis.

## 4.2 Estimating parameters based on measured abundances

In Article II we estimate abundances of titanium, iron and magnesium for a sample of stars in the cluster M30. We then use those to estimate $T_0$, which is a free parameter in certain stellar evolution models. For any value of $T_0$, these models give some prediction of abundances as functions of $T_{\mathrm{eff}}$.

At first glance, this looks like a similar problem to that in Article I: We have a set of spectra for a group of stars, and we estimate some numerical quantity for those stars. It so happens that the quantity is elemental abundances instead of stellar parameters, but it would seem that we should be able to use an approach similar to that described in Sect. 4.1. What prevents us from doing so is that there are no relevant benchmark stars for abundances. If we use the term 'benchmark star' in the strict sense, as a star for which the relevant quantity has been estimated by means independent of spectroscopy, then the only abundance benchmark star is the Sun. We can estimate the abundances of the Sun by studying the composition of primitive meteorites, which formed from the same material as the Sun (Asplund et al., 2021). This prevents us from using the general approach described earlier.

Since we cannot use the more general method, we had to develop a method which is better adapted to the specific research question that we are trying to answer. In Article II itself, this method is essentially described as a finished product: The underlying assumptions are explained, a likelihood function is derived, and a framework for interpreting the likelihood function is given. Here, we try to explain the method in a format that more closely follows the work process, explaining issues that turned up along the way, and some other methods that could have been used, and why we rejected them.

### 4.2.1 Basic problem

The underlying problem that we attempted to solve in Article II can be summarised as:

We have a model of stellar evolution which predicts abundances in M30 as a function[5] of $T_{\mathrm{eff}}$. This model contains a number of free parameters, the most important being $T_0$ and the assumed initial abundance in the cluster. We are interested in finding out $T_0$, while the initial abundance is only interesting to the extent that it is needed to constrain $T_0$. Within a fairly wide range, shifting the initial abundance only adds a constant offset $a^{\mathrm{offset}}$ to the predicted abundances. We have made abundance estimates $a_i$ for stars in M30, for a number of spectral lines. We know $T_{\mathrm{eff}}$ for those stars relatively accurately, but $a_i$ is subject to both random scatter and systematic error. Given this, how well can we constrain $T_0$?

[5] Among our stars there is a one-to-one relationship between $T_{\mathrm{eff}}$ and $\log g$, so for convenience we will only speak of dependence on $T_{\mathrm{eff}}$ throughout this discussion.

We will start by discussing the effect of the systematic errors in the measured abundances, and then explain how we dealt with random scatter and the choice of initial abundance.

### 4.2.2 Parameter dependence in the systematic errors matter

One method which is commonly used in this type of situation is that of *differential analysis*. This is built on the assumption that the systematic errors are approximately constant over the entire range of $T_{\text{eff}}$. Given this, one can simply look at the difference in abundances between the RGB and the TOP, and select the value of $T_0$ for which the model predictions most closely reconstructs this. This would neatly cancel out both the systematic errors in the derived abundances, and the offset $a^{\text{offset}}$.

The problem with this approach is that the underlying assumption is false. In Article I, we showed that the systematic errors in derived quantities, including [M/H], have a strong dependence on the stellar parameters. In fact, since the estimates of [M/H] in Article I were made using multiple iron lines, whose individual quirks can be expected to cancel out to some extent, we should expect the effect to be stronger for the single-line abundances estimated in Article II.

### 4.2.3 Estimating bounds on the size of the systematic error

The reasoning in the last sections puts us in an uncomfortable position: We know that our systematic errors are large, and that they have a non-trivial dependence on $T_{\text{eff}}$. We also have no way of measuring them directly. We only know that our derived abundances $a_i$ are the sum of the true abundance $a_i^{\text{true}}$, some random error $e_i$ and some systematic error function $e_i^{\text{syst}}(T_{\text{eff}})$. That is

$$a_i = a_i^{\text{true}} + e_i^{\text{syst}}(T_{\text{eff}}) + e_i \tag{4.36}$$

To make the problem tractable, we start by attempting to find outer bounds on the systematic error.

As a general rule, the biggest source of systematic error in spectroscopic studies tends to be inaccuracies in the continuum placement. As discussed in more detail in Sects. 3.3 and 4.6, fitting observations to a stellar spectrum requires estimating the continuum level. This cannot be done exactly, and in general one cannot expect the estimates to be symmetrically distributed around the true continuum level. This creates one major source of systematic error.

In our specific case, a second major source of systematic uncertainty is the choice of $v_{\text{mic}}$. This parameter, which is discussed in more detail in Sect. 3.8, is as necessary as $T_{\text{eff}}$ and $\log g$ for making abundance estimates. Unfortunately, unlike $T_{\text{eff}}$ and $\log g$, it cannot be estimated from photometry. In the spectra observed in Article II, there are also no lines that can be used to reliably estimate $v_{\text{mic}}$ separately from the abundance estimates themselves[6]. The only tool left is to use empirical relations, which estimate $v_{\text{mic}}$ as a function of $T_{\text{eff}}$, $\log g$ and possibly [M/H], based on measurements found in the literature. Unfortunately, there are many empirical relations available, and in the process of testing them out, we found that they give very different results. In Sect. 10.1 we discuss in more detail the formulæ that we considered in the early stages of writing the article, and how much they differ in their predictions. For now, it is enough to say that we ended up settling on the 'Sitnova-Mashonkina' relationship, based on Sitnova et al. (2015) and Mashonkina et al. (2017), but this was a fairly arbitrary choice that would have to be taken into account as a major source of systematic uncertainty.

We now make the assumption that if we can put conservative upper bounds on the systematic errors stemming from continuum placement and choice of $v_{\text{mic}}$, they will be enough to also contain the systematic errors stemming from other sources. Given this we can get bounds on the systematic errors for each star by running nine analyses for each star, where each analysis takes some combination of the best, highest and lowest estimates for the continuum level and the choice of $v_{\text{mic}}$. Note that this implicitly assumes that the ranges of continuum placement and $v_{\text{mic}}$ are small enough that the derived abundance is a monotonous function of those parameters. Based on visual inspection we found an outer bound on the error in the continuum placement of $\pm 0.5\%$, and based on the comparisons detailed in Sect. 10.1 we found an outer bound on the error in $v_{\text{mic}}$ of $\pm 0.3\,\text{km/s}$.

### 4.2.4 Estimating what a plausible systematic error could look like

The approach described in Sect. 4.2.3 is sufficient to put a bound on the systematic error for each individual star. In principle, we could use this to estimate the likelihood for our derived abundances given a particular value of $T_0$, allowing the assumed true abundance of each star to lie in whichever position within that bound that maximised the likelihood. This would allow us to put some bounds on $T_0$, but it implicitly uses a very unrealistic model of the systematic errors: While we do know that the systematic errors vary enough that they cannot be treated as constant, we also know that they do not vary over very short scales of $T_{\text{eff}}$. The analysis we just sketched out does not make use of the last piece of knowledge, since it treats it as equally plausible that two stars with nearly-identical stellar parameters have nearly-identical systematic errors, and that they sit on opposite ends of the bounds. We would prefer an analysis that takes into account that while systematic errors are likely to vary over long intervals of $T_{\text{eff}}$, they do not vary on very short scales.

[6] See Sect. 10.1 for a discussion of our attempt to do this.

One way of doing this would be to simply declare that the systematic errors have to fit some slowly-changing function, such as a second-degree polynomial, or a third-degree polynomial, or something else. The problem with this approach is that the choice of function is essentially arbitrary, and it would not be clear how much the results depended on that choice. A more informative approach would be to select a range of functions with increasing complexity, and then somehow penalise the more complex functions when interpreting the results as evidence. That is, if two values of $T_0$ can give the same likelihood for generating the same observed abundances, but to do so one of them requires a much more complex model of the systematic error, then the value with a simpler error model should be treated as more plausible. The best tool we have found for doing this is *Akaike's Corrected Information Criterion*, $\mathrm{AIC_c}$. Formally, the $\mathrm{AIC_c}$ for a particular model in the light of some data is defined as

$$\mathrm{AIC_c} = 2k + \frac{2k\,(k+1)}{n-k-1} - 2\ln\hat{L}, \tag{4.37}$$

where $n$ is the number of data points, $k$ is the number of free parameters in the model, and $\hat{L}$ is the maximum likelihood of the data given the model.

We discuss the criterion in more detail in Sect. 10.2. For now, it is enough to understand the following:

- For each model, the data give some value of $\mathrm{AIC_c}$
- The absolute value of $\mathrm{AIC_c}$ is uninteresting
- The difference $\Delta\mathrm{AIC_c}$ between models can be used to estimate which one of them is closest to reality – for a very specific definition of 'closeness'
- The higher the likelihood of the data given a model, the better
- The more free parameters a model has, the worse

This means that we can get an $\mathrm{AIC_c}$ value for each combination of $T_0$ and assumed error model. In Article II, we display them in a table, with a layout similar to that sketched out in Table 4.2.

| | | | |
|---|---|---|---|
| $T_0$ lowest, simplest error model | $T_0$ second lowest, simplest error model | $\ldots$ | $T_0$ highest, simplest error model |
| $T_0$ lowest, second simplest error model | $T_0$ second lowest, second simplest error model | $\ldots$ | $T_0$ highest, second simplest error model |
| $\vdots$ | $\vdots$ | $\ddots$ | $\vdots$ |
| $T_0$ lowest, most complex error model | $T_0$ second lowest, most complex error model | $\ldots$ | $T_0$ highest, most complex error model |

**Table 4.2.** *Schematic picture of the table of combinations of stellar evolution models and systematic error models from Article II.*

As we progress into more and more complex models of the error, they will eventually be rated very low by the criterion, since the added complexity does not increase explanatory power. As it turned out, we found in Article II that there was a small range of $T_0$ values which were consistently preferred, irrespective of the error model chosen. In the worst case, we could have found that there were a range of $T_0$ values combined with different error models that were rated by $\mathrm{AIC_c}$ as having similar explanatory value. In that case, we would have had to conclude that our study was in fact so sensitive to systematics that no meaningful conclusions could be drawn from it.

The only question remaining is to make a choice of the error models. In Article II we choose to use polynomials of increasing order. There is no strong theoretical motivation for this, and arguably some other choice would have been more optimal. The only real advantage of this is that most people seem to chose polynomials, so it makes our results more likely to be easily comparable to others.

### 4.2.5 Group spectra

Aside from the systematic errors having a dependence on $T_{\mathrm{eff}}$, we found that they depend on S/N. This in turn depends on $T_{\mathrm{eff}}$, for the simple observational reason that the cooler stars are larger and therefore brighter. We could in principle have tried to bake this into our analysis of the systematic errors, but we instead solved it through the technique of creating ‘group spectra’. This method consists of averaging together spectra with similar parameters, effectively turning them into a higher-S/N spectrum with approximately those parameters.

The method assumes that adding together spectra from stars with similar stellar parameters will create a spectrum that behaves in the analysis essentially like a spectrum of a single star with higher S/N. This obviously requires the stellar parameters to be relatively close together – adding together a red giant and a TOP star will produce a spectrum resembling no star that could physically exist. With this method we created group spectra with approximately equal S/N, hopefully cancelling out this source of systematic error.

### 4.2.6 Intrinsic scatter in abundances

In Eq. (4.36) we showed our model of how estimated abundances differ from the true abundances. At the time, we skimmed over the issue that even if a stellar evolution model with some value of $T_0$ is exactly correct, this would not mean that the true abundance of a star is equal to that predicted by the model for a star with that effective temperature. There is some amount of intrinsic scatter among the stars, which needs to be taken into account.

We assume that this intrinsic scatter follows a normal distribution with mean zero around the abundances $a_i^{\text{trend}}(T_0)$ predicted by the stellar evolution models, with some standard deviation $\sigma_{\text{X, int}}$ which is specific to each element. For the standard deviations we chose 0.1 dex for magnesium, which is affected by anticorrelations, and 0.05 dex for titanium and iron, which are not. All of these are probably overestimates for stars that are part of the same cluster, so at worst they make our conclusions seem less certain than they actually are.

Our use of group spectra implicitly assumes that the variations between the spectra are small enough that the fluxes are linear functions of each parameter. Given this assumption, the intrinsic scatter for group spectra should be $\sigma_{\text{X, int}}/\sqrt{n_{\text{star}}}$, where $n_{\text{star}}$ is the number of stars in a group spectrum.

### 4.2.7 Random uncertainty in estimates

In addition to systematic errors and intrinsic scatter, there will be some random uncertainty in the derived abundances, due to pixel noise and similar observational effects. We assume that this random scatter follows a normal distribution with mean zero, and standard deviations $\sigma_i$ equal to the uncertainty reported by SME. However, this required us to decide which SME estimate to choose.

As described in Sect. 4.5, earlier versions of SME estimated uncertainties in derived quantities based on the covariance matrix generated in the fitting, as described in Piskunov and Valenti (1996). More recent versions instead estimate them using the heuristic algorithm described in Piskunov and Valenti (2017). While we believe this algorithm is an improvement in most cases, it is probably not appropriate for single-line abundances. Hence, we rewrote our copy of SME to use the older method.

### 4.2.8 Likelihood function

Taking into account the argument in Sect. 4.2.6, and the fact that the initial abundance is not known, the expression for the derived abundances (4.36) takes the following form:

$$a_i = a_i^{\text{trend}}(T_0) + a^{\text{offset}} + a_i^{\text{int}} + e_i^{\text{syst}} + e_i, \tag{4.38}$$

where $a_i^{\text{trend}}(T_0)$ is the value predicted by the stellar evolution models, $a^{\text{offset}}$ is some $T_{\text{eff}}$-independent offset stemming from the arbitrary choice of initial abundance in the stellar evolution models, $a_i^{\text{int}}$ is drawn from a normal distribution with mean zero and a standard deviation $\sigma_{\text{X, int}}$, $e_i^{\text{syst}}$ is a polynomial in $T_{\text{eff}}$ and $e_i$ is drawn from a normal distribution with mean zero and standard deviation $\sigma_i$.

Given the assumed random scatter described in Sect. 4.2.7, this means that the likelihood for a particular abundance estimate $a_i$ and element $X$ is given by

$$L\left(a_i\right) = \begin{cases} \dfrac{\exp - \frac{a_i^{\text{trend}}\left(T_0\right) + e_i^{\text{syst}} - a_i}{2\left(\sigma_i^2 + \sigma_{\text{X, int}}^2\right)}}{\sqrt{2\pi\left(\sigma_i^2 + \sigma_{\text{X, int}}^2\right)}}, a_i + e_i^{\text{syst}} \in \left[a_i^{\text{min}}, a_i^{\text{max}}\right] \\ \qquad\qquad 0, a_i + e_i^{\text{syst}} \notin \left[a_i^{\text{min}}, a_i^{\text{max}}\right], \end{cases} \tag{4.39}$$

where $\sigma_{\text{X, int}}$ for a group spectrum is rescaled as described in Sect. 4.2.6.

The $a_i$ are assumed to be uncorrelated by construction, since we have already built in correlations in the systematic error model. Hence, the likelihood for the full sequence of estimates is given by

$$L\left(\{a_i\}\right) = \prod_{i=0}^{n} L\left(a_i\right). \tag{4.40}$$

As in Article I, we use the Python module emcee to find the maximum of this function.

### 4.2.9 Recommendations for future studies

Based on these discussions, we can make some recommendations for similar studies.

Our first recommendations are fairly concrete and have to do with $v_{\text{mic}}$. Ideally, one should attempt to observe a spectral range that allows determining $v_{\text{mic}}$ directly. If that is not possible, the uncertainty in $v_{\text{mic}}$ has to be explicitly taken into account in the analysis: As shown in Sect. 10.1, there is no way of indirectly estimating $v_{\text{mic}}$ that is not to a great extent arbitrary. One should also attempt to predict ahead of time which lines are the most sensitive to $v_{\text{mic}}$. In our case, it happily turned out that of the lines we observed, Mg4704 was much less sensitive than the other lines, but we could just as easily have ended up discovering that all lines were highly sensitive, making our analysis essentially useless except as a cautionary example.

When selecting which lines to observe, study what the stellar evolution models actually predict for the elements in questions. The dependence of $a_i$ on $T_{\text{eff}}$ is not interesting in itself. What does matter is the dependence on $T_0$. That is, $\partial a_i / \partial T_0$. This should have as high a value as possible, at least within some range of $T_{\text{eff}}$. As a rough rule of thumb, it is better if it changes very quickly within a small range of $T_{\text{eff}}$, since it is less likely that systematic errors will create the illusion of a sharp shift in a very narrow range, than that they will create an apparent slow shift over a wide range.

### 4.2.10 Reliability of method

The method used in Article II was developed during the process of writing that article, and has never been tested before. When motivating the method, we started by in Sect. 4.2.2 essentially dismissing the method of differential analysis, which has been used for a very long time. This might make a reasonable reader unwilling to trust our results.

Our response is to agree that you should not trust the results of this method until it has been used by several other people, but to point out that if you do not trust this method, you probably should not trust the more common method of differential analysis either. In our analysis we take two sources of systematic error into account, and assume that the systematic error function will look either like a constant offset or something more complicated. Simply ignoring those sources of systematic error and assuming that the systematic error function has to be a constant offset would result in our analysis being equivalent to differential analysis. That is, our method differs from differential analysis essentially in being more pessimistic about how strongly our conclusions can be justified on the basis of our data.

### 4.2.11 Generality of method

As with the method for estimating the performance of pipelines for determining stellar parameters, the method we developed here can be generalised to many other situations. The underlying assumptions can be summarised as:

"We have made measurements of quantities $x$ for a large number of doodads. We have also done something complicated to get estimates of some quantity $y$ for the same doodads. However, we do not actually care about either $x$ or $y$. What we care about is a parameter $a$ in a model that predicts $y = f_a(x)$. We know that our measurements $y$ are subject to a systematic error $y_{\mathrm{err}} = g(x)$. Based on this, how well can we constrain $a$?"

It so happens that in our case the doodads are stars; each $x$ is the stellar parameters for the stars; each $y$ is the derived abundances for stars; and $f_a$ is a stellar evolution model. However, as long as it is possible to derive some absolute bounds on $g(x)$, something similar to the analysis discussed throughout Sect. 4.2 should be possible.

## 4.3 The ExtraTrees algorithm

In Article III we attempted to find out whether it is possible to use machine learning for estimating $[\alpha/\mathrm{Fe}]$ for BP/RP spectra observed with the Gaia. We picked the Extremely Randomised Trees (ExtraTrees) algorithm as representative of machine learning in general, on the assumption that it was close enough to optimal that it would behave qualitatively the same as more sophisticated algorithms. That is, if it was possible to pick out a signal in our data with the ExtraTrees algorithm, another algorithm might pick out a stronger signal, but it would not be able to pick out a signal where the ExtraTrees algorithm only saw noise. Due to space constraints, we did not discuss in the article how the algorithm actually works.

### 4.3.1 Formal definition

The ExtraTrees algorithm is an example of an *ensemble method*. The basic idea behind ensemble methods is fairly simple: Assume that we have some *base estimator*, which is a machine learning algorithm so naïve that it barely performs better than chance. Train very many such base estimators, in a way that ensures that their results are uncorrelated. Then combine their results in some way: If the task is classification, the natural choice is to let them vote, and if the task is regression, a natural choice is to take the average of all outputs. This can give surprisingly accurate results, despite the poor performance of each individual estimator. These algorithms are also highly parallelisable, since each base estimator is trained and runs without input from any of the others (Geurts et al., 2006, Sect. 1).

The base estimator of the ExtraTrees algorithm is the *decision tree* (Geurts et al., 2006, Sect. 2). In general, a decision tree is any algorithm that can be visualised as a tree graph, such that the algorithm starts at the root and then at each node makes a binary choice about how to continue down through the tree based on whether some feature of the data falls above or below a threshold value, and each leaf node is a verdict about what label to assign to the data (Shalev-Shwartz and Ben-David, 2014, Chap. 18). Specifically in the context of machine learning, a decision tree also requires some algorithm for deciding which choice to make at each node. This is typically done recursively, by at each node looking at the subset of the data that could reach the node, and somehow defining a choice of feature and threshold that best separates the labels. In the context of regression, rather than categorisation, 'best' is typically defined as minimising the mean-squared error (MSE). For the ExtraTrees algorithm[7], the choice of feature and threshold is made by initially generating a list of candidate criteria by randomly assigning some treshold to each feature. Out of this list, the node is assigned the criterion that best separates the data (Geurts et al., 2006, Sect. 2.1).

[7] As the name implies, most other algorithms involve less randomisation.

## 4.4 Comparisons to isochrones

In Article I we gauge the performance of the stellar parameter determination pipeline by comparing how close the derived parameters for stars in the cluster M67 are to the isochrone for that cluster. In that article, the comparison is done in $T_{\mathrm{eff}}$-$\log g$ space. In Article II we estimate stellar parameters for stars in the cluster M30 by shifting the stars to the closest point on the isochrone. In that article, the comparison is done in $V_{\mathrm{mag}}$–$(V-I)_{\mathrm{colour}}$ space. In both cases, we make caveats about how this is not as straightforward as it seems, and refer the reader to Valls-Gabaud (2014) for a lengthier discussion. Here, we attempt to describe the problem in more detail.

The fundamental problem is the same in both cases: Isochrones do not exist in a metric space. That is, whether we are working in $T_{\mathrm{eff}}$-$\log g$ space or $V_{\mathrm{mag}}$–$(V-I)_{\mathrm{colour}}$ space, there is no natural definition of the distance between two points. This means that the question of which point on the isochrone is 'closest' to the measured parameters for a star, and how 'close' it is, does not have a well-defined answer. This is most obvious in the case of $T_{\mathrm{eff}}$-$\log g$ space: Asking which is 'closest' of a point 100 K away and a point 0.1 dex away is clearly an apples-to-oranges comparison, and asking for the 'length' of a diagonal covering 100 K and 0.1 dex is simply meaningless, but the argument holds in $V_{\mathrm{mag}}$–$(V-I)_{\mathrm{colour}}$ space as well.

Since we are using the isochrone for parameter estimation, we would like to define distance in such a way that the closest point on the isochrone is that with the highest probability of generating the measurement. Unfortunately this runs into problems. First of all, we need well-defined errors on our measurements. In the case of the $T_{\mathrm{eff}}$ and $\log g$ estimates in Article I we do have uncertainties, but we do not for the $V_{\mathrm{mag}}$ and $(V-I)_{\mathrm{colour}}$ estimates in Article II. The second problem needs to be phrased differently depending on whether we want to use Frequentist or Bayesian language. In Bayesian terms, we would say that we need a prior on the measured parameters. That is, we need to know how likely a generic star is to be at a particular point on the isochrone. Without a detailed physical model of stellar clusters, there is no easy way to get such a prior. One could attempt to get around the problem by choosing a uniform prior, so that the probability density is the same across the entire isochrone. This runs into the problem we described in the beginning: Since this is not a metric space, there is no uniquely defined 'uniform' probability density. That is, there is nothing that tells us that a segment of the isochrone covering $x$ K along the $T_{\mathrm{eff}}$ axis should contain the same amount of probability mass as a segment covering $y$ dex along the $\log g$ axis. The problem is more difficult to express clearly in Frequentist terms, although it is known *a priori* that the problem must be there since Frequentist parameter fitting is equivalent to Bayesian parameter fitting that starts out with a uniform prior and discards all but the maximum of the posterior. One way of expressing it is that the problem is very sensitive to the choice of parametrisation, so that if we arbitrarily start measuring effective temperature in tens of K or surface gravity in tenths of dex, we could get very different results[8].

We do not attempt to find a fundamental solution to this problem. As discussed in Valls-Gabaud (2014), the problem is well-known and unsolved. Instead, we simply note that every time we use isochrones the results should be taken with a few grains of salt.

[8] Note that this strictly speaking applies to all Frequentist parameter fitting. One can always force a Frequentist parameter fit to give a particular result simply feeding it the problem parametrised in a sufficiently perverse way. In most cases, however, there is some unique parametrisation which most people will agree is the sensible one to use. In this case, there is no such natural parametrisation.

## 4.5 Error estimation

SME offers two ways of calculating the statistical errors in the estimated parameter $p$. Neither method is completely ideal, but each one is the more appropriate in certain situations.

The first method estimates the error in each fitted parameter based on the diagonal of the covariance matrix generated by the fitting algorithm. This effectively assumes that the only source of error is pixel noise. This method was the one used in older versions of SME (Piskunov and Valenti, 1996).

The second method uses a heuristic algorithm that tries to take into account the existence of systematic errors. It assumes that the estimates are based on fitting to multiple lines at once, and tries to take into account differing systematic errors from line to line due to uncertainties in the line parameters (Priv. comm. Thomas Nordlander). It works by first picking out a subset of pixels. Initially, each pixel $i$ which is used in the fitting and whose flux has a non-zero derivative with respect to $p$ is selected. Then, out of those the algorithm selects the ones for which the residual $R_i$ between the observed and best-fit flux $F_i$ is less than 5 times the flux uncertainty. As long as the model flux is approximately a linear function of $p$, the change $\Delta p$ needed for the residual to become zero is given by

$$\Delta p = \frac{R_i}{\partial F_i / \partial p} \tag{4.41}$$

All the pixels together give a distribution of $\Delta p$ values. The width of the central 68% is taken as the best estimate of the uncertainty in $p$ (Piskunov and Valenti, 2017). This method is used from SME version 433 onwards. That said, it is easy to restore the older method, by commenting out some lines in the SME code file `sme_solve.pro`.

We used the second method for Article I, since the stellar parameter pipeline that we were testing does make estimates based on multiple lines, as implicitly assumed by the method. However, in Article II we were working with abundances estimated from single lines. Hence, we modified our SME code to restore the first method.

## 4.6 Continuum mask algorithm

As described in Sect. 3.10.2, early versions of SME need a user-defined mask to fit the continuum. In Article I we developed an algorithm for defining that mask. For reasons of space, the description of the algorithm ended up relegated to Appendix D in the article, where it is presented with no real justification. Here, we attempt to clarify why that algorithm is defined as it is.

There are two main sources of error when fitting the continuum. There is the observational problem that each pixel is affected by Poissonian noise. There is also the theoretical problem that our models never have a full line list, meaning that pixels which we think are pure continuum may actually be part of a faint spectral line. If only either one of these error sources existed, and the continuum was simply offset by a constant factor, there would be a single demonstrably correct way of defining the continuum mask.

If we had a complete line list but pixel noise, the solution would be to assign the brightest model pixels to the continuum mask, on the assumption that these are closest to the continuum. What fraction of the total number of pixels should be considered the brightest would need to be tested out empirically, but there would probably be a fairly wide range of viable values. This is essentially how the pipeline attempted to define the mask in versions of the pipeline prior to Article I. This had the problem that occasionally, there would be a fairly deep line inside the continuum mask, which would noticeably lower the fitted continuum relatively to the true level. Presumably, there were also other cases where shallower unknown lines lowered the fitted continuum to an extent not immediately obvious to the eye.

If we had no pixel noise but an incomplete line list, the solution would be to assign the brightest observed pixels to the continuum mask, on the assumption that these are free of missing lines. Again, this would require an arbitrary choice of what fraction should be considered the brightest. If this algorithm was used on a real-world noisy spectrum, it would systematically place the fitted continuum too high, since given a number of genuine continuum pixels, it would preferentially select those where the noise was unusually bright over those where the noise was unusually dim.

In short, the algorithm that would be correct in one idealised case will tend to push the continuum too low when applied to real-world data, and the algorithm that would be correct in the other idealised case would tend to push the continuum too high instead. Hence, we invent a hybrid algorithm which tries to compromise between them. In addition, the algorithm tries to avoid atmospheric emission lines.

Formally the algorithm is defined as follows. We introduce the parameters $c_{\mathrm{model}}$, $c_{\mathrm{obs}}$ and $e_{\mathrm{frac}}$[9]. First, a preliminary fitting is done by fitting a straight line to the brightest points in the spectrum. This hopefully mostly gets rid of

[9] In Article I we use a slightly different notation using $c_{\mathrm{frac}}$ and $c_{wt}$, defined such that $c_{\mathrm{model}} \equiv c_{\mathrm{frac}}$ and $c_{\mathrm{obs}} \equiv c_{\mathrm{frac}} \cdot c_{wt}$. This is hopefully slightly clearer.

the slope of the spectrum. In the next step we assign to the continuum mask those pixels such that

1. The observed intensity is among the $c_{\text{obs}} + e_{\text{frac}}$ brightest, but not among the $e_{\text{frac}}$ brightest
2. The model intensity is among the $c_{\text{model}} + e_{\text{frac}}$ brightest.

In the case that this results in fewer than ten pixels being assigned to the continuum mask in either the left or right one-third of the segment, $c_{\text{obs}}$ and $c_{\text{model}}$ are gradually scaled up until they include ten.

Among the criteria, removing (1) would recover the algorithm that would be optimal under perfect modelling, while removing (2) would recover the algorithm that would be optimal with perfect observations. The parameter $e_{\text{frac}}$ has the effect of allowing us to remove pixels that are affected by emission. Empirically, we got the best results around $c_{\text{obs}} = 0.2$, $c_{\text{model}} = 0.6$, and $e_{\text{frac}} = 0.005$.

## 4.7 Coaddition and cleaning of contaminants

In Article II, we worked with a set of spectra observed with the GIRAFFE spectrograph. These had been observed over multiple observation nights, so that there were around 30 individual spectra for each faint star, and around 5 for the brighter stars. These spectra had low signal to noise, with the S/N reported by the GIRAFFE spectral reduction pipeline going as low as 3.2. Initially we also worked with spectra from Scheutwinkel (2019), which were also badly contaminated by cosmic rays. There was roughly one per 10 Å in the spectra from the individual observation nights, meaning that there would be three per 1 Å in most spectra if we simply averaged the spectra together. Hence, we implemented an algorithm for removing cosmic rays. At the end, this turned out to be unnecessary, since other issues with the files prompted us to download spectra from the ESO Archive Science Portal (Alberto et al., 2019). These had already been cleaned of cosmics, making that part of the algorithm reduntant. For completeness, we describe the algorithm here.

First a barycentric correction is applied to each spectrum to be co-added. This ensures that the spectra have a common wavelength scale. A full radial velocity determination is not done until after the final co-addition, since the individual spectra are too noisy to permit this to be done accurately.

We denote the individual spectra for a star by $k$ and the pixels of the spectrograph by $i$. We denote the number of spectra for star $k$ by $n$ and the number of pixels per spectrum by $N$.

Each spectrum is given a preliminary normalisation, by dividing the pixel fluxes $f_{i,k}$ by the integrated flux $a_k$ over the spectrum. Each normalised flux should then follow an approximate normal distribution centered around the true flux, up to some overall normalisation constant.

We calculate a preliminary weighted average $\overline{f}_{\mathrm{wt}}$ over all spectra for each pixel.

$$\overline{f}_i = \frac{\sum_{k=0}^{n} \frac{f_{i,k}/a_k}{\left(\sigma_{i,k}/a_k\right)^2}}{\sum_{k=0}^{n} \frac{1}{\left(\sigma_{i,k}/a_k\right)^2}} \tag{4.42}$$

where $\sigma_{i,k}$ is the uncertainty in pixel flux $f_{i,k}$.

We then calculate the differences $\Delta f_{i,k} \equiv f_{i,k} - \overline{f}_i$. For pixels unaffected by cosmics, this will follow a normal distribution with mean 0 and standard deviation

$$\sigma_{\overline{f}_{i,k}} = \sqrt{\left(\frac{\sigma_{i,k}}{a_k}\right)^2 + \sigma_{\overline{f}_i}^2 + 2\sigma_{f_{i,k},\overline{f}_i}} \tag{4.43}$$

where

$$\sigma_{\overline{f}_i} = \frac{1}{\sqrt{\sum_{k=0}^{n} \frac{1}{\left(\sigma_{i,k}/a_k\right)^2}}} \tag{4.44}$$

and

$$\sigma_{f_{i,k}\overline{f}_i} = \frac{1}{\sum_{l=0}^{n} \frac{1}{\left(\sigma_{i,l}/a_l\right)^2}} \tag{4.45}$$

Let us now assume that we are willing to have an algorithm for flagging cosmics strict enough that on average $n^{\text{type1}}$ pixels will be flagged in a spectrum entirely lacking cosmics. In that case, we should flag all pixels with fluxes more than $c\sigma_{\overline{f}_{i,k}}$ above $\overline{f}_i$, where

$$c = \sqrt{2}\text{erf}^{-1}\left(1 - \frac{2n^{\text{type1}}}{Nn}\right) \tag{4.46}$$

Since cosmic rays generally affect several adjacent pixels, but not necessarily all of them enough to raise them above this cutoff, we also flag the 5 nearest to either side as well. Note that while this will result in more than $n^{\text{type1}}$ pixels being incorrectly flagged as cosmic rays, this is not a problem: Our reason for imposing $n^{\text{type1}}$ is concern with the bias that comes from truncating the normal distribution followed by non-cosmic pixels. Incorrectly flagging pixels because they happen to be close to an incorrectly flagged pixel does not impose any bias, it only decreases the amount of data.

In the final spectra, each pixel flux is the weighted average over all spectra, but with flagged pixels given zero weight. The uncertainties are calculated according to (4.44), again with flagged pixels treated as though $\sigma_{i,k} = \infty$.

This is implemented in a Python module which itself makes use of the modules SciPy, Astropy (Virtanen et al., 2020; Astropy Collaboration et al., 2013, 2018). Once the coaddition is finished, the radial velocity is estimated by comparison to the star HD 140283, using an algorithm by Ansgar Wehrhahn: First a cross-correlation is used to derive a first guess. Then a $\chi^2$-minimisation is performed, on the assumption that the spectra are now close enough for this not to get stuck in a local minimum (Wehrhahn, 2020).

# 5. Summary of articles

## 5.1 Article I

As described in Sect. 2.3.2, GES contains multiple *nodes* tasked with estimating stellar parameters and abundances from spectra taken with the UVES spectrograph. One of those is the Lund-Uppsala-MPIA-Bordeaux-ANU (LUMBA) node. During the work on Internal Data Release 5 (iDR5), Gregory Ruchti of Lund University was in charge of doing the UVES analyses for LUMBA. Partway through, he fell ill, and the author had to take over.

After iDR5 was finished, we continued work on developing and improving the LUMBA pipeline. This involved developing a method for measuring what was an improvement. This led to Article I, which describes the pipeline. It describes how the pipeline works in enough detail that we believe that it is completely replicable. That is, a dedicated spectroscopist could write a practically equivalent pipeline based on the information in the article. However, for reasons of brevity we had to leave out some details of why the pipeline was written the way it is. This is a bit unfortunate, since in reality it is more likely that a spectroscopist would want to know the general principles of building pipelines than to want to replicate our pipeline in detail.

### 5.1.1 Scientific approach in article

The pipeline attempts to fit model spectra to the observed spectra over wavelength ranges that cover spectral lines sensitive to the stellar parameters, while avoiding wavelengths that are known to be difficult to model. To do the fitting the pipeline uses the spectral synthesis code SME (described in Sect. 3.10.2), together with a grid of stellar atmospheres calculated with the MARCS code (described in Sect. 3.10.1) and the GES line list, described in Heiter et al. (2021).

All of the lines used are sensitive to $T_{\mathrm{eff}}$ and $v_{\mathrm{mic}}$, in ways that in concert is not degenerate with the other parameters. They are also sensitive to the broadening coming from $v_{\mathrm{mac}}$ and $v \sin i$ together, in such a way that we can estimate the overall broadening separately from the other parameters, but not distinguish $v_{\mathrm{mac}}$ and $v \sin i$. This is in practice not a problem, since those parameters are generally not of interest on their own. Finally, each line is either jointly sensitive to either $\log g$ and some metal abundance, or only to the abundance:

- Strong neutral calcium lines: Sensitive to $\log g$ and the calcium abundance
- Weak neutral calcium lines: Sensitive to the calcium abundance
- Strong neutral magnesium lines: Sensitive to $\log g$ and the magnesium abundance
- Weak neutral magnesium lines: Sensitive to the magnesium abundance
- Singly-ionised iron lines: Sensitive to $\log g$ and [M/H]
- Neutral iron lines: Sensitive to [M/H]

Together, this allows the pipeline to estimate $T_{\rm eff}$, $\log g$, [Fe/H], $v_{\rm mic}$, as well as the joint broadening from $v_{\rm mac}$ and $v \sin i$. Note that the pipeline cannot reliably determine $v_{\rm rad}$.

The performance of the pipeline was estimated by testing the pipeline on *benchmark stars* – stars with stellar parameters already known from methods independent of spectroscopy. This was done using the Gaia FGK benchmark stars, which are described in 4.1.1. The development for a reliable method of benchmarking stellar parameter pipelines was probably a more important outcome of the work process than the pipeline itself, but it gets relatively little space in the article. Hence, we describe the method in detail in Sect. 4.1. In brief, the performance of a pipeline is characterised in terms of the parameters $e^{\rm pipe}$, $\sigma^{\rm star}$ and $\sigma^{\rm spec}$. The parameter $e^{\rm pipe}$ describes a systematic offset, which is assumed to be constant for all benchmark stars of a particular type. The parameter $\sigma^{\rm star}$ describes a random scatter which varies from star to star, but not between spectra of a particular star. The parameter $\sigma^{\rm spec}$ describes a random scatter that varies from spectrum to spectrum.

### 5.1.2 Main conclusions in article

Testing the pipeline on the Gaia-ESO FGK benchmark spectra resulted in the estimates of the performance parameters shown in Table 5.1 on page 86. Based on this, we concluded that the pipeline was accurate enough for the type of work it had been designed for, although performance was noticeably worse for red giants.

### 5.1.3 Important conclusions not in the article

During work on the pipeline, we made some useful discoveries about what not to do. This is not the kind of material that gets included in articles, but it is arguably of equal interest to other spectroscopists as the pipeline itself is.

As described in Sect. 3.6.2, when the pipeline first came into our hands it used the Balmer lines H$\alpha$ and H$\beta$ to provide a constraint on $T_{\rm eff}$. It used a relatively complicated algorithm for defining the line mask specifically for these lines. During development we gradually removed both lines from the analysis, and found that this improved the results greatly, removing the major part of our spectrum-to-spectrum scatter. Our initial worries that we would be unable to constrain $T_{\mathrm{eff}}$ without a specifically $T_{\mathrm{eff}}$-sensitive line turned out to be unfounded, since all lines have some $T_{\mathrm{eff}}$ dependence which is often not degenerate with the other parameters. Based on this, we would recommend against any attempts at using Balmer lines for automatic parameter determination, although they may still be used in analyses where the line masks are defined by hand for each spectrum.

As described in Sect. 4.1.9, we originally used the Sun as a benchmark to test the performance of the pipeline. Our assumption was that since the Sun has very well-known stellar parameters, and the spectra have such high quality that they are practically noiseless, this would be the ideal test. This turned out to be a mistake: The Sun has different systematic errors than other stars with almost identical stellar parameters, and so our attempts at tweaking the pipeline to give good results for the Sun resulted in it only giving good results for the Sun. Based on this, we would recommend against ever using a single benchmark star to evaluate the performance of a pipeline, and to use an approach like that described in Sect. 4.1 instead.

As described in Sect. 4.1.7, GES uses a different method for estimating the performance of stellar parameter pipelines. This method implicitly assumes that the systematic error will be almost identical for stars with almost identical stellar parameters. At least for our pipeline, this assumption is not even approximately true. Similar stars can have different systematics, due to quirks unique to each star.

**Table 5.1.** *Best estimates of the error parameters for the stellar parameters $T_{eff}$, $\log g$ and $[M/H]$ for the four types of benchmark star. Offset and spectrum-to-spectrum scatter specific to the Sun are also included.*

| Stellar parameter | Star type | $e^{\mathrm{pipe}}$ | $\sigma^{\mathrm{star}}$ | $\sigma^{\mathrm{spec}}$ |
|---|---|---|---|---|
| $T_{\mathrm{eff}}$ | The Sun | $-25\pm4$ | ... | 10 |
| – | Solar-type | $-30\pm40$ | $80\pm40$ | $20\pm40$ |
| – | F dwarfs | $50\pm40$ | $190\pm40$ | $20\pm40$ |
| – | FGK subgiants | $0\pm20$ | $50\pm20$ | $60\pm40$ |
| – | Red Giants | $50\pm20$ | $30\pm20$ | $70\pm20$ |
| $\log g$ | The Sun | $-0.070\pm0.004$ | ... | 0.01 |
| – | Solar-type | $-0.04\pm0.03$ | $0.07\pm0.03$ | $0.02\pm0.04$ |
| – | F dwarfs | $0.08\pm0.05$ | $0.13\pm0.08$ | $0.08\pm0.06$ |
| – | FGK subgiants | $0.02\pm0.03$ | $0.11\pm0.03$ | $0.15\pm0.03$ |
| – | Red Giants | $0.15\pm0.01$ | $0.14\pm0.02$ | $0.12\pm0.02$ |
| [M/H] | The Sun | $0.008\pm0.003$ | ... | 0.008 |
| – | Solar-type | $0.003\pm0.004$ | $0.020\pm0.002$ | $0.016\pm0.002$ |
| – | F dwarfs | $0.019\pm0.008$ | $0.080\pm0.008$ | $0.034\pm0.008$ |
| – | FGK subgiants | $-0.03\pm0.02$ | $0.05\pm0.02$ | $0.06\pm0.02$ |
| – | Red Giants | $-0.07\pm0.05$ | $0.05\pm0.05$ | $0.06\pm0.05$ |

## 5.2 Article II

As described in Sect. 2.2.2, the elemental abundances seen in the atmospheres of stars in a stellar cluster vary with $T_{\mathrm{eff}}$, due to several processes that cause elements to gradually rise or sink over long time scales. Many of these processes are well understood, but there is an *Additional Mixing or Transport Process* (AddMix) that we still do not understand well. There are models of stellar evolution which attempt to model AddMix while making as few assumptions as possible. This lack of assumptions comes at the cost of introducing a large number of free parameters, which must be estimated empirically. In Article II we attempt to estimate the parameter $T_0$ in such a model.

### 5.2.1 Scientific approach in article

We had theoretically predicted abundances as a function of $T_{\mathrm{eff}}$ for a number of different elements, assuming a range of different values of $T_0$

Our observations consisted of 177 spectra taken with the GIRAFFE spectrograph. We coadded these spectra into group spectra with a $\mathrm{S/N} \approx 100$. The stellar parameters $T_{\mathrm{eff}}$ and $\log g$ were already known from photometry, while $v_{\mathrm{mic}}$ was estimated using an empirical relation. There were five spectral lines that were visible over the entire $T_{\mathrm{eff}}$ range covered by the spectra. We show the lines in Table 5.2. Two of these were left out of the analysis, the Ba4554 since there are no model predictions for that element, and Ti4571 because that line turned out to be saturated, as described in Sect. 3.6.4.

| Line name | Species | Line centre [Å] | Comment |
|---|---|---|---|
| Ba4554 | Ba II | 4554.0290 | Not used |
| Ti4563 | Ti II | 4563.7574 | |
| Ti4571 | Ti II | 4571.9713 | Not used |
| Fe4583 | Fe II | 4583.8292 | |
| Mg4702 | Mg I | 4702.9909 | |

**Table 5.2.** *The spectral lines studied in Article II.*

For the group spectra we derived stellar abundances, using a pipeline which is essentially a simpler version of that described in Article I. We then compared these abundances to those predicted by the models. This comparison required us to take $T_{\mathrm{eff}}$ dependent systematic errors into account. This was a fairly involved process, which is described in Sect. 4.2.

### 5.2.2 Main conclusions in article

We found that $\log_{10}(T_0/[\mathrm{K}])$ is most likely somewhere in the range 6.09-6.20. This has an immediate implication for Paper VI in the same article series, Gruyters et al. (2016). That paper used a value of 6.0 and found an initial abundance for lithium of $A\,(\mathrm{Li})_{\mathrm{init}} = 2.48 \pm 0.10\,\mathrm{dex}$. This was a surprise at the time, since it is quite far below the abundance of 2.72 dex predicted from BBN models. Article I compounds this surprise, since the range 6.09-6.2 corresponds to a starting abundance of 2.42-2.46 dex.

### 5.2.3 Implication for similar studies

During the work on the article, we also learned several things of possible interest to anyone planning a study of this kind.

First of all, many studies assume that systematic errors only affect the measured abundances through a constant, $T_{\mathrm{eff}}$ independent, offset. We found that at least in our case, this assumption is very dubious. We found that in what we considered the best implementation of our analysis, models of the systematic error as a constant and linear offset were about equally compatible with the data (See Sect. 4.2 for details). We also found that relatively small changes in the implementation – such as switching to the continuum fitting algorithm described in Sect. 4.6 – were enough to entirely rule out constant-offset models of the error.

Secondly, the study also started out with the assumption that as long as $T_{\mathrm{eff}}$, $\log g$ and $[\mathrm{M/H}]$ are already known from photometry, it is easy to estimate $v_{\mathrm{mic}}$ using some empirical relation. For example, this is the approach taken by Paper VI. We discovered that in fact, there is a wide range of possible empirical relations to choose from. We show the full range considered in Appendix 10.1. This choice of relation has a strong impact on abundances on the cool end of the temperature range. We also concluded that the relation used in Paper VI is probably not appropriate: It was estimated based on Pop I stars, while M30 is made of Pop II stars. When we used it on our spectra it gave physically impossible derived abundances at the cool end, in the sense that the abundance trend did not level out at any point.

Finally, we found a bug in the Linux libraries for SME versions 503, 520, 536 and 542. The bug has since been reported and fixed. To the best of our knowledge, the effect of the bug is slight, but we recommend any author whose work depends on Linux SME of version 542 or earlier to download a more recent version and verify this by rerunning at least part of the analysis.

### 5.2.4 1D-LTE, 1D-NLTE and 3D-LTE

There was one part of the study that was originally intended to be a fairly major part of the article, but which unfortunately turned out to be infeasible. As described in Sect. 3.8, SME by default uses the approximations of one-dimensional stellar atmospheres (1D), and *local thermodynamic equilibrium* (LTE). This is computationally efficient, but gives worse results than a full 3D-NLTE analysis. While we did not have the ability to do a full 3D-NLTE analysis, we did have the ability to make either a 3D-LTE analysis or a 1D-NLTE analysis. Hence, we planned to run some lines in 3D-LTE and some in 1D-NLTE, depending on which was most suitable for that element, by the criteria below. A a sanity check we would also compare to the results of taking the other option, as well as simple 1D-LTE.

For the 3D-LTE calculation, we did spectrum synthesis using the SCATE code on the STAGGER grid of 3D hydrodynamic model atmospheres (Hayek et al., 2011; Magic et al., 2013). We sampled the temporal evolution by using 20 snapshots, selected at regular intervals, from each hydrodynamic simulation.

For titanium and iron, we could not initially make a definite statement that either 3D-LTE or 1D-NLTE was better. For the Mg4702 line, we did expect 1D-NLTE to be much better. In fact, 3D-LTE might actually be worse for that line than mere 1D-LTE (Nordlander, T. et al., 2017). The reason for this is that Mg I, unlike Ti II and Fe II, takes very little energy to ionise. This makes it much more sensitive to the ultraviolet (UV) flux from deeper layers in the atmosphere. Since this energy transfer is not local, it violates the LTE assumption. By Planck's law the UV flux is strongly temperature-dependent. Since 3D models tend to have steeper temperature gradients than 1D models, they have to take NLTE into account to simulate this species sensibly (Thomas Nordlander, Priv. comm).

As it turned out, the 3D calculations did not work. When we applied 3D corrections to the derived abundances, we found that they did not level out at the cool end of the temperature range, similarly to our results using a poorly-chosen method for estimating $v_{\mathrm{mic}}$. We believe the underlying cause is line saturation, as described in Sect. 3.6.4. Because of this, we had to move the 3D results from the main body of the article to Appendix D.

## 5.3 Article III

The Gaia space observatory described in Sect. 2.3.1 is primarily designed for astrometry. However, it does have the ability to measure very low-resolved spectra, using the Blue Photometer (BP) and Red Photometer (RP). The combined BP/RP-spectra have a resolution varying from 13 to 85 over the spectral range, too low for individual spectra lines to be visible.

The conventional spectroscopic methods described in Chap. 3 all assume that the spectra have distinct spectral lines. Once the resolution is poor enough that spectral lines blend into each other and the surrounding continuum, they become inapplicable. It will still be possible to infer stellar parameters like $T_{\mathrm{eff}}$, which affect the overall shape of the spectrum, but one could assume that the spectra will contain no information about elemental abundances. A counter-argument to this assumption would be that even if an individual spectral line is unresolved, an elemental abundance affects the depth of all lines of that element, as well as the continuum. As long as the element has large numbers of lines, this should affect the overall shape of the spectrum in some measurable way.

In practise, it has already been shown that BP/RP spectra can be used to estimate the overall metallicity $[\mathrm{M}/\mathrm{H}]$, if not individual abundances (Liu et al., 2012). In Article III, we investigate the possibility of going one step further and measuring the $\alpha$-element abundance $[\alpha/\mathrm{Fe}]$. This is not the same as measuring individual elemental abundances, but it is enough to sort stars into different populations.

It would in principle be possible to do this analogously with the more conventional spectral synthesis described in Sect. 3.7. One could generate model spectra, and then fit those to the observed spectrum. The problem is that since the analysis would have to cover a very wide wavelength interval, it would be necessary to know how the overall normalisation changes over the spectrum. Instead of simply knowing the resolution of the BP and RP, it would be necessary to have a detailed model of the Gaia instrumental profile, as well as the effects of interstellar extinction. We do have such a model[1], but it is known not to be perfect, to the extent that we did not trust our ability to do so reliably.

In short, we know that there should be some deterministic relationship between stellar parameters and the observed spectra, but we do not have a reliable explicit model of that relationship. This is the situation that machine learning methods excel at dealing with. As long as there is a subset of Gaia spectra that do have known $[\alpha/\mathrm{Fe}]$, they can be used to train a model to figure out the relationship between spectra and $[\alpha/\mathrm{Fe}]$.

[1] Information about the model of the Gaia instrumental profile will be publicly released as part of Gaia DR3. All statements made here about the instrument model are based on personal communications with Rosanna Sordo.

### 5.3.1 Scientific approach in article

We selected the ExtraTrees algorithm as a representative of machine learning methods in general. The method is described in more detail Sect. 4.3. The important assumption behind this choice is that the performance of the ExtraTrees algorithm is sufficiently close to the best algorithms that if it revealed itself to be completely incapable of determining $[\alpha/\mathrm{Fe}]$, no better algorithm would be likely to manage either.

We initially tested training an ExtraTrees model on a sample of spectra with stellar parameters known from Data Release 2 of the GALAH survey (the survey is described in Sect. 2.3.2). We used cross-validation to verify that it could approximately recreate the parameters of a sample statistically similar to its training sample. Then we applied the model to a sample of Gaia spectra without known parameters. This showed a realistic Galactic structure, with a $[\mathrm{Fe}/\mathrm{H}]$-rich, $\alpha$-poor Galactic Disk, and vice-versa for the Halo. This was made into a Gaia *Image of the Week* (Gavel et al., 2020).

However, while this did show that the model gave $[\alpha/\mathrm{Fe}]$ estimates that were correct on average, it could not tell us whether it was actually using the causal effect of $[\alpha/\mathrm{Fe}]$ on the spectrum, or merely the correlation of $[\alpha/\mathrm{Fe}]$ with other parameters that have stronger effects on the spectrum. For models using the ExtraTrees algorithm, it is in general very difficult to get a clear idea of what features of the data it has actually learned to use – they are sometimes referred to as *black box models* for that reason.

To test this, we generated a grid of synthetic spectra, which lacked the correlations of the GALAH sample. These spectra were not – for the reasons explained above – expected to look exactly like genuine, observed spectra. However, they were realistic enough that we believed they could reveal whether the information needed by the ExtraTrees algorithm was even present in the spectra, and second whether the model trained on observed data was actually seeing the causal effect of $[\alpha/\mathrm{Fe}]$ or the correlation with other parameters.

We trained a model on the synthetic spectra and again used cross-validation to verify that it could approximately recreate the parameters of a sample statistically similar to its training sample. However, when we applied it to the GALAH sample we found it could not give useful $[\alpha/\mathrm{Fe}]$ estimates, even if it did relatively well for other stellar parameters. Similarly, the model trained on the GALAH sample could not give useful estimates for the synthetic sample.

This showed that the models had not learned to use the causal effect of $[\alpha/\mathrm{Fe}]$ on the spectra, but some indirect correlation with other parameters. At the most trivial, it could be that the model was simply using the Galactic trend of $[\alpha/\mathrm{Fe}]$ as a function of $[\mathrm{Fe}/\mathrm{H}]$.

To test this possibility, we acquired a sample from the Gaia-Enceladus structure (the structure is described in Sect. 2.2.5). This sample is observationally similar to the other Gaia sample, but has a different trend of $[\alpha/\mathrm{Fe}]$ as a function of $[\mathrm{Fe}/\mathrm{H}]$. We found that the estimates by a model trained on the

GALAH sample did not follow the true trend of Gaia-Enceladus, but that they also did not simply follow the overall Galactic trend.

This again showed that the model was relying on indirect correlations, but also verified that it did not only use the correlation with [Fe/H]. Even though we had no additional means of finding out what those correlations were, it showed that the estimates were non-trivial enough to be potentially useful.

### 5.3.2 Main conclusions in article

It is possible to estimate $[\alpha/\mathrm{Fe}]$ based on Gaia BP/RP spectra using machine learning methods such as ExtraTrees, but there are several caveats that need to be kept in mind: While a model trained on the GALAH sample will produce $[\alpha/\mathrm{Fe}]$ that are on average fairly close to the correct values, it does not appear to do so by using the causal effect of $[\alpha/\mathrm{Fe}]$ on the shape of the spectrum. Instead, it uses the correlation between $[\alpha/\mathrm{Fe}]$ and other parameters, which do affect the spectrum. This means that while the method does give $[\alpha/\mathrm{Fe}]$ estimates that are good on average, it cannot distinguish two stars that only differ in $[\alpha/\mathrm{Fe}]$.

For cool dwarf stars it might be possible to actually use the causal effect of $[\alpha/\mathrm{Fe}]$ on the spectrum, since they have strong titanium oxide molecular lines which very noticeably affect the spectrum. Unfortunately, those stars are inherently less bright, meaning that those observed by Gaia sample a relatively small fraction of the Galactic volume.

# 6. Final summary and outlook

The field of Galactic archaeology depends crucially on the techniques of stellar spectroscopy. In many cases, these techniques suffer from systematic errors that are of comparable size to the physical effects that current studies are looking at. This means that the standard methods of parameter fitting and hypothesis testing that are taught in basic statistics courses may not be appropriate, since those methods implicitly assume that systematic errors are negligible compared to statistical errors. In addition the systematic errors are often ill-conditioned, meaning that they vary in non-trivial ways. This means that it may also not be appropriate to use some of the standard methods that have been developed specifically within the field of stellar spectroscopy, since those methods assume that the systematic errors have relatively trivial behaviour. Instead, it is often necessary to develop statistical analysis tools appropriate to the specific research question.

In Articles I and II we developed statistical tools for two relatively common cases: evaluating the performance of a pipeline for parameter estimation, and estimating model parameters based on spectroscopically derived quantities. We have also sketched out how those tools can be used for seemingly very different problems.

We believe that in the coming decades, the field of stellar spectroscopy will have to devote more resources to developing statistical tools, if it is to keep delivering reliable results. Thanks to technological improvements, the field is getting access to increasingly larger volumes of data, which inherently makes systematic errors a larger part of the total error in any analysis.

In addition to conventional spectroscopy, the field is increasingly adopting the use of machine learning. In Article III we look at the possibility of using machine learning to extract $\alpha$ abundances from Gaia spectra, which have too low resolution for individual spectral lines to be visible. We find that it is possible to do, but there are several caveats that have to be kept in mind when interpreting the results. In general, we recommend testing machine learning methods in situations where it seems like they could plausibly be useful, since this requires a relatively small investment of effort while potentially having large payoffs.

# 7. Contributions to included papers

## 7.1 Article I

Gregory Ruchti and Pieter Gruyters originally branched off the LUMBA UVES stellar parameter pipeline from the LUMBA Giraffe stellar parameter pipeline, which had been written by Karin Lind. Gregory was in the process of preparing it for use for internal Data Release 5 (iDR5) of Gaia-ESO when he suddenly fell ill and had to hand it over to me. I used the pipeline in iDR5 and the subsequent iDR6, after which we decided to publish an article documenting it.

I continued the work of making the pipeline suitable for UVES spectra. I developed the framework for testing the pipeline. The article was written by me with input from the coauthors.

## 7.2 Article II

Andreas Korn and Pieter Gruyters defined and executed the observational project, as part of the article series *Atomic diffusion and mixing in old stars*. A preliminary analysis of the data had already been performed as a Masters' project by Kilian Scheutwinkel under supervision of Andreas Korn, and working closely with Pieter Gruyters.

I created the algorithm for coadding and cleaning the spectra. I branched off a pipeline for analysing our Giraffe spectra, using the LUMBA UVES abundance pipeline as a basis. The LUMBA UVES abundance pipeline had in turn been branched off by Gregory Ruchti from the original LUMBA Giraffe abundance pipeline written by Karin Lind. I defined the statistical framework for interpreting the results. The article was written by me with input from the coauthors, with Thomas Nordlander writing parts of Sect. 4.4.2.

## 7.3 Article III

Rene Andrae and Morgan Fouesneau defined the project and did an initial test. I did the analysis and developed the tools for evaluating the results. The article was written by me with input from the coauthors.

# 8. Svensk sammanfattning

Denna avhandling handlar om galaxarkeologi, och fokuserar på utveckling av statistiska metoder för spektroskopiska studier inom galaxarkeologi.

Galaktisk arkeologi är det forskningsområde som försöker rekonstruera Vintergatans historia. De två huvudsakliga verktygen för detta är *astrometriska* och kemiska studier. Astrometri är studiet av var astronomiska objekt är och hur de rör sig. Detta kan låta enkelt, och är det rent konceptuellt, men i praktiken är det mycket svårt att genomföra för stjärnor som inte råkar ligga mycket nära Solen. Den första storskaliga astrometriska studien gjordes först i slutet av 1900-talet, med HIPPARCOS-satelliten. I skrivande stund pågår en mycket större studie med rymdteleskopet Gaia, som i och med *Early Data Release 3* har mätt position i tre dimensioner och hastighet i två dimensioner för 1,5 miljarder stjärnor. Kemiska studier ser på ämnena i stjärnornas yttre atmosfärer. Detta görs i huvudsak genom studier av stjärnors *spektra*: mätningar av stjärnljusets intensitet som funktion av våglängd. Stjärnspektra innehåller *absorptionslinjer* som reflekterar stjärnans kemiska sammansättning, samt dess andra fysikaliska egenskaper. Eftersom alla stjärnor ursprungligen fötts i hopar av kemiskt mer eller mindre homogen gas så kan spektroskopiska studier identifiera stjärnor som har ett gemensamt ursprung, även efter att stjärnhopen upplösts och medlemmarna spridits slumpmässigt genom Galaxen. Astrometriska studier kan då användas tillsammans med dessa kemiska data för att rekonstruera Galaxens historia längre bakåt i tiden än vad som skulle vara möjligt med något enskilt av dessa verktyg.

Tack vare flera tekniska framsteg inom observationell stjärnspektroskopi så har vi idag tillgång till mycket större mängder data än som tidigare varit möjligt. Med *multifiberspektrografer* så går det att observera hundratals spektra under samma observationstid som tidigare bara hade räckt till en stjärna. Samtidigt så har teknisk utveckling också lett till framsteg inom teoretisk stjärnspektroskopi, eftersom snabbare processorer och större datorminnen gör det möjligt att använda modeller som tidigare hade varit beräkningsmässigt ohanterliga. Förbättringarna inom observation och teori leder dock till nya problem som måste lösas. Större mängder data innebär generellt sett att de slumpmässiga felen i analyser sjunker, medan de systematiska felen förblir desamma. Mer sofistikerade modeller innebär samtidigt att spektroskopister kommer leta efter allt mindre effekter i sina data. Det gör att det blir allt mer viktigt att ta fram statistiska metoder som på ett realistiskt sett tar hänsyn till de systematiska fel som finns i datat.

I var och en av artiklarna i denna avhandling så försöker vi lösa något problem inom galaxarkeologi. I två av artiklarna så utvecklar vi även en generell, och generaliserbar, metod för att lösa problem av liknande typ.

I Artikel I så dokumenterar vi en kod för att utifrån spektra tagna med spektrografen UVES ta fram stjärnparametrarna effektivtemperatur, ytgravitation, metallicitet, mikroturbulens och makroturbulens. Vi använder en uppsättning spektra från stjärnor med kända stjärnparametrar för att uppskatta storleken på de systematiska och statistiska felen i kodens uppskattningar. Samtidigt så utvecklar vi en generell statistisk metod för att med denna sorts testspektra uppskatta felen i analyskoder av denna typ. I avhandlingen så skissar vi också hur denna metod kan generaliseras för att testa verktyg som utifrån någon typ av data uppskattar någon typ av parameter, och det finns testfall för vilka de korrekta parametervärdena är kända.

I Artikel II så uppskattar vi grundämneshalter från spektra av stjärnor i stjärnhopen M30 för att uppskatta parametern $T_0$, som ingår i så kallade *AddMix*-modeller: Trots att stjärnor i stjärnhopar föds nästan kemiskt identiska så finns det flera processer som får sammansättningen i den yttre atmosfären att förändras över tid. Till exempel så får tyngdkraften tyngre partiklar att sjunka, medan strålningstrycket lyfter mindre genomskinliga partiklar. Modeller av stjärnutveckling som tar med alla dessa processer förutspår att halterna av grundämnen i stjärnhopar borde variera med temperaturen – vilket de också gör, men mycket mindre än modellerna förutser. AddMix-modeller antar att det finns någon ytterligare process som vi ännu inte identifierat som på något sätt blandar gasen i atmosfärens olika lager och på så sätt jämnar ut halterna. Eftersom processen inte är känd så innehåller dessa modeller flera fria parametrar, som genom olika studier har begränsats mer och mer. I denna artikel så försöker vi att för M30 begränsa parametern $T_0$, som beskriver hur stark AddMix-processen är. Vi kommer fram till att $\log_{10}(T_0/\,[\mathrm{K}])$ ligger någonstans i intervallet 6.09-6.20. Samtidigt så utvecklar vi en generell metod för att utifrån halter uppskattade från stjärnhopsspektra uppskatta parametrar i stjärnutvecklingsmodeller, som tar hänsyn till att haltuppskattningarna har temperaturberoende systematiska fel. I avhandlingen så skissar vi också hur denna metod kan generaliseras för att uppskatta någon parameter i modeller som beskriver hur någonting mätbart beror på en variabel, där det är känt att mätvärdena har ett systematiskt fel som beror på variabeln.

I Artikel III så testar vi om det är möjligt att med hjälp av maskininlärning utifrån Gaia-satellitens BP/RP-spektra ta fram uppskattningar av halten av alfaprocessgrundämnen: Gaia-satelliten har en viss förmåga att mäta spektra, men spektrumen har såpass låg upplösning att det inte går att urskilja enskilda absorptionslinjer, vilket gör att vanliga spektroskopiska metoder inte går att använda. Samtidigt så kan man vänta sig att formen hos Gaia-spektrumen till viss grad borde påverkas om väldigt många absorptionslinjer ändras samtidigt, även om de var och en inte går att urskilja. Det är dock svårt att modellera hur, eftersom instrumentprofilen för Gaia-spektrumen till viss del inte är känd.

Vi har också tack vare den spektroskopiska studien GALAH tillgång till alfahalten för en delmängd av Gaia-spektrumen. Vi testar därför om det går att träna en modell byggd på ExtraTrees-algoritmen till att uppskatta alfahalter för Gaia-spektra, utifrån de observerade spektrumen och alfahalterna, utan att ha en explicit modell för spektrumens beroende på alfahalterna. Vi kommer fram till att det i princip går, men att det finns väsentliga begränsningar i hur resultaten kan användas. Även om alfahaltsuppskattningarna i genomsnitt ligger nära sanningen så är de i huvudsak baserade på alfahaltens korrelation med andra egenskaper hos stjärnan, som i sin tur har en effekt på spektrumet, och inte alfahalternas direkta effekt på spektrumet. Det innebär att metoden inte går att använda för att särskilja två stjärnor som enbart skiljer sig i att de har olika alfahalter.

# 9. Acknowledgements

First, I want to thank my supervisors, Andreas Korn and Ulrike Heiter. You took me in after I'd realised I didn't actually like theoretical physics and would rather do astronomy instead. You've guided me for the last five years, and you've been admirably patient with my tendency to get way more interested in research tools than actual research questions, and with my firm conviction that the glass is always half empty.

I also want to thank Martin Sahlén, one of very few other people I've met who also thinks statistics is interesting. Without your course *Statistical Inference for Physics & Astronomy*, this thesis would have looked completely different. (Any errors, omissions, or dodgy uses of Akaike's information criterion remain my own).

I also want to say a few words in memory of Gregory Ruchti. We only met briefly, when I took over his work on Gaia-ESO due to his sudden illness, but that ended up being the chance event that shaped all my subsequent work. He did the work of branching off the UVES pipeline from the GIRAFFE pipeline, and would have been a coauthor on Article I if not for the illness that took him away in 2019.

I want to thank my fellow PhD students, who I've worked with for the last few years. In particular: Thanks to Christian for helping maintain my sanity; Thanks to Terese for working beside me during the most chaotic period of teaching; Thanks to Ansgar for being my office-mate and for giving the best and most succinct summary of my research findings; Thanks to Samuel, for also being my office-mate and showing me that there are more Sci-Fi movies than Star Wars.

Thanks to the two RPG and grand strategy circles that have been getting together regularly for the last few years. Thanks to Anders, Laura, Seméli, David and, lately, Andromeda. Thanks to Frazze, Staffan, Andreas and Johannes. We've had several adventures together, and made the Corona year much less bad than it could have been.

Thanks to Kollektivet Katlagrottan, for being my home during most of my time in Uppsala. Thanks, Johan, Simon and Jojjo. You've been good housemates.

I also want to thank my family. Thanks to my parents, Hillevi and Kai-Mikael. You've both been PhD students, and could share your experiences when I needed it. Thanks to my sister, Vanda. You gave an invaluable outside perspective to the slightly unhinged subculture of academia. Thanks to my grandmother, Margareta. Your home has been a nice place to visit for tea,

food and conversation whenever I needed to get away from PhDing. May the pandemic go away so we can meet again.

# 10. Appendices

## 10.1 Estimating $v_{\rm mic}$ through empirical relations

As mentioned in Sect. 3.5.1, during the work on Article II it was necessary to estimate the microturbulence parameter $v_{\rm mic}$ before the abundances could be estimated. In the article we settled on the formulæ described in Sitnova et al. (2015) and Mashonkina et al. (2017), which we jointly dub 'Sitnova-Mashonkina'. The relationship in Sitnova et al. (2015) applies to dwarf stars, while that in Mashonkina et al. (2017) applies to giant stars. Both have the form

$$v_{\rm mic}^{\text{sit-mash}} = \xi_0 + a\,(T_{\rm eff}/10000\,[\rm K]) + b\log g + c\,[\rm Fe/H] \tag{10.1}$$

where $\xi_0$ and the coefficients differ depending on the type of star, as shown in Table 10.1. We arbitrarily chose to put the dividing line between giant and dwarf at $\log g = 3.5\,\rm dex$. We judged that Sitnova-Mashonkina was the best relationship out of those we found, since it was estimated based on stars similar to those in M30, but we want to show all the other relationships that we also considered and rejected. If nothing else, to give the reader a candid look at the number of researcher degrees of freedom in this type of analysis.

**Table 10.1.** *Values of the coefficients in Eq.* (10.1)*, depending on the type of star.*

| Parameter | Dwarf | Giant |
|---|---|---|
| | ($\log g > 3.5\,\rm dex$) | ($\log g \leq 3.5\,\rm dex$) |
| $\xi_0$ | -0.21 | 1.47 |
| a | 5.6 | 4.90 |
| b | - 0.43 | - 0.47 |
| c | 0.06 | - 0.08 |

In Article I, we estimated starting values for $v_{\rm mic}$ using a relation which was at the time used within Gaia-ESO, which we dub 'GES old'. Since $v_{\rm mic}$ was a free parameter in that analysis, and only needed a starting value somewhere in the right ballpark. This relationship starts with comparing the other stellar parameter estimates to two reference values

$$T_{\rm ref} = 5500\,\rm K \tag{10.2}$$

$$\log g_{\rm ref} = 4.2\,\rm dex \tag{10.3}$$

The microturbulence is then estimated as

$$v_{\rm mic}^{\text{ges old}} = a_{i,1} + a_{i,2}\,(T_{\rm eff} - T_{\rm ref}) + a_{i,3}\,(T_{\rm eff} - T_{\rm ref})^2 \tag{10.4}$$

**Table 10.2.** *Parameters in the 'GES old' formula.*

| Parameter | $T_{\text{eff}} > T_{\text{ref}}$ and $\log g > \log g_{\text{ref}}$ | Otherwise |
|---|---|---|
| $a_{i,1}$ | 1.1 | 1.1 |
| $a_{i,2}$ | $1.6 \cdot 10^{-4}$ | $1.0 \cdot 10^{-4}$ |
| $a_{i,3}$ | 0 | $4.0 \cdot 10^{-7}$ |

where the values of the coefficients depend on the stellar parameters, as shown in Table 10.2.

The 'GES old' formula was also used in Gruyters et al. (2016), the sixth article in the article series. However, Gaia-ESO has since replaced this with a more accurate relation, which we dub 'GES new'. This introduces three reference stellar parameters and six additional parameters. The formula is

$$\begin{aligned} v_{\text{mic}}^{\text{ges new}} = \xi_0 &+ a_{i,1}\left(T_{\text{eff}} - T_{\text{ref}}\right) + a_{i,2}\left(T_{\text{eff}} - T_{\text{ref}}\right)^2 + \\ &+ b_{i,1}\left(\log g - \log g_{\text{ref}}\right) + b_{i,2}\left(\log g - \log g_{\text{ref}}\right)^2 + \\ &+ c_{i,1}\left([\text{Fe/H}] - [\text{Fe/H}]_{\text{ref}}\right) + c_{i,2}\left([\text{Fe/H}] - [\text{Fe/H}]_{\text{ref}}\right)^2 \end{aligned} \tag{10.5}$$

where $\xi_0$, the three reference parameters and the six coefficients depend on whether the stellar parameters fall above certain thresholds, as shown in Table 10.3.

**Table 10.3.** *Parameters in the 'GES new' formula.*

| Parameter | $\log g \geq 3.5$ or $T_{\text{eff}} \geq 5200\,\text{K}$ | $T_{\text{eff}} < 5200\,\text{K}$ |
|---|---|---|
| $\xi_0$ | 1.10 | 1.47 |
| $T_{\text{ref}}$ | 5787. | 4798. |
| $a_{i,1}$ | $6.04 \cdot 10^{-4}$ | $4.58 \cdot 10^{-4}$ |
| $a_{i,2}$ | $1.45 \cdot 10^{-7}$ | $2.16 \cdot 10^{-7}$ |
| $\log g_{\text{ref}}$ | 4.14 | 2.38 |
| $b_{i,1}$ | $-3.33 \cdot 10^{-1}$ | $-5.08 \cdot 10^{-1}$ |
| $b_{i,2}$ | $9.77 \cdot 10^{-2}$ | $-7.71 \cdot 10^{-2}$ |
| $[\text{Fe/H}]_{\text{ref}}$ | -0.33 | -0.71 |
| $c_{i,1}$ | $6.94 \cdot 10^{-2}$ | $2.20 \cdot 10^{-1}$ |
| $c_{i,2}$ | $3.12 \cdot 10^{-2}$ | $5.45 \cdot 10^{-2}$ |

However, both Eq. (10.4) and Eq. (10.5) were calibrated on Pop I stars, and probably are not appropriate for M30. A better choice might be the $v_{\text{mic}}$ relation found in Gruyters et al. (2014), the fifth paper in the article series. This was based on NGC6752, which is at least more similar to M30. The relation, which we dub 'Paper V', is a simple linear fit to the $v_{\text{mic}}$ estimated for the stars in that article, and has the formula

$$v_{\text{mic}}^{\text{Paper V}} = 1.3 + 6.25 \cdot 10^{-4}\left(T_{\text{eff}} - 5200\,\text{K}\right) \tag{10.6}$$

Another possible choice we considered was to ourselves make a quadratic fit to the estimated $v_{\text{mic}}$ in Korn et al. (2007), the first paper in the article series.

We dub this choice ‘Paper I’, and it gave the formula

$$v_{\mathrm{mic}}^{\mathrm{Paper\ I}} = 1.74 + 3.15 \cdot 10^{-4} T_{\mathrm{eff}}^{\mathrm{shift}} + 1.22 \cdot 10^{-7} \left(T_{\mathrm{eff}}^{\mathrm{shift}}\right)^2 \tag{10.7}$$

where for numerical reasons we have introduced

$$T_{\mathrm{eff}}^{\mathrm{shift}} = T_{\mathrm{eff}} - 5608\,\mathrm{K}$$

Finally, we also tried the approach of directly estimating $v_{\mathrm{mic}}$. We used the lines shown in Table 10.4. We then made a linear fit to the stars with $\log g < 3.0$, since the lines were barely perceptible in the warmer stars. We dub this choice ‘Direct’ and it gave the formula

$$v_{\mathrm{mic}}^{\mathrm{direct}} = -0.0486 \log g^{\mathrm{shift}} + 1.75\,\mathrm{km/s} \tag{10.8}$$

where for numerical reasons we have introduced

$$\log g^{\mathrm{shift}} = \log g - 2.39\,\mathrm{dex} \tag{10.9}$$

It is not clear how accurate this formula would be for the dwarf stars, given that it extrapolates far outside of the temperature range of the data that it is based on. Pragmatically, this might not be a big problem: In the context of Article II the parameter $v_{\mathrm{mic}}$ is not interesting in itself. It is only a tool for estimating abundances. Since the spectral lines used for the abundance estimate are much less $v_{\mathrm{mic}}$-sensitive at the hot than the cold end, it is not very important to have accurate estimates for dwarfs. Even so, we decided that formula was based on so little data that it was better to use the Sitnova-Mashonkina formula.

Figures 10.1 and 10.2 show the $v_{\mathrm{mic}}$ predicted by Eqs. (10.1)-(10.7) for each group spectrum. Figure 10.1 shows the $v_{\mathrm{mic}}$ as functions of $T_{\mathrm{eff}}$ while Fig. 10.2 shows them as functions of $\log g$. We show both since it might be tempting to try to evaluate the reliability of the formulæ by eye based on how quickly they postulate that $v_{\mathrm{mic}}$ changes, and we want to emphasise that this depends on the choice of parametrisation. The figures also show the $v_{\mathrm{mic}}$ measured using the lines in Table 10.4. The estimates made using the formulæ cover a range of about $0.6\,\mathrm{km/s}$, which is the basis for our assumption in Article II that $v_{\mathrm{mic}}$ can only be estimated to within $\pm 0.3\,\mathrm{km/s}$. The attempted measurements using $v_{\mathrm{mic}}$-sensitive lines stay within the reasonable region for the hotter stars, but become essentially meaningless at low temperatures. In many cases values are missing entirely since the fitting failed to converge.

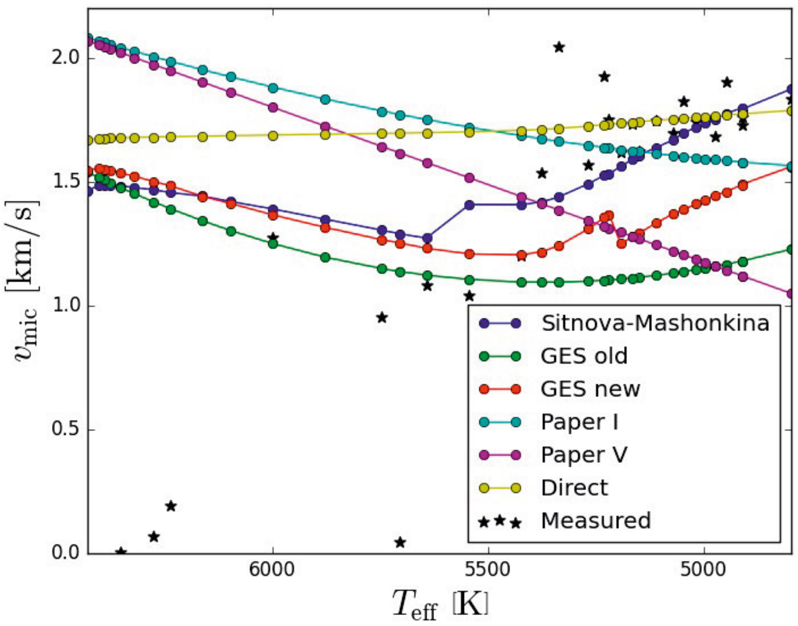

*Figure 10.1.* Estimated $v_{\rm mic}$ using Eqs. (10.1)-(10.7) as a function of $T_{\rm eff}$, together with measurements made using the lines listed in Table 10.4. Since not all of the formulæ gives $v_{\rm mic}$ as a function of $T_{\rm eff}$ alone, we use markers to show the $v_{\rm mic}$ of the individual group spectra.

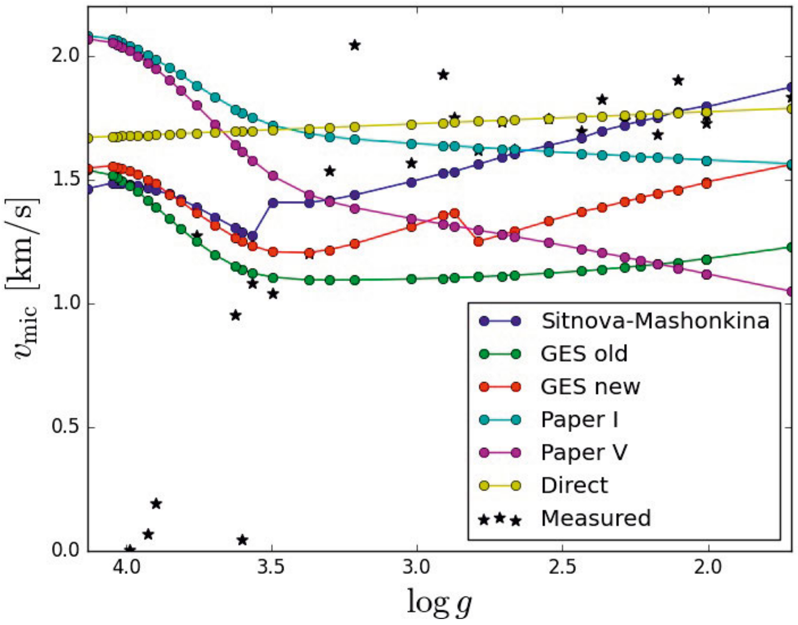

*Figure 10.2.* Same data as in Fig. 10.1 above, but $v_{\rm mic}$ shown as a function of $\log g$.

**Table 10.4.** *Iron lines used in the attempted measurement of $v_{mic}$ in Article II.*

| Line centre [Å] | Species | Line mask [Å] | | Segment mask [Å] | |
|---|---|---|---|---|---|
| | | start | end | start | end |
| 4547.0169 | Fe1 | 4546.80 | 4547.20 | 4530.00 | 4553.99 |
| 4547.8470 | Fe1 | 4547.50 | 4548.20 | – | – |
| 4555.8868 | Fe2 | 4555.00 | 4556.80 | 4554.00 | 4563.00 |
| 4576.3329 | Fe2 | 4576.00 | 4576.90 | 4573.00 | 4577.99 |
| 4580.0531 | Fe2 | 4579.60 | 4580.20 | 4578.00 | 4584.99 |
| 4582.8296 | Fe2 | 4582.50 | 4583.00 | – | – |
| 4583.8292 | Fe1 | 4583.50 | 4584.30 | – | – |
| 4592.6511 | Fe1 | 4592.00 | 4593.00 | 4588.50 | 4599.49 |
| 4595.3584 | Fe1 | 4595.10 | 4595.60 | – | – |
| 4598.1169 | Fe1 | 4597.80 | 4598.40 | – | – |
| 4602.9407 | Fe1 | 4602.40 | 4603.50 | 4599.50 | 4604.99 |
| 4607.6451 | Fe1 | 4607.40 | 4607.80 | 4605.00 | 4608.99 |
| 4611.2788 | Fe1 | 4610.90 | 4611.60 | 4610.00 | 4612.99 |
| 4619.2880 | Fe1 | 4619.00 | 4619.70 | – | – |
| 4620.5128 | Fe2 | 4620.15 | 4620.85 | 4618.00 | 4627.99 |
| 4625.0450 | Fe1 | 4624.65 | 4625.35 | – | – |
| 4626.1515 | Fe1 | 4625.80 | 4626.45 | – | – |
| 4629.3310 | Fe1 | 4628.95 | 4630.55 | 4628.00 | 4631.49 |
| 4632.9114 | Fe1 | 4632.50 | 4633.30 | 4631.50 | 4641.99 |
| 4637.5031 | Fe1 | 4637.20 | 4638.30 | – | – |
| 4643.4634 | Fe1 | 4643.10 | 4643.70 | 4642.00 | 4644.00 |
| 4647.4342 | Fe1 | 4647.00 | 4647.80 | 4645.00 | 4658.99 |
| 4654.4978 | Fe1 | 4654.05 | 4654.90 | – | – |
| 4656.4523 | Fe1 | 4656.20 | 4656.70 | – | – |
| 4661.9703 | Fe1 | 4661.70 | 4662.10 | 4659.00 | 4664.99 |
| 4666.7495 | Fe2 | 4666.50 | 4666.90 | 4665.00 | 4672.49 |
| 4668.1341 | Fe1 | 4667.90 | 4668.40 | – | – |
| 4678.8457 | Fe1 | 4678.50 | 4679.00 | 4675.00 | 4681.99 |
| 4691.4116 | Fe1 | 4691.10 | 4691.70 | 4687.00 | 4699.99 |
| 4710.2833 | Fe1 | 4710.00 | 4710.60 | 4700.00 | 4719.99 |
| 4727.3946 | Fe1 | 4727.00 | 4727.70 | 4720.00 | 4738.99 |
| 4728.5457 | Fe1 | 4728.30 | 4728.90 | – | – |
| 4731.4476 | Fe2 | 4731.00 | 4731.00 | – | – |
| 4733.5914 | Fe1 | 4733.20 | 4734.00 | – | – |
| 4736.7729 | Fe1 | 4736.40 | 4737.10 | – | – |
| 4741.5294 | Fe1 | 4741.20 | 4741.90 | 4739.00 | 4743.99 |
| 4745.7998 | Fe1 | 4745.50 | 4746.20 | 4744.00 | 4752.00 |

## 10.2 Comments on our use of AIC$_c$

In Article II we use Akaike's Information Criterion to compare a set of different models, as described in Sect. 4.2. In the article text we quickly describe the information criterion, but only gesture vaguely at why we chose that method and reject others. Earlier drafts of the article did include a more detailed explanation, but we were forced to conclude that it entered domains of statistical esoterica that would be of very little interest to most readers. We direct the reader to Burnham and Anderson (2003) for a full discussion, but in this appendix we attempt an intuitive explanation of the criterion and why we prefer it to other tools.

Say that we have a sample of measurements, and a set of hypotheses which claim to explain the measurements. We want to make some quantitative statement about of how probable each hypothesis is, in light of the data. Unfortunately, there is no known statistical method that will allow us to do that, and it is not clear that one exists even in principle. What does exist is a collection of statistical methods that each answers a slightly different question. One basic reason why this is so, is that when using statistical analysis to test scientific hypotheses, there is the step where the hypotheses is translated into statistical language – as assertions that the measurements are samples drawn from a statistical distribution. Whatever the merits of the underlying scientific hypotheses, these statistical models are all almost certain to be false: Whatever statistical distribution reality draws samples from, there is no reason why, out of the set of all mathematical functions that exist, it would chose those that can be specified in a few lines. Hence, it is not clear that the Fisher-Pearson method taught in most basic statistics courses – of attempting to 'falsify' the hypotheses by testing if their likelihood falls below a set threshold – is particularly meaningful. After all, they are known *a priori* to be false. Simply by drawing enough samples, we can always get the likelihood of our observations to fall below any significance threshold. Similarly, the Bayesian approach taught in more advanced statistics courses – of using the likelihood to turn prior probabilities on each hypothesis into posterior probabilities – may not be meaningful either. After all, for the reasons already stated, the prior probability on each hypothesis should be zero.

The information criterion takes as given that each statistical hypothesis is false, and simply tries to estimate how close it is to the true distribution, which is unknown and probably unknowable. 'Closeness' in this case being defined as the degree of information lost when using the model distribution to model the true distribution. Note that this is not a replacement for either Frequentism or Bayesianism. The information criterion can be derived from either Frequentist or Bayesian grounds, and in Article II our discussion implicitly uses a Bayesian framework. It is however, different from the most common types of Frequentist and Bayesian hypothesis testing.

All this said, it is possible to criticise this approach. While the information criterion is the best in terms of out-of-sample error – that is, it would be optimal for guessing what the result would be if we observed more stars in M30 – it is not clear that it is the best for testing truth of scientific hypotheses. In fact, Burnham and Anderson (2003) almost certainly would disagree with our using the information criterion for hypothesis testing. Even so, they concede that making inferences based on a single model selected by the information criterion as close to the truth is at least "not terrible" (Burnham and Anderson, 2003, Sect 8.9). This is not exactly a ringing endorsement, but gives some support to our view that our method is the so-far least bad approach of handling the type of research problem in Article II.

# References

Adams, F. C., Bodenheimer, P., and Laughlin, G. (2005). M dwarfs: planet formation and long term evolution. *Astronomische Nachrichten*, 326(10):913–919.

Alberto, M., Magda, A., Nausicaa, D., Vincenzo, F., Nathalie, F., Olivier, H., Uwe, L., Mubashir, K. A., Laura, M., Joerg, R., Martino, R., Devendra, S., Chiara, S., Malgorzata, S., Felix, S., Ignacio, V., and Stefano, Z. (2019). The New Science Portal and the Programmatic Interfaces of the ESO Science Archive. In Teuben, P. J., Pound, M. W., Thomas, B. A., and Warner, E. M., editors, *Astronomical Data Analysis Software and Systems XXVII*, volume 523 of *Astronomical Society of the Pacific Conference Series*, page 433.

Asplund, M., Amarsi, A. M., and Grevesse, N. (2021). The chemical make-up of the Sun: A 2020 vision. *arXiv e-prints*, page arXiv:2105.01661.

Asplund, M., Nordlund, Å., Trampedach, R., Allende Prieto, C., and Stein, R. F. (2000). Line formation in solar granulation. I. Fe line shapes, shifts and asymmetries. *Astronomy & Astrophysics*, 359:729–742.

Astropy Collaboration, Price-Whelan, A. M., Sipőcz, B. M., Günther, H. M., Lim, P. L., Crawford, S. M., Conseil, S., Shupe, D. L., Craig, M. W., Dencheva, N., Ginsburg, A., Vand erPlas, J. T., Bradley, L. D., Pérez-Suárez, D., de Val-Borro, M., Aldcroft, T. L., Cruz, K. L., Robitaille, T. P., Tollerud, E. J., Ardelean, C., Babej, T., Bach, Y. P., Bachetti, M., Bakanov, A. V., Bamford, S. P., Barentsen, G., Barmby, P., Baumbach, A., Berry, K. L., Biscani, F., Boquien, M., Bostroem, K. A., Bouma, L. G., Brammer, G. B., Bray, E. M., Breytenbach, H., Buddelmeijer, H., Burke, D. J., Calderone, G., Cano Rodríguez, J. L., Cara, M., Cardoso, J. V. M., Cheedella, S., Copin, Y., Corrales, L., Crichton, D., D'Avella, D., Deil, C., Depagne, É., Dietrich, J. P., Donath, A., Droettboom, M., Earl, N., Erben, T., Fabbro, S., Ferreira, L. A., Finethy, T., Fox, R. T., Garrison, L. H., Gibbons, S. L. J., Goldstein, D. A., Gommers, R., Greco, J. P., Greenfield, P., Groener, A. M., Grollier, F., Hagen, A., Hirst, P., Homeier, D., Horton, A. J., Hosseinzadeh, G., Hu, L., Hunkeler, J. S., Ivezić, Ž., Jain, A., Jenness, T., Kanarek, G., Kendrew, S., Kern, N. S., Kerzendorf, W. E., Khvalko, A., King, J., Kirkby, D., Kulkarni, A. M., Kumar, A., Lee, A., Lenz, D., Littlefair, S. P., Ma, Z., Macleod, D. M., Mastropietro, M., McCully, C., Montagnac, S., Morris, B. M., Mueller, M., Mumford, S. J., Muna, D., Murphy, N. A., Nelson, S., Nguyen, G. H., Ninan, J. P., Nöthe, M., Ogaz, S., Oh, S., Parejko, J. K., Parley, N., Pascual, S., Patil, R., Patil, A. A., Plunkett, A. L., Prochaska, J. X., Rastogi, T., Reddy Janga, V., Sabater, J., Sakurikar, P., Seifert, M., Sherbert, L. E., Sherwood-Taylor, H., Shih, A. Y., Sick, J., Silbiger, M. T., Singanamalla, S., Singer, L. P., Sladen, P. H., Sooley, K. A., Sornarajah, S., Streicher, O., Teuben, P., Thomas, S. W., Tremblay, G. R., Turner, J. E. H., Terrón, V., van Kerkwijk, M. H., de la Vega, A., Watkins, L. L., Weaver, B. A., Whitmore, J. B., Woillez, J., Zabalza, V., and Astropy Contributors (2018). The Astropy Project: Building an Open-science Project and Status of the v2.0 Core Package. *The Astrophysical Journal*, 156(3):123.

Astropy Collaboration, Robitaille, T. P., Tollerud, E. J., Greenfield, P., Droettboom, M., Bray, E., Aldcroft, T., Davis, M., Ginsburg, A., Price-Whelan, A. M., Kerzendorf, W. E., Conley, A., Crighton, N., Barbary, K., Muna, D., Ferguson, H., Grollier, F., Parikh, M. M., Nair, P. H., Unther, H. M., Deil, C., Woillez, J., Conseil, S., Kramer, R., Turner, J. E. H., Singer, L., Fox, R., Weaver, B. A., Zabalza, V., Edwards, Z. I., Azalee Bostroem, K., Burke, D. J., Casey, A. R., Crawford, S. M., Dencheva, N., Ely, J., Jenness, T., Labrie, K., Lim, P. L., Pierfederici, F., Pontzen, A., Ptak, A., Refsdal, B., Servillat, M., and Streicher, O. (2013). Astropy: A community Python package for astronomy. *Astronomy & Astrophysics*, 558:A33.

Barkat, Z., Reiss, Y., and Rakavy, G. (1974). Stars in the Mass Range $7 \lesssim M/M_{\odot} \lesssim 10$ AS Candidates for Pulsar Progenitors. *Astrophysical Journal*, 193:L21.

Begelman, M. C., Volonteri, M., and Rees, M. J. (2006). Formation of supermassive black holes by direct collapse in pre-galactic haloes. *Monthly Notices of the Royal Astronomical Society*, 370(1):289–298.

Bertulani, C. A. (2013). *Nuclei in the Cosmos*. World Scientific.

Blanco-Cuaresma, S., Soubiran, C., Jofré, P., and Heiter, U. (2014). Gaia FGK benchmark stars: High resolution spectral library. *Astronomy & Astrophysics*, 566.

Bromm, V., Yoshida, N., Hernquist, L., and McKee, C. F. (2009). The formation of the first stars and galaxies. *Nature*, 459(7243):49–54.

Buder, S., Sharma, S., Kos, J., Amarsi, A. M., Nordlander, T., Lind, K., Martell, S. L., Asplund, M., Bland-Hawthorn, J., Casey, A. R., de Silva, G. M., D'Orazi, V., Freeman, K. C., Hayden, M. R., Lewis, G. F., Lin, J., Schlesinger, K. J., Simpson, J. D., Stello, D., Zucker, D. B., Zwitter, T., Beeson, K. L., Buck, T., Casagrande, L., Clark, J. T., Čotar, K., da Costa, G. S., de Grijs, R., Feuillet, D., Horner, J., Kafle, P. R., Khanna, S., Kobayashi, C., Liu, F., Montet, B. T., Nandakumar, G., Nataf, D. M., Ness, M. K., Spina, L., Tepper-García, T., Ting, Y.-S., Traven, G., Vogrinčič, R., Wittenmyer, R. A., Wyse, R. F. G., Žerjal, M., and Galah Collaboration (2021). The GALAH+ survey: Third data release. *Monthly Notices of the Royal Astronomical Society*.

Burnham, K. P. and Anderson, D. R. (2003). *Model Selection and Multimodel Inference*. Springer-Verlag New York, 2 edition.

Busso, M., Gallino, R., and Wasserburg, G. J. (1999). Nucleosynthesis in Asymptotic Giant Branch Stars: Relevance for Galactic Enrichment and Solar System Formation. *Annual Review of Astronomy and Astrophysics*, 37:239–309.

Carroll, B. W. and Ostlie, D. A. (2014). *An Introduction to Modern Astrophysics*. Pearson Limited, 2nd edition.

Coc, A. and Vangioni, E. (2017). Primordial nucleosynthesis. *International Journal of Modern Physics E*, 26(8):1741002.

Cohen, J. G. (1978). Abundances in globular cluster red giants. I. M3 and M13. *The Astrophysical Journal*, 223:487–508.

Cyburt, R. H. and Pospelov, M. (2012). Resonant Enhancement of Nuclear Reactions as a Possible Solution to the Cosmological Lithium Problem. *International Journal of Modern Physics E*, 21(1):1250004–1–1250004–13.

Dekker, H., D'Odorico, S., Kaufer, A., Delabre, B., and Kotzlowski, H. (2000). Design, construction, and performance of UVES, the echelle spectrograph for the

UT2 Kueyen Telescope at the ESO Paranal Observatory. In *Proc. SPIE 4008, Optical and IR Telescope Instrumentation and Detectors*, volume 4008.

Fanning, D. W. (2015). Coyote's guide to IDL programming. `http://www.idlcoyote.com/`.

Fields, B. D. (2011). The primordial lithium problem. *Annual Review of Nuclear and Particle Science*, 61(1):47–68.

Foreman-Mackey, D., Hogg, D. W., Lang, D., and Goodman, J. (2013). emcee: The MCMC Hammer. *arXiv*, 125:306.

Forveille, Thierry, Kotak, Rubina, Shore, Steve, and Tolstoy, Eline (2018). Gaia data release 2. *Astronomy & Astrophysics*, 616:E1.

Freeman, K. C., Bland-Hawthorn, J., Survey Management Group, and GALAH team (2013). The HERMES GALAH survey: overview.

Freiburghaus, C., Rosswog, S., and Thielemann, F. K. (1999). *r*-process in neutron star mergers. *The Astrophysical Journal*, 525(2):L121–L124.

Gaia Collaboration. Gaia data release 2 (Gaia DR2). `https://www.cosmos.esa.int/web/gaia/dr2`. Page last changed: 2018-04-27.

Gaia Collaboration. Gaia early data release 3 (Gaia EDR3). `https://www.cosmos.esa.int/web/gaia/earlydr3`.

Gaia Collaboration (2016a). Gaia data release 1 - summary of the astrometric, photometric, and survey properties. *Astronomy & Astrophysics*, 595:A2.

Gaia Collaboration (2016b). The Gaia mission. *Astronomy & Astrophysics*, 595:A1.

Gaia Collaboration (2018). Gaia data release 2 - Summary of the contents and survey properties. *Astronomy & Astrophysics*, 616:A1.

Gaia Collaboration (2021). Gaia early data release 3 - Summary of the contents and survey properties. *Astronomy & Astrophysics*, 649:A1.

Gavel, A., Korn, A. J., Andrae, R., and Fouesneau, M. (2020). The chemical trace of Galactic stellar populations as seen by Gaia. `https://www.cosmos.esa.int/web/gaia/iow_20200320`.

Genel, S., Vogelsberger, M., Springel, V., Sijacki, D., Nelson, D., Snyder, G., Rodriguez-Gomez, V., Torrey, P., and Hernquist, L. (2014). Introducing the Illustris Project: the evolution of galaxy populations across cosmic time. *Monthly Notices of the Royal Astronomical Society*, 445(1):175–200.

Geurts, P., Ernst, D., and Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning*, 63:3–42.

Gilmore, G., Randich, S., Asplund, M., Binney, J., Bonifacio, P., Drew, J., Feltzing, S., Ferguson, A., Jeffries, R., Micela, G., Negueruela, I., Prusti, T., Rix, H.-W., Vallenari, A., Alfaro, E., Allende-Prieto, C., Babusiaux, C., Bensby, T., Blomme, R., Bragaglia, A., Flaccomio, E., François, P., Irwin, M., Koposov, S., Korn, A., Lanzafame, A., Pancino, E., Paunzen, E., Recio-Blanco, A., Sacco, G., Smiljanic, R., Eck, S. V., and Walton, N. (2012). The Gaia-ESO Public Spectroscopic Survey. *The Messenger*, 147:25–31.

Goldstein, H., Poole, C., and Safko, J. (2002). *Classical Mechanics*. Pearson Education International, 3rd edition.

Gratton, R. G., Bonifacio, P., Bragaglia, A., Carretta, E., Castellani, V., Centurion, M., Chieffi, A., Claudi, R., Clementini, G., D'Antona, F., Desidera, S., François, P., Grundahl, F., Lucatello, S., Molaro, P., Pasquini, L., Sneden, C., Spite, F., and

Straniero, O. (2001). The O-Na and Mg-Al anticorrelations in turn-off and early subgiants in globular clusters. *Astronomy & Astrophysics*, 369:87–98.

Gray, D. F. (2008). *The Observation and Analysis of Stellar Photospheres*. Cambridge University Press, 3rd edition.

Gruyters, P., Lind, K., Richard, O., Grundahl, F., Asplund, M., Casagrande, L., Charbonnel, C., Milone, A., Primas, F., and Korn, A. J. (2016). Atomic diffusion and mixing in old stars - VI. The lithium content of M30. *Astronomy & Astrophysics*, 589:A61.

Gruyters, P., Nordlander, T., and Korn, A. J. (2014). Atomic diffusion and mixing in old stars - V. A deeper look into the globular cluster NGC2. *Astronomy & Astrophysics*, 567:A72.

Gustafsson, B., Edvardsson, B., Eriksson, K., Jørgensen, U. G., Nordlund, Å., and Plez, B. (2008). A grid of MARCS model atmospheres for late-type stars I. Methods and general properties. *Astronomy & Astrophysics*, 486:951–970.

Haeussler, B. (2020). *Very Large Telescope Paranal Science Operations FLAMES User Manual*. European Southern Observatory, Karl-Schwarzschild Str. 2, D-85748 Garching bei München, 106.2 edition.

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., and Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, 585(7825):357–362.

Harris Geospatial Solutions, I. (2021). End user license agreement for IDL® 8.8 and ENVI® 5.6 & ENVI® photogrammetry module 5.6. `https://www.l3harrisgeospatial.com/Company/Legal/EULA`.

Hayek, W., Asplund, M., Collet, R., and Nordlund, Å. (2011). 3D LTE spectral line formation with scattering in red giant stars. *Astronomy & Astrophysics*, 529:A158.

Heiter, U. (2017). Galactic chemical evolution. Compendium for the course *The Physics of Stars*.

Heiter, U., Lind, K., Bergemann, M., Asplund, M., Mikolaitis, Š., Barklem, P. S., Masseron, T., de Laverny, P., Magrini, L., Edvardsson, B., Jönsson, H., Pickering, J. C., Ryde, N., Bayo Arán, A., Bensby, T., Casey, A. R., Feltzing, S., Jofré, P., Korn, A. J., Pancino, E., Damiani, F., Lanzafame, A., Lardo, C., Monaco, L., Morbidelli, L., Smiljanic, R., Worley, C., Zaggia, S., Randich, S., and Gilmore, G. F. (2021). Atomic data for the Gaia-ESO Survey. *Astronomy & Astrophysics*, 645:A106.

Helmi, A., Babusiaux, C., Koppelman, H. H., Massari, D., Veljanoski, J., and Brown, A. G. A. (2018). The merger that led to the formation of the Milky Way's inner stellar halo and thick disk. *Nature*, 563:85–88.

Høg, E., Bässgen, G., Bastian, U., Egret, D., Fabricius, C., Großmann, V., Halbwachs, J. L., Makarov, V. V., Perryman, M. A. C., Schwekendiek, P., Wagner, K., and Wicenec, A. (1997). The TYCHO Catalogue. *Astronomy & Astrophysics*, 323:L57–L60.

Høg, E., Fabricius, C., Makarov, V. V., Urban, S., Corbin, T., Wycoff, G., Bastian, U., Schwekendiek, P., and Wicenec, A. (2000). The Tycho-2 catalogue of the 2.5 million brightest stars. *Astronomy & Astrophysics*, 355:L27–L30.

Holgate, S. A. (2009). *Understanding Solid State Physics*. CRC Press.

Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95.

Jedamzik, K. and Pospelov, M. (2009). Big Bang nucleosynthesis and particle dark matter. *New Journal of Physics*, 11(10):105028.

Kiselman, D. (1999). Effects of NLTE and Granulation on LiBeB Abundance Determinations. In Ramaty, R., Vangioni-Flam, E., Cassé, M., and Olive, K., editors, *LiBeB Cosmic Rays, and Related X- and Gamma-Rays*, volume 171 of *Astronomical Society of the Pacific Conference Series*, page 85.

Kordopatis, G., Gilmore, G., Steinmetz, M., Boeche, C., Seabroke, G., Siebert, A., Zwitter, T., Binney, J., de Laverny, P., Recio-Blanco, A., Williams, M., Piffl, T., Enke, H., Roeser, S., Bijaoui, A., Wyse, R., Freeman, K., Munari, U., Carillo, I., Anguiano, B., Burton, D., Campbell, R., Cass, C., Fiegert, K., M. Hartley, Q. P., Reid, W., Ritter, A., Russell, K., Stupart, M., Watson, F., Bienaymé, O., Bland-Hawthorn, J., Gerhard, O., Gibson, B., Grebel, E., Helmi, A., Navarro, J., Conrad, C., Famaey, B., Faure, C., Just, A., Kos, J., Matijevic, G., McMillan, P., Minchev, I., Scholz, R., Sharma, S., Siviero, A., de Boer, E. W., and Žerjal, M. (2013). The Radial Velocity Experiment (RAVE): Fourth Data Release. *The Astronomical Journal*, 146.

Korn, A. J., Grundahl, F., Richard, O., Mashonkina, L., Barklem, P. S., Collet, R., Gustafsson, B., and Piskunov, N. (2007). Atomic diffusion and mixing in old stars. I. Very Large Telescope FLAMES-UVES observations of stars in NGC 6397. *The Astrophysical Journal*, 671(1):402–419.

Laughlin, G., Bodenheimer, P., and Adams, F. C. (1997). The end of the main sequence. *The Astrophysical Journal*, 482(1):420–432.

Lemson, G. and Virgo Consortium, t. (2006). Halo and Galaxy Formation Histories from the Millennium Simulation: Public release of a VO-oriented and SQL-queryable database for studying the evolution of galaxies in the ΛCDM cosmogony. *arXiv e-prints*, pages astro–ph/0608019.

Levenberg, K. (1944). A method for the solution of certain non-linear problems in least squares. *Quarterly of Applied Mathematics*, 2(2):164–168.

Liu, C., Bailer-Jones, C. A. L., Sordo, R., Vallenari, A., Borrachero, R., Luri, X., and Sartoretti, P. (2012). The expected performance of stellar parametrization with Gaia spectrophotometry. *Monthly Notices of the Royal Astronomical Society*, 426(3):2463–2482.

MacKay, D. J. C. (2003). *Information Theory, Inference and Learning Algorithms*. Cambridge University Press.

Magic, Z., Collet, R., Asplund, M., Trampedach, R., Hayek, W., Chiavassa, A., Stein, R. F., and Nordlund, A. (2013). The Stagger-grid: A grid of 3D stellar atmosphere models - I. Methods and general properties. *Astronomy & Astrophysics*, 557:A26.

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

Marquardt, D. W. (1963). An algorithm for least-squares estimation of nonlinear parameters. *Journal of the Society for Industrial and Applied Mathematics*, 11(2):431–441.

Mashonkina, L., Jablonka, P., Pakhomov, Y., Sitnova, T., and North, P. (2017). The formation of the Milky Way halo and its dwarf satellites; a NLTE-1D abundance analysis - I. homogeneous set of atmospheric parameters. *Astronomy & Astrophysics*, 604:A129.

Michaud, G. (1970). Diffusion Processes in Peculiar A Stars. *The Astrophysical Journal*, 160:641.

Ness, M., Rix, H.-W., Hogg, D. W., Casey, A. R., Holtzman, J., Fouesneau, M., Zasowski, G., Geisler, D., Shetrone, M., Minniti, D., Frinchaboy, P. M., and Roman-Lopes, A. (2018). Galactic doppelgängers: The chemical similarity among field stars and among stars with a common birth origin. *The Astrophysical Journal*, 853(2).

Nomoto, K. and Hashimoto, M.-A. (1986). Late stages of massive star evolution and nucleosynthesis. *Progress in Particle and Nuclear Physics*, 17:267–285.

Nordlander, T., Amarsi, A. M., Lind, K., Asplund, M., Barklem, P. S., Casey, A. R., Collet, R., and Leenaarts, J. (2017). 3D NLTE analysis of the most iron-deficient star, SMSS0313-6708. *Astronomy & Astrophysics*, 597:A6.

Pasquini, L., Avila, G., Blecha, A., Cacciari, C., Cayatte, V., Colless, M., Damiani, F., de Propris, R., Dekker, H., di Marcantonio, P., Farrell, T., Gillingham, P., Guinouard, I., Hammer, F., Kaufer, A., Hill, V., Marteaud, M., Modigliani, A., Mulas, G., North, P., Popovic, D., Rossetti, E., Royer, F., Santin, P., Schmutzer, R., Simond, G., Vola, P., Waller, L., and Zoccali, M. (2002). Installation and commissioning of FLAMES, the VLT Multifibre Facility. *The Messenger*, 110:1–9.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Pence, W. D., Chiappetti, L., Page, C. G., Shaw, R. A., and Stobie, E. (2010). Definition of the Flexible Image Transport System (FITS), version 3.0. *Astronomy & Astrophysics*, 524:A42.

Perez, F. and Granger, B. E. (2007). IPython: A system for interactive scientific computing. *Computing in Science Engineering*, 9(3):21–29.

Perryman, M. A. C., Lindegren, L., Kovalevsky, J., Hog, E., Bastian, U., Bernacca, P. L., Creze, M., Donati, F., Grenon, M., Grewing, M., van Leeuwen, F., van der

Marel, H., Mignard, F., Murray, C. A., Le Poole, R. S., Schrijver, H., Turon, C., Arenou, F., Froeschle, M., and Petersen, C. S. (1997). The Hipparcos Catalogue. *Astronomy & Astrophysics*, 500:501–504.

Piau, L., Beers, T. C., Balsara, D. S., Sivarani, T., Truran, J. W., and Ferguson, J. W. (2006). From first stars to the Spite plateau: A possible reconciliation of halo stars observations with predictions from Big Bang nucleosynthesis. *The Astrophysical Journal*, 653(1):300–315.

Piskunov, N. E. and Valenti, J. A. (1996). Spectroscopy made easy: A new tool for fitting observations with synthetic spectra. *Astronomy & Astrophysics Supplement*, 118:595–603.

Piskunov, N. E. and Valenti, J. A. (2017). Spectroscopy Made Easy: Evolution. *Astronomy & Astrophysics*, 597.

Proffitt, C. R. and Michaud, G. (1991). Diffusion and Mixing of Lithium and Helium in Population II Dwarfs. *The Astrophysical Journal*, 371:584.

Randich, S., Gilmore, G., and Gaia-ESO Consortium (2013). The Gaia-ESO Large Public Spectroscopic Survey. *The Messenger*, 154:47–49.

Richard, O., Michaud, G., and Richer, J. (2001). Iron convection zones in B, A, and F stars. *The Astrophysical Journal*, 558(1):377–391.

Richard, O., Michaud, G., and Richer, J. (2005). Implications of *wmap* observations on Li abundance and stellar evolution models. *The Astrophysical Journal*, 619(1):538–548.

Richard, O., Michaud, G., Richer, J., Turcotte, S., Turck-Chieze, S., and VandenBerg, D. A. (2002). Models of metal-poor stars with gravitational settling and radiative accelerations. I. Evolution and abundance anomalies. *The Astrophysical Journal*, 568(2):979–997.

Richer, J., Michaud, G., and Turcotte, S. (2000). The evolution of AmFm stars, abundance anomalies, and turbulent transport. *The Astrophysical Journal*, 529(1):338–356.

Scheutwinkel, K. H. (2019). Spectroscopy of the globular cluster M30. Master's thesis, Uppsala University.

Schneider, P. (2015). *Extragalactic Astronomy and Cosmology*. Springer-Verlag, 2nd edition.

scikit-learn developers (2020). User guide. `https://scikit-learn.org/stable/user_guide.html`.

Shalev-Shwartz, S. and Ben-David, S. (2014). *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press.

Shurtleff, R. and Derringh, E. (1989). The most tightly bound nucleus. *American Journal of Physics*, 57(6):552–552.

Sitnova, T., Zhao, G., Mashonkina, L., Chen, Y., Liu, F., Pakhomov, Y., Tan, K., Bolte, M., Alexeeva, S., Grupp, F., Shi, J.-R., and Zhang, H.-W. (2015). Systematic non-LTE study of the $-2.6 \leq [\mathrm{Fe/H}] \leq 0.2$ F and G dwarfs in the Solar neighborhood. I. stellar atmosphere parameters. *The Astrophysical Journal*, 808(2):148.

Spite, F. and Spite, M. (1982). Abundance of lithium in unevolved stars and old disk stars : Interpretation and consequences. *Astronomy & Astrophysics*, 115:357–366.

Steinmetz, M., Guiglion, G., McMillan, P. J., Matijevič, G., Enke, H., Kordopatis, G., Zwitter, T., Valentini, M., Chiappini, C., Casagrande, L., Wojno, J., Anguiano, B., Bienaymé, O., Bijaoui, A., Binney, J., Burton, D., Cass, P., de Laverny, P., Fiegert, K., Freeman, K., Fulbright, J. P., Gibson, B. K., Gilmore, G., Grebel, E. K., Helmi, A., Kunder, A., Munari, U., Navarro, J. F., Parker, Q., Ruchti, G. R., Recio-Blanco, A., Reid, W., Seabroke, G. M., Siviero, A., Siebert, A., Stupar, M., Watson, F., Williams, M. E. K., Wyse, R. F. G., Anders, F., Antoja, T., Birko, D., Bland-Hawthorn, J., Bossini, D., García, R. A., Carrillo, I., Chaplin, W. J., Elsworth, Y., Famaey, B., Gerhard, O., Jofre, P., Just, A., Mathur, S., Miglio, A., Minchev, I., Monari, G., Mosser, B., Ritter, A., Rodrigues, T. S., Scholz, R.-D., Sharma, S., and and, K. S. (2020a). The sixth data release of the Radial Velocity Experiment (Rave). II. Stellar atmospheric parameters, chemical abundances, and distances. *The Astronomical Journal*, 160(2):83.

Steinmetz, M., Matijevič, G., Enke, H., Zwitter, T., Guiglion, G., McMillan, P. J., Kordopatis, G., Valentini, M., Chiappini, C., Casagrande, L., Wojno, J., Anguiano, B., Bienaymé, O., Bijaoui, A., Binney, J., Burton, D., Cass, P., de Laverny, P., Fiegert, K., Freeman, K., Fulbright, J. P., Gibson, B. K., Gilmore, G., Grebel, E. K., Helmi, A., Kunder, A., Munari, U., Navarro, J. F., Parker, Q., Ruchti, G. R., Recio-Blanco, A., Reid, W., Seabroke, G. M., Siviero, A., Siebert, A., Stupar, M., Watson, F., Williams, M. E. K., Wyse, R. F. G., Anders, F., Antoja, T., Birko, D., Bland-Hawthorn, J., Bossini, D., García, R. A., Carrillo, I., Chaplin, W. J., Elsworth, Y., Famaey, B., Gerhard, O., Jofre, P., Just, A., Mathur, S., Miglio, A., Minchev, I., Monari, G., Mosser, B., Ritter, A., Rodrigues, T. S., Scholz, R.-D., Sharma, S., and and, K. S. (2020b). The sixth data release of the Radial Velocity Experiment (RAVE). I. Survey description, spectra, and radial velocities. *The Astronomical Journal*, 160(2):82.

Steinmetz, M., Zwitter, T., Siebert, A., Watson, F. G., Freeman, K. C., Munari, U., Campbell, R., Williams, M., Seabroke, G. M., Wyse, R. F. G., Parker, Q. A., Bienaymé, O., Roeser, S., Gibson, B. K., Gilmore, G., Grebel, E. K., Helmi, A., Navarro, J. F., Burton, D., Cass, C. J. P., Dawe, J. A., Fiegert, K., Hartley, M., Russell, K. S., Saunders, W., Enke, H., Bailin, J., Binney, J., Bland-Hawthorn, J., Boeche, C., Dehnen, W., Eisenstein, D. J., Evans, N. W., Fiorucci, M., Fulbright, J. P., Gerhard, O., Jauregi, U., Kelz, A., Mijović, L., Minchev, I., Parmentier, G., Peñarrubia, J., Quillen, A. C., Read, M. A., Ruchti, G., Scholz, R.-D., Siviero, A., Smith, M. C., Sordo, R., Veltz, L., Vidrih, S., von Berlepsch, R., Boyle, B. J., and Schilbach, E. (2006). The Radial Velocity Experiment (RAVE): First data release. *The Astronomical Journal*, 132(4):1645–1668.

Taylor, J. R. (1997). *An Introduction to Error Analysis*. University Science Books, 2nd edition.

Taylor, M. B. (2005). TOPCAT & STIL: Starlink Table/VOTable Processing Software. In Shopbell, P., Britton, M., and Ebert, R., editors, *Astronomical Data Analysis Software and Systems XIV*, volume 347 of *Astronomical Society of the Pacific Conference Series*, page 29.

Trotta, R. (2017). Bayesian Methods in Cosmology. *arXiv e-prints*, page arXiv:1701.01467.

Valls-Gabaud, D. (2014). Bayesian isochrone fitting and stellar ages. *EAS Publications Series*, 65:225–265.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., and SciPy 1.0 Contributors (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272.

Vogelsberger, M., Genel, S., Springel, V., Torrey, P., Sijacki, D., Xu, D., Snyder, G., Nelson, D., and Hernquist, L. (2014). Introducing the Illustris project: simulating the coevolution of dark and visible matter in the universe. *Monthly Notices of the Royal Astronomical Society*, 444:1518–1547.

von Toussaint, U. (2011). Bayesian inference in physics. *Reviews of Modern Physics*, 83:943–999.

Wang, E. X., Nordlander, T., Asplund, M., Amarsi, A. M., Lind, K., and Zhou, Y. (2021). 3D NLTE spectral line formation of lithium in late-type stars. *Monthly Notices of the Royal Astronomical Society*, 500(2):2159–2176.

Wehrhahn, A. (2020). Continuum and radial velocity determination. `https://github.com/AWehrhahn/SME/tree/master/src/pysme`.

Acta Universitatis Upsaliensis

*Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology* 2052

Editor: The Dean of the Faculty of Science and Technology

ACTA
UNIVERSITATIS
UPSALIENSIS
UPPSALA
2021

Distribution: publications.uu.se
urn:nbn:se:uu:diva-446479