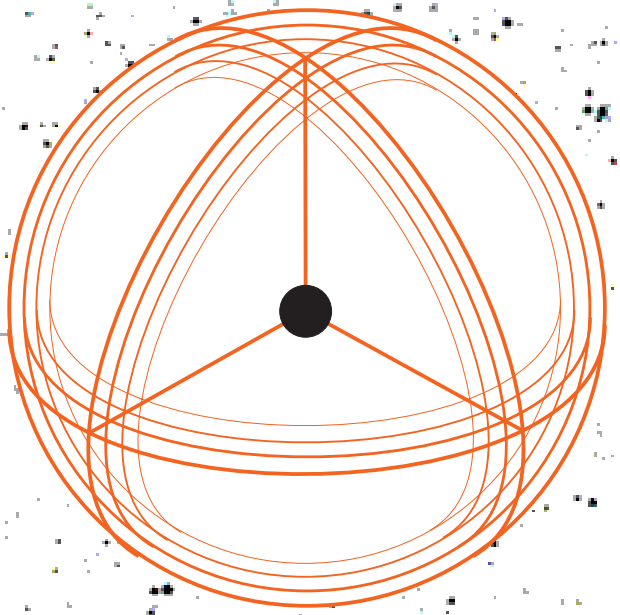
INVERSE PROBLEMS IN ASTEROSEISMOLOGY



EARL PATRICK BELLINGER

International Max Planck Research School for
Solar System Science at the University of Göttingen

Inverse Problems in Asteroseismology

Dissertation
for the award of the degree

"Doctor of Philosophy" (Ph.D.)
Division of Mathematics and Natural Sciences
of the Georg-August-Universität Göttingen

within the Ph.D. Programme in Computer Science (PCS)
of the Georg-August-University School of Science (GAUSS)

submitted by

Earl Patrick Bellinger

from Albany, New York, USA

Göttingen, 2018

Thesis Committee

Dr. ir. Saskia Hekker
Max-Planck-Institut für Sonnensystemforschung, Göttingen, Germany
Stellar Astrophysics Centre, Aarhus University, Denmark

Prof. Dr. Sarbani Basu
Department of Astronomy, Yale University, New Haven, CT, USA

Prof. Dr. Laurent Gizon
Max-Planck-Institut für Sonnensystemforschung, Göttingen, Germany
Institut für Astrophysik, Georg-August-Universität Göttingen, Germany

Prof. Dr. Ramin Yahyapour
Institut für Informatik, Georg-August-Universität Göttingen, Germany
Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen, Germany

Members of the Examination Board

Reviewer: **Prof. Dr. Ramin Yahyapour**
Institut für Informatik, Georg-August-Universität Göttingen, Germany
Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen, Germany

Second Reviewer: **Prof. Dr. Laurent Gizon**
Max-Planck-Institut für Sonnensystemforschung, Göttingen, Germany
Institut für Astrophysik, Georg-August-Universität Göttingen, Germany

Third Reviewer: **Prof. Dr. Yvonne Elsworth, FRS**
School of Physics and Astronomy, University of Birmingham, United Kingdom

Further members of the Examination Board:

Prof. Dr. Carsten Damm
Institut für Informatik, Georg-August-Universität Göttingen, Germany

Jun. Prof. Dr. Ing. Marcus Baum
Institut für Informatik, Georg-August-Universität Göttingen, Germany
Fakultät für Informatik und Mathematik, Universität Passau, Germany

Prof. Dr. Sarbani Basu
Department of Astronomy, Yale University, New Haven, CT, USA

Dr. ir. Saskia Hekker
Max-Planck-Institut für Sonnensystemforschung, Göttingen, Germany
Stellar Astrophysics Centre, Aarhus University, Denmark

Date of the oral examination: May 16, 2018

Bibliografische Information der Deutschen Nationalbibliothek

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über `http://dnb.d-nb.de` abrufbar.

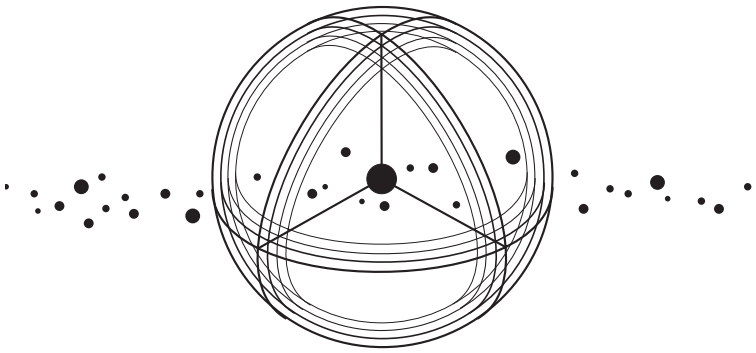
ISBN 978-3-944072-61-6

uni-edition GmbH 2018
http://www.uni-edition.de

Printed in Germany

"Equipped with his five senses, man explores the universe around him and calls the adventure Science."

— Edwin Hubble, 1929



ABOUT THE COVER. An artist's depiction of a star's interior being revealed through its oscillations. The background shows stars from the constellation of Cygnus, which was observed for four continuous years by the *Kepler* space observatory. The cover was created for use in this thesis by K. Casey Shea.

Contents

List of Figures

Summary

Asteroseismology allows us to probe the internal structure of stars through their global modes of oscillation. Thanks to missions such as the NASA *Kepler* space observatory, we now have high-quality asteroseismic data for nearly 100 solar-type stars. This presents an opportunity to measure the core structures of these stars as well as their ages, masses, radii, and other fundamental parameters.

This thesis is primarily concerned with two inverse problems in asteroseismology. The first is to estimate the fundamental parameters of stars from observations using evolutionary arguments. This is inverse to the forward problem of simulating the theoretical evolution of a star, given the initial conditions. We solve this problem using supervised machine learning in Chapter 2. We find ages, masses, and radii of stars with uncertainties (in the sense of precision) better than 6%, 2%, and 1%, respectively. We furthermore use unsupervised machine learning to quantify how each kind of observation of a star is related to its fundamental parameters in Chapter 3.

The second problem is to infer the structure of a star from its frequencies of pulsation using asteroseismic arguments. This is inverse to the forward problem of calculating the theoretical pulsation frequencies for a known stellar structure. Solving this problem presents an opportunity to test the quality of stellar evolution models, as we may then directly compare the asteroseismic structure of a star against theoretical predictions. We solve this problem in Chapter 4. Applying this technique to the solar-type stars in 16 Cygni, we find that while the structure of the 1.03 solar-mass star 16 Cyg B is in good agreement with theoretical expectations, the more massive 16 Cyg A differs in its internal structure from best-fitting evolutionary models.

These inverse problems are both *ill-posed* in the sense that (I) a solution may not exist within the confines of the current theory; (II) if there is a solution, it may not be unique, as many solutions may be consistent with the data; and/or (III) the solutions may be unstable with respect to small fluctuations in the input data. Therefore, care must be put into determining possible solutions and applying regularization where necessary.

Chapter 1 introduces this thesis with the history and theory of stellar structure, evolution, and pulsation; and emphasizes the role that variable star astronomy played in shaping our understanding of stellar evolution. It also contains the kernels of stellar structure, an introduction to ill-posed inverse problems, and a discussion of some computational issues for the algorithms used to solve these problems.

Zusammenfassung

Die Asteroseismologie erlaubt es uns, die innere Struktur der Sterne durch Messungen ihrer globalen Schwingungsmoden zu untersuchen. Dank Missionen wie dem Weltraumteleskop *Kepler* der NASA verfügen wir heute über qualitativ hochwertige asteroseismische Daten von fast 100 sonnenähnlichen Sternen. Dies bietet die Möglichkeit, das Innere dieser Sterne sowie deren Alter, Masse, Radien und andere fundamentale Parameter zu bestimmen.

Diese Doktorarbeit beschäftigt sich in erster Linie mit zwei inversen Problemen der stellaren Astrophysik. Das erste Problem besteht darin, die fundamentalen Parameter eines Sterns aus seinen Beobachtungen mit Hilfe von Argumenten der Sternevolution zu schätzen. Dieses Problem ist invers zu dem Vorwärtsproblem der Simulation der theoretischen Sternentwicklung unter bestimmten Anfangsbedingungen. Mit Hilfe von Methoden des überwachten maschinellen Lernens wird dieses Problem in Kapitel 2 gelöst. So ermitteln wir Alter, Masse und Radien mit einer Unsicherheit von weniger als 6%, 2% und 1%. In Kapitel 3 verwenden wir Methoden des unüberwachten maschinellen Lernens, um zu quantifizieren wie genau sich die fundamentalen Parametern eines Sterns durch die Kombination verschiedener Arten der Sternbeobachtung bestimmen lassen.

Das zweite Problem besteht darin, die Struktur eines Sterns aus seinen Pulsationsfrequenzen abzuleiten, wobei nur asteroseismische Argumente verwendet werden. Dieses Problem ist invers zu dem Vorwärtsproblem der Berechnung der theoretischen Pulsationsfrequenzen einer bekannten Sternstruktur. Die Lösung dieses Problems bietet die Möglichkeit, die Qualität unserer Modelle der Sternentwicklung zu testen, da wir so die asteroseismische Struktur eines Sterns direkt mit theoretischen Vorhersagen vergleichen können. Dieses Problem wird in Kapitel 4 gelöst. Wendet man diese Technik auf die beiden sonnenähnlichen Sterne des Systems 16 Cygni an, so stellt man fest, dass die Struktur des $1,03$ Sonnenmassensterns 16 Cyg B in guter Übereinstimmung mit den theoretischen Vorhersagen ist, während sich der massivere Stern 16 Cyg A in seiner inneren Struktur von den am besten passenden Evolutionsmodellen unterscheidet.

Diese inversen Probleme sind im mathematischen Sinne inkorrekt gestellt, sodass (I) eine Lösung innerhalb der Grenzen der aktuellen Theorie möglicherweise nicht existiert; (II) wenn es eine Lösung gibt, muss sie nicht eindeutig sein, da viele Lösungen mit den Daten konsistent sein können; und/oder (III) die Lösungen können in Bezug auf kleinere Schwankungen der Ausgangsdaten insta-

bil sein. Daher wird viel Sorgfalt darauf verwendet, die Menge der möglichen Lösungen zu bestimmen und bei Bedarf eine Regularisierung vorzunehmen.

Kapitel 1 leitet diese Arbeit mit der Geschichte und Theorie der Sternstruktur und -evolution ein. Der Schwerpunkt liegt hierbei auf der Theorie der stellaren Pulsationen und wie sie dazu beigetragen hat, unser Verständnis der Sternevolution zu formen. Des Weiteren enthält es Ableitungen der Integralkerne der stellaren Struktur, eine kurze Einführung in die mathematisch inkorrekt gestellten inversen Probleme, und eine Diskussion über einige numerische Schwierigkeiten bezüglich des maschinellen Lernens und der statistischen Algorithmen die verwendet werden, um diese Probleme zu lösen.

1

Introduction

1.1 Variable Stars

Points of light in the night sky are not constant but rather they are *variable*: they dim or brighten over time. Some of these variations are periodic: they dim and brighten again with a kind of regularity. This fact may have been known as early as the time of the ancient Egyptians, who, over 3,200 years ago, recorded in their calendars the 2.85-day period of the so-called "Demon Star," Algol (e.g., Jetsu and Porceddu 2015). Periodic variables were not known in the Western world however until around the 17th century, after the German pastor David Fabricius and his son observed the reappearance of a faded object that they had previously assumed to be a nova (e.g., Catelan and Smith 2015). This object was named *Mira*, Latin for 'Wonderful' (Hevelius 1662).

Regardless of variability, it was still not yet known at this time what these points of light in the sky actually were. Extrapolating from Copernicus (1543), the Italian philosopher Giordano Bruno (1584) was the among the first in the Western world to suggest that these lights are in fact *stars* not unlike our own Sun. Though the 17th century began with Bruno being burned at the stake for this heresy (e.g., De Lucca 1998), the recognition of this viewpoint fortunately became commonplace over the following centuries due to the efforts of figures such as Kepler (1609), Galileo (1610), Newton (1686), Huygens (1698), Bessel (1838), and Secchi (1877).

The field of research into periodic variable stars arguably began in the year 1638 when the Frisian astronomer Johannes Holwarda measured the period of Mira to be about 11 months long (e.g., Hoffleit 1997). Algol itself was not rediscovered in the West as being variable until 1667, although others may have seen it without noting it as such (e.g., Bolt et al. 2007). Throughout this and the following century, astronomers such as Hevelius (1671) and Flamsteed (1725) made remarks about a number of stars that seemed to appear, disappear, or otherwise change in brightness; but they did not study them further (e.g., Pigott 1786).

In the 18th century, the English astronomer Edward Pigott and his distant cousin, the short-lived and deaf John Goodricke, calculated the period of Algol as 2.865 days—a few minutes shorter than the present-day observed value (Goodricke 1783, 1784, Baron et al. 2012). They also discovered another variable

star, β Lyrae, whose symmetric light curve resembled Algol's (Pigott 1785). Pigott assembled these and ten "undoubtedly changeable" others—along with 38 more candidates—into the first-ever catalog of variable stars (Pigott 1786).

To explain the variability, the English polymath John Michell used statistical arguments to reason that stars likely group together and form systems, with "the odds against the contrary opinion being many million millions to one" (Michell 1767). The light coming from stars could be then eclipsed, with stars or other objects (planets, moons) regularly passing in front of one another in our line of sight to block the light from reaching our eyes.

At first, Pigott and Goodricke posited that the variability of Algol was caused by eclipses, as Michell had proposed (Goodricke 1783). However, within three years they changed their interpretation, then attributing its variability to "rotation of the star on its axis, having fixed spots that vary only in their size" (Goodricke 1786). This idea may have seemed attractive due to their knowledge of sunspots, which had been known in the Eastern world since at least the time of the Babylonians, though not rediscovered in the West until 150 years prior when Fabricius and his son turned their telescopes to the Sun following their discovery of Mira (Fabricius 1611).

Pigott and Goodricke also discovered two other periodic variable stars, η Aquilae and δ Cephei (Pigott 1785, Goodricke 1786). These stars earned a new name—*Cepheid* variable stars—as the manner in which their light changed over time was noticeably different from Algol's. Rather than quickly dipping and brightening again every so often, these stars appear to change continuously (for a visual comparison, see Figure 1.1). Unlike with Algol, they offered no explanations for Cepheid-type variability. Goodricke died that year at the age of 21, having been elected a Fellow of the Royal Society only days prior, but never learning of the honor.

For a long time thereafter, the discovery of variable stars slowed. Less than ten new variables were discovered in the following 60 or so years. These new variables were published by the German astronomer Friedrich Wilhelm Argelander (Argelander 1844), to whom the variable star naming convention[1] is owed. The only other major advance in the first half of the 19th century was the development of the least squares method, which improved period estimates (e.g., Zsoldos 1994).

In the second half of the 19th century, the fields of astronomical spectroscopy and dry plate astrophotography were born. These technologies proved a great aid for the discovery and analysis of variable stars, and even revealed the existence of several new classes of variable stars. By 1865, the number of known variable stars had more than doubled, going up to 123 (Chambers 1865). In the next 30 years, that number quadrupled with over 300 new discoveries (e.g., Hoffleit 1997). Nearly 50 variable stars were discovered in the year 1896 alone, the majority of which being Mira-type variables, 19 of which were found by the

[1] Starting with the letter R and the name of the constellation where it is found (e.g., R Lyrae), then repeating with double letters when the alphabet is exhausted (e.g., RR Lyrae).

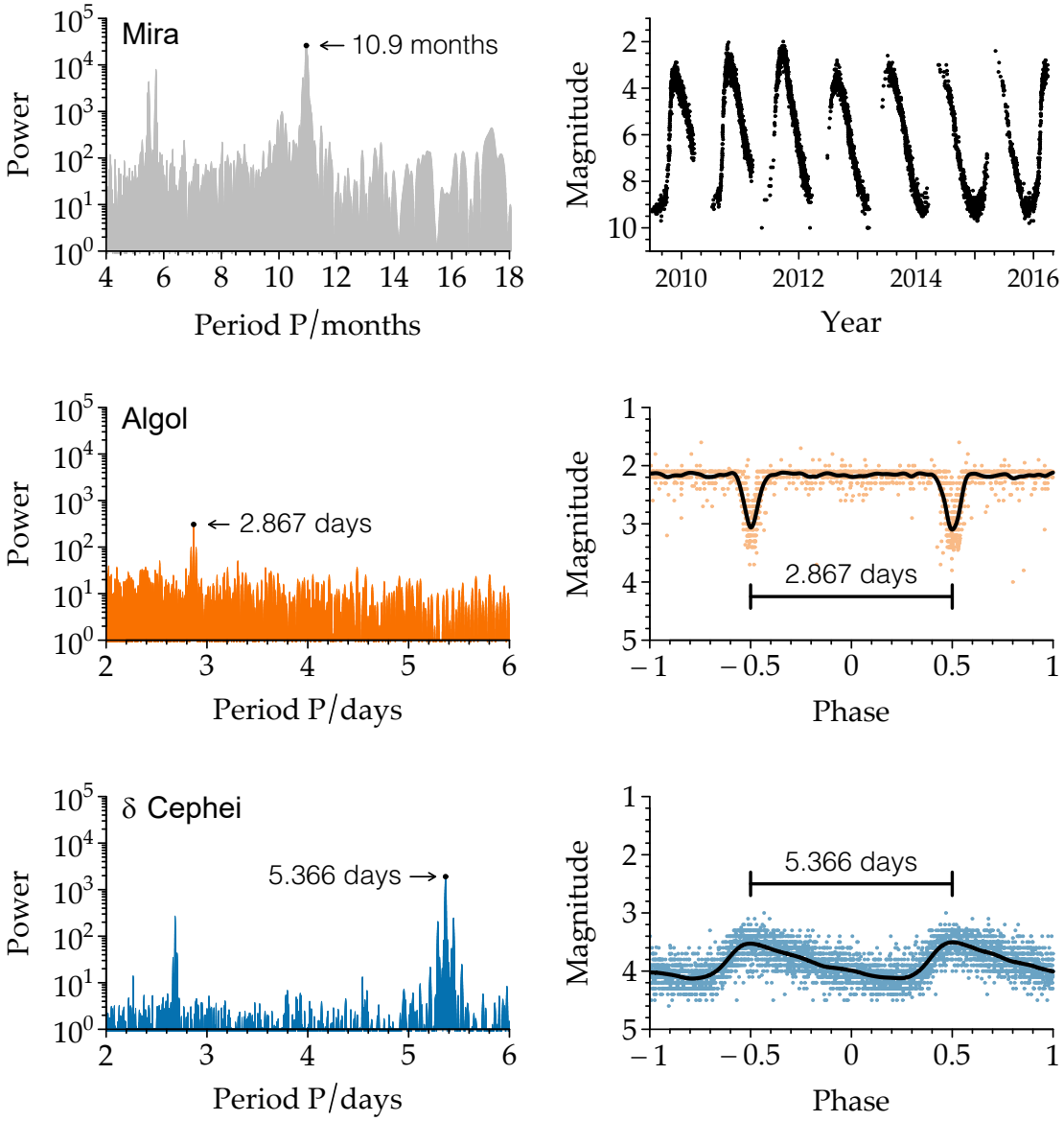


FIGURE 1.1. Modern-day periodograms and light curves for Mira (o Ceti), Algol (β Persei), and δ Cephei—the "prototypes" for the first three discovered classes of periodic variable stars. The light curves for Algol and δ Cephei are phased by their period. Mira has a long and somewhat irregular period. Unlike the other two, which are constantly changing in brightness, the light from Algol is generally stable with occasional quick dips. *Data acquired from the American Association of Variable Star Observers* (AAVSO, Kafka 2017).

Harvard "computer" Williamina P. Fleming. By the end of the 19th century, the number of known variable stars grew to at least 2000 (e.g., Samus' et al. 2017).

The latter half of the 19th century also marked the beginning of a change in attitude toward astronomical research. In addition to cataloging the sky, researchers began seeking rigorous physical foundations to understand the nature of the Sun and the stars. Applying techniques from the recently-born field of thermodynamics, figures such as William Thomson (a.k.a. Lord Kelvin), Julius Robert Mayer, Hermann von Helmholtz and others worked to determine the ages of stars and identify the sources of their energy. In particular, they offered the explanation that gravitational energy can be converted into heat via either contraction or the infall of meteoric material. For example, Helmholtz demonstrated that the Sun could be powered by contracting merely 380 feet each year (e.g., Arny 1990). Now called the Kelvin-Helmholtz mechanism, this was the only known form of stellar heating at the time. Applying it to the study of the Earth and Sun, Kelvin found that the solar system must be at most millions of years old (e.g., Kelvin 1895), much younger than the currently accepted age of about 4.57 billion years.[2]

The calculations that Helmholtz and Kelvin made required details of the structure of the Sun, and so to carry them out, they created the first polytropic models of stellar structure (e.g., Arny 1990). These models are characterized by the internal pressure depending only on the density of the stellar material. Much as is still done today, they considered a sphere where gravity forces are in balance against pressure forces. However, they erroneously assumed that all energy in the Sun is transported by convection.

It was also around this time that the idea stars might pulsate was first given serious attention. Lord Kelvin was the first to state the equations of non-radial pulsation for chemically homogeneous "spheroids of incompressible liquid" (Thomson 1863). Though this work makes no explicit mention of stars, it was thought at this time that stars might be entirely liquid (e.g., Arny 1990). However, it was argued for a long time thereafter that stars could not possibly pulsate non-radially, as these modes of oscillation would be damped out by viscous forces (e.g., Pekeris 1938). Figure 1.2 shows some of the configurations that a star could take under the pulsation hypothesis.

After completing his Ph.D. at the University of Göttingen, the German astrophysicist August Ritter wrote a series of 19 papers over an 11-year span laying out theory of stellar structure (1878–1889, e.g., Ritter 1880). Ritter had the insight to treat stars as an ideal gas, and derived a relationship between the mass of a star and its luminosity. Ritter also developed here the radial theory of stellar pulsations, including the important result connecting the period of stellar pulsation to the mean density of the star. Since the source of stellar variability

[2] A devout Christian, Lord Kelvin used these results to doubt Charles Darwin's recently-published theory of biological evolution, which requires an older Earth (Darwin 1859).

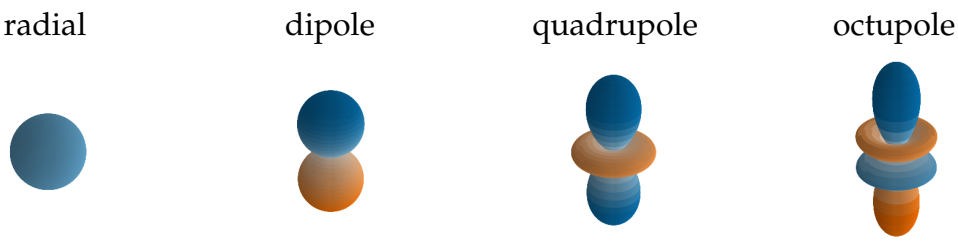


FIGURE 1.2. Radial and non-radial stellar pulsations for a non-rotating star. Mathematically, these show $r(\theta, \phi) = \mathrm{Re}\,|Y_\ell(\theta, \phi)|$ in spherical polar coordinates for $\ell = 0, 1, 2, 3$, where Y_ℓ is the solution to Laplace's equation on a sphere—special functions known as *spherical harmonics*. The sign of $\mathrm{Re}(Y_\ell)$ is indicated by color. Pulsations with $\ell = 0$ correspond to the entire star moving toward or away from the center without horizontal motions, i.e., radial pulsations.

was still an open puzzle, Ritter conjectured that stars might be radial pulsators. Unfortunately, this work was largely ignored.[3]

In an attempt to understand the temperature of the Sun, the American theoretical astrophysicist and Yale alumnus J. Homer Lane continued work on polytropes (e.g., Lane 1870). Lane discovered the curious fact that stars have a negative heat capacity: i.e., when they lose energy, they contract and heat up. Ritter rederived Lane's Law and used it to develop the first physically-motivated (albeit incorrect) theory of stellar evolution: that a star begins its life as a diffuse gaseous mass, which at first contracts and heats; eventually, the star transforms into a liquid, and then undergoes a long period of cooling.

At the end of that century, the German astronomer Hermann Carl Vogel used spectroscopic measurements to firmly establish that Algol is an eclipsing binary, thereby confirming Goodricke's initial speculation (Vogel 1889, Frost 1908). Vogel taught his methods to the Russian astronomer Aristarkh Bélopolsky, who then took spectra of the Cepheid stars δ Cephei and η Aquilae. Though at this time eclipses were widely thought to be the most likely the source of Cepheid variability, Bélopolsky argued that the radial velocity variations of these stars were inconsistent with the eclipse hypothesis (Bélopolsky 1897, 1895).

> *"The times of minimum brightness and the times for which the velocity in the line of sight is zero do not coincide. For this reason the changes in the brightness of the star cannot be explained as the result of eclipses, and some other explanation must be sought."*
>
> — Aristarkh Apollonovich Bélopolsky
> *Researches on the spectrum of the variable star η Aquilae* (1897)

Several alternative theories for Cepheid variability arose over the years. So-called "veil theories" suggested that clouds could rapidly form and evaporate,

[3] In his influential textbook *An Introduction to the Study of Stellar Structure*, Nobel laureate Chandrasekhar (1939) characterized this body of work as "a classic, the value of which has never been adequately recognized," and noted that in these works Ritter worked out "almost the entire foundation for the mathematical theory of stellar structure."

serving to block the source of the light for a short time (e.g., Brester 1889). English astronomer Henry Plummer, later President of the Royal Astronomical Society, suggested that Cepheids are radial pulsators (Plummer 1914). Others maintained the eclipsing binary hypothesis (e.g., Duncan 1909) with some even claiming that Bélopolsky's measurements had in fact proven it (e.g., Brunt 1913).

Regardless of the cause of their blinking, Cepheid stars gained near-immediate fame throughout astronomical circles and beyond following the discovery by American astronomer Henrietta Swan Leavitt (another Harvard 'computer') that "the brighter variables have the longer periods" (Leavitt 1908, 1912). Now known as the Cepheid Period-Luminosity Relation or the *Leavitt Law*, this enabled measurement of vast cosmic distances via comparison of observed brightnesses with those expected from Cepheid periods. This discovery thus established Cepheids as standard candles—the first to be discovered—and was quickly put to use in mapping the structure of the Universe (see Figure 1.3).

Around this time, the then-unknown Danish astronomer Ejnar Hertzsprung was working to combine spectroscopy of stars with parallax distance measurements. He found that stars form two distinct groups: "Riesen" (giants) and "Zwerge" (dwarfs). Hertzsprung published this work in a photographic journal with little impact (Hertzsprung 1905, 1907). It did however get the attention of Karl Schwarzschild, director of the Göttingen Observatory, who then appointed him to a position there (e.g., Bolt et al. 2007). Hertzsprung went on to discover that the pole star Polaris is also a Cepheid-type variable[4] (Hertzsprung 1911) and furthermore concluded that Cepheids are giant stars (Hertzsprung 1913).

The director of the Princeton Observatory, Henry Norris Russell, a much more influential astronomer at the time, also came to the same conclusions as Hertzsprung (e.g., Russell 1913a,b). Plotting the absolute magnitudes of more stars against their spectral type (see Figure 1.4), Russell showed that there was a main diagonal where dwarfs lived, an upper corner where red giants lived, and a lower corner lacking any stars "except for one star[5] whose spectrum is very doubtful." Russell argued that this confirmed Ritter's theory of evolution.

The following year, Harlow Shapley wrote a seminal paper laying out the collective arguments against the eclipsing binary hypothesis of Cepheid variable stars (Shapley 1914). First, Bélopolsky had already shown that the brightness and radial velocity variations did not coincide. Second, the periods of some Cepheids are themselves variable. Third, the shapes of the light curves for some Cepheids change from cycle to cycle (e.g., Curtiss 1905). And lastly, "the best argument," since Hertzsprung and Russell had just shown that Cepheids are giant stars, the companion star would need to be inside of the Cepheid in order for eclipses to explain the observed behavior—a ridiculous hypothesis. Shapley concluded that Cepheid variability is most likely due to pulsation.[6]

[4] Hence, Caesar is as constant as a variable star (Shakespeare 1599).

[5] This would later be recognized the first-discovered white dwarf (e.g., Schatzman 1958).

[6] It is interesting to note here that John Michell had posed both the theory of earthquakes (Michell 1759) and the explanation of stellar variability in terms of eclipsing stars (Michell 1767), but probably never imagined that stars quake, too.

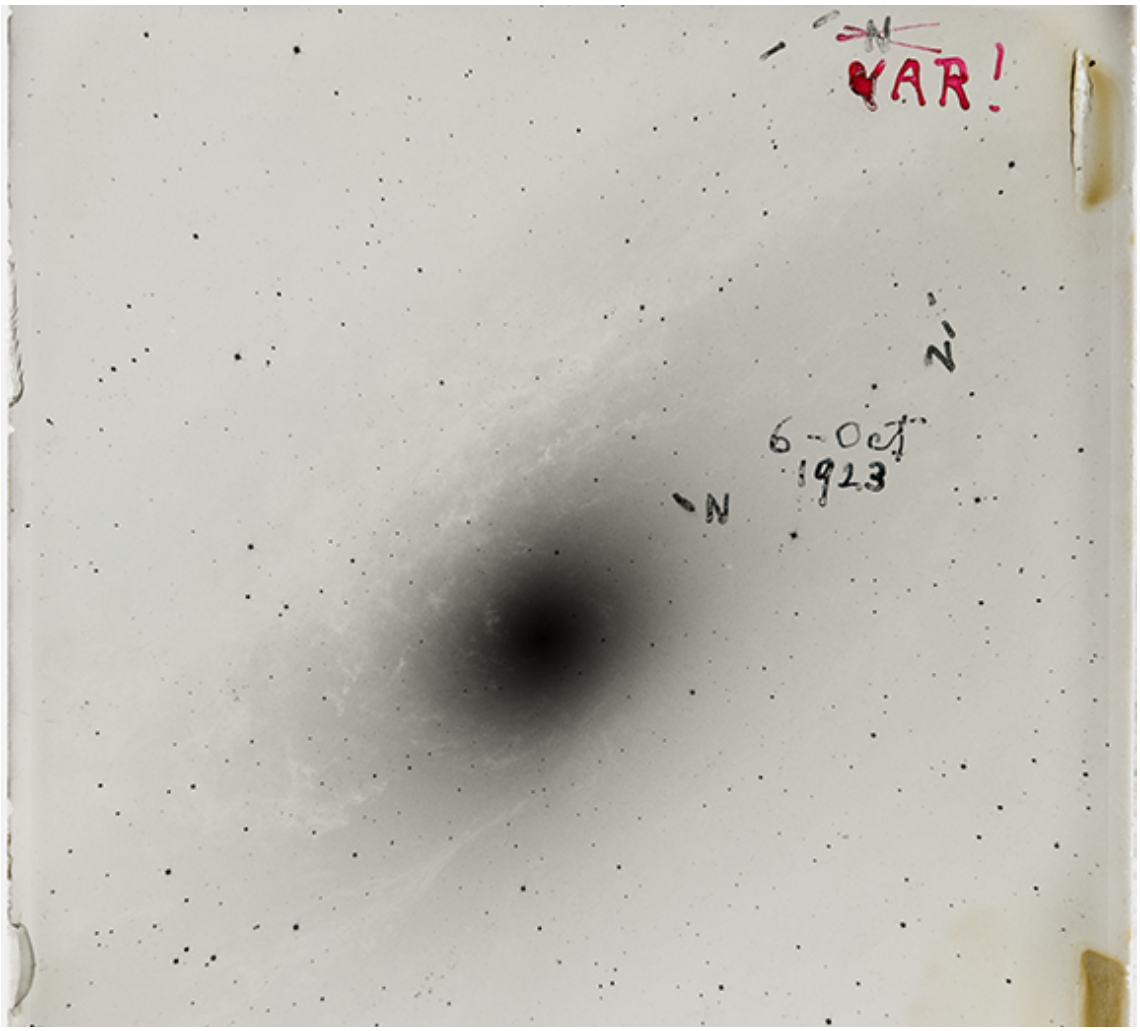


FIGURE 1.3. Edwin Hubble's photographic plate showing the discovery of a Cepheid variable star in the Andromeda Galaxy (M31). In a series of 17 papers, Harlow Shapley used the Leavitt Law to estimate the distance to globular clusters and map out the size of the Galaxy, finding that it was substantially larger than previously estimated (Shapley 1918). In 1920, Shapley engaged in the "Great Debate" of astronomy, in which he argued that the Milky Way comprised the entirety of the Universe (Shapley and Curtis 1921). Soon thereafter, Edwin Hubble used this same technique to measure the distance to the spiral nebulae M31 and M33 (Hubble 1925, see image). Finding that they were extremely distant, Hubble proved that these nebulae were in fact galaxies external to our Milky Way—instantly expanding the calculated size of the Universe by a factor of 100,000. Hubble sent these results to Shapley, who, upon viewing them, is said to have remarked: *"Here is the letter that has destroyed my Universe."* Edwin Hubble subsequently used the Leavitt Law to estimate the distances to several more Cepheid-host galaxies (Hubble 1929). Combining these distances with measurements of the speeds at which those galaxies are receding from us, Hubble measured the rate of cosmic expansion, and thus the age of the Universe. Variable star enthusiasts can celebrate October 6 as "VAR! Day" (see image). *(Image reprinted with permission from Carnegie Observatories.)*

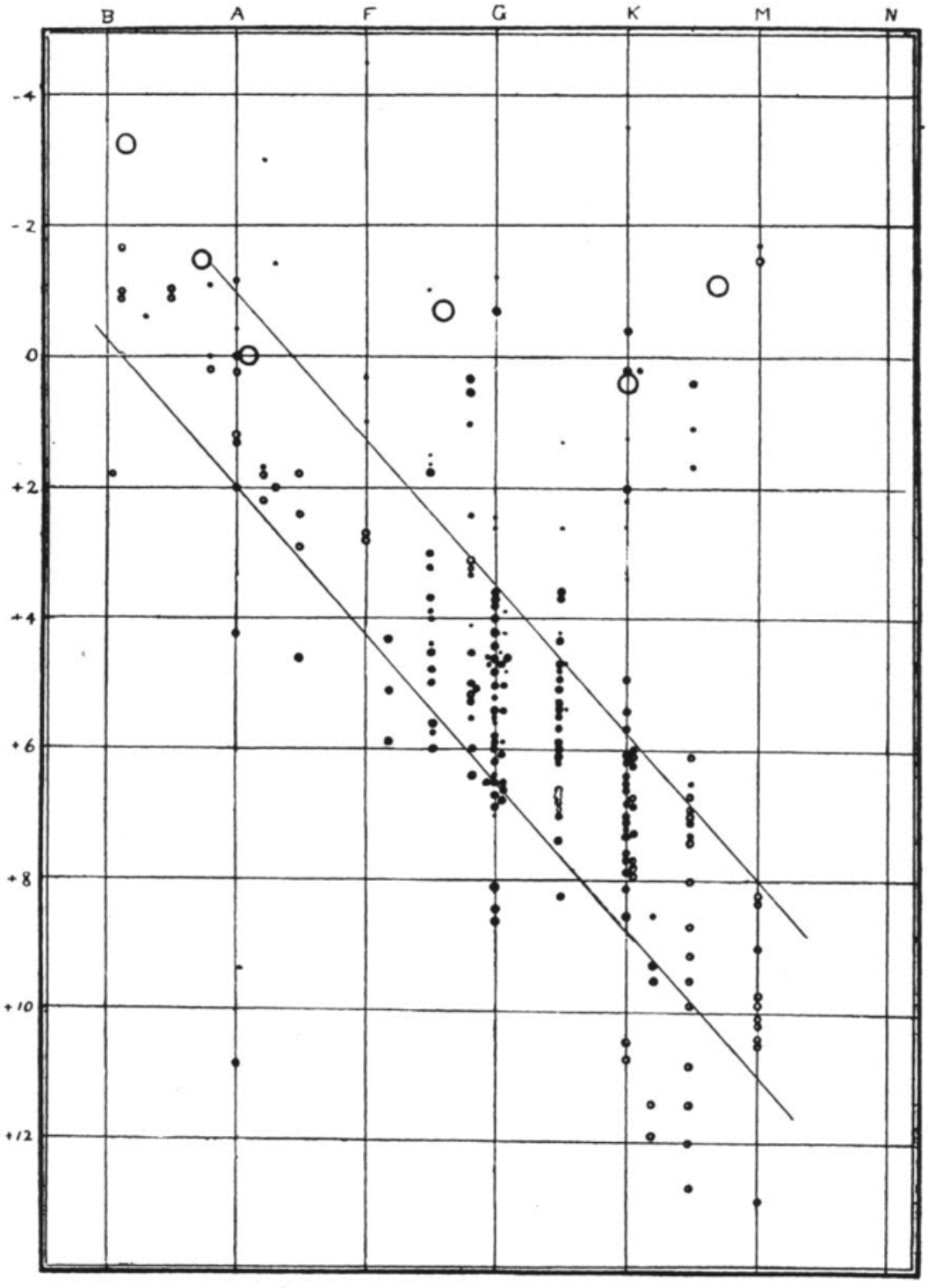


FIGURE 1.4. One of the first Hertzsprung-Russell diagrams, showing the absolute magnitude of stars against their spectral type. Luminosity increases upward; temperature increases leftward. Dwarf stars reside on the diagonal—the *main sequence*—and giant stars occupy the upper right corner. *(Figure reprinted with permission from Russell 1914.)*

> *"Cepheid variables are not binary systems... the explanation of their light-changes can much more likely be found in a consideration of internal or surface pulsations of isolated stellar bodies."*
>
> — Harlow Shapley
> *On the Nature and Cause of Cepheid Variation* (1914)

Thus the pulsation hypothesis was born. But the theory had its doubters. There was no real proof yet—only very strong evidence that the eclipsing binary hypothesis was wrong—and no known mechanism for the pulsation. Many, including the eminent star formation theorist James Jeans, rejected the idea of stellar pulsations, Jeans himself arguing that Cepheid variation is rather caused by repeating explosions (e.g., Jeans 1919). Moreover, many aspects of stellar theory still had major flaws. It was still not yet discovered how stars really get their energy, nor how they transport it throughout the interior, nor what they are made of, nor what state of matter they are in, nor how they evolve. Jeans himself in fact still held the view that stars are liquid (e.g., Jeans 1928).

The modern view of the stars really began to take hold in the early 20th century with the work of Arthur Eddington. Building upon earlier works by Schwarzschild (1906) and Sampson (1895), Eddington developed the first models of radiative transport in stellar interiors (e.g., Eddington 1916). Combating the view that stellar energy is transported entirely by convection, Eddington worked out the balance between radiative pressure—the outward pressure exerted by the enormous numbers of photons streaming through the star—with the inward pressure exerted by the gaseous stellar material. This led to the creation of his "standard model"—a purely radiative star. This treatment complicated stellar models greatly, as the internal structure then depended on the opacity and mean molecular weight of the stellar matter, which were unknown (e.g., Arny 1990).

The following year, Eddington provided the mechanism for Cepheid variability (Eddington 1917). Applying thermodynamics to the study of the interior, Eddington argued qualitatively that Cepheids pulsate due to an internal heat engine: repeated expansion and collapse due to cyclical ionization and recombination of atoms. The following year, he numerically calculated the periods of his stellar models using a linear adiabatic treatment of stellar pulsation, and found good agreement with observations (Eddington 1918). Though further confirmations would come later, this was already strong evidence for the pulsation hypothesis.

Eddington then went on to use observations of Cepheids to dispute the Kelvin-Helmholtz mechanism as being the sole source of stellar longevity (Eddington 1920). If stars survive on contraction alone, he argued, then their rate of rotation should speed up relatively rapidly due to the conservation of angular momentum. This was not what had been observed. Similarly, if the pulsation hypothesis is true, then their period of pulsation should change in accordance with changes to their mean density.

> *"Now, on the contraction hypothesis the change of density must amount to at least 1 per cent. in 40 years. The corresponding change of period should be very easily detectable. For δ Cephei the period ought to decrease 40 seconds annually. Now δ Cephei has been under careful observation since 1785, and it is known that the change of period, if any, must be very small. S. Chandler found a decrease of period of 1/20 second per annum... I hope the dilemma is plain... Only the inertia of tradition keeps the contraction hypothesis alive—or rather, not alive, but an unburied corpse."*
>
> — Sir Arthur Stanley Eddington
> *The Internal Constitution of the Stars* (1920)

Eddington furthermore rederived Ritter's mass-luminosity relation, and upon applying the relation to stars of spectral types B and A, found that these "dwarf" stars are even more massive than the giant stars (e.g., Eddington 1924). This too was difficult to reconcile with the prevailing theory of stellar evolution.

Eddington therefore sought another explanation. During Albert Einstein's "miracle year," Einstein had given his famous equivalence of mass and energy, $E = mc^2$ (Einstein 1905). In 1920, the English chemist and Nobel laureate Francis Aston showed that the mass of one helium atom was approximately 1% less than the sum of four hydrogen atoms (Aston 1920). At this time, it was still assumed that the solar composition was similar to that of the Earth; the amount of hydrogen in the Sun was therefore thought to be relatively small. Nevertheless, and despite lacking an exact mechanism, Eddington used these two developments to speculate that the Sun and stars survive via hydrogen fusion (Eddington 1920).

> *"A star is drawing on some vast reservoir of energy by means unknown to us. This reservoir can scarcely be other than the sub-atomic energy which, it is known, exists abundantly in all matter; we sometimes dream that man will one day learn how to release it and use it for his service... The atoms of all elements are built of hydrogen atoms bound together, and presumably have at one time been formed from hydrogen; the interior of a star seems as likely a place as any for the evolution to have occurred; whenever it did occur a great amount of energy must have been set free; in a star a vast quantity of energy is being set free which is hitherto unaccounted for."*
>
> — Sir Arthur Stanley Eddington
> *The Internal Constitution of the Stars* (1920)

Within five years, Harlow Shapley's Ph.D. student Cecilia Payne showed that hydrogen is about a million times more prevalent in the Sun and stars than on the Earth (Payne 1925). Within two years, the Göttinger physicist Friedrich Hund discovered quantum tunnelling, which gives atomic nuclei a probability of penetrating the Coulomb barrier and achieving thermonuclear fusion (Hund 1927, Nimtz and Clegg 2009). The following year, George Gamow brought this concept to the astrophysical community (Gamow 1928), and Eddington's speculation was proved. Eddington calculated new stellar models that included hy-

drogen burning, and found that this mechanism could power the Sun for billions of years (Eddington 1926).

This was not the end of the story, however. Though hydrogen fusion was now known to fuel the stars, there were still major discrepancies between theory and observation. Using the assumption that the stellar interior is chemically homogeneous, George Gamow calculated evolutionary tracks and found that his models failed to become giant stars (Gamow 1938, see also Figure 1.5). He furthermore found that he could not reproduce the mass-luminosity relation.

The solution came that same year, though it would not be widely recognized until long after. Discarding Gamow's mixing hypothesis, the Estonian astrophysicist Ernst Öpik realized that hydrogen fusion could continue burning in a shell after it had been exhausted in the core. Applying this insight, Öpik succeeded in hand-calculating stellar models that evolve from the main sequence up the red giant branch (Öpik 1938). Thus, the major features of the H-R diagram were explained. Unfortunately, it would be decades before this solution was rediscovered using digital computers (e.g., Arny 1990). Although there was still much to do about the evolution beyond the red giant branch—and although debates continue to this day over why stars actually become giants (e.g., Eggleton and Faulkner 1981, Renzini et al. 1992, Weiss 1983, Yahil and van den Horn 1985, Applegate 1988, Whitworth 1989, 1991, Sugimoto and Fujimoto 2000, etc.)—this essentially captured the first phases in the modern picture of stellar evolution.

There was still one more major hitch that needed to be reconciled. Around the same time that these issues were being resolved, the German-born American astronomer Edward Arthur Fath discovered that δ Scuti—a star with much resemblance to the Cepheids—has more than one period of pulsation (Fath 1935). This brought serious challenges to the theory of stellar pulsation, as the second period measured was inconsistent with the mean density of the star (Sterne 1938, 1940).

> "*One is practically forced to the conclusion that the existence of the pair of periods would be inconsistent with the pulsation theory... If the [second period] is correct, the pulsation theory is seriously jeopardized.*"
>
> — Theodore Eugene Sterne
> *The Secondary Variation of δ Scuti* (1938)

Sterne's argument rested on the longstanding assumption that these modes of pulsation needed to be purely radial in nature. Challenging this view, Pekeris (1938) continued Lord Kelvin's work from 75 years prior to further flesh out the mathematics of non-radial stellar pulsations, only now dealing with heterogeneous chemical compositions—a much more difficult problem. Cowling (1941) used this description to calculate the non-radial pulsation frequencies of a stellar model (though his attention was toward binary interactions). Such calculations would prove invaluable in the decades to come, as it would be applied to a much more familiar star: the Sun.

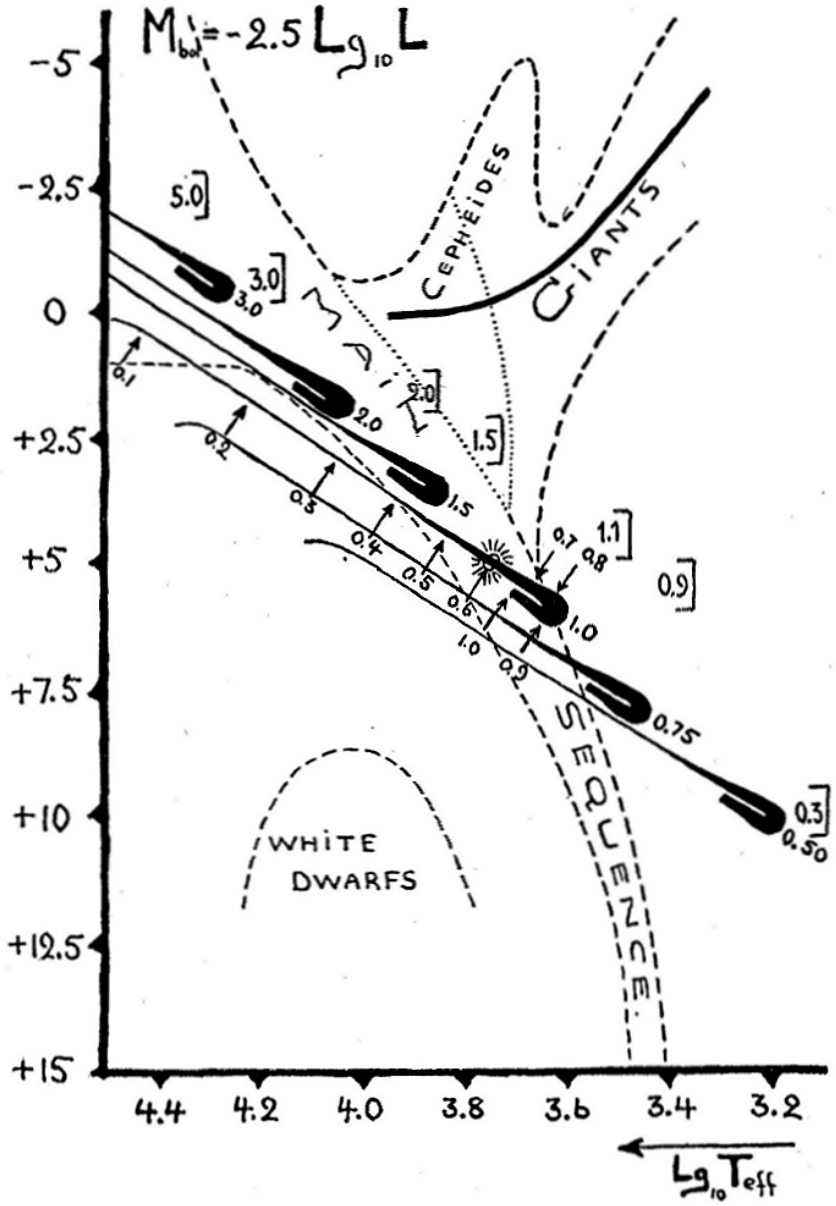


FIGURE 1.5. Historical theoretical Hertzsprung-Russell diagram showing the evolution of stars with initial masses spanning from 0.5 to 3 solar masses. The thickness of each track indicates the time spent at that stage of evolution. The arrowed numbers indicate the amount of hydrogen. The numbers in brackets indicate masses obtained via the mass-luminosity relation. Unlike modern evolutionary tracks, the stars simulated here fail to evolve from dwarfs into giant stars. *(Figure reprinted with permission from Gamow 1938.)*

1.1.1 Helioseismology

The theory that stars pulsate—and that they can pulsate non-radially—was most definitively confirmed with the discovery in the 1960s and 1970s that our own Sun is in fact such a pulsator. Obviously, the nature of solar pulsations are of a different character than the ones discovered in other stars to have gone unnoticed for so long.

Already in 1916, the 23-year old Canadian solar astronomer Harry Plaskett had found variations in Doppler velocity measurements of the solar surface from a spectroscopic investigation into the solar rotation rate (Plaskett 1916). Whether these variations were intrinsic to the Sun, or, for example, effects from the Earth's atmosphere were unknown until the work by Hart (1954, 1956). Many regard the publication of a "preliminary report" by Caltech researchers Robert Leighton, Robert Noyes and George Simon as the birth of helioseismology (Leighton et al. 1962). In this paper, Leighton and colleagues demonstrated that the Sun has multi-periodic variations on the order of about 5 minutes (see also Figure 1.6). They were prescient in their speculation that these variations could be used to determine detailed properties of the Sun, or at least its atmosphere. Frazier (1968) and others furthermore gave evidence that solar oscillations may not merely be confined to the solar atmosphere, but may instead probe deep into the star.

In the early 70s, Ulrich (1970) and Leibacher and Stein (1971) argued that the oscillations are standing acoustic waves trapped below the solar photosphere, and showed that theoretical periods of this description match the observations. Deubner (1975) and Rhodes et al. (1977) found that the relationship between the spatial and temporal frequencies of the oscillations are in similar agreement with expectations, giving further credence to the theory. Goldreich and Keeley (1977) provided a mechanism for the origination of solar oscillations by showing that acoustic waves can be stochastically excited by turbulent convection, which is the dominant source of energy transport in the solar envelope. Claverie et al. (1979) and Grec et al. (1980) made the first identifications of low-degree modes in the Sun, which pass through the entire star, thereby confirming the global nature of the oscillations (see Figures 1.7 and 1.8).

The Sun vibrates in a superposition of a great number of low-amplitude modes simultaneously. Multiple modes of the same spherical degree ℓ (recall Figure 1.2) can be excited simultaneously. These modes are distinguished by their radial order n, i.e., the number of nodes (zero crossings) between the center and the surface. Additionally, the rotation of the Sun splits each non-radial mode of oscillation into a multiplet of $2\ell + 1$ modes, which can be distinguished by their azimuthal order m, i.e., the number of nodes along the equator.

Whereas Cepheid and RR Lyrae stars oscillate in low-order ($n \leqslant 3$) radial ($\ell = 0$) modes, the Sun and other solar-type stars oscillate in high-order ($n \lessapprox 40$) modes of both radial and non-radial ($\ell \geqslant 0$) character, though so far observations of modes with $\ell \geqslant 4$ have only been confirmed in the Sun, which is made possible by resolving the solar disk. Classical pulsators like Mira, Cepheid, RR Lyrae,

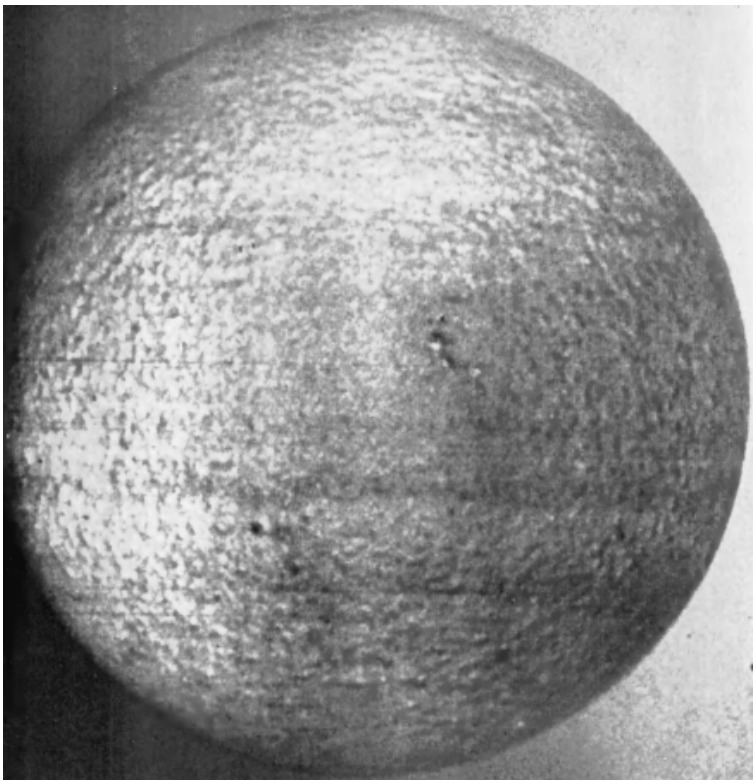


FIGURE 1.6. Velocity fields in the solar atmosphere revealed by Doppler imaging. *(Figure reprinted with permission from Leighton et al. 1962.)*

and δ Scuti stars are intrinsically unstable to their oscillations: they are self-excited by their configuration (e.g., Samadi et al. 2015). Solar-like oscillators, on the other hand, pulsate in stable modes which are both driven and damped by turbulent convection in their outer envelopes. Detailed reviews and overviews of global helioseismology have been given by, e.g., Christensen-Dalsgaard (2002), Kosovichev (1999, 2011), and Basu (2016).

Tassoul (1980) provided asymptotic descriptions for oscillation modes of high radial order ($n \gg \ell$) as seen in the Sun. Mode frequencies of the same spherical degree are equally spaced by a quantity known as the large frequency separation, denoted $\Delta\nu$, which is related to the stellar mean density and the inverse sound travel time through the star. Modes differing by a spherical degree of two (e.g., $\ell = 0$ and $\ell = 2$) and a radial order difference of one (e.g., $n = 21$ and $n = 20$) are spaced by the small frequency separation ($\delta\nu$). This quantity is related to the sound-speed gradient, and its measurement provides a good diagnostic of main-sequence age. The ratios of these quantities are also useful, because they are insensitive to near-surface layers of the star where several assumptions used to calculate theoretical mode frequencies break down (e.g., Roxburgh and Vorontsov 2003). To good approximation, these quantities vary little from one radial order to the next, and hence serve as a good summary of the frequency spectrum. In the early 1980s, Christensen-Dalsgaard & Gough applied this asymptotic description to oscillation modes calculated from a so-

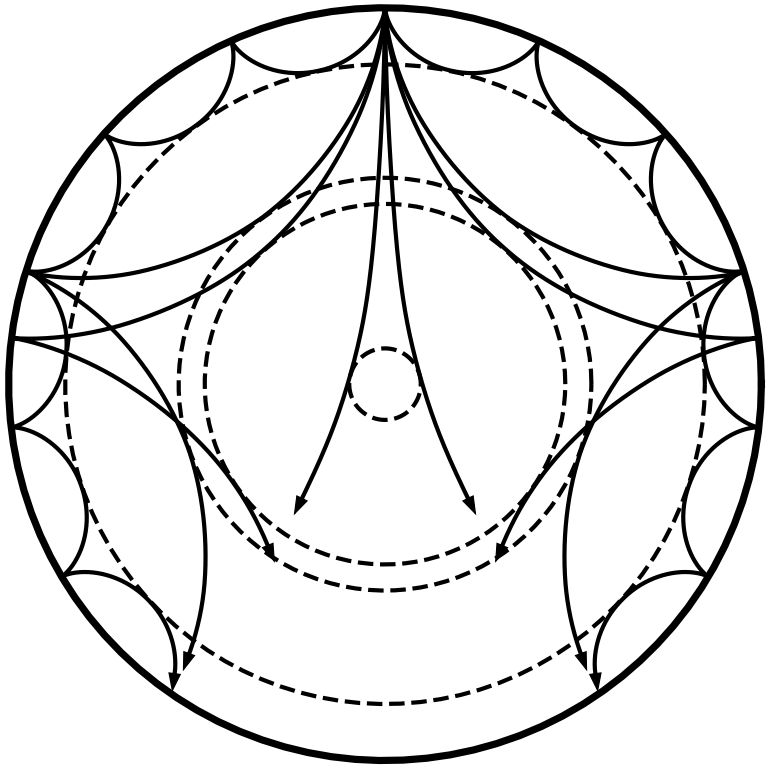


FIGURE 1.7. Ray diagram showing the paths of oscillation modes as they propagate through the interior of a solar model. The innermost circle shows the lower turning point of a quadrupole ($\ell = 2$) oscillation mode. Such kinds of modes are observable in the Sun and other stars exhibiting solar-like oscillations. The other modes are $\ell = 20$, 25, and 75, which have so far only ever been observed in the Sun. *(Figure adapted with permission from Warrick Ball [private communication] using the procedure given by Giles 2000.)*

lar model and were able to show that the model was in agreement with the observations (e.g., Christensen-Dalsgaard 2002).

Of course, helioseismic data nowadays are of superb quality. Figure 1.9 shows a power spectrum from data obtained by the Michelson Doppler Imager (MDI) instrument onboard the Solar and Heliospheric Observatory (SOHO), a €1 billion NASA/ESA space mission launched in 1995. With such data, thousands of solar oscillation modes have been resolved with high precision (e.g., Rhodes et al. 1997).

Helioseismic Inversions

Many of the confirmations of global helioseismology have come through the comparison of observations to a theoretical models constructed to match the properties of the Sun. Such models can be constructed for example via evolutionary modelling; I will discuss the creation of such models in more detail in Section 1.2. However, even to this day, no solar model matches solar oscilla-

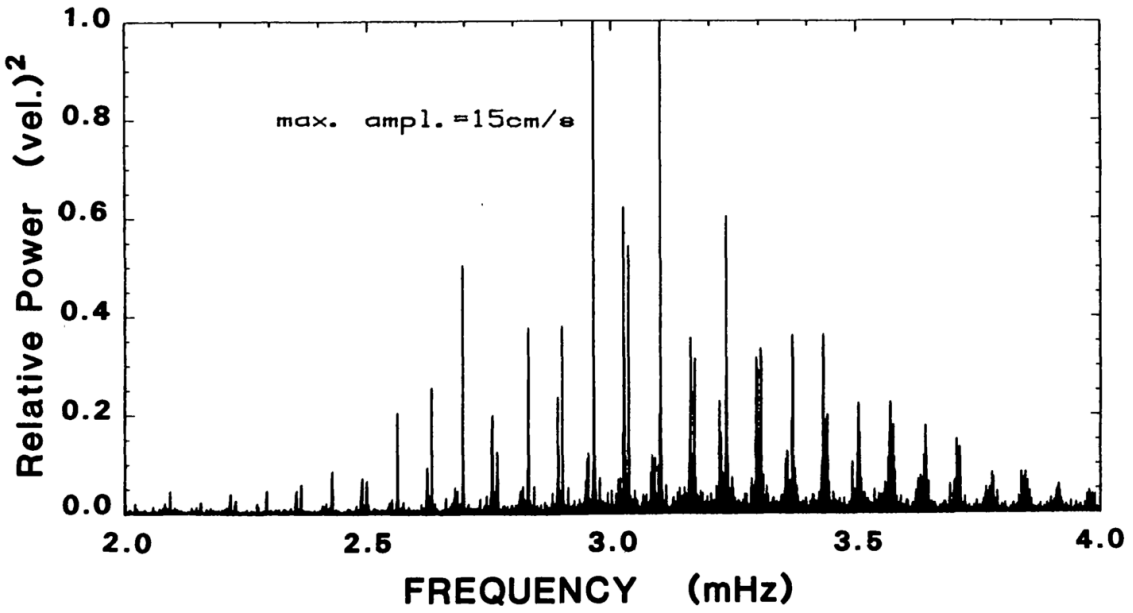


FIGURE 1.8. Power spectrum of the Sun from 3 months of observations showing its 5-minute (3 mHz) oscillations (Claverie et al. 1981). Each peak corresponds to an individual mode of oscillation. *(Figure reprinted with permission from the review article of Deubner and Gough 1984.)*

tion data exactly (e.g., Christensen-Dalsgaard and Gough 1980). The question thus arose as to whether these global oscillation modes could be used to make model-independent measurements of the solar interior, in terms of both its structure and its internal rotation rate (e.g., Christensen-Dalsgaard and Gough 1976, Gough 1981). This would need to be answered in the context of inverse theory.

> *"The astrophysicists' task is not merely to produce a theoretical model of the Sun that is not obviously at variance with observation, but to learn what the internal structure actually of the Sun is, and to understand why it is so."*
>
> — Douglas Owen Gough, FRS
> *Seismic observations of the solar interior* (1991)

The forward problem of global helioseismology is to calculate the oscillation mode frequencies for a given model of solar structure (or solar rotation). The inverse to this problem is then to calculate the structure (or internal rotation profile) from the mode frequencies. The inverse problem is ill-posed because different stellar structures can support the same oscillation pattern, including ones that are clearly nonphysical. Furthermore, unless care is taken, small errors to the input data can lead to large errors in the inversion result. I will discuss ill-posed problems in more detail in Section 1.4.

In the late 1960s, geophysicists George Backus and James Gilbert developed a stable method for inferring the structure of the Earth from seismic measurements

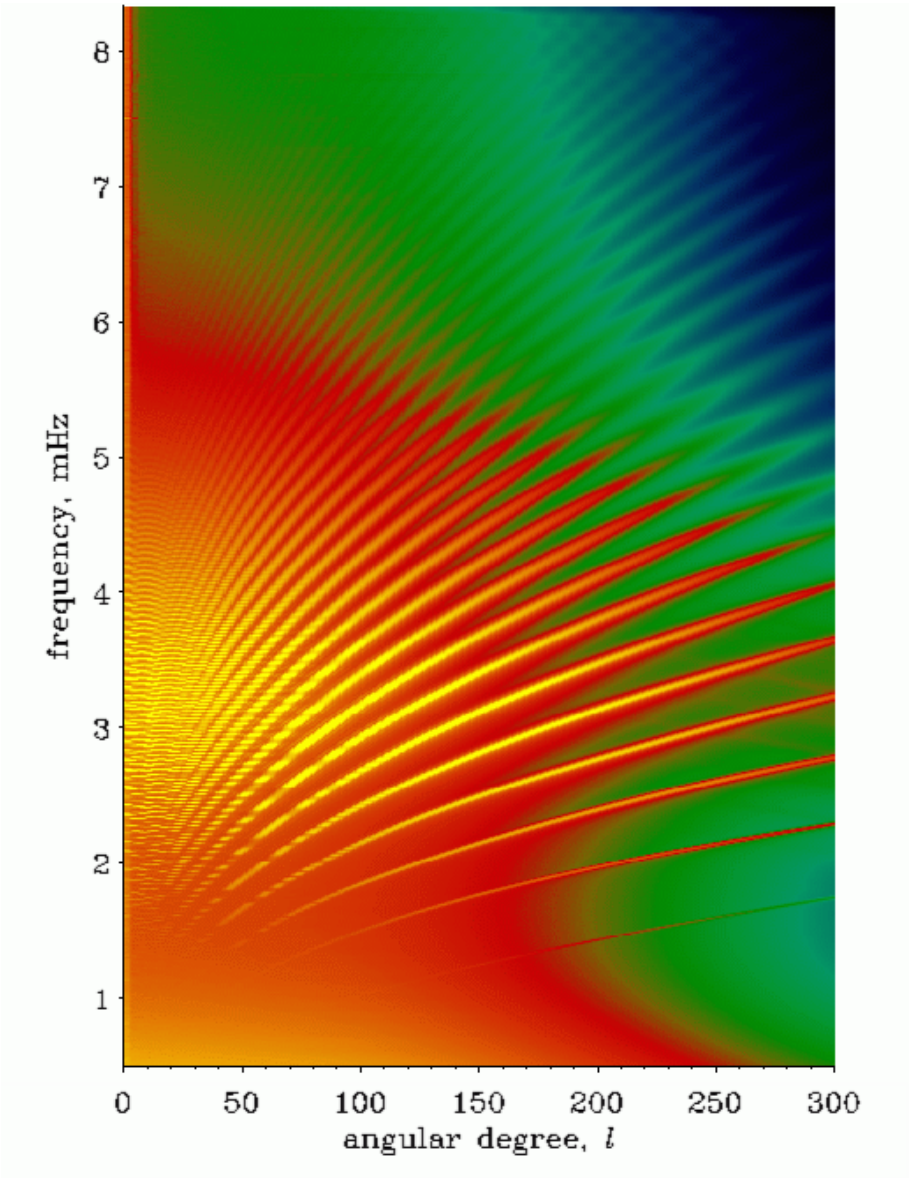


FIGURE 1.9. Solar power spectrum showing helioseismic oscillation mode frequencies as a function of spherical degree as observed by MDI over a time span of 144 days. The acoustic oscillation modes of the Sun form the ridges of high power. *(Figure reprinted with permission from Rhodes et al. 1997.)*

(Backus and Gilbert 1968, 1970). This method came to be known as the Gilbert–Backus method or the method of Optimally Localized Averages (OLA) and has been adapted for use and widely applied in helioseismology.

The idea of OLA is as follows. When comparing the model frequencies to the observed frequencies, there are differences, indicating that the structure (or rotation profile) of the model must differ from the structure of the Sun. If an oscillation mode were only sensitive to one region of the star, then a difference in frequency for that mode would indicate a difference in structure in that region. However, this is not the case: oscillation modes are sensitive to multiple locations in the solar interior, and so it is not possible to disentangle the cause of discrepancy based on only one mode.

The sensitivities of mode frequencies to perturbations in the structure of the star are called kernels. I provide the kernels of stellar structure in Section 1.3.2. The OLA method works by combining the modes in such a way that their combination—the averaging kernel—is only sensitive to one region in the star. When the combination of frequencies corresponding to that combination of modes differs between the model and the star, then the structure must differ in that region. Thus, one can then work out the structure in the locations in the interior where it is possible to construct an averaging kernel.

By the mid-80s, it became possible to invert frequency splittings and infer the internal rotation rate of the Sun (Duvall et al. 1984, see also e.g. Schou et al. 1998, Howe 2009). The following year, the internal solar sound speed profile was deduced via inversion of an asymptotic description known as Duvall's Law, which assumes that the mode frequencies depend exclusively on the speed of sound (Christensen-Dalsgaard et al. 1985). Soon thereafter, full inversions—which separate the influence on mode frequencies of, e.g., sound speed from density—were used to determine the acoustic structure of the majority of the solar interior (Gough 1985, see also e.g. Dziembowski et al. 1990, Gough and Thompson 1991, Gough and Toomre 1991, Antia and Basu 1994, Basu et al. 2009).

Inversions for helioseismic structure have revealed many aspects of the solar interior, such as the depth of the convection zone (e.g., Christensen-Dalsgaard et al. 1991, Basu and Antia 1997), the helium abundance in the solar envelope (e.g., Däppen et al. 1991, Basu 1998), the equation of state of the solar plasma (Basu and Christensen-Dalsgaard 1997), and the efficiency of element diffusion (Christensen-Dalsgaard et al. 1993). Rotation inversions have shown that the Sun rotates differentially, having a latitudinally-dependent rotation rate in the convective outer envelope, and rotating as a solid body in the radiative interior (e.g., Howe 2009). These zones are separated by a shear layer that is referred to as the tachocline (Spiegel and Zahn 1992). Finally, investigations based on helioseismic inversions have been instrumental in resolving longstanding issues such as the solar neutrino problem (e.g., Bahcall et al. 1998), for which four Nobel prizes have been awarded. A detailed review of results that have been obtained via helioseismic inversion has been given by Basu (2016).

1.1.2 Asteroseismology

As our Sun is not thought of as being particularly exceptional, it was obviously expected that other stars similar to the Sun should also exhibit solar-like oscillations (e.g., Christensen-Dalsgaard 1984). In addition to oscillations in solar-like stars, Christensen-Dalsgaard and Frandsen (1983) further predicted that low-mass giant stars should harbor these kinds of oscillations as well, as these stars also have convective envelopes. Moreover, these stars harbor *mixed modes*: modes that behave like acoustic oscillations in the envelope and gravity mode oscillations in the core (e.g., Dziembowski et al. 2001). However, due to the very small amplitudes of the solar oscillations (on the order of 10 cm/s, recall Figure 1.8), their discovery in other stars posed a long-standing challenge.

Already in the late 1980s detections of solar-like oscillations were being claimed (Gelly et al. 1986). These were not however confirmed in follow-up studies (e.g., Innis et al. 1991). Throughout the 1990s there were more claims of detections in other stars, which mainly served to place upper limits on their amplitudes (e.g., Brown and Gilliland 1990, Brown et al. 1991, Pottasch et al. 1992, Edmonds and Cram 1995). Finally, in the 2000s, firm detections of solar-like oscillations in other stars were made, such as in the nearest star, the solar-type star Alpha Centauri (Bouchy and Carrier 2001); the subgiant star β Hyi (Bedding et al. 2001); and the giant stars α Uma (Buzasi et al. 2000) and η Hya (Frandsen et al. 2002). The field of solar-like asteroseismology was born, but in its infancy. With the coming space missions, it would soon undergo a revolution.

The first space-based observations came from the NASA *Wide-Field Infrared Explorer* (WIRE), which had failed in its nominal mission, but was fortunately able to be repurposed into an asteroseismology mission (Buzasi 2000). After one month of observation, space photometry yielded solar-like oscillations in the very bright giant star Alpha Ursae Majoris (Buzasi et al. 2000), and soon thereafter, in Alpha Centauri as well (Schou and Buzasi 2001).

The first purposefully dedicated space asteroseismology mission was the Canadian *Microvariability and Oscillations of STars* telescope (MOST, Walker et al. 2003, duration 2003–2014). Though MOST was not sensitive enough to detect oscillations in solar-type stars, Barban et al. (2006, 2007) did detect radial-mode oscillations in the red giant ϵ Oph using 28 days of MOST observations. Studying this same star from the ground, Hekker et al. (2006) was able to detect non-radial pulsations. The detection of solar-like oscillations in red giants represents a great confirmation of stellar theory. A detailed review on oscillations in red giants has been given by Hekker and Christensen-Dalsgaard (2017).

Soon afterwards came the European/French space mission *Convection, Rotation and planetary Transits* (CoRoT, Baglin et al. 2006, duration 2006–2012), which was able to detect solar-like oscillations in solar-type stars (e.g., Deheuvels et al. 2010). Among other successes, CoRoT was particularly valuable for the study of solar-like oscillations in red giant stars, where oscillations in hundreds of these stars were detected (e.g., De Ridder et al. 2009, Hekker et al. 2009).

Kepler

By far the best asteroseismology mission to date has been the *Kepler* space observatory (Koch et al. 2010, duration 2009–2013). The data yield from *Kepler* has been enormous; here I will largely restrict discussion to solar-type stars which are relevant for this thesis. For detailed reviews and textbooks on asteroseismology, see e.g. Aerts et al. 2010, Christensen-Dalsgaard 2012, Hekker 2013, Chaplin and Miglio 2013, and Basu and Chaplin 2017.

Kepler targeted approximately 150,000 main sequence stars in a fixed field of view around the constellations of Cygnus, Lyra and Draco. Short-cadence and long-cadence targets were observed every 58.89 seconds and every 29.4 minutes, respectively. Several pipelines were created in preparation of processing the expected asteroseismic yield. For example, several groups created pipelines for the automated retrieval of $\Delta\nu$ and ν_{max} from *Kepler* time series (e.g., Huber et al. 2009, Mosser and Appourchaux 2009, Hekker et al. 2010, Mathur et al. 2010). For detailed stellar modelling, Metcalfe et al. (2009) created the Asteroseismic Modelling Portal (AMP), which fits evolutionary models to the observed asteroseismic data using genetic programming. In a hare-and-hound exercise, Stello et al. (2009b) found that the radius determinations from the expected asteroseismic data from *Kepler* are five to ten times better than without.

After launch, the quality of *Kepler* data for asteroseismology was immediately evident, revealing clear signatures of non-radial oscillations in several stars within one month of data collection (Gilliland et al. 2010, Chaplin et al. 2010). For the majority of stars, only the global properties such as $\Delta\nu$ and ν_{max} are able to be resolved. Even with just these quantities, however, it is possible to infer information about the stars. For example, by assumption of homology with the Sun, one can scale oscillation data from solar values to estimate the properties of stars, such as their masses and radii (e.g., Kjeldsen and Bedding 1995). This presents the opportunity for "ensemble asteroseismology." Chaplin et al. (2011, 2014) and Serenelli et al. (2017) used these and other approaches to find the masses, ages, radii, and other fundamental parameters for hundreds of main sequence and subgiant stars observed by *Kepler*. In addition, several groups have also worked on improvements to the solar scaling relations (e.g., Mosser et al. 2013, Sharma et al. 2016, Guggenberger et al. 2016, 2017, Viani et al. 2017).

For the best targets, interferometric and spectroscopic measurements have been obtained to complement the asteroseismic data (e.g., Bruntt et al. 2010, 2012, Mathur et al. 2012, White et al. 2013). These measurements provide the tightest determinations of stellar parameters and the best tests to stellar theory. Comparing these data, Huber et al. (2012) found good agreement between radii determined via interferometry and asteroseismology.

The perhaps best solar-like stars observed by *Kepler* are the solar analogs 16 Cygni A and B. These stars form a hierarchical triple system, with 16 Cyg A being orbited by a red dwarf, and 16 Cyg B being orbited by a Jovian planet. Metcalfe et al. (2012) "peak bagged" these stars (i.e., resolved their frequencies) and found clear detections of $\ell \leqslant 3$ modes (see Figure 1.10). They used AMP to

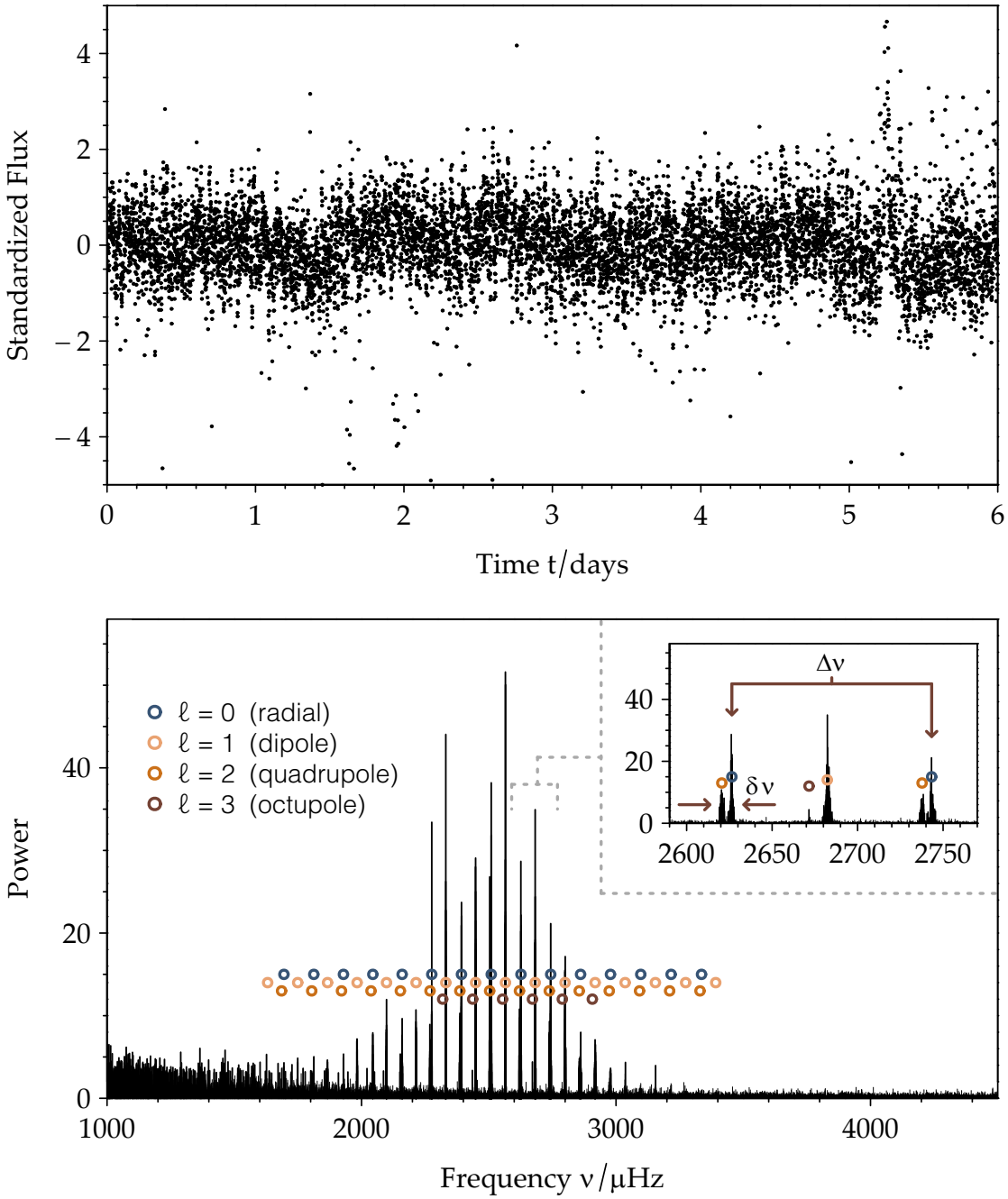


FIGURE 1.10. Light curve (top) and power spectrum of 16 Cyg B (bottom) as obtained from the *Kepler* spacecraft. The power spectrum shows 56 detected oscillation modes, each labelled by their spherical degree (*cf.* Figure 1.2). The power excess is roughly Gaussian in shape and centered around a value of $\nu_{max} \simeq 2550$ μHz. The inset figure shows a zoom into the power spectrum with example large ($\Delta\nu \simeq 117$ μHz) and small ($\delta\nu \simeq 6$ μHz) frequency separations. *Data from the Kepler Asteroseismic Science Operations Center (KASOC 2018).*

determine the evolutionary parameters of these stars, finding a common age of 6.8 Gyr and common initial chemical compositions, which supports the conatality hypothesis of binary star formation. Davies et al. (2015) used rotational splittings of the non-radial modes to infer the inclination angles and rotation rates of these stars, in both cases finding a rotation rate of approximately 23 days.

For approximately 100 solar-like stars observed by *Kepler*, the data have been good enough for dozens of individual mode frequencies to be resolved. These stars form the *Kepler* Ages (Davies et al. 2016) and *Kepler* LEGACY projects (Lund et al. 2017), the former of which comprises 35 planet-host candidates. Silva Aguirre et al. (2015, 2017) determined the fundamental parameters of these stars using pipelines created by different groups, finding roughly broad agreement. Verma et al. (2014b, 2017) used seismic glitch analysis to determine the base of the convection zone and helium abundances for the LEGACY sample. These are the stars analyzed in the coming Sections and Chapters.

A discussion of the *Kepler* mission would be incomplete without a mention of exoplanets. *Kepler* was primarily a plunt-hunting mission, and a very successful one. Within *Kepler* data researchers found a plethora of rocky planets, super Earths, and gas giants (e.g., Pál et al. 2008, Batalha et al. 2011, Borucki et al. 2012, Marcy et al. 2014). Additionally, *Kepler* data were used to find that hot Jupiters are common (Pál et al. 2008), and that many stellar-planetary systems are misaligned (Huber et al. 2013), bringing into question theories of planet formation. Of course, asteroseismology is of great aid to the characterization of exoplanets, since the determination of exoplanetary parameters usually depends strongly on the ability to determine the parameters of the host star (see Figure 1.11).

Following the failure of its reaction wheels, *Kepler* was repurposed into the wandering K2 mission, which is now in its final stages (Howell et al. 2014, duration 2013–2018). This year, NASA's *Transiting Exoplanet Survey Satellite* mission will launch (TESS, Ricker et al. 2010, expected 2018–2020). ESA's *Planetary Transits and Oscillations of stars* mission (PLATO, Rauer et al. 2014, expected 2026–2030) is planned for launch in eight years. We analyze the anticipated yields of these missions for Sun-like stars in Chapter 3 (Angelou et al. 2017).

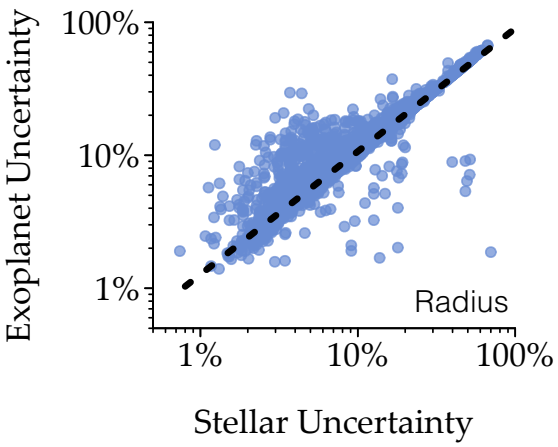


FIGURE 1.11. Uncertainty in the determination of exoplanetary radii as a function of the uncertainty in the determination of the radius of their host star for nearly 2,400 exoplanets detected using the transit method, which will also be the method of choice for finding exoplanets in the forthcoming TESS mission. *Data acquired from exoplanets.org (Han et al. 2014).*

Asteroseismic Inversions

Asteroseismic structure inversions are more difficult to perform than in helioseismology for two main reasons.

Mode set. The mode sets available in asteroseismology are much more limited. Due to cancellation effects, only low-degree modes have been observed so far in stars other than the Sun, and so only dozens rather than thousands of mode frequencies are available. It is only possible to build well-localized averaging kernels in locations where there is a sufficient number of mode lower turning points, as these are the regions where the modes spend most of their time (recall Figure 1.7). Consequently, asteroseismic inversions using only low-degree modes are generally only capable of making localized probes of the stellar core. This limitation also rules out the possibility of using techniques such as Regularized Least Squares, which fit the entire internal profile simultaneously (see, e.g., Basu and Chaplin 2017).

Furthermore, mode frequencies depend on multiple variables of stellar structure. When trying to determine one from asteroseismic information, one must therefore control for other influences. With limited information, this becomes more difficult. In helioseismology, the most common pair of variables is the speed of sound c and the stellar density ρ, denoted the (c, ρ) kernel pair.

Mass and radius. The masses and radii of stars are not known to anywhere near the precision for the Sun. This creates difficulties because the kernel functions are derived with respect to a reference model, which is assumed to have the correct mass and radius. Without accounting for this effect, the results of the inversion results will be offset by the differences in mass and volume (Basu 2003). Furthermore, the mode frequencies themselves scale with the mass and volume of the star.

Already in the early 1990s, before the first confirmed asteroseismic detections, Gough and Kosovichev (1993) considered the prospect of performing asteroseismic inversions to determine stellar structure. In this work, Gough and Kosovichev simulated data sets for a 1.1 solar mass model that they thought might be likely to be obtained from a future mission. They used a solar model as reference. Their work was on the one hand pessimistic—assuming only $\ell \leqslant 2$ modes would be available, having mode uncertainties of 0.1 μHz—and on the other optimistic, assuming that more than 60 modes would be observed. In comparison, the perhaps best *Kepler* solar-type target, 16 Cyg B, has approximately 56 detected modes (though the exact amounts are disputed), 11 of which being $\ell = 3$ modes, with uncertainties ranging from 0.04 μHz up to 5 μHz.

Gough and Kosovichev were able to form four well-localized averaging kernels at target radii 0.05, 0.15, 0.25, and 0.35. They simultaneously estimated the

difference in mass per volume between the two models while performing the inversion. Surface effects were not considered.

In this work it was already realized that inversions with helium as the second variable could be the most promising route. The helium kernels only have amplitude in ionization zones, which are located near to the stellar surface and would require higher-degree modes to resolve anyway. Basu et al. (2001) showed that when using the (c, ρ) kernel pair with expected asteroseismic data, only one averaging kernel can be formed.

Some other early attempts with similar setups and results have been reviewed by Basu (2003). These works all used mode sets that they thought would be available from future missions: PRISMA, MOST, MONS, and *Eddington*. Unfortunately, PRISMA, MONS, and *Eddington* were not funded, and MOST did not detect any oscillations in solar-like stars. It is only now with the CoRoT and *Kepler* missions that the data are good enough to measure internal stellar structure. We invert *Kepler* data to infer the internal structure of 16 Cyg A and B in Chapter 4 (Bellinger et al. 2017b).

Several other kinds of inverse problems have been worked on using asteroseismic data. Instead of inverting for the full density profile, Reese et al. (2012) introduced an OLA-based technique for estimating stellar mean density. They applied the technique to the Sun, α Cen B, and two stars observed by CoRoT. They found that they could estimate mean densities this way to an accuracy of 0.5%. However, it performed no better than estimating mean densities using the Kjeldsen et al. (2008) surface term corrected solar scaling relation.

Buldgen et al. (2015a,b) extended this work by creating kernels for the acoustic radius and two age indicators: the integral of the sound speed derivative, and a weighted square of the isothermal sound speed derivative. They applied these techniques to 16 Cyg A and B, and, when combining them with interferometric radii, found masses and ages for these stars that were inconsistent with evolutionary modelling (Buldgen et al. 2016a,b).

In addition to the global properties of stars, inversions for stellar rotation rates have also had success. Deheuvels et al. (2012, 2014), Di Mauro et al. (2016), and Triana et al. (2017) inverted frequency splittings to obtain the core and envelope rotation rates of several sub- and red-giant stars. They found, in agreement with theoretical expectations, that the cores of these stars rotate more rapidly than their outer layers.

Layout of Thesis

In this section, we saw that the study of pulsating stars has been a primary driver in the development of the theory of stellar evolution. Helioseismic inversions have revealed the structure of the Sun and shown that it is very close (though not identical to) the structure predicted by theoretical models. Asteroseismology has confirmed many details predicted by stellar evolution, and asteroseismic inversions show promise for leading to future improvements to evolutionary theory.

For the interested reader, the following texts contain more details: Ledoux and Walraven (1958) give a thorough overview of variable stars up until the 1950s; Arny (1990) gives the history of stellar evolution, including later phases of evolution which are not covered here; Basu (2016) gives the history of solar oscillations; Bolt et al. (2007) contains an encyclopedia of biographies for astronomers; and Catelan and Smith (2015) give a general overview and history of variable stars.

The remainder of the thesis is organized as follows. The following two sections (1.2, 1.3) give the theoretical background on stellar structure, evolution, and pulsation. These enable us to pose and solve the forward problems of simulating the evolution of a star and calculating its frequencies of oscillation. In Section 1.3, I furthermore state the kernel functions of stellar structure, which allow us to calculate the differences in mode frequencies between a pair of stellar models of differing structure. In the final section of the introduction (Section 1.4), I state more formally the inverse problems of asteroseismology that are considered in this thesis, and give some indication of their difficulty.

In Chapter 2, we perform evolution inversions to determine stellar ages and other fundamental parameters using machine learning (Bellinger et al. 2016). In Chapter 3, we use unsupervised machine learning to determine which observations are useful for constraining which properties of stellar models (Angelou et al. 2017). In Chapter 4, we determine the asteroseismic structure of two stars, in one case finding agreement with evolutionary modelling, but in another not (Bellinger et al. 2017b). Finally, at the end I give what I assess to be the future prospects for this line of research.

1.2 Stellar Structure & Evolution

In this section, I will provide a summary of background information on the theory of stellar structure and evolution, with a focus toward the creation of evolutionary models of solar-like stars. This will allow us to state the evolution inverse problem: i.e., given observations of a star, to determine its age and evolutionary history. Stellar evolution is a well-established field with a rich history and many seminal works on the topic. Textbooks overviewing the underpinnings of stellar structure and evolution are numerous and include works by Eddington (1926), Chandrasekhar (1939), Schwarzschild (1958), Collins (1989), Kippenhahn and Weigert (1990), Hansen and Kawaler (1994), Salaris and Cassisi (2005), Pols (2011), Kippenhahn et al. (2012), and Brown (2015). The following makes heavy use of these works, along with calculations using the stellar evolution code *Modules for Experiments in Stellar Astrophysics* (MESA, Paxton et al. 2011, 2013, 2015, 2018).

Positing that a star begins as an initially homogeneous cloud of mostly hydrogen that collapses under its own weight until the conditions are ripe for fusion to sustain it, stellar evolution is the collection of physical processes that cause the star to vary over time from this state. Reposition in terms of luminosity, radius, density, and color—diagnostics that are visible from the stellar surface—are then predicted from the ensemble of processes that cause the star to transform.

Many such processes are known. Nuclear fusion causes adjustment to the elemental abundances in the core or within shells inside the star. Gravitational settling causes heavier elements to sink inward, and radiative levitation selectively resists this sinking. Convection induces chemical mixing, which leads to chemical discontinuities when the boundaries of convective zones recede, and dredge-up events when an enveloping convective zone deepens into an area of disparate composition. Stars rotate, and this similarly causes material to mix. Magnetic fields, binary accretion, thermohaline mixing, and other processes may affect the evolution of stars as well.

This collection of processes—of which only a subset is "canonically" employed in stellar modelling—has been very successful at explaining both the occupations of stars in the Hertzsprung-Russell and Color-Magnitude diagrams, and in predicting the pulsations of stars as well. Asteroseismic theory, visited in detail in the section following this one, is capable of determining the character of the stellar oscillations during each stage in a star's life, as well as predicting their corresponding periods. For solar-type stars, these predictions are within seconds of their measured values.

Assumptions

The standard theory describing the evolution of a single star follows from a number of basic assumptions:

1. *Stars can be treated as a fluid.* I make this assumption so that we may describe stars using the equations of fluid dynamics rather than considering the motions of individual particles. The fluid approximation is likely a good description for the majority of the stellar interior, but it breaks down above the stellar photosphere.

2. *Stars are isolated in space.* I ignore companions and, consequently, the effects of mass transfer and tidal interactions.

3. *Stars are spherically symmetric.* I will describe the structure of a star from its core to its surface using only one coordinate (e.g. radius, but in practice, some quantity that varies monotonically with radius). I ignore rotation, which would distort the star. While all stars rotate, many (such as the Sun) rotate slowly enough that the effects of rotation on their structure can be considered negligible.

4. *Stars are self-gravitating.* I include the effects of a gravitational field, but I ignore electric and magnetic fields.

5. *Stars are dynamically stable.* Clearly, stars are pulsating; that is the main subject of this thesis. However, the pulsation timescale is usually much shorter than the evolutionary timescale. These will be treated in detail in the next section.

6. *Stars keep their mass.* Stars are observed to lose their mass through, for example, stellar winds. However, isolated main-sequence stars lose very little mass. For example, the Sun loses only about one part in 10^{13} of its mass each year (Krasinsky and Brumberg 2004).

From these assumptions, we may now formulate equations for the structure of a star.

Stellar Structure

By the structure of a star, I mean the *mechanical* (density ρ, pressure P), *thermal* (temperature T, adiabatic exponents Γ), and *chemical* (relative abundances of hydrogen X, helium Y, and heavy elements Z obeying $X + Y + Z = 1$) profiles from the core to the 'surface.' The equations of stellar structure consist of three conservation equations—conservation of mass, momentum, and energy—and the temperature equation. These macrophysical equations are supplemented with 'microphysics,' numerical inputs for necessary ingredients such as nuclear reaction rates. I will present most of the equations of stellar structure essentially without derivation. In order to give the reader an idea of the arguments used, however, I will provide derivations based on geometry and basic physics for the conservation of mass and the conservation of (linear) momentum.

It is natural to consider these quantities spatially (i.e., in one dimension, by the stellar radius). However, the radius of a star changes considerably over its

lifetime, growing from a dwarf to a giant and then becoming a dwarf again. On the other hand, the mass of a star, at least in the main-sequence phase, is very stable. The Sun, for example, loses only 10^{-14} of its mass per year through fusion and the solar wind. Therefore, I will here cast the equations using mass as the independent variable. In practice, stellar evolution codes often use a more complex variable which is even more stable than mass.

Conservation of Mass. Geometrically speaking, the mass m contained within a sphere spanning from the centerpoint ($r = 0$) to a radius of r is given by

$$m(r) = \int_0^r 4\pi x^2 \rho(x)\, dx. \tag{1.1}$$

Differentiating this equation, and dropping arguments, we arrive at

$$\boxed{\frac{dr}{dm} = \frac{1}{4\pi r^2 \rho}} \tag{1.2}$$

which, as we will see, is the continuity equation in the absence of flows. □

Conservation of Momentum. The state of balance between gravity and a pressure-gradient force is called hydrostatic support (also known as hydrostatic equilibrium or hydrostatic balance) and is a special case of conservation of momentum. The equation can be derived from either Newton's laws of motion, the Navier–Stokes equations, or from general relativity. Here I show the former.

Consider a small fluid parcel inside of the star whose base is located at radius r having height dr and a constant area A. The parcel has three forces acting upon it: downward and upward forces from pressure, and a downward force from gravity. The upward force on the parcel is

$$F_{\text{upward}}(r) = A \cdot \underbrace{P(r)}_{\text{pressure below}} \tag{1.3}$$

and the combined downward force is

$$F_{\text{downward}}(r) = -\left(A \cdot \underbrace{P(r+dr)}_{\text{pressure above}} + \underbrace{A \cdot \rho(r) g(r) \cdot dr}_{\text{gravity}} \right). \tag{1.4}$$

When these forces are balanced, i.e. $F_{\text{upward}} = F_{\text{downward}}$, the parcel is said to be in hydrostatic equilibrium. Thus, we have

$$0 = -A \left(\underbrace{\overbrace{P(r)}^{dP} - P(r+dr)}_{F_{\text{downward}}} - \rho(r) g(r) \cdot dr \right) \tag{1.5}$$

which then gives us

$$\frac{dP}{dr} = -\rho g. \tag{1.6}$$

We may then apply the equation of conservation of mass (1.2) and obtain

$$\boxed{\frac{dP}{dm} = -\frac{Gm}{4\pi r^4}} \tag{1.7}$$

where $G = 6.67408 \times 10^{-8}\ g^{-1}\, cm^3\, s^{-2}$ is the gravitational constant. □

These two conservation equations give us the mechanical structure of the star—the pressure and density throughout the stellar interior. Assuming a constant temperature, the ratio of these quantities gives us the speed at which acoustic waves propagate in the star:

$$u = P/\rho. \tag{1.8}$$

This quantity is known as the *squared isothermal speed of sound* and will be important in the following investigations.

Conservation of Energy. The flow of energy l throughout the stellar interior is given by

$$\boxed{\frac{dl}{dm} = \epsilon_{nuc} - \epsilon_\nu + \epsilon_g} \tag{1.9}$$

where ϵ_{nuc} is the energy generated by nuclear reactions, ϵ_ν is the energy lost by neutrinos, and ϵ_g is the gravitational energy from expansion or compression:

$$\epsilon_g = -T\frac{\partial s}{\partial t} \tag{1.10}$$

where s is the specific entropy. The nuclear energy generation rates are supplied externally. Here I use the rates from the *Nuclear Astrophysics Compilation of Reaction Rates* (NACRE, Angulo et al. 1999). The neutrino energy loss rates can be calculated using the formulas given by Itoh et al. (1996).

Computing the ϵ_g term requires an equation of state (EOS). This too is supplied externally. For the low-mass stars considered here, I use the *Opacity Project at Livermore* EOS (OPAL, Rogers and Nayfonov 2002). The EOS relates the pressure, density, and temperature of the stellar matter to each other in a thermodynamically-consistent manner. The adiabatic exponents Γ, introduced by Chandrasekhar, give these relations as follows:

$$\Gamma_1 = \left(\frac{\partial \ln P}{\partial \ln \rho}\right)_{ad} \tag{1.11}$$

$$\frac{\Gamma_2}{\Gamma_2 - 1} = \left(\frac{\partial \ln P}{\partial \ln T}\right)_{ad} = \frac{1}{\nabla_{ad}} \tag{1.12}$$

$$\Gamma_3 - 1 = \left(\frac{\partial \ln T}{\partial \ln \rho}\right)_{ad} \tag{1.13}$$

which are related to each other as:

$$\frac{\Gamma_1}{\Gamma_3 - 1} = \frac{\Gamma_2}{\Gamma_2 - 1}. \tag{1.14}$$

The first adiabatic exponent describes how the compression of a layer changes the pressure in that layer, which, as we will see, is important for determining dynamical stability, i.e., stellar pulsations. In particular, in an anisotropic ideal gas, the speed at which acoustic waves propagate—the adiabatic speed of sound[7]—can be defined as

$$c = \sqrt{\Gamma_1 u}. \tag{1.15}$$

The second adiabatic exponent describes how changes in pressure impact upon the temperature, which is important for determining stability against convection. In an ideal monoatomic gas, the adiabatic exponents all equal 5/3.

Temperature Equation. The temperature throughout the star is given by

$$\boxed{\frac{dT}{dm} = -\frac{Gm}{4\pi r^4}\frac{T}{P}\nabla_T} \tag{1.16}$$

where ∇_T is a dimensionless temperature gradient:

$$\nabla_T = \frac{d\ln T}{d\ln P} \tag{1.17}$$

whose form depends on the mode of energy transport. In the case of pure radiation,

$$\nabla_T = \nabla_{rad} = \frac{3}{64\pi\sigma G}\frac{\kappa l P}{m T^4}. \tag{1.18}$$

where $\sigma = 5.670367 \cdot 10^{-5}$ erg cm^{-2} s^{-1} K^{-4} is the Stefan-Boltzmann constant and κ is the opacity of the stellar matter, which is also supplied externally. Here I use the OPAL opacities (Iglesias and Rogers 1996).

The conductive temperature gradient is negligible for our purposes, though it is relevant e.g. in white dwarfs. The convective temperature gradient comes from both the adiabatic gradient of the assumed EOS (*cf.* Equation 1.12) and the specific treatment of convection, which I will discuss later in this section.

We thus have four coupled differential equations (1.2, 1.7, 1.9, 1.16) that govern stellar structure. In order to solve them, we will need four boundary conditions.

[7] Not to be confused with the speed of light.

Boundary Conditions

The first boundary is at the central point in the star, where $m = 0$. Here we have

$$m = 0, \qquad r = 0, \qquad l = 0. \tag{1.19}$$

The second boundary is at the stellar surface. This is where the mass equals the total mass, $m = M$; and where the radius equals the total radius, $r = R$. A simple option is to assume that the temperature and pressure vanish at the surface, i.e.

$$T(r = R) = 0, \qquad P(r = R) = 0. \tag{1.20}$$

These are known as *zero-boundary* conditions and we will make use of them later when calculating variational pulsation mode frequencies (see Section 1.3.1). They are unrealistic, however, as even the interstellar medium has a non-zero temperature.

A more sophisticated option is to call the surface the region where majority of the radiation escapes from the star, i.e., the photosphere. Here I will use a standard Eddington gray atmosphere, which gives the total luminosity and effective temperature

$$l(r = R) = L, \qquad T(r = R) = T_{\rm eff} \tag{1.21}$$

following the Stefan-Boltzmann Law for blackbody radiation:

$$L = 4\pi R^2 \sigma T_{\rm eff}^4 \tag{1.22}$$

where σ is again the Stefan-Boltzmann constant. Finally, the pressure at the surface is given by

$$P(r = R) = \frac{2}{3}\frac{GM}{R^2}\frac{1}{\bar{\kappa}} \tag{1.23}$$

where $\bar{\kappa}$ is the Rosseland mean opacity.

Stellar Evolution

For a star to *evolve*, it must change over time. With the exception of one term for the gravitational energy from expansion or compression (Equation 1.10), the equations of stellar structure feature no time derivatives; they describe a static star. The equations of stellar structure may be supplemented with time-dependent evolution equations describing the internal transport or modification of chemical species.

> **Nuclear reactions.** Energy generation on the main sequence stems predominately from the conversion of hydrogen atoms (H) into helium atoms (He). The net reaction is
>
> $$4\ {}^1\mathrm{H} \rightarrow {}^4\mathrm{He} + 2e^+ + 2\nu_e \tag{1.24}$$

where e^+ is a positron and ν_e is a neutrino. Earth-based detections of neutrinos matching the predicted solar output essentially confirm this description. The evolution due to nuclear reactions can be given as:

$$\boxed{\frac{\partial X_i}{\partial t} = \frac{m_i}{\rho}\left(\sum_j r_{ji} - \sum_k r_{ik}\right)} \tag{1.25}$$

where X_i is the i^{th} isotope, m_i is the mass of that isotope, and $r_{i,j}$ is the rate at which X_i is formed from X_j. As mentioned, these rates must be supplied externally; here I've chosen to use the NACRE rates.

Diffusion. The processes of element diffusion and the gravitational settling of helium and heavy elements can be included via the diffusion equation:

$$\boxed{\frac{\partial X_i}{\partial t} = D_i \frac{\partial^2 X_i}{\partial m^2}} \tag{1.26}$$

where D_i is the diffusion coefficient for isotope X_i. Diffusion coefficients must also be externally supplied; a common choice are those of Thoul et al. 1994.

Convection. According to the Schwarzschild criterion (e.g., Schwarzschild 1958), a region is unstable to convection when the radiative gradient exceeds the adiabatic gradient:

$$\nabla_{rad} > \nabla_{ad} \tag{1.27}$$

(*cf.* Equations 1.12 and 1.18). Here I will treat convection using the standard Böhm-Vitense (1958) mixing length theory, which approximates the effects of convection by assuming that convective elements travel to some characteristic length ℓ_m before mixing the transported material with their newfound surroundings. The mixing length is controlled by a free parameter α_{MLT}, which is scaled by the local pressure scale height:

$$\ell_m = \alpha_{MLT} \cdot H_p \tag{1.28}$$

$$H_p = -\left(\frac{d \ln P}{dr}\right)^{-1}. \tag{1.29}$$

There is no *a priori* choice for α_{MLT}. Generally, α_{MLT} is either fixed to a value that has been calibrated to the observed characteristics of the Sun, which we shall address later in this section; or fit on a star-by-star basis (Chapter 2).

Convection is an efficient mixer. We can model the changes to chemical abundances due to convection as a diffusion process:

$$\boxed{\frac{\partial X_i}{\partial t} = \frac{\partial}{\partial m}\left(D_{conv}\frac{\partial X_i}{\partial m}\right)} \tag{1.30}$$

where $D_{conv} \propto v_c \cdot \ell_m$, with v_c being the convective velocity.

Convective zones can be extended beyond their normal boundaries via convective overshooting. Overshooting is similarly controlled by a free parameter α_{ov}, which extends the boundary by $\alpha_{ov} \cdot H_p$. Like α_{MLT}, the overshooting parameter has no predefined value. While it is not uncommon to exclude the effects of overshooting altogether, α_{ov} can also be determined from a fit to a stellar population (e.g., Gallart et al. 2005) or on a star-by-star basis (Chapter 2).

Calculations generally proceed as follows. First, the equations of stellar structure are solved for a given composition. Then, time is advanced, and a new composition is computed using the evolution equations. The equations of stellar structure are then solved again for the new composition, and the procedure is repeated. Henyey et al. (1959) introduced an efficient scheme to solve these equations based on iterative application of the Newton-Raphson method. We will now solve these equations and model the evolution of the stars.

Solar Calibration

We may begin our calculations by calibrating an evolutionary track to the observed properties of the Sun (e.g. Christensen-Dalsgaard 1982) in accordance with the recommended nominal solar values adopted by the IAU (Mamajek et al. 2015). The standard gravitational parameter of the Sun $\mu_\odot$ is known to very high precision from planetary orbits:

$$\mu_\odot = GM_\odot = 1.3271244 \cdot 10^{26}\ \mathrm{cm^3\ s^{-2}}.$$

The gravitational constant may be determined experimentally; this then yields the solar mass. Next, the Earth-Sun distance as well as direct observations give the solar radius. Solar irradiance measurements give the solar luminosity. Spectroscopy gives the composition the solar photosphere; I use the mixture as measured by Grevesse and Sauval (1998, hereinafter GS98) which gives good agreement with helioseismology. Finally, radiometric dating of meteorites gives the age of the solar system. Putting this all together, the Sun has the following characteristics:

$$\begin{aligned} \text{mass } M_\odot &= 1.988475 \cdot 10^{33}\ \mathrm{g} \\ \text{radius } R_\odot &= 6.957 \cdot 10^{10}\ \mathrm{cm} \\ \text{luminosity } L_\odot &= 3.828 \cdot 10^{33}\ \mathrm{erg\ s^{-1}} \\ \text{effective temperature } T_{\mathrm{eff},\odot} &= 5772\ \mathrm{K} \\ \text{heavy mass fraction } (Z/X)_\odot &= 0.02293 \\ \text{age } \tau_\odot &= 4.572 \cdot 10^9\ \mathrm{yr}. \end{aligned} \tag{1.31}$$

These are the values that must be reproduced in our solar calibration. We will achieve this by altering the initial chemical composition and the efficiency of convective mixing (recall Equation 1.28) until these values are reproduced at

the solar age. Since the Sun is an isolated main-sequence star, its mass has been presumably very stable throughout its lifetime. The initial mass of the calibration can therefore remain fixed at the solar value. Finally, we only need to check that e.g. the luminosity and radius are matched, since R, L, and $T_{\rm eff}$ are related through the Stefan-Boltzmann Law (Equation 1.22).

We therefore have the following optimization problem: we wish to tune the initial helium abundance Y_0, initial metallicity Z_0, and mixing length parameter $\alpha_{\rm MLT}$ of a solar-mass track such that we minimize $\log_{10}(L/L_\odot)$, $\log_{10}(R/R_\odot)$, and [Fe/H] at the solar age, where [Fe/H] is defined as

$$[\mathrm{Fe/H}] \equiv \log_{10}\left(\frac{Z}{X}\right)_* - \log_{10}\left(\frac{Z}{X}\right)_\odot . \tag{1.32}$$

We may achieve solar calibration by, e.g., iterative application of Newton's rule:

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \mathbf{J}_t^{-1}\mathbf{f}(\mathbf{x}_t) \tag{1.33}$$

where (dropping the MLT and 0 subscripts)

$$\mathbf{x}_t = \left(Y_t, \quad Z_t, \quad \alpha_t\right) \tag{1.34}$$

$$\mathbf{f}(\mathbf{x}_t) = \left(\log_{10}\{L_t/L_\odot\}, \quad \log_{10}\{R_t/R_\odot\}, \quad [\mathrm{Fe/H}]_t\right) \tag{1.35}$$

$$\mathbf{J}_t = \begin{pmatrix} \dfrac{\partial \log_{10}\{L_t/L_\odot\}}{\partial Y} & \dfrac{\partial \log_{10}\{L_t/L_\odot\}}{\partial Z} & \dfrac{\partial \log_{10}\{L_t/L_\odot\}}{\partial \alpha} \\ \dfrac{\partial \log_{10}\{R_t/R_\odot\}}{\partial Y} & \dfrac{\partial \log_{10}\{R_t/R_\odot\}}{\partial Z} & \dfrac{\partial \log_{10}\{L_t/L_\odot\}}{\partial \alpha} \\ \dfrac{\partial [\mathrm{Fe/H}]_t}{\partial Y} & \dfrac{\partial [\mathrm{Fe/H}]_t}{\partial Z} & \dfrac{\partial [\mathrm{Fe/H}]_t}{\partial \alpha} \end{pmatrix}. \tag{1.36}$$

Here t refers to the $t^{\rm th}$ iteration, and the partial derivatives are to be calculated numerically (i.e. by running tracks with small changes to those parameters). It may also be prudent to enforce some box constraints, for example: $0.23 \leqslant Y_0 \leqslant 0.33$, $0 < Z_0 < 0.05$, $1 \leqslant \alpha_{\rm MLT} \leqslant 3$. When supplied with a reasonable initial guess, this scheme eventually converges onto a set of parameters that reproduce the observed solar values:

$$\begin{array}{lll} Y_0 \simeq 0.273 & & \log_{10}(L/L_\odot) \simeq 0 \\ Z_0 \simeq 0.019 & \Rightarrow & \log_{10}(R/R_\odot) \simeq 0 \\ \alpha_{\rm MLT} \simeq 1.84 & & [\mathrm{Fe/H}] \simeq 0. \end{array} \tag{1.37}$$

These initial values, as well as the observed values of the Sun (Equations 1.31) are the ones that will need to be reproduced when we later perform evolutionary inversions on degraded Sun-as-a-star data (see Chapter 2), where they are all either unknown or highly uncertain.

We may now inspect the structure of our solar model. Figure 1.12 shows some aspects of the mechanical, thermal, and chemical structure of the model.

Helioseismology has revealed that these profiles are exceptionally close to the actual interior of the Sun (see, e.g., Basu 2016).

A few points are worthy of note here. The first adiabatic exponent is close to 5/3 (i.e., nearly the conditions of an ideal gas) for the majority of the solar interior and only deviates from this value close to the solar surface. The helium abundance Y in the solar core is maximal due to nearly 5 Gyr of hydrogen-to-helium fusion. Helium is now the dominant element in the core, with the fractional hydrogen abundance being reduced to 0.344. Throughout the convection zone, which extends from $\sim 0.7\ r/R$ to the solar surface, the helium abundance has a constant value of 0.279 due to convective mixing. This value is somewhat higher than the protosolar value of 0.273 due to element diffusion.

The density ranges from around 150 g/cm^3 in the core to less than that of water in the outer half of the star, with the mean density of the Sun being about a hundredth of the core density. The pressure in the solar core falls off more rapidly than the density, which causes the speed of sound to rise temporarily when moving away from the centerpoint. Furthermore, since $u \propto T/\mu$, where T is the temperature and μ is the mean molecular weight, the speed of sound in the solar core is related to the age of the Sun via the increased abundance of helium.

We may additionally inspect the resulting evolutionary path of the solar-calibrated track. Figure 1.13 shows the past and future evolution of our Sun, assuming that the theory of stellar evolution is approximately correct; and Figure 1.14 shows the chemical evolution of the solar core. The Sun is currently on the main sequence; after several billion years, it will cross the sub-giant branch, climb the red giant branch (RGB), reach the tip of the RGB, and then fall onto the red clump (RC). The configurations of the star at these points in its evolution are shown in Figure 1.15.

Subsequent to these stages is the asymptotic giant branch (AGB, shell-helium & shell-hydrogen burning) phase, followed by the (misnomered) planetary nebula phase in which the outer layers of the Sun will be shed. The Earth and the terrestrial planets of the solar system will almost certainly be consumed or burnt to the point of inhabitability by this point. The Sun will then cool nearly indefinitely as a white dwarf—until, after trillions of years, it will finally settle as a black dwarf.

This is the fairly typical path of a low-mass star and looks roughly the same for stars of solar composition with masses $0.2 \lessapprox M/M_\odot \lessapprox 1.2$, with the amount of time taken through this sequence being inversely related to the stellar mass. Outside of this range, less massive stars are fully convective and so their evolution can be quite different. Even less massive objects ($M \lessapprox 0.1\ M_\odot$) never achieve hydrogen fusion, and as such, never enter the main sequence. More massive stars ($M/M_\odot \gtrapprox 1.2$) sustain a convective core on the main sequence, and exhibit a feature known as the Henyey hook when leaving it. Stars more massive than $\sim 2.2\ M_\odot$ do not undergo a helium flash on the red giant branch; instead, they gently begin helium burning. Finally, stars with a final mass (i.e., after the loss of

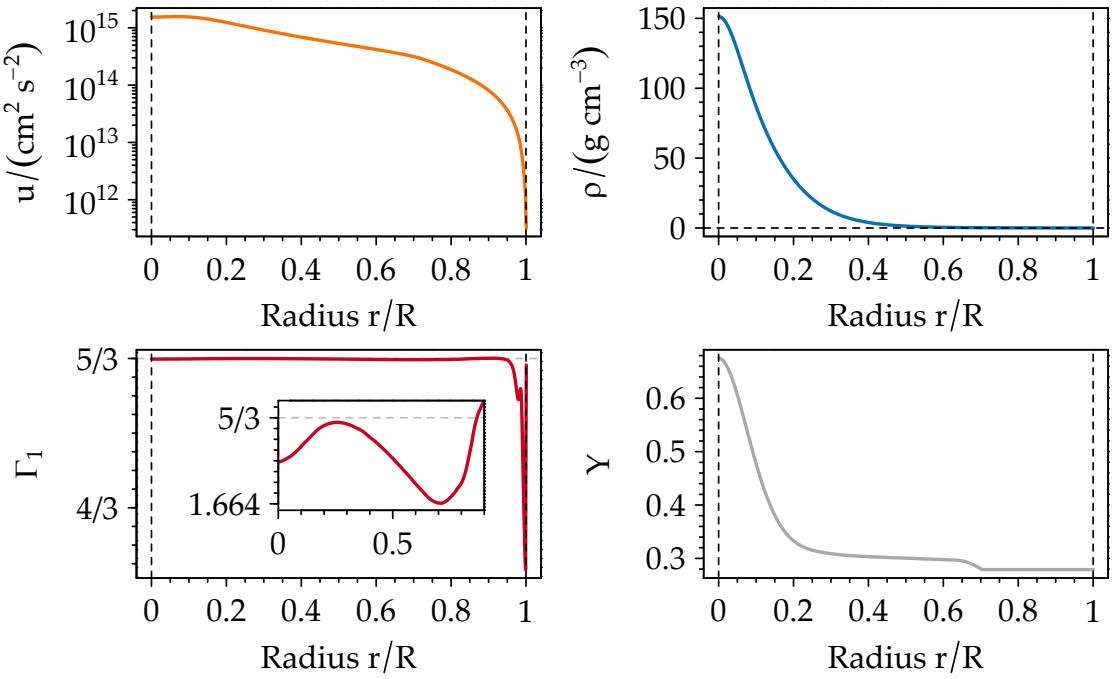


FIGURE 1.12. Squared isothermal sound speed (top left), density (top right), first adiabatic exponent (bottom left), and fractional helium abundance (bottom right) profiles for a solar model.

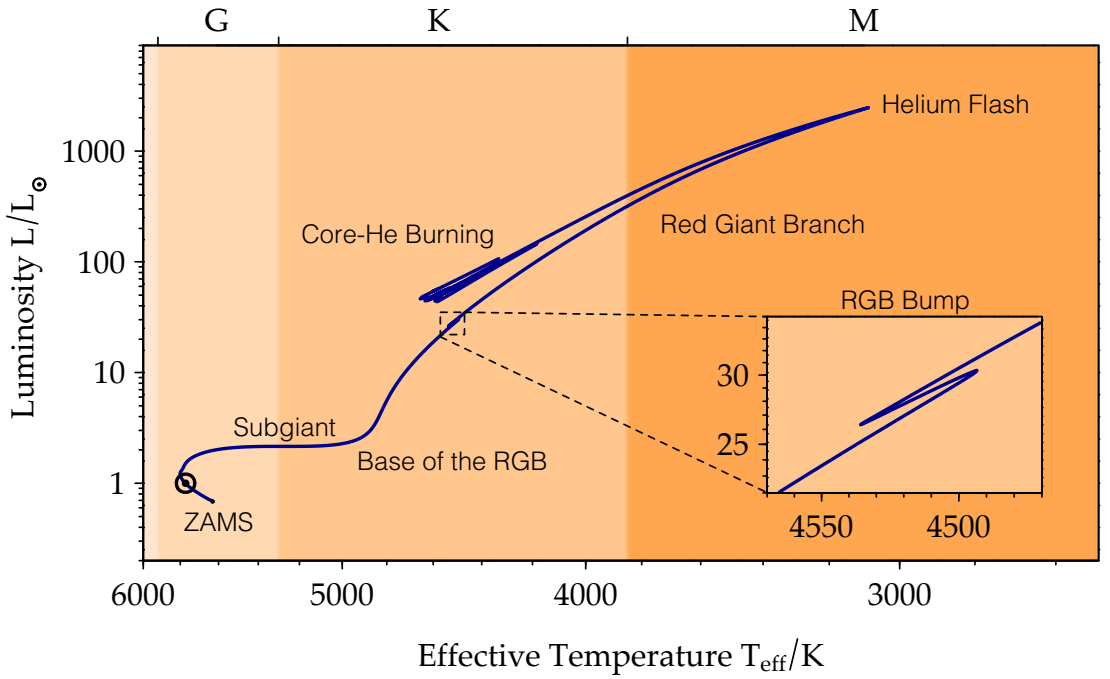


FIGURE 1.13. Hertzsprung-Russell diagram showing the evolution of the Sun. The background colors correspond to spectral type (F, G, K, M). The position of the Sun is indicated with the solar symbol (⊙).

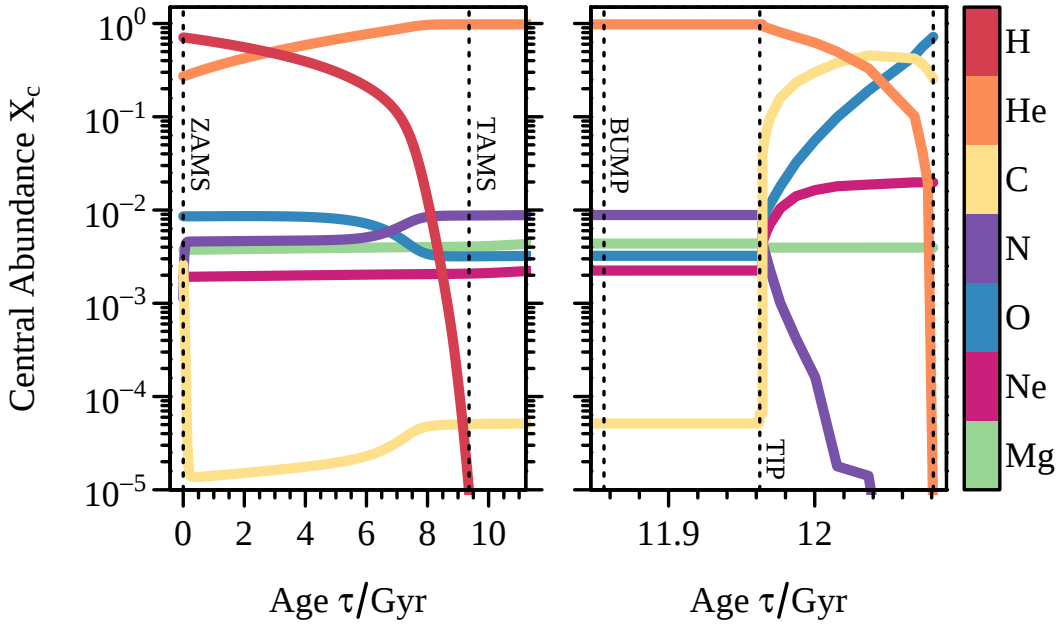


FIGURE 1.14. The past and future chemical evolution of the core of our Sun. The left panel shows the main sequence evolution, from the zero-age main sequence (ZAMS) to the terminal-age main sequence (TAMS). The right panel shows the evolution from the red giant luminosity bump through to the tip of the red giant branch and eventually to core-helium exhaustion. The core composition does not change throughout the majority of the subgiant and red giant phases.

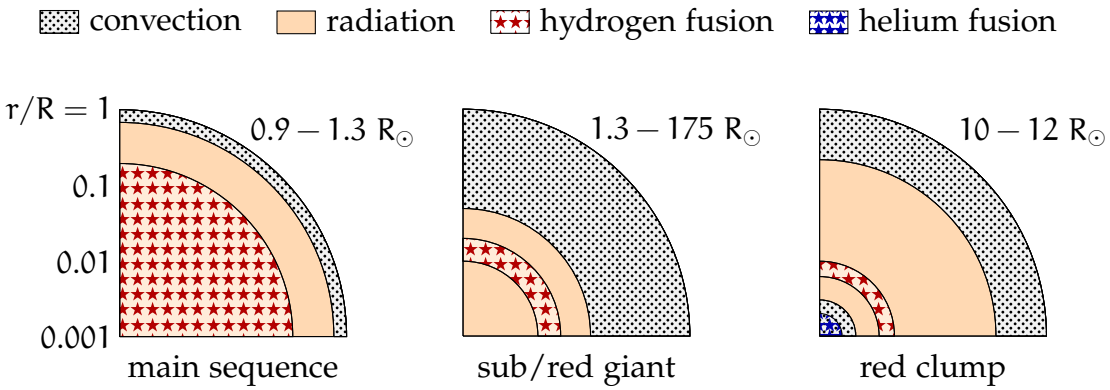


FIGURE 1.15. The configuration of the solar interior as the Sun evolves. The present Sun is a main-sequence star with a radiative core where hydrogen fusion is synthesizing helium. The outer ~ 30% of the Sun by radius transports energy by convection. When the Sun depletes its supply of core hydrogen in ~ 5 Gyr, it will continue burning hydrogen in a shell outside of the core. For the next ~ 2.5 Gyr, the inert helium core will contract while the convective envelope deepens as the Sun puffs up into a giant star. The Sun will then reach the tip of the red giant branch, where helium in the highly degenerate core will suddenly undergo a flash ignition. The Sun will subsequently become a red clump star, where it will continue fusing hydrogen in a radiative shell while simultaneously fusing helium in its convective core.

mass in the later stages of evolution) above about 1.44 $M_\odot$ (the Chandrasekhar limit) do not become white and black dwarfs; they rather explode in a supernova, enriching the interstellar medium with heavy mass elements. It is to these stars that we owe our astronomical heritage.

In this thesis, I am mainly focused on the study of stars in their first and longest-lived phase of evolution: the main sequence. Currently ongoing work is the application of these techniques developed herein to those later stages of evolution.

Evolutionary Paths

The last investigation of this section is focused toward gaining an intuition for what kinds of (in this case: low-mass, main sequence) stars are theoretically possible under the above assumptions. This is the forward problem of stellar evolution. Figure 1.16 shows evolutionary tracks for stars under non-solar conditions that I generated by varying the free parameters of stellar evolution from their solar-calibrated values, one at a time. Notice that adjustments to different parameters have similar impacts on the resulting evolution of the star. Thus it is very difficult, at least on the basis of the position in the H-R diagram, to determine the characteristics of a star. As we will see in Section 1.4, determining the evolutionary characteristics of a star from observations forms the first of the two inverse problems that are considered in this thesis.

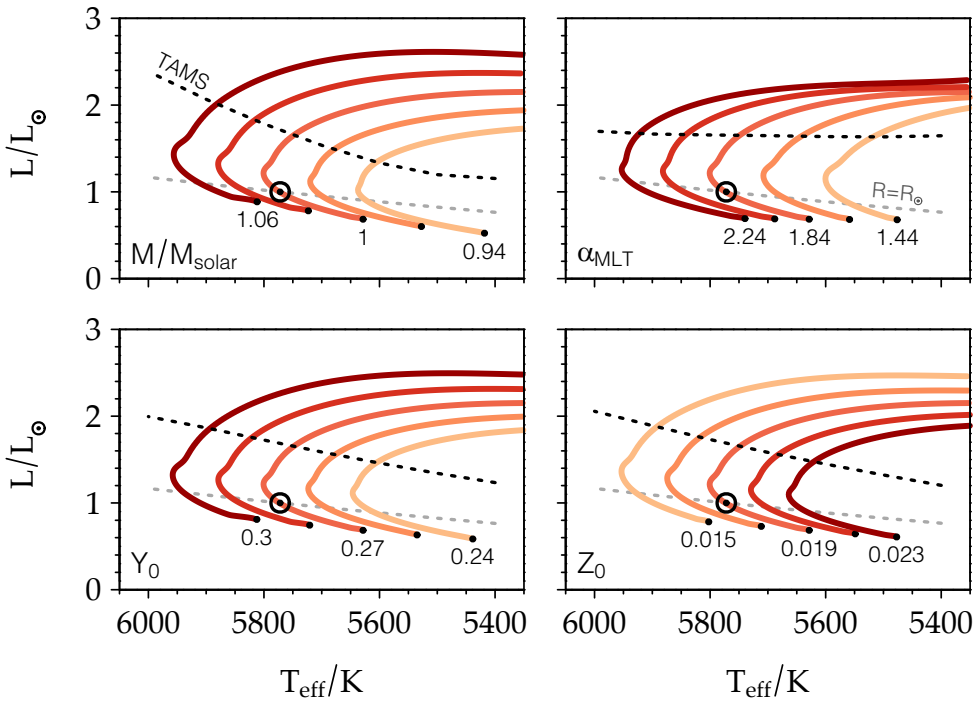


FIGURE 1.16. Theoretical Hertzsprung-Russell diagrams showing the main-sequence and sub-giant phases for evolutionary tracks varied in initial mass (top left), mixing length parameter (top right), initial helium abundance (bottom left) and initial metallicity (bottom right). Aside from the parameter being varied, the remaining parameters are kept fixed at the solar-calibrated values. For each track, ZAMS is marked with a black dot. The solar radius is indicated by the gray dotted line (recall Equation 1.22). Core-hydrogen exhaustion (TAMS, $X_c \sim 10^{-5}$) is indicated by the black dotted line. The color of the track darkens as the parameter under consideration increases. Notice that unlike the other parameters, an increase to the initial metallicity decreases the effective temperature. The H-R diagram is degenerate in that the sense that the same point can be reached by evolutionary tracks with different input parameters.

1.3 Theory of Stellar Pulsations

The purpose of this section is to give the reader a sufficient background summary on non-radial stellar pulsations in order to be able to understand the remainder of this thesis. I draw heavily here from the numerous textbooks that have been written on stellar pulsations, which include works by Eddington (1926), Rosseland (1949), Unno et al. (1979), Cox (1980), Aerts et al. (2010), and Basu and Chaplin (2017). Additionally, the long reviews by Ledoux and Walraven (1958), Gough (1993), and Basu (2016) were valuable references. I will perform calculations in this section using the *Aarhus adiabatic oscillation package* (ADIPLS, Christensen-Dalsgaard 2008).

Observations of stellar pulsations grant a new kind of insight into the behavior of stars. Whereas classical measurements of stars probe the stellar surface, observations of stellar pulsations, which traverse the stellar interior, bring deeper information to light. Measurements of stellar pulsations provide stringent tests on the processes of stellar evolution, as the frequencies of pulsation profoundly depend on the predicted stellar structure. Stars exhibiting solar-like oscillations are particularly valuable for this pursuit. These stars vibrate in a superposition of a great number of oscillation modes simultaneously, and each mode that can be observed provides additional information that can be used to constrain stellar models.

The pulsation hypothesis of stellar variability is supported by the fact that the theoretical pulsations of stellar models generally match the observed pulsations of stars. Furthermore, theoretically predicted pulsations in stars that were previously not observed to be variable (such as red giants) have been now overwhelmingly confirmed. That being said, while the agreement with models is very good, it is not perfect. In this section, I will outline the theory of stellar pulsations, thereby allowing us to calculate the time-independent adiabatic pulsation frequencies of our stellar models. I will compare the frequencies of my solar-calibrated model to measurements of the Sun. I will furthermore present the kernel functions of stellar structure, which quantify how changes to the stellar structure translate into changes in pulsation frequencies. This will allow me to state the structure inverse problem: i.e., the problem of determining a star's structure using only asteroseismic arguments.

Assumptions

I again begin with my assumptions. In addition to the assumptions for stellar structure, I assume:

1. *The stellar structure is nearly static.* I ignore all time derivatives (including velocities) in the equilibrium structure of the star. Thus, I am considering only time-independent pulsation frequencies. Clearly, stars evolve over

time—the entire preceding section was based on that fact. That said, the evolutionary timescale in the stars considered here (billions of years) is far greater than the pulsation timescale (minutes).

2. *The pulsations are linear perturbations the static stellar structure.* I ignore non-linear perturbations. This assumption should hold when the pulsation amplitudes are much smaller than the speed of sound. As we've seen, solar oscillations have amplitudes around 10 cm/s, whereas the speed of sound at the surface of the solar-calibrated model is on the order of 10 km/s.

3. *The pulsations are adiabatic.* I ignore the transfer of energy between the oscillations and the equilibrium stellar structure. This assumption should hold to good approximation when the pulsation time-scale is much smaller than the thermal timescale. With pulsation periods on the order of minutes, this is true for the majority of the stellar interior. However, this assumption too breaks down near to the stellar surface. Furthermore, without consideration of non-adiabatic effects, we will be unable to predict mode amplitudes, and we will not be able to determine whether the modes are excited (e.g., Samadi et al. 2015).

4. *The stellar material is inviscid.* I ignore internal friction. Although the viscosity of the solar core is similar to that of honey ($\sim 100\ \mathrm{cm}^2/\mathrm{s}$, e.g., Fox and Kerr 2000), the Reynolds numbers throughout the solar interior are large enough to justify this assumption. However, this assumption does break down in convection zones, where turbulent viscosity damps the oscillations.

Here and in the previous section I have made several assumptions that are violated in the near-surface layers of stars, or in locations where energy is transported by convection. These violations will cause errors in the predicted mode frequencies. I will introduce a correction to deal with these errors later in the section.

Fluid Dynamics

Given a static stellar structure, we consider a small perturbation that displaces all quantities (density, pressure, etc.) from equilibrium. For example, the stellar density at position $\vec{r}$ and time t is

$$\text{(Eulerian perturbation)} \qquad \rho(\vec{r}, t) = \rho_0(\vec{r}) + \rho'(\vec{r}, t) \tag{1.38}$$

$$\text{(Lagrangian perturbation)} \qquad \delta\rho(\vec{r}) = \rho'(\vec{r}_0) + \vec{\xi} \cdot \nabla\rho_0(\vec{r}) \tag{1.39}$$

where ρ_0 is the equilibrium density, ρ' is the perturbed density, and $\vec{\xi} \equiv \vec{r} - \vec{r}_0$ is the displacement in space. Here I have made use of the assumption that the equilibrium structure does not depend on time. The perturbation induces a velocity field $\vec{v}$ given by

$$\vec{v}(\vec{r}, t) = \frac{\partial}{\partial t}\vec{\xi}(\vec{r}, t). \tag{1.40}$$

This velocity field is then controlled by the following equations:

The continuity equation. As we've seen previously, the equation of continuity is a statement of mass conservation (*cf.* Equation 1.2). It states that mass cannot teleport through the star, but rather must travel through it continuously. The equation can be given as

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \vec{v}) = 0 \tag{1.41}$$

where $\nabla\cdot$ is the divergence vector operator. Substituting the perturbed quantities (Equation 1.38) into Equation 1.41, we get

$$\frac{\partial}{\partial t}\left[\rho_0(\vec{r}) + \rho'(\vec{r}, t)\right] + \nabla \cdot \left\{ \left[\rho_0(\vec{r}) + \rho'(\vec{r}, t)\right] \frac{\partial \vec{\xi}}{\partial t} \right\} = 0. \tag{1.42}$$

As we have assumed the equilibrium structure to be static, the corresponding time derivatives vanish. Integrating with respect to time, we then obtain

$$\boxed{\rho' + \nabla \cdot \left(\rho_0 \vec{\xi}\right) = 0} \tag{1.43}$$

i.e., the perturbed equation of continuity. □

The equation of motion. To first order, the general Navier–Stokes momentum equation can be expressed as

$$\rho \left(\frac{\partial}{\partial t} + \vec{v} \cdot \nabla \right) \vec{v} = -\nabla P + \mu \nabla^2 \vec{v} + \frac{1}{3} \mu \nabla \left(\nabla \cdot \vec{v} \right) + \rho \vec{g} \tag{1.44}$$

where μ is the viscosity of the stellar material and $\vec{g}$ is the gravitational acceleration. Since I have assumed that the stellar viscosity is negligible, we can obtain

$$\rho \left(\frac{\partial}{\partial t} + \vec{v} \cdot \nabla \right) \vec{v} = -\nabla P + \rho \vec{g} \tag{1.45}$$

Notice that this equation at equilibrium is the familiar equation of hydrostatic support (1.6):

$$0 = -\nabla P_0 + \rho_0 \vec{g}_0. \tag{1.46}$$

Substituting the perturbations into Equation (1.45) and dropping all higher-order terms, we find the perturbed equation of motion:

$$\boxed{\rho_0 \frac{\partial^2 \vec{\xi}}{\partial t^2} = -\nabla P' - \rho_0 \nabla \Phi' - \rho' \nabla \Phi_0}. \tag{1.47}$$

Here I have introduced the gravitational potential Φ, the negative gradient of which is the gravitational acceleration:

$$\vec{g} = -\nabla \Phi \qquad \text{and} \qquad \Phi(\vec{r}, t) = -G \int_V \frac{\rho}{|\vec{r} - \vec{x}|} \, d^3\vec{x} \tag{1.48}$$

where V is the volume of the star at equilibrium.

Poisson's equation. Gauss's law for gravity gives that

$$\nabla \cdot \vec{g} = -4\pi G\rho. \tag{1.49}$$

After substituting the gravitational potential and the Eulerian perturbations, we obtain the perturbed Poisson equation to describe the gravitational field:

$$\boxed{\nabla^2 \Phi' = 4\pi G\rho'}. \tag{1.50}$$

The energy equation. The energy equation completes the system by thermodynamically connecting pressure to density. Since I have assumed adiabatic pulsations, the energy equation can be given as

$$\frac{\partial P}{\partial t} + \vec{v} \cdot \nabla P = c^2 \left(\frac{\partial \rho}{\partial t} + \vec{v} \cdot \nabla \rho \right) \tag{1.51}$$

where c is again the adiabatic speed of sound (*cf.* Equation 1.15). Substituting the Lagrangian perturbation, we obtain the perturbed energy equation

$$\boxed{P' + \vec{\xi} \cdot \nabla P_0 = c_0^2 \left(\rho' + \vec{\xi} \cdot \nabla \rho_0 \right)}. \tag{1.52}$$

Symmetry

Now I will apply the assumption of symmetry and consider only oscillatory solutions on a sphere. I separate the displacement vector into radial and horizontal components

$$\vec{\xi} = \xi_r \hat{a}_r + \vec{\xi}_h, \qquad \vec{\xi}_h = \xi_\theta \hat{a}_\theta + \xi_\phi \hat{a}_\phi \tag{1.53}$$

where $\hat{a}$ are unit vectors in indicated directions. The radial component of the displacement, for example, can now be expressed as

$$\xi_r(r, \theta, \phi, t) = \xi_r(r) Y_\ell(\theta, \phi) \exp\{-i\omega t\} \tag{1.54}$$

where θ and ϕ are latitude and longitude, Y_ℓ is Laplace's spherical harmonic for degree ℓ (*cf.* Figure 1.2), i is the imaginary unit, and $\omega = 2\pi\nu$ is the cyclic frequency. When ω^2 is real, the solution is oscillatory; when it is imaginary, the solution either grows or delays. Substituting the spherical, symmetric, harmonic variables into the previous equations (1.43, 1.47, 1.50, 1.52) and dropping subscripts for unperturbed quantities, after some manipulations we may find

$$\frac{d\xi_r}{dr} = -\left(\frac{2}{r} + \frac{1}{\Gamma_1 P} \frac{dP}{dr} \right) \xi_r + \frac{1}{\rho c^2} \left(\frac{S_\ell^2}{\omega^2} - 1 \right) P' - \frac{\ell(\ell+1)}{\omega^2 r^2} \Phi' \tag{1.55}$$

$$\frac{dP'}{dr} = \rho \left(\omega^2 - N^2 \right) \xi_r + \frac{1}{\Gamma_1 P} \frac{dP}{dr} P' + \rho \frac{d\Phi'}{dr} \tag{1.56}$$

$$\frac{1}{r^2} \frac{d}{dr} \left(r^2 \frac{d\Phi'}{dr} \right) = -4\pi G \left(\frac{P'}{c^2} + \frac{\rho}{g} \xi_r N^2 \right) + \frac{\ell(\ell+1)}{r^2} \Phi' \tag{1.57}$$

as well as

$$\vec{\xi}_h(r,\theta,\phi,t) = \sqrt{4\pi}\,\xi_h(r)\left(\frac{\partial Y_\ell}{\partial\theta}\,\hat{a}_\theta + \frac{1}{\sin\theta}\frac{\partial Y_\ell}{\partial\phi}\,\hat{a}_\phi\right)\exp\{-i\omega t\} \tag{1.58}$$

$$\xi_h(r) = \frac{1}{r\omega^2}\left(\frac{1}{\rho}P' - \Phi'\right). \tag{1.59}$$

Here I have introduced the *Brunt-Väisälä* and *Lamb* squared frequencies:

$$N^2 = g\left(\frac{1}{\Gamma_1}\frac{d\ln P}{dr} - \frac{d\ln\rho}{dr}\right) \tag{1.60}$$

$$S_\ell^2 = \frac{\ell(\ell+1)c^2}{r^2} \tag{1.61}$$

which give the regions in the star where modes of different character can propagate. The former, N^2, describes where g-modes can propagate, so called because their restoring force is gravity. The latter, S_ℓ^2, depending on the spherical degree ℓ, describes where p-modes can propagate, called as such because their restoring force is the pressure gradient. These cavities are visualized in Figure 1.17. Here it can be appreciated that g-modes and convection are two sides of the same coin: when $N^2 < 0$ the fluid is unstable to convection; otherwise, the fluid is unstable to g-mode oscillations.

This system of equations (1.55–1.57) constitutes a fourth order boundary eigenvalue problem. Equipped with suitable boundary conditions, we may numerically calculate the eigenfunctions $\vec{\xi}$ (see Figure 1.18) and their corresponding eigenfrequencies ω for a given model of stellar structure. This is the forward problem of stellar pulsation.

Some Properties of Solar-like Oscillations

As we have seen in the first section, oscillation modes of the same spherical degree ℓ can differ in their radial order n and be excited simultaneously with different frequencies. For solar-type stars, it is currently possible to resolve frequencies for modes of low spherical degree ($0 \leqslant \ell \leqslant 3$) and 'high' radial order ($8 \leqslant n \leqslant 31$). The frequency range where oscillation power is maximum, called by ν_{max}, generally corresponds to around $n = 20$ or so. This region of power is proportional to (and obviously lower than) the acoustic cut-off frequency, i.e., the upper frequency bound for oscillations to be reflected back into the star rather than being lost to space:

$$\nu_{max} \propto \nu_{ac} \propto \frac{g}{\sqrt{T_{eff}}}. \tag{1.62}$$

For the Sun, $\nu_{max,\odot} \simeq 3090$ μHz (~ 5.4 minutes) and $\nu_{ac,\odot} \simeq 5000$ μHz (~ 3.3 minutes). Since we lack a proper theoretical treatment of convective transport, which both excites and damps the oscillation modes, we are unable to theoretically predict the amplitudes of the oscillations of our solar model. In lieu of this, we may

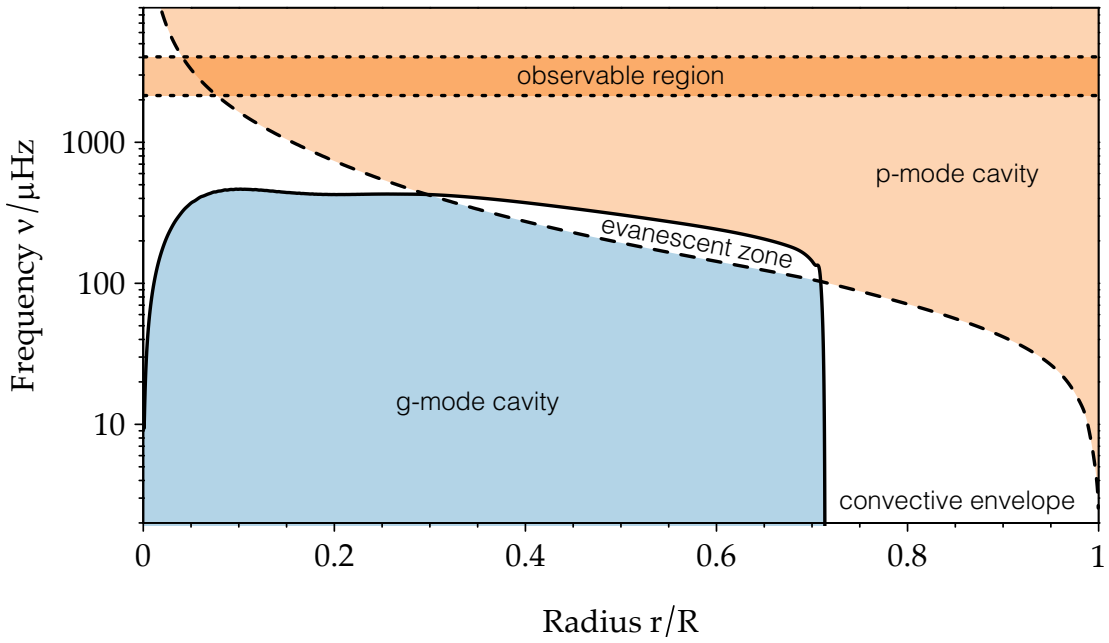


FIGURE 1.17. Propagation diagram for a solar model. The blue-shaded area shows the Brunt-Väisälä region where g-modes can propagate (*cf.* Equation 1.60). The orange-shaded area shows the $\ell = 1$ Lamb region where dipolar p-modes can propagate (*cf.* Equation 1.61). Modes are exponentially damped in the evanescent zone; nevertheless, modes of similar frequency can couple in this region, giving rise to mixed modes. The observable region is a few $\Delta\nu$ around ν_{max}; thus, only p-modes are expected to be observed in this range at this stage of evolution.

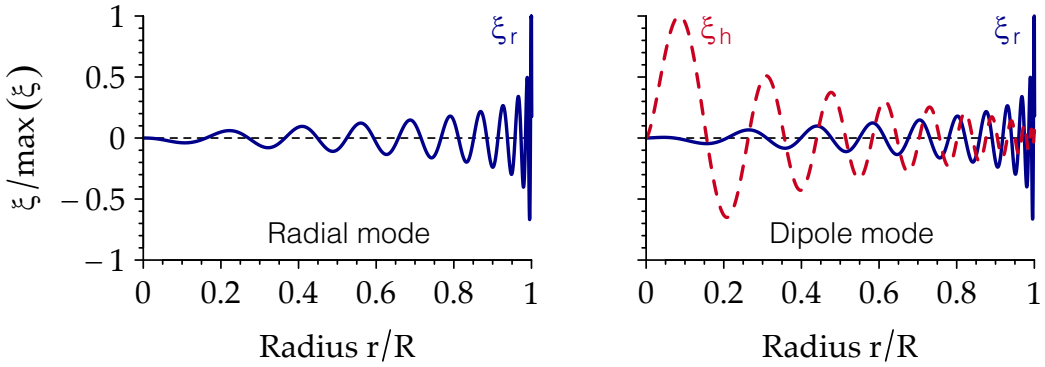


FIGURE 1.18. Radial (blue) and horizontal (red) normalized eigenfunctions for radial ($\ell = 0$, left) and dipolar ($\ell = 1$, right) oscillation modes, both having radial order $n = 20$. The radial displacement of the two modes are quite similar, being only slightly offset in the interior and basically identical in the envelope. The horizontal displacement has zero crossings when the radial displacement is maximal, and vice versa. Radial modes lack horizontal displacement by definition.

try to predict the general region where oscillations with the greatest amplitudes are to be expected by scaling from the observed solar values (e.g., Kjeldsen and Bedding 1995):

$$\frac{\nu_{max,*}}{\nu_{max,\odot}} = \left(\frac{M_*}{M_\odot}\right)\left(\frac{R_*}{R_\odot}\right)^{-2}\left(\frac{T_{eff,*}}{T_{eff,\odot}}\right)^{-\frac{1}{2}} \tag{1.63}$$

and likewise for the acoustic cutoff frequency.

Tassoul (1980) considered oscillation modes in the asymptotic limit of high radial order ($n \gg \ell$) and found that theoretical mode frequencies form a pattern. In particular, adjacent modes of the same spherical degree are approximately equally spaced, which agrees with the observations that we saw in Figures 1.8 and 1.10. The pattern of frequencies can be summarized to first-order approximation as

$$\nu_{n,\ell} \simeq \Delta\nu\left(n + \frac{\ell}{2} + \epsilon\right) \tag{1.64}$$

where $\nu_{n,\ell}$ is the frequency of mode (n, ℓ) and ϵ is a phase shift ($\epsilon_\odot \simeq 1.6$). The spacing $\Delta\nu$ is called the *large frequency separation* and is related to the inverse sound travel time and proportional to the root mean density of the star (Ulrich 1986, Kjeldsen and Bedding 1995):

$$\Delta\nu \simeq \left(2\int \frac{dr}{c}\right)^{-1} \propto \left(\frac{M}{R^3}\right)^{1/2}. \tag{1.65}$$

Since the large frequency separation gives the spacing between modes of different orders, it can be calculated empirically with

$$\Delta\nu_{n,\ell} = \nu_{n,\ell} - \nu_{n-1,\ell}. \tag{1.66}$$

Calculating the average large frequency separation of the Sun for radial modes using data from the Birmingham Solar Oscillations Network (*BiSON*, Broomhall et al. 2009) we can obtain

$$\Delta\nu_\odot = 134.8693 \pm 0.0042\ \mu\text{Hz}. \tag{1.67}$$

This presents an opportunity to test the quality of our solar model. We can calculate the large frequency for our solar-calibrated model either using the inverse sound travel time, or using the frequencies themselves. In the former case, we obtain $\Delta\nu = 136.2970\ \mu$Hz. In the latter, $\Delta\nu = 136.2208\ \mu$Hz.

On the one hand, these model values differ by only about one percent from the solar values, which is quite good by astrophysical standards. On the other hand, when considering the precision with which $\Delta\nu_\odot$ can be calculated, this is a highly significant $\sim 300\sigma$ difference. This difference arises due to our ill treatment of the stellar surface, which we will address later in this section.

A higher-order expansion of the asymptotic expression additionally gives a term known as the *small frequency separation*, the spacing between modes adjacent in frequency and whose spherical degree differs by two (Tassoul 1980):

$$\delta\nu_{n,\ell} = \nu_{n,\ell} - \nu_{n-1,\ell+2} \simeq -(4\ell + 6)\frac{\Delta\nu}{4\pi^2\nu_{n,\ell}}\int \frac{dc}{dr}\frac{dr}{r}. \tag{1.68}$$

As we can see, the small frequency separation is sensitive to the sound speed gradient, and is therefore a good proxy for the conditions in the stellar core, where the sound speed gradient changes sign (*cf.* Figure 1.12). This makes $\delta\nu$ a diagnostic of main-sequence age. We will make use of these relations to infer the properties of stars in Chapter 2, and use computational methods to further understand what properties of stars they reflect in Chapter 3. The average small frequency separation between solar oscillation modes with ($\ell = 0,\ \ell = 2$) is

$$\delta\nu_\odot \simeq 8.957 \pm 0.059\ \mu\text{Hz} \tag{1.69}$$

and for our solar model, $\delta\nu = 8.939\ \mu$Hz, which is good agreement.

A Direct Comparison

We have just compared our solar model against the asymptotic properties of the solar oscillations, finding good agreement with the small frequency separation but less good agreement with the large frequency separation. We may now test the quality of our solar model more directly by comparing the individual pulsation mode frequencies themselves to those observed in the Sun. This comparison is shown in Figure 1.19.

Immediately it can be seen that there are systematic discrepancies between the model and the actual mode frequencies on the order of 10 μHz, i.e., tenths of a percent, which is a difference in period of about 1 to 2 seconds. In particular, the disagreement gets worse with increasing frequency. This phenomenon is called the *surface effect* and has arisen from our improper modelling of the near-surface layers (e.g., Christensen-Dalsgaard 1984). The large frequency separation is also sensitive to surface effects, which is why our model $\Delta\nu$ differed so significantly from the observed value.

It is noteworthy that, because all of the waves propagate essentially radially in the near-surface layers (*cf.* Figures 1.7 and 1.18), the surface term is a function of frequency alone and is independent of the spherical degrees of the modes. The surface effect is thus often dealt with by introducing a correction that increases with frequency. The Ball and Gizon (2014) treatment of the surface term fits coefficients $\mathbf{a}$ to the differences between observed and model frequencies according to

$$\delta\nu_{n,\ell} = \frac{1}{I_{n,\ell}}\left[a_1\left(\frac{\nu_{n,\ell}}{\nu_{ac}}\right)^{-1} + a_2\left(\frac{\nu_{n,\ell}}{\nu_{ac}}\right)^{3}\right] \tag{1.70}$$

where ν_{ac} is the acoustic cutoff frequency, with $\nu_{ac,\odot} \approx 5000$, and $I_{n,\ell}$ is the normalized mode inertia:

$$I_{n,\ell} = \frac{4\pi}{M}\frac{\int \rho\left(|\xi_r|^2 + \ell(\ell+1)|\xi_h|^2\right) r^2\, dr}{|\xi_r(r=R)|^2 + \ell(\ell+1)|\xi_h(r=R)|^2}. \tag{1.71}$$

However, Figure 1.19 further shows that even after correcting for the surface term, differences remain. This implies that even beyond the near-surface layers, the structure of the Sun differs from the model.

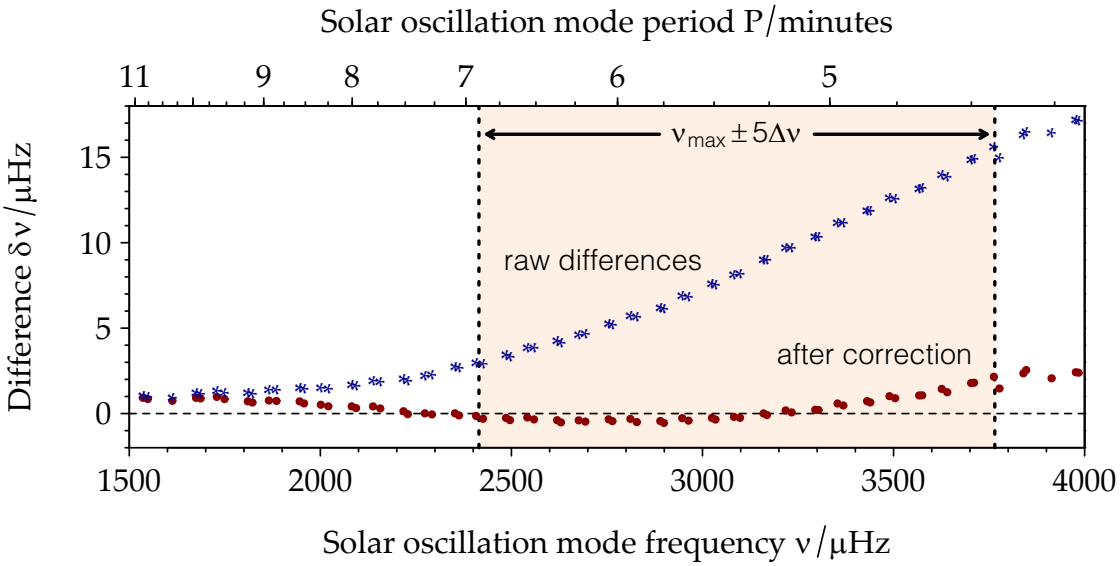


FIGURE 1.19. Differences in oscillation frequencies between the Sun and the best-fitting solar model, in the sense of (model − Sun). Even after correcting for the surface term, substantial differences remain. Being that solar frequencies are measured on the order of one part in a thousand, the uncertainties are too small to be visible at this resolution. The offset at zero is likely due to the assumed solar radius differing from the helioseismic radius. The shaded region indicates what the frequency range of the Sun might be if it were a field star observed by *Kepler*.

This motivates the inverse approach. We have seen that evolutionary theory can produce a model that agrees with the overall properties of the Sun. However, a detailed inspection of the mode frequencies of the model reveals significant disagreement between theory and observation, even after applying corrections. We wish to deduce the actual structure of the Sun and the stars using only asteroseismic arguments: i.e., to find the structure that will pulsate identically. This problem of deducing the structure of a star from its oscillation frequencies is inverse to the problem of deducing the oscillation frequencies from a given stellar structure. In order to pose the inverse problem in a manner that we can solve, however, it is convenient to first make some slight adjustments to our statement of the respective forward problem.

1.3.1 The Relative Forward Problem

The forward problem of asteroseismology is to calculate the seismic frequencies of a stellar model. However, it is not clear how one would go about solving the inverse problem corresponding to this forward problem. Instead, we restate the forward problem as the problem of calculating the frequency *differences* with respect to another model—one with a different structure. That is: by comparing the differences in structure of two models, what will be the differences in their frequencies? I call this the relative forward problem of asteroseismology.

The benefit of posing the problem in this way is that it facilitates the inverse problem, which is to ask: by comparing the frequencies of the two models, what is the difference in their structure? Thus, since we are able to observe frequencies of real stars, we may substitute a star for one of the models, and hence measure the structure of a star.

To give a concrete example, I have calibrated another solar model using different assumptions on the physics of the stellar interior. In particular, this second model differs in that it does not include the effects of elemental diffusion and gravitational settling (i.e., $\mathbf{D}$ is the null matrix in Equation 1.26). This model has the same mass, radius, luminosity, metallicity, and age as the diffusion model—yet it differs in internal structure (see Figure 1.20). The differences in internal structure then give rise to differences in oscillation mode frequencies.

In order to state the relative forward problem, I will first put the oscillation equations in their so-called *variational formulation*, and then linearize the variational frequencies around a reference model. The end result will be a Fredholm integral equation relating the relative differences in oscillation mode frequencies to the relative differences in structure, which will then be a suitable starting point for the inverse analysis.

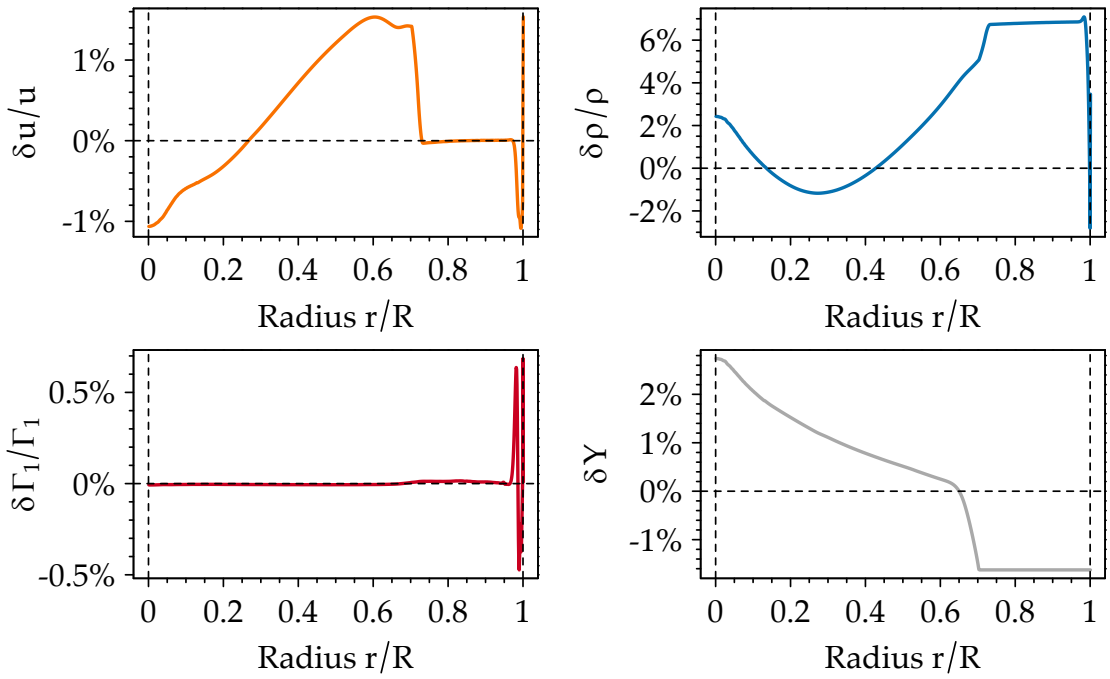


FIGURE 1.20. Relative differences in isothermal sound speed (top left), density (top right), the first adiabatic exponent (bottom left), and helium abundance (bottom right) as a function of radius between two solar-calibrated models with differing input physics (*cf.* Figure 1.12). Although the models have the same overall properties (e.g. mass & age); they differ structurally and chemically throughout their interiors.

Variational Frequencies

The perturbed hydrodynamical equations (1.55–1.57) feature derivatives of the displacement vector. Since we have sought only periodic solutions, we have

$$\vec{\xi}(t) = \vec{\xi} \cdot \exp\{i\omega t\} \qquad \Rightarrow \qquad \frac{\partial \vec{\xi}}{\partial t} = -i\omega\vec{\xi}. \tag{1.72}$$

Combining the perturbed equations, we can arrive at (e.g., Unno et al. 1979)

$$-\omega^2\rho\vec{\xi} = \nabla\left(c^2\rho\nabla\cdot\vec{\xi} + \nabla P\cdot\vec{\xi}\right) - \vec{g}\,\nabla\cdot\left(\rho\vec{\xi}\right) + \rho\vec{g}' \tag{1.73}$$

where I have dropped the subscripts on the unperturbed quantities. This equation relates the cyclic frequency ω to the properties of the stellar structure. Recalling Equation (1.48), we can substitute the perturbed gravitational potential with

$$\vec{g}' = -\nabla\Phi' = G\nabla\int_V \frac{\rho'}{|\vec{r}-\vec{x}|}\,d^3\vec{x} = -G\nabla\int_V \frac{\nabla\cdot\left(\rho\vec{\xi}\right)}{|\vec{r}-\vec{x}|}\,d^3\vec{x}. \tag{1.74}$$

where the latter substitution makes use of the perturbed equation of continuity (Equation 1.43). Thus, all terms in the right hand side of Equation (1.73) are functions of $\vec{\xi}$, and so it is an eigenvalue problem of the form

$$\mathcal{L}(\vec{\xi}_i) = -\omega_i^2\vec{\xi}_i \tag{1.75}$$

with $\mathcal{L}$ being the linear integro-differential operator satisfying that equation. Now $\vec{\xi} \equiv \vec{\xi}_i$ is the displacement eigenfunction for the mode with label $i \equiv (n, \ell)$ and $\omega \equiv \omega_i$ is its corresponding eigenfrequency. Chandrasekhar (1964) showed that when $\rho = P = 0$ at the outer boundary, this eigenvalue problem is Hermitian, i.e.,

$$\langle\vec{\xi}, \mathcal{L}(\vec{\eta})\rangle = \langle\mathcal{L}(\vec{\xi}), \vec{\eta}\rangle \tag{1.76}$$

where $\langle\cdot\rangle$ denotes the inner product defined by

$$\langle\vec{\xi}_i, \vec{\eta}_i\rangle = \int_V \rho\vec{\xi}_i^*\cdot\vec{\eta}_i\,d^3\vec{r} = 4\pi\int \rho\,(\xi_r^*\eta_r + \ell(\ell+1)\xi_h^*\eta_h)\,r^2\,dr. \tag{1.77}$$

Here * is the complex conjugate and $\vec{\eta}$ is any (suitably regular) vector function of stellar structure. This is useful because then squared mode frequencies are real and may be calculated via

$$-\omega_i^2 = \frac{\langle\vec{\xi}_i, \mathcal{L}(\vec{\xi}_i)\rangle}{\langle\vec{\xi}_i, \vec{\xi}_i\rangle} \tag{1.78}$$

where $\vec{\xi}_i$ is an eigenvector of the problem and ω_i^2 is a real eigenvalue. A further property is that the eigenvectors of the problem are orthogonal. Finally, we have the variational principle: perturbations to an eigenvector result in only second-order perturbations to the corresponding eigenvalue. Frequencies calculated using Equations (1.78) are referred to as variational frequencies.

Linearization Around a Reference Model

We now seek to linearize the problem around a reference model. We consider a small perturbation to the eigenfrequency, call it $\delta\omega^2$, to the eigenfunction, $\delta\vec{\xi}$, and to the operator, $\delta\mathcal{L}$:

$$\Big(\mathcal{L}+\delta\mathcal{L}\Big)\Big(\vec{\xi}+\delta\vec{\xi}\Big)=-\Big(\omega+\delta\omega\Big)^2\Big(\vec{\xi}+\delta\vec{\xi}\Big). \tag{1.79}$$

After perturbing all the components from Equation (1.73), we can find (e.g., Antia and Basu 1994)

$$\begin{aligned}\delta\mathcal{L}(\vec{\xi}) = &\frac{\nabla\rho}{\rho}\delta c^2\nabla\cdot\vec{\xi}+\nabla\left(\delta c^2\nabla\cdot\vec{\xi}+\delta\vec{g}\cdot\vec{\xi}\right)+\delta\vec{g}\,\nabla\cdot\vec{\xi}\\ &+\nabla\left(\frac{\delta\rho}{\rho}\right)c^2\nabla\cdot\vec{\xi}-G\nabla\int_V\frac{\nabla\cdot\left(\delta\rho\vec{\xi}\right)}{|\vec{r}-\vec{\chi}|}\,d^3\vec{\chi}.\end{aligned} \tag{1.80}$$

Expanding Equation (1.79), we find at the first order

$$\mathcal{L}(\delta\vec{\xi})+\delta\mathcal{L}(\vec{\xi})=-\omega^2\delta\vec{\xi}-2\omega\delta\omega\vec{\xi}. \tag{1.81}$$

Taking the product of both sides with $(\rho\vec{\xi}^*)$ and integrating, we obtain

$$\begin{aligned}&\int_V\rho\vec{\xi}^*\cdot\mathcal{L}(\delta\vec{\xi})\,d^3\vec{r}+\int_V\rho\vec{\xi}^*\cdot\delta\mathcal{L}(\vec{\xi})\,d^3\vec{r}\\ =&-\omega^2\int_V\rho\vec{\xi}^*\cdot\delta\vec{\xi}\,d^3\vec{r}-2\omega\delta\omega\int_V\rho\vec{\xi}^*\cdot\vec{\xi}\,d^3\vec{r}.\end{aligned} \tag{1.82}$$

Since $\mathcal{L}$ is Hermitian, the first term on both sides cancel to give

$$\delta\omega=-\frac{1}{2\omega}\frac{\langle\vec{\xi},\delta\mathcal{L}(\vec{\xi})\rangle}{\langle\vec{\xi},\vec{\xi}\rangle}. \tag{1.83}$$

Now plugging $\delta\mathcal{L}$ from Equation (1.80) into Equation (1.83) and assuming that $\delta P=0$ at the outer boundary (e.g., Lynden-Bell and Ostriker 1967), one may use integration by parts to obtain, quite generally, a Fredholm integral relation for each mode of oscillation i:

$$\boxed{\frac{\delta\omega_i}{\omega_i}=\int K_i^{(f_1,f_2)}\frac{\delta f_1}{f_1}+K_i^{(f_2,f_1)}\frac{\delta f_2}{f_2}\,dr}. \tag{1.84}$$

Here f_1 and f_2 are two variables of stellar structure (e.g., sound speed and density), and δf_1 and δf_2 are the differences with respect to another model. Relative differences in the frequencies $\delta\omega_i/\omega_i$ of mode $i\equiv(n,\ell)$ between two models relate to relative differences in physical quantities of those models via a pair of *kernel functions* $\vec{K}_i$.

Equation (1.84) is the central equation of this thesis, as this is the equation that we will use to infer the internal structures of stars. In particular, we will determine the stellar structure profile f_1 of a star (for some choice of $\vec{f}$, discussed later) by deducing the relative difference with a best-fitting evolutionary model $\delta f_1/f_1$ via inversion of this equation. This is the structure inversion problem, which we will revisit in Section 1.4 and Chapter 4. For now, we will continue by inspecting the kernel functions in detail.

1.3.2 Stellar Structure Kernels

We have seen in Equation (1.84) that perturbations to the stellar structure translate into perturbations in oscillation mode frequencies, and kernel functions quantify that response. The kernels for any given pair of stellar structure variables can be calculated by transforming Equation (1.83) into an equation in the form of Equation (1.84). Because the variables of stellar structure are not independent, kernels must be given with respect to (at least) two variables simultaneously. Here I will give the kernels for the following pairs: (c, ρ), (c^2, ρ), (Γ_1, ρ), and (u, Y).

Kernel Pair c, ρ

The kernels for the sound speed and density, i.e. $(f_1, f_2) = (c, \rho)$ of Equation (1.84), can be found as (*cf.* Gough and Thompson 1991)

$$\omega^2 \mathcal{S} K_i^{(c,\rho)} = r^2 \rho c^2 \chi^2 \tag{1.85}$$

$$\begin{aligned}\omega^2 \mathcal{S} K_i^{(\rho,c)} = & -\frac{1}{2}\left(\xi_r^2 + L^2 \xi_h^2\right) r^2 \rho \omega^2 \\ & + \frac{1}{2}\rho c^2 \chi^2 r^2 - G m \rho \left(\chi + \frac{1}{2}\xi_r \frac{d \ln \rho}{dr}\right) \xi_r \\ & - 4\pi G \rho r^2 \int_r^R \left(\chi + \frac{1}{2}\xi_r \frac{d \ln \rho}{ds}\right) \xi_r \rho \, ds \\ & + G m \rho \, \xi_r \frac{d\xi_r}{dr} + \frac{1}{2} G \left(m \frac{d\rho}{dr} + 4\pi r^2 \rho^2\right) \xi_r^2 \\ & - \frac{4\pi G}{2\ell + 1} \rho \Bigg[(\ell + 1) r^{-\ell} \left(\xi_r - \ell \xi_h\right) \int_0^r \left(\rho \chi + \xi_r \frac{d\rho}{ds}\right) s^{\ell+2} \, ds \\ & \qquad - \ell r^{\ell+1} \left(\xi_r + (\ell + 1) \, \xi_h\right) \int_r^R \left(\rho \chi + \xi_r \frac{d\rho}{ds}\right) s^{-(\ell-1)} \, ds \Bigg]\end{aligned} \tag{1.86}$$

where I have introduced the dilatation

$$\chi = \frac{d\xi_r}{dr} + 2\frac{\xi_r}{r} - \ell(\ell + 1)\frac{\xi_h}{r} \tag{1.87}$$

and $\mathcal{S}$ is a quantity proportional to the energy of the mode

$$\mathcal{S} = \int \rho \left(\xi_r^2 + \ell(\ell + 1)\xi_h^2\right) r^2 \, dr. \tag{1.88}$$

Kernel Pair c^2, ρ

Since all kernel pairs must satisfy Equation (1.84), it is straightforward to transform kernel pair (c, ρ) to kernel pair (c^2, ρ). We have that

$$\int K_i^{(c,\rho)} \frac{\delta c}{c} + K_i^{(\rho,c)} \frac{\delta \rho}{\rho} \, dx = \int K_i^{(c^2,\rho)} \frac{\delta c^2}{c^2} + K_i^{(\rho,c^2)} \frac{\delta \rho}{\rho} \, dx. \tag{1.89}$$

We may expand the sound speed perturbation as

$$\frac{\delta c^2}{c^2} = \frac{2c\delta c}{c^2} = 2\frac{\delta c}{c} \tag{1.90}$$

hence we have

$$K_i^{(c^2,\rho)} = \frac{1}{2} K_i^{(c,\rho)} \tag{1.91}$$

$$K_i^{(\rho,c^2)} = K_i^{(\rho,c)}. \tag{1.92}$$

It is instructive at this point to inspect some kernels and see what they actually look like. Figures 1.21 and 1.22 show Equations (1.91) and (1.92) for various different oscillation modes of a solar model. These kernels tell us how perturbations to the relevant physical variables would translate into perturbations of the respective oscillation mode frequencies. The figure additionally shows more kernel pairs, some of which will be also derived in this section.

Kernel Pair Γ_1, ρ

Kernel functions for the first adiabatic exponent and density may be transformed from (c^2, ρ) kernels via (e.g. Reese et al. 2014, Equations 104-105):

$$K_i^{(\Gamma_1,\rho)} = K_i^{(c^2,\rho)} \tag{1.93}$$

$$K_i^{(\rho,\Gamma_1)} = K_i^{(\rho,c^2)} - K_i^{(c^2,\rho)} + \frac{Gm\rho}{r^2} \int_{s=0}^{r} \frac{\Gamma_1 \chi^2 s^2}{2\mathcal{S}\omega^2}\, ds + \rho r^2 \int_{s=r}^{R} \frac{4\pi G\rho}{s^2} \left(\int_{t=0}^{s} \frac{\Gamma_1 \chi^2 t^2}{2\mathcal{S}\omega^2}\, dt \right) ds. \tag{1.94}$$

Kernel Pair u, Y

Using additional assumptions, for example under assumption of the EOS, we may formulate kernels for other quantities such as the fractional helium abundance. For each mode i we wish to obtain the pair of kernel functions for the isothermal sound speed (recall Equation 1.15) and helium abundance Y

$$\vec{K}_i^{(2)} = \left[K_i^{(u,Y)}, K_i^{(Y,u)} \right] \tag{1.95}$$

via conversion from the kernel pair of (Γ_1, ρ)

$$\vec{K}_i^{(1)} = \left[K_i^{(\rho,\Gamma_1)}, K_i^{(\Gamma_1,\rho)} \right]. \tag{1.96}$$

We can expand the perturbation to the first adiabatic exponent as

$$\frac{\delta\Gamma_1}{\Gamma_1} = \Gamma_{1,\rho}\frac{\delta\rho}{\rho} + \Gamma_{1,P}\frac{\delta P}{P} + \Gamma_{1,Y}\delta Y \tag{1.97}$$

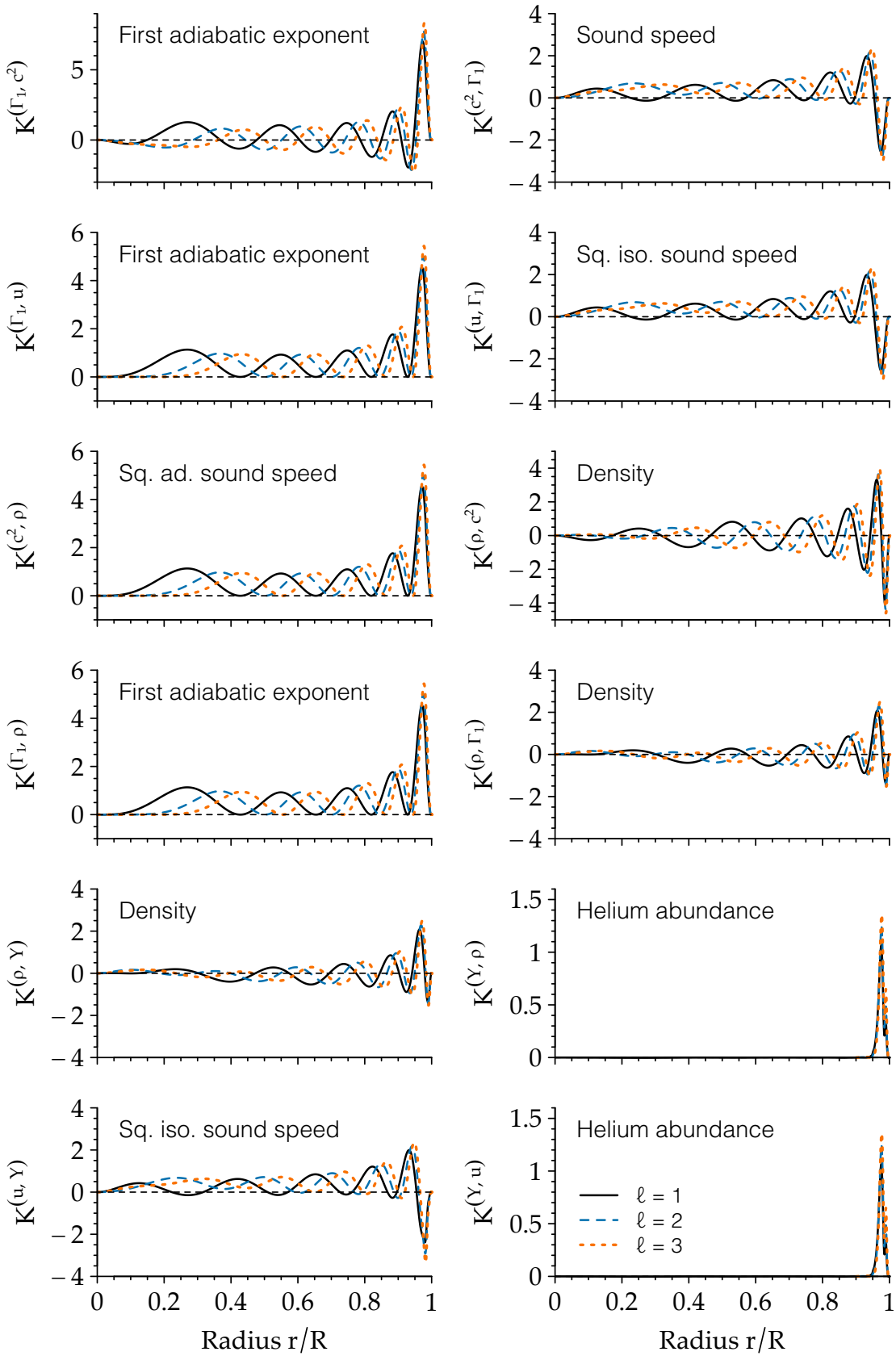


FIGURE 1.21. Pairs of kernel functions for modes with the same radial order $n = 5$ and different spherical degrees $\ell = 1, 2, 3$.

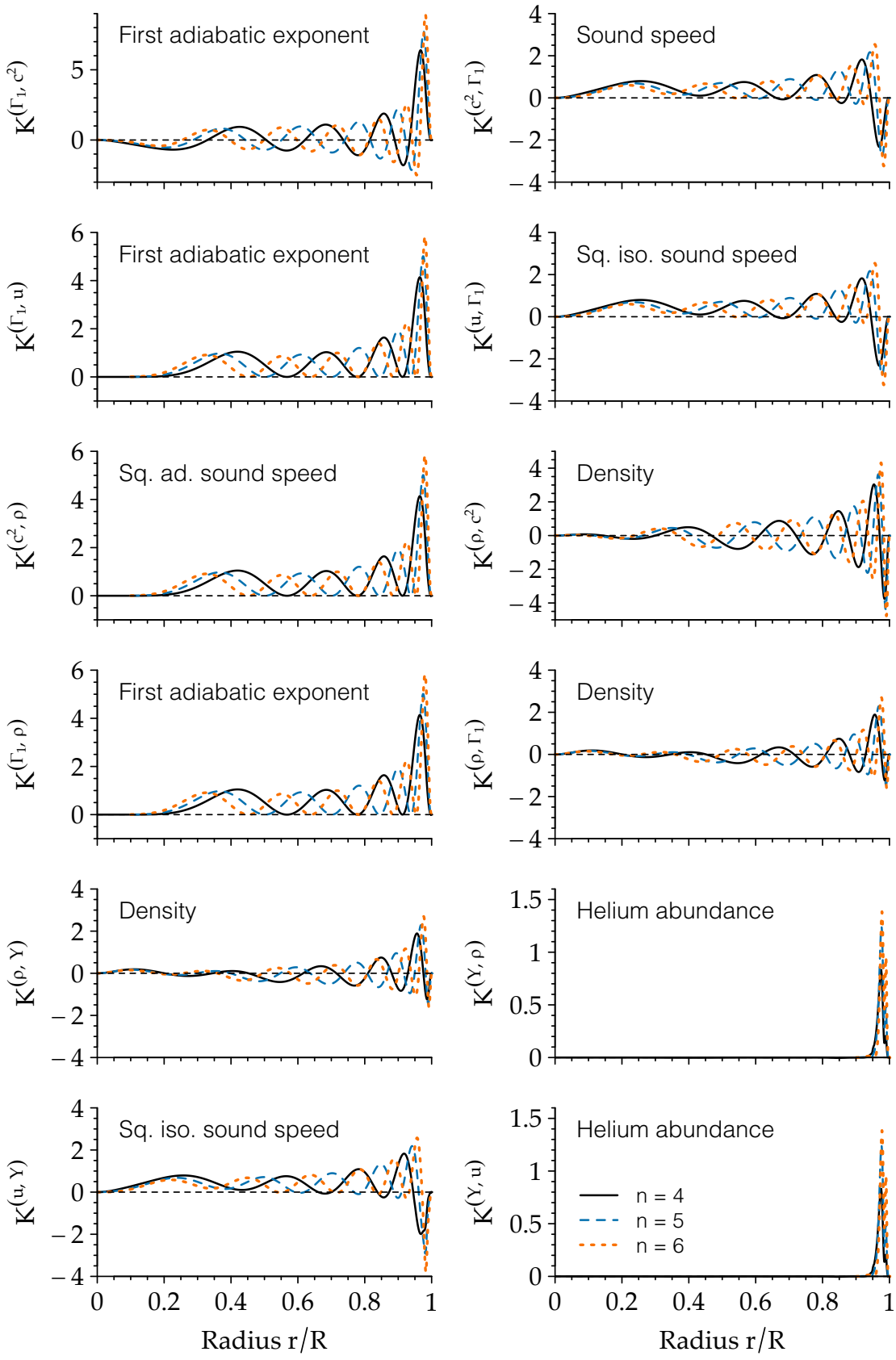


FIGURE 1.22. Pairs of kernel functions for modes with the same spherical degree $\ell = 2$ and different radial order $n = 4, 5, 6$.

where I have introduced the quantities

$$\Gamma_{1,\rho} \equiv \left(\frac{\partial \ln \Gamma_1}{\partial \ln \rho}\right)_{P,Y} \qquad \Gamma_{1,P} \equiv \left(\frac{\partial \ln \Gamma_1}{\partial \ln P}\right)_{\rho,Y} \qquad \Gamma_{1,Y} \equiv \left(\frac{\partial \ln \Gamma_1}{\partial Y}\right)_{\rho,P} \tag{1.98}$$

which are calculated from the assumed EOS. There are two formulations of these kernels that appear in the literature: the Thompson and Christensen-Dalsgaard (2002) formulation and the Kosovichev (1999) formulation. For the sake of completeness, I show both here.

Thompson–JCD Formulation. This kernel pair may be calculated with (Thompson and Christensen-Dalsgaard 2002, their Equation A9)

$$K_i^{(Y,u)} = \Gamma_{1,Y} \cdot K_i^{(\Gamma_1,\rho)} \tag{1.99}$$

$$K_i^{(u,Y)} = \Gamma_{1,P} \cdot K_i^{(\Gamma_1,\rho)} - P \cdot \frac{d}{dr}\left(\frac{\psi_i}{P}\right) \tag{1.100}$$

where $\psi(r)$ is the solution to the system of differential equations

$$\frac{\rho}{r^2 P}\psi_i = \frac{1}{4\pi G} \cdot \frac{d}{dr}\left(\frac{F_i}{r^2\rho} - \frac{1}{r^2\rho} \cdot \frac{d\psi_i}{dr}\right) \tag{1.101}$$

$$F_i(r) = (\Gamma_{1,P} + \Gamma_{1,\rho}) \cdot K_i^{(\Gamma_1,\rho)} + K_i^{(\rho,\Gamma_1)} \tag{1.102}$$

with boundary conditions

$$\psi(r = 0) = \psi(r = R) = 0. \tag{1.103}$$

In order to calculate these kernels, we must first solve Equation (1.101) for ψ numerically. As it is a system of second-order differential equations, we must first massage it into a first-order system. We may integrate both sides of Equation (1.101) to obtain

$$\frac{d\psi_i}{dr} = F_i - 4\pi G r^2 \rho \int_{s=r}^{R} \frac{\rho}{s^2 P}\psi_i \, ds. \tag{1.104}$$

I use this approach here in this thesis.

Kosovichev Formulation. First let (Kosovichev 1999, his Equations 40; 43-45; 48)

$$U = \frac{4\pi\rho r^3}{m} \qquad V = \frac{Gm\rho}{rP} \tag{1.105}$$

$$A = \left(\begin{bmatrix} V & -V \\ 0 & -U \end{bmatrix} + \begin{bmatrix} -V & 0 \\ U & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -\Gamma_{1,\rho} & 1 \end{bmatrix}^{-1} \begin{bmatrix} 1 & 0 \\ \Gamma_{1,P} & 0 \end{bmatrix} \right) = \begin{bmatrix} 0 & -U \\ V & U \end{bmatrix} \tag{1.106}$$

$$B = \left(\begin{bmatrix} -V & 0 \\ U & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -\Gamma_{1,\rho} & 1 \end{bmatrix}^{-1} \begin{bmatrix} -1 & 0 \\ 0 & \Gamma_{1,Y} \end{bmatrix} \right) = \begin{bmatrix} V & 0 \\ -U & 0 \end{bmatrix} \tag{1.107}$$

$$C = \left(\begin{bmatrix} 1 & 0 \\ -\Gamma_{1,\rho} & 1 \end{bmatrix}^{-1} \begin{bmatrix} 1 & 0 \\ -\Gamma_{1,P} & 0 \end{bmatrix} \right) = \begin{bmatrix} 1 & 0 \\ \Gamma_{1,\rho} + \Gamma_{1,P} & 0 \end{bmatrix} \tag{1.108}$$

$$D = \left(\begin{bmatrix} 1 & 0 \\ -\Gamma_{1,\rho} & 1 \end{bmatrix}^{-1} \begin{bmatrix} -1 & 0 \\ 0 & \Gamma_{1,Y} \end{bmatrix} \right) = \begin{bmatrix} -1 & 0 \\ -\Gamma_{1,\rho} & \Gamma_{1,Y} \end{bmatrix}. \tag{1.109}$$

The kernels can be expressed in matrix form

$$\vec{K}_i^{(2)} = D^T \vec{K}^{(1)} - B^T \vec{w} \tag{1.110}$$

with $\vec{w}$ being the solution of the differential equation

$$\frac{d}{d\ln r}[\vec{w}] = -A^T \vec{w} - C^T \vec{K}^{(1)} \tag{1.111}$$

having boundary conditions

$$\frac{\delta\rho}{\rho} w_1 + \frac{\delta m}{m} w_2 = 0 \text{ at } r = 0 \text{ and } r = R. \tag{1.112}$$

By substitution of these matrices, we have that $\vec{w}$ is the solution to

$$\frac{dw_1}{d\ln r} = -\frac{4\pi\rho r^3}{m} w_2 - K_i^{(\rho,\Gamma_1)} - \left(\Gamma_{1,\rho} + \Gamma_{1,P}\right) K_i^{(\Gamma_1,\rho)} \tag{1.113}$$

$$\frac{dw_2}{d\ln r} = \frac{Gm\rho}{rP} w_1 + \frac{4\pi\rho r^3}{m} w_2. \tag{1.114}$$

Since these derivatives are with respect to a logarithmic quantity, and recalling the identity

$$\frac{dx}{d\ln y} = y\frac{dx}{dy} \tag{1.115}$$

we cast Equation (1.111) into a useful form as a linear system of first-order differential equations

$$\frac{dw_1}{dr} = -\frac{4\pi\rho r^2}{m} w_2 - \frac{1}{r}\left[K_i^{(\rho,\Gamma_1)} + \left(\Gamma_{1,\rho} + \Gamma_{1,P}\right) K_i^{(\Gamma_1,\rho)}\right] \tag{1.116}$$

$$\frac{dw_2}{dr} = \frac{Gm\rho}{r^2 P} w_1 + \frac{4\pi\rho r^2}{m} w_2 \tag{1.117}$$

with the boundary conditions of Equation (1.112), which without loss of generality may be transformed into

$$w_1(r = 0) = w_2(r = R) = 0. \tag{1.118}$$

Finally we may calculate the kernels using this $\vec{w}$ by substituting the matrices above into Equation (1.110) to get

$$K_i^{(u,Y)} = -K_i^{(\rho,\Gamma_1)} - \Gamma_{1,\rho} \cdot K_i^{(\Gamma_1,\rho)} + \frac{Gm\rho}{rP} w_1 - \frac{4\pi\rho r^3}{m} w_2 \tag{1.119}$$

$$K_i^{(Y,u)} = \Gamma_{1,Y} \cdot K_i^{(\Gamma_1,\rho)}. \tag{1.120}$$

These last kernels—the (u, Y) kernel pair—are especially valuable for the following analysis. An inspection of their form (Figures 1.21 and 1.22) reveals that the Y kernels only have amplitude in ionization zones, which are located near to the stellar surface. As we will see later, this implies that it will be possible to isolate the effects of differences in mode frequencies to differences in internal isothermal sound speeds.

Testing the Forward Formulation

We may now compare the actual frequency differences between the two solar models to the differences that we get through the kernel equation (Equation 1.84). The top pair of plots in Figure 1.23 shows this comparison for the (c^2, ρ) and (u, Y) kernel pairs. Here I have shown the comparison using the set of modes (i.e., the n, ℓ labels) that have been observed in 16 Cyg B. As we have seen previously, the differences again increase as a function of frequency due to surface effects. We therefore modify Equation (1.84) to take this phenomenon into account by including the Ball and Gizon (2014) surface term:

$$\boxed{\frac{\delta\nu_i}{\nu_i} = \int_0^R \left[K_i^{(f_1,f_2)} \frac{\delta f_1}{f_1} + K_i^{(f_2,f_1)} \frac{\delta f_2}{f_2} \right] dr + \frac{F(\nu_i)}{I_i}} \tag{1.121}$$

where $F(\nu_i)$ is adapted from the surface term of Equation (1.70)

$$F(\nu_i) = a_1 \left(\frac{\nu_i}{\nu_{ac}} \right)^{-2} + a_2 \left(\frac{\nu_i}{\nu_{ac}} \right)^{2}. \tag{1.122}$$

Figure 1.23 shows that after applying the surface term correction, the agreement between the exact differences and those obtained through the kernels is much better. In other words, through the use of the stellar structure kernels, we can translate differences in structure to differences in pulsation frequency.

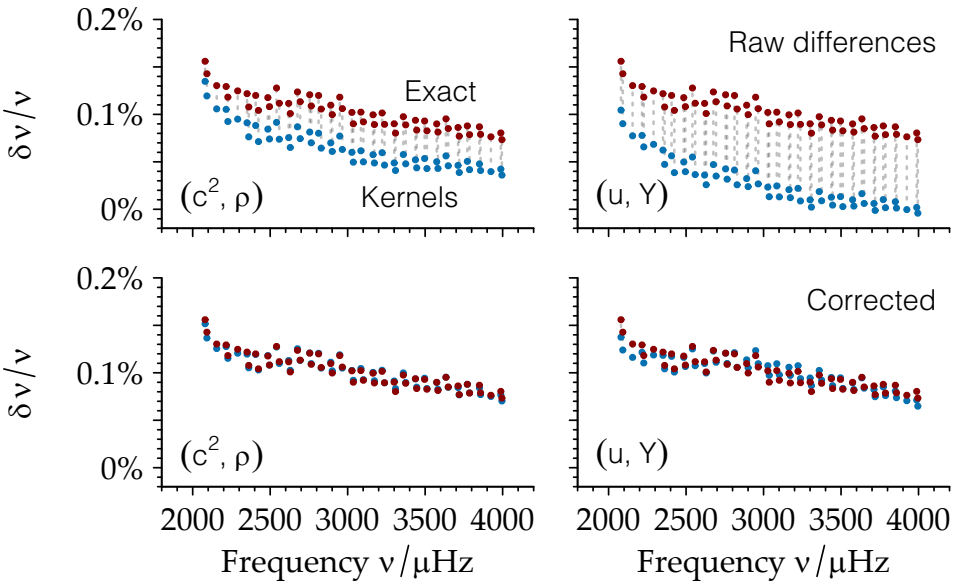


FIGURE 1.23. Top: Relative frequency differences between two solar models using the 16 Cyg B mode set. The points in red are the exact differences; the points in blue are the differences obtained through Equation (1.84) using (c^2, ρ) kernels (left) and (u, Y) kernels (right). Bottom: the same, but also including the surface-term corrections of Equation (1.121).

1.4 Inverse Problems

In this section, I will provide a general summary of inverse problems, with particular attention toward those that are posed and solved in the subsequent chapters of this thesis. Several textbooks discuss inverse problems and their solutions. In writing this section, I have made use of the textbooks by Basu and Chaplin (2017), Kirsch (2011) and Neto and Neto (2012). Additionally, I have found the reviews by Tenorio (2001), Gough and Thompson (1991), and Reese (2018) helpful.

So far we have concerned ourselves with discussions of *forward problems*. These can be thought of as problems where we have a theory, we input some initial conditions, and we compute the result deterministically. The two topics of the previous chapters have been the theory of stellar evolution and the theory of stellar pulsation. In the case of evolution, we supplied the initial conditions (mass, initial composition, mixing length parameter, etc.), and then applied the theory to simulate what such a star would be like at each given time in the future. In the case of pulsation, we supplied a static stellar structure, and then applied the theory to calculate the corresponding frequencies of oscillation. Now we wish to go in the opposite direction (see Figure 1.24).

> *"The cause is hidden, but the result is known."*
>
> — Ovid
> *Metamorphoses* (8 AD)

In the case of evolution, given the observation of a star (e.g., its luminosity, or pulsation data), we wish to determine its overall properties (e.g., mass, radius, age) and evolutionary history (initial composition and so on) using the theory of

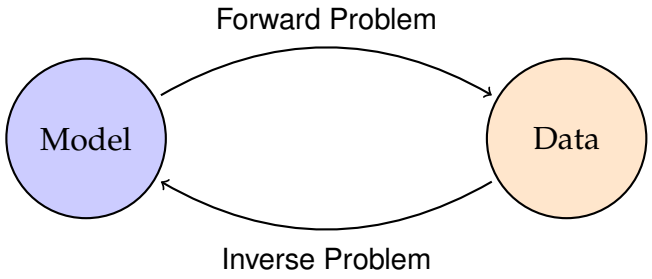


FIGURE 1.24. A schematic for the relationship between forward and inverse problems. In the forward problem, we use the theory or a model to generate data, such as the types of information that could be observed about a system. In the inverse problem, we seek to reconstruct all the possibilities that are consistent with that observed data.

evolution. In the case of pulsation, given the observed oscillation frequencies, we wish to determine the stellar structure that supports those oscillations using the theory of stellar pulsation. These are the inverse problems of asteroseismology that form this thesis.

The difficulty in solving these problems comes in part from the fact that they are *ill-posed*. At the beginning of the 20th century, the French mathematician Jacques Hadamard (1902) gave his definition for what constitutes a well-posed problem. Hadamard believed that problems worth consideration should have the properties that

1. a solution exists (existence),
2. the solution is unique (uniqueness), and
3. the solution changes continuously with changes to the input (stability).

A problem that fails to meet one or more of these criteria is then said to be ill-posed.

> "*The respect for Hadamard was so great that incorrectly posed problems were considered 'taboo' for generations of mathematicians, until comparatively recently it became clear that there are a number of quite meaningful problems, the so-called 'inverse problems,' which are nearly always unstable with respect to fluctuations of input data.*"
>
> — H. Allison
> *Inverse Unstable Problems and Some of Their Applications* (1979)

An example of a well-posed problem is: given the formula for a line and some coordinates, calculate the corresponding points on the line. The inverse of this problem—calculating the formula of a line given points belonging to it—also happens to be well-posed. Suppose however that we only have one point. Then the uniqueness condition is not satisfied, as infinitely many lines pass through that point. Suppose instead that we have multiple points, but one of the points does not actually belong to the line. Then the existence condition is not satisfied, as no one line passes through all the points.

One of the most famous inverse problems is the question from mathematician Mark Kac: "Can One Hear the Shape of a Drum?" (Kac 1966). In a response article entitled "You Can't Hear the Shape of a Drum," Gordon and Webb (1996) produced two different drums with the same eigenfrequencies. The solution to the problem therefore lacks uniqueness, and so it is ill-posed.

The solutions to physical inverse problems often lack uniqueness. At a basic level, measurements are nearly always uncertain, and therefore the solution is uncertain. Less obvious however is that two distinct sets of initial conditions can often lead to the same observables (i.e., the forward function is non-injective, see Figure 1.25). This is sometimes referred to as degeneracy. The evolution inverse problem has the additional issue that there are observations of stars (the Sun

is an example) that cannot yet be fully reproduced by any evolutionary model (i.e., the forward function is non-surjective, see again Figure 1.25). This is one reason why we separate the two inverse problems, and use the solution from the evolution inversion as the starting point for the structure inversion.

The word "inverse" is especially appropriate because inverse problems can often be stated as finding the inverse of a forward function, operator, or matrix. For example, if we have a model M that takes initial conditions x and produces data $y = M(x)$, then the inverse problem is to determine $x = M^{-1}(y)$ from observations of y. This is where the condition of stability often runs into problems.

As an example, consider a simple theory defined by the following linear system of equations:

$$x_1 + x_2 = y_1 \tag{1.123}$$

$$x_1 + (1+\epsilon)x_2 = y_2 \tag{1.124}$$

where ϵ is an arbitrarily small number. The values y_1 and y_2 are then observed in nature, each coming up to be $y_1 = y_2 = 2$. We now seek the "initial conditions" $\mathbf{x} \equiv (x_1, x_2)$ to explain this observation. The solution is clearly $\mathbf{x} = (2, 0)$. Now consider that y_2 was instead measured to be $2 + \epsilon$. The solution then changes to $\mathbf{x} = (1, 1)$. Recall however that ϵ was chosen to be arbitrarily small. Thus, an arbitrarily small change to the measurement has completely changed the solution. To be even more concrete, if we let $\epsilon = 10^{-10}$ and modify y_2 to be, say, $2 + 10^{-5}$, then we obtain $\mathbf{x} \simeq (-99998, 10000)$. The system is unstable.

In matrix notation, this system corresponds to

$$\mathbf{M}\mathbf{x} = \mathbf{y}, \qquad \mathbf{M} = \begin{bmatrix} 1 & 1 \\ 1 & 1+\epsilon \end{bmatrix}. \tag{1.125}$$

Here our model is the nearly singular matrix $\mathbf{M}$, we have observed the data $\mathbf{y}$, and we've sought the initial conditions $\mathbf{x} = \mathbf{M}^{-1}\mathbf{y}$. When the condition number

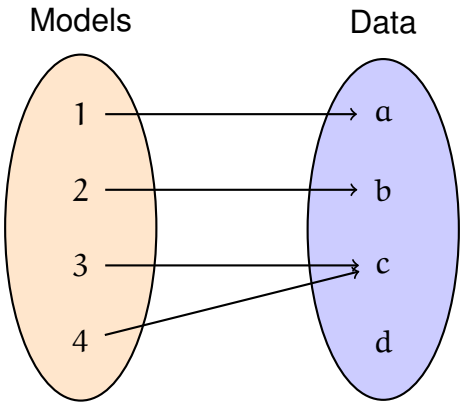


FIGURE 1.25. Physical systems are often *non-injective* in the sense that two systems may have different internal conditions but the same external observables. Here models 3 and 4 share the same set of observables c. This system is also *non-surjective* because the fourth set of observations is not produced by any model.

$\kappa(\mathbf{M}) = \|\mathbf{M}\| \, \|\mathbf{M}^{-1}\|$ is large, the problem is said to be ill-conditioned. When $\kappa = \infty$, the problem is ill-posed. For this particular system,

$$\lim_{\epsilon \to 0} \kappa(\mathbf{M}) = \infty. \tag{1.126}$$

The kernel functions that we derived in the previous section are nearly linearly dependent across the different modes, and so the structure inversion problem is ill-conditioned. As we will see later, such problems are generally dealt with by enforcing stability or regularity conditions, i.e., regularization (e.g., Tikhonov 1977, Tenorio 2001).

1.4.1 Evolution Inversions

With the equations of Section 1.2 and some chosen initial conditions, we can simulate the life of a star, and at each step of the way, determine what observations of that star would yield. Thus we have a forward model M which is parameterized by initial conditions $\mathbf{x}$ and time τ, and yields data $\mathbf{y}$:

$$M(\mathbf{x}, \tau) = \mathbf{y} \tag{1.127}$$

$$\mathbf{x} = [M, Y_0, Z_0, \alpha_{MLT}, \ldots] \tag{1.128}$$

$$\mathbf{y} = [L, T_{eff}, [Fe/H], \boldsymbol{\nu}, \ldots]. \tag{1.129}$$

We now seek to interpret observations of a star in the context of the theory of stellar evolution. In other words, we seek the inverse function:

$$M^{-1}(\mathbf{y}) = [\mathbf{x}, \tau]. \tag{1.130}$$

Of course, we can also seek a function that outputs additional quantities at the present age, such as the radius if it has not been observed. There are several approaches that have been taken to solve this problem, which I will now review.

Scaling Relations

A simple approach to estimate stellar properties is to "scale" them from solar values using the equations of stellar structure and pulsation. While such an approach does not solve the full evolution inversion problem, it shares a common goal of estimating (a more limited set of) properties such as the stellar mass.

A simple example comes from the Stefan-Boltzmann law (Equation 1.22). Replacing this equation with ratios with respect to the solar values, we may obtain

$$\frac{R_*}{R_\odot} = \left(\frac{L_*}{L_\odot}\right)^{-2} \left(\frac{T_{eff,*}}{T_{eff,\odot}}\right)^4 \tag{1.131}$$

from which we can estimate an unknown stellar radius R_* from a measured stellar luminosity L_* and effective temperature $T_{eff,*}$. In principle, this relation works; in practice, the luminosities of most stars are unknown, and effective temperatures are measured rather imprecisely ($\gtrapprox 50$ K uncertainty).

The same kind of manipulation can be used on the asymptotic equations of stellar pulsation to obtain stellar masses and radii. From manipulation of Equations (1.62) and (1.66) we find (e.g., Kjeldsen and Bedding 1995):

$$\frac{R_*}{R_\odot} = \left(\frac{\nu_{\max,*}}{\nu_{\max,\odot}}\right)\left(\frac{\Delta\nu_*}{\Delta\nu_\odot}\right)^2\left(\frac{T_{\text{eff},*}}{T_{\text{eff},\odot}}\right)^{\frac{1}{2}} \tag{1.132}$$

$$\frac{M_*}{M_\odot} = \left(\frac{\nu_{\max,*}}{\nu_{\max,\odot}}\right)^3\left(\frac{\Delta\nu_*}{\Delta\nu_\odot}\right)^4\left(\frac{T_{\text{eff},*}}{T_{\text{eff},\odot}}\right)^{\frac{3}{2}} \tag{1.133}$$

which hold to decent approximation. Viani et al. (2017) recently pointed out that the $\nu_{\max}$ scaling relation can be improved by including a term for the mean molecular weight.

As stars evolve into giants, the assumption of homology breaks down more and more, leading to systematic errors as high as 15% (e.g., Gaulme et al. 2016). By comparison of theoretical red giant model mode frequencies with those given by the scaling relations, Guggenberger et al. (2016, 2017) developed metallicity-dependent and mass-dependent corrections to the $\Delta\nu$ scaling relation.

These scaling relations do not tell us about the age or evolution of the star. We saw previously that the small frequency separation probes the sound speed gradient, which is then an indicator on the main sequence of the conditions in the core, and therefore main-sequence age. The so-called C–D diagram shows the core-hydrogen abundance and stellar mass as a function of the frequency separations (Christensen-Dalsgaard 1984, see also Figure 1.26). If all stars had the solar abundances and solar mixing length, it would suffice to look up their mass and core-hydrogen abundance in this diagram. Since they do not, a more sophisticated approach is required.

Repeated Forward Modelling

A more involved approach to determining the properties of stars is through repeated forward modelling. Such an approach can also be applied to non-solar-like stars (e.g., evolved stars) where homology relations break down. These methods still make no attempt to determine the function M^{-1}. Though there are variations, they instead try to optimize the result of the forward operator against the observations:

$$[\hat{\mathbf{x}}, \hat{\tau}] = \underset{[\mathbf{x},\tau]}{\arg\min}\ [M(\mathbf{x},\tau) - \mathbf{y}]^T\, \mathbf{\Sigma}_\mathbf{y}^{-1}\, [M(\mathbf{x},\tau) - \mathbf{y}] \tag{1.134}$$

where $\hat{\cdot}$ means the optimal $\cdot$, and $\mathbf{\Sigma}_\mathbf{y}$ is the covariance matrix for the observations. There are several drawbacks with this approach:

Speed. This approach can be prohibitively slow, especially if new models need to be computed for each input, or if multiple input parameters are being optimized. This is often dealt with by applying additional assumptions to simplify the problem. For example, the mixing length parameter

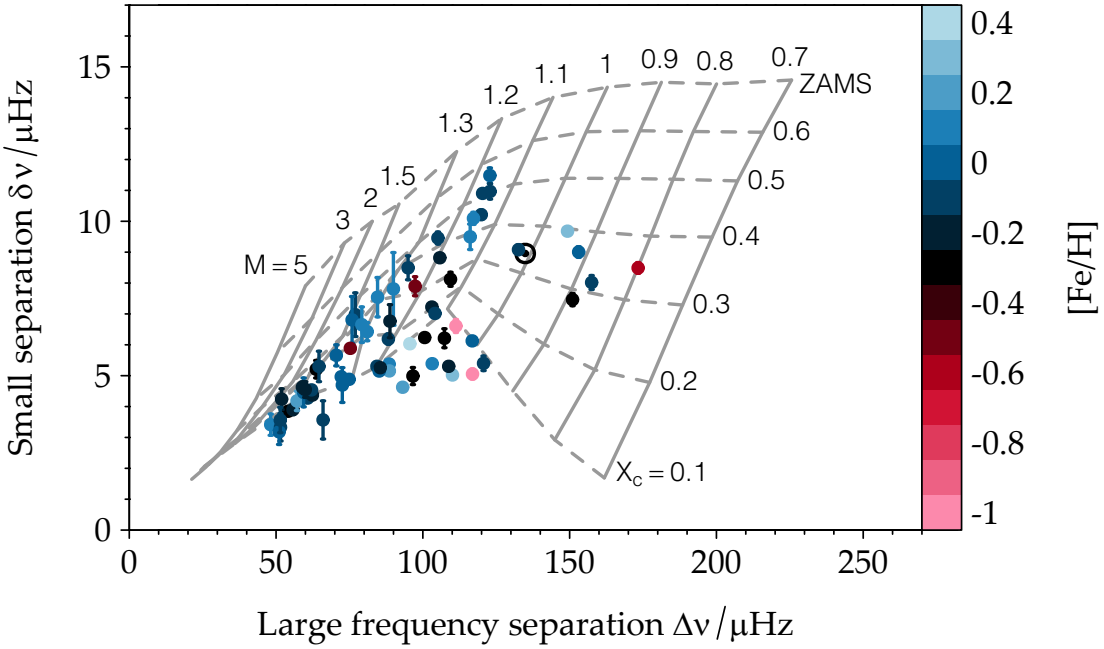


FIGURE 1.26. The C–D diagram. The small frequency separation is a proxy for core hydrogen abundance (X_c, dashed lines) through the sound speed gradient, and the large frequency separation is a proxy for stellar mass (M, solid lines) through the mean density. The gray lines are evolutionary simulations varied in their initial mass and evolved along the main sequence. The frequencies of the models have been calculated using GYRE (Townsend and Teitler 2013). Stars with $M \gtrapprox 1.8\ M_\odot$ do not have convective envelopes on the main sequence and are therefore not theoretically predicted to harbor solar-like oscillations. The points are LEGACY stars observed by *Kepler*, colored by their metallicity (Lund et al. 2017). Many of the stars fall off the diagram, thus illustrating its limitations as a look-up table for stellar properties. *Figure adapted from Bellinger et al. 2017a.*

can be kept fixed to the solar-calibrated value (e.g., Silva Aguirre et al. 2015, 2017). Another simplification is to calculate the initial helium abundance from the initial metallicity by assuming a galactic chemical evolution law (e.g., Silva Aguirre et al. 2015, 2017). This is usually achieved by fitting a line through to two points: the primordial helium abundance from models of Big Bang nucleosynthesis $[Y_p = 0.2463, Z_p = 0]$ (e.g., Coc et al. 2014) and the calibrated initial solar mixture, e.g., $[Y_{0,\odot} = 0.273, Z_{0,\odot} = 0.019]$, so $\Delta Y/\Delta Z \simeq 1.4$. The optimization is then performed over a limited set of input parameters (e.g., $[M, Z_0]$) and potentially on a pre-computed grid of models as well. However, the end result then has (typically unpropagated) systematic errors.

Local Minima. Commonly, iterative numerical optimization algorithms such as Levenberg–Marquardt (1944, 1963) and Nelder–Mead (1965) are applied for this task (e.g., Lebreton and Goupil 2014, Appourchaux et al.

2015). These approaches can have difficulty finding global minima of the solution.

There are also no currently known theoretical bounds on the complexity of a Nelder-Mead search (Singer and Singer 1999). It is however known that this algorithm scales poorly to high dimensions (e.g., Chen et al. 2015).

Redundancy. This approach implicitly assumes that each bit of observable information provides a fully independent constraint to the stellar model, and weights each observation only by its uncertainty. In reality, the observations have some degree of redundancy with respect to the aspects of the model that they constrain (Angelou & Bellinger et al. 2017, see also Chapter 3). Matching such an aspect of the model is then arbitrarily upweighted. Some practitioners deal with this problem by applying *ad hoc* weightings (e.g., Paxton et al. 2013).

We therefore seek an approach that naturally avoids these problems.

Random Forest Regression

In recent years, machine learning techniques have become increasingly popular for solving inverse problems (e.g., Rosasco et al. 2005, Fai et al. 2017, Adler and Öktem 2017). Some applications include automatic photograph coloration (Larsson et al. 2016), image reconstruction (e.g., Schlemper et al. 2017), and medical imaging (e.g., Prato and Zanni 2008, Jin et al. 2017). In fact, supervised learning itself can be viewed as an inverse problem (Vito et al. 2005).

In Chapter 2 we propose a solution to the evolution inversion problem based on machine learning. In particular, we use the variant of random forest regression (Breiman 2001) known as extremely randomized trees (Geurts et al. 2006) to learn the function M^{-1} from a dense grid of evolutionary simulations. Ensemble tree-based algorithms are known to be quick to train (especially because the task is 'embarrassingly' parallelizable), quick to predict (when the number of trees is not very large), and to have very good predictive performance (e.g., Caruana and Niculescu-Mizil 2006). Furthermore, the bootstrap aggregation ("bagging") that is performed helps with problem degeneracy and dimensionality (e.g., Skurichina and Duin 2002). Random forests can suffer from reduced performance if the number of redundant variables is large (Louppe 2014), however there are strategies to deal with this drawback (Tuv et al. 2009).

Louppe (2014) derived the worst-case time complexity of training extremely randomized trees to be $\mathcal{O}(MKN^2)$, where M is the number of trees, N is the number of samples, and K is the number of features that is randomly drawn at each node. In Chapter 2, we cross-validate M and find satisfactory performance at $M = 256$. The parameter K varies between 2 and 9, depending on the types of observations available for a given star.

To obtain the posterior distribution of solutions for an observed star with measurement uncertainties, we pass random instances of the observations perturbed by their uncertainties through the trained network. We have to choose

how many random instances that we will use. This number should be chosen such that the sample distribution converges to a reasonable degree to the population distribution. A useful way to quantify the differences in distributions is the Kullback-Leibler (KL) divergence, also known as relative entropy:

$$D_{KL}(P\|Q) = \int_{-\infty}^{\infty} p(x) \log \frac{p(x)}{q(x)} \, dx \tag{1.135}$$

where P and Q are two continuous random variables and p and q are their respective densities (Kullback and Leibler 1951). A low relative entropy indicates similarity.

We seek to determine how many random samples we need to generate in order for our posterior distributions to converge to a reasonable degree to their actual distributions. A proxy for this would be to determine the KL divergence between the normal distribution and sample normal distributions of varying sizes. Figure 1.27 shows an example of a standard normal distribution ψ and sample normal densities with different sample sizes. The figure furthermore shows the KL divergence of these sample normal distributions as a function of sample size, averaged over 1,000 random trials. The distribution converges around 10,000 samples. Thus, we propagate 10,000 random instances of the measurement uncertainty through the random forest. Applying the technique fleshed out in detail in Chapter 2 to 94 stars observed by *Kepler*, we find the estimates shown in Figure 1.28.

1.4.2 Structure Inversions

By solving the evolution inverse problem, we can obtain an evolutionary model for a given observed star. However, regardless of the technique used, the mode frequencies of best-fitting models generally fail to match one or more mode frequencies of the star—even after correcting surface effects. This implies that the structure of the star differs from the structure of the model. This is the starting

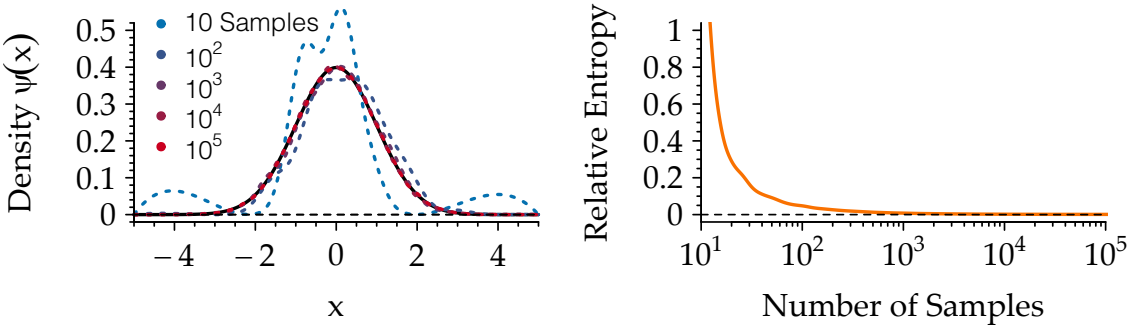


FIGURE 1.27. Left: Normal density distribution (black line) and example sample normal distributions for various sample sizes (dashed lines). Right: Average divergence of sample normal distributions from the standard normal distribution as a function of sample size.

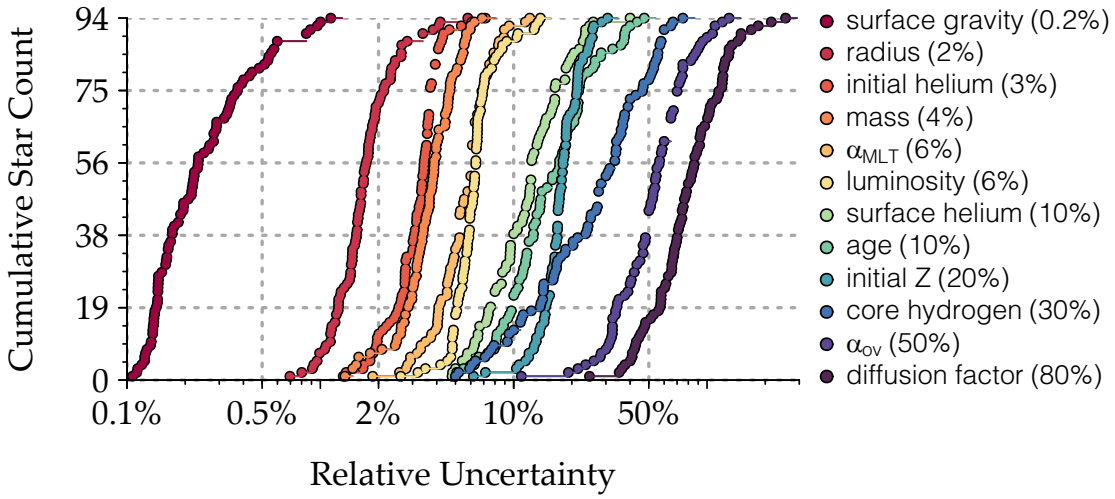


FIGURE 1.28. Cumulative distribution functions showing the relative uncertainties in estimated stellar parameters for 94 main-sequence stars. Each type of measurement is sorted by uncertainty. The numbers in parentheses in the legend give the median uncertainty. *Figure adapted from Bellinger et al. 2017a.*

point for the structure inversion problem. We seek to invert Equation (1.121) to infer $f_1(r)$ from observed mode frequencies, for some choice of f_1, by deducing the difference in f_1 between the best-fitting model and the star. This problem is difficult for multiple reasons:

Degeneracy. As the kernels reveal, a modification to the structure anywhere in a stellar model may cause several or all of its pulsation modes to shift in their frequency of oscillation, and each frequency may shift in a different way. Modifications to different locations in the stellar interior may also cause the same change to the frequency of a mode.

Furthermore, the mode frequencies are a function of multiple structural quantities. When trying to infer f_1, we must ensure that the results are not unduly influenced by f_2. With the present quality of asteroseismic data, this restricts us to kernel pairs with $f_2 = Y$ (recall Section 1.3.2).

Information Content. Whereas we are trying to measure a continuous function, which in principle may contain infinite information, we have only a finite set of mode frequencies with which to do it.

Furthermore, we will only be able to form well-localized averaging kernels in regions where a sufficient number of lower turning points are situated (recall Figure 1.7). This rules out some inversion methods.

Stability. The kernel functions are nearly linearly dependent, and so the problem is ill-conditioned. Even if the measurements of the mode frequencies were certain, an exact fit to mode frequencies yields highly oscillatory, non-physical solutions (see, e.g., Dziembowski et al. 1990).

Surface Effects. All of the modes are sensitive to the outermost layers of the star, where our assumptions break down (recall Sections 1.2 and 1.3). Thus, we must take special care to suppress surface effects. However, there may be additional surface effects that the present treatment do not suppress. The treatment of the surface term may furthermore erroneously subtract off more than just surface effects.

Uniqueness. The solutions are not unique. From any solution to the inverse problem, a different solution can be generated (see Gough and Thompson 1991 for a discussion).

As discussed in the first section, inversion of asteroseismic data presents some novel challenges over helioseismic inversions (e.g., Basu 2014). Unlike in helioseismology, in which the solar mass and radius are known to high precision, the masses and radii of solar-type oscillating stars are generally uncertain by at least a percent (see e.g., White et al. 2013, Silva Aguirre et al. 2015, Bellinger et al. 2016, see also Figure 1.28). Although seemingly small, such uncertainties in stellar mass and radius are generally about two orders of magnitude greater than the uncertainties in oscillation mode frequencies. The number of observed oscillation modes is also much smaller, and the inner radii at which these modes turn around is much more limited as well.

The most 'obvious' way to invert Equation (1.121) would be via a least squares fit to the entire unknown profile. That is: replace the functions to be estimated by linear basis functions (e.g., cubic B-splines, de Boor 1972), and then select the coefficients of the basis functions such that the residuals are minimized (e.g., Basu and Chaplin 2017). However, this approach yields oscillatory and nonphysical solutions. One can then seek a regularized solution by applying, e.g., the O'Sullivan penalty (O'Sullivan et al. 1986). This is a fruitful approach in global helioseismology (e.g., Dziembowski et al. 1990), where there is enough information to resolve the majority of the solar interior, to disentangle f_1 from f_2, and to suppress the surface term. For stars, however, there is just not enough information in current observational data for this technique to work.

The technique of Optimally Localized Averages (OLA, Backus and Gilbert 1968, 1970) provides a path forward. As discussed in Section 1.1, the idea of OLA is to linearly combine the modes in such a way that their combination is only sensitive to perturbations in one region in the star. Then, if the frequencies of that combination differ between model and star, then the structure of the star differs in that location.

There are two variants of OLA that appear in the literature: Multiplicative OLA (MOLA), which is based on the original Backus–Gilbert formulation; and Subtractive OLA (SOLA, Pijpers and Thompson 1992, 1994), which was introduced in helioseismology to reduce computational costs. Whereas MOLA requires a matrix inversion at each radius where an averaging kernel is sought (which, as we will see, is computationally intensive), SOLA can use the same matrix inversion for all target radii. This speed-up comes at the cost of an addi-

tional free parameter. We use SOLA to solve the structure inversion problem in Chapter 4.

To solve the SOLA problem, we must find the coefficients c that form the linear combination corresponding to (I) a well-localized averaging kernel, (II) a small cross-term kernel, (III) a reasonably suppressed surface term, and (IV) suitably small uncertainties. We thus seek to find the coefficients c that minimize

$$\int \left(\sum_i c_i K_i^{(f_1,f_2)} - T(r; r_0, \Delta) \right)^2 dr + \beta \int \left(\sum_i K_i^{(f_2,f_1)} \right)^2 dr + \mu \sum_{i,j} c_i c_j E_{i,j} \quad (1.136)$$

where β is a parameter controlling the cross-term kernel, μ is a parameter controlling the data uncertainties, and E is the error covariance matrix. The function we wish the averaging kernel at the target radius r_0 to approximate is called the "target kernel," which I have denoted T. It may be chosen for example to resemble a localized Gaussian.

Minimizing this functional amounts to solving the matrix equation $\mathbf{A}\boldsymbol{x} = \mathbf{b}$ that is shown in Equation (1.141), where $\mathbf{A}$ is a symmetric $(N+3) \times (N+3)$ matrix with N being the number of observed modes. In this matrix I have introduced

$$\begin{aligned} \mathcal{A}_{i,j} = &\int K_i^{(f_1,f_2)} \cdot K_j^{(f_1,f_2)} \, dr \\ &+ \beta \int K_i^{(f_2,f_1)} \cdot K_j^{(f_2,f_1)} \, dr \; + \; \mu E_{i,j} \end{aligned} \quad (1.137)$$

$$y_i = \int K_i^{(f_1,f_2)}(r) \cdot T(r; r_0, \Delta) \, dr. \quad (1.138)$$

Furthermore, I have introduced the Lagrange multipliers λ_1, λ_2, and λ_3 to normalize the averaging kernel and to suppress the surface term. Given choices of the parameters β, μ, and Δ, the matrix $\mathbf{A}$ may be inverted to yield $\mathbf{A}^{-1}\mathbf{b} = \mathbf{x}$, from which we may deduce $\mathbf{c}(r_0)$ and hence $f_1(r_0)$. Rabello-Soares et al. (1998, 1999) examined the influence of each of these parameters (β, μ, Δ) on the inversion result. In Chapter 4 we introduce a heuristic algorithm to choose these parameters. For further details on OLA inversions in helio/asteroseismology, see e.g., Basu and Chaplin (2017).

As discussed earlier, the matrix $\mathbf{A}$ is ill-conditioned, and so special care must be taken when trying to obtain the least-squares solution for $\mathbf{x}$ from Equation (1.141). Since $\mathbf{A}$ is symmetric, we can use the LDL^T decomposition (e.g., Banerjee and Roy 2014), which gives

$$\mathbf{A} = \mathbf{L}\mathbf{D}\mathbf{L}^T \quad (1.139)$$

$$
\underbrace{\begin{array}{c c}
 & \begin{array}{cccccc} j=1 & \dots & N & N+1 & N+2 & N+3 \end{array} \\
\begin{array}{c} i=1 \\ \vdots \\ N \\ 1+N \\ 2+N \\ 3+N \end{array} &
\begin{pmatrix}
\mathcal{A}_{1,1} & \cdots & \mathcal{A}_{1,N} & \int K_1^{(f_1,f_2)}\,dr & (\nu_1/\nu_{ac})^{-2}/I_1 & (\nu_1/\nu_{ac})^{2}/I_1 \\
\vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\
\mathcal{A}_{N,1} & \cdots & \mathcal{A}_{N,N} & \int K_N^{(f_1,f_2)}\,dr & (\nu_N/\nu_{ac})^{-2}/I_N & (\nu_N/\nu_{ac})^{2}/I_N \\
\int K_1^{(f_1,f_2)}\,dr & \cdots & \int K_N^{(f_1,f_2)}\,dr & 0 & 0 & 0 \\
(\nu_1/\nu_{ac})^{-2}/I_1 & \cdots & (\nu_N/\nu_{ac})^{-2}/I_N & 0 & 0 & 0 \\
(\nu_1/\nu_{ac})^{2}/I_1 & \cdots & (\nu_N/\nu_{ac})^{2}/I_N & 0 & 0 & 0
\end{pmatrix}
\end{array}}_{A}
\underbrace{\begin{pmatrix} c_1 \\ \vdots \\ c_N \\ \lambda_1 \\ \lambda_2 \\ \lambda_3 \end{pmatrix}}_{x}
=
\underbrace{\begin{pmatrix} y_1 \\ \vdots \\ y_N \\ 1 \\ 0 \\ 0 \end{pmatrix}}_{b}
\tag{1.141}
$$

where $\mathbf{D}$ is a square diagonal matrix with entries $\mathbf{D} = \mathrm{diag}(d_1, d_2, \ldots d_{N+3})$; and $\mathbf{L}$ is a lower unitriangular matrix, i.e. a matrix of the form

$$\begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ L_{2,1} & 1 & 0 & \cdots & 0 \\ L_{3,1} & L_{3,2} & 1 & \ddots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ L_{n,1} & L_{n,2} & \cdots & L_{n,m-1} & 1 \end{pmatrix}. \tag{1.140}$$

Substituting the LDL^T decomposition of $\mathbf{A}$ into our matrix equation, we get

$$\begin{aligned} &\mathbf{L}\mathbf{D}\mathbf{L}^\mathsf{T}\mathbf{x} = \mathbf{b} \\ &\quad\Rightarrow\ \mathbf{x} \simeq \mathbf{L}\mathbf{D_0}^{-1}\mathbf{L}^\mathsf{T}\mathbf{b}. \end{aligned} \tag{1.142}$$

Since $\mathbf{A}$ is ill-conditioned and hence many of its entries are very nearly zero, I have introduced the pseudo-inverse for the diagonal matrix $\mathbf{D_0}$, which gives

$$\mathbf{D_0}^{-1} = \mathrm{diag}\,(\delta_1, \delta_2, \ldots \delta_{N+3}) \qquad \text{where} \qquad \delta_i = \begin{cases} 1/d_i & \text{if } |d_i| > t \\ 0 & \text{otherwise} \end{cases} \tag{1.143}$$

where t is a small threshold (e.g., machine precision). In this work, I calculate the LDL^T decomposition using CHOLMOD (Chen et al. 2008). The cost to obtain this solution is as follows:

- LDL^T decomposition: $\mathcal{O}(N^3)$ (Krishnamoorthy and Menon 2013)
- conversion and inversion of the diagonal matrix: $\mathcal{O}(N)$
- multiplication of the matrix factors: $\mathcal{O}(N^6)$ (although there are more efficient algorithms, e.g., Coppersmith and Winograd 1990)

where I have here made use of the fact that the matrix is square. Hence, the total time complexity is dominated by the final step, yielding $\mathcal{O}(N^6)$.

1.5 Summary of Thesis

To conclude the introduction, I will now summarize the ten most important aspects of this thesis:

Chapter 2 (Bellinger et al. 2016)

1. We introduce a new method based on machine learning for precisely determining the ages, masses, radii, and other properties of main sequence stars within seconds. We test this method extensively, including cross-validation, hare-and-hound exercises, on the Sun, and on well-studied stars.

2. We apply this method to measure properties of solar-like stars whose frequencies have been resolved using data from *Kepler*. We find age, mass, and radius estimates with uncertainties on the order of 6%, 2%, and 1%, respectively.

3. We use this method to recover a diffusion–mass relation, which demonstrates the promise of using this approach to empirically uncover relationships in stellar physics.

Chapter 3 (Angelou & Bellinger et al. 2017)

4. We systematically investigate the properties of stellar models and determine which kinds of observations of stars are important for constraining unobservable aspects of stars, such as their ages. We find that metallicity measurements are independent and indispensable constraints to stellar models. We furthermore quantify the increase in uncertainty for each stellar parameter that arises from increases in uncertainty of the observational data.

5. We analyze the expected asteroseismic yield of the forthcoming space missions TESS and PLATO for solar-like stars. We find that with typical TESS data, we will be able to determine the mass and radius of a Sun-like star to better than 5% uncertainty. This precision will be indispensable in the search for Earth twins.

Chapter 4 (Bellinger et al. 2017b)

6. We introduce an algorithm for inverting asteroseismic data to measure stellar structure, which takes care of imprecise radius and mass estimations and includes the automated determination of inversion parameters.

7. We apply our method of asteroseismic structure inversions to measure the internal isothermal speeds of sound in the cores of the solar twins 16 Cyg A and B.

8. We find that in the case of 16 Cyg B, the asteroseismic structure of the star is in good agreement with the best-fitting evolutionary model. In the case of 16 Cyg A, however, we find less agreement.

Future Prospects

9. We solve the structure inverse problem for 18 more stars, finding even greater disagreements with theoretical models of solar interiors, even when considering a variety of physics inputs. These results seem to indicate that there are improvements needed in our understanding of stellar physics.

10. We follow the evolution of the stellar structure kernels past core hydrogen exhaustion and into the sub-giant phase of evolution. We find much greater sensitivity to the deep stellar core, indicating there may soon be the prospect of learning more about the deep interior of another star than we even know about our own Sun.

2

Fundamental Parameters of Main Sequence Stars in an Instant with Machine Learning

The contents of this chapter were authored by E. P. Bellinger, G. C. Angelou, S. Hekker, S. Basu, W. H. Ball, and E. Guggenberger and published in October of 2016 in *The Astrophysical Journal*, 830 (1), 31.[1]

Chapter Summary

Owing to the remarkable photometric precision of space observatories like *Kepler*, stellar and planetary systems beyond our own are now being characterized *en masse* for the first time. These characterizations are pivotal for endeavors such as searching for Earth-like planets and solar twins, understanding the mechanisms that govern stellar evolution, and tracing the dynamics of our Galaxy. The volume of data that is becoming available, however, brings with it the need to process this information accurately and rapidly. While existing methods can constrain fundamental stellar parameters such as ages, masses, and radii from these observations, they require substantial computational efforts to do so.

We develop a method based on machine learning for rapidly estimating fundamental parameters of main-sequence solar-like stars from classical and asteroseismic observations. We first demonstrate this method on a hare-and-hound exercise and then apply it to the Sun, 16 Cyg A & B, and 34 planet-hosting candidates that have been observed by the *Kepler* spacecraft. We find that our estimates and their associated uncertainties are comparable to the results of other methods, but with the additional benefit of being able to explore many more stellar parameters while using much less computation time. We furthermore use this method to present evidence for an empirical diffusion-mass relation. Our method is open source and freely available for the community to use.[2]

[1] Contribution statement: The work of this chapter was carried out by me; the text was mainly written by me, with contributions from G. C. Angelou, in collaboration with the other authors.

[2] The source code for all analyses and for all figures appearing in this chapter can be found electronically at `https://github.com/earlbellinger/asteroseismology` (Bellinger 2016).

2.1 Introduction

In recent years, dedicated photometric space missions have delivered dramatic improvements to time-series observations of solar-like stars. These improvements have come not only in terms of their precision, but also in their time span and sampling, which has thus enabled direct measurement of dynamical stellar phenomena such as pulsations, binarity, and activity. Detailed measurements like these place strong constraints on models used to determine the ages, masses, and chemical compositions of these stars. This in turn facilitates a wide range of applications in astrophysics, such as testing theories of stellar evolution, characterizing extrasolar planetary systems (e.g. Campante et al. 2015, Silva Aguirre et al. 2015), assessing galactic chemical evolution (e.g. Chiappini et al. 2015), and performing ensemble studies of the Galaxy (e.g. Chaplin et al. 2011, Miglio et al. 2013, Chaplin et al. 2014).

The motivation to increase photometric quality has in part been driven by the goal of measuring oscillation modes in stars that are like our Sun. Asteroseismology, the study of these oscillations, provides the opportunity to constrain the ages of stars through accurate inferences of their interior structures. However, stellar ages cannot be measured directly; instead, they depend on indirect determinations via stellar modelling.

Traditionally, to determine the age of a star, procedures based on iterative optimization (hereinafter IO) seek the stellar model that best matches the available observations (Brown et al. 1994). Several search strategies have been employed, including exploration through a pre-computed grid of models (i.e. grid-based modelling, hereinafter GBM; see Gai et al. 2011, Chaplin et al. 2014); or *in situ* optimization (hereinafter ISO) such as genetic algorithms (Metcalfe et al. 2014), Markov-chain Monte Carlo (Bazot et al. 2012), or the downhill simplex algorithm (Paxton et al. 2013; see e.g. Silva Aguirre et al. 2015 for an extended discussion on the various methods of dating stars). Utilizing the detailed observations from the *Kepler* and CoRoT space telescopes, these procedures have constrained the ages of several field stars to within 10% of their main-sequence lifetimes (Silva Aguirre et al. 2015).

IO is computationally intensive in that it demands the calculation of a large number of stellar models (see Metcalfe et al. 2009 for a discussion). ISO requires that new stellar tracks are calculated for each target, as they do not know *a priori* all of the combinations of stellar parameter values that the optimizer will need for its search. They furthermore converge to local minima and therefore need to be run multiple times from different starting points to attain global coverage. GBM by way of interpolation in a high-dimensional space, on the other hand, is sensitive to the resolution of each parameter and thus requires a very fine grid of models to search through (see e.g. Quirion et al. 2010, who use more than five million models that were varied in just four initial parameters). Additional dimensions such as efficiency parameters (e.g. overshooting or mixing length parameters) significantly impact on the number of models needed and hence the search times for these methods. As a consequence, these approaches typically

use, for example, a solar-calibrated mixing length parameter or a fixed amount of convective overshooting. Since these values in other stars are unknown, keeping them fixed therefore results in underestimations of uncertainties. This is especially important in the case of atomic diffusion, which is essential when modelling the Sun (see e.g. Basu and Antia 1994), but is usually disabled for stars with $M/M_{\odot} > 1.4$ because it leads to the unobserved consequence of a hydrogen-only surface (Morel and Thévenin 2002).

These concessions have been made because the relationships connecting observations of stars to their internal properties are non-linear and difficult to characterize. Here we will show that through the use of machine learning, it is possible to avoid these difficulties by capturing those relations statistically and using them to construct a regression model capable of relating observations of stars to their structural, chemical, and evolutionary properties. The relationships can be learned using many fewer models than IO methods require, and can be used to process entire stellar catalogs with a cost of only seconds per star.

To date, only about a hundred solar-like oscillators have had their frequencies resolved, allowing each of them be modelled in detail using costly methods based on IO. In the forthcoming era of TESS (Ricker et al. 2015) and PLATO (Rauer et al. 2014), however, seismic data for many more stars will become available, and it will not be possible to dedicate large amounts of supercomputing time to every star. Furthermore, for many stars, it will only be possible to resolve *global* asteroseismic quantities rather than individual frequencies. Therefore, the ability to rapidly constrain stellar parameters for large numbers of stars by means of global oscillation analysis will be paramount.

In this work, we consider the constrained multiple-regression problem of inferring fundamental stellar parameters from observable quantities. We construct a random forest of decision tree regressors to learn the relationships connecting observable quantities of main-sequence (MS) stars to their zero-age main-sequence (ZAMS) histories and current-age structural and chemical attributes. We validate our technique by inferring the parameters of simulated stars in a hare-and-hound exercise, the Sun, and the well-studied stars 16 Cyg A and B. Finally, we conclude by applying our method on a catalog of *Kepler* objects-of-interest (hereinafter KOI; Davies et al. 2016).

We explore various model physics by considering stellar evolutionary tracks that are varied not only in their initial mass and chemical composition, but also in their efficiency of convection, extent of convective overshooting, and strength of gravitational settling. We compare our results to the recent findings from GBM (Silva Aguirre et al. 2015), ISO (Metcalfe et al. 2015), interferometry (White et al. 2013), and asteroseismic glitch analyses (Verma et al. 2014b) and find that we obtain similar estimates but with orders-of-magnitude speed-ups.

2.2 Method

We seek a multiple-regression model capable of characterizing observed stars. To obtain such a model, we build a matrix of evolutionary simulations and use machine learning to discover relationships in the stellar models that connect observable quantities of stars to the model quantities that we wish to predict. The matrix is structured such that each column contains a different stellar quantity and each row contains a different stellar model. We construct this matrix by extracting models along evolutionary sequences (see Appendix 2.6.1 for details on the model selection process) and summarizing them to yield the same types of information as the stars being observed. Although each star (and each stellar model) may have a different number of oscillation modes observed, it is possible to condense this information into only a few numbers by leveraging the fact that the frequencies of these modes follow a regular pattern (for a review of solar-like oscillations, see Chaplin and Miglio 2013). Once the machine has processed this matrix, one can feed the algorithm a catalogue of stellar observations and use it to predict the fundamental parameters of those stars.

The observable information obtained from models that can be used to inform the algorithm may include, but is not limited to, combinations of temperatures, metallicities, global oscillation information, surface gravities, luminosities, and/or radii. From these, the machine can learn how to infer stellar parameters such as ages, masses, core hydrogen and surface helium abundances. If luminosities, surface gravities, and/or radii are not supplied, then they may be predicted as well. In addition, the machine can also infer evolutionary parameters such as the initial stellar mass and initial chemical compositions as well as the mixing length parameter, overshoot coefficient, and diffusion multiplication factor needed to reproduce observations, which are explained in detail below.

2.2.1 Model Generation

We use the open-source 1D stellar evolution code *Modules for Experiments in Stellar Astrophysics* (MESA; Paxton et al. 2011) to generate main-sequence stellar models from solar-like evolutionary tracks varied in initial mass M, helium Y_0, metallicity Z_0, mixing length parameter $\alpha_{\rm MLT}$, overshoot coefficient $\alpha_{\rm ov}$, and diffusion multiplication factor D. The diffusion multiplication factor serves to amplify or diminish the effects of diffusion, where a value of zero turns it off and a value of two doubles all velocities. The initial conditions are varied in the ranges $M \in [0.7, 1.6]$ $M_\odot$, $Y_0 \in [0.22, 0.34]$, $Z_0 \in [10^{-5}, 10^{-1}]$ (varied logarithmically), $\alpha_{\rm MLT} \in [1.5, 2.5]$, $\alpha_{\rm ov} \in [10^{-4}, 1]$ (varied logarithmically), and $D \in [10^{-6}, 10^{2}]$ (varied logarithmically). We put a cut-off of 10^{-3} and 10^{-5} on $\alpha_{\rm ov}$ and D, respectively, below which we consider them to be zero and disable them. The initial parameters of each track are chosen in a quasi-random fashion so as to populate the initial-condition hyperspace as homogeneously and rapidly as possible (shown in Figure 2.1; see Appendix 2.6.2 for more details).

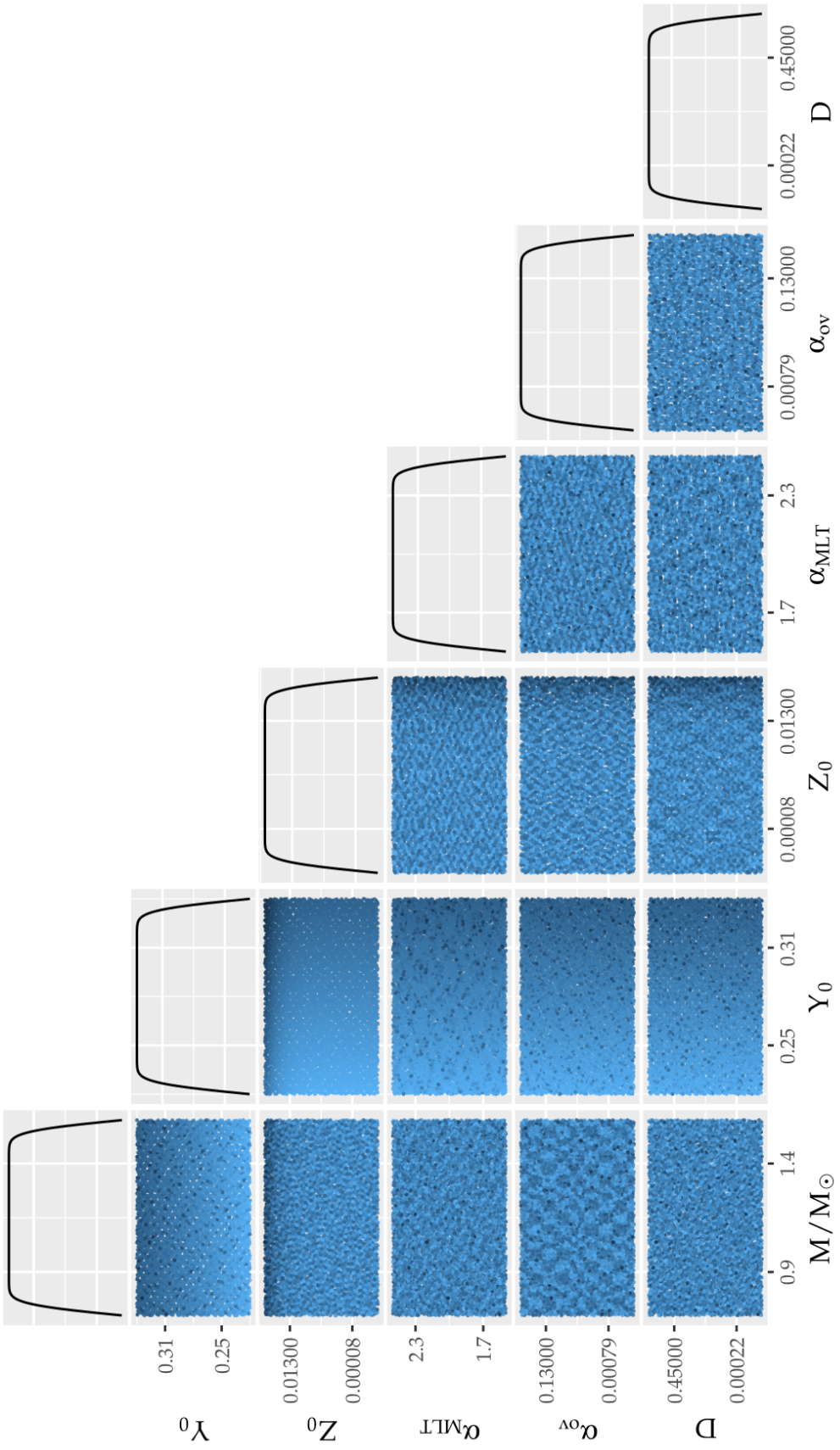


FIGURE 2.1. (Caption on other page.)

We use MESA version r8118 with the Helmholtz-formulated equation of state that allows for radiation pressure and interpolates within the 2005 update of the OPAL EOS tables (Rogers and Nayfonov 2002). We assume a Grevesse and Sauval (1998) solar composition for our initial abundances and opacity tables. Since we restrict our study to the main sequence, we use an eight-isotope nuclear network consisting of ^{1}H, ^{3}He, ^{4}He, ^{12}C, ^{14}N, ^{16}O, ^{20}Ne, and ^{24}Mg. We use a step function for overshooting and set a scaling factor $f_0 = \alpha_{ov}/5$ to determine the radius $r_0 = H_p \cdot f_0$ inside the convective zone at which convection switches to overshooting, where H_p is the pressure scale height. The overshooting parameter applies to all convective boundaries and is kept fixed throughout the course of a track's evolution, so a non-zero value does not imply that the model has a convective core at any specific age. All pre-main-sequence (PMS) models are calculated with a simple photospheric approximation, after which an Eddington $T - \tau$ atmosphere is appended on at ZAMS. We call ZAMS the point at which the nuclear luminosity of the models make up 99.9% of the total luminosity. We calculate atomic diffusion with gravitation settling and without radiative levitation on the main sequence using five diffusion class representatives: ^{1}H, ^{3}He, ^{4}He, ^{16}O, and ^{56}Fe (Burgers 1969).[3] Following their most recent measurements, we correct the defaults in MESA of the gravitational constant ($G = 6.67408 \times 10^{-8}$ g^{-1} cm^3 s^{-2}; Mohr et al. 2016), the gravitational mass of the Sun ($M_\odot = 1.988475 \times 10^{33}$ g $= \mu G^{-1} = 1.32712440042 \times 10^{11}$ km s^{-1} G^{-1}, where μ is the standard gravitational parameter; Pitjeva 2015), and the solar radius ($R_\odot = 6.95568 \times 10^{10}$ cm; Haberreiter et al. 2008).

Each track is evolved from ZAMS to either an age of $\tau = 16$ Gyr or until terminal-age main sequence (TAMS), which we define as having a fractional core hydrogen abundance (X_c) below 10^{-3}. Evolutionary tracks with efficient heavy-element settling can develop discontinuities in their surface abundances if they lack sufficient model resolution. We implement adaptive remeshing by recomputing any track with abundance discontinuities in its surface layers using finer spatial and temporal resolutions (see Appendix 2.6.3 for details). Running stellar physics codes in a batch mode like this requires care, so we manually

FIGURE 2.1. Scatterplot matrix (lower panels) and density plots (diagonal) of evolutionary track initial conditions considered. Mass (M), initial helium (Y_0), initial metallicity (Z_0), mixing length parameter (α_{MLT}), overshoot (α_{ov}), and diffusion multiplication factor (D) were varied in a quasi-random fashion to obtain a low-discrepancy grid of model tracks. Points are colored by their initial hydrogen $X_0 = 1 - Y_0 - Z_0$, with blue being high X_0 ($\approx 78\%$) and black being low X_0 ($\approx 56\%$). The parameter space is densely populated with evolutionary tracks of maximally different initial conditions.

[3] The atomic number of each representative isotope is used to calculate the diffusion rate of the other isotopes allocated to that group; see Paxton et al. (2011).

inspect multiple evolutionary diagnostics to ensure that proper convergence has been achieved.

2.2.2 Calculation of Seismic Parameters

We use the ADIPLS pulsation package (Christensen-Dalsgaard 2008) to compute p-mode oscillations up to spherical degree $\ell = 3$ below the acoustic cut-off frequency. We use on average of around 4,000 points per stellar model and therefore have adequate resolution to calculate frequencies without remeshing. We denote any frequency separation S as the difference between a frequency ν of spherical degree ℓ and radial order n and another frequency, that is:

$$S_{(\ell_1,\ell_2)}(n_1, n_2) \equiv \nu_{\ell_1}(n_1) - \nu_{\ell_2}(n_2). \tag{2.1}$$

The large frequency separation is then

$$\Delta\nu_\ell(n) \equiv S_{(\ell,\ell)}(n, n-1) \tag{2.2}$$

and the small frequency separation is

$$\delta\nu_{(\ell,\ell+2)}(n) \equiv S_{(\ell,\ell+2)}(n, n-1). \tag{2.3}$$

Near-surface layers of stars are poorly-modeled, which induces systematic frequency offsets (see e.g. Rosenthal et al. 1999). The ratios between the large and small frequency separations (Equation 2.4), and also between the large frequency separation and five-point-averaged frequencies (Equation 2.5) have been shown to be less sensitive to the surface term than the aforementioned separations and are therefore valuable asteroseismic diagnostics of stellar interiors (Roxburgh and Vorontsov 2003). They are defined as

$$r_{(\ell,\ell+2)}(n) \equiv \frac{\delta\nu_{(\ell,\ell+2)}(n)}{\Delta\nu_{(1-\ell)}(n+\ell)} \tag{2.4}$$

$$r_{(\ell,1-\ell)}(n) \equiv \frac{dd_{(\ell,1-\ell)}(n)}{\Delta\nu_{(1-\ell)}(n+\ell)} \tag{2.5}$$

where

$$\begin{aligned} dd_{0,1}(n) \equiv \frac{1}{8}\big[&\nu_0(n-1) - 4\nu_1(n-1) + 6\nu_0(n) \\ &- 4\nu_1(n) + \nu_0(n+1)\big] \end{aligned} \tag{2.6}$$

$$\begin{aligned} dd_{1,0}(n) \equiv -\frac{1}{8}\big[&\nu_1(n-1) - 4\nu_0(n) + 6\nu_1(n) \\ &- 4\nu_0(n+1) + \nu_1(n+1)\big]. \end{aligned} \tag{2.7}$$

Since the set of radial orders that are observable differs from star to star, we collect global statistics on $\Delta\nu_0$, $\delta\nu_{0,2}$, $\delta\nu_{1,3}$, $r_{0,2}$, $r_{1,3}$, $r_{0,1}$, and $r_{1,0}$. We mimic the range of observable frequencies in our models by weighting all frequencies

by their position in a Gaussian envelope centered at the predicted frequency of maximum oscillation power ν_{max} and having full-width at half-maximum of $0.66 \cdot \nu_{max}{}^{0.88}$ as per the prescription given by Mosser et al. (2012). We then calculate the weighted median of each variable, which we denote with angled parentheses (e.g. $\langle r_{0,2} \rangle$). We choose the median rather than the mean because it is a robust statistic with a high breakdown point, meaning that it is much less sensitive to the presence of outliers (for a discussion of breakdown points, see Hampel 1971, who attributed them to Gauss). This approach allows us to predict the fundamental stellar parameters of any solar-like oscillator with multiple observed modes irrespective of which exact radial orders have been detected. Illustrations of the methods used to derive the frequency separations and ratios of a stellar model are shown in Figure 2.2.

2.2.3 Training the Random Forest

We train a random forest regressor on our matrix of evolutionary models to discover the relations that facilitate inference of stellar parameters from observed quantities. A schematic representation of the topology of our random forest regressor can be seen in Figure 2.3. Random forests arise in machine learning through the family of algorithms known as CART, i.e. Classification and Regression Trees. There are several good textbooks that discuss random forests (see e.g. Hastie et al. 2009, Chapter 15). A random forest is an ensemble regressor, meaning that it is composed of many individual components that each perform statistical regression, and the forest subsequently averages over the results from each component (Breiman 2001). The components of the ensemble are decision trees, each of which learns a set of decision rules for relating observable quantities to stellar parameters. An ensemble approach is preferred because using only a single decision tree that is able to see all of the training data may result in a regressor that has memorized the training data and is therefore unable to generalize to as yet unseen values. This undesirable phenomenon is known in machine learning as over-fitting, and is analogous to fitting n data points using a degree n polynomial: the fit will work perfectly on the data that was used for fitting, but fail badly on any unseen data. To avoid this, each decision tree in the forest is given a random subset of the evolutionary models and a random subset of the observable quantities from which to build a set of rules relating observed quantities to stellar parameters. This process, known as statistical bagging (Hastie et al. 2009, Section 8.7), prevents the collection of trees from becoming over-fit to the training data, and thus results in a regression model that is capable of generalizing the information it has learned and predicting values for data on which it has not been trained.

Feature Importance

The CART algorithm uses information theory to decide which rule is the best choice for inferring stellar parameters like age and mass from the supplied in-

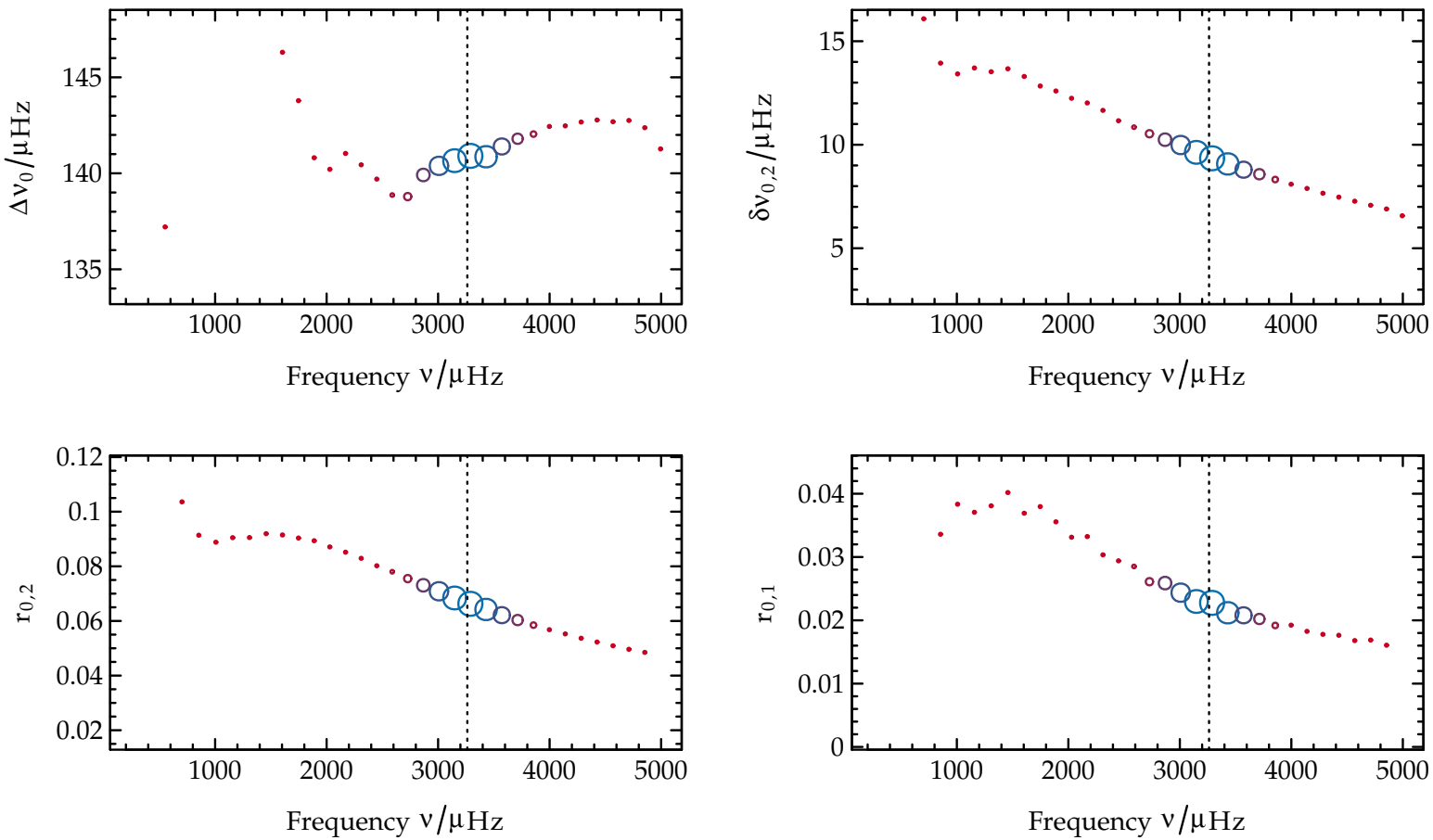


FIGURE 2.2. Calculation of seismic parameters for a stellar model. The large and small frequency separations $\Delta\nu_0$ (top left) and $\delta\nu_{0,2}$ (top right) and frequency ratios $r_{0,2}$ (bottom left) and $r_{0,1}$ (bottom right) are shown as a function of frequency. The vertical dotted line in these bottom four plots indicates ν_{max}. Points are sized and colored proportionally to the applied weighting, with large blue symbols indicating high weight and small red symbols indicating low weight.

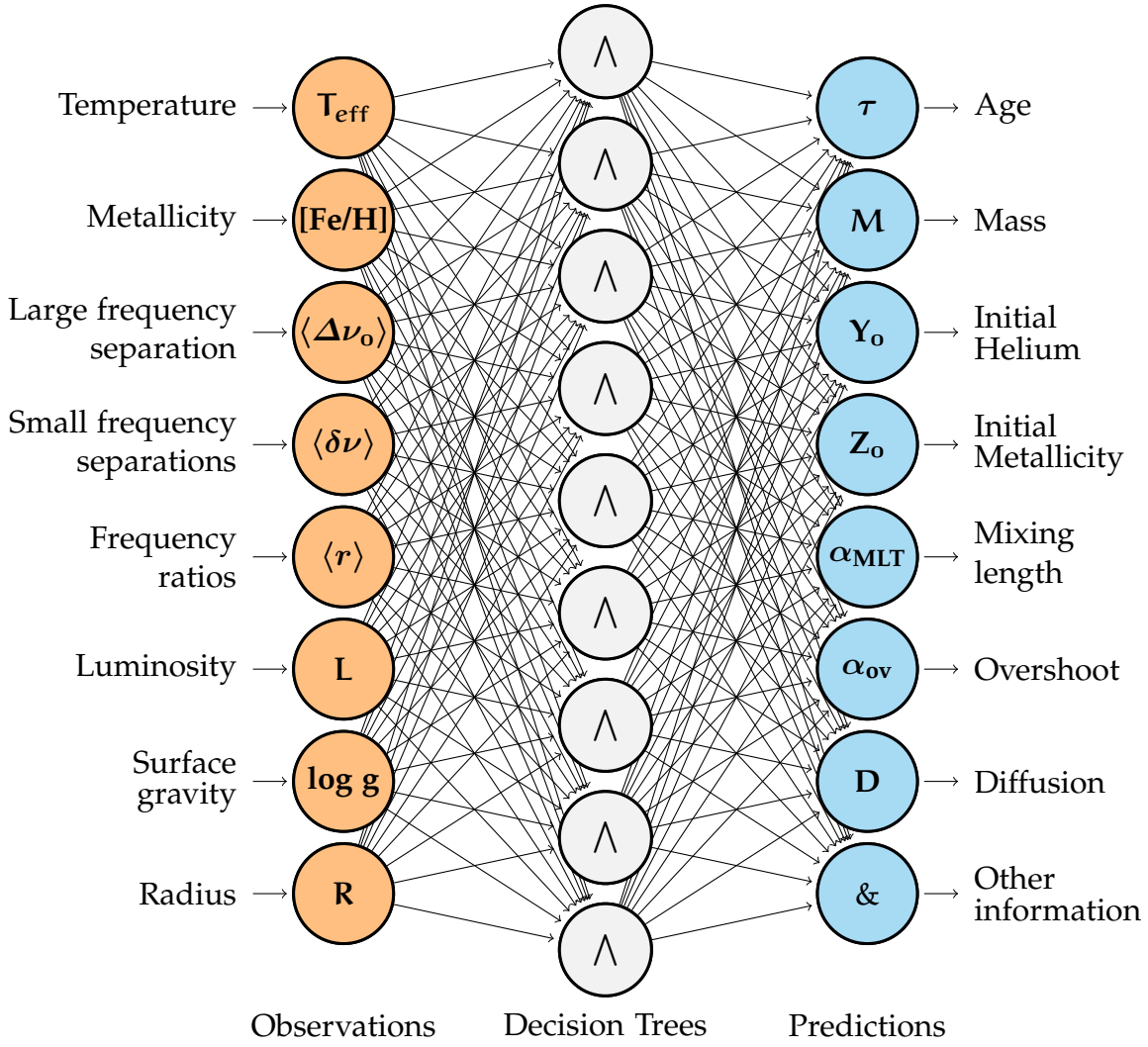


FIGURE 2.3. A schematic representation of a random forest regressor for inferring fundamental stellar parameters. Observable quantities such as T_{eff} and [Fe/H] and global asteroseismic quantities like $\langle \Delta\nu \rangle$ and $\langle \delta\nu_{0,2} \rangle$ are input on the left side. These quantities are then fed through to some number of hidden decision trees, which each independently predict parameters like age and mass. The predictions are then averaged and output on the right side. All inputs and outputs are optional. For example, surface gravities, luminosities, and radii are not always available from observations (e.g. with the KOI stars, see Section 2.3.3 below). In their absence, these quantities can be predicted instead of being supplied. In this case, those nodes can be moved over to the "prediction" side instead of being on the "observations" side. Also, in addition to potentially unobserved inputs like stellar radii, other interesting model parameters can be predicted as well, such as core hydrogen mass fraction or surface helium abundance.

formation (Hastie et al. 2009, Chapter 9). At every stage, the rule that creates the largest decrease in mean squared error (MSE) is crafted. A rule may be, for

example, "all models with $L < 0.4\ L_\odot$ have $M < 1\ M_\odot$." Rules are created until every stellar model that was supplied to that particular tree is fully explained by a sequence of decisions. We moreover use a variant on random forests known as *extremely* randomized trees (Geurts et al. 2006), which further randomize attribute splittings (e.g. split on L) and the location of the cut-point (e.g. split on $0.4\ L/L_\odot$) used when creating decision rules.

The process of constructing a random forest presents an opportunity for not only inferring stellar parameters from observations, but also for understanding the relationships that exist in the stellar models. Each decision tree explicitly ranks the relative "importance" of each observable quantity for inferring stellar parameters, where importance is defined in terms of both the reduction in MSE after defining a decision rule based on that quantity and the number of models that use that rule. In machine learning, the variables that have been measured and are supplied as inputs to the algorithm are known as "features." Figure 2.4 shows a feature importance plot, i.e. distributions of relative importance over all of the trees in the forest for each feature used to infer stellar parameters. The features that are used most often to construct decision rules are metallicity and temperature, which are each significantly more important features than the rest. The importance of [Fe/H] is due to the fact that the determinations of quantities like the Z_0 and D depend nearly entirely on it (see also Angelou et al. 2017). Note that importance does not indicate indispensability: an appreciable fraction of decision rules being made based off of one feature does not mean that another forest without that feature would not perform just as well. That being said, these results indicate that the best area to improve measurements would be in metallicity determinations, because for stars being predicted using this random forest, less precise values here means exploring many more paths and hence arriving at less certain predictions.

For many stars, stellar quantities such as radii, luminosities, surface gravities, and/or oscillation modes with spherical degree $\ell = 3$ are not available from observations. For example, the KOI data set (see Section 2.3.3 below) lacks all of this information, and the hare-and-hound exercise data (see Section 2.3.1 below) lack all of these except luminosities. We therefore must train random forests that predict those quantities instead of using them as features. We show the relative importance for the remaining features that were used to train these forests in Figure 2.5. When $\ell = 3$ modes and luminosities are omitted, effective temperature jumps in importance and ties with [Fe/H] as the most important feature.

Advantages of CART

We choose random forests over any of the many other non-linear regression routines (e.g. neural networks, support vector regression, etc.) for several reasons. First, random forests perform *constrained* regression; that is, they only make predictions within the boundaries of the supplied training data (see e.g. Hastie et al. 2009, Section 9.2.1). This is in contrast to other methods like neural networks, which ordinarily perform unconstrained regression and are therefore not pre-

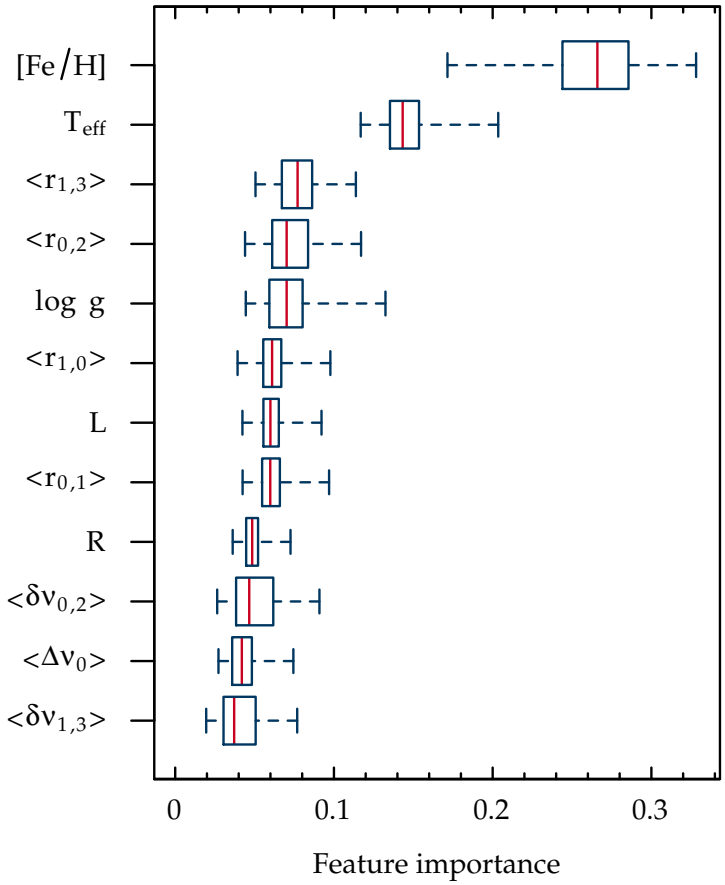


FIGURE 2.4. Box-and-whisker plots of relative importance for each observable feature in inferring fundamental stellar parameters as measured by a random forest regressor grown from a grid of evolutionary models. The boxes display the first (16%) and third (84%) quartile of feature importance over all trees, the center line indicates the median, and the whiskers extend to the most extreme values.

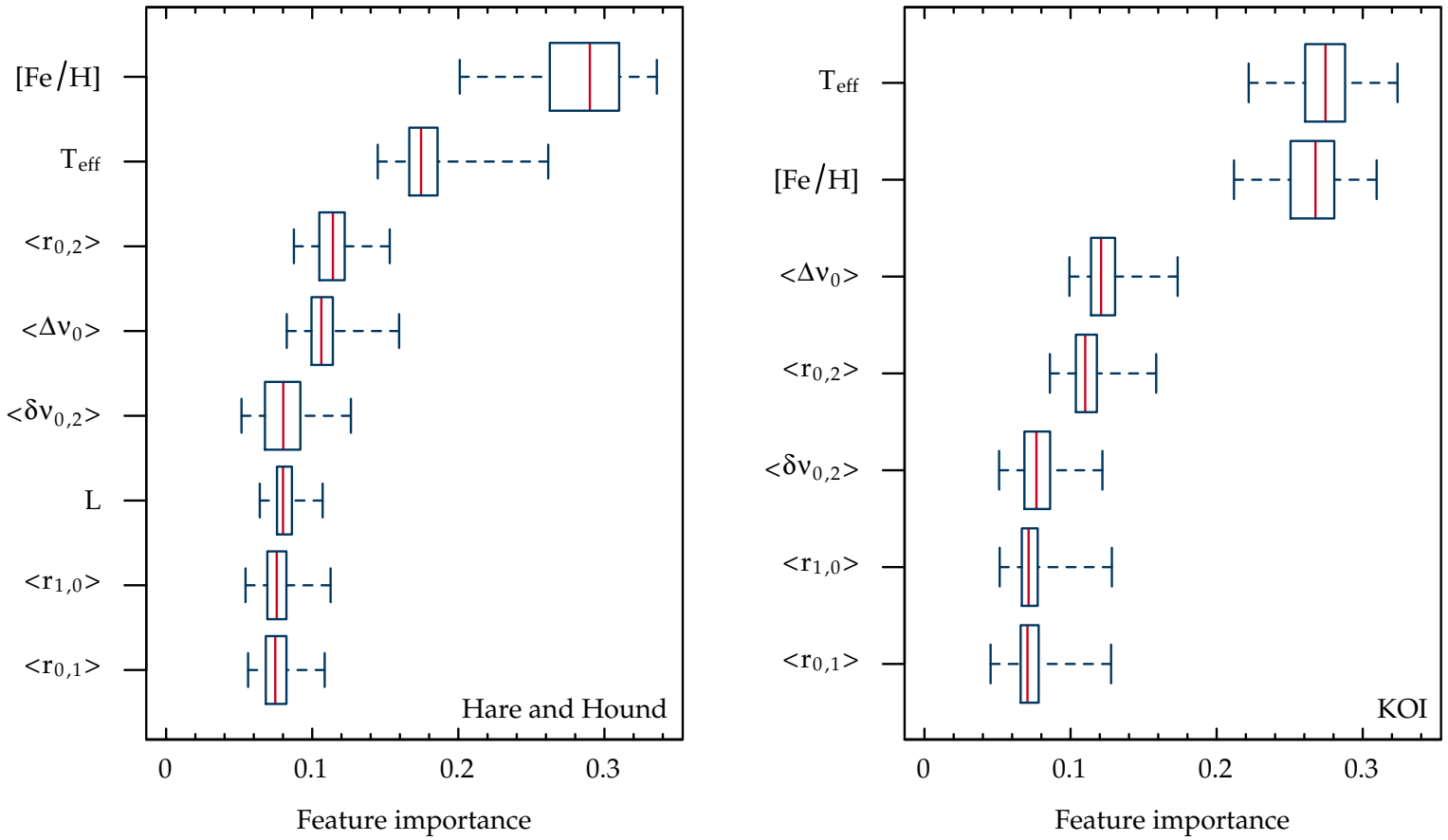


FIGURE 2.5. Box-and-whisker plots of relative importance for each feature in measuring fundamental stellar parameters for the hare-and-hound exercise data (left), where luminosities are available; and the *Kepler* objects-of-interest (right), where they are not. Octupole ($\ell = 3$) modes have not been measured in any of these stars, so $\langle \delta\nu_{1,3} \rangle$ and $\langle r_{1,3} \rangle$ from evolutionary modelling are not supplied to these random forests. The boxes are sorted by median importance.

vented from predicting non-physical quantities such as negative masses or from violating conservation requirements.

Secondly, due to the decision rule process that is explained below, random forests are insensitive to the scale of the data. Unless care is taken, other regression methods will artificially weight some observable quantities like temperature as being more important than, say, luminosity, solely because temperatures are written using larger numbers (e.g., 5777 vs. 1, see for example section 11.5.3 of Hastie et al. 2009 for a discussion). Consequently, solutions obtained by other methods will change if they are run using features that are expressed using different units of measure. For example, other methods will produce different regressors if trained on luminosity values expressed in solar units verses values expressed in ergs, whereas random forests will not. Commonly, this problem is mitigated in other methods by means of variable standardization and through the use of Mahalabonis distances (Mahalanobis 1936). However, these transformations are arbitrary, and handling variables naturally without rescaling is thus preferred.

Thirdly, random forests take only seconds to train, which can be a large benefit if different stars have different features available. For example, some stars have luminosity information available whereas others do not, so a different regressor must be trained for each. In the extreme case, if one wanted to make predictions for stars using all of their respectively observed frequencies, one would need to train a new regressor for each star using the subset of simulated frequencies that correspond to the ones observed for that star. Ignoring the difficulties of surface-term corrections and mode identifications, such an approach would be well-handled by random forest, suffering only a small hit to performance from its relatively small training cost. On the other hand, it would be infeasible to do this on a star-by-star basis with most other routines such as deep neural networks, because those methods can take days or even weeks to train.

And finally, as we saw in the previous section, random forests provide the opportunity to extract insight about the actual regression being performed by examining the importance of each feature in making predictions.

Uncertainty

There are three separate sources of uncertainty in predicting stellar parameters. The first is the systematic uncertainty in the physics used to model stars. These uncertainties are unknown, however, and hence cannot be propagated. The second is the uncertainty belonging to the observations of the star. We propagate measurement uncertainties σ into the predictions by perturbing all measured quantities $n = 10,000$ times with normal noise having zero mean and standard deviation σ. We account for the covariance between asteroseismic separations and ratios by recalculating them upon each perturbation.

The final source is regression uncertainty. Fundamentally, each parameter can only be constrained to the extent that observations are able to bear infor-

mation pertaining to that parameter. Even if observations were error-free, there still may exist a limit to which information gleaned from the surface may tell us about the physical qualities and evolutionary history of a star. We quantify those limits via cross-validation: we train the random forest on only a subset of the simulated evolutionary tracks and make predictions on a held-out validation set. We randomly hold out a different subset of the tracks 25 times to serve as different validation sets and obtain averaged accuracy scores.

We calculate accuracies using several scores. The first is the explained variance score V_e:

$$V_e = 1 - \frac{\mathrm{Var}\{y - \hat{y}\}}{\mathrm{Var}\{y\}} \tag{2.8}$$

where y is the true value we want to predict from the validation set (e.g. stellar mass), $\hat{y}$ is the predicted value from the random forest, and Var is the variance, i.e. the square of the standard deviation. This score tells us the extent to which the regressor has reduced the variance in the parameter it is predicting. The value ranges from negative infinity, which would be obtained by a pathologically bad predictor; to one for a perfect predictor, which occurs if all of the values are predicted with zero error.

The next score we consider is the residuals of each prediction, i.e. the absolute difference between the true value y and the predicted value $\hat{y}$. Naturally, we want this value to be as low as possible. We also consider the precision of the regression $\hat{\sigma}$ by taking the standard deviation of predictions across all of the decision trees in the forest. Finally, we consider these scores together by calculating the distance of the residuals in units of precision, i.e. $|\hat{y} - y|/\hat{\sigma}$.

Figure 2.6 shows these accuracies as a function of the number of evolutionary tracks used in the training of the random forest. Since the residuals and standard deviations of each parameter are incomparable, we normalize them by dividing by the maximum value. We also consider the number of trees in the forest and the number of models per evolutionary track. In this work, we use 256 trees in each forest, which we have selected via cross-validation by choosing a number of trees that is greater than the point at which we saw that the explained variance was no longer increasing greatly; see Appendix 2.6.4 for an extended discussion.

When supplied with enough stellar models, the random forest reduces the variance in each parameter and is able to make precise inferences. The forest has very high predictive power for most parameters, and as a result, essentially all of the uncertainty when predicting quantities such as stellar radii and luminosities will stem from observational uncertainty. However, for some model parameters—most notably the mixing length parameter—there is still a great deal of variance in the residuals. Prior to the point where the regressor has been trained on about 500 evolutionary tracks, the differences between the true and predicted mixing lengths actually have a greater variance than just the true mixing lengths themselves. Likewise, the diffusion multiplication factor is difficult to constrain because a star can achieve the same present-day [Fe/H] by either having a large initial non-hydrogen abundance and a large diffusion mul-

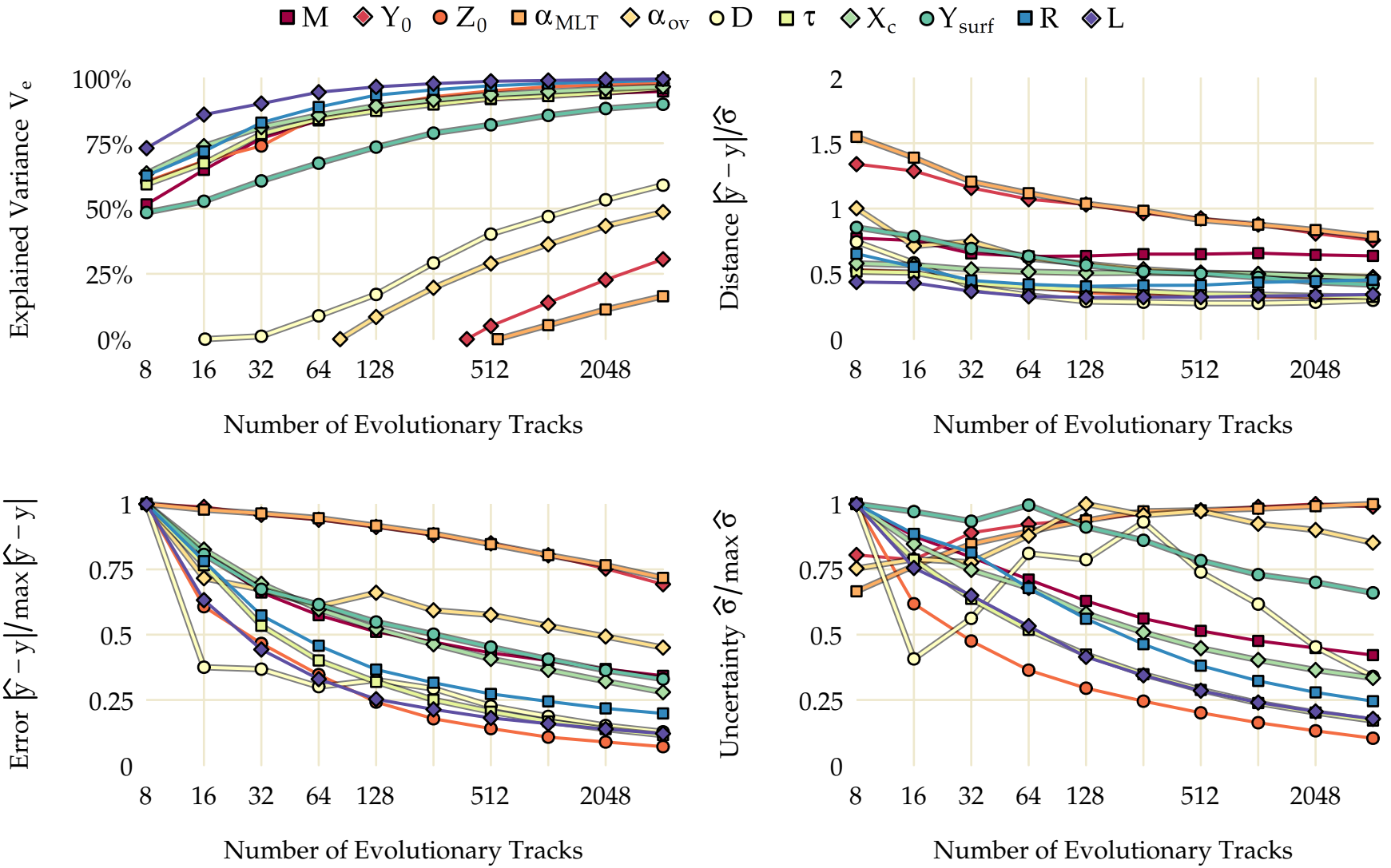


FIGURE 2.6. Evaluations of regression accuracy. Explained variance (top left), accuracy per precision distance (top right), normalized absolute error (bottom left), and normalized uncertainty (bottom right) for each stellar parameter as a function of the number of evolutionary tracks used in training the random forest. These results use 64 models per track and 256 trees in the random forest.

tiplication factor, or by having the same initial [Fe/H] as present [Fe/H] but with diffusion disabled. These difficult-to-constrain parameters will therefore be predicted with substantial uncertainties regardless of the precision of the observations.

2.3 Results

We perform three tests of our method. We begin with a hare-and-hound simulation exercise to show that we can reliably recover parameters. We then move to the Sun and the solar-like stars 16 Cyg A & B, which have been the subjects of many investigations; and we conclude by applying our method to 34 *Kepler* objects-of-interest. In each case, we train our random forest regressor on the subset of observational data that is available for the stars being processed. In the case of the Sun and 16 Cygni, we know very accurately their radii, luminosities, and surface gravities. For other stars, we will predict this information instead of supplying it.

2.3.1 Hare and Hound

We performed a blind hare-and-hound exercise to evaluate the performance of our predictor. Author S.B. prepared twelve models varied in mass, initial chemical composition, and mixing length parameter with only some models having overshooting and only some models having atomic diffusion included. The models were evolved without rotation using the Yale rotating stellar evolution code (YREC; Demarque et al. 2008), which is a different evolution code than the one that was used to train the random forest. Effective temperatures, luminosities, [Fe/H] and ν_{max} values as well as $\ell = 0, 1, 2$ frequencies were obtained from each model. Author G.C.A. perturbed the "observations" of these models according to the scheme devised by Reese et al. (2016). Appendix 2.6.5 lists the true values and the perturbed observations of the hare-and-hound models. The perturbed observations and their uncertainties were given to author E.P.B., who used the described method to recover the stellar parameters of these models without being given access to the true values. Relative differences between the true and predicted ages, masses, and radii for these models are plotted against their true values in Figure 2.7. The method is able to recover the true model values within uncertainties even when they have been perturbed by noise. We do not compare the predicted mixing length parameter, overshooting parameter or diffusion multiplication factor the interpretation of these parameters depends on how they have been defined and their precise implementation.

2.3.2 The Sun and the 16 Cygni System

To ensure confidence in our predictions on *Kepler* data, we first degrade the frequencies of the Sun at solar minimum that were obtained by the Birmingham

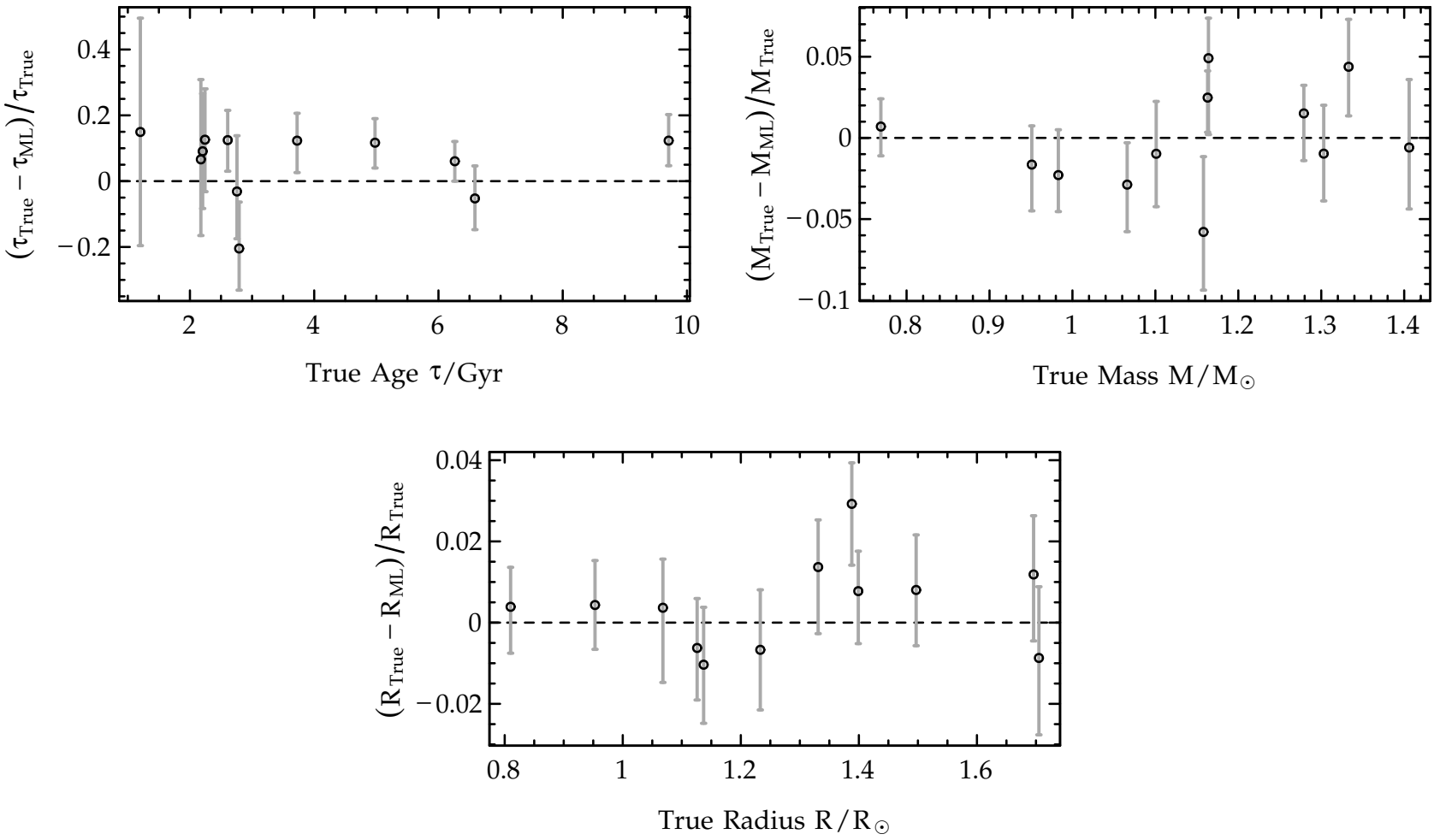


FIGURE 2.7. Relative differences between the predicted and true values for age (top left), mass (top right), and radius (bottom) as a function of the true values in a hare-and-hound simulation exercise.

Solar-Oscillations Network (BiSON; Davies et al. 2014a) to the level of information that is achievable by the spacecraft. We also degrade the Sun's uncertainties of other observations by applying 16 Cyg B's uncertainties of effective temperature, luminosity, surface gravity, metallicity, ν_{max}, radius, and radial velocity. Finally, we perturb each value with random Gaussian noise according to its uncertainty to reflect the fact that the measured value of an uncertain observation is not *per se* the true value. We use the random forest whose feature importances were shown in Figure 2.4 to predict the values of the Sun; i.e. the random forest trained on effective temperatures, metallicities, luminosities, surface gravities, radii, and global asteroseismic quantities $\langle \Delta\nu_0 \rangle$, $\langle \delta\nu_{0,2} \rangle$, $\langle \delta\nu_{1,3} \rangle$, $\langle r_{0,2} \rangle$, $\langle r_{1,3} \rangle$, $\langle r_{0,1} \rangle$, and $\langle r_{1,0} \rangle$. We show in Figure 2.8 the densities for the predicted mass, initial composition, mixing length parameter, overshoot coefficient, and diffusion multiplication factor needed for fitting an evolutionary model to degraded data of the Sun as well as the predicted solar age, core hydrogen abundance, and surface helium abundance. As discussed in Section 2.2.3, these densities show the distributions resulting from running $10,000$ different noise perturbations fed through the random forest. Relative uncertainties $\epsilon = 100 \cdot \sigma/\mu$ are also indicated, where μ is the mean and σ is the standard deviation of the quantity being predicted. Our predictions are in good agreement with the known values (see also Table 2.1 and Table 2.2, and *cf*. Equation 1.37).

Several parameters show multimodality due to model degeneracies. For example, two solutions for the initial helium are present. This is because it covaries with the mixing length parameter: the peak of higher Y_0 corresponds to the peak of lower α_{MLT} and vice versa. Likewise, high values of surface helium correspond to low values of the diffusion multiplication factor.

Effective temperatures, surface gravities, and metallicities of 16 Cyg A and B were obtained from Ramírez et al. (2009); radii and luminosities from White et al. (2013); and frequencies from Davies et al. (2015). We obtained the radial velocity measurements of 16 Cyg A and B from Nidever et al. (2002) and corrected frequencies for Doppler shifting as per the prescription in Davies et al. (2014b). We tried with and without line-of-sight corrections and found that it did not affect the predicted quantities or their uncertainties. We use the same random forest as we used for the degraded solar data to predict the parameters of these stars. The initial parameters—masses, chemical compositions, mixing lengths, diffusion multiplication factors, and overshoot coefficients—for 16 Cygni as predicted by machine learning are shown in Table 2.1, and the predicted current parameters—age, surface helium and core hydrogen abundances—are shown in Table 2.2. For reference we also show the predicted solar values from these inputs there as well. These results support the hypothesis that 16 Cyg A and B were co-natal; i.e. they formed at the same time with the same initial composition.

We additionally predict the radii and luminosities of 16 Cyg A and B instead of using them as features. Figure 2.9 shows our inferred radii, luminosities and surface helium abundances of 16 Cyg A and B plotted along with the

values determined by interferometry (White et al. 2013) and an asteroseismic estimate (Verma et al. 2014b). Here again we find excellent agreement between our method and the measured values.

Metcalfe et al. (2015) performed detailed modelling of 16 Cyg A and B using the Asteroseismic Modeling Portal (AMP), a genetic algorithm for matching individual frequencies of stars to stellar models. They calculated their results without heavy-element diffusion (i.e. with helium-only diffusion) and without overshooting. In order to account for systematic uncertainties, they multiplied the spectroscopic uncertainties of 16 Cyg A and B by an arbitrary constant $C = 3$. Therefore, in order to make a fair comparison between the results of our method and theirs, we generate a new matrix of evolutionary models with those same conditions and also increase the uncertainties on [Fe/H] by a factor of C. In Figure 2.10, we show probability densities of the predicted parameters of 16 Cyg A and B that we obtain using machine learning in comparison with the results obtained by AMP. We find the values and uncertainties agree well. To perform their analysis, AMP required more than $15,000$ hours of CPU time to model 16 Cyg A and B using the world's 10th fastest supercomputer, the Texas Advanced Computing Center Stampede (TOP500 2015). Here we have obtained comparable results in roughly one minute on a computing cluster with 64 2.5 GHz cores using only global asteroseismic quantities and no individual frequencies. Although more computationally expensive than our method, detailed optimization codes like AMP do have advantages in that they are additionally able to obtain detailed structural models of stars.

2.3.3 *Kepler* Objects of Interest

We obtain observations and frequencies of the KOI targets from Davies et al. (2016). We use line-of-sight radial velocity corrections when available, which was only the case for KIC 6278762 (Latham et al. 2002), KIC 10666592 (Maldonado et al. 2013), and KIC 3632418 (Gontcharov 2006). We use the random forest whose feature importances were shown in Figure 2.5 to predict the fundamental parameters of these stars; that is, the random forest that is trained on effective temperatures, metallicities, and asteroseismic quantities $\langle \Delta\nu_0 \rangle$, $\langle \delta\nu_{0,2} \rangle$, $\langle r_{0,2} \rangle$, $\langle r_{0,1} \rangle$, and $\langle r_{1,0} \rangle$. The predicted initial conditions—masses, chemical compositions, mixing lengths, overshoot coefficients, and diffusion multiplication factors—are shown in Table 2.3; and the predicted current conditions—ages, core hydrogen abundances, surface gravities, luminosities, radii, and surface helium abundances—are shown in Table 2.4. Figure 2.11 shows the fundamental parameters obtained from our method plotted against those obtained by Silva Aguirre et al. (2015, hereinafter KAGES). We find good agreement across all stars.

Although still in statistical agreement, the median values of our predicted ages are systematically lower and the median values of our predicted masses are systematically higher than those predicted by KAGES. We conjecture that these discrepancies arise from differences in input physics. We vary the efficiency of

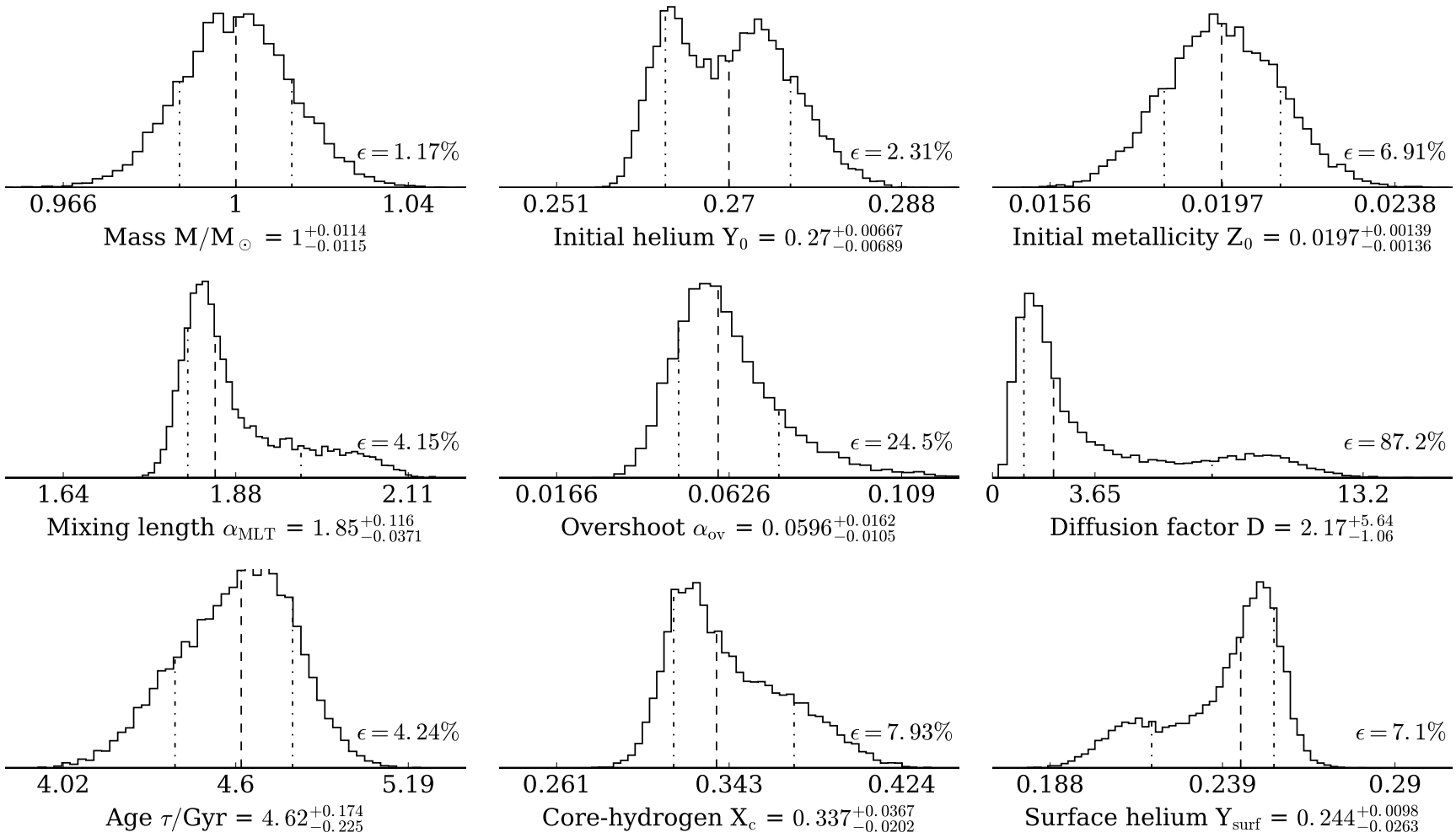


FIGURE 2.8. Predictions from machine learning of initial (top six) and current (bottom three) stellar parameters for degraded solar data. Labels are placed at the mean and 3σ levels. Dashed and dot-dashed lines indicate the median and quartiles, respectively. Relative uncertainties ϵ are shown beside each plot. Note that the overshoot parameter applies to all convective boundaries and is not modified over the course of evolution, so a non-zero value does not imply a convective core.

TABLE 2.1. Means and standard deviations for predicted initial stellar parameters of the Sun (degraded data) and 16 Cyg A and B.

Name	$M/M_\odot$	Y_0	Z_0	α_{MLT}	α_{ov}	D
Sun	1.00 ± 0.012	0.270 ± 0.0062	0.020 ± 0.0014	1.88 ± 0.078	0.06 ± 0.015	3.7 ± 3.18
16 Cyg A	1.08 ± 0.016	0.262 ± 0.0073	0.022 ± 0.0014	1.86 ± 0.077	0.07 ± 0.028	0.9 ± 0.76
16 Cyg B	1.03 ± 0.015	0.268 ± 0.0065	0.021 ± 0.0015	1.83 ± 0.069	0.11 ± 0.029	1.9 ± 1.57

TABLE 2.2. Means and standard deviations for predicted current-age stellar parameters of the Sun (degraded data) and 16 Cyg A and B.

Name	τ/Gyr	X_c	Y_{surf}
Sun	4.6 ± 0.20	0.34 ± 0.027	0.24 ± 0.017
16 Cyg A	6.9 ± 0.40	0.06 ± 0.024	0.246 ± 0.0085
16 Cyg B	6.8 ± 0.28	0.15 ± 0.023	0.24 ± 0.017

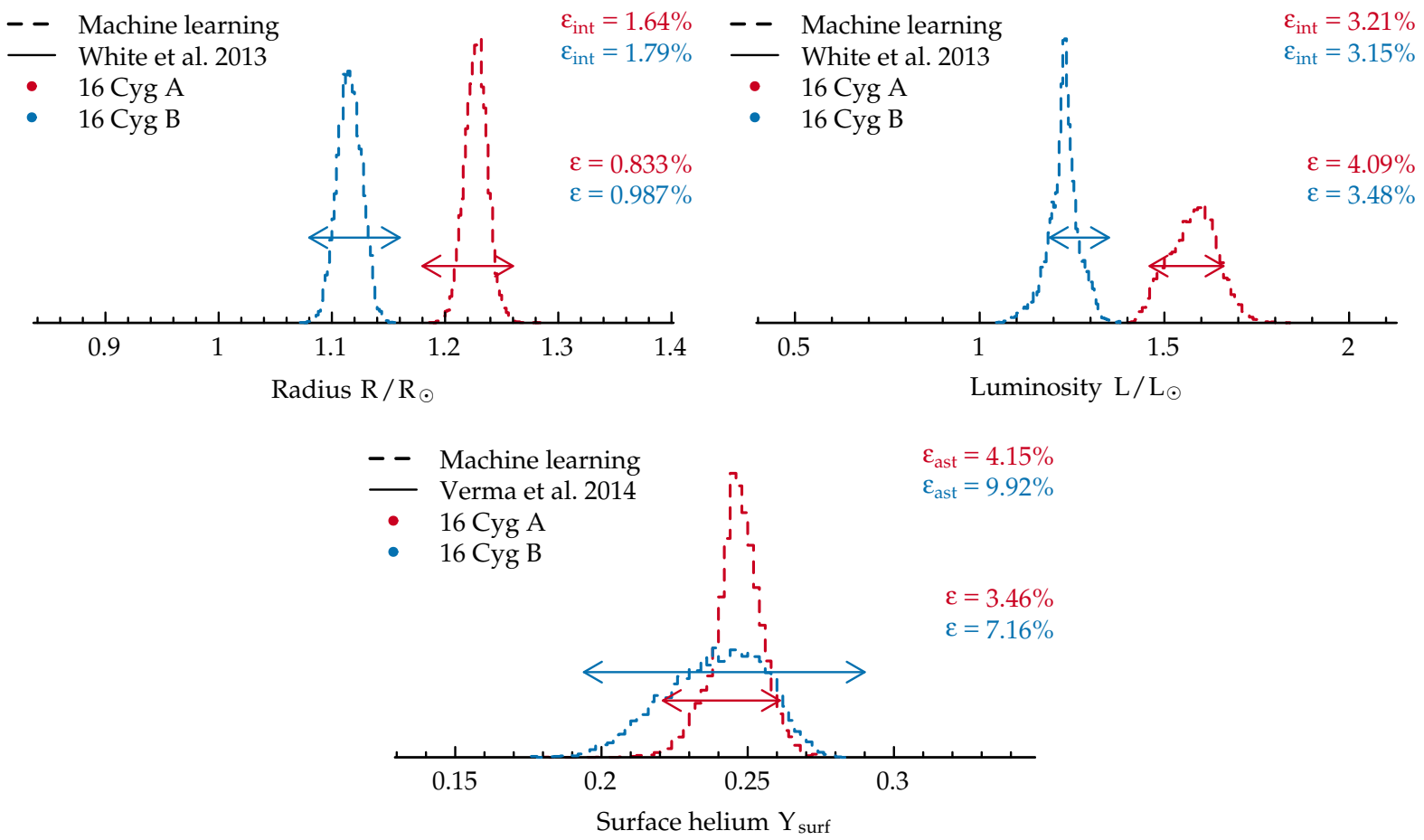


FIGURE 2.9. Probability densities for predictions of 16 Cyg A (red) and B (blue) from machine learning of radii (top left), luminosities (top right), and surface helium abundances (bottom). Relative uncertainties ϵ are shown beside each plot. Predictions and 2σ uncertainties from interferometric ("int") measurements and asteroseismic ("ast") estimates are shown with arrows.

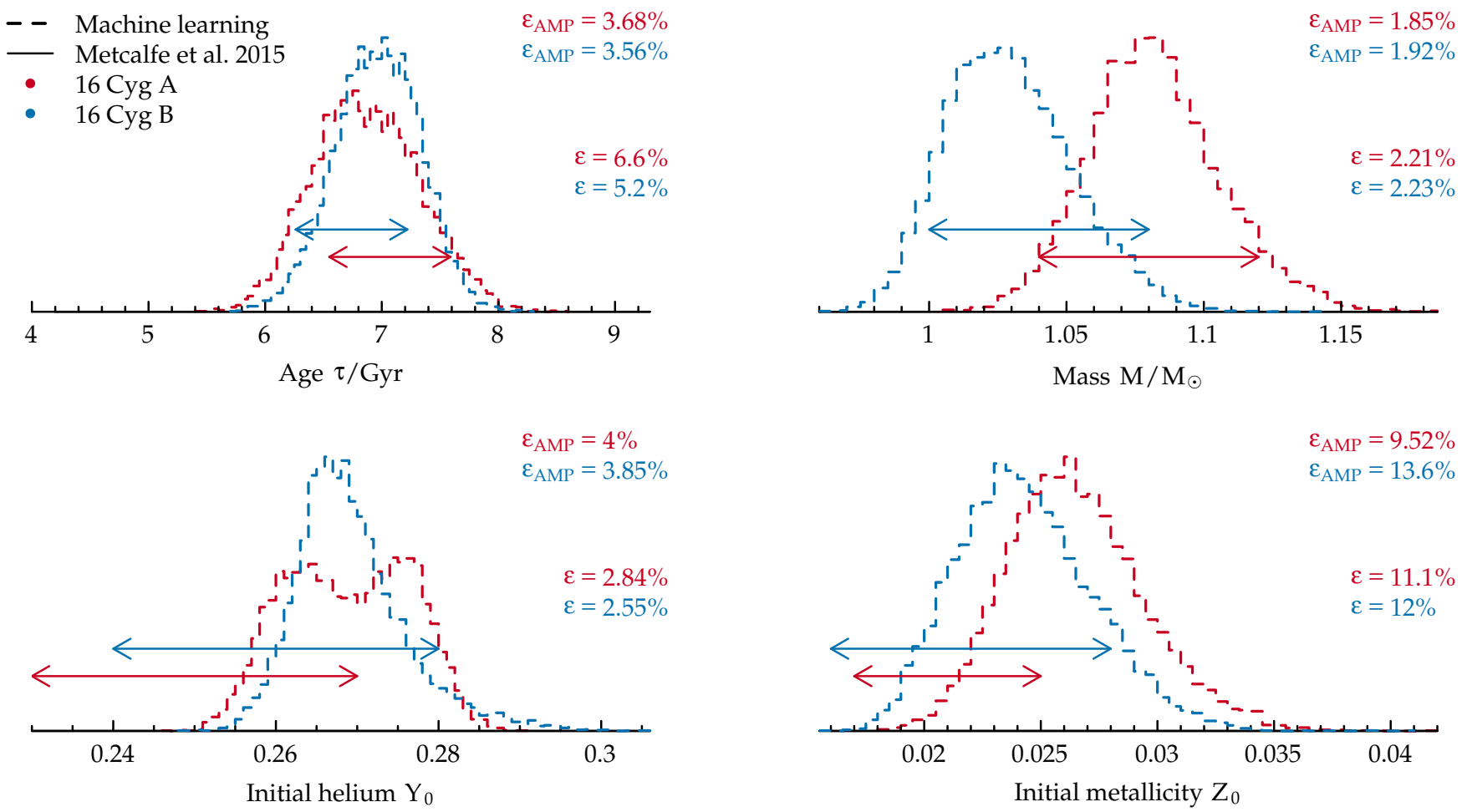


FIGURE 2.10. Probability densities showing predictions from machine learning of fundamental stellar parameters for 16 Cyg A (red) and B (blue) along with predictions from AMP modelling. Relative uncertainties are shown beside each plot. Predictions and 2σ uncertainties from AMP modelling are shown with arrows.

diffusion, the extent of convective overshooting, and the value of the mixing length parameter to arrive at these estimates, whereas the KAGES models are calculated using fixed amounts of diffusion, without overshoot, and with a solar-calibrated mixing length. Models with overshooting, for example, will be more evolved at the same age due to having larger core masses. Without direct access to their models, however, the exact reason is difficult to pinpoint.

We find a significant linear trend in the *Kepler* objects-of-interest between the diffusion multiplication factor and stellar mass needed to reproduce observations ($P = 0.0001$ from a two-sided t-test with $N - 2 = 32$ degrees of freedom). Since the values of mass and diffusion multiplication factor are uncertain, we use Deming regression to estimate the coefficients of this relation without regression dilution (Deming 1943). We show the diffusion multiplication factors as a function of stellar mass for all of these stars in Figure 2.12. We find that the diffusion multiplication factor linearly decreases with mass, i.e.

$$D = (8.6 \pm 1.94) - (5.6 \pm 1.37) \cdot M/M_{\odot} \tag{2.9}$$

and that this relation explains observations better than any constant factor (e.g., $D = 1$ or $D = 0$).

2.4 Discussion

The amount of time it takes to make predictions for a star using a trained random forest can be decomposed into two parts: the amount of time it takes to calculate perturbations to the observations of the star (see Section 2.2.3), and the amount of time it takes to make a prediction on each perturbed set of observations. Hence we have

$$t = n(t_p + t_r) \tag{2.10}$$

where t is the total time, n is the number of perturbations, t_p is the time it takes to perform a single perturbation, and t_r is the random forest regression time. We typically see times of $t_p = (7.9 \pm 0.7) \cdot 10^{-3}$ (s) and $t_r = (1.8 \pm 0.4) \cdot 10^{-5}$ (s). We chose a conservative $n = 10,000$ for the results presented here, which results in a time of around a minute per star. Since each star can be processed independently and in parallel, a computing cluster could feasibly process a catalog containing millions of objects in less than a day. Since $t_r \ll t_p$, the calculation depends almost entirely on the time it takes to perturb the observations.[4] There is also the one-time cost of training the random forest, which takes less than a minute and can be reused without retraining on every star with the same information. It does need to be retrained if one wants to consider a different combination of input or output parameters.

There is a one-time cost of generating the matrix of training data. We ran our simulation generation scheme for a week on our computing cluster and

[4] Our perturbation code uses an interpreted language (R), so if needed, there is still room for speed-up.

TABLE 2.3. Means and standard deviations for initial conditions of the KOI data set inferred via machine learning. The values obtained from degraded solar data predicted on these quantities are shown for reference.

KIC	$M/M_\odot$	Y_0	Z_0	α_{MLT}	α_{ov}	D
3425851	1.15 ± 0.053	0.28 ± 0.020	0.015 ± 0.0028	1.9 ± 0.23	0.06 ± 0.057	0.5 ± 0.92
3544595	0.91 ± 0.032	0.270 ± 0.0090	0.015 ± 0.0028	1.9 ± 0.10	0.2 ± 0.11	4.9 ± 4.38
3632418	1.39 ± 0.057	0.267 ± 0.0089	0.019 ± 0.0032	2.0 ± 0.12	0.2 ± 0.14	1.1 ± 1.01
4141376	1.03 ± 0.036	0.267 ± 0.0097	0.012 ± 0.0025	1.9 ± 0.12	0.1 ± 0.11	4.0 ± 4.09
4143755	0.99 ± 0.037	0.277 ± 0.0050	0.014 ± 0.0026	1.77 ± 0.033	0.37 ± 0.071	13.4 ± 5.37
4349452	1.22 ± 0.056	0.28 ± 0.012	0.020 ± 0.0043	1.9 ± 0.17	0.10 ± 0.090	7.3 ± 8.82
4914423	1.19 ± 0.048	0.274 ± 0.0097	0.026 ± 0.0046	1.8 ± 0.11	0.08 ± 0.043	2.3 ± 1.6
5094751	1.11 ± 0.038	0.274 ± 0.0082	0.018 ± 0.0030	1.8 ± 0.11	0.07 ± 0.041	2.3 ± 1.39
5866724	1.29 ± 0.065	0.28 ± 0.011	0.027 ± 0.0058	1.8 ± 0.13	0.12 ± 0.086	7.0 ± 8.38
6196457	1.31 ± 0.058	0.276 ± 0.005	0.032 ± 0.0050	1.71 ± 0.050	0.16 ± 0.055	5.7 ± 2.34
6278762	0.76 ± 0.012	0.254 ± 0.0058	0.013 ± 0.0017	2.09 ± 0.069	0.06 ± 0.028	5.3 ± 2.23
6521045	1.19 ± 0.046	0.273 ± 0.0071	0.027 ± 0.0044	1.82 ± 0.074	0.12 ± 0.036	3.2 ± 1.31
7670943	1.30 ± 0.061	0.28 ± 0.017	0.021 ± 0.0045	2.0 ± 0.23	0.06 ± 0.064	1.0 ± 2.55
8077137	1.23 ± 0.070	0.270 ± 0.0093	0.018 ± 0.0028	1.8 ± 0.14	0.2 ± 0.11	2.9 ± 2.08
8292840	1.15 ± 0.079	0.28 ± 0.010	0.016 ± 0.0049	1.8 ± 0.15	0.1 ± 0.12	11. ± 10.7
8349582	1.23 ± 0.040	0.271 ± 0.0069	0.043 ± 0.0074	1.9 ± 0.12	0.11 ± 0.060	2.5 ± 1.11
8478994	0.81 ± 0.022	0.272 ± 0.0082	0.010 ± 0.0012	1.91 ± 0.054	0.21 ± 0.068	17. ± 9.74
8494142	1.42 ± 0.058	0.27 ± 0.010	0.028 ± 0.0046	1.70 ± 0.064	0.10 ± 0.051	1.6 ± 1.65
8554498	1.39 ± 0.067	0.272 ± 0.0082	0.031 ± 0.0032	1.70 ± 0.077	0.14 ± 0.079	1.7 ± 1.17
8684730	1.44 ± 0.030	0.277 ± 0.0075	0.041 ± 0.0049	1.9 ± 0.14	0.29 ± 0.094	15.2 ± 8.81
8866102	1.26 ± 0.069	0.28 ± 0.013	0.021 ± 0.0048	1.8 ± 0.15	0.08 ± 0.070	5. ± 7.48
9414417	1.36 ± 0.054	0.264 ± 0.0073	0.018 ± 0.0028	1.9 ± 0.13	0.2 ± 0.1	2.2 ± 1.68
9592705	1.45 ± 0.038	0.27 ± 0.010	0.029 ± 0.0038	1.72 ± 0.064	0.12 ± 0.056	0.6 ± 0.47
9955598	0.93 ± 0.028	0.27 ± 0.011	0.023 ± 0.0039	1.9 ± 0.10	0.2 ± 0.13	2.2 ± 1.76
10514430	1.13 ± 0.053	0.277 ± 0.0046	0.021 ± 0.0039	1.78 ± 0.059	0.30 ± 0.097	4.7 ± 1.77
10586004	1.31 ± 0.078	0.274 ± 0.0055	0.038 ± 0.0071	1.8 ± 0.13	0.2 ± 0.13	4.3 ± 3.99
10666592	1.50 ± 0.023	0.30 ± 0.013	0.030 ± 0.0032	1.8 ± 0.11	0.06 ± 0.043	0.2 ± 0.14
10963065	1.09 ± 0.031	0.264 ± 0.0083	0.014 ± 0.0025	1.8 ± 0.11	0.05 ± 0.027	3.1 ± 2.68
11133306	1.11 ± 0.044	0.272 ± 0.0099	0.021 ± 0.0040	1.8 ± 0.16	0.04 ± 0.033	5. ± 5.75
11295426	1.11 ± 0.033	0.27 ± 0.010	0.025 ± 0.0036	1.81 ± 0.084	0.05 ± 0.035	1.3 ± 0.87
11401755	1.15 ± 0.039	0.271 ± 0.0057	0.015 ± 0.0023	1.88 ± 0.055	0.33 ± 0.071	3.8 ± 1.81
11807274	1.32 ± 0.079	0.276 ± 0.0097	0.024 ± 0.0051	1.77 ± 0.083	0.11 ± 0.066	5.4 ± 5.61
11853905	1.22 ± 0.055	0.272 ± 0.0072	0.029 ± 0.0050	1.8 ± 0.12	0.18 ± 0.086	3.3 ± 1.85
11904151	0.93 ± 0.033	0.265 ± 0.0091	0.016 ± 0.0030	1.8 ± 0.13	0.05 ± 0.029	3.1 ± 2.09
Sun	1.00 ± 0.0093	0.266 ± 0.0035	0.018 ± 0.0011	1.81 ± 0.032	0.07 ± 0.021	2.1 ± 0.83

TABLE 2.4. Means and standard deviations for current-age conditions of the KOI data set inferred via machine learning. The values obtained from degraded solar data predicted on these quantities are shown for reference.

KIC	τ/Gyr	X_c	log g	$L/L_\odot$	$R/R_\odot$	Y_{surf}
3425851	3.7 ± 0.76	0.14 ± 0.081	4.234 ± 0.0098	2.7 ± 0.16	1.36 ± 0.022	0.27 ± 0.026
3544595	6.7 ± 1.47	0.31 ± 0.078	4.46 ± 0.016	0.84 ± 0.068	0.94 ± 0.020	0.23 ± 0.023
3632418	3.0 ± 0.36	0.10 ± 0.039	4.020 ± 0.0076	5.2 ± 0.25	1.91 ± 0.031	0.24 ± 0.021
4141376	3.4 ± 0.67	0.38 ± 0.070	4.41 ± 0.011	1.42 ± 0.097	1.05 ± 0.019	0.24 ± 0.022
4143755	8.0 ± 0.80	0.07 ± 0.022	4.09 ± 0.013	2.3 ± 0.12	1.50 ± 0.029	0.17 ± 0.023
4349452	2.4 ± 0.78	0.4 ± 0.10	4.28 ± 0.012	2.5 ± 0.14	1.32 ± 0.022	0.22 ± 0.043
4914423	5.2 ± 0.58	0.06 ± 0.032	4.162 ± 0.0097	2.5 ± 0.16	1.50 ± 0.022	0.24 ± 0.023
5094751	5.3 ± 0.67	0.07 ± 0.039	4.209 ± 0.0082	2.2 ± 0.13	1.37 ± 0.017	0.23 ± 0.024
5866724	2.4 ± 0.96	0.4 ± 0.12	4.24 ± 0.017	2.7 ± 0.13	1.42 ± 0.022	0.23 ± 0.038
6196457	4.0 ± 0.73	0.18 ± 0.061	4.11 ± 0.022	3.3 ± 0.21	1.68 ± 0.041	0.24 ± 0.016
6278762	10.3 ± 0.96	0.35 ± 0.026	4.557 ± 0.0084	0.34 ± 0.022	0.761 ± 0.0061	0.19 ± 0.023
6521045	5.6 ± 0.370	0.027 ± 0.0097	4.122 ± 0.0055	2.7 ± 0.15	1.57 ± 0.025	0.22 ± 0.019
7670943	2.3 ± 0.59	0.32 ± 0.088	4.234 ± 0.0099	3.3 ± 0.23	1.44 ± 0.025	0.26 ± 0.029
8077137	4.4 ± 0.96	0.08 ± 0.052	4.08 ± 0.016	3.7 ± 0.24	1.68 ± 0.044	0.22 ± 0.031
8292840	3.4 ± 1.48	0.3 ± 0.14	4.25 ± 0.023	2.6 ± 0.20	1.34 ± 0.026	0.19 ± 0.049
8349582	6.7 ± 0.53	0.02 ± 0.012	4.16 ± 0.012	2.2 ± 0.12	1.52 ± 0.016	0.23 ± 0.015
8478994	4.6 ± 1.75	0.50 ± 0.055	4.55 ± 0.012	0.51 ± 0.036	0.79 ± 0.014	0.21 ± 0.022
8494142	2.8 ± 0.52	0.18 ± 0.067	4.06 ± 0.018	4.5 ± 0.32	1.84 ± 0.043	0.24 ± 0.029
8554498	3.7 ± 0.79	0.09 ± 0.060	4.04 ± 0.015	4.1 ± 0.20	1.86 ± 0.043	0.25 ± 0.018
8684730	3.0 ± 0.38	0.24 ± 0.065	4.06 ± 0.046	4.1 ± 0.53	1.9 ± 0.11	0.17 ± 0.040
8866102	1.9 ± 0.71	0.4 ± 0.11	4.27 ± 0.014	2.8 ± 0.16	1.36 ± 0.024	0.24 ± 0.039
9414417	3.1 ± 0.31	0.09 ± 0.030	4.016 ± 0.0058	5.0 ± 0.32	1.90 ± 0.032	0.21 ± 0.026
9592705	3.0 ± 0.38	0.05 ± 0.026	3.973 ± 0.0087	5.7 ± 0.37	2.06 ± 0.035	0.26 ± 0.015
9955598	7.0 ± 0.98	0.37 ± 0.035	4.494 ± 0.0061	0.66 ± 0.041	0.90 ± 0.013	0.25 ± 0.020
10514430	6.5 ± 0.89	0.06 ± 0.022	4.08 ± 0.014	2.9 ± 0.17	1.62 ± 0.026	0.22 ± 0.021
10586004	4.9 ± 1.39	0.12 ± 0.090	4.09 ± 0.041	3.1 ± 0.27	1.71 ± 0.070	0.24 ± 0.021
10666592	2.0 ± 0.24	0.15 ± 0.036	4.020 ± 0.0066	5.7 ± 0.33	1.98 ± 0.018	0.29 ± 0.014
10963065	4.4 ± 0.58	0.16 ± 0.054	4.292 ± 0.0070	2.0 ± 0.1	1.24 ± 0.015	0.22 ± 0.029
11133306	4.1 ± 0.84	0.22 ± 0.079	4.319 ± 0.0096	1.7 ± 0.11	1.21 ± 0.019	0.22 ± 0.036
11295426	6.2 ± 0.78	0.09 ± 0.036	4.283 ± 0.0059	1.65 ± 0.095	1.26 ± 0.016	0.24 ± 0.012
11401755	5.6 ± 0.630	0.037 ± 0.0053	4.043 ± 0.0071	3.4 ± 0.19	1.69 ± 0.026	0.21 ± 0.026
11807274	2.8 ± 1.05	0.3 ± 0.11	4.17 ± 0.024	3.5 ± 0.22	1.57 ± 0.038	0.22 ± 0.035
11853905	5.7 ± 0.78	0.04 ± 0.020	4.11 ± 0.011	2.7 ± 0.16	1.62 ± 0.030	0.23 ± 0.022
11904151	9.6 ± 1.43	0.08 ± 0.037	4.348 ± 0.0097	1.09 ± 0.06	1.07 ± 0.019	0.21 ± 0.026
Sun	4.6 ± 0.16	0.36 ± 0.012	4.439 ± 0.0038	1.01 ± 0.041	1.000 ± 0.0066	0.245 ± 0.0076

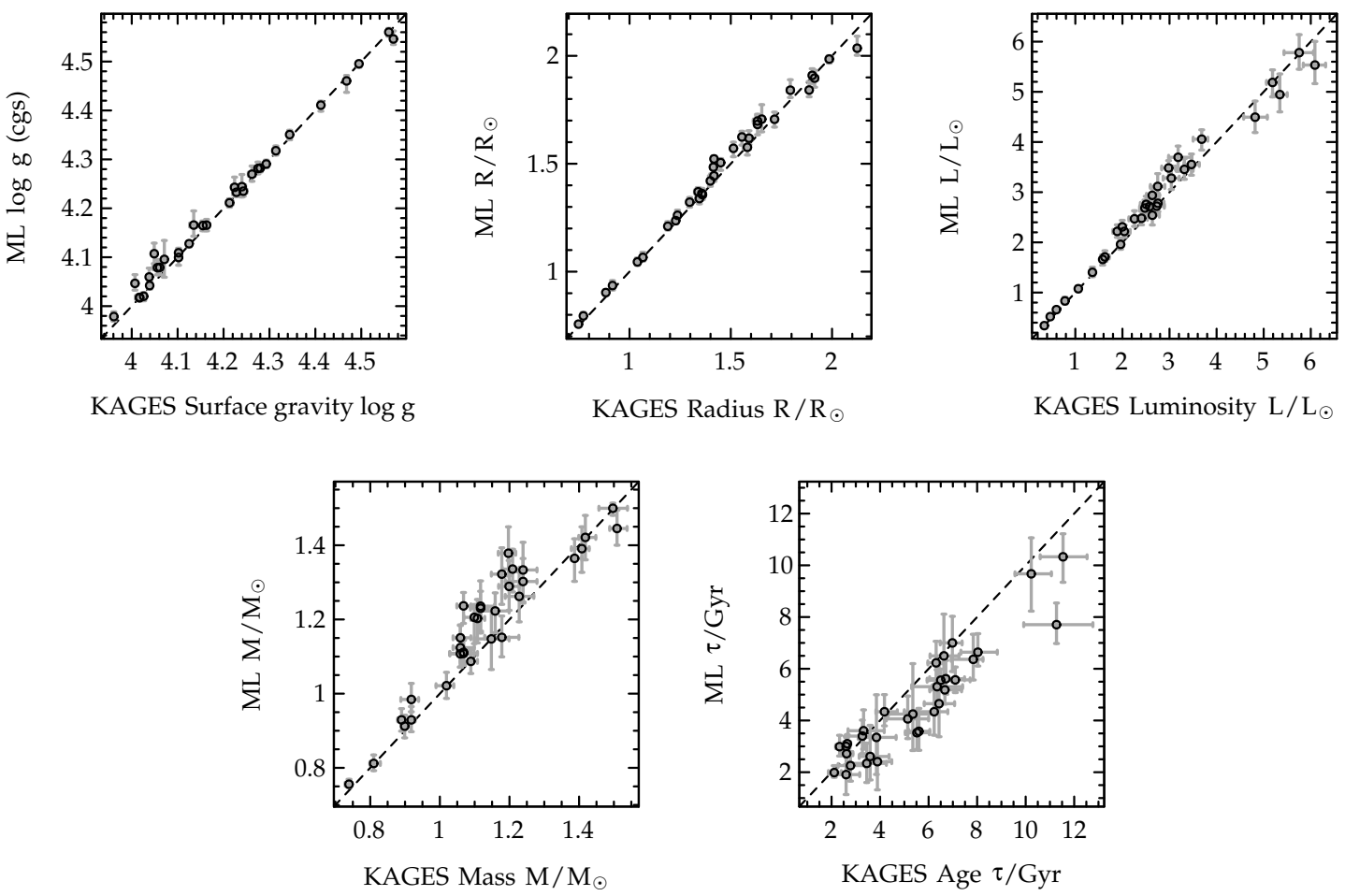


FIGURE 2.11. Predicted surface gravities, radii, luminosities, masses, and ages of 34 *Kepler* objects-of-interest plotted against the suggested KAGES values. Medians, 16% quantiles, and 84% quantiles are shown for each point. A dashed line of agreement is shown in all panels to guide the eye.

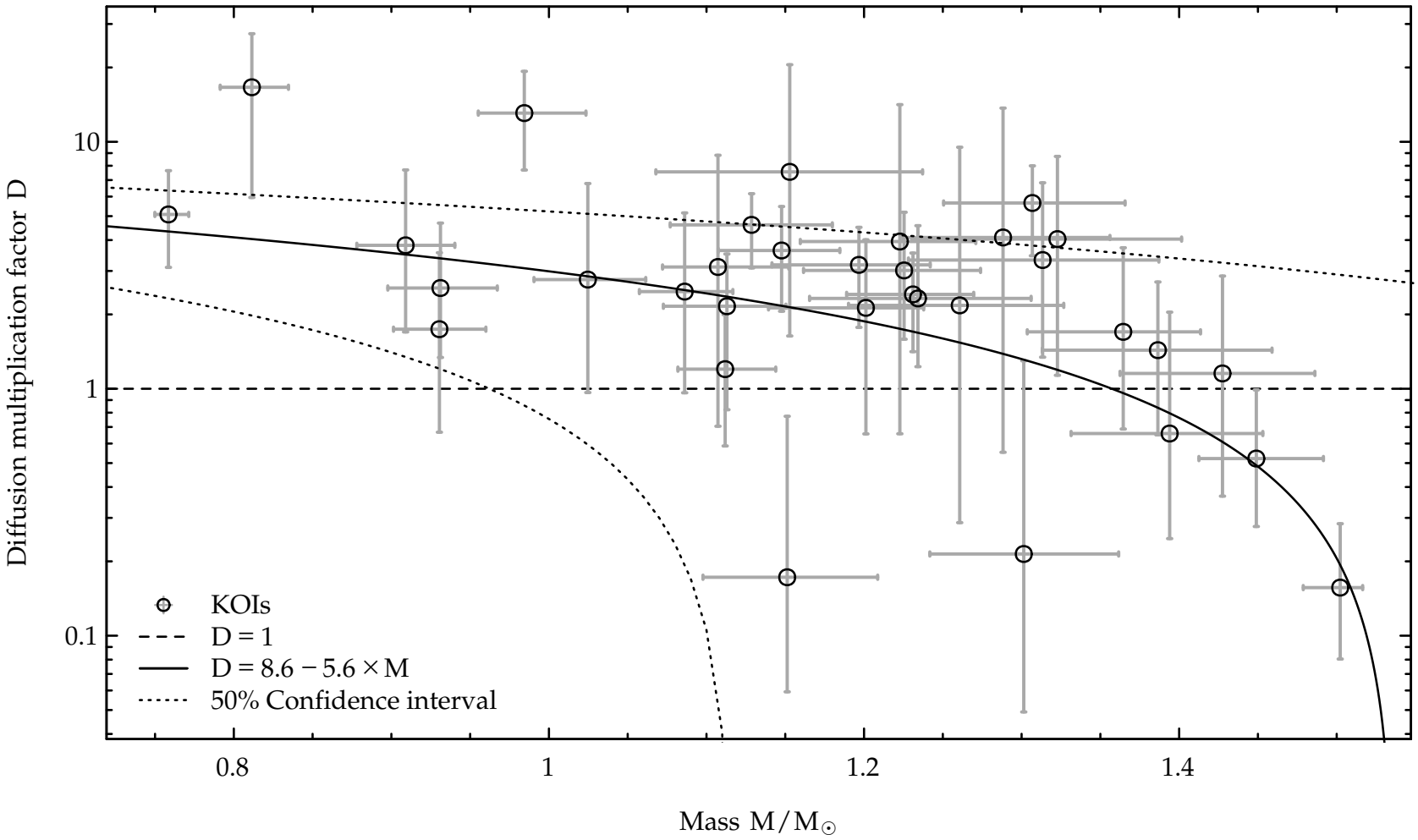


FIGURE 2.12. Logarithmic diffusion multiplication factor as a function of stellar mass for 34 *Kepler* objects-of-interest. The solid line is the line of best fit from Equation (2.9) and the dashed lines are the 50% confidence interval around this fit.

obtained 5,325 evolutionary tracks with 64 models per track, which resulted in a 123 MB matrix of stellar models. This is at least an order of magnitude fewer models than the amount that other methods use. Furthermore, this is in general more tracks than is needed by our method: we showed in Figure 2.6 that for most parameters—most notably age, mass, luminosity, radius, initial metallicity, and core hydrogen abundance—one needs only a fraction of the models that we generated in order to obtain good predictive accuracies. Finally, unless one wants to consider a different range of parameters or different input physics, this matrix would not need to be calculated again; a random forest trained on this matrix can be re-used for all future stars that are observed. Of course, our method would still work if trained using a different matrix of models, and our grid should work with other grid-based modelling methods.

Previously, Pulone and Scaramella (1997) developed a neural network for predicting stellar age based on the star's position in the Hertzsprung-Russell diagram. More recently, Verma et al. (2016) have worked on incorporating seismic information into that analysis as we have done here. Our method provides several advantages over these approaches. Firstly, the random forests that we use perform constrained regression, meaning that the values we predict for quantities like age and mass will always be non-negative and within the bounds of the training data, which is not true of the neural networks-based approach that they take. Secondly, using *averaged* frequency separations allows us to make predictions without need for concern over which radial orders were observed. Thirdly, we have shown that our random forests are very fast to train, and can be retrained in only seconds for stars that are missing observational constraints such as luminosities. In contrast, deep neural networks are computationally intensive to train, potentially taking days or weeks to converge depending on the breadth of network topologies considered in the cross-validation. Finally, our grid is varied in six initial parameters—M, Y_0, Z_0, $\alpha_{\rm MLT}$, $\alpha_{\rm ov}$, and D, which allows our method to explore a wide range of stellar model parameters.

2.5 Conclusions

Here we have considered the constrained multiple-regression problem of inferring fundamental stellar parameters from observations. We created a grid of evolutionary tracks varied in mass, chemical composition, mixing length parameter, overshooting coefficient, and diffusion multiplication factor. We evolved each track in time along the main sequence and collected observable quantities such as effective temperatures and metallicities as well as global statistics on the modes of oscillations from models along each evolutionary path. We used this matrix of stellar models to train a machine learning algorithm to be able to discern the patterns that relate observations to fundamental stellar parameters. We then applied this method to hare-and-hound exercise data, the Sun, 16 Cyg A and B, and 34 planet-hosting candidates that have been observed by *Kepler* and rapidly obtained precise initial conditions and current-age values of

these stars. Remarkably, we were able to empirically determine the value of the diffusion multiplication factor and hence the efficiency of diffusion required to reproduce the observations instead of inhibiting it *ad hoc*. A larger sample size will better constrain the diffusion multiplication factor and determine what other variables are relevant in its parameterization. This is work in progress.

The method presented here has many advantages over existing approaches. First, random forests can be trained and used in only seconds and hence provide substantial speed-ups over other methods. Observations of a star simply need to be fed through the forest—akin to plugging numbers into an equation—and do not need to be subjected to expensive iterative optimization procedures. Secondly, random forests perform non-linear and non-parametric regression, which means that the method can use orders-of-magnitude fewer models for the same level of precision, while additionally attaining a more rigorous appraisal of uncertainties for the predicted quantities. Thirdly, our method allows us to investigate wide ranges and combinations of stellar parameters. And finally, the method presented here provides the opportunity to extract insights from the statistical regression that is being performed, which is achieved by examining the relationships in stellar physics that the machine learns by analyzing simulation data. This contrasts the blind optimization processes of other methods that provide an answer but do not indicate the elements that were important in doing so.

We note that the predicted quantities reflect a set of choices in stellar physics. Although such biases are impossible to propagate, varying model parameters that are usually kept fixed—such as the mixing length parameter, diffusion multiplication factor, and overshooting coefficient—takes us a step in the right direction. Furthermore, the fact that quantities such as stellar radii and luminosities—quantities that have been measured accurately, not just precisely—can be reproduced both precisely and accurately by this method, gives a degree of confidence in its efficacy.

The method we have presented here is currently only applicable to main-sequence stars. We intend to extend this study to later stages of evolution.

Acknowledgements

The research leading to the presented results has received funding from the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013) / ERC grant agreement no 338251 (StellarAges). This research was undertaken in the context of the International Max Planck Research School for Solar System Research. S.B. acknowledges partial support from NSF grant AST-1514676 and NASA grant NNX13AE70G. W.H.B. acknowledges research funding by Deutsche Forschungsgemeinschaft (DFG) under grant SFB 963/1 "Astrophysical flow instabilities and turbulence" (Project A18).

Software

Analysis in this chapter was performed with python 3.5.1 libraries scikit-learn 0.17.1 (Pedregosa et al. 2011), NumPy 1.10.4 (Van Der Walt et al. 2011), and pandas 0.17.1 (McKinney 2010) as well as R 3.2.3 (R Core Team 2014) and the R libraries magicaxis 1.9.4 (Robotham 2015), RColorBrewer 1.1-2 (Neuwirth 2014), parallelMap 1.3 (Bischl and Lang 2015), data.table 1.9.6 (Dowle et al. 2015), lpSolve 5.6.13 (Berkelaar and others 2015), ggplot2 2.1.0 (Wickham 2016), GGally 1.0.1 (Schloerke et al. 2014), scales 0.3.0 (Wickham 2015), deming 1.0-1 (Therneau 2014), and matrixStats 0.50.1 (Bengtsson 2015).

2.6 Appendix

2.6.1 Model Selection

To prevent statistical bias towards the evolutionary tracks that generate the most models, i.e. the ones that require the most careful calculations and therefore use smaller time-steps, or those that live on the main sequence for a longer amount of time; we select $n = 64$ models from each evolutionary track such that the models are as evenly-spaced in core hydrogen abundance as possible. We chose 64 because it is a power of two, which thus allows us to successively omit every other model when testing our regression routine and still maintain regular spacings.

Starting from the original vector of length n of core hydrogen abundances $\mathbf{x}$, we find the subset of length m that is closest to the optimal spacing $\mathbf{b}$, where[5]

$$b_i = X_T + (i-1) \cdot \frac{X_Z - X_T}{m-1}, \qquad i = 1, \ldots, m \tag{2.11}$$

with X_Z being the core hydrogen abundance at ZAMS and X_T being that at TAMS. To obtain the closest possible vector to $\mathbf{b}$ from our data $\mathbf{x}$, we solve a transportation problem using integer optimization (Delmotte 2014). First we set up a cost matrix $\mathbf{C}$ consisting of absolute differences between the original abundances $\mathbf{x}$ and the ideal abundances $\mathbf{b}$:

$$\mathbf{C} = \begin{bmatrix} |b_1 - x_1| & |b_1 - x_2| & \ldots & |b_1 - x_n| \\ |b_2 - x_1| & |b_2 - x_2| & \ldots & |b_2 - x_n| \\ \vdots & \vdots & \ddots & \vdots \\ |b_m - x_1| & |b_m - x_2| & \ldots & |b_m - x_n| \end{bmatrix}. \tag{2.12}$$

We then require that exactly m values are selected from $\mathbf{x}$, and that each value is selected no more than one time. Simply selecting the closest data point to each ideally-separated point will not work because this could result in the same point being selected twice; and selecting the second closest point in that situation

[5] This equation has been corrected from the original publication.

does not remedy it because a different result could be obtained if the points were processed in a different order.

We denote the optimal solution matrix by $\hat{\mathbf{S}}$, and find it by minimizing the cost matrix subject to the following constraints:

$$\begin{aligned} \hat{\mathbf{S}} = \arg\min_{\mathbf{S}} & \sum_{ij} S_{ij} C_{ij} \\ \text{subject to } & \sum_{j} S_{ij} \leqslant 1 \quad \text{for all } i = 1 \dots n \\ \text{and } & \sum_{i} S_{ij} = 1 \quad \text{for all } j = 1 \dots m. \end{aligned} \tag{2.13}$$

The indices of $\mathbf{x}$ that are most near to being equidistantly-spaced are then found by looking at which columns of $\hat{\mathbf{S}}$ contain ones, and we are done. The solution is visualized in Figure 2.13.

2.6.2 Initial Grid Strategy

The initial conditions of a stellar model can be viewed as a six-dimensional hyperrectangle with dimensions M, Y_0, Z_0, $\alpha_{\rm MLT}$, $\alpha_{\rm ov}$, and D. In order to vary all of these parameters simultaneously and fill the hyperrectangle as quickly as possible, we construct a grid of initial conditions following a quasi-random point generation scheme. This is in contrast to linear or random point generation schemes, over which it has several advantages.

A linear grid subdivides all dimensions in which initial quantities can vary into equal parts and creates a track of models for every combination of these subdivisions. Although in the limit such a strategy will fill the hyperrectangle of initial conditions, it does so very slowly. It is furthermore suboptimal in the sense that linear grids maximize redundant information, as each varied quantity is tried with the exact same values of all other parameters that have been considered already. In a high-dimensional setting, if any of the parameters are irrelevant to the task of the computation, then the majority of the tracks in a linear grid will not contribute any new information.

A refinement on this approach is to create a grid of models with randomly varied initial conditions. Such a strategy fills the space more rapidly, and furthermore solves the problem of redundant information. However, this approach suffers from a different problem: since the points are generated at random, they tend to "clump up" at random as well. This results in random gaps in the parameter space, which are obviously undesirable.

Therefore, in order to select points that do not stack, do not clump, and also fill the space as rapidly as possible, we generate Sobol numbers (Sobol 1967) in the unit 6-cube and map them to the parameter ranges of each quantity that we want to vary. Sobol numbers are a sequence of m-dimensional vectors $x_1 \dots x_n$ in the unit hypercube I^m constructed such that the integral of a real function f in that space is equivalent in the limit to that function evaluated on those numbers,

FIGURE 2.13. A visualization of the model selection process performed on each evolutionary track in order to obtain the same number of models from each track. The blue crosses show all of the models along the evolutionary track as they vary from ZAMS to TAMS in core hydrogen abundance and the red crosses show the models selected from this track. The models were chosen via linear transport such that they satisfy Equation (2.13). For reference, an equidistant spacing is shown with black points.

that is,

$$\int_{I^m} f = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} f(x_i) \tag{2.14}$$

with the sequence being chosen such that the convergence is achieved as quickly as possible. By doing this, we both minimize redundant information and furthermore sample the hyperspace of possible stars as uniformly as possible. Figure 2.14 visualizes the different methods of generating multidimensional grids: linear, random, and the quasi-random strategy that we took. This method applied to initial model conditions was shown in Figure 2.1 with 1- and 2D projection plots of the evolutionary tracks generated for our grid.

2.6.3 Adaptive Remeshing

When performing element diffusion calculations in MESA, the surface abundance of each isotope is considered as an average over the outermost cells of the model. The number of outer cells N is chosen such that the mass of the surface is more than ten times the mass of the $(N+1)^{th}$ cell. Occasionally, this approach can lead to a situation where surface abundances change dramatically and dis-

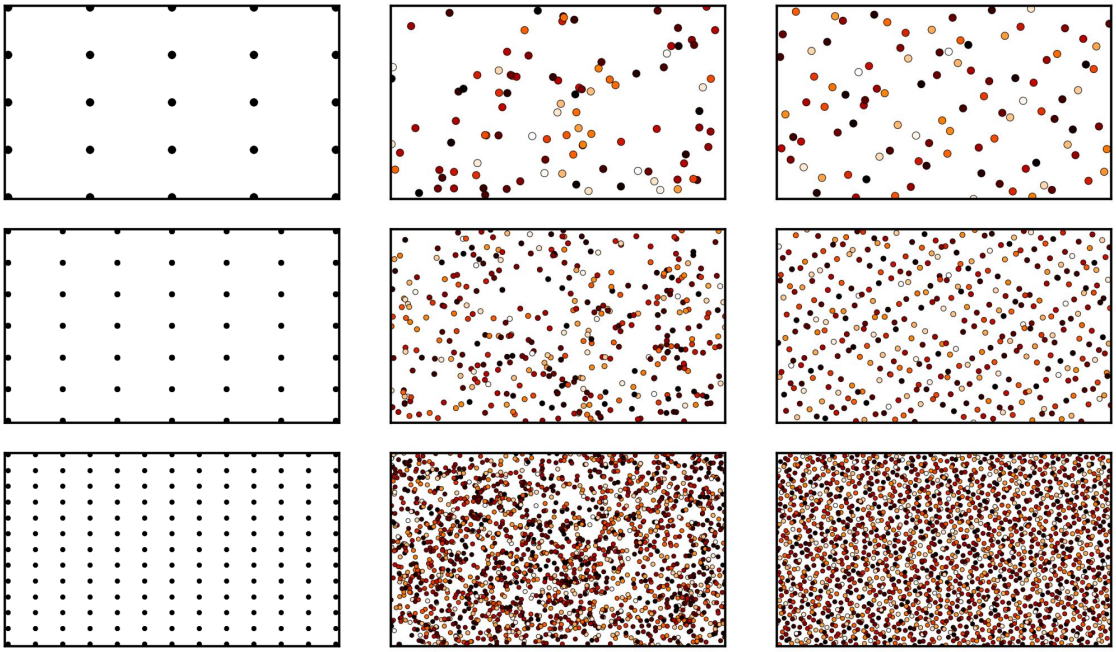


FIGURE 2.14. Results of different methods for generating multidimensional grids portrayed via a unit cube projected onto a unit square. Linear (left), random (middle), and quasi-random (right) grids are generated in three dimensions, with color depicting the third dimension, i.e., the distance between the reader and the screen. From top to bottom, all three methods are shown with 100, 400, and 2000 points generated, respectively.

continuously in a single time-step. These abundance discontinuities then propagate as discontinuities in effective temperatures, surface gravities, and radii. An example of such a difficulty can be seen in Figure 2.15.

Instead of being a physical reality, these effects arise only when there is insufficient mesh resolution in the outermost layers of the model. We therefore seek to detect these cases and re-run any such evolutionary track using a finer mesh resolution. We consider a track an outlier if its surface hydrogen abundance changes by more than 1% in a single time-step. We iteratively re-run any track with outliers detected using a finer mesh resolution, and, if necessary, smaller time-steps, until convergence is reached. The process and a resolved track can also be seen in Figure 2.15.

Some tracks still do not converge without surface abundance discontinuities despite the fineness of the mesh or the brevity of the time-steps, and are therefore not included in our study. These troublesome evolutionary tracks seem to be located only located in a thin ridge of models having sufficiently high stellar mass ($M > M_\odot$), a deficit of initial metals ($Z_0 < 0.001$) and a specific inefficiency of diffusion ($D \simeq 0.01$). A visualization of this can be seen in Figure 2.16.

2.6.4 Evaluating the Regressor

In training the random forest regressor, we must determine how many evolutionary tracks N to include, how many models M to extract from each evolutionary track, and how many trees T to use when growing the forest. As such it is useful to define measures of gauging the accuracy of the random forest so that we may evaluate it with different combinations of these parameters.

By far the most common way of measuring the quality of a random forest regressor is its so-called "out-of-bag" (OOB) score (see e.g. Section 3.1 of Breiman 2001). While each tree is trained on only a subset (or "bag") of the stellar models, all trees are tested on all of the models that they did not see. This provides an accuracy score representing how well the forest will perform when predicting on observations that it has not seen yet. We can then use the scores defined in Section 2.2.3 to calculate OOB scores.

However, such an approach to scoring is too optimistic in this scenario. Since a tree can get models from every simulation, predicting the parameters of a model when the tree has been trained on one of that model's neighbors leads to an artificially inflated OOB score. This is especially the case for quantities like stellar mass, which do not change along the main sequence. A tree that has witnessed neighbors on either side of the model being predicted will have no error when predicting that model's mass, and hence the score will seem artificially better than it should be.

Therefore, we opt instead to build validation sets containing entire tracks that are left out from the training of the random forest. We omit models and tracks in powers of two so that we may roughly maintain the regular spacing that we have established in our grid of models (refer back to Appendices 2.6.2 and 2.6.1 for details).

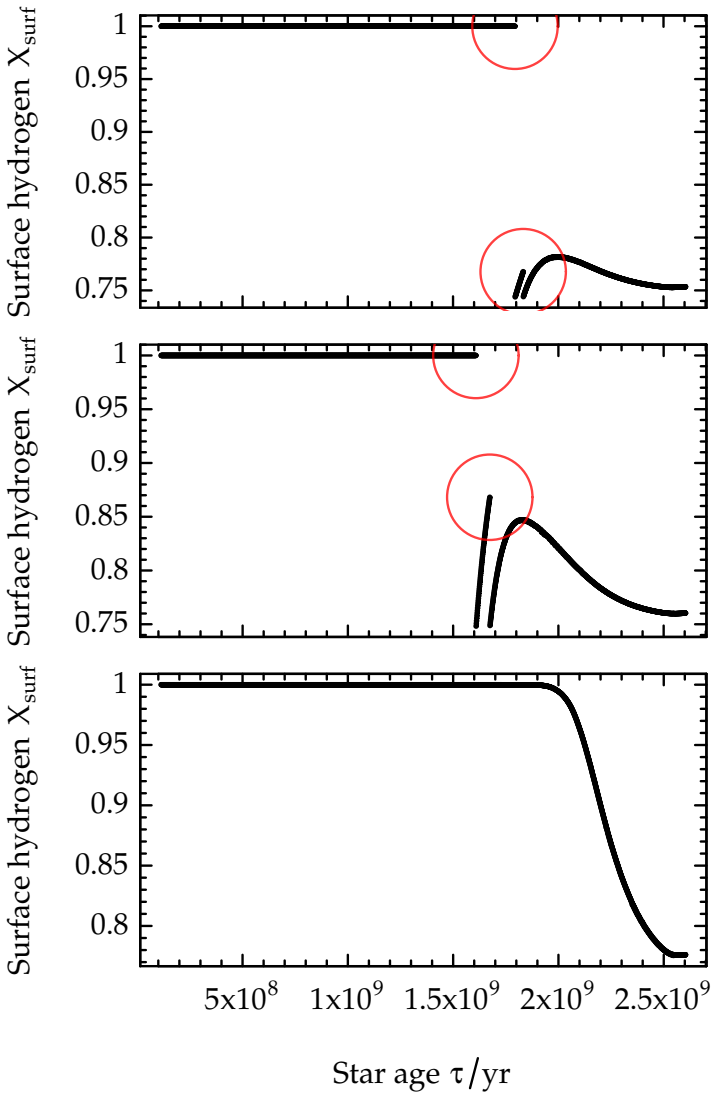


FIGURE 2.15. Three iterations of surface abundance discontinuity detection and iterative remeshing for an evolutionary track. The detected discontinuities are encircled in red. The third iteration has no discontinuities and so this track is considered to have converged.

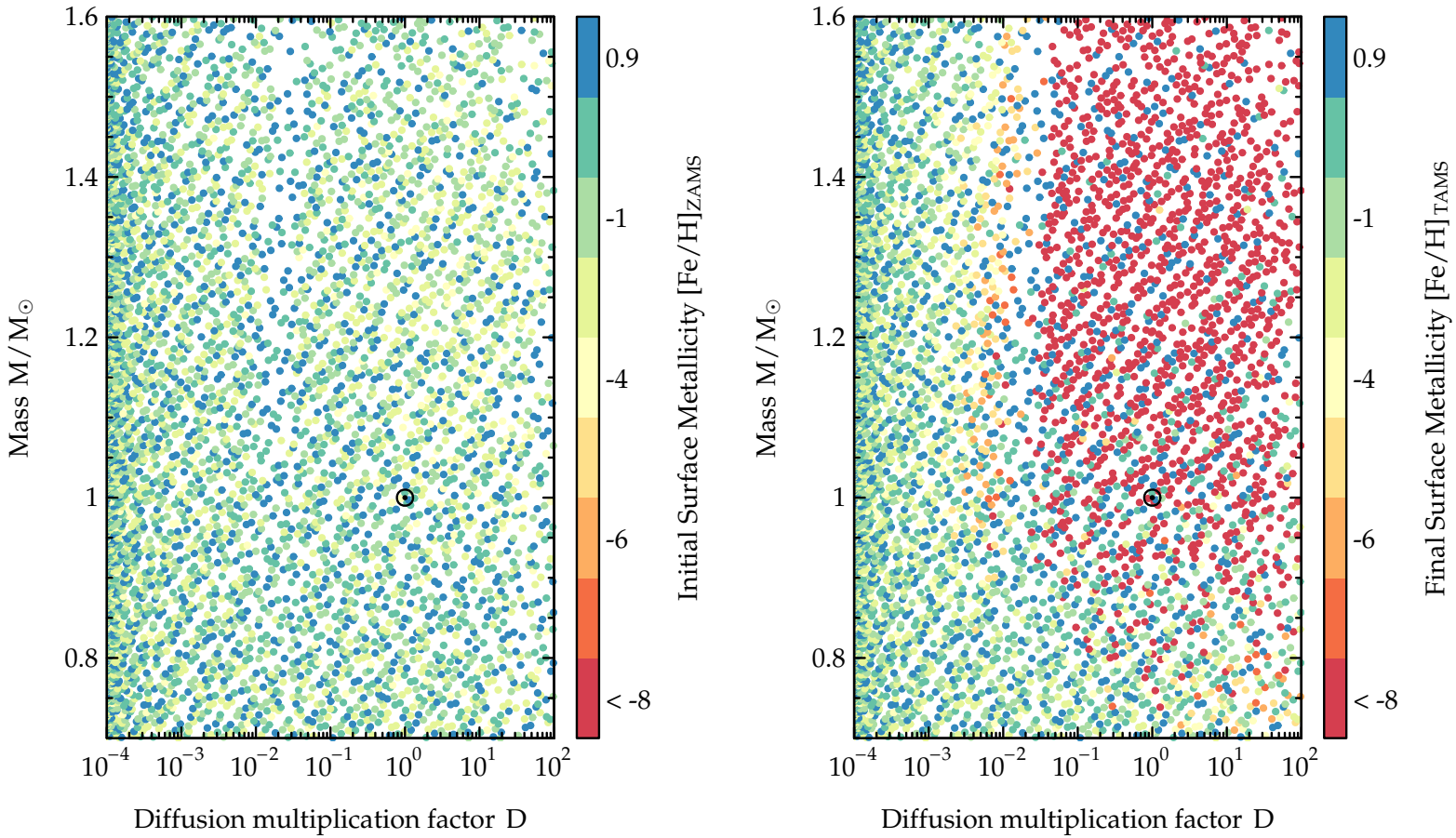


FIGURE 2.16. Stellar mass as a function of diffusion multiplication factor colored by initial surface metallicity (left) and final surface metallicity (right). A ridge of missing points indicating unconverged evolutionary tracks can be seen around a diffusion multiplication factor of 0.01. Beyond this ridge, tracks that were initially metal-poor end their main-sequence lives with all of their metals drained from their surfaces.

only four models per track, which results in a random forest trained on only a few thousand models.

2.6.5 Hare and Hound

Table 2.5 lists the true values of the hare-and-hound exercise performed here, and Table 2.6 lists the perturbed inputs that were supplied to the machine learning algorithm.

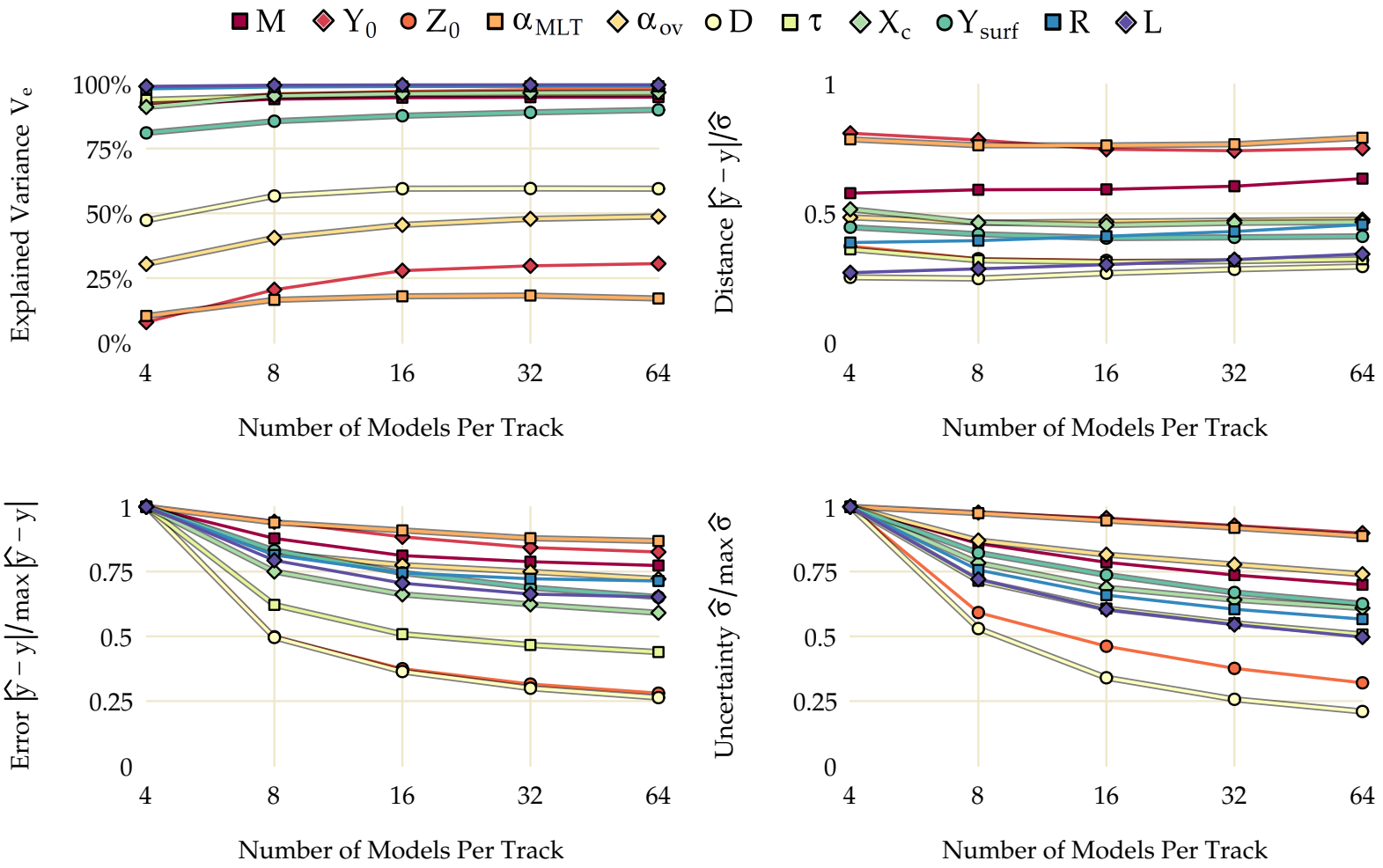


FIGURE 2.17. Explained variance (top left), accuracy per precision distance (top right), normalized absolute error (bottom left), and normalized standard deviation of predictions (bottom right) for each stellar parameter as a function of the number of models per evolutionary track.

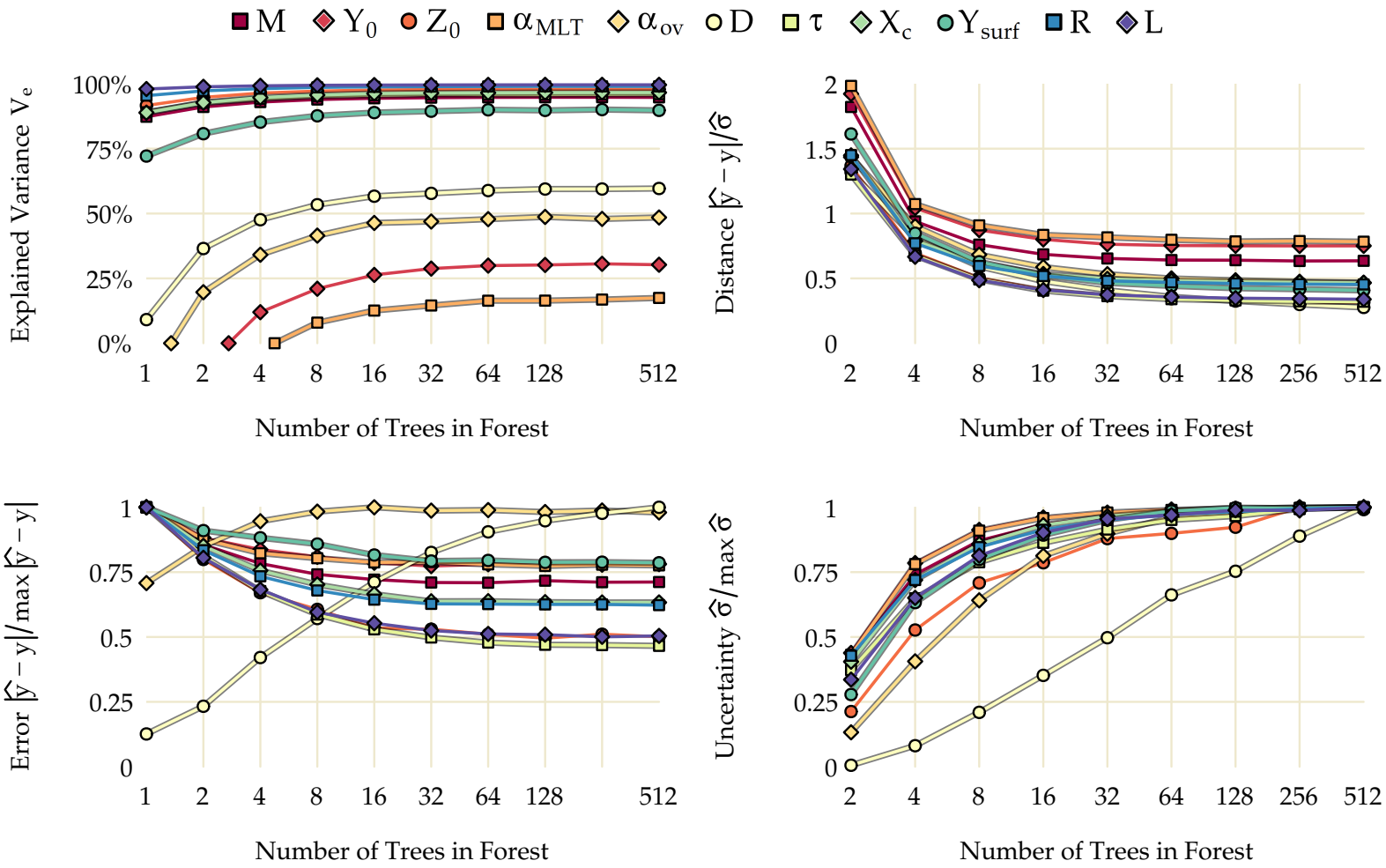


FIGURE 2.18. Explained variance (top left), accuracy per precision distance (top right), normalized absolute error (bottom left), and normalized model uncertainty (bottom right) for each stellar parameter as a function of the number of trees used in training the random forest.

TABLE 2.5. True values for the hare-and-hound exercise.

Model	$R/R_\odot$	$M/M_\odot$	τ	T_{eff}	$L/L_\odot$	[Fe/H]	Y_0	ν_{max}	α_{ov}	D
0	1.705	1.303	3.725	6297.96	4.11	0.03	0.2520	1313.67	No	No
1	1.388	1.279	2.608	5861.38	2.04	0.26	0.2577	2020.34	No	No
2	1.068	0.951	6.587	5876.25	1.22	0.04	0.3057	2534.29	No	No
3	1.126	1.066	2.242	6453.57	1.98	-0.36	0.2678	2429.83	No	No
4	1.497	1.406	1.202	6506.26	3.61	0.14	0.2629	1808.52	No	No
5	1.331	1.163	4.979	6081.35	2.18	0.03	0.2499	1955.72	No	No
6	0.953	0.983	2.757	5721.37	0.87	-0.06	0.2683	3345.56	No	No
7	1.137	1.101	2.205	6378.23	1.92	-0.31	0.2504	2483.83	No	No
8	1.696	1.333	2.792	6382.22	4.29	-0.07	0.2555	1348.83	No	No
9	0.810	0.769	9.705	5919.70	0.72	-0.83	0.2493	3563.09	No	No
10	1.399	1.164	6.263	5916.71	2.15	0.00	0.2480	1799.10	Yes	Yes
11	1.233	1.158	2.176	6228.02	2.05	0.11	0.2796	2247.53	Yes	Yes

TABLE 2.6. Supplied (perturbed) inputs for the hare-and-hound exercise.

Model	T_{eff}	$L/L_\odot$	[Fe/H]	ν_{max}
0	6237 ± 85	4.2 ± 0.12	-0.03 ± 0.09	1398 ± 66
1	5806 ± 85	2.1 ± 0.06	0.16 ± 0.09	2030 ± 100
2	5885 ± 85	1.2 ± 0.04	-0.05 ± 0.09	2630 ± 127
3	6422 ± 85	2.0 ± 0.06	-0.36 ± 0.09	2480 ± 124
4	6526 ± 85	3.7 ± 0.11	0.14 ± 0.09	1752 ± 89
5	6118 ± 85	2.2 ± 0.06	0.04 ± 0.09	1890 ± 101
6	5741 ± 85	0.8 ± 0.03	0.06 ± 0.09	3490 ± 165
7	6289 ± 85	2.0 ± 0.06	-0.28 ± 0.09	2440 ± 124
8	6351 ± 85	4.3 ± 0.13	-0.12 ± 0.09	1294 ± 67
9	5998 ± 85	0.7 ± 0.02	-0.85 ± 0.09	3290 ± 179
10	5899 ± 85	2.2 ± 0.06	-0.03 ± 0.09	1930 ± 101
11	6251 ± 85	2.0 ± 0.06	0.13 ± 0.09	2360 ± 101

3

On the Statistical Properties of the Lower Main Sequence

The contents of this chapter were authored by G. C. Angelou, E. P. Bellinger, S. Hekker, and S. Basu and published in April of 2017 in *The Astrophysical Journal*, 839 (2), 116.[1]

Chapter Summary

Astronomy is in an era where all-sky surveys are mapping the Galaxy. The plethora of photometric, spectroscopic, asteroseismic and astrometric data allows us to characterize the comprising stars in detail. Here we quantify to what extent precise stellar observations reveal information about the properties of a star, including properties that are unobserved, or even unobservable. We analyze the diagnostic potential of classical and asteroseismic observations for inferring stellar parameters such as age, mass and radius from evolutionary tracks of solar-like oscillators on the lower main sequence. We perform rank correlation tests in order to determine the capacity of each observable quantity to probe structural components of stars and infer their evolutionary histories. We also analyze the principal components of classic and asteroseismic observables to highlight the degree of redundancy present in the measured quantities and demonstrate the extent to which information of the model parameters can be extracted. We perform multiple regression using combinations of observable quantities in a grid of evolutionary simulations and appraise the predictive utility of each combination in determining the properties of stars. We identify the combinations that are useful and provide limits to where each type of observable quantity can reveal information about a star. We investigate the accuracy with which targets in the upcoming TESS and PLATO missions can be characterized. We demonstrate that the combination of observations from GAIA and PLATO will allow us to tightly constrain stellar masses, ages and radii with machine learning for the purposes of galactic and planetary studies.

[1] Contribution statement: The work and writing of this chapter were done in equal parts between G. C. Angelou and myself, under the supervision of S. Hekker and S. Basu.

3.1 Introduction

The main sequence is generally considered the most well-understood phase of stellar evolution. Our Sun is a main-sequence star, and its proximity provides a wealth of constraints to the physics that may occur in low-mass counterparts during this phase (e.g., Basu et al. 2015, Basu 2016). Core-hydrogen burning stars are long-lived and hence numerous: indeed, the majority of the stars for which we can resolve parallaxes reside on the main sequence (Gaia Collaboration et al. 2016). Additionally, many stars of this type display stochastic or "solar-like" oscillations that serve to reveal the stellar interior (see, for example, Chaplin and Miglio 2013 for a review on solar-like oscillators). Main-sequence stars are important astrophysical laboratories for testing theories of stellar physics, structure, and evolution; and are a testbed for general physical theories such as nuclear fusion, diffusion, and convection (e.g., Basu and Antia 1994, Spruit et al. 1990).

Despite all of this, however, the ages of main-sequence stars remain uncertain to at least 10%. This uncertainty stems not only from observational imprecision, but also from the inability of observations to fully constrain stellar parameters. Recently, Bellinger & Angelou et al. (2016, BA1 hereinafter) showed that even for stellar models without observational uncertainties, some model attributes of stars—such as their initial helium abundance or efficiency of convection—could not be fully resolved via global information that can be gleaned from their surfaces.

It is well-known that different observable quantities of stars constrain different model properties. For example, in the now-famous Christensen-Dalsgaard diagram (C–D diagram, the so-called "asteroseismic HR diagram"), in which the large frequency separation is plotted against the small frequency separation (Appendix 3.9.1), the large frequency separation covaries with the mass of the star and the small frequency separation covaries with its core-hydrogen abundance. Hence, observing one of these quantities sheds light on its unobservable counterpart. However, to date, a systematic investigation of the extent to which each observable quantity constrains each model property has not been performed.

The equations dictating stellar structure and evolution, and the corresponding microphysics that these equations respond to, give rise to emergent behaviors that are difficult to characterize through examination of the constituting ingredients themselves. To elucidate these opaque relationships, we seek to determine the extent to which observable stellar properties are capable of constraining the internal structures, chemical mixtures, and evolutionary histories of stars. Here we employ the methodology of exploratory data science, a statistical philosophy by which underlying structure in data—simulated or otherwise—can be unearthed.

BA1 used machine learning to build a statistical description of main-sequence stellar evolution. They trained a random forest (RF) of decision trees to learn the relationships that exist between model input parameters and their resultant observable quantities. The technique was developed with particular focus on

the determination of stellar ages. Ages are essential for understanding stellar evolution, characterizing extrasolar planetary systems and advancing models of galactic chemical evolution. Notably, the RF developed by BA1 was able to accurately predict stellar properties such as radii and luminosities using other information collected from the stars in their sample. This illustrates that there is redundant information in the stellar quantities, and that there exist model covariances between these quantities that can be characterized and exploited.

The philosophy employed in BA1 is a departure from the standard practice of stellar model fitting. Ordinarily, stellar parameters of observed stars are sought via χ^2-minimization. The difference in approaches give rise to two points that motivate this paper:

1. Methods based on χ^2-minimization assume that each bit of observed information contributes to the objective of constraining the model properties of a star in an exact proportion to how precisely it has been measured. However, two quantities may be measured independently with no measured covariance, and yet still provide redundant information about the star. The result of such a minimization procedure will therefore be a model that is biased towards that redundant information. The RF developed in BA1, on the other hand, uses the process of statistical bagging to mitigate overfitting of the data (see also Hastie et al. 2009). Here we demonstrate the degree to which the observables carry redundant information about the star.

2. The optimization searches of iterative model finding procedures provide solutions but do not indicate the elements that were important in doing so. The use of regression requires that the observables correlate with those model parameters that we wish to infer. We therefore identify to what extent each observable constrains each model property, and how well the observables must be measured to achieve a desired precision from the regression.

The method developed in BA1 makes use of an artificial intelligence strategy known as supervised learning. The RF that they train seeks relations in evolutionary simulations that enable model properties to be inferred as precisely as possible. Although the RF performs the analysis quickly, precisely, and automatically; supervised machine learning strategies do not provide much insight into how the end result is obtained. The algorithm essentially produces a formula for inferring stellar properties from observations, but one that is too complex for people to use analytically by hand.

Here we incorporate a complementary strategy. We use the counterpart of supervised learning—*unsupervised learning*—to explicitly uncover the relations between observable properties of stars and their model parameters. Hence, BA1 is of a strictly practical nature: stellar parameters can be inferred rapidly without regard for the how or why; and this paper is aimed to further an understanding of the processes actually involved in such a deduction.

In this study we draw heavily from the work presented in BA1. Our analysis initially focuses on elucidating the inherent statistical properties of the grid of stellar models used to train the BA1 RF. We determine the relationships and covariances between a chosen subset of stellar parameters and asteroseismic quantities (see Table 3.1). We carry out simultaneous rank correlation tests on the chosen parameters and identify the necessary, dispensable, and irrelevant information for determining each stellar property. Then, using principal component analysis we reduce the dimensionality of the observable quantities and identify to what extent they reveal information of the model parameters. We subsequently shift the focus of our analysis to how the grid properties are used by the RF and how the choices in the parameters impact on the precision of the regression. We train RFs using all combinations of observable quantities in our dataset. The purpose of this is two-fold: first, it is often the case that we wish to quickly characterize a star from a few easily observed quantities—the Hertzsprung-Russell (HR) diagram serves as the classic example. Training and scoring all possible RF combinations provides a means to *quantify* the utility and predictive power of classical and asteroseismic parameters for inferring stellar properties. Secondly, it provides insight into the relationships determined by machine learning algorithms. Finally, we identify the observational accuracy required to satisfactorily constrain key stellar parameters. We investigate the observable quantities independently as well as consider the measurements expected from the upcoming TESS and PLATO missions.

3.2 Stellar Models and Parameters

We used *Modules for Experiments in Stellar Astrophysics* (MESA, Paxton et al. 2011) to generate a grid of stellar evolutionary sequences initially for the purpose of training a random forest. The tracks are varied in initial mass M, helium Y_0, metallicity Z_0, mixing length parameter α_{MLT}, overshoot coefficient α_{ov}, and atomic diffusion multiplication factor D (see BA1 Section 2.1 for details). Initial model parameters were chosen in a quasi-random fashion from the parameter ranges listed in Table 3.2. In total 5325 evolutionary tracks were evolved from ZAMS to either an age of $\tau = 15$ Gyr or until terminal-age main sequence (TAMS), which we define as having a fractional core-hydrogen abundance X_c below 10^{-3}. We conduct our analysis on a subset of stellar models chosen from each sequence so not to bias our statistics towards longer lived stars or numerically challenging evolutionary tracks. Details of the choice of input physics, grid generation strategy, and model selection procedure are further outlined in BA1. In addition to computing the stellar structure we post process each model with the ADIPLS pulsation package (Christensen-Dalsgaard 2008). P-mode oscillations up to spherical degree $\ell = 3$ below the acoustic cut-off frequency are computed, and from these, frequency separations and separation ratios calculated (see Appendix 3.9.1 for mathematical definitions).

Qty	Definition	Unit
Model Input Parameters		
M	Initial mass	$M_\odot$
Y_0	Initial helium mass fraction	
Z_0	Initial metal mass fraction	
α_{MLT}	Mixing length parameter	
α_{ov}	Overshoot parameter	
D	Diffusion efficiency factor	
Stellar Attributes		
τ	Age	yr
τ_{MS}	Normalized main-sequence lifetime	
M_{cc}	Convective core mass	$M_\odot$
X_{surf}	Surface hydrogen mass fraction	
Y_{surf}	Surface helium mass fraction	
X_c	Central hydrogen mass fraction	
L	Luminosity	$L_\odot$
R	Radius	$R_\odot$
Classical Observables		
[Fe/H]	Surface metallicity	
$\log g$	Logarithmic surface gravity	
T_{eff}	Effective temperature	K
Asteroseismic Observables		
ν_{max}	Frequency of maximum oscillation power	μHz
$\langle\Delta\nu_0\rangle$	Large frequency separation ($\ell = 0$)	μHz
$\langle\delta\nu_{02}\rangle$	Small frequency separation ($\ell = 0, 2$)	μHz
$\langle\delta\nu_{13}\rangle$	Small frequency separation ($\ell = 1, 3$)	μHz
$\langle r_{02}\rangle$	Frequency separation ratio ($\ell = 0, 2$)	
$\langle r_{13}\rangle$	Frequency separation ratio ($\ell = 1, 3$)	
$\langle r_{01}\rangle$	Frequency average ratio ($\ell = 0, 1$)	
$\langle r_{10}\rangle$	Frequency average ratio ($\ell = 1, 0$)	

TABLE 3.1. Definitions of the quantities analyzed in this study separated into four parts: model input parameters, stellar attributes, classical observables, and asteroseismic observables. Asteroseismic definitions are in Appendix 3.9.1. Angled parenthesis indicate the quantity is a calculated weighted median.

There are many quantities that could be included in the current analysis. The 25 parameters we have selected to investigate are listed in Table 3.1. They comprise key asteroseismic and structural quantities and reflect our focus on characterizing the relationships between observable quantities (observables hereinafter) and those variables that allow us to generate detailed stellar models.

We consider two parameters not included in the RF training data. BA1 elected to omit the frequency of maximum oscillation power, ν_{max} (Equation 3.19), in their regression model[2]. This quantity displays a strong correlation with $\langle \Delta\nu_0 \rangle$ (see Figure 3.2 or Hekker et al. 2009, Stello et al. 2009a) and thus offers very little additional information when frequencies are known. We include it in the current analysis because ν_{max} is the simplest global asteroseismic parameter to extract from time-series observations, and because recent work by Themeßl et al. (private communication) indicates that the ν_{max} scaling relation more accurately reproduces stellar parameters in well-constrained binary systems than the $\langle \Delta\nu_0 \rangle$ relation (Equation 3.20). This is despite the fact that $\langle \Delta\nu_0 \rangle$ can be measured more precisely and that the relation can be corrected for temperature and metallicity dependencies (Equation 3.21) to yield greater accuracy (Guggenberger et al. 2016, Sharma et al. 2016).

To complement τ, we have also added normalized main-sequence age, τ_{MS}, which describes how parameters change as a function of stellar evolution. Many low-mass stars in the grid do not reach the terminal-age main sequence (TAMS) before their evolution is stopped. Their main-sequence lifetime is estimated by linearly extrapolating the rate at which the central hydrogen is depleted,

$$\tau_{TAMS} = \frac{\tau_{last}}{1 - (X_{c,\,last}/X_{c,\,init})} \tag{3.1}$$

where τ_{TAMS} is the TAMS age, τ_{last} is the age of the last model in the track, $X_{c,\,last}$ is the corresponding core-hydrogen abundance for that model and $X_{c,\,init}$ is the core-hydrogen abundance of the initial model in that track. For the longest-lived stars we find such an extrapolation is within about 25% of the true TAMS age. The uncertainty in the extrapolation for these stars stems from the fact we

Parameter	Min Value	Max Value	Variation
Mass	0.7	1.6	linear
Y_0	0.22	0.34	linear
Z_0	10^{-5}	10^{-1}	logarithmic
α_{MLT}	1.5	2.5	linear
α_{ov}	10^{-4}	1	logarithmic
D	10^{-6}	10^{2}	logarithmic

TABLE 3.2. Ranges and sampling strategy for the initial model parameters in the BA1 grid.

[2] ν_{max} does have some role in the algorithm developed by BA1, as it is responsible for the location of the Gaussian envelope used to weight and derive averaged/median frequency separations.

only capture the hydrogen depletion in the early part of the main sequence i.e., when $X_{c,\,last} > 0.3$. Estimating the TAMS age in this manner, however, will not impact our conclusions. Large discrepancies are limited to a small number of tracks (192) and differences between the true and extrapolated ages are reduced as $X_{c,\,last} \to 0$. Main sequence lifetime provides insight into the general correlations that develop as a function of main-sequence stellar evolution. Thus it is the monotonicity of τ_{MS} within a given track that is key. The stellar age parameter, on the other hand, is useful for exploring correlations across the whole parameter space.

3.3 Rank Correlation Test

We begin our analysis with a rank correlation test, the purpose of which being to understand the statistical properties of the collective lower main sequence. This is distinct from typical analyses that focus on the evolutionary properties within individual stellar tracks or chemically homogeneous isochrones. By identify-

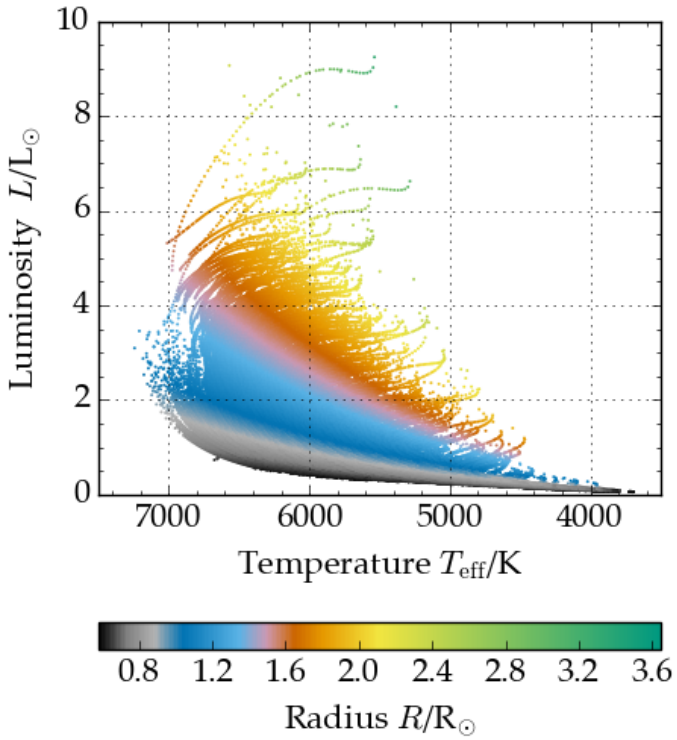


FIGURE 3.1. Hertzsprung-Russell diagram for those tracks in the truncated grid (see text for details). Here each model is coloured by stellar radius.

[2] As principal component analysis is the eigensolution of the correlation (or covariance) matrix, the first eigenvalue indicates the maximum variance in the variables that can be accounted for by a linear model with a single underlying 'factor.' Ordering the parameters in this way demonstrates the direction of the first principal component (PC_1) vector. Figure 3.2 thus offers a visual representation of principal component analysis which we employ in Section 3.4.

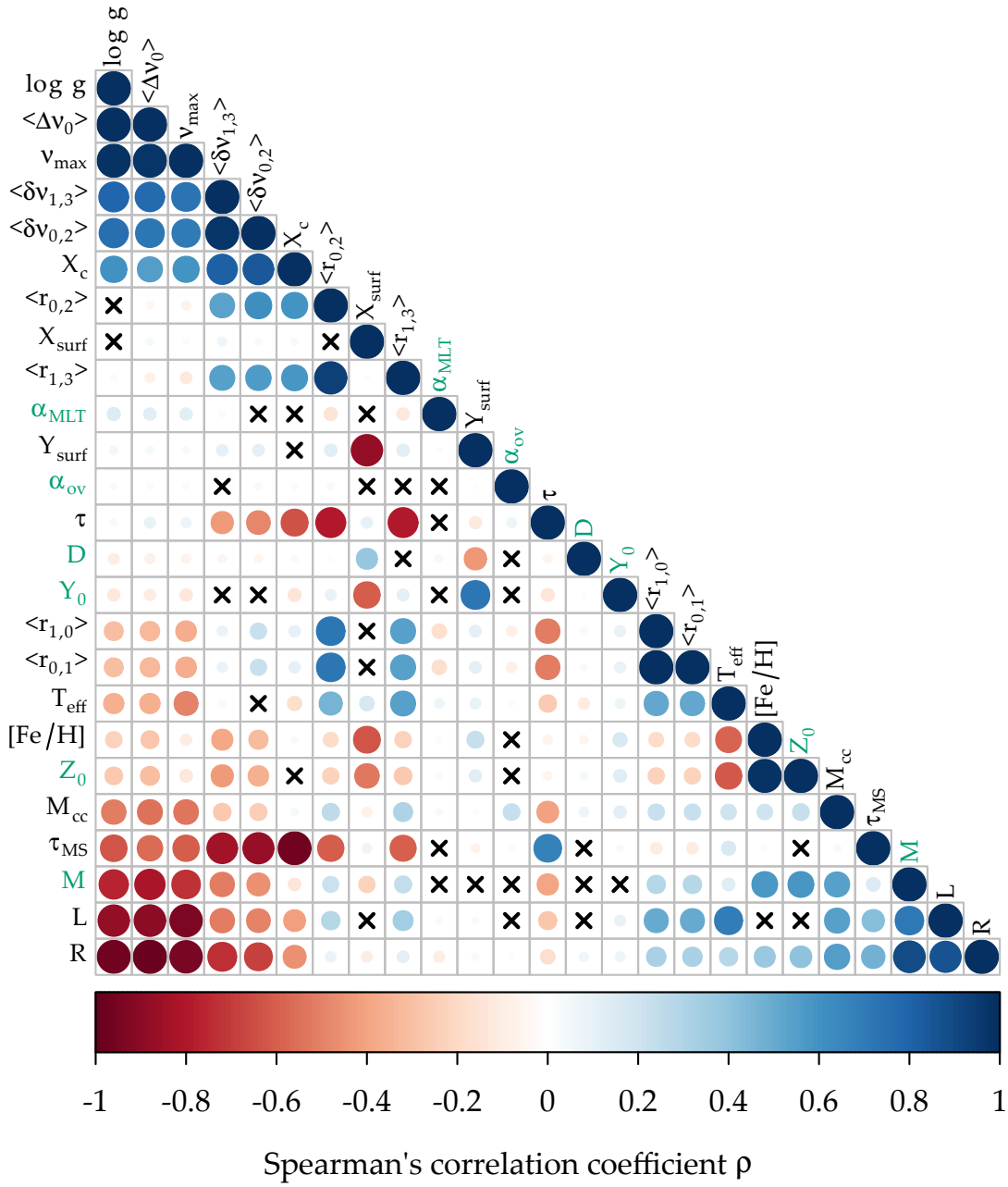


FIGURE 3.2. Spearman rank correlation matrix comprising various stellar and asteroseismic parameters. The quantities are as described in Table 3.1 with model input parameters marked in green. The size and the color of each circle both indicate the magnitude of the Spearman coefficient with red and blue denoting negative and positive correlations respectively. The presence of a cross indicates that the two parameters have failed our significance test; i.e., the correlation is indistinguishable from nil. The variables are ordered according to their correlation with the first eigensolution of the correlation matrix[2].

ing correlations present across the entire parameter space we reveal exploitable relationships available to model fitting and regression methods.

Since many quantities (see Table 3.1) are known to vary in a highly non-linear fashion, we opt to study *rank* statistics. In particular, we replace each quantity by its rank, i.e., an integer representing how big or small a particular quantity is compared to the other models; and calculate Spearman's correlation coefficient ρ between all variables. We further calculate the significance of these correlations (p-values) using the Spearman ρ test. We adopt a conservative significance cut-off of $\alpha = 10^{-5}$ and use the Bonferroni correction to account for the fact that we are making multiple (625) comparisons (e.g., Dunnett 1955).

This analysis allows us to determine whether quantities vary monotonically in the same direction ($\rho \approx 1$), i.e. both increasing or both decreasing; monotonically apart ($\rho \approx -1$), i.e. one increases while the other decreases; or neither ($\rho \approx 0$)[3]. When $|\rho|$ is nearly one, the information from one parameter can be used to determine information about the other. Therefore, this is a valuable tool for probing the relationships that exist in and across evolutionary tracks and determining which model properties can be inferred from which observable quantities.

In the current analysis, we are strictly interested in the relationships expected from the observational data. We apply cuts to the grid computed by BA1 as it spans a wide parameter range[4]. The full set of tracks in the BA1 grid includes models with temperatures exceeding the limit in which solar-like oscillations are thought to develop ($T_{\rm eff} \approx 6700$ K, i.e., the approximate surface temperature beyond which the stellar envelopes are radiative rather than convective). Evolutionary tracks in the training grid with more than half of the constituent models having $T_{\rm eff} > 6700$ K are excluded from the rank correlation analysis. Note that the grid will still contain models with $T_{\rm eff} > 6700$ K if more than half the models in a track display temperatures below this cutoff; there is some chance we may observe such stars. Likewise, we omit tracks where high atomic-diffusion rates significantly drain metals from the surface, i.e., tracks where more than half the models display surface-hydrogen mass fractions > 0.95. The dearth of stars observed at zero metallicity indicates that there are some physical processes not included in our models (e.g., radiative levitation or turbulent diffusion) which inhibit the unabated flow of metals from the stellar surface. This is a common result in models of high-mass stars that include gravitational settling and therefore the process is *ordinarily* suppressed once $M \gtrsim 1.1\ M_\odot$. Metal depletion may also

[3] Spearman's ρ is equivalent to Pearson's r on ranked quantities. We note also that $\rho = 0$ does not necessarily indicate a relationship does not exist; simply that the relationship is not monotonic. A parabolic function for example would result in $\rho = 0$.

[4] When training a RF for the purposes of characterizing stellar systems, sampling the parameter space well beyond the expected ranges of each quantity is prudent. RFs do not extrapolate—doing so would be undesirable anyway—so characterizing a star requires that all of its observations are firmly within the boundaries of the grid used to train the RF. Doing this furthermore avoids pre-conceived biases in the analysis: it allows the observations to dictate the interesting regions of the parameter space rather than limiting the ranges to the values we *expect* the parameters to take.

arise in cases when settling is made to operate extremely efficiently. The removal of these sequences reduces the BA1 training set from 5325 to 2010 evolutionary tracks (truncated grid hereinafter) for the current analysis. In Figure 3.1 we plot the truncated grid in the HR diagram and color the models according to radius.

Figure 3.2 shows the results of the correlation analysis for the truncated grid. We defer correlation analysis on the full grid of models to Appendix 3.9.3. Care is needed when interpreting Figure 3.2. First, it is important to remember that correlation is not transitive[5] (Langford et al. 2001), i.e.,

$$\mathrm{Corr}(X, Y) \wedge \mathrm{Corr}(Y, Z) \not\Rightarrow \mathrm{Corr}(X, Z) \tag{3.2}$$

even when the correlations are due to causative relationships (Veresoglou and Rillig 2015). In fact one can only draw inference on the direction of $\mathrm{Corr}(X, Z)$ in cases when

$$\rho^2_{X,Y} + \rho^2_{Y,Z} > 1 \tag{3.3}$$

(transitive criterion hereinafter).

Second, recall that these correlations hold only for the main sequence. During the main sequence there is generally a positive correlation between, say, L and $T_{\rm eff}$. This relationship will change as the stars evolve further beyond the main-sequence turnoff.

Third, save for correlations with $\tau_{\rm MS}$, the relationships presented here do not describe how parameters correlate internally throughout an evolutionary track. Rather, they describe how they correlate across *all* tracks. For example, as a star ascends the main sequence, luminosity increases and therefore one may expect a strong positive correlation between τ and L. The fact that we report a negative correlation is because higher-mass stars are shorter lived – thus high L corresponds to a lower τ when the whole parameter space is considered. This correlation is in fact stronger in the analysis of the complete grid used in BA1 which we report in Appendix 3.9.3, as our grid truncation preferentially selects against higher-mass stars. Furthermore we note that some initial model variables (M, Y_0, Z_0, $\alpha_{\rm MLT}$, $\alpha_{\rm ov}$ and D; all indicated in green) correlate with other parameters. This would not be the case if we reported correlations within tracks, as these parameters do not change within a given track.

It should be noted that there is some bias present in the grid as the low-mass stars are not computed to the end of their main-sequence lifetime. The strengths of some correlations would change had we considered evolution beyond the age of the Universe.

3.3.1 Interpreting the Correlations

Having set the general context in which to interpret Figure 3.2, we highlight some statistical features of the lower main sequence that can be extracted:

[5] This is irrespective of whether one is using Pearson's r, Spearman's ρ or Kendall's τ.

- Most pairs of parameters with $|\rho| \approx 1$ correspond to well known main-sequence and/or asteroseismic relations. Pairs displaying strong correlations include:

  $\langle \Delta\nu_0 \rangle - \log g$; $\langle \Delta\nu_0 \rangle - \nu_{max}$; $\log g - \nu_{max}$;
  $\langle \Delta\nu_0 \rangle - R$; $\log g - R$; $M - R$;
  $L - R$; $\langle \delta\nu_{02} \rangle - X_c$.

- Figure 3.1 illustrates why T_{eff} and its correlations with R and L are weaker than those listed above. Many of the tracks evolve past the main-sequence turn off before exhausting their core-hydrogen abundance. The change in morphology of the HR diagram and resultant increase in radius impacts on the monotonicity of the respective correlations.

- The mass of the convective core (M_{cc}) displays a moderate negative correlation with age whereas it barely registers a relationship with τ_{MS}. It is the higher-mass and hence shorter-lived stars that preferentially develop convective cores. A negative correlation with age is therefore according to expectations. In stars that burn hydrogen radiatively no correlation will develop between M_{cc} and τ_{MS}. In those stars that burn convectively, the size of the convective core will grow but then recede as the CNO-burning region becomes more centrally condensed. These two factors lead to an (essentially) null result between M_{cc} and τ_{MS}.

- The correlations between τ and the ratios $\langle r_{02} \rangle$ and $\langle r_{13} \rangle$ are stronger than the correlation between τ and X_c. The grid comprises large ranges in mass and metallicity and hence stars at different ages can possess the same X_c, thereby weakening the strength of that correlation. Conversely, as one might expect, τ_{MS} exhibits a stronger relationship with X_c than the ratios.

- The small frequency separations and the asteroseismic frequency ratios strongly correlate with both τ and X_c. The large frequency separation, however, demonstrates a much stronger correlation with X_c than it does with τ. The rate at which stars burn their central fuel will largely depend on their mass, thus the models can attain the same density (which is proportional to the large frequency separation) at a range of ages. Both τ_{MS} and X_c are evolutionary variables and display the expected correlations with $\langle \Delta\nu_0 \rangle$.

- We lack the necessary information to constrain some of the initial model variables. Indeed [Fe/H] provides some constraints on the diffusion efficiency factor D, but there is much degeneracy: a model can attain the same surface Y starting with a low [Fe/H] and low diffusion rate as a track with a high [Fe/H] and high diffusion rate. It is possible that fitting for the base of the convective envelope through seismic analysis of the acoustic glitch signal (Mazumdar et al. 2014, Verma et al. 2014a) could help further constrain these parameters.

Figure 3.2 immediately reveals information about the relationships utilized in the machine learning algorithms. For those parameter pairs that failed the significance test, neither is likely to feature in the regression model that predicts the other, except in a circumstance where a subset of models exhibit a trend that is absent from the general case of all the models being considered together. Conversely, where possible, the regressor will attempt to draw on information from pairs that display the strongest correlations. Quantities such as radius illustrate that there is indeed redundant information in independently measured parameters. This is useful as the observables measured, and their corresponding accuracy, will vary from survey to survey. If a key piece of datum is missing or unreliable, a new regression model can be trained using an appropriate substituted quantity in its place. This requires that the redundant information in the observables are treated correctly, if however they are not, then they will lead to biases in model finding procedures. We explore this point further in the next section.

3.4 Principal Component Analysis

Past studies, particularly Brown et al. (1994), have argued that redundancies and covariances in the stellar observables should be taken into account during any model fitting procedure. They demonstrated one particular method (singular

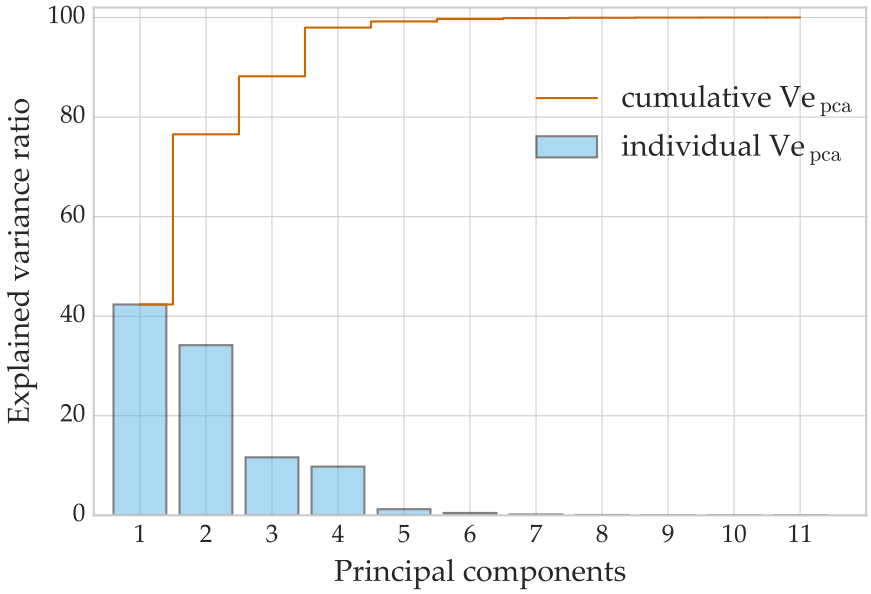


FIGURE 3.3. The explained variance (Ve_{pca}) and cumulative Ve_{pca} of the principal components comprising the *observable* quantities listed in Table 3.1. The figure demonstrates that 98% of the variance in the 11 observational parameters can be explained by four independent components and 99.2% of the variance explained when a fifth component is considered. The Ve_{pca} of each component is also presented in the second column of Table 3.14.

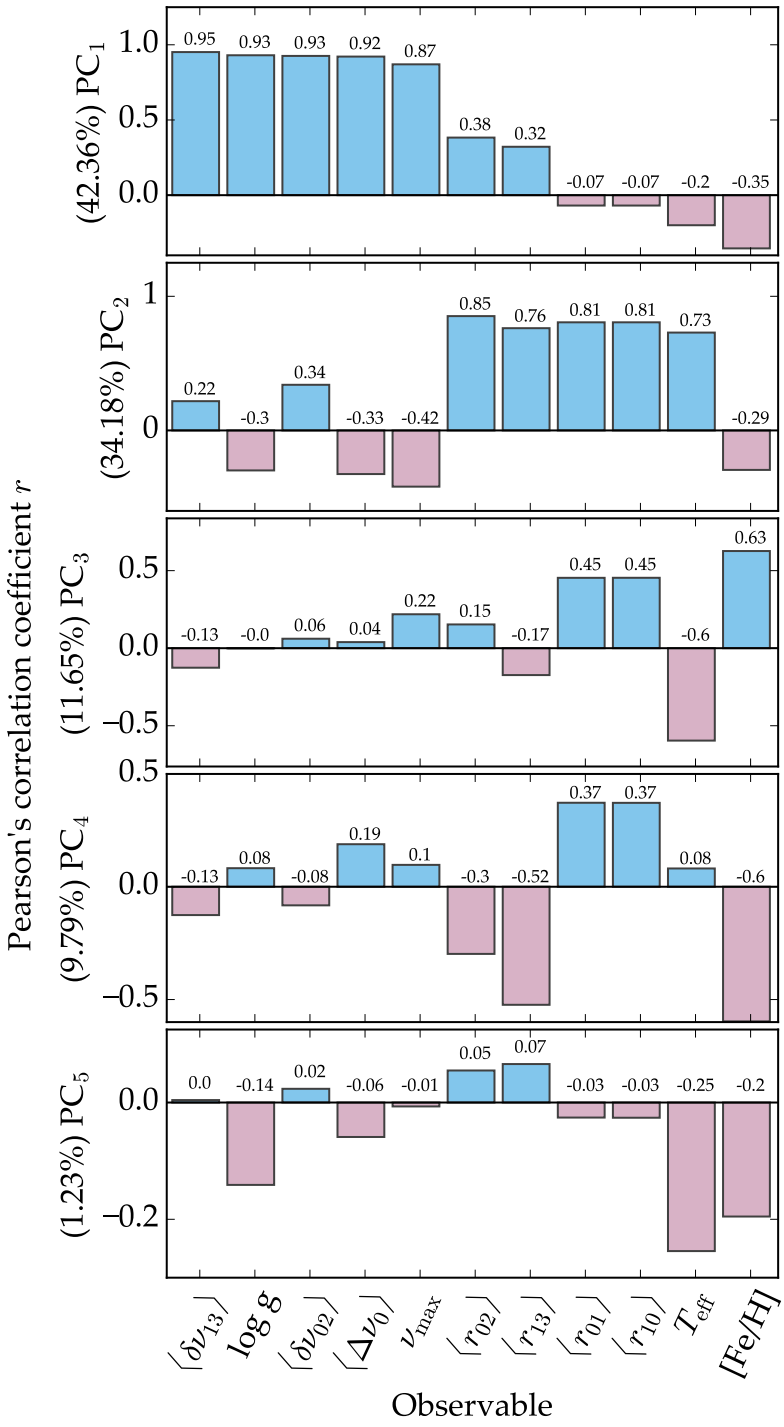


FIGURE 3.4. Pearson correlation strength between the first five principal components and the stellar observables. Quantities are ordered according to their correlation strength with the first principal component. Strong correlations indicate that much of the variance of the quantity is captured by the given PC. Note that the ordinate axes in this figure are on different scales.

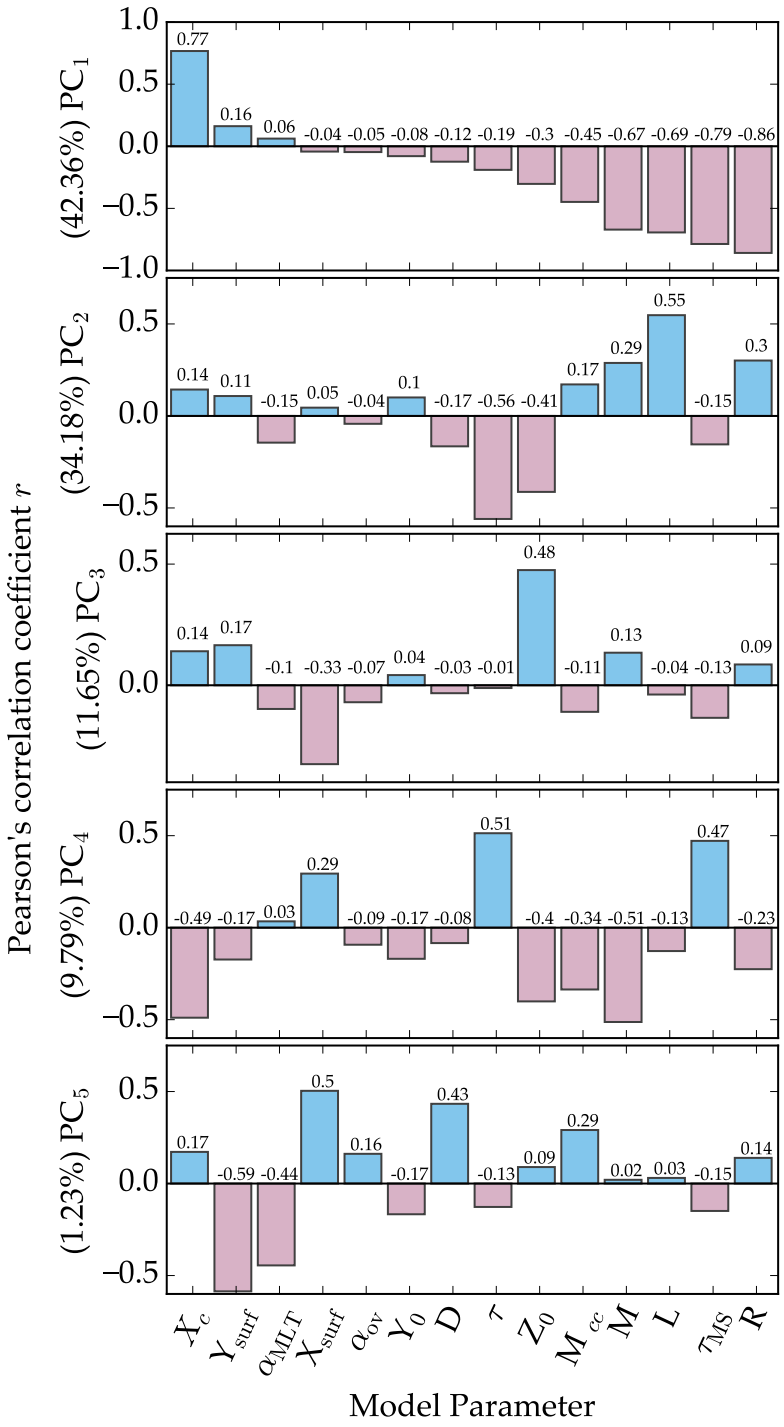


FIGURE 3.5. Pearson correlation strength between the first five principal components and the model parameters (*cf.* Figure 3.4).

value decomposition, SVD hereinafter) of avoiding such biases. In the previous section we identified correlations present in the lower main sequence. Here we demonstrate the degree of redundant information contained in the observables by applying dimensionality reduction. We perform principal component analysis (PCA) in order to discover latent structure in observable stellar quantities such that they may be related more directly—and without redundancy—to parameters of stellar modelling. Through the principal components (PCs) we quantify the extent to which the observables capture information of the model parameters.

A natural strategy for dealing with high-dimensional data is to reduce the dimensionality in search of *latent variables*; i.e., hidden variables that are more useful than the original quantities under consideration. Principal component analysis (PCA) is a technique to transform data into a sequence of orthogonal, and hence independent, linear combinations of the variables. Each successive component is constructed to maximize the residual variance from the original data whilst remaining orthogonal to the previous components. It is a linear transformation in which the change of basis captures the variance contained in original data. If parameters in the data are highly correlated, then PCA can potentially produce a lower-dimensional representation without significant loss of the information. The method can therefore introduce a new set of variables capable of revealing the underlying structure of an originally high-dimensional space.

PCA belongs to a family of matrix decomposition techniques that also include methods such as non-negative matrix factorization and independent components analysis as well as variations such as sparse PCA and kernel PCA. It has previously been employed in an astrophysical context (Baldner and Basu 2008, Murtagh and Heck 1987) along with SVD (Brown et al. 1994, Metcalfe et al. 2009) to handle correlated errors in observational data. The PCs in this work are calculated from the eigensolution of the correlation matrix, the results of which are not scale invariant. We note that PCA can be interpreted as the singular value decomposition of a data matrix in cases where the columns have first been centered by their means. Thus SVD analysis[6] is an alternative method for extracting the PCs (see also Appendix 3.9.7). We indeed compare both methods as a check on our methodology and find that the magnitude of PC scores are identical although the direction (sign) of the vector may differ on occasion.

3.4.1 Explained Variance of the Principal Components

We perform PCA on 11 classical and asteroseismic observables listed in Table 3.1. The chosen parameters reflect the quantities typically extracted[7] from stars in the *Kepler* (Koch et al. 2004, Borucki et al. 2010) field. Our analysis focuses on

[6] This method is in fact more numerically stable but more computationally expensive for extracting PCs.

[7] Radius and luminosity are in some cases observable, but not ubiquitously available in the pre-GAIA era. We concede that the inclusion of $\ell = 3$ modes is an optimistic assumption.

the truncated grid of models[8] (see Section 3.3). The truncated grid reduces our matrix to size 128640×11 on which we perform the PCA (there are $340,800$ models in the full BA1 grid).

The PCs throughout this analysis are calculated from the eigendecomposition of observables in the correlation matrix. Here we wish to explain the variance in the data values rather than their rankings. We employ Pearson's r in the computation of the correlation matrix for the PCA analysis rather than Spearman's ρ. This allows us to transform freely back and forth between the original data space and the space of Pearson PCs.

A given data matrix $\mathbf{X}$ (grid) is of size $n \times p$ where n is the number of models and p is the number of observable parameters. Each entry x_{np} in $\mathbf{X}$ is centered and scaled such that

$$\bar{x}_{np} = (x_{np} - \hat{x_n})/\sigma_{x_n} \tag{3.4}$$

where $\bar{x}_{np}$ is the centered and scaled value, x_{np} is the original entry, $\hat{x}_n$ is the mean of the particular parameter and σ_{x_n} is its standard deviation. With all variables having zero mean and unit variance ($\bar{\mathbf{X}}$), our analysis is equivalent to performing eigendecomposition on the covariance matrix[9]. We compute $\mathbf{\Sigma}$, the matrix of Pearson's r coefficients, between all entries in $\bar{\mathbf{X}}$; and compute the eigenvalues and eigenvectors of $\mathbf{\Sigma}$ to determine the PCs. The eigenvalues, λ_i, of $\mathbf{\Sigma}$ indicate the absolute variance explained by the eigenvectors. We use this to compute the fraction of variance explained by the eigenvector in the dataset, $\mathrm{Ve}_{\mathrm{pca}}$, such that:

$$\mathrm{Ve}_{\mathrm{pca}}\ (\mathrm{PC}_i) = \frac{\lambda_i}{\sum_{i=1}^{p} \lambda_i}, \tag{3.5}$$

where the number of observables in the data matrix, p, is equivalent to the number of principal components we extract.

The $\mathrm{Ve}_{\mathrm{pca}}$ and the cumulative explained variance of the PCs are reported in Figure 3.3 (see also the second column in Table 3.14). Remarkably, we find that 99.2% of the variance in our 11-dimensional observable space can be explained by a space of five components. Hence, observable stellar quantities are clearly highly redundant in what they reveal, as only five dimensions contain original information about the star.

Further insight into the PCs can be gained through correlation analysis between the transformed data (i.e., data matrix projected onto the new PC features) and the original data matrix of observables. Any observable that correlates with a PC contributes to the linear combination of parameters that comprise that PC – the PC is capturing part of the variance in that observable/dimension. Multiple parameters that simultaneously have a large fraction of their variance explained by the same PC, must therefore carry redundant information about the star[10].

[8] To extract a robust interpretation of the PCs we consider different subsets of the BA1 grid (see Appendix 3.9.4).

[9] We are essentially performing the eigendecompostion of the normalized covariance matrix.

[10] The correlation analysis is in general similar to reporting the PC loadings. In PCA loadings are the elements of the eigenvector scaled by the square roots of the respective eigenvalues.

In Figure 3.4 we quantify, through Pearson's r coefficient, the extent to which each observable correlates with the first five PCs. The parameters in the top panel of Figure 3.4 are ordered by their correlation with the first principal component. PC_1 accounts for a significant fraction of the variance in the observables ($Ve_{pca} = 42.36\%$). The top panel of Figure 3.4 reveals that this component correlates very strongly ($r > 0.85$) with ν_{max}, $\langle \Delta\nu_0 \rangle$, $\langle \delta\nu_{02} \rangle$, $\langle \delta\nu_{13} \rangle$, and $\log g$. The strong correlations imply that the basis vector captures most of the variance across the five parameters simultaneously and points to a common latent variable.

3.4.2 Interpreting the Principal Components

In Figure 3.4 and Figure 3.5 we plot the results of correlation analysis between all parameters in the grid and the transformed observables (PCs). The figures offer a quantitative overview of the PCs allowing us to identify what interpretable features the PCs have captured. We have seen that Figure 3.4 demonstrates the extent to which each observable correlates with the first five PCs, similarly Figure 3.5 demonstrates how the principal components correlate with the model parameters. The corresponding correlation coefficients between the parameters and *all* PCs are listed in Tables 3.15 & 3.16.

Any interpretation of the PCs based on Figures 3.4 and 3.5 are only valid for the truncated grid of models to which this PCA has been applied. For results on other sub grids we refer the reader to Appendices 3.9.4 and 3.9.6. We draw upon the figures for generality in the discussion section (Section 3.7).

Information about *direct* correlations between parameters can be extracted from PCA which further helps with interpreting the underlying features. Any two parameters that correlate with a given principal component and meet the transitive criterion will be positively correlated if they both have the same sign with respect to the PC, and negatively correlated if their signs differ.

As is often the case with PCA, the first few principal components can be interpreted as describing the large-scale physical behavior of the system. We interpret that the underlying feature that PC_1 captures is straightforwardly the stellar radius. This is the physical property that has the greatest impact on the observables. From PC_1 in Figures 3.4 and 3.5 we can infer (from the transitive criterion) that as a star evolves along the main sequence, i.e., τ_{MS} increases or X_c decreases, radius (and for the most part L) will increase. The consequence of increasing radius being ν_{max}, $\langle \Delta\nu_0 \rangle$, $\langle \delta\nu_{02} \rangle$, $\langle \delta\nu_{13} \rangle$, $\log g$ all decrease and thus their variance is explained by PC_1. We note that this PC also correlates with M as stars with larger M will have larger radii.

PC_2 can be interpreted as a 'core-surface' feature. PC_2 correlates strongly with different combinations of seismic ratios and small frequency separations. With strong weightings from the core it is no surprise that PC_2 features a

The elements of the eigenvector are coefficients that indicate the weighting of the original data parameters that combine to form that PC. As we have centred and scaled the data before performing the PCA, the correlation coefficients are equivalent to the loadings.

moderate-to-strong correlation with τ. This direction of maximal variance comprises information from all the observables and correlates with (mostly) all the dependent model variables further suggesting some form of time evolution. The information from the surface is provided by T_{eff}. There is a degree to which the variance in T_{eff} is captured by the time-evolutionary aspect of this component. However PC_2 also displays a moderate correlation with the time-independent Z_0 and thus there is a second aspect to PC_2. Z_0 dictates the temperature at the surface through opacities and nuclear burning in the core.

PC_3 appears to have the role of capturing the more extreme models in the grid. In the truncated grid the correlations with [Fe/H] and T_{eff} suggest that the focus of this PC to account for the variance in the observations imparted by low-metallicity models.

PC_4 appears to be a secondary 'core-surface' feature much like PC_2. It uses surface information, in this case [Fe/H], in conjunction with some information from the core in the form of the $\langle r_{02} \rangle$ and $\langle r_{13} \rangle$ ratios.

PC_5 encapsulates the mixing processes that impact upon the surface abundances of the star but it is only required to explain a small fraction of the total variance in the data.

3.4.3 Inferring Stellar Parameters

The dimensionality reduction achieved by the PCA quantifies the degree of redundancy in the stellar observables alluded to by Figure 3.2. However, we also wish to quantify the extent to which the observed stellar properties constrain the internal structures and chemical mixtures of the star, i.e., the model properties.

In our application of RF regression the machine tries to fit for each model parameter, the success of which we can appraise (see Section 3.5). Here we conduct a more fundamental evaluation: how well can we capture the variance in the model parameters simply by explaining the variance in the observed data? In other words: having removed the redundancies, to what extent is information of the model parameters encoded in the observables? We hence devise a score, Λ, such that:

$$\Lambda(X) = \sum_{i=1}^{p} r(X, PC_i)^2 \tag{3.6}$$

where X is the parameter of interest, p is the number of PCs (11 in our case) and $r(X, PC_i)$ is the Pearson coefficient between the parameter and the PC. As we centred and scaled our data before computing the correlation matrix and extracting the PCs, the $\Lambda(X)$ score is equivalent to summing the square of the PC loadings. The square of each loading indicates the variation in an observable that is explained by the component. A useful property of having scaled our data is that $\Lambda(X) = 1$ for each of our observables. We demonstrate these properties further in Appendix 3.9.7.

In Figure 3.5 we projected the parameter space of our model quantities onto the PC space. Whilst these are not the optimum vectors to explain our model

parameters, that is not their purpose; we instead wish to determine what we can learn about the model quantities by understanding the observables. As the square of the correlation coefficients (loadings) will indicate the fraction of explained variance for the parameter by a given PC, determining the $\Lambda(X)$ score for the model parameters gives an indication of the extent the model data are retrievable from the observables.

In Table 3.3 we list the Λ score for each of the model parameters in Table 3.1. Parameters with larger Λ scores have much of their variance captured by the linear models used to explain the observables. We expect to be able to infer parameters such as R, L and τ_{MS} with a great deal of confidence through regression. Parameters with intermediate values of Λ (τ, M_{cc}) we can expect to recover with some success by employing more sophisticated modelling, however, it is not clear that there is enough information contained in the observables to always do so. In cases with the lowest values of Λ, such as the initial model parameters α_{MLT}, Y_0 and α_{ov}, explaining the variance in the observables does not explain the variance in the model parameters. New observables that provide independent information about the star are required to recover these parameters with higher confidence. Fitting the acoustic glitch for example may (eventually) provide constraints on the degree of convective envelope overshoot or atomic diffusion (Verma et al. 2017).

Parameter	Λ_{param}
R	0.97
L	0.96
X_c	0.94
τ_{MS}	0.93
M	0.91
τ	0.79
Z_0	0.73
M_{cc}	0.61
Y_{surf}	0.50
X_{surf}	0.48
α_{MLT}	0.38
Y_0	0.31
D	0.29
α_{ov}	0.08

TABLE 3.3. The Λ score is a sum of the squares of $r_{PC,\,param}$. Any parameter with high Λ is explained well by a linear model and can be confidently inferred. We have insufficient information to constrain those parameters with the lowest Λ.

3.5 Quantifying the Utility of Stellar Observables

There is certainly value and a degree of intuition in dimensionality reduction. PCA has demonstrated the significant information redundancy in our data. It has also allowed us to identify information from the model parameters manifested in the observables, and indicated to what extent those parameters can be extracted. We now turn to another strategy of exploratory data science, which is to let machine learning algorithms fit complicated models to the data. As we shift our focus from *what* information is present to *how* it can be exploited, we transition from unsupervised to supervised learning methods.

In the PCA we determined orthogonal vectors that are the best fit to the observables. Here we utilize a RF to perform non-parametric, multiple regression in order to create the best functions capable of inferring each stellar parameter. With this particular form of supervised learning the relationships between observables and model parameters remain hidden. Though some insight into the regression function can be gained through examination of the feature importances, the tree topology makes further interpretation difficult. We thus seek to elucidate the RF's decision making processes by appraising how well different combinations of parameters can predict the quantities in Table 3.1.

This approach not only illustrates the RF's ability to recover missing observational data, say for a rapid stellar evolution calculation, but also systematically *quantifies* the usefulness of each parameter in predicting all other quantities in the limit of perfect information. It is analogous to a seismic inversion in that it demonstrates the inherent uncertainty with which information can be reconstructed from the available observations. Whereas PCA serves to remove the redundant stellar information in the parameters, the analysis here is designed to highlight them.

Using the full grid of BA1 models, we perform multiple regression on every unique combination of observables in Table 3.1. We omit those combinations that contain the quantity we are training for and include models with R and L as observables, resulting in the calculation of $49,153$ RFs.

We divide the full grid into a testing ($\approx 15,000$ models) and training set as per the method ascribed in Appendix D of BA1 so not to over-estimate the performance of the regression. We perform two-fold cross-validation on each RF and, as in BA1, measure their success on the test data with an explained variance score, V_e:

$$V_e = 1 - \frac{\mathrm{Var}\{y - \hat{y}\}}{\mathrm{Var}\{y\}}. \tag{3.7}$$

Here y is the *true* value we want to predict, $\hat{y}$ is the predicted value from the random forest, and Var is the variance. This score tells us the extent to which the regressor has reduced the variance in the parameter it is predicting with a score of one implying that the model predicts all values with zero error. This is a different but equivalent definition by which to measure the same quantity in Equation (3.5). We have adopted the same notation as BA1 for evaluating the RF which we use to distinguish from the definition used in the PCA (Section 3.4).

We also provide a measure of the 'typical' error in the predictions, $\mu(\epsilon)$, which is calculated by averaging the absolute difference (ϵ) between the predicted and true values for each parameter. More formally:

$$\mu(\epsilon) = \frac{1}{n}\sum_{i=1}^{n} |\hat{y}_i - y_i|, \tag{3.8}$$

where n is the number of models in the test data. Through $\mu(\epsilon)$ we provide an indicative error associated with the regression model, over the whole parameter space, and in units of the quantity of interest.

The best combinations of parameters for inferring each quantity of interest are listed in Table 3.5. We present combinations of up to five parameters after which there is negligible improvement to the predictions. We mark with a dash the occasions where the regressor is unable to produce a positive V_e score. It is important to remember that while a score of one implies a perfect predictor, any $V_e < 1$ implies there is still *some* error in the model. We thus opt for truncation rather than rounding when listing the scores. Predictions of the seismic quantities are omitted here. They strongly co-vary and are easily recovered when other seismic parameters are known; they are discussed separately in Section 3.5.4. Their strong covariances also mean that many of the ratios and separations used in the regression models are interchangeable (e.g., $\langle r_{02}\rangle$ for $\langle r_{13}\rangle$ or $\langle r_{01}\rangle$ for $\langle r_{10}\rangle$) resulting in negligible differences to our two scores.

Many of the RFs we trained do not provide a satisfactory regression model for the quantity we are training for. Below we provide a deeper analysis for some of the more interesting results, focusing primarily on the predictions of ages and surface abundances.

3.5.1 Ages

The current exercise allows us to evaluate the theoretical limit in which parameter pairs, such as those used in the C–D diagram, can constrain stellar ages. Recall that there are six initial model parameters varied simultaneously in the BA1 grid. Describing a six dimensional parameter space with two quantities invariably leads to degenerate solutions for age and necessarily high uncertainties. The parameter pairs that offer similarly the best constraints on τ are listed in Table 3.4. The combination of $\langle r_{02}\rangle$ and ν_{max} marginally provide the best probe, explaining the largest fraction of the variance and inferring ages with uncertainty $\mu(\epsilon) = \pm 642$ Myr.

This is in comparison to $\mu(\epsilon) = \pm 701$ Myr for $\langle \Delta\nu_0\rangle$ and $\langle \delta\nu_{02}\rangle$ as per the C–D diagram. In Table 3.4 we also include results from regression calculated with the PCs and find they perform comparably well. The results here omit any uncertainty stemming from the surface effect suggesting that the $\langle r_{02}\rangle$ and ν_{max} pair are indeed the preferable choice.

It is clear from Tables 3.5 and 3.4 how important the small frequency separation and frequency ratios are for the determination of stellar ages on the MS.

TABLE 3.4. The best two-parameter combinations of observables for constraining stellar age. Below the dividing horizontal line we include the best spectroscopic pair for comparison as well as $\log g$ – $\langle \Delta\nu_0 \rangle$ to highlight the necessity of the small frequency separation in determining stellar ages. The BA1 grid is varied in six dimensions and with such a high-dimensional parameter space the quantities in the C–D diagram (fifth row) constrain age with 'typical' uncertainty of 701 Myr.

Parameters		V_e	$\mu(\epsilon)$ [Gyr]
$\langle r_{02} \rangle$	ν_{max}	0.844	0.642
$\langle r_{02} \rangle$	$\log g$	0.833	0.683
$\langle r_{13} \rangle$	ν_{max}	0.827	0.711
$\langle r_{02} \rangle$	$\langle \Delta\nu_0 \rangle$	0.825	0.694
$\langle \Delta\nu_0 \rangle$	$\langle \delta\nu_{02} \rangle$	0.824	0.701
$\langle r_{02} \rangle$	$\langle \delta\nu_{02} \rangle$	0.821	0.701
PC_2	PC_8	0.788	0.767
PC_2	PC_4	0.776	0.762
$\log g$	$\langle \Delta\nu_0 \rangle$	0.481	1.29
$\log g$	T_{eff}	0.321	1.53

TABLE 3.5. The best combinations of observables for constraining the non-seismic parameters in Table 3.1. For each combination we provide the V_e score (Equation 3.7) and $\mu(\epsilon)$ score (Equation 3.8, given in the units indicated by the predicted quantity column).

Predicted	One Parameter			Two Parameters			Three Parameters			Four Parameters			Five Parameters		
Quantity	Observable	V_e	$\mu(\epsilon)$	Observables	V_e	$\mu(\epsilon)$	Observables	V_e	$\mu(\epsilon)$	Observables	V_e	$\mu(\epsilon)$	Observables	V_e	$\mu(\epsilon)$
$R/R_\odot$	$\langle\Delta\nu_0\rangle$	0.955	0.046	$\langle\Delta\nu_0\rangle$, ν_{max}	0.985	0.027	$\langle\Delta\nu_0\rangle$, ν_{max}, T_{eff}	0.999	0.009	$\langle\Delta\nu_0\rangle$, ν_{max}, T_{eff}, log g	0.999	0.008	$\langle\Delta\nu_0\rangle$, ν_{max}, T_{eff}, log g, $\langle r_{10}\rangle$	0.999	0.008
log g	$\langle\Delta\nu_0\rangle$	0.86	0.046	T_{eff}, ν_{max}	0.999	0.004	T_{eff}, ν_{max}, [Fe/H]	0.999	0.003	T_{eff}, ν_{max}, [Fe/H], $\langle r_{13}\rangle$	0.999	0.002	T_{eff}, ν_{max}, [Fe/H], $\langle r_{02}\rangle$, $\langle r_{13}\rangle$	0.999	0.002
$L/L_\odot$	T_{eff}	0.739	1.583	T_{eff}, $\langle\Delta\nu_0\rangle$	0.993	0.254	T_{eff}, $\langle\Delta\nu_0\rangle$, ν_{max}	0.999	0.13	T_{eff}, $\langle\Delta\nu_0\rangle$, ν_{max}, $\langle r_{10}\rangle$	0.999	0.136	T_{eff}, $\langle\Delta\nu_0\rangle$, ν_{max}, log g, $\langle r_{10}\rangle$	0.999	0.135
T_{eff}/K	[Fe/H]	0.298	1216	log g, ν_{max}	0.989	104	log g, ν_{max}, $\langle r_{01}\rangle$	0.991	95	log g, ν_{max}, $\langle r_{01}\rangle$, $\langle\delta\nu_{13}\rangle$	0.992	96	log g, ν_{max}, $\langle r_{01}\rangle$, $\langle\Delta\nu_0\rangle$, $\langle\delta\nu_{13}\rangle$	0.992	96
Z_0	[Fe/H]	0.927	0.003	[Fe/H], $\langle\delta\nu_{02}\rangle$	0.96	0.002	[Fe/H], T_{eff}, $\langle\Delta\nu_0\rangle$	0.982	0.001	[Fe/H], T_{eff}, $\langle\Delta\nu_0\rangle$, $\langle r_{13}\rangle$	0.986	0.001	[Fe/H], T_{eff}, $\langle\Delta\nu_0\rangle$, $\langle r_{01}\rangle$, $\langle r_{13}\rangle$	0.987	0.001
$M/M_\odot$	$\langle\Delta\nu_0\rangle$	0.348	0.157	$\langle\Delta\nu_0\rangle$, log g	0.857	0.072	$\langle\Delta\nu_0\rangle$, T_{eff}, ν_{max}	0.982	0.022	$\langle\Delta\nu_0\rangle$, log g, ν_{max}, T_{eff}	0.986	0.02	$\langle\Delta\nu_0\rangle$, log g, ν_{max}, T_{eff}, $\langle r_{10}\rangle$	0.982	0.024
τ_{MS}	$\langle\delta\nu_{02}\rangle$	0.543	0.147	$\langle r_{02}\rangle$, $\langle r_{01}\rangle$	0.846	0.077	$\langle\Delta\nu_0\rangle$, ν_{max}, $\langle r_{13}\rangle$	0.957	0.038	$\langle r_{02}\rangle$, ν_{max}, $\langle r_{10}\rangle$, T_{eff}	0.977	0.025	$\langle r_{02}\rangle$, ν_{max}, $\langle r_{10}\rangle$, T_{eff}, [Fe/H]	0.981	0.021
X_c	$\langle\delta\nu_{02}\rangle$	0.508	0.113	ν_{max}, $\langle r_{13}\rangle$	0.842	0.062	ν_{max}, $\langle r_{13}\rangle$, $\langle\Delta\nu_0\rangle$	0.958	0.031	ν_{max}, $\langle r_{13}\rangle$, $\langle\Delta\nu_0\rangle$, $\langle r_{10}\rangle$	0.978	0.023	ν_{max}, $\langle r_{13}\rangle$, $\langle\Delta\nu_0\rangle$, log g, $\langle r_{10}\rangle$	0.979	0.022
τ (Gyr)	$\langle r_{02}\rangle$	0.645	0.995	$\langle r_{02}\rangle$, ν_{max}	0.844	0.642	$\langle r_{13}\rangle$, ν_{max}, $\langle r_{10}\rangle$	0.907	0.468	$\langle r_{02}\rangle$, T_{eff}, $\langle r_{01}\rangle$, $\langle\Delta\nu_0\rangle$	0.931	0.332	$\langle r_{02}\rangle$, ν_{max}, $\langle r_{01}\rangle$, T_{eff}, [Fe/H]	0.943	0.282
X_{surf}	[Fe/H]	0.655	0.051	[Fe/H], log g	0.772	0.041	[Fe/H], $\langle\Delta\nu_0\rangle$, ν_{max}	0.895	0.027	[Fe/H], $\langle\Delta\nu_0\rangle$, T_{eff}, $\langle r_{02}\rangle$	0.928	0.024	[Fe/H], $\langle\Delta\nu_0\rangle$, T_{eff}, $\langle r_{02}\rangle$, ν_{max}	0.936	0.022
$M_{cc}/M_\odot$	—	—	—	$\langle r_{13}\rangle$, $\langle\delta\nu_{02}\rangle$	0.679	0.015	$\langle r_{13}\rangle$, ν_{max}, $\langle r_{10}\rangle$	0.862	0.009	$\langle r_{13}\rangle$, ν_{max}, $\langle r_{10}\rangle$, T_{eff}	0.908	0.007	$\langle r_{13}\rangle$, ν_{max}, $\langle r_{10}\rangle$, T_{eff}, [Fe/H]	0.928	0.006
Y_{surf}	[Fe/H]	0.597	0.052	[Fe/H], log g	0.736	0.041	[Fe/H], $\langle\Delta\nu_0\rangle$, ν_{max}	0.887	0.025	[Fe/H], $\langle\Delta\nu_0\rangle$, $\langle r_{02}\rangle$, T_{eff}	0.916	0.024	[Fe/H], $\langle\Delta\nu_0\rangle$, $\langle r_{02}\rangle$, T_{eff}, ν_{max}	0.927	0.022
Y_0	—	—	—	$\langle\Delta\nu_0\rangle$, ν_{max}	0.077	0.027	$\langle\Delta\nu_0\rangle$, ν_{max}, [Fe/H]	0.471	0.02	$\langle\Delta\nu_0\rangle$, ν_{max}, [Fe/H], log g	0.536	0.019	$\langle\Delta\nu_0\rangle$, ν_{max}, [Fe/H], log g, $\langle\delta\nu_{13}\rangle$	0.625	0.017
α_{ov}	—	—	—	$\langle r_{13}\rangle$, $\langle r_{02}\rangle$	0.231	0.089	$\langle r_{13}\rangle$, $\langle r_{10}\rangle$, ν_{max}	0.44	0.075	$\langle r_{13}\rangle$, $\langle r_{10}\rangle$, ν_{max}, T_{eff}	0.524	0.068	$\langle r_{13}\rangle$, $\langle r_{10}\rangle$, ν_{max}, T_{eff}, [Fe/H]	0.55	0.067
D	—	—	—	[Fe/H], $\langle\delta\nu_{02}\rangle$	0.022	5.393	[Fe/H], T_{eff}, $\langle\Delta\nu_0\rangle$	0.295	4.483	[Fe/H], T_{eff}, $\langle r_{02}\rangle$, $\langle\Delta\nu_0\rangle$	0.446	3.706	[Fe/H], T_{eff}, $\langle r_{02}\rangle$, log g, $\langle r_{10}\rangle$	0.519	3.333
[Fe/H]	—	—	—	ν_{max}, log g	0.179	2.777	ν_{max}, log g, $\langle r_{02}\rangle$	0.273	2.439	ν_{max}, log g, $\langle r_{02}\rangle$, $\langle r_{10}\rangle$	0.309	2.312	ν_{max}, log g, $\langle r_{02}\rangle$, $\langle r_{01}\rangle$, $\langle r_{13}\rangle$	0.312	2.277
α_{MLT}	—	—	—	— —	—	—	T_{eff}, ν_{max}, $\langle\delta\nu_{13}\rangle$	0.069	0.234	T_{eff}, ν_{max}, [Fe/H], $\langle r_{02}\rangle$	0.201	0.211	T_{eff}, ν_{max}, $\langle r_{01}\rangle$, [Fe/H], $\langle r_{13}\rangle$	0.229	0.207

If we limit the combinations to the classical observables, we find that $\log g$ and $T_{\rm eff}$ can explain just 32.1% of the variance in τ with uncertainty $\mu(\epsilon) = \pm 1.5$ Gyr across the whole grid. The introduction of the large separation offers little improvement. The parameter pair $\log g$ and $\langle \Delta\nu_0 \rangle$ explain 48.1% of the variance with $\mu(\epsilon) = \pm 1.29$ Gyr. If we permit the RF to draw upon five observables for its regression model, some of the degeneracy in τ is lifted. The last column in Table 3.5 indicates that the RF can reduce the average uncertainty in predicting τ such that $\mu(\epsilon) = \pm 282$ Myr.

3.5.2 Abundances

The small frequency separations and separation ratios are integral for the determination of ages. However, the feature importances in BA1 (their Figure 5) indicate that the RF relies predominately on $T_{\rm eff}$ and [Fe/H] to infer other model parameters. Table 3.5 confirms how important measuring [Fe/H] is for characterizing stars. This quantity is preferentially selected in the many RFs and their regression models, whilst [Fe/H] itself cannot be determined from the other observables with any degree of confidence. [Fe/H] is an indispensable piece of independent information.

Accurate determination of [Fe/H] is paramount for inferring many of the current-age stellar attributes. [Fe/H] also features prominently in the retrodiction of the initial model parameters but these quantities are characterized by

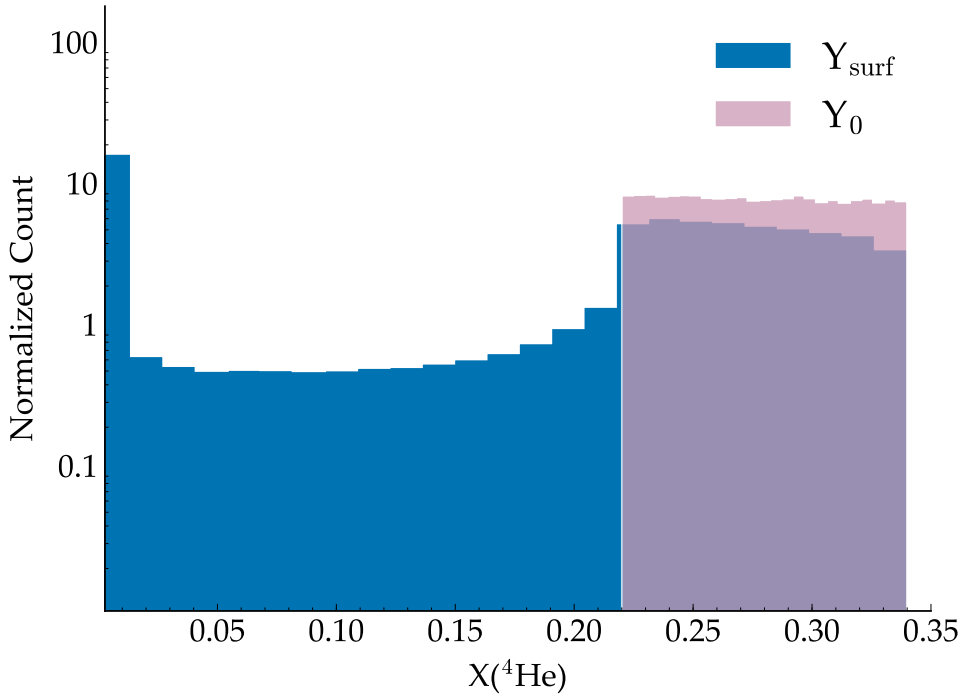


FIGURE 3.6. Distributions of $Y_{\rm surf}$ and Y_0 in the BA1 grid.

large uncertainties. Foremost, we have no observable that satisfactorily constrains diffusion; D demonstrates an average uncertainty spanning three orders of magnitude. This in turn introduces uncertainty in retrodicting the initial metal content.

Predictions for Z_0 at first glace appear to be robust; we report V_e and $\mu(\epsilon) = \pm 0.001$. However we contend that a reported error of $\mu(\epsilon) = \pm 0.001$ is not all that insightful given that the grid is sampled down to $Z_0 = 10^{-5}$. Z_0 is sampled logarithmically and takes a small (linear) range in values. In such cases a relative error is a more useful measure of performance than an absolute difference.

In Table 3.6 we devise a series of measures that better appraise the performance of the RF in predicting abundances. We report the average absolute difference as per Table 3.5 [$\mu(\epsilon)$], the maximum absolute difference [$\max(\epsilon)$] and the median absolute difference [$\tilde{\epsilon}$]. We also consider the average relative error [$\mu(\eta)$], the maximum relative error [$\max(\eta)$] and median relative error [$\tilde{\eta}$], where the relative error is a percentage defined as

$$\eta = \frac{|\hat{y}_i - y_i|}{|y_i|} \cdot 100. \tag{3.9}$$

We find $\mu(\eta) = 125\%$ in the retrodiction of metallicity. We attribute the seemingly large uncertainty to the bias imparted by extreme models that have undergone significant diffusion – we report a maximum relative error of 9000%. With less sensitivity to the outlying metal-depleted models, the median relative uncertainty, $\tilde{\eta} = 13.5\%$, offers the most appropriate measure of error in the regression. Likewise, the extreme $\mu(\eta)$ and $\max(\eta)$ scores for Y_{surf} also stem from models with high diffusion leading to very small non-zero abundances by which we normalize.

It is interesting to compare the regressor's ability to infer Y_{surf} and Y_0 abundances. We find that Y_{surf} can be well fit ($V_e = 0.927$) with $\mu(\epsilon) = \pm 0.022$. In contrast, the initial abundance, Y_0, cannot be confidently retrodicted ($V_e = 0.625$) yet results in a smaller average error [$\mu(\epsilon) = \pm 0.017$]. This initially surprising result can be understood through examination of the respective parameter distributions in the BA1 grid (Figure 3.6). The grid is uniformly sampled in initial

TABLE 3.6. Different measures of uncertainty in predicting stellar abundances with the RF. See text for definitions and motivations.

Error Measure	Y_{surf}	Y_0	Z_0
$\mu(\epsilon)$	0.02	0.017	0.001
$\text{Max}(\epsilon)$	0.25	0.09	0.037
$\tilde{\epsilon}$	0.016	0.02	0.00019
$\mu(\eta)$ [%]	10^{13}	8.92	124.5
$\text{Max}(\eta)$ [%]	10^{14}	40.34	9052
$\tilde{\eta}$ [%]	10.88	7.68	13.5

helium with $Y_0 \in [0.22, 0.34]$. Atomic diffusion acts to drain helium from the surface layers and in fact, in some models, completely depletes this species from the envelope. The surface helium abundance of a stellar model can thus attain values in the larger range $Y_{surf} \in [0.0, 0.34]$. In a uniform distribution, such as we have for Y_0, the largest theoretical uncertainty is

$$\max\left(\frac{\sigma^2(Y_0)}{Y_0}\right) = \frac{|b-a|}{|a|} \cdot 100 = 54.51\%, \tag{3.10}$$

where a and b are the respective minimum and maximum values in our parameter range. This means that if the regressor was unable to explain *any* of the variance in this quantity and was randomly choosing Y_0 values from the initial distribution, the worst relative uncertainty we would expect is 54.51%. The fact that we do go someway to predicting this quantity results in $\mu(\eta) \approx 8\%$ and more accurate inferences than for Y_{surf}.

3.5.3 Other Results

We mention briefly other interesting results from the approximately $50,000$ RFs not necessarily reported in Table 3.5. Stellar masses can be accurately inferred from spectroscopic measurements. The combination of $\log g$, T_{eff} and [Fe/H] constrains mass equally well as the pair $\langle \Delta\nu_0 \rangle$ – $\log g$. Both combinations explain 86% of the variance in mass with $\mu(\epsilon) = \pm 0.07\ M_\odot$. With six degrees of freedom in the BA1 grid, we cannot determine mass to an accuracy better than $\mu(\epsilon) = \pm 0.02\ M_\odot$. Whilst all observables correlate with M, they do not contain sufficient information to separate out the redundant structures that are possible by tweaking the other initial model parameters. We in fact find no improvement in our regression for M beyond three parameters[11].

If required, the RF can determine T_{eff} with high accuracy. Although this is almost certainly always an input for the RF, with two or more observables T_{eff} can be determined with $\mu(\epsilon) \approx 100$ K – an uncertainty comparable to typical spectroscopic errors. If one of L or R are provided as an input to the RF, a factor of two reduction in the uncertainty is achieved with $\mu(\epsilon) \lesssim 50$ K. Furthermore, our testing of the RF (not included here) indicates that if both L and R are provided as observables the Stefan-Boltzmann law is recovered with $\mu(\epsilon) = 4$ K.

3.5.4 Seismic Quantities

We did not include the predictions for the seismic parameters in Table 3.5 as they often carry redundant information. Indeed we accomplish little by reporting how the different combinations of ratios and separations can be used to recover each other. We thus opt to analyze the seismic parameters separately, where we can employ discretion to present useful comparisons and highlight noteworthy results.

[11] Numerics accounts for the differences in the third decimal place for scores in Table 3.5.

The large frequency separation – $\langle \Delta\nu_0 \rangle$

In lieu of a direct measurement, $\langle \Delta\nu_0 \rangle$ can be estimated from stellar models via an asteroseismic scaling relation (Equation 3.20). Alternatively, it may be inferred from the observables through an empirical power law that relates $\langle \Delta\nu_0 \rangle$ to ν_{max} (Hekker et al. 2009, Stello et al. 2009a). The power law estimates $\langle \Delta\nu_0 \rangle$ within 15% of its measured value (Stello et al. 2009a). We compare the RF's ability to likewise predict $\langle \Delta\nu_0 \rangle$ from ν_{max} in Table 3.7. We also consider two and three parameter combinations for inferring $\langle \Delta\nu_0 \rangle$ with the requirement that they do not comprise the remaining seismic observables.

We find that the RF predicts $\langle \Delta\nu_0 \rangle$ from ν_{max} with $\mu(\eta) \approx 6\%$. These results are based on error free information (cross-validation hence no measurement noise) and the inclusion of ν_{max} from a scaling law. In order to conduct a more faithful comparison with Stello et al. (2009a), we analyze the same data used in the derivation of their power law. Their Table 1 is a compilation of ν_{max} and $\langle \Delta\nu_0 \rangle$ values from the literature. The data are predominately from radial velocity studies and measured with less precision than we have come to expect from *Kepler* timeseries; they provide a robust test of the RF. We feed the RF the quoted ν_{max} measurements and predict associated $\langle \Delta\nu_0 \rangle$ values. We compare our predictions to the $\langle \Delta\nu_0 \rangle$ values from the literature which are used to calculate corresponding ϵ and η scores. Our results are presented in Table 3.8. We omit entries from the Stello et al. (2009a) dataset that are outside the parameter ranges of our training grid. For the remaining 17 stars we find $\mu(\eta) \approx 8\%$ which is comparable to $\mu(\eta) \approx 6\%$ accuracy achieved from cross-validation test (approximately 15,000 stars).

The last column in Table 3.8 indicates that the accuracy from the RF is similar to that of the power law. In addition, we find that parameterizing the RF regression as a function of two observables reduces the uncertainty by a factor of 2–3 (Table 3.7). This hints that the inclusion of a temperature or metallicity dependence may also improve the fit offered by the power law[12].

TABLE 3.7. Combinations of observables that best constrain $\langle \Delta\nu_0 \rangle$.

Parameters			V_e	$\mu(\epsilon)$ [μHz]	$\mu(\eta)$ [%]
ν_{max}			0.930	7.815	6.11
T_{eff}	ν_{max}		0.990	3.09	2.46
$\log g$	ν_{max}		0.990	2.95	2.34
$\log g$	T_{eff}		0.990	2.92	2.31
[Fe/H]	ν_{max}		0.991	2.81	2.24
T_{eff}	[Fe/H]	ν_{max}	0.995	1.67	2.13
$\log g$	[Fe/H]	ν_{max}	0.995	1.65	2.11

[12] Symbolic regression will help determine whether, in this case, the fitting by the RF has a sensible functional form that can be straightforwardly expressed by two independent variables. This

TABLE 3.8. Predictions of $\langle \Delta\nu_0 \rangle$ for stars listed in Stello et al. (2009a). Results pertain to a random forest trained with ν_{max} as the only input. Predictions are compared to literature values from the sources listed in Table 1 of Stello et al. (2009a). The RF performs as well as the power-law relation (10-15%) even on data measured with less precision than stars observed by *Kepler*.

Star	ν_{max} (μHz)	$\langle \Delta\nu_0 \rangle_{lit}$ (μHz)	$\langle \Delta\nu_0 \rangle_{pred}$ (μHz)	ϵ (μHz)	η (%)
τ Cet	4500	170	171	1	1
α Cen B	4100	161	184	22	14
Sun	3100	135	138	3	2
ι Hor	2700	120	136	16	14
γ Pav	2600	120	122	1	1
α Cen A	2400	106	124	18	17
HD 175726	2000	97	100	3	3
μ Ara	2000	90	100	10	11
HD 181906	1900	88	97	10	11
HD 49933	1760	86	101	15	18
HD 181420	1500	75	76	1	1
β Vir	1400	72	77	5	8
μ Her	1200	57	63	7	12
β Hyi	1000	57	57	0	0
Procyon	1000	55	57	2	4
η Boo	750	40	45	5	13
ν Ind	320	25	23	3	10

Analysis of recent *Kepler* data yields a similar result. In Figure 3.7 we present the percentage error in our predictions of 467 stars measured by *Kepler* as reported in Table 1 of Chaplin et al. (2014). We analyze stars for which ν_{max}, $\langle \Delta\nu_0 \rangle$ have been measured from the oscillation spectra along with T_{eff} as determined by Pinsonneault et al. (2012) based on Sloan Digital Sky Survey (SDSS) photometry. Results from the *Kepler* sample confirm that predictions for $\langle \Delta\nu_0 \rangle$ are improved with the inclusion of T_{eff} (lavender distribution). The blue distribution indicates that $\langle \Delta\nu_0 \rangle$ is systematically overestimated when the RF only has access to information from ν_{max} – a bias that may very well be present in the power-law fit. With the inclusion of T_{eff} our predictions become more accurate and precise with the bias from the single parameter function mitigated. We do not quite reproduce the accuracy achieved in the cross validation (Table 3.7) using error free information. Unsurprisingly, measurement uncertainty, which we do not consider here, does not permit the accuracy attained in the ideal case.

result seems reasonable as the additional information is likely providing a better handle on the stellar mass.

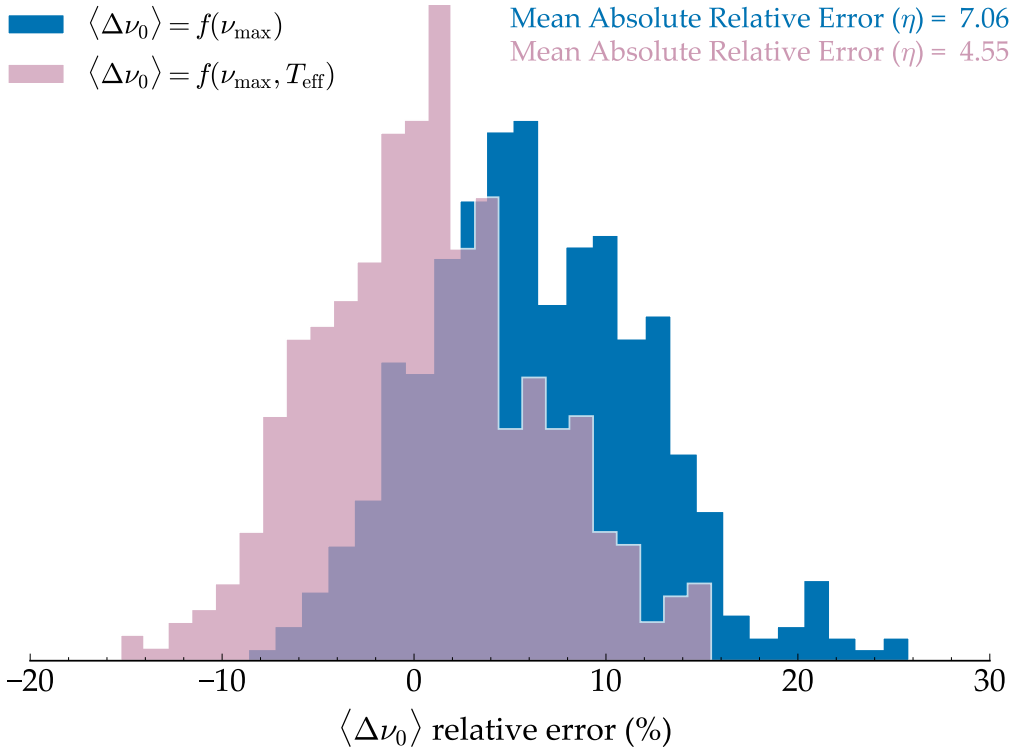


FIGURE 3.7. Relative error (%) in the predictions for $\langle \Delta\nu_0 \rangle$ for 467 stars reported in Chaplin et al. (2014). The blue colored distribution indicates the error in the predictions from the random forest using $\nu_{\max}$ as the only input observation whilst the distribution marked in lavender are the results from providing $\nu_{\max}$ and $T_{\rm eff}$. In the calculations we employ the effective temperatures determined from Pinsonneault et al. (2012) based on SDSS photometry.

The frequency of maximum oscillation power – $\nu_{\max}$

Currently we are unable to predict the frequency of maximum oscillation power from first principles. Brown et al. (1991) and Kjeldsen and Bedding (1995) showed that this quantity does scale with the acoustic cut-off frequency and can thus be estimated via the Equation (3.19) scaling relation. It is therefore expected that Table 3.9 indicates that $\nu_{\max}$ is best inferred from $\log g$ and $T_{\rm eff}$. These are the two observables that correlate strongest those parameters used to calculate $\nu_{\max}$ in the training grid.

The small frequency separation – $\langle \delta\nu_{02} \rangle$

The small frequency separation is an indispensable piece of independent information for determining stellar age. In the asymptotic limit (Tassoul 1980)

$$\langle \delta\nu_{13} \rangle = \frac{5}{3} \langle \delta\nu_{02} \rangle \tag{3.11}$$

and as Table 3.10 demonstrates, the RF recovers $\langle \delta\nu_{02} \rangle$ in the unlikely case that it is not extracted but $\langle \delta\nu_{13} \rangle$ is. If we disregard combinations that include the

TABLE 3.9. Combinations of observables that best constrain ν_{max}.

Parameters		V_e	$\mu(\epsilon)$ [μHz]
$\langle\Delta\nu_0\rangle$		0.923	7.88
$\log g$	[Fe/H]	0.888	9.99
$\log g$	$\langle\Delta\nu_0\rangle$	0.954	5.38
T_{eff}	$\langle r_{10}\rangle$	0.960	5.11
[Fe/H]	$\langle\Delta\nu_0\rangle$	0.992	2.90
T_{eff}	$\langle\Delta\nu_0\rangle$	0.992	2.84
$\log g$	T_{eff}	0.999	0.83

seismic ratios, which also contain information of the local small frequency separation, we lack sufficient information to satisfactorily constrain $\langle\delta\nu_{02}\rangle$. Clearly much of the evolutionary aspect of this quantity can be explained though parameters that correlate with main-sequence lifetime e.g., $\log g$, $\langle\Delta\nu_0\rangle$, ν_{max} and T_{eff}. However the associated errors of $\mu(\epsilon) > 1.0$ μHz can correspond to large age uncertainties for main sequence stars ($\eta > 10\%$).

3.6 Quantifying the Required Measurement Accuracy of Stellar Observables

In the previous section we used RF regression to appraise how well combinations of observables constrain other stellar parameters. The $\approx 50,000$ RFs were evaluated using cross-validation. The tests are a pure measure of the regressor's performance as we have error-free information that we attempt to reproduce (withheld models). As we have already alluded to, like all procedures that seek

TABLE 3.10. Combinations of observables, without the asteroseismic ratios, that best constrain $\langle\delta\nu_{02}\rangle$.

Parameters			V_e	$\mu(\epsilon)$ [μHz]
$\langle\delta\nu_{13}\rangle$			0.944	0.66
$\log g$			0.542	2.08
$\langle\delta\nu_{13}\rangle$	$\langle r_{10}\rangle$		0.987	0.320
T_{eff}	ν_{max}		0.776	1.40
$\log g$	T_{eff}		0.775	1.40
$\log g$	ν_{max}		0.772	1.41
$\log g$	$\langle\Delta\nu_0\rangle$		0.723	1.54
T_{eff}	$\langle\Delta\nu_0\rangle$		0.720	1.58
$\log g$	[Fe/H]		0.720	1.59
$\log g$	[Fe/H]	$\langle\Delta\nu_0\rangle$	0.861	1.06
$\log g$	ν_{max}	$\langle\Delta\nu_0\rangle$	0.860	1.09

to infer stellar parameters, we must also consider the consequences of measurement uncertainty in our method.

Measurement uncertainty will impact the RF results in a manner that is different to model finding algorithms. Consider an iterative model finding procedure in which we seek an optimum model for a set of observations. We can typically expect T_{eff} as a constraint with an associated uncertainty of $\sigma = 100$ K. The minimization algorithm will identify a set of candidate models, many with quite different structures. Hence the uncertainty in T_{eff} will impact all stellar quantities simultaneously. The RF, on the other hand, builds a statistical description of stellar evolution by calculating a regression model for each individual parameter from the training data. The BA1 method requires that each input observable is perturbed with random Gaussian noise according to its measurement uncertainty. Monte Carlo perturbations are performed $10,000$ times and each instantiation evaluated by the RF to yield individual density distributions for each stellar parameter. Thus the uncertainty in T_{eff}, or any observable for that matter, will only impact on the predictions of each parameter in proportion to the degree to which it features in that parameter's regression model.

The methodology, combined with the speed of the RF, provides a tractable means to asses how the individual measurement uncertainty of an observable will impact upon each predicted stellar quantity. We hence determine how accurately the observables must be measured in order to achieve a desired precision from the RF.

We train a RF on the observables listed in Table 3.11. We take the (approximate) solar value of each observable as our measurement and consider 'observational uncertainties' (σ) within the ranges specified in Table 3.11. We first perturb the measurement values with Gaussian noise assuming the minimum σ values listed. We produce $10,000$ instantiations for that set of σ values, ensuring each perturbed observable remains within the limits of our training grid. We evaluate stellar parameters and determine detailed distributions for that set of uncertainties. We repeat the process increasing the σ for a single observable always keeping the σ values of the other observables at their minimum. We draw 50 σ values for each observable sampling their specified ranges evenly. We produce probability density distributions for 250 sets of σ values, the results of which are summarized in Figure 3.8.

TABLE 3.11. Central values and uncertainty ranges used for predicting the Sun in Figure 3.8.

Quantity	Value	Min(σ)	Max(σ)
T_{eff} (K)	5777	10	500
log g	4.43812	0.00013	1.0
[Fe/H]	0.0	0.05	0.2
$\langle \Delta\nu_0 \rangle$ (μHz)	136.0	0.5	10
$\langle \delta\nu_{02} \rangle$ (μHz)	9.0	0.5	5

In Figure 3.8 we plot the median value (solid line) and the 68% confidence interval (shaded region) for M, τ, L and R as a function of the uncertainty applied to each observable. The figure is organised such that each row (and color) corresponds to the observable that has had its uncertainty increased and each column corresponds to the model parameter of interest. In this Figure, the left axis indicates the predicted value from the RF and the right axis indicates the relative error with reference to the true values of the Sun. The horizontal dotted grey lines mark the reference value in each case whilst the dotted vertical lines indicate a typical uncertainty for the perturbed observable.

The particular RF we have trained does not significantly rely on $T_{\rm eff}$ in its regression model for M, τ or R. As the radius is supplemented by the seismic quantities, any uncertainty in $T_{\rm eff}$ is propagated as uncertainty in the luminosity. We find a typical uncertainty of 100 K corresponds to an error of ± 0.2 $L/L_{\odot}$ at the 68% confidence level.

The inference on solar mass is affected once $\delta \log g > 0.03$. However, even at unreasonably large values of $\delta \log g = 1$, the uncertainties for mass and age remained relatively constrained by additional seismic information. We find that L and R are far more reliant on $\log g$ in their regression function with uncertainties in these quantities growing significantly once $\delta \log g > 0.1$.

The feature importances in BA1 indicate that [Fe/H] is used most often by the RF in crafting its decision rules. The four stellar parameters we investigate here indeed all rely on information from [Fe/H], however, they are supplemented by seismic information which helps to constrain the uncertainty in their predictions. It is the model parameters such as the mixing length, degree of overshoot and initial metallicty that become much less certain as we increase $\sigma([\mathrm{Fe/H}])$ (not shown here).

The seismic diagnostics are very sensitive to the stellar structure, and hence also those parameters we use to characterize a star (M, τ, L and R). We have seen how reliant the RF is on the seismic diagnostics in the regression models, allowing us to still predict the structural properties with relatively good precision in the face of large spectroscopic uncertainties. Without accurate measurement of $\langle \Delta\nu_0 \rangle$ the uncertainty in structure parameters increase significantly. Whilst the uncertainty in $\langle \delta\nu_{02} \rangle$ does introduce some small uncertainty in M, L and R, as expected, its accuracy significantly impacts upon our ability constrain stellar age.

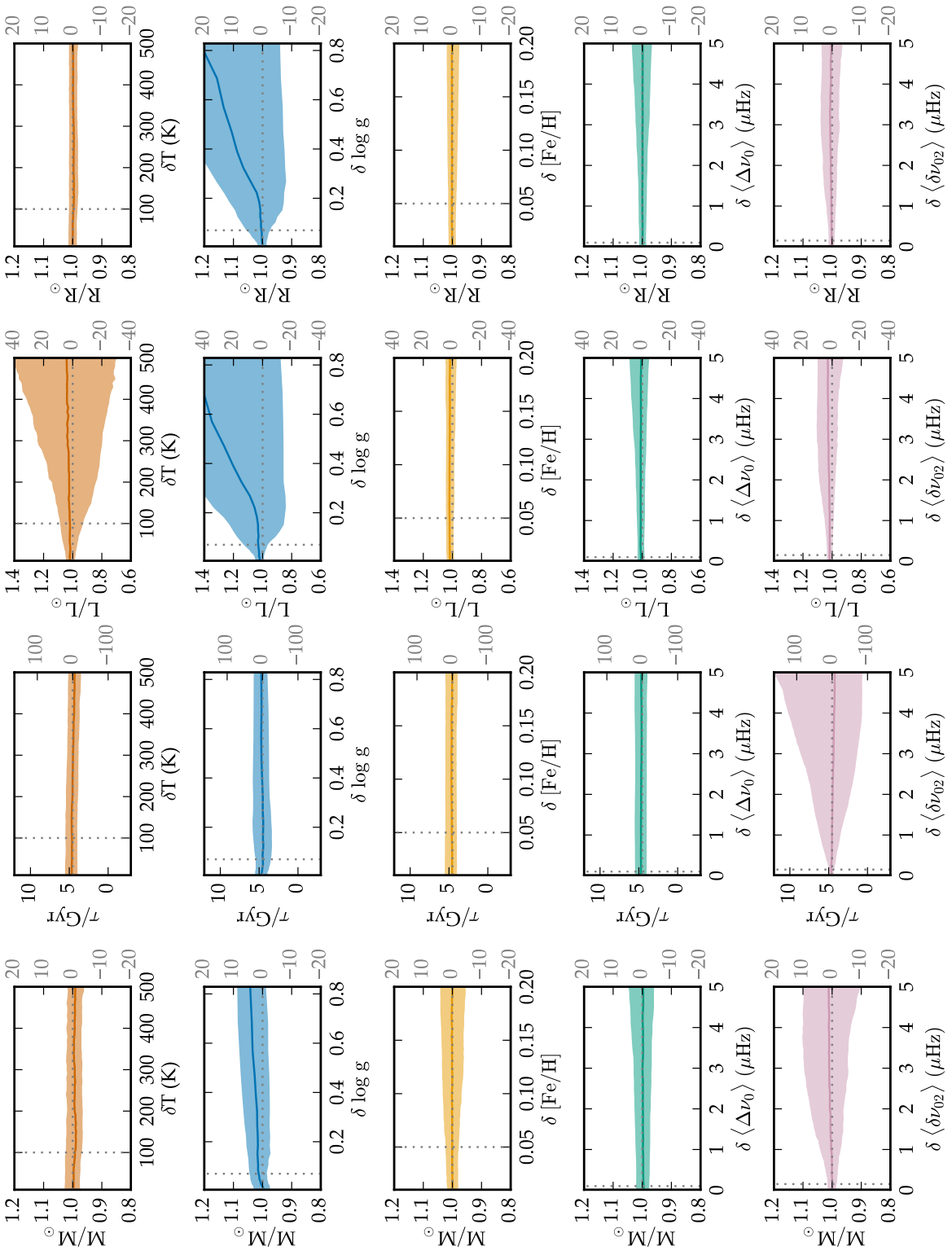


FIGURE 3.8. (Caption on other page.)

3.7 Discussion

Advances in stellar evolution theory are usually sought through refinement of the standard canonical model. In this classical approach, observations reveal behaviour that cannot be explained by the current stellar theory, a model is constructed, analysis of the resultant predictions are carried out and conclusions on the efficacy of that model drawn. In this study we adopted a complementary approach: an exploratory based method whereby we performed statistical analysis of models covering a large range of known physics. Rather than first develop a new model to evaluate, we explored the current paradigm to quantify existing relationships and draw new conclusions.

Some of the techniques employed in this analysis are over 100 years old and in many areas of research are powerful standalone tools. They have rarely featured in the field of stellar modelling. Here we comment briefly on the timing of our manuscript which we attribute to two main factors: the advent of supervised machine learning techniques and modern computing resources.

Random forests are an integral part of the present analysis and are a modern technology. They help place the use of statistical methods in stellar evolution in a wider practical context. Elucidating both the relationships found by RF and the exploitable information inherent in the model data provided motivation for the use of techniques such as PCA and correlation analysis. The RF further facilitated the application of these methods due to the requirement that the models be cast into a comprehensive evolutionary matrix; something that is not strictly necessary for grid based searches.

Our approach shares similarities to that taken by Brown et al. (1994) although we differ in methodology. Since their work, we have seen the necessary increase in computing power and the success of the *Kepler* and CoRoT space missions. The statistical analysis here requires a well sampled grid of stellar models both with structure and oscillations computed. It cost a week of modern supercomputing time to generate the matrix upon which these operations are performed. Evaluating and training approximately $50,000$ RFs itself is also a computationally expensive endeavour.

FIGURE 3.8. Predictions for the solar mass, age, luminosity and radius as a function of the uncertainties applied to key observables. In each panel we have perturbed the quantity on the abscissa in isolation, centred around the measured value listed in Table 3.11 and with the uncertainties in the ranges specified therein. We indicate the median predicted value (solid line) and the 68% confidence interval (shaded region). The dotted horizontal lines mark the zero point or true value in each panel and the vertical line indicates a typical observational uncertainty for the perturbed quantity.

3.7.1 Features of the Dataset

It is not clear *a priori* through inspection of the equations of stellar structure, if and how any two emergent quantities of the models co-vary. There are, of course, combinations of parameters whose covariances are well-founded in stellar theory, but there exist quantities whose diagnostic power remain under-utilized and could in fact offer additional insight into the underlying models. Bringing such relationships to light over the collective lower main sequence is a key aim of our statistical investigation. The correlations in the truncated grid (Figure 3.2) and full BA1 grid (3.10) reveal the relationships that can be utilized to constrain each of the quantities listed in Table 3.1. Many of the model properties that we wish to infer correlate with several observables simultaneously. This indicates that the observables carry redundant information about the star. In addition, observables co-vary amongst themselves. During iterative model searches some of the covariances, such as between the seismic ratios, are taken into account. However, for example, it is possible to obtain independent measurements of $\nu_{\rm max}$, $\langle\Delta\nu_0\rangle$, and $\log g$. Treating these as independent degrees of freedom without considering model covariances then biases the fit towards the parameters to which these quantities pertain and can result in a solution that is overfit.

We determined the degree of degeneracy in the observables through PCA dimensionality reduction. As mentioned previously, RF regression falls under the umbrella of *supervised* learning, whereas PCA is a form of *unsupervised* learning. The difference is that in supervised learning, there is a correct answer that the algorithm is trying to understand how to reproduce. In the case of unsupervised learning, the machine attempts to directly infer properties of data without any help from the supervisor. Hence, regression and classification analyses are forms of supervised learning, whereas cluster and factor analyses are examples of unsupervised learning. In the case of supervised learning there is a clear measure of success in the resultant model. There is a desired output that the inputs try to match. The efficacy can be quantified and evaluated via, say, cross-validation or information-theoretic metrics. Unsupervised learning methods simply try to identify features and in the case of PCA these features are not necessarily interpretable.

The PCA in Section 3.4 focused on the truncated grid. It comprises 11 stellar observables of all which carry information on the model properties to varying degrees. We found that 99.2% of the variance in the observables could be explained by five components with nearly 98% of the data are explained by four components. It could be argued that PC_5 explains noise rather than features, however, we found that PC_5 displays distinct enough correlations (i.e., with near surface physics) that it warrants inclusion in our analysis. The clear dimensionality reduction, from 11 observables to five PCs, highlights the value in performing PCA: had we found comparable contributions from each component, we would have instead confirmed a clear dominance from higher order relations and an inadequacy of an approach based on linear analysis.

Our primary goal in Section 3.4 was to reduce the dimensionality of the observables. We initially considered regions of the parameter space where observations have shown stars to occupy. Following on from the rank correlation tests in Section 3.3 we applied PCA to a truncated version of the BA1 grid. However, the results of the PCA depend on the properties of the data and will change depending on features such as the parameter ranges and number of models in the grid. For example performing PCA on the full set of evolutionary tracks (340,800 models) demands that components are dedicated to explaining variance in (wider) unobserved regions of the parameter space. In order to demonstrate that our interpretations of the PCs are robust, we repeated the PCA on four different subsets of the BA1 grid. We made cuts to the mass and metallicity ranges on the training data the results of which are included in Appendix 3.9.6 by means of qualitative correlation plots.

The PCs of the respective grids explain a similar percentage of the variance in each grid: PC_1 accounts for approximately 40% of the variance, PC_2 approximately 35% etc., with more than 75% of the variance in the observables explained by the first two PCs. We interpret this result as the PCA capturing essentially the same five inherent 'features' in the observables. It follows that the choices in grid size and parameter range have only a small effect on the explained variances. Analysis of all four grids helps further illustrate that there is redundant information carried in some observables, particularly the seismic separations and ratios. Varying the parameter ranges changes the correlations between the PCs and observables (loadings) yet the PCs still explain a similar percentage of the variance in each case. Due to the information redundancies the PCs can be constructed such that same features are captured with different linear combinations of the observables. How exactly a PC is constructed in a particular grid will depend on the amount of variance in the observables imparted by the chosen parameter ranges.

With respect to the independent model parameters, it is no surprise that in general PC_1 is strongly correlated with the stellar mass (M) and and PC_2 with initial metallicity (Z_0). These are the principal determinants of stellar evolution in that order and both impact upon the stellar structure independently. In the two grids where we have cut the mass and metallicity ranges we find that the loading of $T_{\rm eff}$ is larger in PC_1. This is because in the more solar-like tracks $T_{\rm eff}$ is a strongly monotonic function of evolution. The surface aspect of PC_2 is then supplemented with some information from $\log g$ and [Fe/H].

Reducing the dimensionality of the observables and relating them back to the model parameters without redundancy aided with the interpretation of the PCs. Whilst it is useful to have the observables so succinctly described, it does not provide insight into the model parameters we wish to infer. We thus condensed the information from the correlation plots into a Λ score which is the sum of the square of the correlation coefficients between the model parameters and the PCs (determined for the observables). Squaring the correlation coefficients is equivalent to the squaring the PC loadings of the centered and scaled

observables. The score is a means to quantify the extent to which information from the model parameters, dependent and independent, are encoded in the observables. We calculated Λ scores for all four grids upon which PCA was performed (Appendix 3.9.7) and indeed found mostly consistent results. We note some differences arise in the initial model parameters such as α_{MLT} and α_{OV} which reflect their underlying distributions from the choices in grid truncation. The above analyses can be applied to any combination of observables and model parameters to gauge their utility.

3.7.2 Exploiting the Inherent Relationships

Understanding the inherent properties of the collective lower main sequence is the first step in elucidating the BA1 RF regression. The statistical analysis quantified what information was present in the training data for the RF to exploit. We illustrated why the available data permit BA1 to predict parameters such as M, R and L with such high precision and why initial model parameters such as D and α_{MLT} remain uncertain in comparison. Whilst Section 3.3 and Section 3.4 demonstrated the breadth of information available to the RF, in Section 3.5 we determined how the information could best be used.

RFs are amongst the most powerful tools in mathematics for non-linear regression. The BA1 RF uses the observables, creating a set of decision rules that reduce the variance in the parameter it is fitting. Whilst feature importances provide some insight into this process as a whole it does not provide specific details for the individual parameters. By performing non-parametric multiple regression with every combination of observable in our grid, we demonstrated how the correlations in Figure 3.2 could best be exploited and best combined to reveal the most information about each stellar quantity. Two of the observables, [Fe/H] and $\langle \delta\nu_{02} \rangle$ (or as a ratio), are of vital importance in model fitting procedures as they provide indispensable pieces of independent information that cannot be inferred from other quantities.

We in effect invert the observations for the model parameters based on functions learnt from the training data. Thus we can determine the relative importance of each observable for inferring the model parameters. We, in addition, provide a precision with which we can determine each model parameter *directly* from the information contained in the observables. The attainable precision is a function of the number of initial model parameters that are varied and the model degeneracy in the data. For example, with perfect information from the observables, the six dimensions in the BA1 grid limits our inference on mass to $\mu(\epsilon) = 0.02\ M_{\odot}$.

Many of the Tables in Section 3.5 demonstrated an important property of the RF. In the case of missing or unreliable measurements of an observable, the RF can draw upon information redundancies in the data to determine new regression rules for the model parameters. In principle, such redundancies can lead to biases and overfitting in iterative model finding methods. During such search procedures the best fitting stellar model is the one that best matches all

of the observations but each observation only bares on some parts of the model, and observations can contain redundant information.

Through statistical bagging and multiple regression the RF is less likely to overfit. These underlying methodologies are the reason why in Section 3.6 many of the parameters we inferred remained well constrained despite large uncertainties in some of the observables. In statistical bagging different subsets of the training grid are sent to different nodes. Each node will use information theory to create a set of decision trees to explain the parameter of interest. The nodes will differ in their rules and choice of parameters. Thus the uncertainty in an observable will only impact on the parameter we infer to the extent to which the observable is used in the rules. Take the example from Figure 3.8 where with a 5 μHz uncertainty in $\langle \Delta\nu_0 \rangle$ the RF still predicts the solar properties albeit with slightly less confidence. The other observables help constrain the predictions.

Part of the analysis in Section 3.5 demonstrated the best possible (average) precision in which we can hope to infer stellar parameters. Our error analysis in Section 3.6 is an extension of this. Rather than assume perfect information we determined the measurement accuracy required of the observables to attain a desired precision from the RF. Our analysis focused on the Sun and is indicative of solar-like analogues. In Table 3.6 we saw some of the large uncertainties associated with retrodicting abundances in low-metallicity stars. We have greater degeneracy with the efficiency of diffusion and the initial abundances. These large error scores by no means indicate that the RF is incapable of characterizing low-metallicity stars. Rather it is an honest appraisal of stellar uncertainties when we do not make assumptions of the initial abundance say through a dY/dZ chemical evolution "law" or a fixed diffusion efficiency. Our error analysis here does not take into account covariances and was designed to investigate the impact on an observable-by-observable basis. A more detailed error analysis and the associated issues at low metallicity form the focus of a forthcoming paper.

3.7.3 Implications for the TESS and PLATO missions

The NASA TESS mission (Ricker et al. 2015) and ESA's PLATO (Rauer et al. 2014) herald a new age for the space-based photometry and the detection of planetary transits. Due to launch in 2018 and 2025 respectively, their common primary science mission is to identify terrestrial planets around bright stars. The pre-selection of bright targets will ensure that the stellar hosts can be further analyzed with spectroscopy and it is expected that many of the planet candidates will be suitable for atmospheric follow-up (ideally) with the James Webb Space Telescope. As was the case with the *Kepler* and *CoRoT* missions, the photometric time-series observations will prove useful to asteroseismology. In the case of PLATO the study of the stellar structure through asteroseismology is a key science goal in the mission design (Rauer et al. 2014).

TESS will monitor photometric variations of $> 10^5$ low-mass main-sequence stars. Under its 'step and stare' pointing strategy, fields will be monitored for

periods ranging from one month to one year depending primarily on their ecliptic latitude. With its two minute and 30 minute cadences, TESS will be able to detect small rocky planets around solar like stars at $\leqslant$ 7th magnitude. It is expected to detect of the order 1,700 planets with sub-Neptune masses (Campante et al. 2016) and will identify many more larger planets around dimmer targets. The asteroseismic potential of TESS has been rigorously investigated by Campante et al. (2016). Their analysis of the expected TESS photometry indicates the presence of an oscillation power excess in low-mass main-sequence stars when there is no systematic noise present in the data. With an expected systematic noise level of 60 ppm $hr^{1/2}$ from the mission, their analysis indicates a detectable power-excess in F-dwarfs as well as sub giants and red giants – this owing to the higher luminosity and hence larger mode amplitudes in these stars. For a majority of stars the 27 day pointing is insufficient to extract detailed asteroseismic diagnostics such as mode frequencies or separations. Rather, the seismic information will be limited to the determination of ν_{max} in stars where the power-excess is detected. As a consequence, masses and radii for the TESS targets are to be determined using a combination of GAIA data, the $\nu_{max} - \langle\Delta\nu_0\rangle$ power law (Hekker et al. 2009, Stello et al. 2009a) , asteroseismic scaling relations and grid-based searches.

The number of small planet detections from the PLATO mission is expected to eclipse the number found by *Kepler* and TESS by up to three orders of magnitude. In addition, the PLATO pointing strategy will allow for the measurement of oscillation frequencies in $>$ 80,000 dwarf and subgiant stars with magnitudes less than 11. In total the mission will provide approximately one million light curves for stars with brightness $\leqslant$ 13th magnitude (Rauer et al. 2014). In many stars modes up to spherical degree $\ell = 3$ will be detected with typical frequency uncertainties in the range 0.1 – 0.3 μHz. The second major science goal of PLATO is to probe stellar structure and evolution by asteroseismology and provide support to exoplanet science through determining

- stellar masses with an accuracy of better than 10%,
- stellar radii accurate to 1–2%, and
- ages of solar-like stars accurate to 10%.

Here we treat the 'Sun as a star' in order to quantify how well we can characterize target systems observed by the upcoming space missions and to determine the prospect of meeting the accuracy requirements. In Table 3.12 we indicate the observables the missions are likely to provide. We degrade the corresponding solar data according to the expected uncertainty from the respective measurements. As GAIA is complete down to 20th magnitude we have assumed that distances and hence luminosities will be available for all targets in these missions. We consider data for TESS targets assuming both 60 ppm $hr^{1/2}$ and no systematic noise in the photometry. Thus in the case of the latter we anticipate that an oscillation power excess can be extracted for a solar-like star and ν_{max}

determined. The large and small frequency separations for the PLATO data are determined by degrading a subset of solar frequencies using the method described in BA1. We take a conservative approach in this calculation and assume that the $\ell = 3$ modes are not extracted.

Figure 3.9 shows our predictions for masses, radii, ages, initial helium and metallicity for a 'Sun-as-a-star' exercise. In each panel we indicate the median of the probability density distribution and the corresponding uncertainty from the 16% and 84% confidence intervals for the parameter we are predicting. In addition we determine the relative error which we define as $\epsilon = 100 \cdot \sigma/\mu$ where μ is the mean and σ is the standard deviation of the distributions. In Appendix 3.9.8 we further demonstrate the impact of the measurement uncertainty on the prediction of each quantity as per Figure 3.8.

Although we can expect accurate mass determinations for targets in both missions, the supplementary seismic data from PLATO allows us to improve the precision with which we determine mass by approximately a factor of two. This is despite the fact the RF has identified a less-likely but not impossible (slightly) younger, higher-mass, higher-metallicity solution from the PLATO data (we find bimodalities for most quantities predicted with the PLATO observables). In the case of TESS, the absence of the large frequency separation leads to greater uncertainty. One of the methods discussed by Campante et al. (2016) for the mass determination of TESS targets is to use the power law linking ν_{max} to $\langle \Delta\nu_0 \rangle$ (which has been shown to be accurate to 10–15%) and apply the asteroseismic scaling laws (Equations 3.19 and 3.20). In Section 3.5.4 we demonstrated that the random forest exploits further information from temperature or metallicity measurements to improve the accuracy of the ν_{max} – $\langle \Delta\nu_0 \rangle$ relation. Thus we expect the accuracy with which we predict mass from TESS data to represent an upper limit to that attainable by applying the power-law and scaling relations.

The assumption of GAIA distances and hence stellar luminosities ensure that radii can be determined for targets in both missions; the seismology is essentially redundant for the inference of the stellar radius. We note that the relative error for PLATO in our 'Sun-as-a-star' test is a factor of two higher than the 1–2% expected by the consortium. This is a consequence of having identified bimodal solutions. Their target accuracy can likely be met if the uncertainties in the measurements are further reduced and a unimodal solution found.

The analysis in Section 3.5.4 has highlighted the necessity of the small frequency separation in order to tightly constrain the ages of field stars. The predictions for age in Figure 3.9 are therefore as expected. The inclusion of oscillation frequencies and determination of the small frequency separation (and ratios) from PLATO data result in age uncertainties for solar-like stars to within the 10% level. Without information from the core, ages for TESS targets remain largely unconstrained and consistent with the accuracy typically expected when dating field stars spectroscopically.

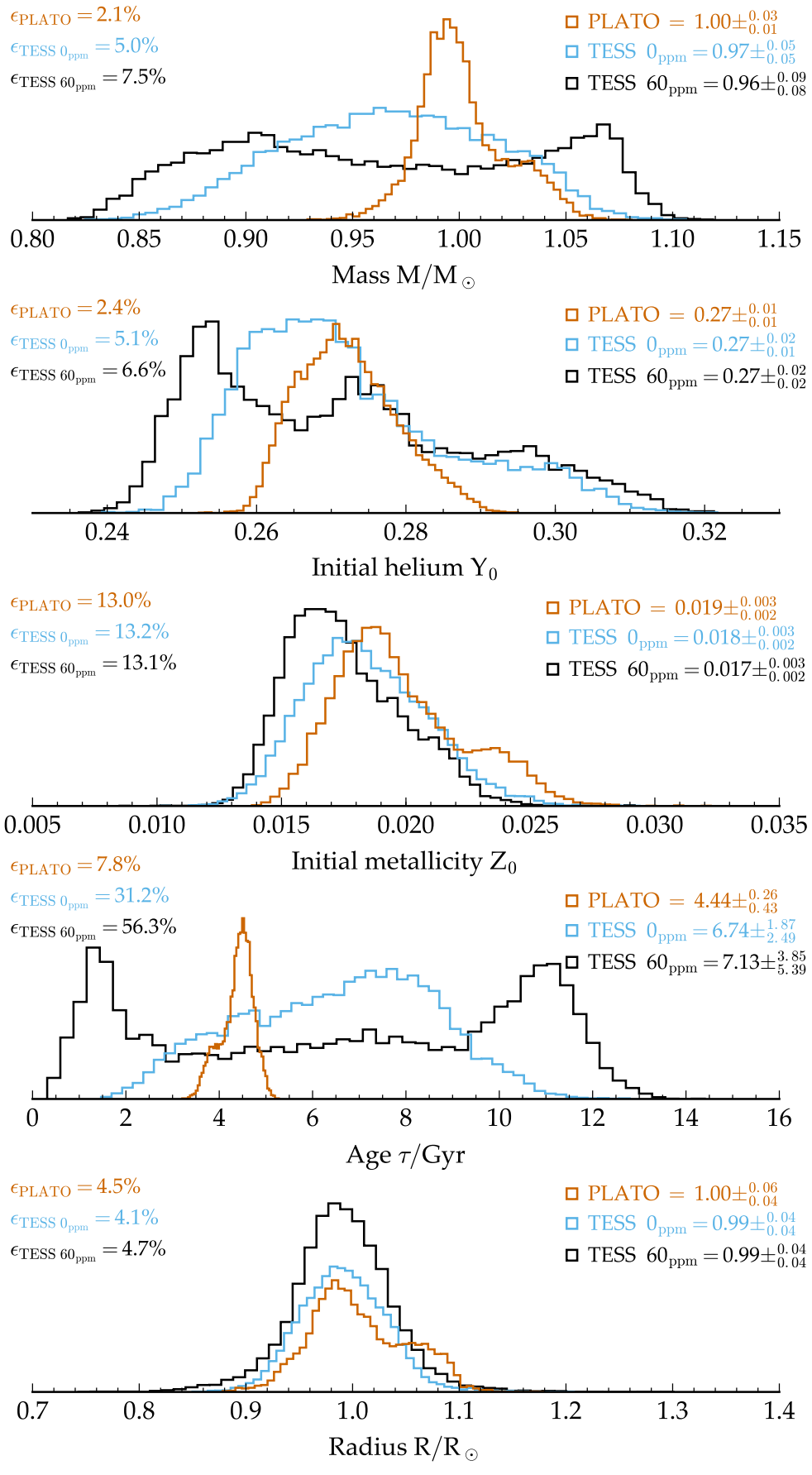


FIGURE 3.9. (Caption on other page.)

FIGURE 3.9. Predictions for the 'Sun as a star' using observations expected for targets from TESS (assuming two different systematic noise levels) and PLATO space missions. In each panel we list the median with uncertainties (84%–50% confidence intervals and 50%–16% confidence intervals) for the quantities as well as the relative error in our prediction.

TABLE 3.12. Solar data degraded to the level expected for sun-like stars in: the TESS catalogue assuming systematic noise of 60 ppm $hr^{1/2}$ from the mission, TESS assuming no systematic noise and from PLATO. For each set of observables we include the feature importances from the random forest used in characterizing the 'Sun as a star.' Note that in the case of the expected PLATO data we have perturbed a subset of frequencies according to their distance from ν_{max}. The numbers reported for the separations and ratios are thus the respective means and standard deviations of 10,000 perturbations to the data which we evaluate to determine our parameter distributions.

	TESS (60 ppm $hr^{1/2}$)			TESS (0 ppm $hr^{1/2}$)			PLATO		
Parameter	Value	Uncertainty	Importance	Value	Uncertainty	Importance	Value	Uncertainty	Importance
T_{eff} (K)	5778	100	29.3%	5778	100	26.7%	5778	100	16.2%
[Fe/H]	-0.014	0.021	34.3%	-0.014	0.021	33.4%	-0.014	0.021	27.9%
log g	4.43	0.07	18.5%	4.43	0.07	12.4%	4.43	0.07	8.8
L ($L/L_\odot$)	0.98	0.04	18.0%	0.98	0.04	16.7%	0.98	0.04	7.8%
ν_{max} (μHz)	–	–	–	3093	100	10.8%	–	–	–
$\langle\Delta\nu_0\rangle$ (μHz)	–	–	–	–	–	–	134.81	0.05	6.4%
$\langle\delta\nu_{02}\rangle$ (μHz)	–	–	–	–	–	–	9.02	0.15	7.1%
$\langle r_{01}\rangle$	–	–	–	–	–	–	0.0226	0.0005	7.4%
$\langle r_{10}\rangle$	–	–	–	–	–	–	0.0227	0.0005	7.3%
$\langle r_{02}\rangle$	–	–	–	–	–	–	0.0668	0.0011	11.1%

3.8 Conclusions

In this work we examined the processes that allow random forest regression to rapidly and accurately infer stellar parameters (Bellinger et al. 2016). We shed light on the inherent properties of the model training data that the algorithm can exploit.

- We demonstrated that there is a large amount of information redundancy in the stellar parameters which is integral to the efficacy of the random forest algorithm. Through statistical bagging, the random forest creates sets of decision rules using different combinations of observables to infer a given quantity. The methodology results in robust predictions and includes the ability to compensate for data that are missing or unreliable.
- We illustrated the behaviour of parameters across the collective lower main sequence with the relationships that arise (e.g., age – luminosity) different to those that develop internally along an evolutionary track. This is the inherent information the random forest draws upon in its regression.
- We found the parameter pairs that exhibit the strongest correlations correspond to well known asteroseismic and main-sequence relations.
- The random forest works well in cases when there is sufficient information and sufficient redundancy. Through principal component analysis we quantified the degree of degeneracy in the observables. Our analysis demonstrated that 99.2% of the variance in the 11 stellar observables could be explained by five principal components.
- The observables we have considered only carry five pieces of independent information. During iterative model searches it is common that independently determined parameters such as ν_{max}, $\langle\Delta\nu_0\rangle$, and $\log g$ are treated as independent degrees of freedom. The composition of the principal components indicate that by not considering their model covariances, any fit is biased towards the common stellar information to which these parameters pertain.
- We devised a score which allows us to rank the degree to which model parameters can be inferred from the observables. Radius, luminosity, and main-sequence lifetime can be extracted with confidence, however, the initial model parameters such as α_{MLT}, Y_0 and α_{ov} are not sufficiently constrained by the observables and cannot be inferred directly from the data. Our analysis can be extended in a straightforward manner to model parameters and observables not considered here.
- Having elucidated the statistical properties of the training data, we sought to better understand how the random forest uses the data in its decision making rules. By performing non-parametric multiple regression with every combination of observable in our grid we determined:

1. which observables are the most important/useful for each model parameter,
2. the minimum set of observables that satisfactorily constrain each model parameter, and
3. the precision with which we can determine each model parameter *directly* from the information contained in the observables.

- We examined the quantities on a parameter by parameter basis and here highlight the results for mass and age. In a grid of stellar evolution models varied in six initial parameters we find that the average error in predicting mass across the grid is ± 0.02 $M_\odot$ and ± 282 Myr for age. The average error in age increases by a factor of three when we are limited to information from only two observables such as in the Christensen-Dalsgaard diagram. Three parameters are sufficient for constraining mass whereas we require five observables to determine age.

- We determined whether the random forest could reproduce the well-known power law that relates $\langle \Delta\nu_0 \rangle$ to ν_{max} and found that additional information from T_{eff} or [Fe/H] reduces the average error in the relation by a factor of two.

- We investigated the measurement accuracy required of the observables to attain a desired precision from the random forest. The processes of statistical bagging and multiple regression help mitigate the impact of large spectroscopic errors as the random draws upon complementary seismic information when devising its decision rules. The results confirm that [Fe/H] and $\langle \delta\nu_{02} \rangle$ are indispensable independent pieces of information for model fitting algorithms.

- Finally, we determined the accuracy and precision with which we can expect to characterize solar-like stars observed by the upcoming TESS and PLATO space missions. In both cases masses can be accurately inferred and measurements from GAIA will ensure that radii are well constrained. Oscillation frequencies will not be detectable in most low-mass main sequence stars observed by TESS. In contrast, the availability of the small frequency separation for PLATO targets will permit accurately determined stellar ages.

Acknowledgements

The research leading to the presented results has received funding from the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013) / ERC grant agreement no 338251 (StellarAges). E.B. undertook this research in the context of the International Max Planck Research School for Solar System Research. S.B. acknowledges partial support from NSF grant AST-1514676 and NASA grant NNX13AE70G. We thank Alexey

Mints and the anonymous referee for their useful comments and discussions which helped improve this manuscript.

Software

Stellar models were calculated with *Modules for Experiments in Stellar Astrophysics* r8118 (MESA, Paxton et al. 2011) and stellar oscillations with the ADIPLS pulsation package 0.2 (Christensen-Dalsgaard 2008). Analysis in this manuscript was performed with python 3.5.1 libraries scikit-learn 0.17.1 (Pedregosa et al. 2011), NumPy 1.11.0 (Van Der Walt et al. 2011), matplotlib 1.5.1 (Hunter 2007), biokit 0.3.2 (Cokelaer 2016) and pandas 0.19.0 (McKinney 2010) as well as R 3.3.2 (R Core Team 2014) and the R libraries magicaxis 2.0.0 (Robotham 2015), RColorBrewer 1.1-2 (Neuwirth 2014), parallelMap 1.3 (Bischl and Lang 2015), data.table 1.9.6 (Dowle et al. 2015), ggplot2 2.1.0 (Wickham 2016), GGally 1.2.0 (Schloerke et al. 2014), scales 0.4.0 (Wickham 2015) and Corrplot 0.77.

3.9 Appendix

3.9.1 Seismic Definitions

We denote any frequency separation S as the difference between a frequency ν of spherical degree ℓ and radial order n and another frequency:

$$S_{(\ell_1,\ell_2)}(n_1, n_2) \equiv \nu_{\ell_1}(n_1) - \nu_{\ell_2}(n_2). \tag{3.12}$$

The large-frequency separation is defined as

$$\Delta\nu_\ell(n) \equiv S_{(\ell,\ell)}(n, n-1) \tag{3.13}$$

and the small-frequency separation is

$$\delta\nu_{(\ell,\ell+2)}(n) \equiv S_{(\ell,\ell+2)}(n, n-1). \tag{3.14}$$

Roxburgh and Vorontsov (2003) have demonstrated that taking the ratio of the *local* large and small-frequency separations reduces the systematic offset introduced from improper modelling of the near-surface super-adiabatic region. This ratio is defined as:

$$r_{(\ell,\ell+2)}(n) \equiv \frac{\delta\nu_{(\ell,\ell+2)}(n)}{\Delta\nu_{(1-\ell)}(n+\ell)}. \tag{3.15}$$

In addition, it was shown that the frequency-dependent offset can be somewhat mitigated by constructing ratios from five-point frequency separations and the *local* large separation:

$$r_{(\ell,1-\ell)}(n) \equiv \frac{dd_{(\ell,1-\ell)}(n)}{\Delta\nu_{(1-\ell)}(n+\ell)} \tag{3.16}$$

where the five point separations are defined as:

$$dd_{0,1} \equiv \frac{1}{8}\big[\nu_0(n-1) - 4\nu_1(n-1) + 6\nu_0(n) - 4\nu_1(n) + \nu_0(n+1)\big] \tag{3.17}$$

$$dd_{1,0} \equiv -\frac{1}{8}\big[\nu_1(n-1) - 4\nu_0(n) + 6\nu_1(n) - 4\nu_0(n+1) + \nu_1(n+1)\big]. \tag{3.18}$$

We calculate dozens of oscillation frequencies per star with the mode sets available dependent on the internal structure of an individual model. We thus determine a single representative value by following the prescription of Mosser et al. (2012). In order to mimic how the oscillation spectra would appear in an observational data, we weight all frequencies by their position in a Gaussian envelope with full-width at half-maximum of $0.66 \cdot \nu_{max}{}^{0.88}$ and centered at the predicted frequency of maximum oscillation power ν_{max}. We then calculate the weighted median of each variable, which we denote with angled parentheses (e.g. $\langle r_{1,0} \rangle$).

3.9.2 Asteroseismic Scaling Relations

$$\nu_{max} \approx \frac{M/M_\odot (T_{eff}/T_{eff,\odot})^{3.5}}{L/L_\odot} \nu_{max,\odot} \tag{3.19}$$

$$\Delta\nu \approx \frac{(M/M_\odot)^{0.5} (T_{eff}/T_{eff,\odot})^{3}}{(L/L_\odot)^{0.75}} \Delta\nu_\odot \tag{3.20}$$

Guggenberger et al. (2016) have shown that a metallicity-dependent correction is required for the Equation (3.20) scaling relation. The $\Delta\nu_\odot$ term can be replaced with a more appropriate reference value which can be calcuated according to:

$$\Delta\nu_{ref} = A \cdot e^{\lambda T_{eff}/10^4 K} \cdot (\cos\Big(\omega \cdot T_{eff}/10^4 K + \phi\Big)) + B, \tag{3.21}$$

and where the unkown terms are listed in Table 3.13.

3.9.3 Correlation Plot

The full BA1 grid introduces some biases in our correlation analysis, particularly from tracks with calculated with high-mass and/or high-diffusion. Correlation analysis with all models included are presented in Figure 3.10.

TABLE 3.13. Parameters of the correction function.

A	0.64·[Fe/H] + 1.78 μHz
λ	−0.55·[Fe/H] + 1.23
ω	22.21 rad/K
ϕ	0.48·[Fe/H] + 0.12
B	0.66·[Fe/H] + 134.92 μHz

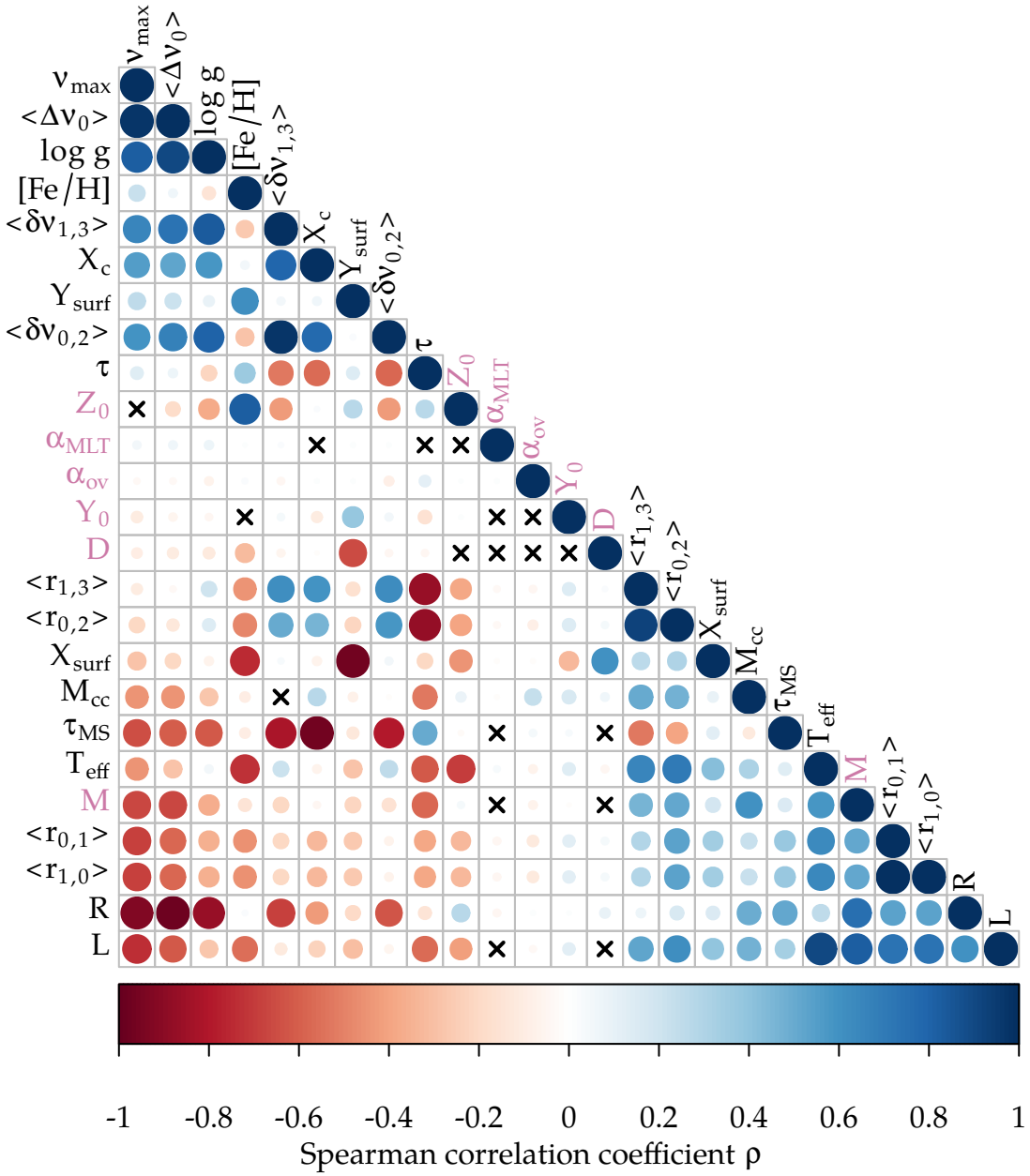


FIGURE 3.10. Spearman rank correlation matrix comprising various stellar and asteroseismic parameters. The quantities are as described in Table 3.1 with model input parameters marked in purple above. The complete grid of models are considered here. The size and color of each circle indicates the sign and magnitude of the Spearman rank coefficient, ρ, between two variables. All correlations are significant excepting the entries indicated with a cross. The variables are ordered by the first principal component of the correlation matrix.

A major difference that arises between Figure 3.2 and Figure 3.10 is in the ordering of variables. Recall that we report the quantities according to the first principal component of the correlation matrix. Different combinations of variables are required to maximise the variance of each principal component in the new parameter space. Although the PCA analysis in Figures 3.11 and Figures 3.12 rely on Pearson rather than Spearman correlations, they do demonstrate the difference in the composition of the PCs in each grid.

We also find differences in the correlations that pertain to current surface abundance parameters. Consider the pair M – Y_{surf}. In Figure 3.10 we find a small but non-negligible negative correlation. The reason being that higher mass tracks diffuse the helium from their surface more efficiently than low-mass stars. Without the influence of these stars in our sample, our significance test yields a null correlation in Figure 3.2; the expected result from a quasi-random distribution of initial abundances.

Two interesting features emanating from our grid selection relates to the parameter pairs $\langle \delta\nu_{02} \rangle$ – T_{eff} and $\langle r_{02} \rangle$ – $\log g$. We find a null correlation between $\langle \delta\nu_{02} \rangle$ – T_{eff} in truncated grid however this emerges as a small positive correlation when the full grid is considered. In Section 3.5.1 we discussed the redundancy in the C–D diagram when projecting stellar models varied in six dimensions into a two-dimensional parameter space. Thus the null correlation arising from the truncated grid reflects the fact there many combinations of (primarily) mass and metallicity and hence temperature at a given age. The full grid, however, consists of a large number of hot short-lived stars that impart a noticeable trend.

A similar argument applies to $\langle r_{02} \rangle$ – $\log g$. There are a great number of combinations of $\langle \Delta\nu_0 \rangle$ and $\langle \delta\nu_{02} \rangle$ for a given $\langle r_{02} \rangle$ thus in the truncated grid no correlation with $\log g$ is registered. Once again the number of massive short-lived stars bias this previous null correlation.

Finally we note two minor results. Some pairs of parameters in the truncated grid which report null correlations in Figure 3.2, show very weak correlations in Figure 3.10. We refer to L – α_{ov} and α_{MLT} – $\langle \delta\nu_{02} \rangle$ as cases in point. The correlations remain very weak in the current analysis and the larger sample size has introduced a minor trend that in this case passes our conservative significance criterion. We note also that most variables display a much stronger correlation with age in the full grid.

3.9.4 Principal Component Analysis Explained Variance

The PCs and their correlations will change depending on the number of dimensions included in the grid and the range of values each parameter takes; the PCs identify vectors of maximal variance. Our aim is to determine whether the PCs capture fundamental features ubiquitously encoded in the observables. Thus, we wish to investigate the information inherent to the dimensions and mitigate the impact of parameter ranges on our PCs. In order to provide a more robust interpretation we have calculated the PCs and their correlations with four different considerations given to the BA1 grid:

Grid A The full BA1 training grid;

Grid B The truncated grid;

Grid C A grid where more than half the models in each track possess metallicities of $[Fe/H] > -2$; and

Grid D A grid with masses limited to $M < 1.2\ M_{\odot}$.

Qualitative correlations between the stellar parameters and the PCs in each grid are presented in Figures 3.11 and Figures 3.12.

3.9.5 PCA Correlation Analysis

Figures 3.4 and 3.5 demonstrate the correlation strengths between our stellar parameters and the first five PCs. In Tables 3.15 and 3.15 we list the coefficients between all parameters and all PCs. The table is useful for determining whether the transitive criterion applies to parameters within a given PC. It also aids in the calculation of the Λ scores in Section 3.4.3.

TABLE 3.14. Percentage of the variance explained by each principal component. We report the explained variance percentages for the complete grid of training models (Grid A) and for the truncated set (Grid B, see Section 3.3) that better encompasses the observational parameter space. In each case we consider the grid with and without the inclusion of ν_{max} which is estimated using the Kjeldsen and Bedding (1995) scaling relations rather than calculated from first principle equations. We also consider the explained variances when limits are placed on the metallicity (Grid C) and mass (Grid D) ranges of the models. These grids are used in Section 3.7 to help interpret the PCs.

	ν_{max} Included				ν_{max} Excluded	
Component	Grid A	Grid B	Grid C	Grid D	Grid A	Grid B
PC_1	41.79	42.36	42.49	42.74	40.89	41.47
PC_2	36.12	34.18	37.49	35.89	36.52	33.65
PC_3	9.17	11.65	9.39	10.25	8.99	12.21
PC_4	7.69	9.79	7.69	6.89	8.27	10.58
PC_5	4.23	1.23	2.14	3.36	4.55	1.36
PC_6	0.54	0.48	0.41	0.53	0.48	0.51
PC_7	0.25	0.18	0.24	0.18	0.16	0.12
PC_8	0.12	0.08	0.09	0.10	0.10	0.09
PC_9	0.05	0.03	0.04	0.04	0.03	0.02
PC_{10}	0.02	0.01	0.01	0.01	0.01	0.00
PC_{11}	0.01	0.00	0.00	0.00	–	–

TABLE 3.15. Pearson's r coefficients between the principal components and observables in the truncated grid.

	$\log g$	T_{eff}	[Fe/H]	$\langle\Delta\nu_0\rangle$	$\langle\delta\nu_{02}\rangle$	$\langle r_{02}\rangle$	$\langle r_{01}\rangle$	$\langle\delta\nu_{13}\rangle$	$\langle r_{13}\rangle$	$\langle r_{10}\rangle$	ν_{max}
PC_1	0.93	-0.20	-0.35	0.92	0.93	0.38	-0.07	0.95	0.32	-0.07	0.87
PC_2	-0.30	0.73	-0.29	-0.33	0.34	0.85	0.81	0.22	0.76	0.81	-0.42
PC_3	0.00	-0.60	0.63	0.04	0.06	0.15	0.45	-0.13	-0.17	0.45	0.22
PC_4	0.08	0.08	-0.60	0.19	-0.08	-0.30	0.37	-0.13	-0.52	0.37	0.10
PC_5	0.14	0.25	0.20	0.06	-0.02	-0.05	0.03	0.00	-0.07	0.03	0.01
PC_6	0.11	-0.01	-0.04	0.01	-0.11	0.09	0.00	-0.12	0.06	0.00	0.04
PC_7	-0.09	0.03	0.01	0.05	-0.03	-0.01	0.00	-0.01	0.03	0.00	0.09
PC_8	0.02	-0.02	0.00	0.00	-0.04	-0.05	0.02	0.03	0.05	0.02	-0.01
PC_9	0.01	0.01	0.00	-0.04	0.02	-0.02	0.00	-0.02	0.01	0.00	0.03
PC_{10}	0.00	0.00	0.00	0.02	0.02	-0.01	0.00	-0.02	0.01	0.00	-0.01
PC_{11}	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	-0.01	0.00

TABLE 3.16. Pearson's r coefficients between the principal components and model parameters in the truncated grid.

	M	Y	Z	α_{MLT}	α_{ov}	D	τ	τ_{MS}	X_c	M_{cc}	X_{surf}	Y_{surf}	R	L
PC_1	-0.67	-0.08	-0.30	0.06	-0.05	-0.12	-0.19	-0.72	0.77	-0.45	-0.04	0.16	-0.86	-0.69
PC_2	0.29	0.10	-0.41	-0.15	-0.04	-0.17	-0.56	-0.26	0.14	0.17	0.05	0.11	0.30	0.55
PC_3	0.13	0.04	0.48	-0.10	-0.07	-0.03	-0.01	-0.04	0.14	-0.11	-0.33	0.17	0.09	-0.04
PC_4	-0.51	-0.17	-0.40	0.03	-0.09	-0.08	0.51	0.48	-0.49	-0.34	0.29	-0.17	-0.23	-0.13
PC_5	-0.02	0.17	-0.09	0.44	-0.16	-0.43	0.13	0.13	-0.17	-0.29	-0.50	0.59	-0.14	-0.03
PC_6	-0.02	-0.20	0.08	-0.01	-0.15	-0.10	0.28	0.07	-0.02	-0.25	0.05	-0.10	-0.14	-0.03
PC_7	0.09	0.05	-0.05	0.14	0.11	-0.15	-0.10	0.03	0.10	0.30	-0.14	0.17	0.18	0.38
PC_8	-0.11	0.08	-0.06	-0.25	0.05	-0.07	0.03	-0.03	-0.02	-0.12	-0.05	0.08	-0.07	-0.09
PC_9	0.21	-0.35	0.22	0.20	-0.06	-0.08	-0.22	-0.12	0.15	0.04	-0.01	-0.10	0.06	0.07
PC_{10}	-0.17	0.26	-0.14	-0.15	0.02	0.06	0.17	0.12	-0.10	-0.01	0.01	0.06	-0.09	-0.04
PC_{11}	-0.01	-0.01	-0.01	0.01	0.01	0.00	0.00	0.00	0.00	-0.02	0.01	-0.01	0.00	-0.01

3.9.6 PC correlations with different grids

In Section 3.4.2 we presented the correlation strengths between the PCs and observables (Figure 3.4) and the PCs and the model parameters (3.5). Here we perform the same analysis with the different subsets of the BA1 grid described in Appendix 3.9.4. In order to compare the results for each grid, in Figures 3.11 and 3.12 we employ a correlation plot rather than the quantitative bar chart used in Section 3.4.2. This allows an inspection of the qualitative behaviour of the PCs in each case. We find a similar explained variance from the corresponding PCs in each grid. This suggests that the PCs capture essentially the same inherent features in model data and that the PCs are not due to the number of models in our analysis or the chosen parameter ranges.

3.9.7 Λ Analysis

The data matrix of observables $\mathbf{X}$ is size $n \times p$ where n is the number of training models and p the number of parameters. We centre and scale the entries according to the mean and standard deviation of each parameter. The resultant matrix, $\bar{\mathbf{X}}$, therefore has the property that for each parameter, p, $\mu(p) = 0$ and $\sigma(p) = 1$. We compute the correlation matrix, R, for the matrix $\bar{\mathbf{X}}$:

$$\begin{aligned} \mathbf{R} &= \mathrm{Corr}(\bar{\mathbf{X}}) \\ &= \bar{\mathbf{X}}\bar{\mathbf{X}}^{\top}. \end{aligned} \tag{3.22}$$

As the correlation and covariance matrices are symmetric we calculate the eigendecomposition of R such that:

$$\mathbf{R} = \mathbf{V}\mathbf{L}\mathbf{V}^{\top}, \tag{3.23}$$

where $\mathbf{V}$ a matrix of eigenvector columns and $\mathbf{L}$ a diagonal matrix of eigenvalues. The eigenvectors specify the principal axes of the data and the eigenvalues indicate the amount of variance there is in the data in the direction of the corresponding eigenvector. We can define the projection matrix $\mathbf{P}$ such that we project/transform our data into the new space

$$\mathbf{P} = \bar{\mathbf{X}}\mathbf{V}. \tag{3.24}$$

The correlation matrix is a special case of the covariance matrix in that the former is normalised. For generality let us consider the covariance matrix, such that the original data matrix was centred but not scaled ($\hat{\mathbf{X}}$), then

$$\begin{aligned} \mathbf{C} &= \mathrm{Cov}(\hat{\mathbf{X}}) \\ &= \tfrac{1}{n-1}\hat{\mathbf{X}}\hat{\mathbf{X}}^{\top} \\ &= \mathbf{V}\mathbf{L}\mathbf{V}^{\top}, \end{aligned} \tag{3.25}$$

where we divide by (n-1) to unbias to covariance (the covariance entries will have different scales).

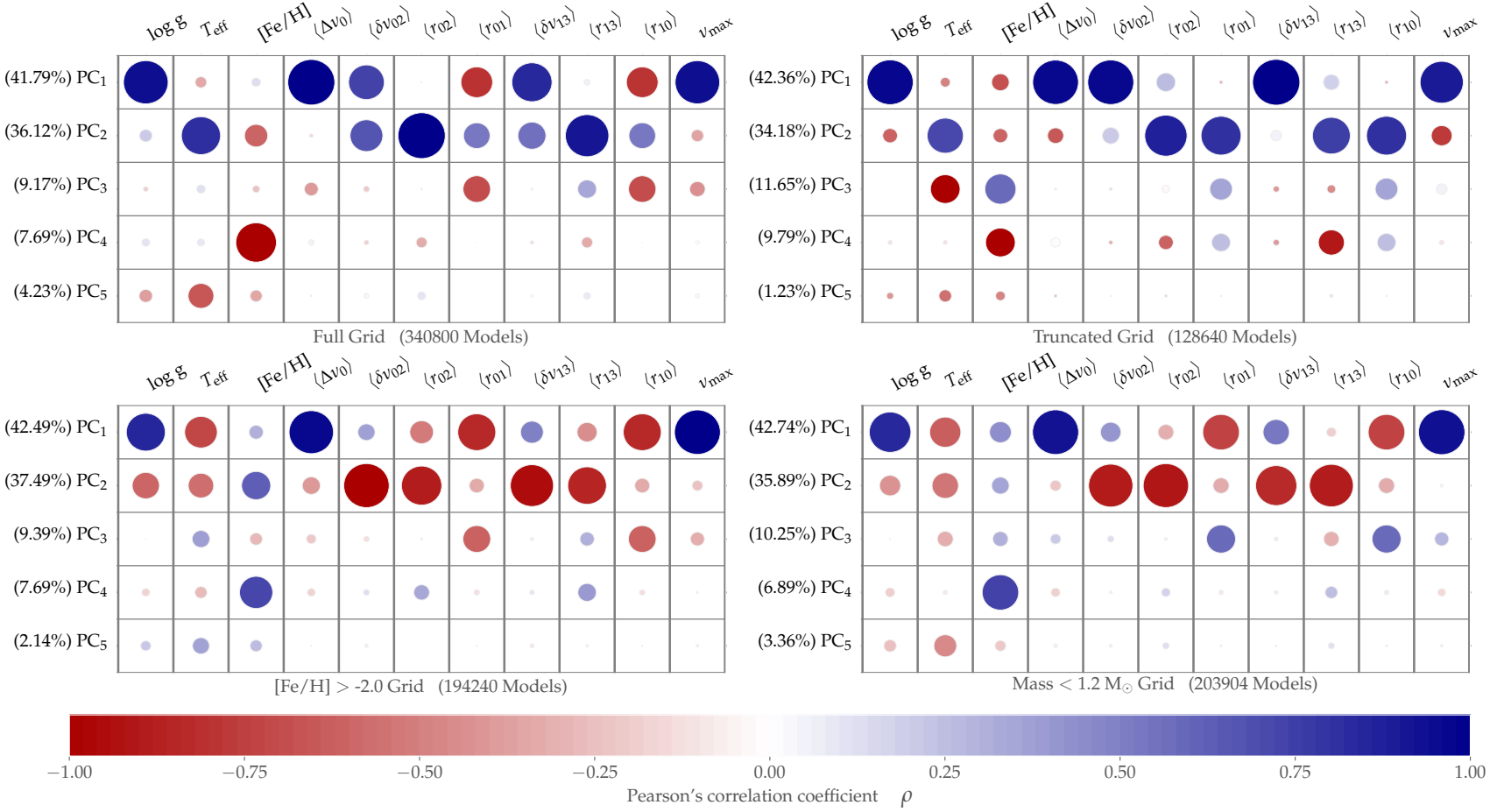


FIGURE 3.11. Pearson correlation matrices relating the principal components back to the stellar observables in each of the four grids described in Appendix 3.9.4.

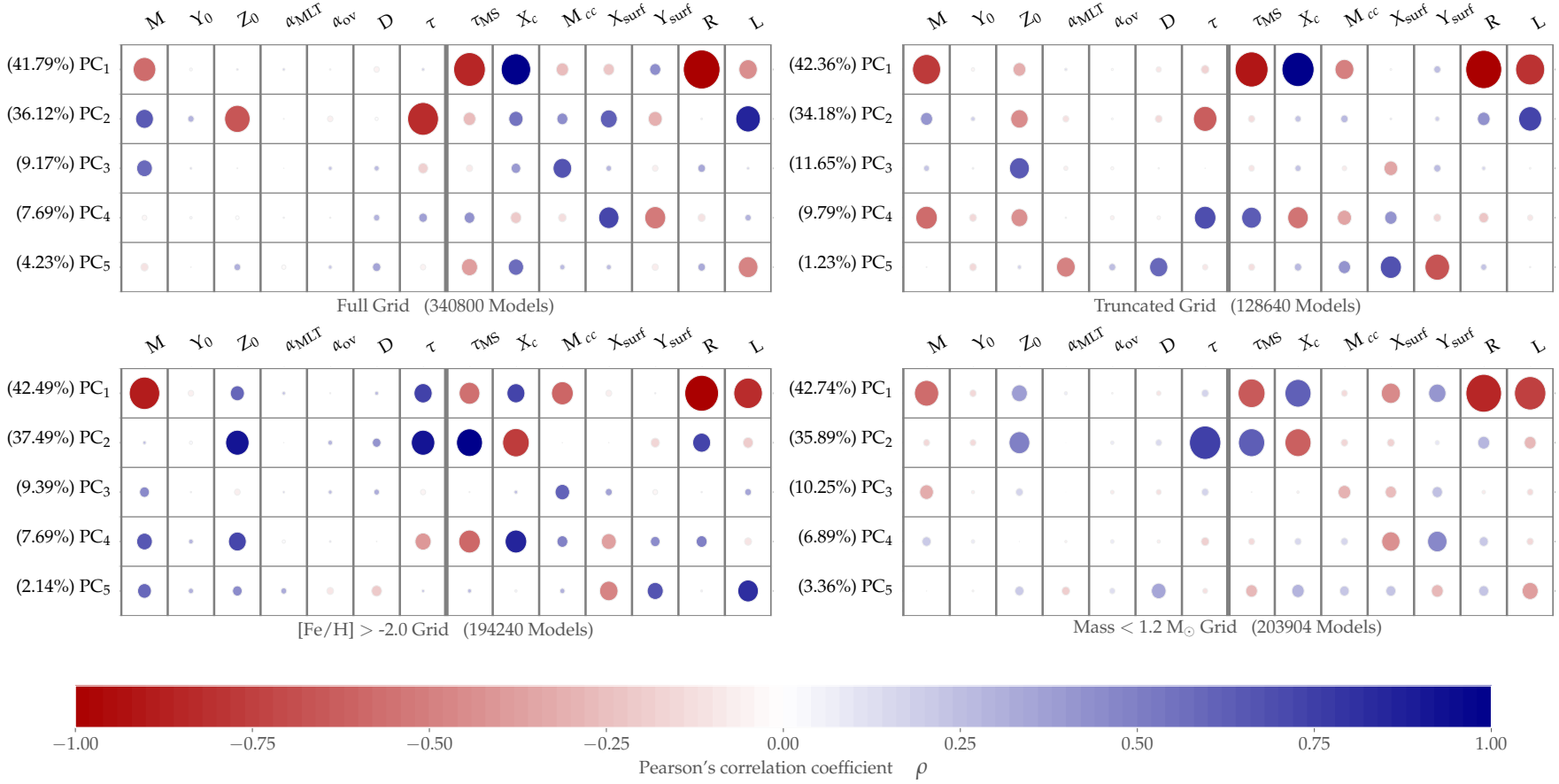


FIGURE 3.12. Pearson correlation matrices relating the principal components back to the model quantities in each of the four grids described in Appendix 3.9.4.

Alternatively and equivalently, we may extract our PCs through SVD of $\hat{\mathbf{X}}$ such that:

$$\hat{\mathbf{X}} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top \tag{3.26}$$

where $\mathbf{U}$ is the left matrix of singular orthogonal vectors with dimensions $n \times n$, $\mathbf{\Sigma}$ is a diagonal matrix of singular values with dimensions $n \times p$, and $\mathbf{V}^\top$ is the right matrix of singular orthogonal vectors with diemsnions $p \times p$. The diagonal elements of $\mathbf{\Sigma}$ assign a relative importance to each vector whereas the vectors of $\mathbf{V}$ are the principal directions/axes. As the matricies $\mathbf{U}$ and $\mathbf{V}$ comprise orthogonal components they have the property

$$\begin{aligned} \mathbf{U}^\top\mathbf{U} &= \mathbf{I}_{n\times n} \\ \mathbf{V}^\top\mathbf{V} &= \mathbf{I}_{p\times p}. \end{aligned} \tag{3.27}$$

We note also that

$$(\mathbf{A}\cdot\mathbf{B}\cdot\mathbf{C})^\top = \mathbf{C}^\top\cdot\mathbf{B}^\top\cdot\mathbf{A}^\top \tag{3.28}$$

$$\implies (\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top)^\top = (\mathbf{V}\mathbf{\Sigma}\mathbf{U}^\top) \tag{3.29}$$

as $\mathbf{\Sigma}$ is a diagonal matrix.

We can reconstruct the eigendecomposition of the covariance matrix from the SVD:

$$\begin{aligned} \frac{1}{n-1}\hat{\mathbf{X}}\hat{\mathbf{X}}^\top &= \tfrac{1}{n-1}(\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top)(\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top)^\top \\ &= \tfrac{1}{n-1}(\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top)(\mathbf{V}\mathbf{\Sigma}\mathbf{U}^\top) \end{aligned} \tag{3.30}$$

and from our identities in Equation (3.27)

$$\frac{1}{n-1}\hat{\mathbf{X}}\hat{\mathbf{X}}^\top = \mathbf{U}\frac{\mathbf{\Sigma}^2}{n-1}\mathbf{U}^\top. \tag{3.31}$$

We therefore find that the square roots of the eigenvalues of $\mathbf{C}$ are the singular values of $\bar{\mathbf{X}}$ and that the vectors in the right singular matrix, $\mathbf{V}$, are the principal directions/axes. The projection matrix can be calculated from the SVD such that

$$\begin{aligned} \mathbf{P} &= \hat{\mathbf{X}}\mathbf{V} \\ &= \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top\mathbf{V} \\ &= \mathbf{U}\mathbf{\Sigma}. \end{aligned} \tag{3.32}$$

The PCA loadings are the columns of $\mathbf{L}$ which implies that

$$\mathbf{L} = \mathbf{V}\frac{\mathbf{\Sigma}}{\sqrt{n-1}}. \tag{3.33}$$

We can see that the loadings are the eigenvectors scaled by the square roots of the respective eigenvalues. With these definitions we can compute the cross-covariance matrix between original variables and the standardized projection

matrix. To calculate the standardized PC scores for $\mathbf{P}$ we require each column of $\mathbf{U}$ to have unit variance. As $\mathbf{\Sigma}$ is diagonal it is simply a scaling matrix and can be dropped here yielding:

$$\frac{1}{n-1}\mathbf{X}^\top(\sqrt{n-1}\mathbf{U}) = \frac{1}{\sqrt{n-1}}\mathbf{V}\mathbf{\Sigma}\mathbf{U}^\top\mathbf{U} \tag{3.34}$$

$$= \frac{1}{\sqrt{n-1}}\mathbf{V}\mathbf{\Sigma} \tag{3.35}$$

$$= \mathbf{L}. \tag{3.36}$$

We find that the covariance matrix between the standardized PCs and original variables is in fact given by the loadings. In Section 3.4.1 we computed the *correlations* between the observables and their PCs rather than the covariances, requiring that the observables are normalized by their standard deviation. As we centred and scaled our data prior to performing the PCA, their values are unity and our correlation analysis is therefore equivalent to reporting the loadings.

The correlation analysis allowed us to project the model data onto the PC space and determine the 'equivalent' loadings for each parameter. Through the λ score we can therefore determine to what extent the variance in the model data is captured by the PCs. In Table 3.17 we compare the results of the analysis for each grid. We find similar results for most parameters with differences in some of the initial model parameters due to their underlying distributions as a result of the grid truncations.

3.9.8 Impact of Uncertainties for Upcoming Photometric Space Missions

Below we demonstrate the impact of measurement uncertainty on the prediction of parameters from the upcoming TESS (Figure 3.13) and PLATO (Figure 3.14) space missions. We produce probability density distributions for 250 sets of σ values for each parameter we predict. The ranges for each parameter from which we draw our σ values are listed in Table 3.18. We restrict out observables to those we are likely to possess from the respective missions. In each figure we plot the median value (solid line) and the 68% confidence interval (shaded region).

TABLE 3.17. The Λ score is a sum of the squares of $r(X, PC_i)$ indicating the variance explained for a given parameter. These scores are by definition unity for our observables.

	Λ_{param}			
Parameter	Grid A	Grid B	Grid C	Grid D
R	0.97	0.97	0.98	0.97
L	0.93	0.96	0.93	0.95
X_c	0.93	0.94	0.93	0.94
τ_{MS}	0.93	0.93	0.93	0.94
M	0.91	0.91	0.92	0.88
τ	0.74	0.79	0.78	0.76
Z_0	0.76	0.73	0.78	0.80
M_{cc}	0.58	0.61	0.68	0.41
Y_{surf}	0.48	0.50	0.55	0.54
X_{surf}	0.50	0.48	0.53	0.55
α_{MLT}	0.02	0.38	0.04	0.06
Y_0	0.10	0.31	0.27	0.09
D	0.13	0.29	0.22	0.21
α_{ov}	0.10	0.08	0.11	0.12

TABLE 3.18. Central solar values and uncertainty ranges used for predictions in Figures 3.13 and 3.14.

	TESS			PLATO		
Quantity	Value	Min(σ)	Max(σ)	Value	Min(σ)	Max(σ)
T_{eff} (K)	5777	10	500	5777	10	500
log g	4.44	0.0001	1.0	4.44	0.0001	1.0
[Fe/H]	0.0	0.05	0.5	0.0	0.05	0.5
L	1.0	0.001	10	1.0	0.001	10
ν_{max}	3050	10	500	–	–	–
$\langle\Delta\nu_0\rangle$ (μHz)	–	–	–	136.0	0.5	50
$\langle\delta\nu_{02}\rangle$ (μHz)	–	–	–	9.0	0.5	5

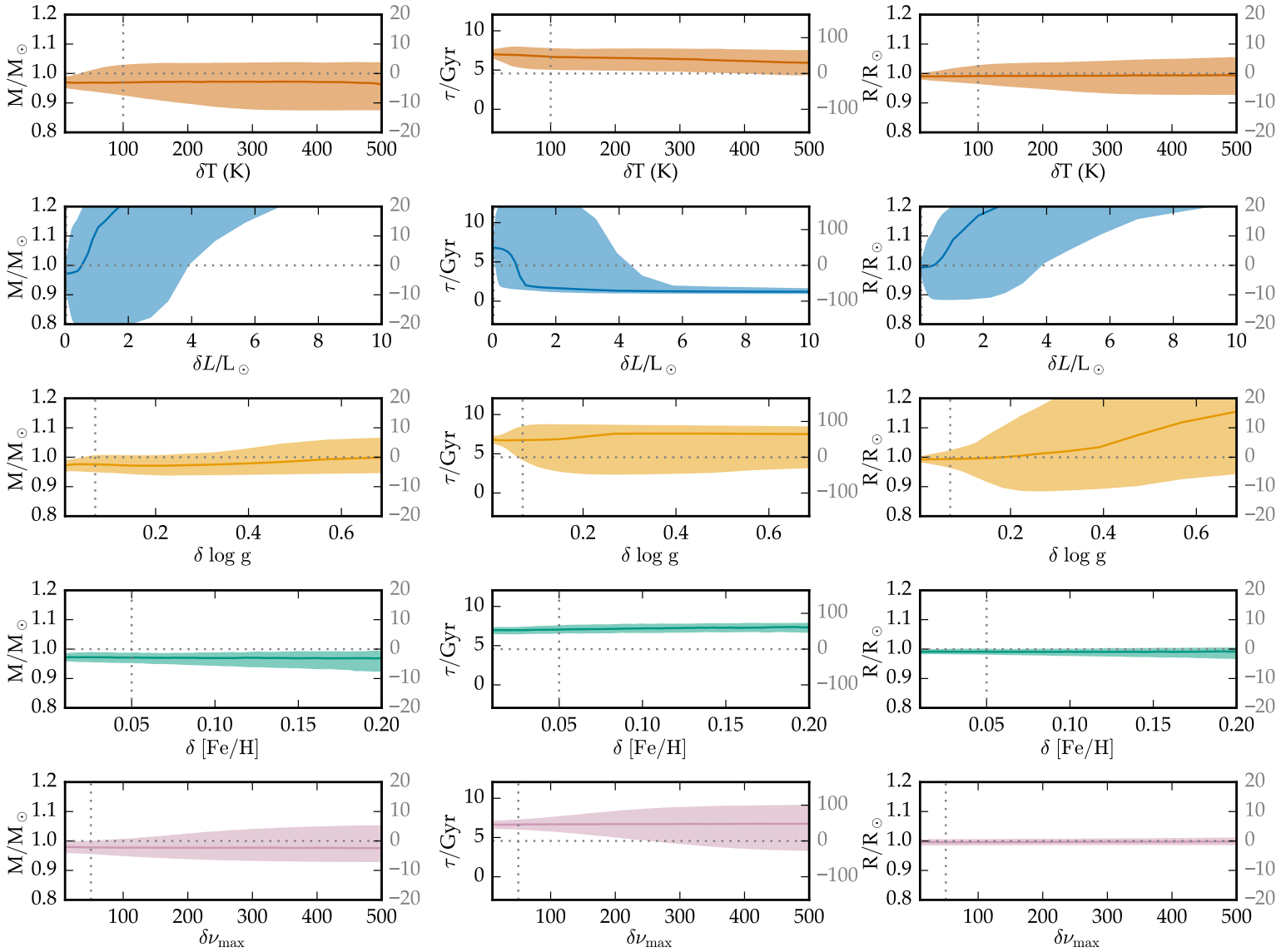


FIGURE 3.13. (Caption on other page.)

FIGURE 3.13. Predictions for the solar mass, age, luminosity and radius as a function of the uncertainties applied to key observables. In each panel we have perturbed the quantity on the abscissa in isolation, centred around the measured value listed in Table 3.18 and with the uncertainties in the ranges specified therein. We indicate the median predicted value (solid line) and the 68% confidence interval (shaded region). Here the observables comprise those expected from the TESS space mission assuming that the p-mode power excess can be extracted.

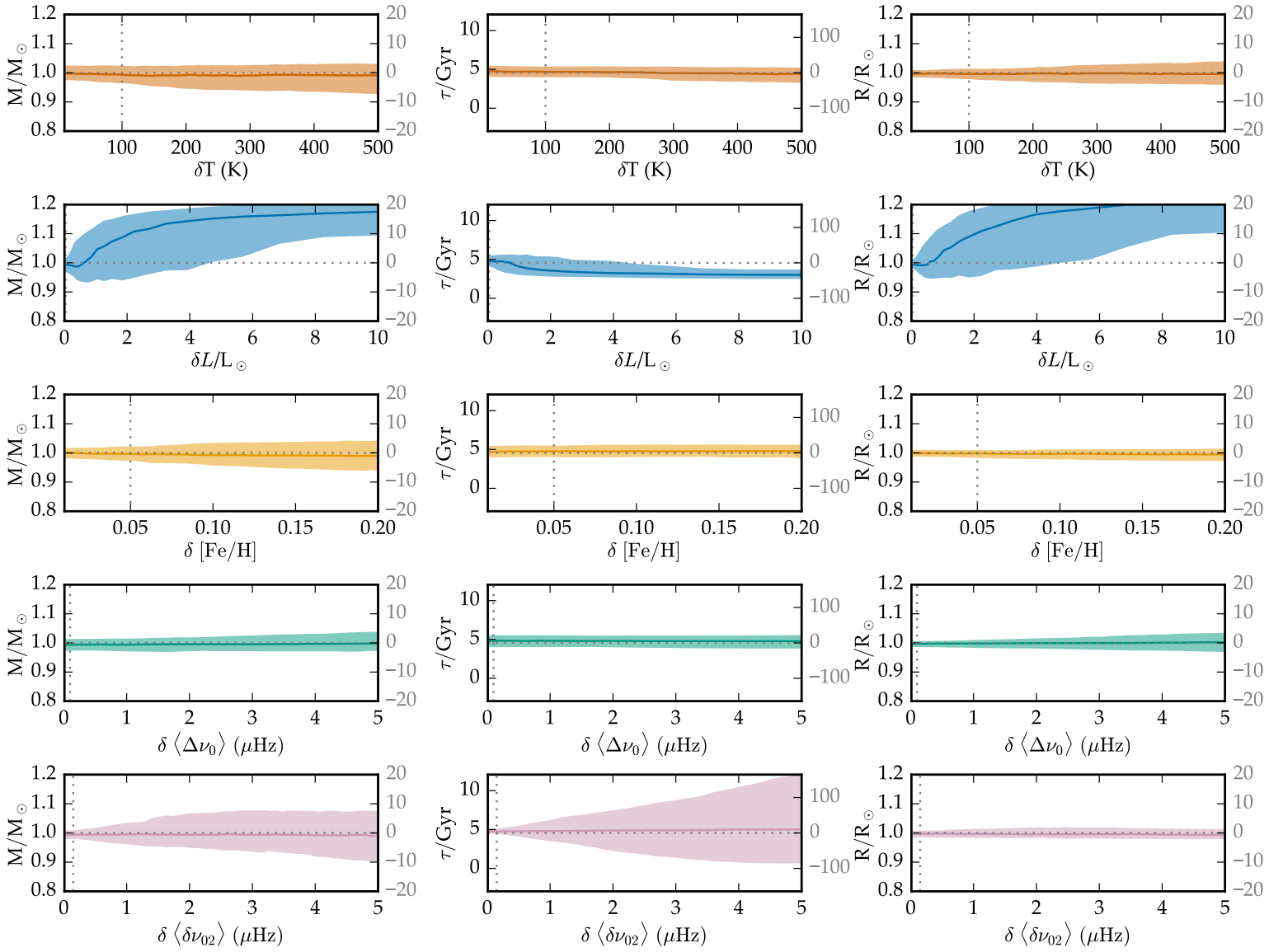


FIGURE 3.14. The same as Figure 3.13, but for PLATO.

4

Model-Independent Measurement of Internal Stellar Structure in 16 Cygni A & B

The contents of this chapter were authored by E. P. Bellinger, S. Basu, S. Hekker, and W. H. Ball and published in December of 2017 in *The Astrophysical Journal*, 851 (2), 80.[1]

Chapter Summary

We present a method for measuring internal stellar structure based on asteroseismology that we call "inversions for agreement." The method accounts for imprecise estimates of stellar mass and radius as well as the relatively limited oscillation mode sets that are available for distant stars. By construction, the results of the method are independent of stellar models. We apply this method to measure the isothermal sound speeds in the cores of the solar-type stars 16 Cyg A and B using asteroseismic data obtained from *Kepler* observations. We compare the asteroseismic structure that we deduce against best-fitting evolutionary models and find that the sound speeds in the cores of these stars exceed those of the models.

[1] Contribution statement: The work of this chapter was carried out and written by me, under the supervision of S. Basu and S. Hekker and in collaboration with W. H. Ball.

4.1 Introduction

The detection and study of internal waves in stars—asteroseismology—provides a unique view into stellar interiors. As the structure of a star dictates the varieties and frequencies of its normal modes of oscillation, asteroseismic data can be used to set limits on the conditions inside a star. This is usually achieved by evolving stellar models, and the structure of the best-fitting model is then assumed to be a proxy for the structure of the star. However, theoretical pulsation frequencies of even the best stellar models have significant discrepancies with observations, implying that the structure of the star differs from the structure of the model. This is true for the Sun and other stars alike. A way to proceed from this point would be to quantify what internal conditions do support the oscillations that have been observed. This problem is the inverse of determining the mode frequencies of a known stellar structure, and is thus known as a *structure inversion*. Structure inversions are of value because their results are independent of models. However, the structure inversion problem is ill-posed in the sense described by Hadamard (1902) and therefore difficult to solve, especially given the relatively limited data that are available for other stars. Consequently, structure inversions for internal properties such as the sound-speed profile have thus far been restricted to the Sun and other bodies within the solar system. In this paper, we present results of structure inversions performed to probe core structure in other stars. More specifically, we invert measured p-mode frequencies to deduce the squared isothermal sound speed ($u \equiv P/\rho$, where P is pressure and ρ is density), in the cores of the two solar analogs 16 Cyg A and 16 Cyg B. We achieve this by introducing an algorithm that we call "inversions for agreement" that works with the available data.

Helioseismic inversions, i.e. inversions for the Sun, have revealed that sound-speed profiles of solar-calibrated evolutionary models differ by only fractions of a percent from the actual structure of the Sun—a rare triumph of accuracy by astrophysical standards. Furthermore, even before all flavors of solar neutrinos could be detected, helioseismic inversions were instrumental in showing that the solar neutrino problem was external to solar modeling (e.g. Antia and Chitre 1997, Bahcall et al. 1998). Additionally, the importance of some physical processes in stellar physics have been revealed by helioseismic inversions as well. For example, by comparing solar models with and without diffusion and gravitational settling of helium and heavy elements, Christensen-Dalsgaard et al. (1993) showed that it is important to take these effects into account (see also Figure 20 of Basu 2016), and it has now become common practice to include these processes when modeling other solar-like stars. Hence, structure inversions are useful for verifying and improving models both within stellar physics and beyond.

The stars we wish to study with structure inversions are pulsating solar-type stars observed by *Kepler*. They are cool dwarf stars on the main sequence that pulsate in pure p-modes and show no signs of mode mixing (for a review of solar-like oscillations, see, e.g., Chaplin and Miglio 2013). The precise mea-

surement of pulsation frequencies in these and other similar stars has enabled estimates of their ages, masses, and radii to better than 15%, 4%, and 2%, respectively (Silva Aguirre et al. 2015, 2017, Bellinger et al. 2016, 2017a, Angelou et al. 2017). The solar-type stars belonging to the triple system of 16 Cygni are two of the most well-studied stars in this field. Though stellar models of these stars match the overall characteristics of the stars, such as their radii, luminosities, temperatures, and metallicities; an inspection of their mode frequencies reveals significant disagreements. Figure 4.1 shows a comparison of mode frequencies between models (Silva Aguirre et al. 2017, models *GOE*) and observations (Davies et al. 2015) of 16 Cyg A and B, with Sun-as-a-star data shown for reference. Clear differences can be seen between the mode frequencies of the evolutionary models and the measured mode frequencies of the stars.

The most conspicuous difference between the oscillations of stars and stellar models is an offset that increases with frequency. This offset arises due to inadequacies in modeling the effects of convection in the near-surface layers (see, e.g., Christensen-Dalsgaard 1984) as well as neglected treatment of pulsation-convection interaction (Houdek et al. 2017). These are collectively known as "surface effects," and the offset they produce is usually called the "surface term." For modes of low spherical degree ℓ, the surface term is a function of frequency alone. There are a number of methods for correcting the disparities imposed by surface effects, such as those given by Kjeldsen et al. (2008), Ball and Gizon (2014, hereinafter BG14), and Sonoi et al. (2015). Each of these methods work by assuming that the frequency offset due to the surface term has a particular form that can be fitted to the frequency differences and subtracted off. Even after correction for the surface term, however, differences remain. Figure 4.2 shows the remaining discrepancies between mode frequencies of models and observations of 16 Cygni after subtracting off the two-term "BG14-2" surface effect. More than half of the surface-term corrected mode frequencies still have significant differences with the observed values. Moreover, the disparities are most significant in the radial and dipole modes, which probe the deep interior of the star.

Since mode frequencies of models produced by stellar evolution codes have significant differences with respect to observations even after correction for the surface term, we pursue the use of inversion techniques to make more direct determinations of stellar structure.

4.1.1 The Inversion Problem

Structure inversions can be posed as the problem of deducing small differences in structure between a star and a sufficiently close reference model by comparison of their mode frequencies. The basic problem is the same as the structure inversion problem for the Sun (for reviews of solar structure inversions, see for example Kosovichev 1999, Basu 2016). The dependence of mode frequencies on the radial structure of a star is nonlinear and involves unobservable displacement eigenfunctions. However, the oscillation equations are, to first order, a set of Hermitian eigenvalue equations (Chandrasekhar 1964), and hence they can be

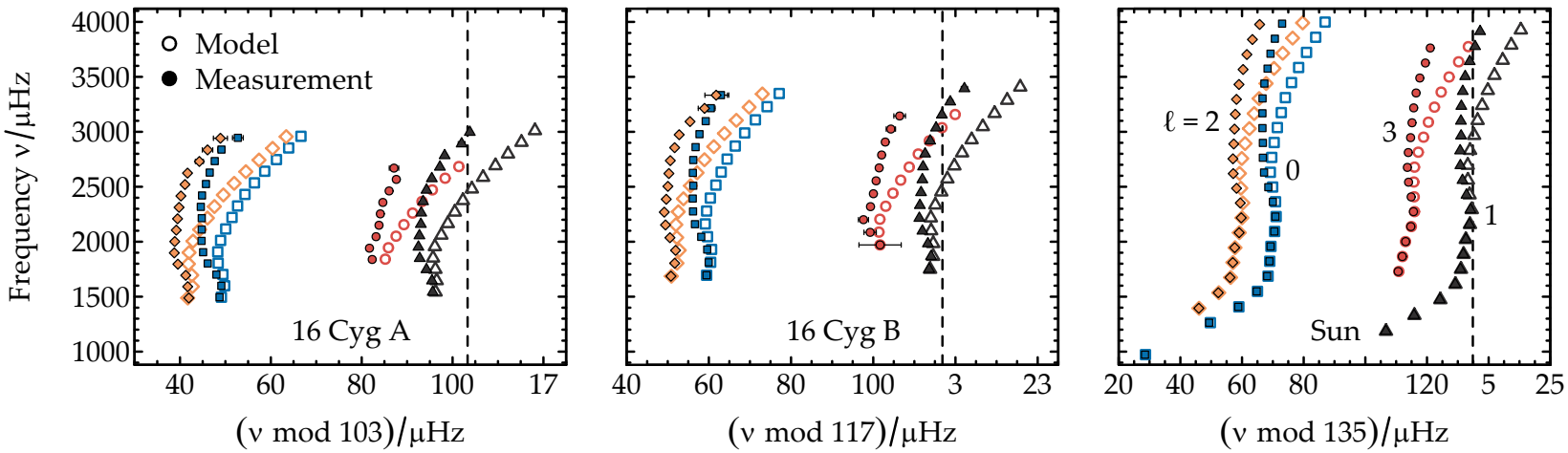


FIGURE 4.1. Échelle diagrams comparing *GOE* evolutionary models of 16 Cyg A (left) and B (center) to frequencies extracted from *Kepler* data. For reference, the right panel shows the solar model Model S (Christensen-Dalsgaard et al. 1996) in comparison with low-degree frequencies of the quiet Sun from BiSON data (Davies et al. 2014a). The dashed line indicates the large frequency separation ($\Delta\nu$). Open symbols are model frequencies and filled symbols are observed frequencies. Spherical degrees ℓ are indicated with color and shape: 0 (blue squares), 1 (black triangles), 2 (yellow diamonds), and 3 (red circles). Error bars show 1σ uncertainties, which in most cases are not visible. Model frequencies significantly differ from observed frequencies in nearly all cases.

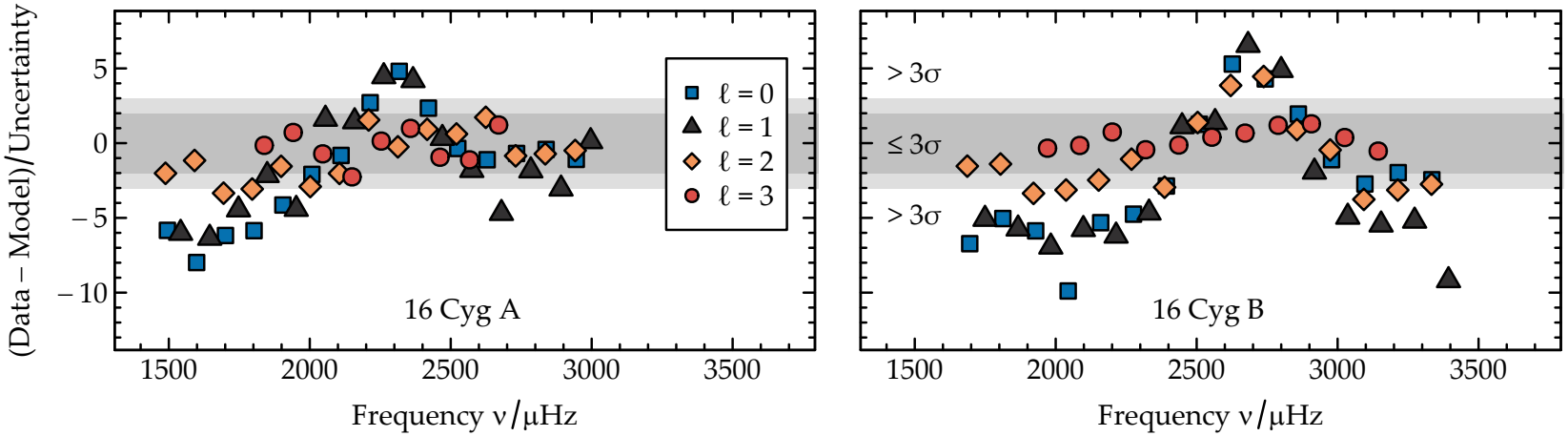


FIGURE 4.2. Differences in oscillation mode frequencies between models and observations after correcting for surface effects. Mode frequencies that lie outside of the shaded regions, demarcating the 2σ and 3σ boundaries, have significant differences that are caused by differences in internal structure.

linearized around a known model using the variational principle. The linearization links the differences in frequencies between the reference model and the star to the differences in their internal structure. A byproduct of the linearization is the fact that the differences must be considered with respect to at least two stellar structure functions simultaneously, as variables such as the sound speed c and density ρ are not independent but rather related through the equations of stellar structure. The equations resulting from the linearization can be written as

$$\mathscr{P}[\nu_i] = \int \mathbf{K}_i(r) \cdot \mathscr{P}[\mathbf{f}(r)]\, dr + \epsilon_i, \quad i \in \mathscr{M} \tag{4.1}$$

where $\mathscr{M}$ is the set of observed modes, $\boldsymbol{\nu}$ are the oscillation frequencies of those modes, $\mathbf{f}$ contains two stellar structure functions (i.e., $f_1(r)$ and $f_2(r)$; e.g. $c(r)$ and $\rho(r)$), r is the fractional stellar radius, and $\mathscr{P}$ is a perturbation operator (in this case, the relative difference operator). Since measurements are uncertain, we include a term $\boldsymbol{\epsilon}$ for the differences between the true and the measured values. Each mode of oscillation i has its own pair of kernels $\mathbf{K}_i$ that relate changes in $\mathbf{f}$ to changes in ν_i. The kernels are derived from the perturbation analysis (see, e.g., Gough and Thompson 1991 or Sec. 6.2. of Basu 2016 for details) and can be computed for a given reference model. Since the eigenproblem is Hermitian, perturbations to the oscillation mode eigenfrequencies do not depend to the first order on perturbations to the mode eigenfunctions. The inverse problem is thus to deduce $\mathbf{f}$ from the data $\boldsymbol{\nu}$, given that the kernels are known. There is no analytic solution to this problem and numerical methods must be employed. In practice, another term must also be added in order to account for the aforementioned surface effects. Although the technique makes use of a reference model, the results are independent; all stellar models within the linear regime produce essentially the same inference about the star (Basu et al. 2000). We expand Equation (4.1) explicitly in the next section.

Like many inverse problems, the structure inversion problem is ill-posed: the solutions are not unique, and they are also unstable with respect to small fluctuations in the oscillation data (see Gough and Thompson 1991 for a discussion). Solutions must therefore be regularized (for a review of statistical regularization, see, e.g., Tenorio 2001). There are two popular ways of inverting Equation (4.1): the Regularized Least Squares (RLS; Tikhonov 1977) fitting method, which attempts to determine the stellar structure functions $\mathbf{f}$ that best fit to the observed data; and (2) the method of Optimally Localized Averages (OLA; Backus and Gilbert 1968), which attempts to make linear combinations of the data that correspond to localized averages of one of the two components of $\mathbf{f}$. Both methods have been used extensively in the case of the Sun. Details of how the inversions are implemented can be found in Basu 2016 and references therein.

In helioseismic investigations, the most common choice of $\mathbf{f}$ is the combination of squared adiabatic sound speed c^2 and density ρ. The kernels for this pair are shown in Figure 4.3. The basic ingredients of helioseismic inversion are the thousands of precisely measured solar mode frequencies whose spherical degrees range up to $\ell \simeq 200$ or higher. Reference models have the same mass,

radius, and age as the Sun. Inversion of helioseismic data yields inferences of solar structure throughout most of the solar interior (see, e.g., Basu et al. 2009).

There are two major difficulties in trying to invert for the structure of other stars. The first difficulty is the lack of data. Even for the best solar-type targets, only about 55 mode frequencies have been able to be measured. Furthermore, due to cancellation effects, we only get data for low-degree modes, usually of degree $\ell = 0 - -2$ and sometimes 3. This limits the regions in the star that we are able to probe, the inversion techniques that we are able to employ, and the pair of stellar structure functions that we are able to use. Second, when compared with the Sun, masses and radii of stars are not known with the same precision. This is problematic because differences in mass and radius between the reference model and the proxy star cause systematic errors in the inversion results (see

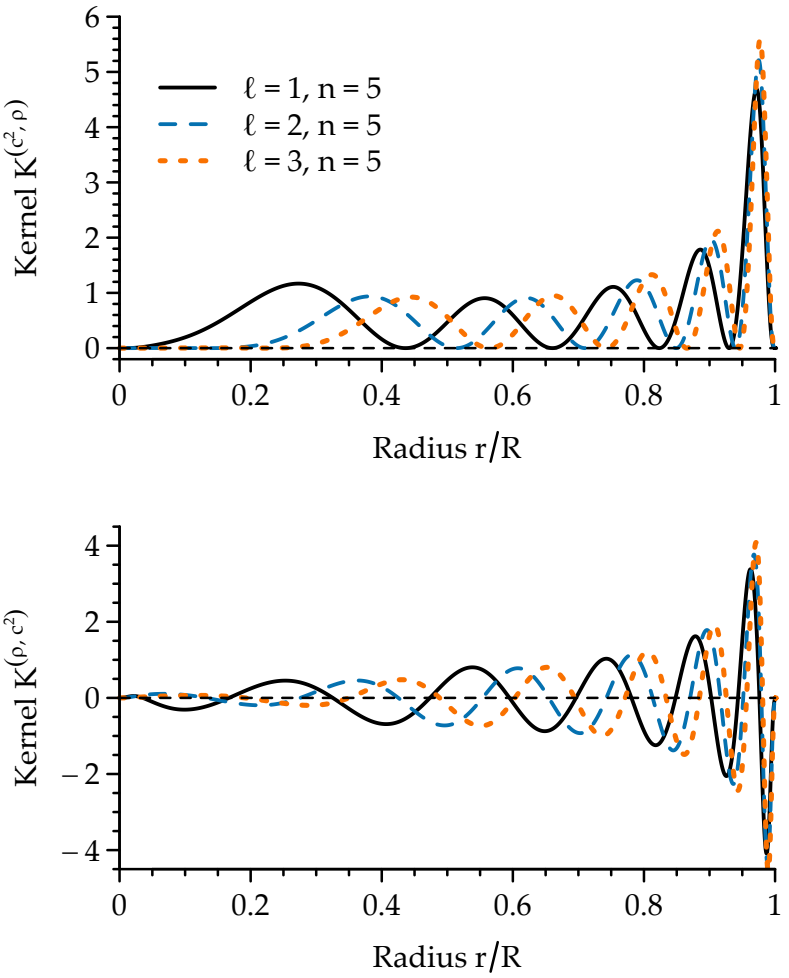


FIGURE 4.3. Kernels for the squared adiabatic sound speed and density, $K^{(c^2,\rho)}$ (top), and the reverse, $K^{(\rho,c^2)}$ (bottom), as a function of fractional radius for oscillation modes of model *GOE* of 16 Cyg A. Kernels are shown for modes with the same radial order n but different spherical degree ℓ (see the legend).

Basu 2003). Most of the time, these quantities are not known independently and need to be determined from the same set of data. Even where independent estimates are available, such as radii from interferometric measurements, the uncertainties are non-negligible. Both the amount of data and the precision to which the stellar mass and radius are known cause difficulties in inversion of asteroseismic data, and therefore the inversion methods need to be modified.

4.1.2 Asteroseismic Inversions

Even before *CoRoT* and *Kepler* detected oscillations in a large number of stars, there were a number of studies that investigated the possibility of inverting asteroseismic p-mode oscillations to determine the core structures of solar-like stars (Gough and Kosovichev 1993, Gough 1998, Berthomieu et al. 2001, Basu et al. 2001, 2002, Basu 2003). Additionally, there was at least one inconclusive study that tried to perform an inversion of seismic data from Procyon A (di Mauro 2004). The theoretical investigations of structure inversions all used mode sets and data uncertainties that were expected to be available from future missions to determine how well the structure differences between the cores of pairs of models could be determined. Unfortunately, the assumptions about the available mode sets and uncertainties were rather optimistic when compared with data available today.

Mode Set

The limited mode set available for stars other than the Sun makes the inversion problem more difficult. The fact that we cannot make resolved-disk observations of other stars generally restricts the detection of modes to $\ell \leqslant 3$. The lower turning points of these modes are within the stellar core; consequently, lacking more shallowly trapped modes, we will be unable to resolve the details of the stellar envelope. Figure 4.4 illustrates this difficulty by comparing the propagation cavities of oscillation modes with different degrees from a solar model. The figure shows lower turning points for low-degree Sun-as-a-star modes obtained by the Birmingham Solar Oscillation Network (BiSON; Davies et al. 2014a) and the $\ell > 3$ modes obtained by the Michaelson Doppler Imager (MDI) mission on board the Solar and Heliospheric Observatory (SOHO, Rhodes et al. 1997). The figure further shows the mode set that would be available if the Sun were a star in the *Kepler* field. Such a restricted mode set eliminates the possibility of using an inversion technique, such as RLS, that requires simultaneous determination of f_1 and f_2 over as large a part of the star as possible. Instead, we are confined to investigations of the stellar core.

Inversions using the OLA method or its variants are most suited for asteroseismic inversions, since OLA allows inversions over a small part of the star. Basu (2003) showed that instead of the (c^2, ρ) pair of variables used in solar inversions, the (u, Y) pair is better suited for asteroseismic structure inversions, where Y is the fractional helium abundance. This is because the kernels for Y

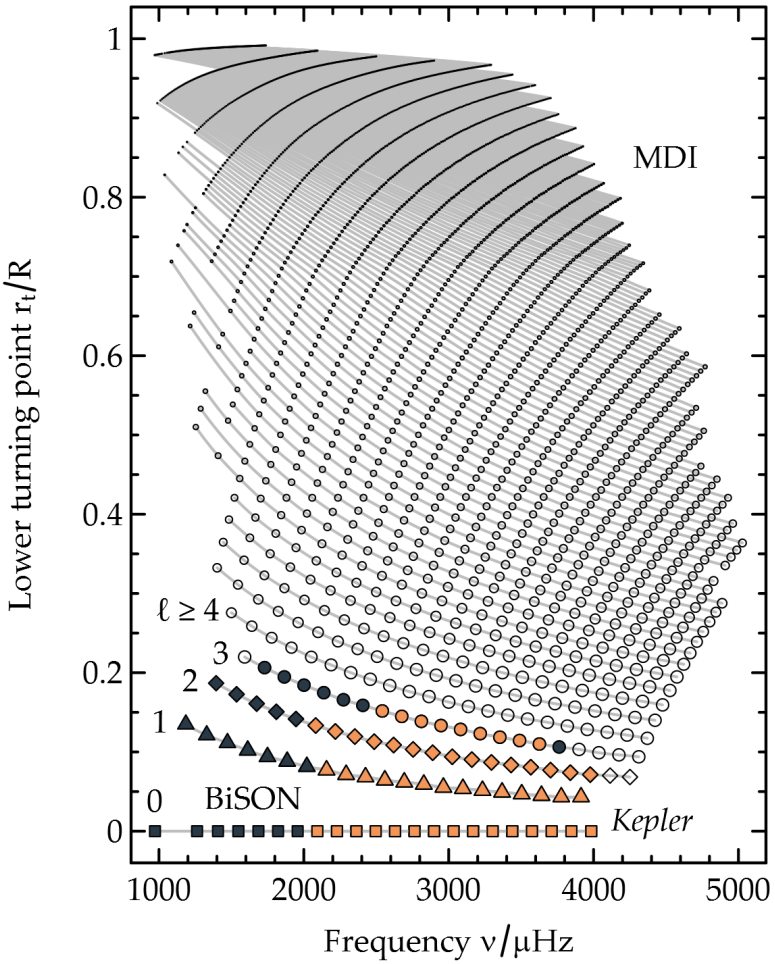


FIGURE 4.4. Lower turning points as a function of frequency for oscillation modes of a solar model with the MDI mode set (all points), BiSON mode set (all filled points) and the 16 Cyg A mode set from *Kepler* (orange filled points). Modes of the same spherical degree are connected by lines, with modes of spherical degree $\ell = 0$, 1, 2, and 3 shown with squares, triangles, diamonds, and circles, respectively. Compared to the Sun, asteroseismology of solar-like oscillators is restricted to low-degree, high-frequency modes.

are nonzero only in the helium ionization zone, as shown in Figure 4.5. Thus from the point of view of Equation (4.1) the data, i.e., the frequency differences, are almost completely determined by differences in u, thereby making u easier to determine. However, in order to derive the kernels for the (u, Y) pair, we have to assume that the EOS of the star is the same as that of the reference model (Dziembowski et al. 1990, Kosovichev 1999, Thompson and Christensen-Dalsgaard 2002). In other words, we are artificially adding information to the system. Basu and Christensen-Dalsgaard (1997) have shown that in the case of the Sun, this results in systematic errors in the inversion result; however, for other stars, we expect the errors caused by data uncertainties to be much larger than the systematic errors caused by an incorrect EOS. Thus, we proceed with this pair of variables.

Mass and Radius

The reduced precision of mass M and radius R estimates for stars other than the Sun also makes the problem more difficult. Frequencies scale as the square root of mean density, i.e., $\nu^2 \propto M/R^3$, so an unaccounted for difference in M and R between the star and the reference model gives rise to additional systematic errors in the inversion result. As these errors are proportional to the uncertainties in M and R, they are much larger than those expected from an incorrect EOS. Solar inversions as well as trial inversions for stellar models have hitherto been performed under the assumption that the mass and radius of the star are known. Having imprecise estimates of the stellar mass and radius means that the mass and radius of the reference model are likely to differ from those of the star. Berthomieu et al. (2001) accounted for this effect in their tests of asteroseismic inversions with pairs of models by adding terms for δM and δR to the inversion procedure. However, they assumed δM and δR to be known exactly, and the impact of uncertainties was not explored in that work.

Another difficulty arises from the fact that the inversion equation and the kernels are usually derived using dimensionless units, with the relative differences in f_1 and f_2 being calculated at constant fractional radii. This raises complications alluded to earlier: the u inversion result itself is also systematically offset by the differences in mass and radius (Basu 2003). In short, since kernels are derived using dimensionless variables, instead of a dimensional u, we actually have $u' \equiv P'/\rho'$, where $'$ denotes a dimensionless variable. It is straightforward to see from the equation governing conservation of mass that $\rho \propto M/R^3$. Likewise, from the equation of hydrostatic support one finds that $P \propto M^2/R^4$. Hence $u' = uR/M$, and so an inversion whose reference model has a different M or R will result in a u profile that differs by

$$\frac{\delta u'}{u'} - \frac{\delta u}{u} = \frac{\delta R}{R} - \frac{\delta M}{M}. \tag{4.2}$$

Thus the inversion procedure must be modified in order to accommodate the reduced precision of mass and radius estimates.

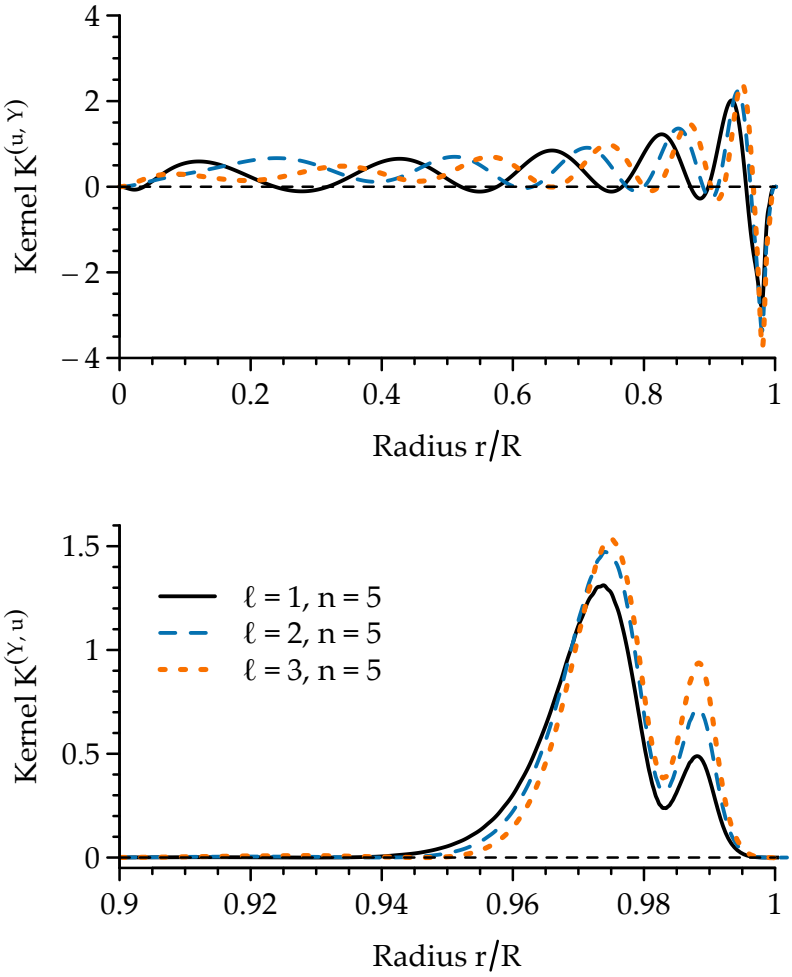


FIGURE 4.5. Kernels for the squared isothermal sound speed and helium abundance, $K^{(u,Y)}$ (top), and the reverse, $K^{(Y,u)}$ (bottom), as a function of fractional radius for oscillation modes of model *GOE* of 16 Cyg A. Notice that in contrast to the $K^{(\rho,c^2)}$ kernels shown in Figure 4.3, the $K^{(Y,u)}$ kernels have very small values ($0 < K(r) < 0.01$) in the interior $r < 0.9$ R.

These difficulties—limited mode sets and the uncertainties in stellar mass and radius estimates—have so far prevented structure inversions from widespread application in other stars. In this paper, we propose a way to circumvent the systematic error that results from the reference model having an incorrect mass and radius by extending the inversion procedure to use multiple reference models spanning the uncertainties in mass and radius. Furthermore, we introduce a new algorithm for the automated determination of inversion parameters. To put it concisely, this algorithm works by selecting the inversion parameters that maximize the agreement in the inversion result from different reference models. We apply this technique to the areas where the limited set of observed asteroseismic modes have resolving power, i.e., in the interior 30% of the star. We first demonstrate the efficacy of the algorithm by inverting the frequency differences between known models to determine that we are capable of producing the correct result. We then apply the method to the solar-type components of the 16 Cyg system with data obtained from the *Kepler* mission.

4.2 Methods

We seek to measure the difference in internal structure between stars and their best-fitting evolutionary models, which we assume to be sufficiently close in structure such that linear perturbation theory applies. We begin by explicitly expanding Equation (4.1) using the (u', Y) kernel pair. Given a set of $\mathscr{M}$ pulsation modes whose frequencies $\boldsymbol{\nu}$ have been measured, e.g.

$$\mathscr{M} = \{(\ell = 0, n = 10), (\ell = 1, n = 12), \ldots\}$$

for each mode of oscillation $i \in \mathscr{M}$ we have an equation relating a frequency perturbation to perturbations in stellar structure:

$$\frac{\delta\nu_i'}{\nu_i'} = \int K_i^{(u',Y)}(r) \cdot \frac{\delta u'}{u'}(r)\, dr + \int K_i^{(Y,u')}(r) \cdot \delta Y(r)\, dr + \frac{F_{\text{surf}}(\nu_i')}{\nu_i' \cdot I_i} + \epsilon_i. \tag{4.3}$$

Here $\delta\nu'$ is the difference in dimensionless oscillation mode frequency in the sense of (model - star), $\delta u'(r)$ is the difference in the dimensionless squared isothermal sound speed between a given stellar model and the star at fractional radius r, and $\delta Y(r)$ is the difference in the helium abundance. We assume the unknown differences between the true and the measured frequencies $\boldsymbol{\epsilon}$ to be independent and normally distributed with zero mean and known standard deviations $\boldsymbol{\sigma}$. The kernel functions $\mathbf{K}^{(u',Y)}$ and $\mathbf{K}^{(Y,u')}$ are known functions of the reference model and serve to relate changes in u' and Y to changes in oscillation mode frequencies. Finally, F_{surf} is a surface term that depends on frequency and is normalized by mode inertiae $\mathbf{I}$. Here we use the BG14-2 surface term, which Schmitt and Basu (2015) showed to be a good choice. This relation has

$$F_{\text{surf}}(\nu'; \nu_{\text{ac}}', \mathbf{a}) = a_1 \left(\frac{\nu'}{\nu_{\text{ac}}'}\right)^{-1} + a_2 \left(\frac{\nu'}{\nu_{\text{ac}}'}\right)^{3} \tag{4.4}$$

where $\mathbf{a}$ are coefficients that must be estimated during the inversion procedure and ν'_{ac} is the dimensionless acoustic frequency cut-off, which, under assumption of ideal gas, can be approximated by scaling from solar values with (Brown et al. 1991)

$$\nu'_{ac} = \nu_{ac,\odot} \cdot \frac{g}{g_\odot} \left(\frac{T_{eff}}{T_{eff,\odot}} \right)^{-1/2} \left(\frac{R^3}{GM} \right)^{1/2} \tag{4.5}$$

with g being the surface gravity of the reference model, T_{eff} its effective temperature, G the gravitational constant, and quantities subscripted with $\odot$ indicating the solar value. The next step is to invert Equation (4.3) to infer $\delta u'/u'(r)$, for which we will use the OLA technique.

4.2.1 Optimally Localized Averages

We invert Equation (4.3) using the OLA method. If, for the sake of argument, the (u', Y) kernel function of an oscillation mode were a δ function located at r_0 and zero elsewhere, and also if the (Y, u') kernel were zero everywhere, then a departure in frequency of this mode from the observed value would demand that $u'(r_0)$ differs between model and star. According to Equation (4.3), the relative difference in $u'(r_0)$ between the model and the star would be proportional to the relative difference in that mode's frequency. The OLA inversion technique works based on this concept.

OLA combines the kernels of the observed modes into an *averaging kernel* $\mathscr{K}$ resembling a localized function that is peaked at a chosen target radius inside the star. This is done via a linear combination of Equation (4.3) over the observed modes, where each mode $i \in \mathscr{M}$ is weighted by a coefficient c_i. If a vector of coefficients $\mathbf{c}$ exists such that an averaging kernel with the desired properties can be formed, the inversion result, i.e., the relative difference in u' between the model and the star, is then given by that same combination of the data. The process that creates the averaging kernel for u' also combines the kernels of Y to create a *cross-term kernel*, $\mathscr{C}$, and a reliable inversion result depends on $\mathscr{C}$ being as small as possible. Under these conditions, and assuming the surface term has been removed, the inversion result corresponds to an average of the underlying true difference weighted by the averaging kernel, i.e.,

$$\left\langle \frac{\delta u'}{u'} \right\rangle (r_0) = \int \mathscr{K}(r, r_0) \cdot \frac{\delta u'}{u'}(r)\, dr \tag{4.6}$$

assuming that $\int \mathscr{K}\, dr = 1$. Of course, the influence of data uncertainties must be controlled as well.

More formally, for a given target radius r_0, the OLA procedure aims to construct an averaging kernel $\mathscr{K}(r)$ that is well-localized around $r = r_0$. Recalling Equation (4.3), OLA proceeds by constructing a linear combination over all the

observed modes:

$$\begin{aligned}\sum_{i\in\mathscr{M}} c_i(r_0)\frac{\delta\nu_i'}{\nu_i'} = &\int \mathscr{K}(r;r_0,\mathbf{c})\cdot\frac{\delta u'}{u'}(r)\,dr\\ &+\int \mathscr{C}(r;r_0,\mathbf{c})\cdot\delta Y(r)\,dr\\ &+\sum_{i\in\mathscr{M}} c_i(r_0)\cdot F_{surf}(\nu_i';\nu_{ac}',\mathbf{a})/\left(\nu_i'\cdot I_i\right)\\ &+\sum_{i\in\mathscr{M}} c_i(r_0)\cdot\epsilon_i \end{aligned} \tag{4.7}$$

where the vector $\mathbf{c}$ are inversion coefficients that will need to be determined for each given r_0 and

$$\mathscr{K}(r;r_0,\mathbf{c}) = \sum_{i\in\mathscr{M}} c_i(r_0)\cdot K_i^{(u,Y)}(r) \tag{4.8}$$

$$\mathscr{C}(r;r_0,\mathbf{c}) = \sum_{i\in\mathscr{M}} c_i(r_0)\cdot K_i^{(Y,u)}(r) \tag{4.9}$$

subject to the constraint that

$$\int \mathscr{K}(r;r_0)\,dr = 1. \tag{4.10}$$

Provided that the averaging kernel is well-localized at the target radius and the cross-term kernel, the surface-term contributions, and the combined data uncertainties are all small; this combination of relative frequency differences gives a localized average of $\delta u'/u'$ at the target radius r_0:

$$\left\langle\frac{\delta u'}{u'}\right\rangle(r_0) = \sum_{i\in\mathscr{M}}\left(c_i(r_0)\cdot\frac{\delta\nu_i'}{\nu_i'}\right). \tag{4.11}$$

Here we have chosen to express relative differences in the sense

$$\frac{\delta q}{q} = \frac{(\text{model}-\text{star})}{\text{model}} = \frac{(q_{ref}-q_{star})}{q_{ref}} \tag{4.12}$$

where q can refer to any quantity. Thus, Equation (4.11) can be redimensionalized using Equation (4.2) to infer u_{star} with

$$u_{star}(r) = \left(1-\frac{\delta u'}{u'}(r)+\frac{\delta R}{R}-\frac{\delta M}{M}\right)\cdot u_{ref}(r). \tag{4.13}$$

We now turn our attention to determining the coefficients $\mathbf{c}$ that make this estimate possible.

4.2.2 Inversion Coefficients Using Subtractive OLA

The optimal inversion coefficients $\hat{\mathbf{c}}$ must strike a balance between forming a well-localized averaging kernel and forming a small cross-term kernel, while still having small uncertainty. In Subtractive OLA (SOLA, Pijpers and Thompson 1992, 1994), the averaging kernel is formed according to a specified well-localized form (the "target kernel"), and the coefficients $\mathbf{c}$ are determined by minimizing the difference between the averaging kernel obtained and the target kernel. This is a fast implementation of the OLA method. It comes at the price of a free parameter in the form of the properties of the target kernel. SOLA determines optimal coefficients $\hat{\mathbf{c}}$ for a given target radius r_0 by solving the optimization problem

$$\hat{\mathbf{c}}(r_0;\beta,\mu,\Delta) = \arg\min_{\mathbf{c}} \left\{ \mathscr{F}(\mathbf{c};r_0,\Delta) + \beta \int \mathscr{C}(r;r_0,\mathbf{c})^2 \,\mathrm{d}r + \mu \sum_{i\in\mathscr{M}} \left(c_i^2 \cdot \sigma_i^2 \right) \right\}$$
$$\text{subject to } \int \mathscr{K}(r;r_0,\mathbf{c}) \,\mathrm{d}r = 1 \quad \text{and} \quad \sum_{i\in\mathscr{M}} c_i \cdot \frac{F_{\text{surf}}(\nu_i';\nu_{\text{ac}})}{\nu_i' \cdot I_i} = 0. \quad (4.14)$$

Here β and μ are parameters that must be chosen to penalize the amplitude of the cross-term kernel and the effect of data uncertainties, respectively. A third parameter, Δ, gives the width of the target kernel(s). The function $\mathscr{F}$ penalizes deviations of the averaging kernel from the target kernel T and can be calculated as

$$\mathscr{F}(\mathbf{c};r_0,\Delta) = \int \left[\mathscr{K}(r;r_0,\mathbf{c}) - T(r;r_0,\Delta)\right]^2 \,\mathrm{d}r. \quad (4.15)$$

The functional form of T can be chosen, e.g. as a modified Gaussian that decays to zero at $r = 0$ but remains peaked at $r = r_0$ (e.g. Rabello-Soares et al. 1999) with

$$T(r;r_0,\Delta) = A \cdot r \cdot \exp\left\{-\mathcal{G}(r;r_0,\Delta)^2\right\} \quad (4.16)$$

$$\mathcal{G}(r;r_0,\Delta) = \frac{r - r_0}{D(r_0,\Delta)} + \frac{D(r_0,\Delta)}{2r_0}. \quad (4.17)$$

The normalization factor A is chosen to ensure $\int T \,\mathrm{d}r = 1$. Since the resolution ultimately depends on the internal sound speed c_s (Thompson 1993), the function D gives the width of the kernels according to variations in c_s and a free parameter Δ that describes a fiducial width as

$$D(r_0,\Delta) = \Delta \cdot \frac{c_s(r_0)}{c_s(r_f)} \quad (4.18)$$

with r_f being an arbitrary reference point (e.g. we choose $r_f = 0.2$, although the result is rather insensitive to the choice). We note that other choices of $\mathscr{F}$, T, $\mathcal{G}$, and D are possible (see, e.g., Gough 1985, Brown et al. 1989), but they will not be explored here.

The SOLA inversion problem can be cast into a system of linear equations with the constraints enforced using Lagrange multipliers. Given choices of β, μ,

and Δ, Equation (4.14) can be solved via matrix inversion, the details of which can be found, for example, in Chapter 10 of Basu and Chaplin 2017. See Rabello-Soares et al. (1999) for a description of how inversion parameters are usually selected in helioseismology. Depending on the data that are available, it may be possible to form zero, one, or more well-localized averaging kernels with correspondingly small cross-term kernels and well-controlled uncertainties at different locations in the stellar interior.

4.2.3 Selecting Inversion Parameters with Multiple Reference Models ("Inversions for Agreement")

It is not clear *a priori* which inversion parameters should be chosen, nor is there a reliable algorithm for their selection. Here we propose an algorithm for selecting inversion parameters based on the following information. First, besides the effects that stem from differences in M and R, inversion results do not otherwise depend on the choice of reference model: with proper selection of inversion parameters, a wide range of reference models are capable of producing the correct inference (Basu et al. 2000). Furthermore, for a given mode set, and setting aside the surface term, the values of the mode frequencies themselves do not play a role in determining the averaging and cross-term kernels. Thus, provided the differences in the kernels between models are small, the same inversion parameters can be used for different models. Instead of performing single-model inversions, we invert using an array of reference models that span the uncertainties in M and R. We simultaneously estimate the inversion parameters and the stellar M and R such that the inferred stellar u profile from the different models are in agreement. We achieve this via repeated iterative optimization with random noise realizations. We constrain M and R with normal priors based on past studies, and set uniform priors on the inversion parameters. We have also tried this procedure with each reference model having its own individual set of inversion parameters (β, μ, Δ) to optimize, and we found that it did not have a substantial impact on the results.

We generate an array of nine reference models that are calibrated to span the 1σ uncertainties in mass and radius for each star whose interior structure we seek to infer. We optimize a vector of five inversion parameters $\boldsymbol{\alpha} = (\beta, \mu, \Delta, M_{\rm star}, R_{\rm star})$ which are shared among the nine models. We take an average among their inferred values of $u_{\rm star}$, and finally we choose the $\boldsymbol{\alpha}$ that minimizes the variance of this average, weighted by the priors on $M_{\rm star}$ and $R_{\rm star}$. Formally, we postulate that the optimal inversion parameters $\hat{\boldsymbol{\alpha}}$ across all of the reference models is

$$\hat{\boldsymbol{\alpha}} = \arg\min_{\boldsymbol{\alpha}} \left\{ \sum_{r_j \in \mathbf{r}_0} \log \mathrm{Var}\left[\tilde{u}\left(r_j; \boldsymbol{\alpha}\right)\right] - \log \Psi(\boldsymbol{\alpha}) \right\} \tag{4.19}$$

where Var is the variance operator, $\mathbf{r}_0$ are the target radii, and $\tilde{u}$ is a vector whose kth element $u_k(r_0; \boldsymbol{\alpha})$ gives the inferred value of $u_{\rm star}$ at target radius r_0 via the

kth reference model using the inversion parameters $\boldsymbol{\alpha}$ (*cf.* Equations. 4.11-4.14). Finally, Ψ is the prior distribution, which in this case has

$$\Psi(\boldsymbol{\alpha}) = \psi\left(M_{\text{star}}; \mu_M, \sigma_M^2\right) \cdot \psi\left(R_{\text{star}}; \mu_R, \sigma_R^2\right) \tag{4.20}$$

with ψ being the normal density function and μ_x and σ_x being the mean and standard deviation of x. In each iteration of the algorithm, each of the non- and redimensionalizations are performed with the current estimate of M_{star} and R_{star}. For example,

$$\frac{\delta\nu'}{\nu'} = \left[\left(\frac{R_{\text{ref}}^{3/2}}{M_{\text{ref}}^{1/2}}\right)\nu_{\text{ref}} - \left(\frac{R_{\text{star}}^{3/2}}{M_{\text{star}}^{1/2}}\right)\nu_{\text{star}}\right] \Big/ \left[\left(\frac{R_{\text{ref}}^{3/2}}{M_{\text{ref}}^{1/2}}\right)\nu_{\text{ref}}\right]. \tag{4.21}$$

In summary, Equation (4.19) says that the optimal inversion parameters are the ones that give the same inference of u_{star} across all the reference models.

Since the inversion results depend on uncertain measurements, we perform repeated trials with random realizations of noise. Specifically, in each trial, we perturb each frequency ν with normal noise according its uncertainty σ_ν, and the mass and radius estimates μ_M and μ_R via their uncertainties σ_M and σ_R. We then use the Nelder–Mead (1965) downhill simplex method to numerically search for the parameters that satisfy Equation (4.19) for that realization of noise. Because each inversion parameter is strictly non-negative and can potentially take on a large range of values, we optimize $\log \boldsymbol{\alpha}$. We stop each trial after either the relative change in the objective function is reduced by less than the square root of the machine precision for double precision floating point numbers ($\sim 10^{-8}$), or a maximum number of 512 iterations is reached. In the majority of cases, the former condition is met. We perform 128 trials and report the averaged results. Finally, we visually inspect the resulting averaging kernels and cross-term kernels to ensure that the averaging kernels are well-localized at the target radii and that the cross-term kernels have small amplitude everywhere.

4.3 Results

4.3.1 Tests on Models

In order to validate our technique, we first apply the method to known models; this allows us to check that the procedure does indeed produce the correct result. Specifically, we determine whether or not we can accurately recover the internal u profiles of the *GOE* models of 16 Cyg A and B using an array of different reference models as reference.

For the test, we generate an array of reference models for each star by calibrating models to their estimated masses ($\pm 1\sigma$, Bellinger et al. 2016), radii ($\pm 1\sigma$, White et al. 2013), ages (Bellinger et al. 2016), luminosities (White et al. 2013), and metallicities (Ramírez et al. 2009). The estimates we use for these stars are given in Table 4.1. We calculate the models using the given mean values of their

ages, luminosities, and metallicities. We construct the models using the MESA stellar evolution code (*Modules for Experiments in Stellar Astrophysics*, Paxton et al. 2011). For each model, we use ADIPLS (*the Aarhus adiabatic oscillation package*, Christensen-Dalsgaard 2008) to calculate the adiabatic oscillation mode frequencies corresponding to the 54 and 56 oscillation modes that have been identified in 16 Cyg A and B, respectively. We use the same treatments of evolution and pulsation that are described in Section 2.1 of Bellinger et al. 2016. None of the reference models have exactly the same mass or radius as the two *GOE* models that we are treating as our proxy stars. We perturb the proxy star frequencies with noise prior to beginning the procedure.

We apply the inversion-for-agreement procedure described in Section 4.2.3. The results are shown in Figure 4.6. The procedure gets the correct result. The uncertainties in $\delta u/u$ are given by the average over the 128 trials. The "uncertainties" in fractional radius r/R are a measure of the resolution of the inversion and are given by the width at half maximum of an average over the averaging kernels of the different trials. The averaging kernels are reasonably well-localized and the cross-term kernels are small everywhere. The averaging kernels placed at $r_0 = 0.3$ begin to develop some amplitude outside of the target region; this is why we do not attempt to probe shallower layers.

4.3.2 Inversions for Stellar Structure

We now apply our structure inversion-for-agreement procedure on asteroseismic data of 16 Cyg A and B. The relative differences with respect to the *GOE* evolutionary models of these stars are shown in Figure 4.7. As the mode sets are the same as in our tests with models, the averaging kernels and cross-term kernels are nearly identical to those shown in Figure 4.6. The results are also tabulated in Tables 4.2 and 4.3. We find that the sound speeds throughout the cores of 16 Cyg A and B exceed those of these evolutionary models.

In the case of 16 Cyg A, each of the individual measurements hovers around a 1σ difference. On the one hand, all of the model sound speeds are found to be lower than in the star, indicating that there are systematic differences between the model and the star. Viewed this way, the overall result is more significant than each of the measurements taken separately. On the other hand, there is covariance between the different measurements, because the different averaging kernels overlap to some degree. Thus, assigning an overall level of statistical significance to these results is challenging.

TABLE 4.1. Fundamental parameters of 16 Cyg A and B.

Name	Mass	Radius	Age	Luminosity	Metallicity
	$\mu_M \pm \sigma_M$	$\mu_R \pm \sigma_R$	τ	L	[Fe/H]
	$[M_\odot]$	$[R_\odot]$	[Gyr]	$[L_\odot]$	(dex)
16 Cyg A	1.080 ± 0.016	1.22 ± 0.02	6.90 ± 0.40	1.56 ± 0.05	0.096 ± 0.026
16 Cyg B	1.030 ± 0.015	1.12 ± 0.02	6.80 ± 0.28	1.27 ± 0.04	0.052 ± 0.021

FIGURE 4.5. Structural inversions for the internal squared isothermal sound-speed profile u of evolutionary models of 16 Cyg A (left) and 16 Cyg B (right). **Top**: actual relative difference $\delta u/u$ between the evolutionary model and a reference model from the corresponding array of reference models for that star (dashed gray line), and the result of the inversion-for-agreement procedure presented here (colored points). The colors serve to associate the inversion results with their respective averaging and cross-term kernels. **Middle**: averaged averaging kernels, sensitive to changes in u', which have been placed at target radii $r_0 = [0.05, 0.1, 0.15, 0.2, 0.25, 0.3]$. **Bottom**: averaged cross-term kernels that are sensitive to changes in helium abundance, whose amplitudes should be small everywhere relative to the averaging kernels. **Insets**: the behavior of the averaging and cross-term kernels closer to the surface, where their amplitudes are small as desired (note the change in axes).

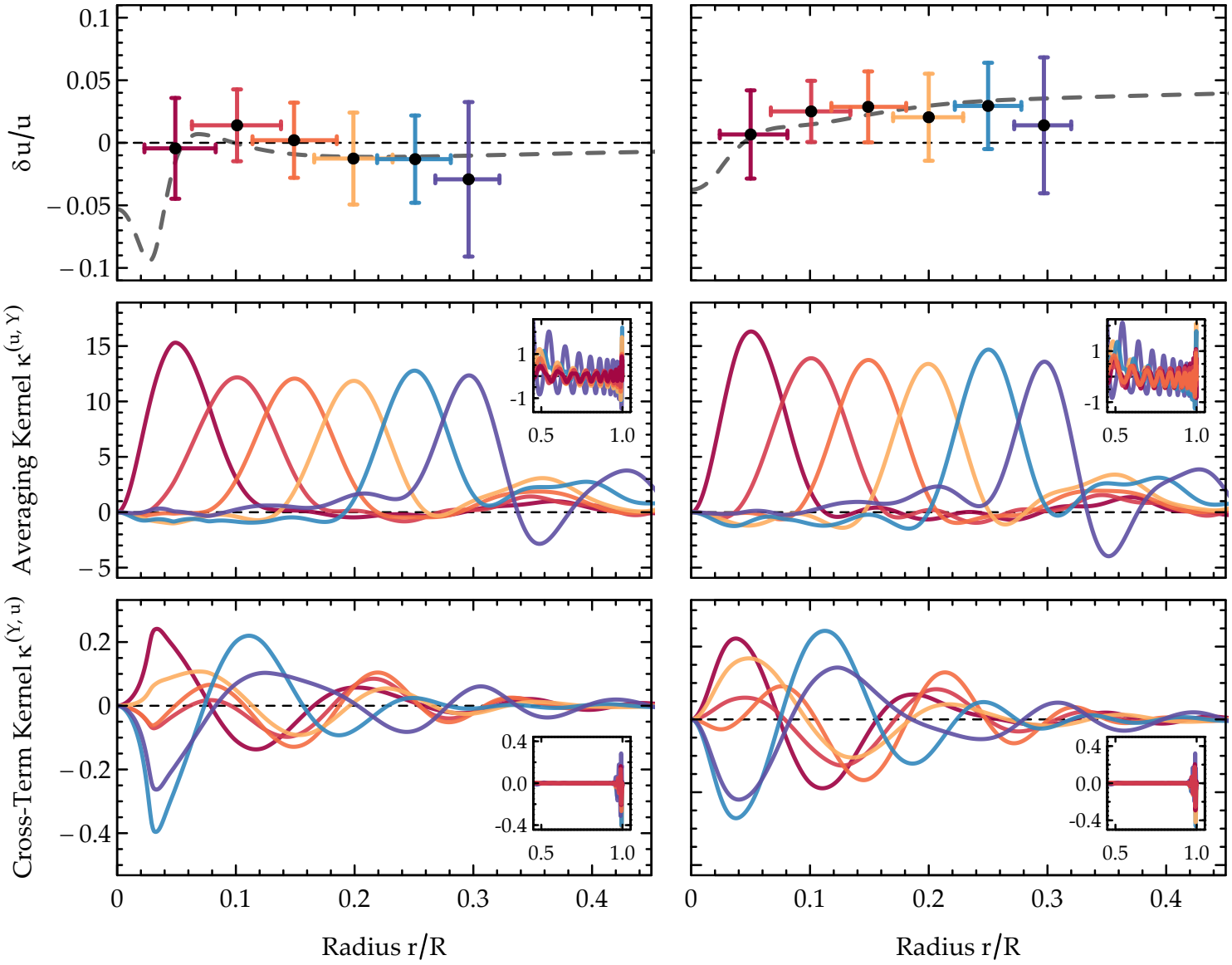


FIGURE 4.6. (Caption on other page.)

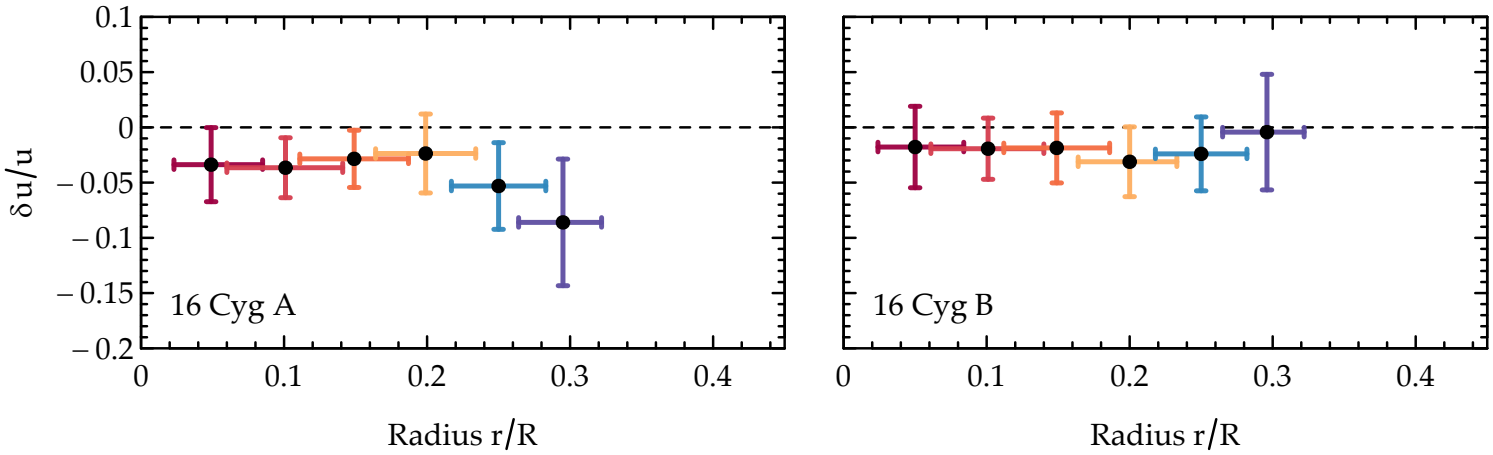


FIGURE 4.7. Structural inversions for the internal squared isothermal sound-speed profile u of 16 Cyg A (left) and 16 Cyg B (right) using the inversion-for-agreement technique introduced in this paper. Results are shown in terms of relative differences with respect to the *GOE* evolutionary models of these stars (*cf.* Equation 4.12). The sound speeds in the cores of 16 Cyg A and B are greater than those of the evolutionary models.

To assess whether the differences may stem from the *GOE* models having wrong masses or radii, we compare the inversion results against other models of different mass and radius. Following Equation (4.2), the spread in sound speeds caused by mass and radius estimates are largest for the models with either a high radius and a low mass, or models with a low radius and high mass. Thus, we show in Figure 4.8 these inversion results against models with masses and radii that differ by 1σ in opposite directions from the mean estimated masses and radii of these stars. In both cases, the models with higher masses and lower radii are preferred. However, while the 16 Cyg B models show roughly broad agreement, the 16 Cyg A models do not agree quite as well.

The isothermal speed of sound depends principally on the inverse of the mean molecular weight μ of the fluid. Fusion alters the core composition and increases μ; thus, with all else equal, older stars will have a lower u in the core. To assess the effect of stellar age in the context of these results, we evolve two models to match the characteristics of 16 Cyg A (*cf.* Table 4.1) with ages of

TABLE 4.2. Results of Inversions for the Squared Isothermal Sound Speed u inside of 16 Cyg A at Different Target Radii r_0 in the Stellar Core

Target radius	Peak of $\mathscr{K}$	Relative u difference	Sq. iso. sound speed
r_0	$r_{max} \pm$ FWHM	$\delta u/u \pm \sigma_\delta$	$u \pm \sigma_u$
[R]	[R]	[w.r.t. model *GOE*]	[10^{15} cm^2 s^{-2}]
0.05	0.049 ± 0.031	-0.033 ± 0.033	1.515 ± 0.049
0.10	0.101 ± 0.041	-0.036 ± 0.027	1.580 ± 0.041
0.15	0.149 ± 0.038	-0.028 ± 0.025	1.404 ± 0.035
0.20	0.199 ± 0.035	-0.023 ± 0.035	1.181 ± 0.041
0.25	0.250 ± 0.033	-0.053 ± 0.039	1.019 ± 0.037
0.30	0.295 ± 0.029	-0.086 ± 0.057	0.910 ± 0.048

TABLE 4.3. Results of inversions for the squared isothermal sound speed u inside of 16 Cyg B.

Target radius	Peak of $\mathscr{K}$	Relative u difference	Sq. iso. sound speed
r_0	$r_{max} \pm$ FWHM	$\delta u/u \pm \sigma_\delta$	$u \pm \sigma_u$
[R]	[R]	[w.r.t. model *GOE*]	[10^{15} cm^2 s^{-2}]
0.05	0.050 ± 0.030	-0.017 ± 0.036	1.485 ± 0.053
0.10	0.101 ± 0.039	-0.019 ± 0.027	1.533 ± 0.041
0.15	0.149 ± 0.037	-0.018 ± 0.031	1.402 ± 0.043
0.20	0.200 ± 0.034	-0.031 ± 0.031	1.216 ± 0.037
0.25	0.250 ± 0.032	-0.024 ± 0.033	1.025 ± 0.033
0.30	0.296 ± 0.028	-0.004 ± 0.052	0.870 ± 0.045

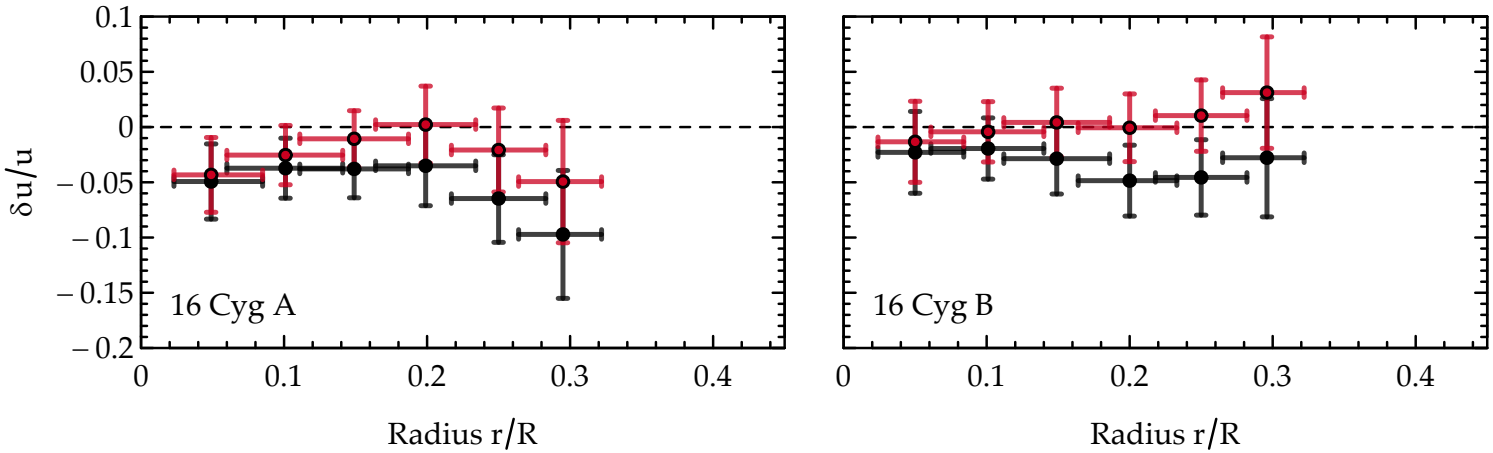


FIGURE 4.8. Relative differences between the inferred sound speeds u of 16 Cyg A (left) and 16 Cyg B (right) shown against a model with a high radius and low mass (black points) and a model with a low radius and high mass (red points).

$\tau = 6$ Gyr and $\tau = 5$ Gyr, which are significantly lower than the estimated age of $\tau = 6.90 \pm 0.40$ Gyr. The relative differences between the core u of 16 Cyg A and these models are shown in Figure 4.9. In the deep core ($r = 0.05$), the young age models have smaller differences when compared with the *GOE* model. However, the differences farther out are not explained with smaller ages. Furthermore, although it seems the inner core is better with the low-age models, frequency combinations such as $r_{0,2}$ (Roxburgh and Vorontsov 2003) indicate that the low-age models are not appropriate. A comparison of $r_{0,2}$ values for these models is shown in Figure 4.10. This may explain why the differences in u worsen just outside the core.

4.4 Discussion and Conclusions

In this paper, we examined the problem of deducing the core structures of solar-like stars based on the frequencies of their normal modes of oscillation. We applied the SOLA inversion technique to infer the radial dependence of the squared isothermal sound speed throughout the interiors of two solar-type main-sequence stars. We inverted using the (u', Y) kernel pair because the influence of the second variable (Y) is very low in the regions of our interest. We presented a new algorithm for the automated determination of inversion parameters that also accounts for imprecise/inaccurate stellar mass and radius estimates. We validated this technique on models, and then applied it to the well-studied stars 16 Cyg A and B. We measured u at several different radii within these stars and compared these values to best-fitting evolutionary models of these stars. We found that the sound speeds in the cores of these stars are greater than in the *GOE* models. This is to our knowledge the first time the radial variation in sound speed has been measured in a star other than the Sun.

In the case of 16 Cyg B, it seems plausible that adjustments to the mass and radius of the *GOE* model may serve to fix the differences that we find. In the case of 16 Cyg A, however, the source of the disparities is more difficult to pinpoint. Lower age models help with the differences in the deeper parts of the core, but do not aid with the differences farther out. Furthermore, the lower age models fail to reproduce the asteroseismic frequency ratios of 16 Cyg A, which effectively rules age out as the culprit. Missing physical processes, incorrect application of known processes, or inadequate inputs in the calculations of the models may therefore be at fault. For example, while the GOE model of 16 Cyg A does not have a convective core at the present age, it did have one during the first 1.75 Gyr of its evolution. As core convection modifies the mean molecular weight, the duration of its existence may leave a footprint in the sound speed. It may then be the case that an incorrect prescription of convection in stellar cores is the cause of these discrepancies.

16 Cyg A and B are stars either on the main sequence or nearly at the main-sequence turnoff. The main sequence is a well-studied phase of evolution, and the different types of observations that are possible for main-sequence stars lead

to estimates of their ages, masses, and radii in a well-known way. Being the first and also the longest-lived stage of evolution, getting the details of the main-sequence evolution right is necessary for also getting the later stages of stellar evolution right as well. Any neglected processes that cause substantial errors on

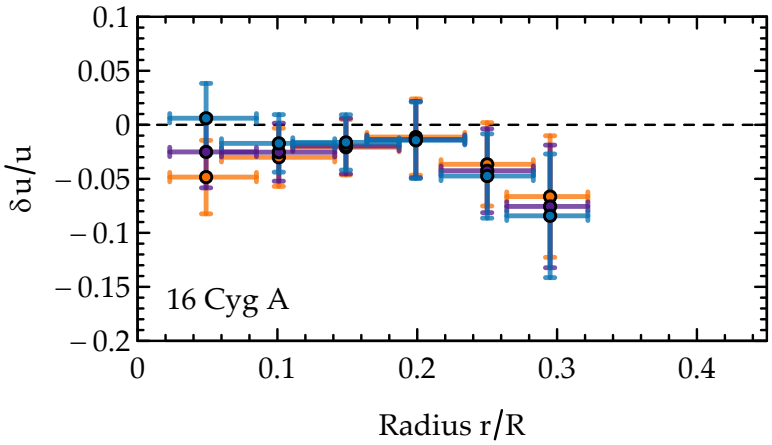


FIGURE 4.9. Relative differences between the isothermal sound speed u in the core of 16 Cyg A and models with lower ages (6 Gyr in purple, 5 Gyr in blue). A model of 16 Cyg A at the mean estimated present age (6.9 Gyr in orange) is shown for reference.

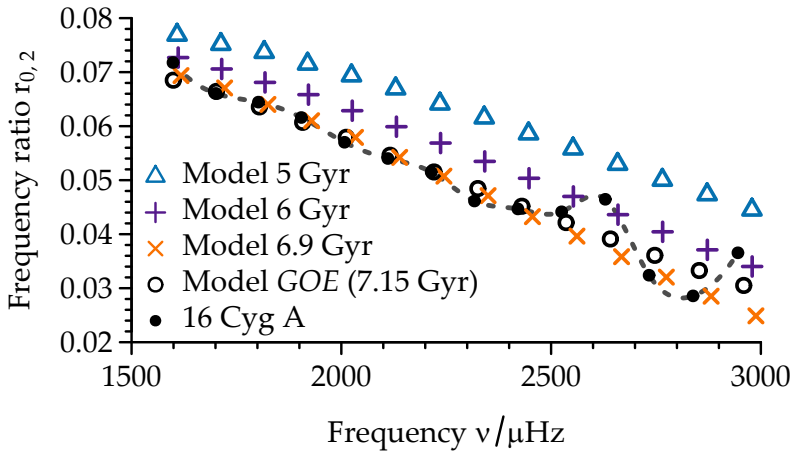


FIGURE 4.10. Ratio of the small frequency separation to the large frequency separation—a core-conditions indicator that is insensitive to surface effects—against mode frequency for asteroseismic data of 16 Cyg A in comparison with models of various ages.

the core structure of main-sequence stars will subsequently propagate into the later stages of evolution.

As is always the case with ill-posed inverse problems, there is no guarantee that the end result will be the true profile of the star. That being said, the procedure has worked well in blind tests on models with known structure. Therefore, some confidence can be put in the results.

Acknowledgements

The research leading to the presented results has received funding from the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013) / ERC grant agreement no 338251 (StellarAges). This research was undertaken in the context of the International Max Planck Research School for Solar System Research. E.P.B. acknowledges support from the National Physical Science Consortium Fellowship. S.B. acknowledges partial support from NSF grant AST-1514676 and NASA grant NNX13AE70G. We thank the anonymous referee for their very helpful report.

Software

R 3.2.3 (R Core Team 2014), magicaxis 2.0.0 (Robotham 2016), kernel calculations by Thompson (2000), MESA (Paxton et al. 2011, 2013, 2015), and ADIPLS (Christensen-Dalsgaard 2008).

Future Prospects

Though stars are, overall, generally considered to be well-understood, a number of open problems remain in asteroseismology and, more widely, the field of stellar astrophysics as a whole. At a basic level, we currently are unable to predict stellar radii from first principles. This is due to the fact that we use time-independent one-dimensional theories of convection in evolutionary models—approximations which are controlled by free parameters. Properly modelling convection in stellar interiors seems to be among the biggest goals in modern theoretical stellar astrophysics. Furthermore, for similar reasons, we generally fail to predict pulsation frequencies of stars, even after making post-hoc corrections for near-surface effects.

Along similar lines, one of the most basic facts about stars (and astronomical bodies in general) is that they rotate. Yet canonical stellar modelling often neglects the effects of rotation, and other similarly 'obvious' phenomena such as magnetic fields. The very long-term future of research into stars may feature fully 3D magnetohydrodynamical stellar modelling, or even a full treatment of every individual particle that make up the star; however, it is clear that we are far away from that point.

In terms of the continuation of the research presented in this thesis, there are a few avenues in particular that I intend to explore in the coming months and years:

Structure inversions of more stars. The next step is to apply the technique developed in Chapter 4 to as many stars as possible. This will allow us to determine whether the theory of stellar evolution produces models with the correct interior structures.

Figures 5.1 and 5.2 show structure inversions for 20 stars from the *Kepler* LEGACY sample (Lund et al. 2017). The reference models have been constructed under four different assumptions of input physics: with/without diffusion, and with/without convective core overshooting. While some stars show broad agreement throughout their interior with evolutionary models (e.g., KIC 5184732), most of the models disagree substantially with the interior structure of the stars. Furthermore, there seems to be no set of input physics considered here that repairs the differences. This indicates that important ingredients may be missing from canonical models of stellar interiors, such as mixing induced by internal rotation.

This work is soon to be submitted to the Astrophysical Journal.

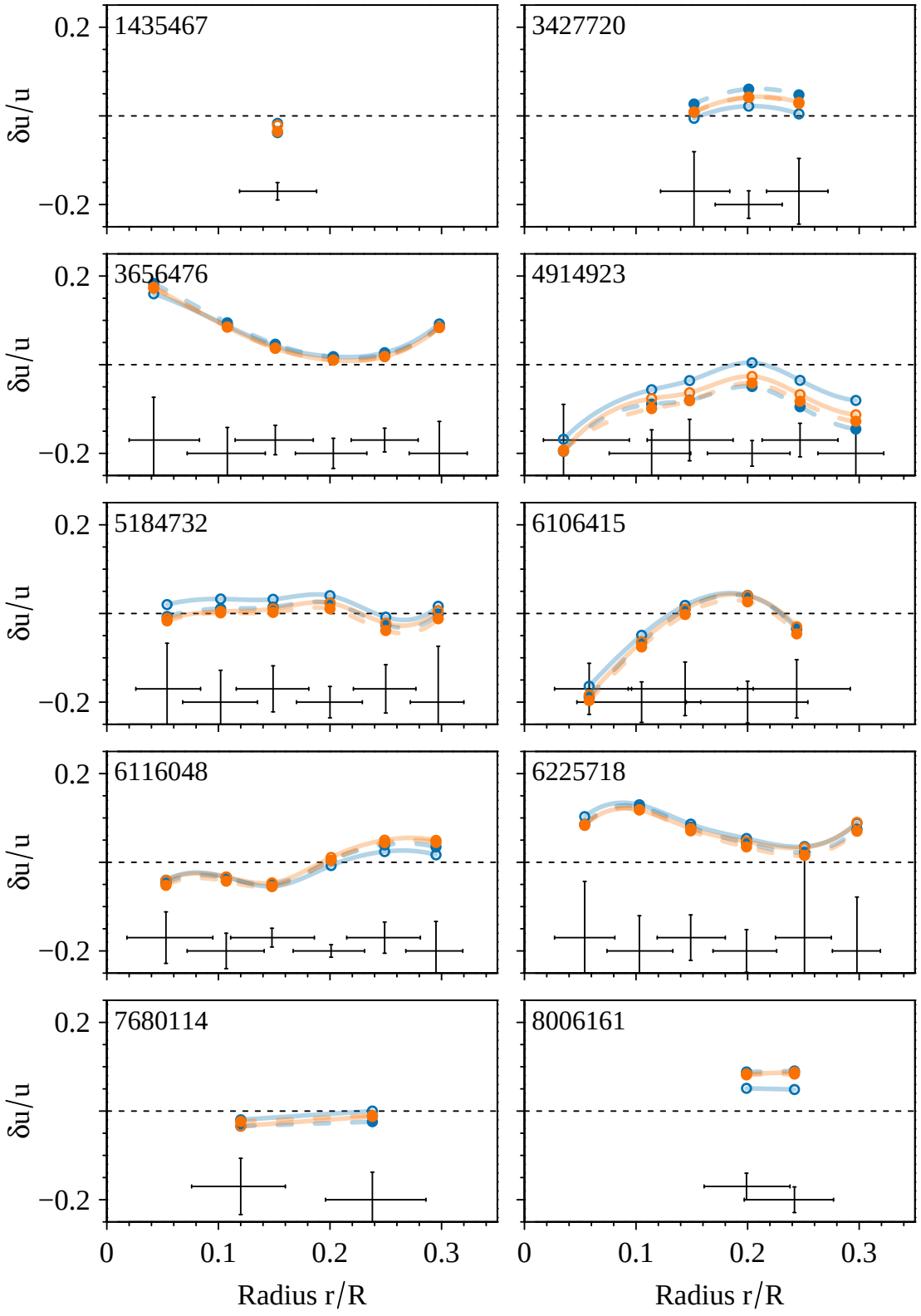


FIGURE 5.1. (Continued in Figure 5.2.)

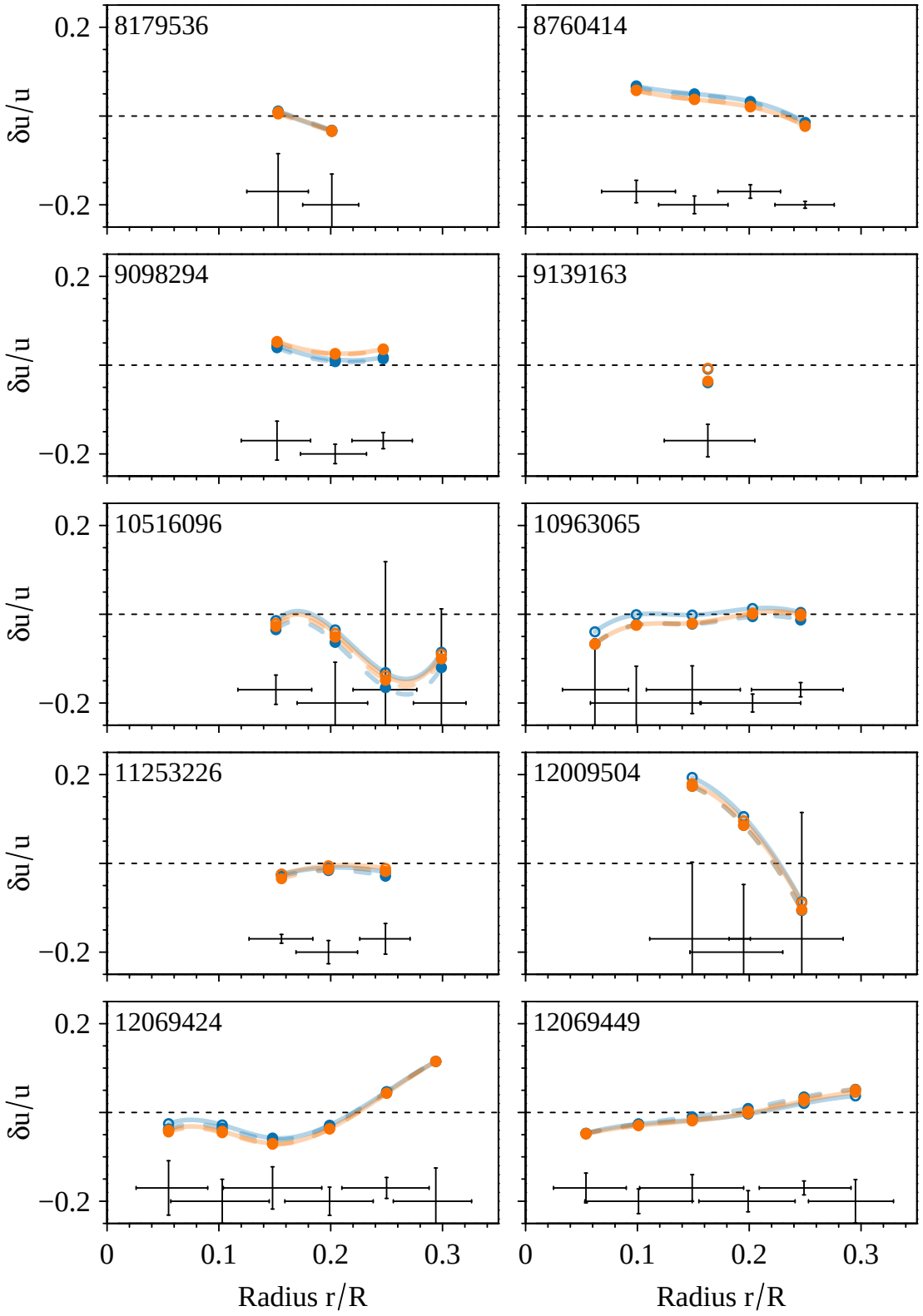


FIGURE 5.2. (Caption on other page.)

FIGURE 5.2. Core sound-speed profiles of LEGACY stars compared against stellar models constructed with different physics inputs: with/without diffusion (orange/blue, respectively) and with/without overshooting (filled/open points, respectively). The quantity $\delta u/u$ is the relative difference in the isothermal speed of sound between the model and the star at that location in the stellar interior. The uncertainties of the inversion results and the widths of the corresponding averaging kernels are shown as error bars in the bottom of each panel, and are vertically offset from one another for visibility.

Evolution inversions of evolved stars. There have been at least an order of magnitude more detections of solar-like oscillations in evolved stars such as red giants than in main-sequence stars. When combined with kinematic information, determining the ages and chemical compositions of a large number of red giant stars will allow us to reconstruct the history of the Galaxy's development.

In Chapter 1 I showed the future evolution of the Sun up through to core helium exhaustion. Current ongoing work is the application of the techniques developed in Chapters 2 and 3 to these later stages of evolution.

Structure inversions of evolved stars. In this thesis, I analyzed main-sequence solar-like oscillators. After stars leave the main sequence, the p-modes in their envelopes mix with the g-modes in their deep interiors to give rise to mixed modes of oscillation. Figure 5.3 shows the evolution of the kernel function for an $\ell = 1$ mixed mode throughout the sub-giant phase of evolution. After obtaining suitable reference models, for example using the technique mentioned in the previous point, I will invert mixed mode frequencies to determine the core structures of sub-giant and eventually red-giant stars. This presents the exciting prospect for potentially learning more about the deep core structure of another star than we know about our own Sun.

Evolution inversions for fundamental constants. A problem of cosmological significance is the measurement of physical constants, and the determination of whether or not they really are constant. The idea of using the Sun to constrain the cosmic variation of the gravitational constant G goes back at least to the time of Dirac (1938). So far, this approach has not been undertaken using other stars. I intend to use the tools discussed in this thesis to measure G as well as other fundamental quantities that impact on stellar evolution and pulsation, such as the fine structure constant (e.g., Adams 2008, Coc et al. 2010). Though the Sun is the star with the best data, observations of a large number stars may be able to be combined into a more sensitive tool for these measurements. Furthermore, the Sun's evolution only covers one third of the history of the Universe, and is therefore insensitive to any earlier variations to these quantities.

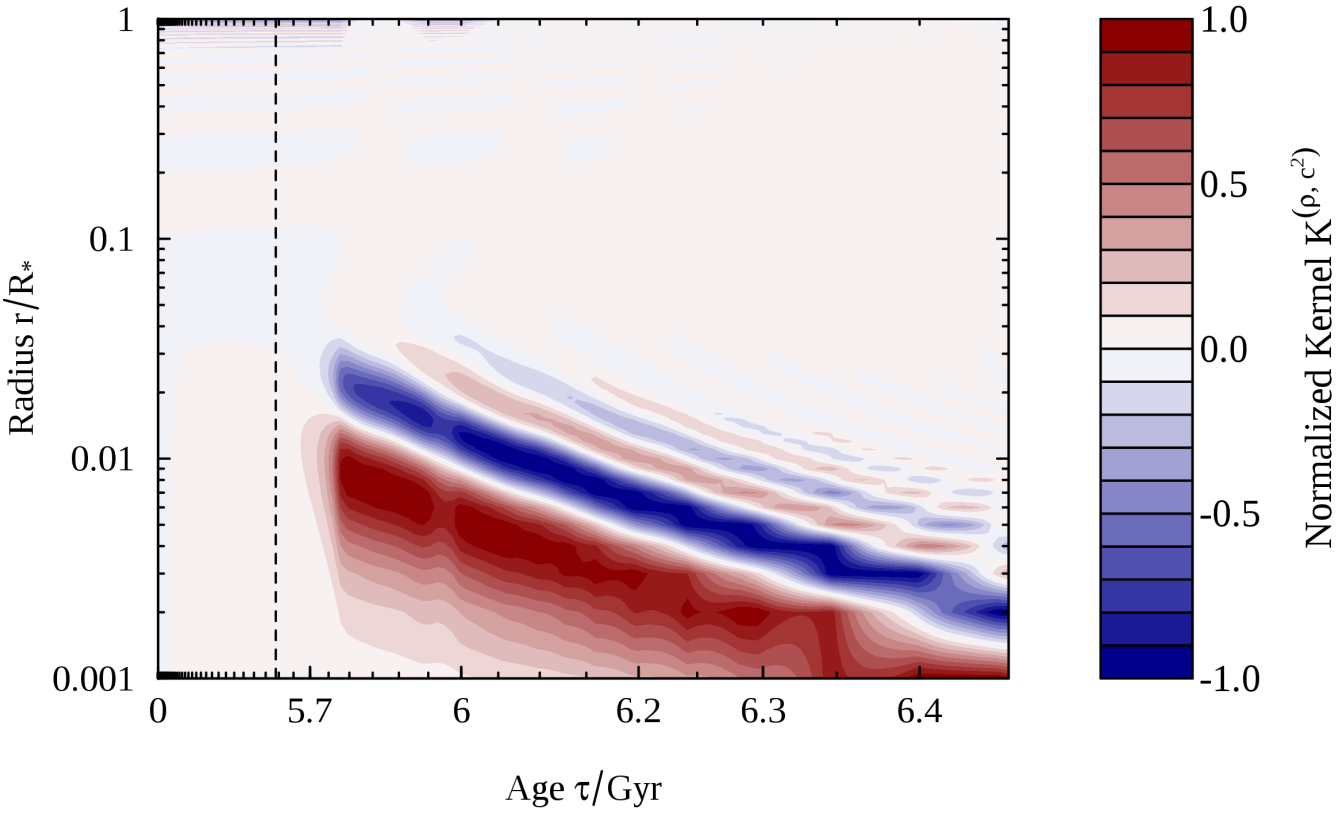


FIGURE 5.3. Evolution of the (ρ, c^2) kernel function for the $(\ell = 1, n = 11)$ mode of a 1.11 $M/M_\odot$ star. The vertical dashed line shows the end of the main sequence (TAMS). As the mode mixes with a g-mode, it develops extreme sensitivity to the deep core structure of the star.

In the longer term, there are other prospects that are quite exciting. Lund et al. (2014) predicted that $\ell = 4$ modes would be observable in 16 Cyg A and B from *Kepler* data. With such data, it would be possible to resolve the sound speed profiles of the observed stars to even shallower layers, which would provide further constraints on theories of the stellar interior. However, recent data releases seem not to have produced any such detections. It does seem feasible within the coming decades that such observations could become available, perhaps through a combination of *Kepler* data with SONG observations (Andersen et al. 2014, Grundahl et al. 2017) and possibly utilizing the forthcoming TESS and PLATO missions.

In this thesis, I used artificial intelligence to assist in solving problems in stellar astrophysics. This is a form of so-called *weak* AI. These tools will only get more powerful with the coming decades. Eventually, we may have *strong* AI, which will be capable of fully driving scientific research. One day, it may be that AI will be able to determine on its own the set of astrophysical laws that are most harmonious with enormous quantities of empirical data.

Bibliography

1. Adams, F. C.: 2008, "Stars in other universes: stellar structure with different fundamental constants", *Journal of Cosmology and Astroparticle Physics* **8**, 010
2. Adler, J. and Öktem, O.: 2017, "Solving ill-posed inverse problems using iterative deep neural networks", *Inverse Problems* **33 (12)**, 124007
3. Aerts, C., Christensen-Dalsgaard, J., and Kurtz, D. W.: 2010, "Asteroseismology", Springer
4. Allison, H.: 1979, "Inverse unstable problems and some of their applications", *The Mathematical Scientist*
5. Andersen, M. F., Grundahl, F., Christensen-Dalsgaard, J., et al.: 2014, "Hardware and software for a robotic network of telescopes - SONG" in *Revista Mexicana de Astronomia y Astrofisica Conference Series*, Vol. 45 of *Revista Mexicana de Astronomia y Astrofisica, vol. 27*, pp 83–86
6. Angelou, G. C., Bellinger, E. P., Hekker, S., and Basu, S.: 2017, "On the Statistical Properties of the Lower Main Sequence", *The Astrophysical Journal* **839**, 116
7. Angulo, C., Arnould, M., Rayet, M., et al.: 1999, "A compilation of charged-particle induced thermonuclear reaction rates", *Nuclear Physics A* **656**, 3
8. Antia, H. M. and Basu, S.: 1994, "Nonasymptotic helioseismic inversion for solar structure", *Astronomy & Astrophysics Supplement Series* 107
9. Antia, H. M. and Chitre, S. M.: 1997, "Helioseismic models and solar neutrino fluxes", *Monthly Notices of the Royal Astronomical Society* **289**, L1
10. Applegate, J. H.: 1988, "Why stars become red giants", *The Astrophysical Journal* **329**, 803
11. Appourchaux, T., Antia, H. M., Ball, W., et al.: 2015, "A seismic and gravitationally bound double star observed by Kepler. Implication for the presence of a convective core", *Astronomy & Astrophysics* **582**, A25
12. Argelander, F.: 1844, "Aufforderung an Freunde der Astronomie, zur Anstellung von eben so interessanten und nützlichen, als leicht auszuführenden Beobachtungen über mehrere wichtige Zweige der Himmelskunde", Schumacher's Jahrbuch für 1844

13. Arny, T.: 1990, "The star makers: A history of the theories of stellar structure and evolution", *Vistas in Astronomy* **33**, 211

14. Aston, F. W.: 1920, "LIX. The Mass-Spectra of Chemical Elements", *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* **39 (233)**, 611

15. Backus, G. and Gilbert, F.: 1968, "The Resolving Power of Gross Earth Data", *Geophysical Journal* **16**, 169

16. Backus, G. and Gilbert, F.: 1970, "Uniqueness in the Inversion of Inaccurate Gross Earth Data", *Philosophical Transactions of the Royal Society of London Series A* **266**, 123

17. Baglin, A., Auvergne, M., Barge, P., et al.: 2006, "Scientific Objectives for a Minisat: CoRoT" in M. Fridlund, A. Baglin, J. Lochard, and L. Conroy (eds.) *The CoRoT Mission Pre-Launch Status - Stellar Seismology and Planet Finding*, Vol. 1306 of *ESA Special Publication*, p. 33

18. Bahcall, J. N., Basu, S., and Pinsonneault, M. H.: 1998, "How uncertain are solar neutrino predictions?", *Physics Letters B* **433**, 1

19. Baldner, C. S. and Basu, S.: 2008, "Solar Cycle Related Changes at the Base of the Convection Zone", *The Astrophysical Journal* **686**, 1349

20. Ball, W. H. and Gizon, L.: 2014, "A new correction of stellar oscillation frequencies for near-surface effects", *Astronomy & Astrophysics* **568**, A123

21. Banerjee, S. and Roy, A.: 2014, "Linear algebra and matrix analysis for statistics", CRC Press

22. Barban, C., Matthews, J., De Ridder, J., et al.: 2006, "Studying solar-like oscillations in red giants: MOST spacebased photometry of epsilon Ophiuchi" in *Proceedings of SOHO 18/GONG 2006/HELAS I, Beyond the spherical Sun*, Vol. 624 of *ESA Special Publication*, p. 30

23. Barban, C., Matthews, J. M., De Ridder, J., et al.: 2007, "Detection of solar-like oscillations in the red giant star ϵ Ophiuchi by MOST spacebased photometry", *Astronomy & Astrophysics* **468**, 1033

24. Baron, F., Monnier, J. D., Pedretti, E., et al.: 2012, "Imaging the Algol Triple System in the H Band with the CHARA Interferometer", *The Astrophysical Journal* **752**, 20

25. Basu, S.: 1998, "Effects of errors in the solar radius on helioseismic inferences", *Monthly Notices of the Royal Astronomical Society* **298**, 719

26. Basu, S.: 2003, "Stellar Inversions", *Astrophysics and Space Science* **284**, 153

27. Basu, S.: 2014, "Studying stars through frequency inversions", Cambridge University Press

28. Basu, S.: 2016, "Global seismology of the Sun", *Living Reviews in Solar Physics* **13**, 2

29. Basu, S. and Antia, H. M.: 1994, "Effects of Diffusion on the Extent of Overshoot Below the Solar Convection Zone", *Monthly Notices of the Royal Astronomical Society* **269**, 1137

30. Basu, S. and Antia, H. M.: 1997, "Seismic measurement of the depth of the solar convection zone", *Monthly Notices of the Royal Astronomical Society* **287**, 189

31. Basu, S. and Chaplin, W.: 2017, "Asteroseismic Data Analysis: Foundations and Techniques", Princeton University Press

32. Basu, S., Chaplin, W. J., Elsworth, Y., New, R., and Serenelli, A. M.: 2009, "Fresh Insights on the Structure of the Solar Core", *The Astrophysical Journal* **699**, 1403

33. Basu, S. and Christensen-Dalsgaard, J.: 1997, "Equation of state and helioseismic inversions.", *Astronomy & Astrophysics* **322**, L5

34. Basu, S., Christensen-Dalsgaard, J., Monteiro, M. J. P. F. G., and Thompson, M. J.: 2001, "Seismology of solar-type stars" in A. Wilson and P. L. Pallé (eds.) *SOHO 10/GONG 2000 Workshop: Helio- and Asteroseismology at the Dawn of the Millennium*, Vol. 464 of *ESA Special Publication*, pp 407–410

35. Basu, S., Christensen-Dalsgaard, J., and Thompson, M. J.: 2002, "SOLA inversions for the core structure of solar-type stars" in B. Battrick, F. Favata, I. W. Roxburgh, and D. Galadi (eds.) *Stellar Structure and Habitable Planet Finding*, Vol. 485 of *ESA Special Publication*, pp 249–252

36. Basu, S., Grevesse, N., Mathis, S., and Turck-Chièze, S.: 2015, "Understanding the Internal Chemical Composition and Physical Processes of the Solar Interior", *Space Science Review* **196**, 49

37. Basu, S., Pinsonneault, M. H., and Bahcall, J. N.: 2000, "How Much Do Helioseismological Inferences Depend on the Assumed Reference Model?", *The Astrophysical Journal* **529**, 1084

38. Batalha, N. M., Borucki, W. J., Bryson, S. T., et al.: 2011, "Kepler's First Rocky Planet: Kepler-10b", *The Astrophysical Journal* **729**, 27

39. Bazot, M., Bourguignon, S., and Christensen-Dalsgaard, J.: 2012, "A Bayesian approach to the modelling of α Cen A", *Monthly Notices of the Royal Astronomical Society* **427**, 1847

40. Bedding, T. R., Butler, R. P., Kjeldsen, H., et al.: 2001, "Evidence for Solar-like Oscillations in β Hydri", *The Astrophysical Journal Letters* **549**, L105

41. Bellinger, E.: 2016, "asteroseismology: Fundamental Parameters of Main-Sequence Stars in an Instant with Machine Learning", zenodo

42. Bellinger, E. P., Angelou, G. C., Hekker, S., et al.: 2016, "Fundamental Parameters of Main-Sequence Stars in an Instant with Machine Learning", *The Astrophysical Journal* **830**, 31

43. Bellinger, E. P., Angelou, G. C., Hekker, S., et al.: 2017a, "Stellar Parameters in an Instant with Machine Learning. Application to Kepler LEGACY Targets" in *Seismology of the Sun and the Distant Stars*, Vol. 160 of *European Physical Journal Web of Conferences*, p. 05003

44. Bellinger, E. P., Basu, S., Hekker, S., and Ball, W. H.: 2017b, "Model-independent Measurement of Internal Stellar Structure in 16 Cygni A and B", *The Astrophysical Journal* **851**, 80

45. Bélopolsky, A.: 1895, "The spectrum of delta Cephei", *The Astrophysical Journal* 1

46. Bélopolsky, A.: 1897, "Researches on the spectrum of the variable star eta Aquilae", *The Astrophysical Journal* 6

47. Bengtsson, H.: 2015, "matrixStats: Methods that Apply to Rows and Columns of Matrices (and to Vectors)", R package version 0.14.2

48. Berkelaar, M. and others: 2015, "lpSolve: Interface to lpSolve v. 5.5 to Solve Linear/Integer Programs", R package version 5.6.12

49. Berthomieu, G., Toutain, T., Gonczi, G., et al.: 2001, "About structure inversions of simulated COROT data for a solar like star" in A. Wilson and P. L. Pallé (eds.) *SOHO 10/GONG 2000 Workshop: Helio- and Asteroseismology at the Dawn of the Millennium*, Vol. 464 of *ESA Special Publication*, pp 411–414

50. Bessel, F. W.: 1838, "Bestimmung der Entfernung des 61sten Sterns des Schwans.", *Astronomische Nachrichten* **16**, 65

51. Bischl, B. and Lang, M.: 2015, "parallelMap: Unified Interface to Parallelization Back-Ends", R package version 1.3

52. Böhm-Vitense, E.: 1958, "Über die Wasserstoffkonvektionszone in Sternen verschiedener Effektivtemperaturen und Leuchtkräfte. Mit 5 Textabbildungen", *Zeitschrift für Astrophysik* **46**, 108

53. Bolt, M., Hockey, T., Palmeri, J., et al.: 2007, "Biographical Encyclopedia of Astronomers", Springer

54. Borucki, W. J., Koch, D., Basri, G., et al.: 2010, "Kepler Planet-Detection Mission: Introduction and First Results", *Science* **327**, 977

55. Borucki, W. J., Koch, D. G., Batalha, N., et al.: 2012, "Kepler-22b: A 2.4 Earth-radius Planet in the Habitable Zone of a Sun-like Star", *The Astrophysical Journal* **745**, 120

56. Bouchy, F. and Carrier, F.: 2001, "P-mode observations on α Cen A", *Astronomy & Astrophysics* **374**, L5

57. Breiman, L.: 2001, "Random Forests", *Machine Learning* **45 (1)**, 5

58. Brester, A.: 1889, "Variable Stars and the Constitution of the Sun", *Nature* **39**, 606

59. Broomhall, A.-M., Chaplin, W. J., Davies, G. R., et al.: 2009, "Definitive Sun-as-a-star p-mode frequencies: 23 years of BiSON observations", *Monthly Notices of the Royal Astronomical Society* **396**, L100

60. Brown, E. F.: 2015, "Stellar Astrophysics", Open Astrophysics Bookshelf

61. Brown, T. M., Christensen-Dalsgaard, J., Dziembowski, W. A., et al.: 1989, "Inferring the sun's internal angular velocity from observed p-mode frequency splittings", *The Astrophysical Journal* **343**, 526

62. Brown, T. M., Christensen-Dalsgaard, J., Weibel-Mihalas, B., and Gilliland, R. L.: 1994, "The effectiveness of oscillation frequencies in constraining stellar model parameters", *The Astrophysical Journal* **427**, 1013

63. Brown, T. M. and Gilliland, R. L.: 1990, "A search for solar-like oscillations in Alpha Centauri A", *The Astrophysical Journal* **350**, 839

64. Brown, T. M., Gilliland, R. L., Noyes, R. W., and Ramsey, L. W.: 1991, "Detection of possible p-mode oscillations on Procyon", *The Astrophysical Journal* **368**, 599

65. Bruno, G.: 1584, "De l'infinito, universo e mondi", English: "On the Infinite Universe and Worlds"

66. Brunt, D.: 1913, "The problem of the Cepheid variables", *The Observatory* **36**, 59

67. Bruntt, H., Basu, S., Smalley, B., et al.: 2012, "Accurate fundamental parameters and detailed abundance patterns from spectroscopy of 93 solar-type Kepler targets", *Monthly Notices of the Royal Astronomical Society* **423**, 122

68. Bruntt, H., Bedding, T. R., Quirion, P.-O., et al.: 2010, "Accurate fundamental parameters for 23 bright solar-type stars", *Monthly Notices of the Royal Astronomical Society* **405**, 1907

69. Buldgen, G., Reese, D. R., and Dupret, M. A.: 2015a, "Using seismic inversions to obtain an indicator of internal mixing processes in main-sequence solar-like stars", *Astronomy & Astrophysics* **583**, A62

70. Buldgen, G., Reese, D. R., and Dupret, M. A.: 2016a, "Constraints on the structure of 16 Cygni A and 16 Cygni B using inversion techniques", *Astronomy & Astrophysics* **585**, A109

71. Buldgen, G., Reese, D. R., Dupret, M. A., and Samadi, R.: 2015b, "Stellar acoustic radii, mean densities, and ages from seismic inversion techniques", *Astronomy & Astrophysics* **574**, A42

72. Buldgen, G., Salmon, S. J. A. J., Reese, D. R., and Dupret, M. A.: 2016b, "In-depth study of 16CygB using inversion techniques", *Astronomy & Astrophysics* **596**, A73

73. Burgers, J. M.: 1969, "Flow Equations for Composite Gases" Technical report, DTIC Document

74. Buzasi, D.: 2000, "Platforms of opportunity: asteroseismology by Piggyback" in R. Pallavicini, G. Micela, and S. Sciortino (eds.) *Stellar Clusters and Associations: Convection, Rotation, and Dynamos*, Vol. 198 of *Astronomical Society of the Pacific Conference Series*, p. 557

75. Buzasi, D., Catanzarite, J., Laher, R., et al.: 2000, "The Detection of Multimodal Oscillations on α Ursae Majoris", *The Astrophysical Journal Letters* **532**, L133

76. Campante, T. L., Barclay, T., Swift, J. J., et al.: 2015, "An Ancient Extrasolar System with Five Sub-Earth-size Planets", *The Astrophysical Journal* **799**, 170

77. Campante, T. L., Schofield, M., Kuszlewicz, J. S., et al.: 2016, "The Asteroseismic Potential of TESS: Exoplanet-host Stars", *The Astrophysical Journal* **830**, 138

78. Caruana, R. and Niculescu-Mizil, A.: 2006, "An Empirical Comparison of Supervised Learning Algorithms" in *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, pp 161–168, ACM, New York, NY, USA

79. Catelan, M. and Smith, H. A.: 2015, "Pulsating Stars", Wiley-VCH

80. Chambers, G. F.: 1865, "A Catalogue of Variable Stars", *Astronomische Nachrichten* **63**, 117

81. Chandrasekhar, S.: 1939, "An Introduction to the Study of Stellar Structure", The University of Chicago Press

82. Chandrasekhar, S.: 1964, "A General Variational Principle Governing the Radial and the Non-Radial Oscillations of Gaseous Masses", *The Astrophysical Journal* **139**, 664

83. Chaplin, W. J., Appourchaux, T., Elsworth, Y., et al.: 2010, "The Asteroseismic Potential of Kepler: First Results for Solar-Type Stars", *The Astrophysical Journal Letters* **713**, L169

84. Chaplin, W. J., Basu, S., Huber, D., et al.: 2014, "Asteroseismic Fundamental Properties of Solar-type Stars Observed by the NASA Kepler Mission", *The Astrophysical Journal Supplement Series* **210**, 1

85. Chaplin, W. J., Kjeldsen, H., Christensen-Dalsgaard, J., et al.: 2011, "Ensemble Asteroseismology of Solar-Type Stars with the NASA Kepler Mission", *Science* **332**, 213

86. Chaplin, W. J. and Miglio, A.: 2013, "Asteroseismology of Solar-Type and Red-Giant Stars", *Annual Review of Astronomy and Astrophysics* **51**, 353

87. Chen, S., Montgomery, J., and Bolufé-Röhler, A.: 2015, "Measuring the curse of dimensionality and its effects on particle swarm optimization and differential evolution", *Applied Intelligence* **42 (3)**, 514

88. Chen, Y., Davis, T. A., Hager, W. W., and Rajamanickam, S.: 2008, "Algorithm 887: CHOLMOD, Supernodal Sparse Cholesky Factorization and Update/Downdate", *ACM Trans. Math. Softw.* **35**, 22:1

89. Chiappini, C., Minchev, I., Anders, F., et al.: 2015, "New Observational Constraints to Milky Way Chemodynamical Models", *Astrophysics and Space Science Proceedings* **39**, 111

90. Christensen-Dalsgaard, J.: 1982, "On solar models and their periods of oscillation", *Monthly Notices of the Royal Astronomical Society* **199**, 735

91. Christensen-Dalsgaard, J.: 1984, "What Will Asteroseismology Teach us" in A. Mangeney and F. Praderie (eds.) *Space Research in Stellar Activity and Variability*, p. 11

92. Christensen-Dalsgaard, J.: 2002, "Helioseismology", *Reviews of Modern Physics* **74**, 1073

93. Christensen-Dalsgaard, J.: 2008, "ADIPLS—the Aarhus adiabatic oscillation package", *Astrophysics and Space Science* **316**, 113

94. Christensen-Dalsgaard, J.: 2012, "Stellar model fits and inversions", *Astronomische Nachrichten* **333**, 914

95. Christensen-Dalsgaard, J., Dappen, W., Ajukov, S. V., et al.: 1996, "The Current State of Solar Modeling", *Science* **272**, 1286

96. Christensen-Dalsgaard, J., Duvall, Jr., T. L., Gough, D. O., Harvey, J. W., and Rhodes, Jr., E. J.: 1985, "Speed of sound in the solar interior", *Nature* **315**, 378

97. Christensen-Dalsgaard, J. and Frandsen, S.: 1983, "Stellar 5 min oscillations", *Solar Physics* **82**, 469

98. Christensen-Dalsgaard, J. and Gough, D. O.: 1976, "Towards a heliological inverse problem", *Nature* **259**, 89

99. Christensen-Dalsgaard, J. and Gough, D. O.: 1980, "Is the sun helium-deficient", *Nature* **288**, 544

100. Christensen-Dalsgaard, J., Gough, D. O., and Thompson, M. J.: 1991, "The depth of the solar convection zone", *The Astrophysical Journal* **378**, 413

101. Christensen-Dalsgaard, J., Proffitt, C. R., and Thompson, M. J.: 1993, "Effects of diffusion on solar models and their oscillation frequencies", *The Astrophysical Journal Letters* **403**, L75

102. Claverie, A., Isaak, G. R., McLeod, C. P., van der Raay, H. B., and Cortes, T. R.: 1979, "Solar structure from global studies of the 5-minute oscillation", *Nature* **282**, 591

103. Claverie, A., Isaak, G. R., McLeod, C. P., van der Raay, H. B., and Roca Cortes, T.: 1981, "Structure of the 5-minute solar oscillations - 1976-1980", *Solar Physics* **74**, 51

104. Coc, A., Ekström, S., Descouvemont, P., et al.: 2010, "Effects of the variation of fundamental constants on Pop III stellar evolution" in *American Institute of Physics Conference Series*, Vol. 1269, pp 21–26

105. Coc, A., Uzan, J.-P., and Vangioni, E.: 2014, "Standard big bang nucleosynthesis and primordial CNO abundances after Planck", *Journal of Cosmology and Astroparticle Physics* **10**, 050

106. Cokelaer, T.: 2016, "Bioinformatics in Python", version 0.3.2

107. Collins, G. W.: 1989, "The Fundamentals of Stellar Astrophysics", W. H. Freeman and Co.

108. Copernicus, N.: 1543, "De revolutionibus orbium coelestium", English: "On the Revolutions of the Heavenly Spheres"

109. Coppersmith, D. and Winograd, S.: 1990, "Matrix multiplication via arithmetic progressions", *Journal of Symbolic Computation* **9 (3)**, 251

110. Cowling, T. G.: 1941, "The non-radial oscillations of polytropic stars", *Monthly Notices of the Royal Astronomical Society* **101**, 367

111. Cox, J. P.: 1980, "Theory of Stellar Pulsation", Princeton University Press

112. Curtiss, R. H.: 1905, "On the light- and velocity-curves of W Sagittarii", *The Astrophysical Journal* 22

113. Däppen, W., Gough, D. O., Kosovichev, A. G., and Thompson, M. J.: 1991, "A New Inversion for the Hydrostatic Stratification of the Sun" in D. Gough and J. Toomre (eds.) *Challenges to Theories of the Structure of Moderate-Mass Stars*, Vol. 388 of *Lecture Notes in Physics, Berlin Springer Verlag*, p. 111

114. Darwin, C. R.: 1859, "On the Origin of Species", John Murray

115. Davies, G. R., Aguirre, V. S., Bedding, T. R., et al.: 2016, "Oscillation frequencies for 35 Kepler solar-type planet-hosting stars using Bayesian techniques and machine learning", *Monthly Notices of the Royal Astronomical Society* **456**, 2183

116. Davies, G. R., Broomhall, A. M., Chaplin, W. J., Elsworth, Y., and Hale, S. J.: 2014a, "Low-frequency, low-degree solar p-mode properties from 22 years of Birmingham Solar Oscillations Network data", *Monthly Notices of the Royal Astronomical Society* **439**, 2025

117. Davies, G. R., Chaplin, W. J., Farr, W. M., et al.: 2015, "Asteroseismic inference on rotation, gyrochronology and planetary system dynamics of 16 Cygni", *Monthly Notices of the Royal Astronomical Society* **446**, 2959

118. Davies, G. R., Handberg, R., Miglio, A., et al.: 2014b, "Why should we correct reported pulsation frequencies for stellar line-of-sight Doppler velocity shifts?", *Monthly Notices of the Royal Astronomical Society* **445**, L94

119. de Boor, C.: 1972, "On calculating with B-splines", *Journal of Approximation Theory* **6 (1)**, 50

120. De Lucca, R.: 1998, "Giordano Bruno: Cause, Principle and Unity: And Essays on Magic", Cambridge University Press

121. De Ridder, J., Barban, C., Baudin, F., et al.: 2009, "Non-radial oscillation modes with long lifetimes in giant stars", *Nature* **459**, 398

122. Deheuvels, S., Bruntt, H., Michel, E., et al.: 2010, "Seismic and spectroscopic characterization of the solar-like pulsating CoRoT target HD 49385", *Astronomy & Astrophysics* **515**, A87

123. Deheuvels, S., Doğan, G., Goupil, M. J., et al.: 2014, "Seismic constraints on the radial dependence of the internal rotation profiles of six Kepler subgiants and young red giants", *Astronomy & Astrophysics* **564**, A27

124. Deheuvels, S., García, R. A., Chaplin, W. J., et al.: 2012, "Seismic Evidence for a Rapidly Rotating Core in a Lower-giant-branch Star Observed with Kepler", *The Astrophysical Journal* **756**, 19

125. Delmotte, F.: 2014, "Sample equidistant points from a numeric vector", StackOverflow

126. Demarque, P., Guenther, D. B., Li, L. H., Mazumdar, A., and Straka, C. W.: 2008, "YREC: the Yale rotating stellar evolution code. Non-rotating version, seismology applications", *Astrophysics and Space Science* **316**, 31

127. Deming, W. E.: 1943, "Statistical Adjustment of Data", Wiley

128. Deubner, F.-L.: 1975, "Observations of low wavenumber nonradial eigenmodes of the sun", *Astronomy & Astrophysics* **44**, 371

129. Deubner, F.-L. and Gough, D.: 1984, "Helioseismology: Oscillations as a Diagnostic of the Solar Interior", *Annual Review of Astronomy and Astrophysics* **22**, 593

130. di Mauro, M. P.: 2004, "Theoretical Aspects of Asteroseismology: Small Steps Towards a Golden Future" in D. Danesy (ed.) *SOHO 14 Helio- and Asteroseismology: Towards a Golden Future*, Vol. 559 of *ESA Special Publication*, p. 186

131. Di Mauro, M. P., Ventura, R., Cardini, D., et al.: 2016, "Internal Rotation of the Red-giant Star KIC 4448777 by Means of Asteroseismic Inversion", *The Astrophysical Journal* **817**, 65

132. Dirac, P. A. M.: 1938, "A New Basis for Cosmology", *Proceedings of the Royal Society of London* **165 (921)**, 199

133. Dowle, M., Srinivasan, A., Short, T., with contributions from R Saporta, S. L., and Antonyan, E.: 2015, "data.table: Extension of Data.frame", R package version 1.9.6

134. Duncan, J. C.: 1909, "The orbits of the Cepheid variables Y Sagittarii and RT Aurigae : with a discussion of the possible causes of this type of stellar variation", *Lick Observatory Bulletin* **5**, 82

135. Dunnett, C. W.: 1955, "A Multiple Comparison Procedure for Comparing Several Treatments with a Control", *Journal of the American Statistical Association* **50 (272)**, 1096

136. Duvall, Jr., T. L., Dziembowski, W. A., Goode, P. R., et al.: 1984, "Internal rotation of the sun", *Nature* **310**, 22

137. Dziembowski, W. A., Gough, D. O., Houdek, G., and Sienkiewicz, R.: 2001, "Oscillations of α UMa and other red giants", *Monthly Notices of the Royal Astronomical Society* **328**, 601

138. Dziembowski, W. A., Pamyatnykh, A. A., and Sienkiewicz, R.: 1990, "Solar model from helioseismology and the neutrino flux problem", *Monthly Notices of the Royal Astronomical Society* **244**, 542

139. Eddington, A. S.: 1916, "On the radiative equilibrium of the stars", *Monthly Notices of the Royal Astronomical Society* **77**, 16

140. Eddington, A. S.: 1917, "The pulsation theory of Cepheid variables", *The Observatory* **40**, 290

141. Eddington, A. S.: 1918, "Stars, Gaseous, On the pulsations of a gaseous star", *Monthly Notices of the Royal Astronomical Society* **79**, 2

142. Eddington, A. S.: 1920, "The Internal Constitution of the Stars", *The Scientific Monthly* **11**, 297

143. Eddington, A. S.: 1924, "On the relation between the masses and luminosities of the stars", *Monthly Notices of the Royal Astronomical Society* **84**, 308

144. Eddington, A. S.: 1926, "The Internal Constitution of the Stars", Cambridge University Press

145. Edmonds, P. D. and Cram, L. E.: 1995, "A Search for Global Acoustic Oscillations on ALPHA-1-CENTAURI and Beta-Hydri", *Monthly Notices of the Royal Astronomical Society* **276**, 1295

146. Eggleton, P. P. and Faulkner, J.: 1981, "Why Do Stars Become Red Giants?" in I. Iben and A. Renzini (eds.) *Physical Processes in Red Giants*, pp 179–182, Springer Netherlands, Dordrecht

147. Einstein, A.: 1905, "Ist die Trägheit eines Körpers von seinem Energieinhalt abhängig?", *Annalen der Physik* **323**, 639

148. Fabricius, J.: 1611, "De maculis in sole observatis narratio", Wittenberga

149. Fai, K., Wei, Q., Carin, L., and Heller, K.: 2017, "An inner-loop free solution to inverse problems using deep neural networks" in *NIPS 2017*

150. Fath, E. A.: 1935, "A photometric study of delta Scuti", *Lick Observatory Bulletin* **17**, 175

151. Flamsteed, J.: 1725, "Historia Coelestis Britannica"

152. Fox, P. A. and Kerr, R. M.: 2000, "Geophysical & Astrophysical Convection", CRC Press

153. Frandsen, S., Carrier, F., Aerts, C., et al.: 2002, "Detection of Solar-like oscillations in the G7 giant star ξ Hya", *Astronomy & Astrophysics* **394**, L5

154. Frazier, E. N.: 1968, "An Observational Study of the Hydrodynamics of the Lower Solar Photosphere", *The Astrophysical Journal* **152**, 557

155. Frost, E. B.: 1908, "Hermann Carl Vogel", *The Astrophysical Journal* **27**, 1

156. Gai, N., Basu, S., Chaplin, W. J., and Elsworth, Y.: 2011, "An In-depth Study of Grid-based Asteroseismic Analysis", *The Astrophysical Journal* **730**, 63

157. Gaia Collaboration, Brown, A. G. A., Vallenari, A., et al.: 2016, "Gaia Data Release 1. Summary of the astrometric, photometric, and survey properties", *Astronomy & Astrophysics* **595**, A2

158. Galilei, G.: 1610, "Sidereus Nuncius", English: "Starry Messenger"

159. Gallart, C., Zoccali, M., and Aparicio, A.: 2005, "The Adequacy of Stellar Evolution Models for the Interpretation of the Color-Magnitude Diagrams of Resolved Stellar Populations", *Annual Review of Astronomy and Astrophysics* **43**, 387

160. Gamow, G.: 1928, "The Quantum Theory of Nuclear Disintegration", *Nature* **122**, 805

161. Gamow, G.: 1938, "Tracks of Stellar Evolution", *Physical Review* **53**, 907

162. Gaulme, P., McKeever, J., Jackiewicz, J., et al.: 2016, "Testing the Asteroseismic Scaling Relations for Red Giants with Eclipsing Binaries Observed by Kepler", *The Astrophysical Journal* **832**, 121

163. Gelly, B., Grec, G., and Fossat, E.: 1986, "Evidence for global pressure oscillations in Procyon and Alpha Centauri", *Astronomy & Astrophysics* **164**, 383

164. Geurts, P., Ernst, D., and Wehenkel, L.: 2006, "Extremely Randomized Trees", *Machine Learning* **63 (1)**, 3

165. Giles, P. M.: 2000, "Time-distance measurements of large-scale flows in the solar convection zone", *Ph.D. thesis*, Stanford University

166. Gilliland, R. L., Brown, T. M., Christensen-Dalsgaard, J., et al.: 2010, "Kepler Asteroseismology Program: Introduction and First Results", *Publications of the Astronomical Society of the Pacific* **122**, 131

167. Goldreich, P. and Keeley, D. A.: 1977, "Solar seismology. II - The stochastic excitation of the solar p-modes by turbulent convection", *The Astrophysical Journal* **212**, 243

168. Gontcharov, G. A.: 2006, "Pulkovo Compilation of Radial Velocities for 35 495 Hipparcos stars in a common system", *Astronomy Letters* **32**, 759

169. Goodricke, J.: 1783, "A Series of Observations on, and a Discovery of, the Period of the Variation of the Light of the Bright Star in the Head of Medusa, Called Algol. In a Letter from John Goodricke, Esq. to the Rev.

Anthony Shepherd, D. D. F. R. S. and Plumian Professor at Cambridge", *Philosophical Transactions of the Royal Society of London* **73**, 474

170. Goodricke, J.: 1784, "On the Periods of the Changes of Light in the Star Algol. In a Letter from John Goodricke, Esq. to the Rev. Anthony Shepherd, D. D. F. R. S. Professor of Astronomy at Cambridge", *Philosophical Transactions of the Royal Society of London* **74**, 287
171. Goodricke, J.: 1786, "A Series of Observations on, and a Discovery of, the Period of the Variation of the Light of the Star Marked Îτ by Bayer, Near the Head of Cepheus. In a Letter from John Goodricke, Esq. to Nevil Maskelyne, D. D. F. R. S. and Astronomer Royal", *Philosophical Transactions of the Royal Society of London* **76**, 48
172. Gordon, C. and Webb, D.: 1996, "You Can't Hear the Shape of a Drum", *American Scientist* **84 (1)**, 46
173. Gough, D.: 1985, "Inverting helioseismic data", *Solar Physics* **100**, 65
174. Gough, D. and Toomre, J.: 1991, "Seismic observations of the solar interior", *Annual Review of Astronomy and Astrophysics* **29**, 627
175. Gough, D. O.: 1981, "A new measure of the solar rotation", *Monthly Notices of the Royal Astronomical Society* **196**, 731
176. Gough, D. O.: 1993, "Linear adiabatic stellar pulsation." in J.-P. Zahn and J. Zinn-Justin (eds.) *Astrophysical Fluid Dynamics - Les Houches 1987*, pp 399–560
177. Gough, D. O.: 1998, "Inversion for the internal structure and rotation of the Sun and other sun-like stars" in H. Kjeldsen and T. R. Bedding (eds.) *The First MONS Workshop: Science with a Small Space Telescope*, p. 33
178. Gough, D. O. and Kosovichev, A. G.: 1993, "Initial asteroseismic inversions" in W. W. Weiss and A. Baglin (eds.) *IAU Colloq. 137: Inside the Stars*, Vol. 40 of *Astronomical Society of the Pacific Conference Series*, p. 541
179. Gough, D. O. and Thompson, M. J.: 1991, "The Inversion Problem", University of Arizona Press
180. Grec, G., Fossat, E., and Pomerantz, M.: 1980, "Solar oscillations - Full disk observations from the geographic South Pole", *Nature* **288**, 541
181. Grevesse, N. and Sauval, A. J.: 1998, "Standard Solar Composition", *Space Science Review* **85**, 161
182. Grundahl, F., Fredslund Andersen, M., Christensen-Dalsgaard, J., et al.: 2017, "First Results from the Hertzsprung SONG Telescope: Asteroseismology of the G5 Subgiant Star μ Herculis", *The Astrophysical Journal* **836**, 142
183. Guggenberger, E., Hekker, S., Angelou, G. C., Basu, S., and Bellinger, E. P.: 2017, "Mitigating the mass dependence in the $\Delta\nu$ scaling relation of red giant stars", *Monthly Notices of the Royal Astronomical Society* **470**, 2069

184. Guggenberger, E., Hekker, S., Basu, S., and Bellinger, E.: 2016, "Significantly improving stellar mass and radius estimates: a new reference function for the $\Delta\nu$ scaling relation", *Monthly Notices of the Royal Astronomical Society* **460**, 4277

185. Haberreiter, M., Schmutz, W., and Kosovichev, A. G.: 2008, "Solving the Discrepancy between the Seismic and Photospheric Solar Radius", *The Astrophysical Journal Letters* **675**, L53

186. Hadamard, J.: 1902, "Sur les problèmes aux dérivés partielles et leur signification physique", *Princeton University Bulletin* **13**, 49

187. Hampel, F. R.: 1971, "A general qualitative definition of robustness", *The Annals of Mathematical Statistics* pp 1887–1896

188. Han, E., Wang, S. X., Wright, J. T., et al.: 2014, "Exoplanet Orbit Database. II. Updates to Exoplanets.org", *Publications of the Astronomical Society of the Pacific* **126**, 827

189. Hansen, C. J. and Kawaler, S. D.: 1994, "Stellar Interiors. Physical Principles, Structure, and Evolution.", Springer-Verlag

190. Hart, A. B.: 1954, "Motions in the Sun at the photospheric level. IV. The equatorial rotation and possible velocity fields in the photosphere", *Monthly Notices of the Royal Astronomical Society* **114**, 17

191. Hart, A. B.: 1956, "Motions in the Sun at the photospheric level. VI. Large-scale motions in the equatorial region", *Monthly Notices of the Royal Astronomical Society* **116**, 38

192. Hastie, T., Tibshirani, R., and Friedman, J.: 2009, "The Elements of Statistical Learning", Springer

193. Hekker, S.: 2013, "CoRoT and Kepler results: Solar-like oscillators", *Advances in Space Research* **52**, 1581

194. Hekker, S., Broomhall, A.-M., Chaplin, W. J., et al.: 2010, "The Octave (Birmingham-Sheffield Hallam) automated pipeline for extracting oscillation parameters of solar-like main-sequence stars", *Monthly Notices of the Royal Astronomical Society* **402**, 2049

195. Hekker, S. and Christensen-Dalsgaard, J.: 2017, "Giant star seismology", *Astronomy & Astrophysics Review* **25**, 1

196. Hekker, S., Kallinger, T., Baudin, F., et al.: 2009, "Characteristics of solar-like oscillations in red giants observed in the CoRoT exoplanet field", *Astronomy & Astrophysics* **506**, 465

197. Hekker, S., Reffert, S., Quirrenbach, A., et al.: 2006, "Precise radial velocities of giant stars. I. Stable stars", *Astronomy & Astrophysics* **454**, 943

198. Henyey, L. G., Wilets, L., Böhm, K. H., Lelevier, R., and Levee, R. D.: 1959, "A Method for Automatic Computation of Stellar Evolution.", *The Astrophysical Journal* **129**, 628

199. Hertzsprung, E.: 1905, "Zur Stralung Der Strerne", *Zeitschrift Fur Wissenschaftliche Photographie, Vol 3, p. 442-449* **3**, 442

200. Hertzsprung, E.: 1907, "Zur Stralung Der Strerne", *Zeitschrift Fur Wissenschaftliche Photographie, Vol 5, p. 86-107* **5**, 86

201. Hertzsprung, E.: 1911, "Nachweis der Veränderlichkeit von α Ursae minoris", *Astronomische Nachrichten* **189**, 89

202. Hertzsprung, E.: 1913, "Über die räumliche Verteilung der Veränderlichen vom δ Cephei-Typus", *Astronomische Nachrichten* **196**, 201

203. Hevelius, J.: 1662, "Mercurius in Sole visus Gedani"

204. Hevelius, J.: 1671, "An Extract of a Letter, Written to the Publisher by the Excellent Johannes Hevelius, Concerning, His Further Observations of the New Star Near the Beak of the Swan; To be Compared with What Was Formerly Published of the Same Argument in Numb. 65. and Numb. 66", *Philosophical Transactions (1665-1678)* **6**, 2197

205. Hoffleit, D.: 1997, "History of the Discovery of Mira Stars", *Journal of the American Association of Variable Star Observers (JAAVSO)* **25**, 115

206. Houdek, G., Trampedach, R., Aarslev, M. J., and Christensen-Dalsgaard, J.: 2017, "On the surface physics affecting solar oscillation frequencies", *Monthly Notices of the Royal Astronomical Society* **464**, L124

207. Howe, R.: 2009, "Solar Interior Rotation and its Variation", *Living Reviews in Solar Physics* **6**, 1

208. Howell, S. B., Sobeck, C., Haas, M., et al.: 2014, "The K2 Mission: Characterization and Early Results", *Publications of the Astronomical Society of the Pacific* **126**, 398

209. Hubble, E.: 1929, "A Relation between Distance and Radial Velocity among Extra-Galactic Nebulae", *Proceedings of the National Academy of Science* **15**, 168

210. Hubble, E. P.: 1925, "Cepheids in spiral nebulae", *The Observatory* **48**, 139

211. Huber, D., Carter, J. A., Barbieri, M., et al.: 2013, "Stellar Spin-Orbit Misalignment in a Multiplanet System", *Science* **342**, 331

212. Huber, D., Ireland, M. J., Bedding, T. R., et al.: 2012, "Fundamental Properties of Stars Using Asteroseismology from Kepler and CoRoT and Interferometry from the CHARA Array", *The Astrophysical Journal* **760**, 32

213. Huber, D., Stello, D., Bedding, T. R., et al.: 2009, "Automated extraction of oscillation parameters for Kepler observations of solar-type stars", *Communications in Asteroseismology* **160**, 74

214. Hund, F.: 1927, "Zur deutung der molekelspektren. I", *Zeitschrift für Physik* **40 (10)**, 742

215. Hunter, J. D.: 2007, "Matplotlib: A 2D graphics environment", *Computing In Science & Engineering* **9 (3)**, 90

216. Huygens, C.: 1698, "Cosmotheoros: The Celestial Worlds discover'd: Or, Conjectures concerning the inhabitants, plants and productions of the worlds in the planets"

217. Iglesias, C. A. and Rogers, F. J.: 1996, "Updated Opal Opacities", *The Astrophysical Journal* **464**, 943

218. Innis, J. L., Isaak, G. R., Speake, C. C., Williams, H. K., and Brazier, R. I.: 1991, "High-precision velocity observations of Procyon A. I - Search for p-mode oscillations from 1988, 1989 and 1990 observations", *Monthly Notices of the Royal Astronomical Society* **249**, 643

219. Itoh, N., Hayashi, H., Nishikawa, A., and Kohyama, Y.: 1996, "Neutrino Energy Loss in Stellar Interiors. VII. Pair, Photo-, Plasma, Bremsstrahlung, and Recombination Neutrino Processes", *The Astrophysical Journal Supplement Series* **102**, 411

220. Jeans, J. H.: 1919, "The problem of the Cepheid variables", *The Observatory* **42**, 88

221. Jeans, J. H.: 1928, "Liquid Stars", *Nature* **121**, 173

222. Jetsu, L. and Porceddu, S.: 2015, "Shifting Milestones of Natural Sciences: The Ancient Egyptian Discovery of Algol's Period Confirmed", *PloS One* **10 (12)**, e0144140

223. Jin, K. H., McCann, M. T., Froustey, E., and Unser, M.: 2017, "Deep convolutional neural network for inverse problems in imaging", *IEEE Transactions on Image Processing* **26 (9)**, 4509

224. Kac, M.: 1966, "Can One Hear the Shape of a Drum?", *The American Mathematical Monthly* **73 (4)**, 1

225. Kafka, S.: 2017, "The American Association of Variable Star Observers (AAVSO)"

226. KASOC: 2018, "Kepler Asteroseismic Science Operations Center"

227. Kelvin: 1895, "The Age of the Earth", *Nature* **51**, 438

228. Kepler, J.: 1609, "Astronomia nova", English: "New Astronomy"

229. Kippenhahn, R. and Weigert, A.: 1990, "Stellar Structure and Evolution", Springer-Verlag

230. Kippenhahn, R., Weigert, A., and Weiss, A.: 2012, "Stellar Structure and Evolution", Springer-Verlag

231. Kirsch, A.: 2011, "An introduction to the mathematical theory of inverse problems", Springer Science & Business Media

232. Kjeldsen, H. and Bedding, T. R.: 1995, "Amplitudes of stellar oscillations: the implications for asteroseismology.", *Astronomy & Astrophysics* **293**, 87

233. Kjeldsen, H., Bedding, T. R., and Christensen-Dalsgaard, J.: 2008, "Correcting Stellar Oscillation Frequencies for Near-Surface Effects", *The Astrophysical Journal Letters* **683**, L175

234. Koch, D. G., Borucki, W., Dunham, E., et al.: 2004, "Overview and status of the Kepler Mission" in J. C. Mather (ed.) *Optical, Infrared, and Millimeter Space Telescopes*, Vol. 5487 of *Proc. SPIE*, pp 1491–1500

235. Koch, D. G., Borucki, W. J., Basri, G., et al.: 2010, "Kepler Mission Design, Realized Photometric Performance, and Early Science", *The Astrophysical Journal Letters* **713**, L79

236. Kosovichev, A. G.: 1999, "Inversion methods in helioseismology and solar tomography", *Journal of Computational and Applied Mathematics* **109**, 1

237. Kosovichev, A. G.: 2011, "Advances in Global and Local Helioseismology: An Introductory Review" in J.-P. Rozelot and C. Neiner (eds.) *Lecture Notes in Physics, Berlin Springer Verlag*, Vol. 832

238. Krasinsky, G. A. and Brumberg, V. A.: 2004, "Secular increase of astronomical unit from analysis of the major planet motions, and its interpretation", *Celestial Mechanics and Dynamical Astronomy* **90**, 267

239. Krishnamoorthy, A. and Menon, D.: 2013, "Matrix inversion using Cholesky decomposition" in *2013 Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA)*, pp 70–72

240. Kullback, S. and Leibler, R. A.: 1951, "On Information and Sufficiency", *The Annals of Mathematical Statistics* **22 (1)**, 79

241. Lane, H. J.: 1870, "On the theoretical temperature of the Sun, under the hypothesis of a gaseous mass maintaining its volume by its internal heat, and depending on the laws of gases as known to terrestrial experiment", *American Journal of Science* **50**, 57

242. Langford, E., Schwertman, N., and Owens, M.: 2001, "Is the Property of Being Positively Correlated Transitive?", *The American Statistician* **55 (4)**, 322

243. Larsson, G., Maire, M., and Shakhnarovich, G.: 2016, "Learning representations for automatic colorization" in *European Conference on Computer Vision*, pp 577–593, Springer

244. Latham, D. W., Stefanik, R. P., Torres, G., et al.: 2002, "A Survey of Proper-Motion Stars. XVI. Orbital Solutions for 171 Single-lined Spectroscopic Binaries", *Astronomical Journal* **124**, 1144

245. Leavitt, H. S.: 1908, "1777 variables in the Magellanic Clouds", *Annals of Harvard College Observatory* **60**, 87

246. Leavitt, H. S.: 1912, "Periods of 25 Variable Stars in the Small Magellanic Cloud", *Harvard College Observatory Circular* **173**, 1

247. Lebreton, Y. and Goupil, M. J.: 2014, "Asteroseismology for "à la carte" stellar age-dating and weighing. Age and mass of the CoRoT exoplanet host HD 52265", *Astronomy & Astrophysics* **569**, A21

248. Ledoux, P. and Walraven, T.: 1958, "Variable Stars", *Handbuch der Physik* **51**, 353

249. Leibacher, J. W. and Stein, R. F.: 1971, "A New Description of the Solar Five-Minute Oscillation", *Astrophysics Letters* **7**, 191

250. Leighton, R. B., Noyes, R. W., and Simon, G. W.: 1962, "Velocity Fields in the Solar Atmosphere. I. Preliminary Report", *The Astrophysical Journal* **135**, 474

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

252. Louppe, G.: 2014, "Understanding Random Forests: From Theory to Practice", *Ph.D. thesis*, University of Liege, Belgium

253. Lund, M. N., Kjeldsen, H., Christensen-Dalsgaard, J., Handberg, R., and Silva Aguirre, V.: 2014, "Detection of $\ell = 4$ and $\ell = 5$ Modes in 12 Years of Solar VIRGO-SPM Data—Tests on Kepler Observations of 16 Cyg A and B", *The Astrophysical Journal* **782**, 2

254. Lund, M. N., Silva Aguirre, V., Davies, G. R., et al.: 2017, "Standing on the Shoulders of Dwarfs: the Kepler Asteroseismic LEGACY Sample. I. Oscillation Mode Parameters", *The Astrophysical Journal* **835**, 172

255. Lynden-Bell, D. and Ostriker, J. P.: 1967, "On the stability of differentially rotating bodies", *Monthly Notices of the Royal Astronomical Society* **136**, 293

256. Mahalanobis, P. C.: 1936, "On the generalized distance in statistics", *Proceedings of the National Institute of Sciences (Calcutta)* **2**, 49

257. Maldonado, J., Villaver, E., and Eiroa, C.: 2013, "The metallicity signature of evolved stars with planets", *Astronomy & Astrophysics* **554**, A84

258. Mamajek, E. E., Prsa, A., Torres, G., et al.: 2015, "IAU 2015 Resolution B3 on Recommended Nominal Conversion Constants for Selected Solar and Planetary Properties", *ArXiv e-prints*

259. Marcy, G. W., Isaacson, H., Howard, A. W., et al.: 2014, "Masses, Radii, and Orbits of Small Kepler Planets: The Transition from Gaseous to Rocky Planets", *The Astrophysical Journal Supplement Series* **210**, 20

260. 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

261. Mathur, S., García, R. A., Régulo, C., et al.: 2010, "Determining global parameters of the oscillations of solar-like stars", *Astronomy & Astrophysics* **511**, A46

262. Mathur, S., Metcalfe, T. S., Woitaszek, M., et al.: 2012, "A Uniform Asteroseismic Analysis of 22 Solar-type Stars Observed by Kepler", *The Astrophysical Journal* **749**, 152

263. Mazumdar, A., Monteiro, M. J. P. F. G., Ballot, J., et al.: 2014, "Measurement of Acoustic Glitches in Solar-type Stars from Oscillation Frequencies Observed by Kepler", *The Astrophysical Journal* **782**, 18

264. McKinney, W.: 2010, "Data structures for statistical computing in Python" in *Proceedings of the 9th Python in Science Conference*, Vol. 445, pp 51–56

265. Metcalfe, T. S., Chaplin, W. J., Appourchaux, T., et al.: 2012, "Asteroseismology of the Solar Analogs 16 Cyg A and B from Kepler Observations", *The Astrophysical Journal Letters* **748**, L10

266. Metcalfe, T. S., Creevey, O. L., and Christensen-Dalsgaard, J.: 2009, "A Stellar Model-fitting Pipeline for Asteroseismic Data from the Kepler Mission", *The Astrophysical Journal* **699**, 373

267. Metcalfe, T. S., Creevey, O. L., and Davies, G. R.: 2015, "Asteroseismic Modeling of 16 Cyg A and B using the Complete Kepler Data Set", *The Astrophysical Journal Letters* **811**, L37

268. Metcalfe, T. S., Creevey, O. L., Doğan, G., et al.: 2014, "Properties of 42 Solar-type Kepler Targets from the Asteroseismic Modeling Portal", *The Astrophysical Journal Supplement Series* **214**, 27

269. Michell, J.: 1759, "LV. Conjectures concerning the cause, and observations upon the phænomena of earthquakes; particularly of that great earthquake of the first November, 1755, which proved so fatal to the city of Lisbon, and whose effects were felt as far as africa and more or less throughout almost all Europe; by the Reverend John Michell, M. A. Fellow of Queen's College, Cambridge", *Philosophical Transactions* **51**, 566

270. Michell, J.: 1767, "An Inquiry into the Probable Parallax, and Magnitude of the Fixed Stars, from the Quantity of Light Which They Afford us, and the Particular Circumstances of Their Situation, by the Rev. John Michell, B. D. F. R. S.", *Philosophical Transactions (1683-1775)* **57**, 234

271. Miglio, A., Chiappini, C., Morel, T., et al.: 2013, "Galactic archaeology: mapping and dating stellar populations with asteroseismology of red-giant stars", *Monthly Notices of the Royal Astronomical Society* **429**, 423

272. Mohr, P. J., Newell, D. B., and Taylor, B. N.: 2016, "CODATA recommended values of the fundamental physical constants: 2014*", *Reviews of Modern Physics* **88 (3)**, 035009

273. Morel, P. and Thévenin, F.: 2002, "Atomic diffusion in star models of type earlier than G", *Astronomy & Astrophysics* **390**, 611

274. Mosser, B. and Appourchaux, T.: 2009, "On detecting the large separation in the autocorrelation of stellar oscillation times series", *Astronomy & Astrophysics* **508**, 877

275. Mosser, B., Elsworth, Y., Hekker, S., et al.: 2012, "Characterization of the power excess of solar-like oscillations in red giants with Kepler", *Astronomy & Astrophysics* **537**, A30

276. Mosser, B., Michel, E., Belkacem, K., et al.: 2013, "Asymptotic and measured large frequency separations", *Astronomy & Astrophysics* **550**, A126

277. Murtagh, F. and Heck, A. (eds.): 1987, "Multivariate Data Analysis", Vol. 131 of *Astrophysics and Space Science Library*

278. Nelder, J. A. and Mead, R.: 1965, "A simplex method for function minimization", *The Computer Journal* **7 (4)**, 308

279. Neto, F. D. M. and Neto, A. J. d. S.: 2012, "An introduction to inverse problems with applications", Springer Science & Business Media

280. Neuwirth, E.: 2014, "RColorBrewer: ColorBrewer Palettes", R package version 1.1-2

281. Newton, I.: 1686, "PhilosophiæNaturalis Principia Mathematica", English: "The Mathematical Principles of Natural Philosophy"

282. Nidever, D. L., Marcy, G. W., Butler, R. P., Fischer, D. A., and Vogt, S. S.: 2002, "Radial Velocities for 889 Late-Type Stars", *The Astrophysical Journal Supplement Series* **141**, 503

283. Nimtz, G. and Clegg, B.: 2009, "Tunneling", Springer Berlin Heidelberg

284. Öpik, E.: 1938, "Stellar Structure, Source of Energy, and Evolution", *Publications of the Tartu Astrofizica Observatory* **30**, C1

285. O'Sullivan, F., Yandell, B. S., and Raynor Jr, W. J.: 1986, "Automatic smoothing of regression functions in generalized linear models", *Journal of the American Statistical Association* **81 (393)**, 96

286. Ovid: 8 AD, "Metamorphoseon libri", English: "Metamorphoses"

287. Pál, A., Bakos, G. Á., Torres, G., et al.: 2008, "HAT-P-7b: An Extremely Hot Massive Planet Transiting a Bright Star in the Kepler Field", *The Astrophysical Journal* **680**, 1450

288. Paxton, B., Bildsten, L., Dotter, A., et al.: 2011, "Modules for Experiments in Stellar Astrophysics (MESA)", *The Astrophysical Journal Supplement Series* **192**, 3

289. Paxton, B., Cantiello, M., Arras, P., et al.: 2013, "Modules for Experiments in Stellar Astrophysics (MESA): Planets, Oscillations, Rotation, and Massive Stars", *The Astrophysical Journal Supplement Series* **208**, 4

290. Paxton, B., Marchant, P., Schwab, J., et al.: 2015, "Modules for Experiments in Stellar Astrophysics (MESA): Binaries, Pulsations, and Explosions", *The Astrophysical Journal Supplement Series* **220**, 15

291. Paxton, B., Schwab, J., Bauer, E. B., et al.: 2018, "Modules for Experiments in Stellar Astrophysics (MESA): Convective Boundaries, Element Diffusion, and Massive Star Explosions", *The Astrophysical Journal Supplement Series* **234**, 34

292. Payne, C. H.: 1925, "Stellar Atmospheres; a Contribution to the Observational Study of High Temperature in the Reversing Layers of Stars", *Ph.D.*

thesis, Radcliffe College

293. Pedregosa, F., Varoquaux, G., Gramfort, A., et al.: 2011, "Scikit-learn: Machine Learning in Python", *Journal of Machine Learning Research* **12**, 2825

294. Pekeris, C. L.: 1938, "Nonradial Oscillations of Stars.", *The Astrophysical Journal* **88**, 189

295. Pigott, E.: 1785, "Observations of a New Variable Star. In a Letter from Edward Pigott, Esq. to Sir H. C. Englefield, Bart. F. R. S. and A. S.", *Philosophical Transactions of the Royal Society of London Series I* **75**, 127

296. Pigott, E.: 1786, "Observations and Remarks on Those Stars Which the Astronomers of the Last Century Suspected to be Changeable. By Edward Pigott, Esq.; Communicated by Sir Henry C. Englefield, Bart. F. R. S. and A. S.", *Philosophical Transactions of the Royal Society of London* **76**, 189

297. Pijpers, F. P. and Thompson, M. J.: 1992, "Faster formulations of the optimally localized averages method for helioseismic inversions", *Astronomy & Astrophysics* **262**, L33

298. Pijpers, F. P. and Thompson, M. J.: 1994, "The SOLA method for helioseismic inversion", *Astronomy & Astrophysics* **281**, 231

299. Pinsonneault, M. H., An, D., Molenda-Żakowicz, J., et al.: 2012, "A Revised Effective Temperature Scale for the Kepler Input Catalog", *The Astrophysical Journal Supplement Series* **199**, 30

300. Pitjeva, E.: 2015, "Determination of the Value of the Heliocentric Gravitational Constant ($GM_\odot$) from Modern Observations of Planets and Spacecraft", *Journal of Physical and Chemical Reference Data* **44 (3)**, 031210

301. Plaskett, H. H.: 1916, "A Variation in the Solar Rotation", *The Astrophysical Journal* **43**, 145

302. Plummer, H. G.: 1914, "Note on the velocity of light and Doppler's principle", *Monthly Notices of the Royal Astronomical Society* **74**, 660

303. Pols, O. R.: 2011, "Stellar Structure and Evolution", Astronomical Institute Utrecht

304. Pottasch, E. M., Butcher, H. R., and van Hoesel, F. H. J.: 1992, "Solar-like oscillations on Alpha Centauri A", *Astronomy & Astrophysics* **264**, 138

305. Prato, M. and Zanni, L.: 2008, "Inverse problems in machine learning: an application to brain activity interpretation" in *Journal of Physics: Conference Series*, Vol. 135, p. 012085

306. Pulone, L. and Scaramella, R.: 1997, "Age estimates of stellar systems by Artificial Neural Networks", in *Neural Nets WIRN VIETRI-96*, pp 231–236, Springer

307. Quirion, P.-O., Christensen-Dalsgaard, J., and Arentoft, T.: 2010, "Automatic Determination of Stellar Parameters Via Asteroseismology of

Stochastically Oscillating Stars: Comparison with Direct Measurements", *The Astrophysical Journal* **725**, 2176

308. R Core Team: 2014, "R: A Language and Environment for Statistical Computing", R Foundation for Statistical Computing
309. Rabello-Soares, M. C., Basu, S., and Christensen-Dalsgaard, J.: 1998, "A Study of the Parameters for Solar Structure Inversion Methods" in S. Korzennik (ed.) *Structure and Dynamics of the Interior of the Sun and Sun-like Stars*, Vol. 418 of *ESA Special Publication*, p. 505
310. Rabello-Soares, M. C., Basu, S., and Christensen-Dalsgaard, J.: 1999, "On the choice of parameters in solar-structure inversion", *Monthly Notices of the Royal Astronomical Society* **309**, 35
311. Ramírez, I., Meléndez, J., and Asplund, M.: 2009, "Accurate abundance patterns of solar twins and analogs. Does the anomalous solar chemical composition come from planet formation?", *Astronomy & Astrophysics* **508**, L17
312. Rauer, H., Catala, C., Aerts, C., et al.: 2014, "The PLATO 2.0 mission", *Experimental Astronomy* **38**, 249
313. Reese, D., Zharkov, S., and Buldgeon, G.: 2014, "InversionKit"
314. Reese, D. R.: 2018, "Stellar Inversion Techniques", *Asteroseismology and Exoplanets: Listening to the Stars and Searching for New Worlds* **49**, 75
315. Reese, D. R., Chaplin, W. J., Davies, G. R., et al.: 2016, "SpaceInn hare-and-hounds exercise: Estimation of stellar properties using space-based asteroseismic data", *Astronomy & Astrophysics* **592**, A14
316. Reese, D. R., Marques, J. P., Goupil, M. J., Thompson, M. J., and Deheuvels, S.: 2012, "Estimating stellar mean density through seismic inversions", *Astronomy & Astrophysics* **539**, A63
317. Renzini, A., Greggio, L., Ritossa, C., and Ferrario, L.: 1992, "Why stars inflate to and deflate from red giant dimensions", *The Astrophysical Journal* **400**, 280
318. Rhodes, Jr., E. J., Kosovichev, A. G., Schou, J., Scherrer, P. H., and Reiter, J.: 1997, "Measurements of Frequencies of Solar Oscillations from the MDI Medium-l Program", *Solar Physics* **175**, 287
319. Rhodes, Jr., E. J., Ulrich, R. K., and Simon, G. W.: 1977, "Observations of nonradial p-mode oscillations on the sun", *The Astrophysical Journal* **218**, 901
320. Ricker, G. R., Latham, D. W., Vanderspek, R. K., et al.: 2010, "Transiting Exoplanet Survey Satellite (TESS)" in *American Astronomical Society Meeting Abstracts #215*, Vol. 42 of *Bulletin of the American Astronomical Society*, p. 459

321. Ricker, G. R., Winn, J. N., Vanderspek, R., et al.: 2015, "Transiting Exoplanet Survey Satellite (TESS)", *Journal of Astronomical Telescopes, Instruments, and Systems* **1 (1)**, 014003

322. Ritter, G. A. D.: 1880, "Untersuchungen über die Höhe der Atmosphäre und die Constitution gasförmiger Weltkörper", *Wiedemann Annalen*

323. Robotham, A.: 2015, "magicaxis: Pretty Scientific Plotting with Minor-Tick and log Minor-Tick Support", R package version 1.9.4

324. Robotham, A.: 2016, "magicaxis: Pretty Scientific Plotting with Minor-Tick and log Minor-Tick Support", R package version 2.0.0

325. Rogers, F. J. and Nayfonov, A.: 2002, "Updated and Expanded OPAL Equation-of-State Tables: Implications for Helioseismology", *The Astrophysical Journal* **576**, 1064

326. Rosasco, L., Caponnetto, A., Vito, E. D., Odone, F., and Giovannini, U. D.: 2005, "Learning, regularization and ill-posed inverse problems" in *Advances in Neural Information Processing Systems*, pp 1145–1152

327. Rosenthal, C. S., Christensen-Dalsgaard, J., Nordlund, Å., Stein, R. F., and Trampedach, R.: 1999, "Convective contributions to the frequencies of solar oscillations", *Astronomy & Astrophysics* **351**, 689

328. Rosseland, S.: 1949, "The Pulsation Theory of Variable Stars", Princeton University Observatory

329. Roxburgh, I. W. and Vorontsov, S. V.: 2003, "The ratio of small to large separations of acoustic oscillations as a diagnostic of the interior of solar-like stars", *Astronomy & Astrophysics* **411**, 215

330. Russell, H. N.: 1913a, ""Giant" and "dwarf" stars", *The Observatory* **36**, 324

331. Russell, H. N.: 1913b, "Notes on the Real Brightness of Variable Stars", *Science* **37**, 651

332. Russell, H. N.: 1914, "Relations Between the Spectra and other Characteristics of the Stars. II. Brightness and Spectral Class", *Nature* **93**, 252

333. Salaris, M. and Cassisi, S.: 2005, "Evolution of Stars and Stellar Populations", Wiley

334. Samadi, R., Belkacem, K., and Sonoi, T.: 2015, "Stellar oscillations - II - The non-adiabatic case" in *EAS Publications Series*, Vol. 73 of *EAS Publications Series*, pp 111–191

335. Sampson, R. A.: 1895, "On the Rotation and Mechanical State of the Sun", *MmRAS* **51**, 123

336. Samus', N. N., Kazarovets, E. V., Durlevich, O. V., Kireeva, N. N., and Pastukhova, E. N.: 2017, "General catalogue of variable stars: Version GCVS 5.1", *Astronomy Reports* **61 (1)**, 80

337. Schatzman, E. L.: 1958, "White dwarfs", Interscience

338. Schlemper, J., Caballero, J., Hajnal, J. V., Price, A., and Rueckert, D.: 2017, "A Deep Cascade of Convolutional Neural Networks for MR Image Reconstruction", *ArXiv e-prints*

339. Schloerke, B., Crowley, J., Cook, D., et al.: 2014, "GGally: Extension to ggplot2", R package version 0.5.0

340. Schmitt, J. R. and Basu, S.: 2015, "Modeling the Asteroseismic Surface Term across the HR Diagram", *The Astrophysical Journal* **808**, 123

341. Schou, J., Antia, H. M., Basu, S., et al.: 1998, "Helioseismic Studies of Differential Rotation in the Solar Envelope by the Solar Oscillations Investigation Using the Michelson Doppler Imager", *The Astrophysical Journal* **505**, 390

342. Schou, J. and Buzasi, D. L.: 2001, "Observations of p-modes in α Cen" in A. Wilson and P. L. Pallé (eds.) *SOHO 10/GONG 2000 Workshop: Helio- and Asteroseismology at the Dawn of the Millennium*, Vol. 464 of *ESA Special Publication*, pp 391–394

343. Schwarzschild, K.: 1906, "Über das Gleichgewicht der Sonnenatmosphäre", *Nachrichten von der Gesellschaft der Wissenschaften zu Göttingen, Mathematisch-Physikalische Klasse* **1906**, 41

344. Schwarzschild, M.: 1958, "Structure and evolution of the stars.", Princeton University Press

345. Secchi, P. A.: 1877, "Le Stelle", English: "The Stars"

346. Serenelli, A., Johnson, J., Huber, D., et al.: 2017, "The First APOKASC Catalog of Kepler Dwarf and Subgiant Stars", *The Astrophysical Journal Supplement Series* **233**, 23

347. Shakespeare, W.: 1599, "The Tragedy of Julius Caesar", The First Folio

348. Shapley, H.: 1914, "On the Nature and Cause of Cepheid Variation", *The Astrophysical Journal* **40**, 448

349. Shapley, H.: 1918, "Studies based on the colors and magnitudes in stellar clusters. VI. On the determination of the distances of globular clusters", *The Astrophysical Journal* 48

350. Shapley, H. and Curtis, H. D.: 1921, "The Scale of the Universe", *Bulletin of the National Research Council, Vol. 2, Part 3, No. 11* **2**, 171

351. Sharma, S., Stello, D., Bland-Hawthorn, J., Huber, D., and Bedding, T. R.: 2016, "Stellar Population Synthesis Based Modeling of the Milky Way Using Asteroseismology of 13,000 Kepler Red Giants", *The Astrophysical Journal* **822**, 15

352. Silva Aguirre, V., Davies, G. R., Basu, S., et al.: 2015, "Ages and fundamental properties of Kepler exoplanet host stars from asteroseismology", *Monthly Notices of the Royal Astronomical Society* **452**, 2127

353. Silva Aguirre, V., Lund, M. N., Antia, H. M., et al.: 2017, "Standing on the Shoulders of Dwarfs: the Kepler Asteroseismic LEGACY Sample. II. Radii, Masses, and Ages", *The Astrophysical Journal* **835**, 173

354. Singer, S. and Singer, S.: 1999, "Complexity analysis of Nelder-Mead search iterations" in *Proceedings of the 1. Conference on Applied Mathematics and Computation*, pp 185–196, PMF–Matematički odjel, Zagreb

355. Skurichina, M. and Duin, R. P. W.: 2002, "Bagging, Boosting and the Random Subspace Method for Linear Classifiers", *Pattern Analysis & Applications* **5 (2)**, 121

356. Sobol, I. M.: 1967, "On the distribution of points in a cube and the approximate evaluation of integrals", *USSR Computational mathematics and mathematical physics* **7**, 86

357. Sonoi, T., Samadi, R., Belkacem, K., et al.: 2015, "Surface-effect corrections for solar-like oscillations using 3D hydrodynamical simulations. I. Adiabatic oscillations", *Astronomy & Astrophysics* **583**, A112

358. Spiegel, E. A. and Zahn, J.-P.: 1992, "The solar tachocline", *Astronomy & Astrophysics* **265**, 106

359. Spruit, H. C., Nordlund, A., and Title, A. M.: 1990, "Solar Convection", *Annual Review of Astronomy and Astrophysics* **28**, 263

360. Stello, D., Chaplin, W. J., Basu, S., Elsworth, Y., and Bedding, T. R.: 2009a, "The relation between $\Delta\nu$ and ν_{max} for solar-like oscillations", *Monthly Notices of the Royal Astronomical Society* **400**, L80

361. Stello, D., Chaplin, W. J., Bruntt, H., et al.: 2009b, "Radius Determination of Solar-type Stars Using Asteroseismology: What to Expect from the Kepler Mission", *The Astrophysical Journal* **700**, 1589

362. Sterne, T. E.: 1938, "The Secondary Variation of δ Scuti", *The Astrophysical Journal* **87**, 133

363. Sterne, T. E.: 1940, "A Note on the Variation of delta Scuti", *Proceedings of the National Academy of Science* **26**, 537

364. Sugimoto, D. and Fujimoto, M. Y.: 2000, "Why Stars Become Red Giants", *The Astrophysical Journal* **538**, 837

365. Tassoul, M.: 1980, "Asymptotic approximations for stellar nonradial pulsations", *The Astrophysical Journal Supplement Series* **43**, 469

366. Tenorio, L.: 2001, "Statistical Regularization of Inverse Problems", *SIAM Review* **43 (2)**, 347

367. Therneau, T.: 2014, "deming: Deming, Thiel-Sen and Passing-Bablock Regression", R package version 1.0-1

368. Thompson, M. J.: 1993, "Seismic Investigation of the Sun's Internal Structure and Rotation" in T. M. Brown (ed.) *GONG 1992. Seismic Investigation of the Sun and Stars*, Vol. 42 of *Astronomical Society of the Pacific Conference*

Series, p. 141

369. Thompson, M. J.: 2000, private communication

370. Thompson, M. J. and Christensen-Dalsgaard, J.: 2002, "On inverting asteroseismic data" in B. Battrick, F. Favata, I. W. Roxburgh, and D. Galadi (eds.) *Stellar Structure and Habitable Planet Finding*, Vol. 485 of *ESA Special Publication*, pp 95–101

371. Thomson, W.: 1863, "Dynamical Problems Regarding Elastic Spheroidal Shells and Spheroids of Incompressible Liquid", *Philosophical Transactions of the Royal Society of London Series I* **153**, 583

372. Thoul, A. A., Bahcall, J. N., and Loeb, A.: 1994, "Element diffusion in the solar interior", *The Astrophysical Journal* **421**, 828

373. Tikhonov, A. N. and Arsenin, V. Y.: 1977, "Solutions of ill-posed problems", Winston

374. TOP500: 2015, "TOP500 Supercomputer Site"

375. Townsend, R. H. D. and Teitler, S. A.: 2013, "GYRE: an open-source stellar oscillation code based on a new Magnus Multiple Shooting scheme", *Monthly Notices of the Royal Astronomical Society* **435**, 3406

376. Triana, S. A., Corsaro, E., De Ridder, J., et al.: 2017, "Internal rotation of 13 low-mass low-luminosity red giants in the Kepler field", *Astronomy & Astrophysics* **602**, A62

377. Tuv, E., Borisov, A., Runger, G., and Torkkola, K.: 2009, "Feature selection with ensembles, artificial variables, and redundancy elimination", *Journal of Machine Learning Research* **10**, 1341

378. Ulrich, R. K.: 1970, "The Five-Minute Oscillations on the Solar Surface", *The Astrophysical Journal* **162**, 993

379. Ulrich, R. K.: 1986, "Determination of stellar ages from asteroseismology", *The Astrophysical Journal Letters* **306**, L37

380. Unno, W., Osaki, Y., Ando, H., and Shibahashi, H.: 1979, "Nonradial oscillations of stars", University of Tokyo Press

381. Van Der Walt, S., Colbert, S. C., and Varoquaux, G.: 2011, "The NumPy array: a structure for efficient numerical computation", *Computing in Science & Engineering* **13 (2)**, 22

382. Veresoglou, S. D. and Rillig, M. C.: 2015, "Evidence-Based Data Analysis: Protecting the World From Bad Code? Comment by Veresoglou and Rillig", *The American Statistician* **69 (3)**, 257

383. Verma, K., Antia, H. M., Basu, S., and Mazumdar, A.: 2014a, "A Theoretical Study of Acoustic Glitches in Low-mass Main-sequence Stars", *The Astrophysical Journal* **794**, 114

384. Verma, K., Faria, J. P., Antia, H. M., et al.: 2014b, "Asteroseismic Estimate of Helium Abundance of a Solar Analog Binary System", *The Astrophysi-*

cal Journal **790**, 138

385. Verma, K., Hanasoge, S., Bhattacharya, J., Antia, H. M., and Krishnamurthi, G.: 2016, "Asteroseismic determination of fundamental parameters of Sun-like stars using multilayered neural networks", *Monthly Notices of the Royal Astronomical Society* **461**, 4206
386. Verma, K., Raodeo, K., Antia, H. M., et al.: 2017, "Seismic Measurement of the Locations of the Base of Convection Zone and Helium Ionization Zone for Stars in the Kepler Seismic LEGACY Sample", *The Astrophysical Journal* **837**, 47
387. Viani, L. S., Basu, S., Chaplin, W. J., Davies, G. R., and Elsworth, Y.: 2017, "Changing the ν_{max} Scaling Relation: The Need for a Mean Molecular Weight Term", *The Astrophysical Journal* **843**, 11
388. Vito, E. D., Rosasco, L., Caponnetto, A., Giovannini, U. D., and Odone, F.: 2005, "Learning from examples as an inverse problem", *Journal of Machine Learning Research* **6**, 883
389. Vogel, H. C.: 1889, "Über die auf dem Potsdamer Observatorium unternommenen Untersuchungen über die Bewegung der Sterne im Visionsradius vermittelst der spectrographischen Methode", *Astronomische Nachrichten* **121**, 241
390. Walker, G., Matthews, J., Kuschnig, R., et al.: 2003, "The MOST Asteroseismology Mission: Ultraprecise Photometry from Space", *Publications of the Astronomical Society of the Pacific* **115**, 1023
391. Weiss, A.: 1983, "On the evolution to red giants", *Astronomy & Astrophysics* **127**, 411
392. White, T. R., Huber, D., Maestro, V., et al.: 2013, "Interferometric radii of bright Kepler stars with the CHARA Array: θ Cygni and 16 Cygni A and B", *Monthly Notices of the Royal Astronomical Society* **433**, 1262
393. Whitworth, A.: 1991, "Why do stars become giants?", *Annales de Physique* **16**, 515
394. Whitworth, A. P.: 1989, "Why red giants are giant", *Monthly Notices of the Royal Astronomical Society* **236**, 505
395. Wickham, H.: 2015, "scales: Scale Functions for Visualization", R package version 0.3.0
396. Wickham, H.: 2016, "ggplot2: elegant graphics for data analysis", Springer
397. Yahil, A. and van den Horn, L.: 1985, "Why do giants puff up?", *The Astrophysical Journal* **296**, 554
398. Zsoldos, E.: 1994, "Three Early Variable Star Catalogues", *Journal for the History of Astronomy* **25**, 92

Publications

Refereed publications

1. **Bellinger, E. P.**, Basu, S., Hekker, S., & Ball, W.: 2017, "Model-independent Measurement of Internal Stellar Structure in 16 Cygni A and B", *The Astrophysical Journal*, 851 (2), 80

2. **Bellinger, E. P.**, Angelou, G. C., Hekker, S., Basu, S., Ball, W., & Guggenberger, E.: 2016, "Fundamental Parameters of Main-Sequence Stars in an Instant with Machine Learning", *The Astrophysical Journal*, 830 (1), 20

3. Angelou, G. C., **Bellinger, E. P.**, Hekker, S., & Basu, S.: 2017, "On the Statistical Properties of the Lower Main Sequence", *The Astrophysical Journal*, 839 (2) 116 (co-first author)

4. Guggenberger, E., Hekker, S., Basu, S., Angelou, G. C., & **Bellinger, E. P.**: 2017, "Mitigating the mass dependence in the $\Delta\nu$ scaling relation of red-giant stars", *Monthly Notices of the Royal Astronomical Society*, 470 (2)

5. Guggenberger, E., Hekker, S., Basu, S., & **Bellinger, E. P.**: 2016 "Significantly improving stellar mass and radius estimates: A new reference function for the $\Delta\nu$ scaling relation", *Monthly Notices of the Royal Astronomical Society*, 461 (2)

6. Glover, M., **Bellinger, E. P.**, Radivojac, P., & Clemmer, D.: 2015, "Penultimate Proline in Neuropeptides", *Analytical Chemistry*, 87 (16), 8466-8472

7. Ji, C., Li, Y., **Bellinger, E. P.**, Li, S., Arnold, R., Radivojac, P., & Tang, H.: 2015, "A maximum-likelihood approach to absolute protein quantification in mass spectrometry", In refereed proceedings of the *6th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics* (pp. 296-305)

8. Ngeow, C. C., Kanbur, S. M., **Bellinger, E. P.**, Marconi, M., Musella, I., Cignoni, M., & Lin, Y. H.: 2012, "Period-luminosity relations for Cepheid variables: from mid-infrared to multi-phase", *Astrophysics and Space Science*, 341 (1), 105-113

Conference proceedings

1. **Bellinger, E. P.**, Angelou, G., Hekker, S., Basu, S., Ball, W., & Guggenberger, E.: 2017, "Fundamental Parameters in an Instant with Machine Learning: Application to Kepler LEGACY Targets", in *Seismology of the Sun and the Distant Stars*, Vol. 60 of *European Physical Journal Web of Conferences*, p. 05003

2. **Bellinger, E. P.**, Wysocki, D., & Kanbur, S. M.: 2015, "Measuring amplitudes of harmonics and combination frequencies in variable stars", in *Communications from the Konkoly Observatory of the Hungarian Academy of Sciences*, 105

3. **Bellinger, E. P.**, Kanbur, S. M., & Ngeow, C. C.: 2012, "New insights into the Cepheid PL Relation through the use of multiphase relations", in proceedings of the *20th Stellar Pulsations Conference*

4. **Bellinger, E. P.**: 2012, "Multiphase Relations of Magellanic Cloud Cepheids", in proceedings of the *2012 National Conference on Undergraduate Research*

5. **Bellinger, E. P.**, Kanbur, S. M., & Ngeow, C. C.: 2011, "Multiphase Comparison of Period-Luminosity Relations for Magellanic Cloud Cepheids", in proceedings of the *9th Pacific Rim Conference on Stellar Astrophysics*, 451 (311)

6. Hekker, S., Elsworth, Y., Basu, S., & **Bellinger, E. P.**: 2017, "Evolutionary states of red-giant stars from grid-based modelling", in *Seismology of the Sun and the Distant Stars*, Vol. 160 of *European Physical Journal Web of Conferences*, p. 05003

7. Reyner, S., **Bellinger, E. P.**, & Kanbur, S. M.: 2012, "The approximation of RR Lyrae and eclipsing binary light curves using cubic polynomials", in proceedings of the *20th Stellar Pulsations Conference*

Technical reports

1. **Bellinger, E. P.**, Conner, D., Mittman, D., Magee, K., & Heventhal, B.: 2012, "CASSIUS: the Cassini Uplink Scheduler", *JPL: NASA*, hdl:2014/43122

Acknowledgements

This thesis represents the culmination of my, by now, nearly ten-year-long fascination with variable stars, which began way back in my first year of university. It would have been all but impossible to chase this dream without the support of many individuals. I would like now to give thanks to all those who have supported me on this journey.

I would first like to thank my doctoral advisors, Dr. ir. Saskia Hekker and Prof. Dr. Sarbani Basu, for their advice, guidance, and good ideas over the past three years. I appreciate the amount they pushed me to make this thesis what it is, and I look back with amazement at all the things I have been given the opportunity to learn about. I am proud of the hard work that they encouraged from me, and I look forward to continued collaboration in the future.

During my studies, I have had the great fortune of being able to lean on the expertise of two post-docs, Dr. George Angelou and Dr. Warrick Ball. Without their help, I would have surely been stuck in the dark for far longer than I was. I want to especially thank George for teaching me about stellar evolution, and to thank Warrick for teaching me about kernels. I hope we will continue to collaborate long into the future!

Next I want to thank the SAGE Group at the Max Planck Institute for Solar System Research and the Department of Astronomy at Yale University for hosting me over these three years. I have greatly enjoyed my stays, the exchange of ideas, and the numerous friendships that I've made in these places. I want to specifically thank Dr. Andrés García Saravia Ortiz de Montellano and Dr. Timo Reinhold for their valued help with this thesis. I also thank the IMPRS scientific coordinator, Dr. Sonja Schuh, and the staff at both Yale University and the MPS for all their assistance. I especially want to thank the IMPRS Student Group, which makes it easy for anyone from anywhere to fit in and make friends.

Special thanks go to the Director of the Max Planck Institute for Solar System Research, Prof. Dr. Laurent Gizon; the Director of the GWDG, Prof. Dr. Ramin Yahyapour; and the Dean of Computer Science, Prof. Dr. Jens Grabowski for helping me to enroll into the Göttingen Ph.D. Programme in Computer Science. Additionally, I thank the remaining members of the examination board, Prof. Dr. Carsten Damm, Jun. Prof. Dr. Ing. Marcus Baum, and Prof. Dr. Yvonne Elsworth, FRS for agreeing to examine this thesis.

I thank the National Physical Science Consortium for their very generous support in the form of a graduate fellowship over five years of my graduate studies. I also thank Dr. Judith E. Devaney Terrill for selecting me for the NPSC

Fellowship, for hosting me at NIST for two summers, and especially for always encouraging a strong scientific mindset.

I have had the privilege and honor of working with and (co-)supervising several wonderful students over the course of my graduate studies. I want to acknowledge: Felix Ahlborn (now a Ph.D. student at the Max Planck Institute for Astrophysics), Kenny Roffo (now employed at the NASA Jet Propulsion Laboratory and pursuing graduate studies at Johns Hopkins University), Marc Hon (finishing up his Ph.D. at the University of New South Wales in Sydney, Australia), and Alejandra Perea Rojas (in the midsts of applying to prestigious universities). I'm proud of you all - keep up the great work!

At the Max Planck Institute for Solar System Research, we started a band called MegaGauß that practices every Monday evening and provides a much needed reprieve from the sometimes rollarcoaster-like nature of academia. I want to thank everyone who has played and participated over the last three years and over the many gigs we had; this list includes over twenty people! With no guarantee of completeness, the band included Abbey Ingram, Alessandro Cilla, Bastian Proxauf, Carla Wiles, ChiJu Wu, Daniel Maase, David Marshall, Fatima Kahil, Felix Mackebrandt, Hans Huybrighs, Holly Waller, Katja Karmrodt, Kenny Roffo, Nils Gottschling, Robin Thor, Sudharshan Saranathan, Dr. Ankit Barik, Dr. David Martin Belda, Dr. Emanuele Papini, Dr. James Kuszlewicz, Dr. Keaton Bell, Dr. Theodosis Chatzistergos, and Dr. Vera Dobos. Special thanks go out to my "other half" of the rhythm section, Helge Mißbach, without whom there would have been no band!

I want to take this opportunity to thank some of the teachers who have encouraged and inspired me over the years. This list includes my high school English, history, and physics teachers: Mr. Nelson, Mr. Kaufman, Mr. Battisti; and several of my college computer science professors: Prof. Vampola, Prof. Graci, and Prof. Dr. Early.

To my 'cohort' in the IMPRS school, Alessandro Cilla and Fatima Kahil, and to my other graduate student friends as well: best of luck with finishing your studies! To my friend K. Casey Shea, thank you for making this amazing thesis cover design for me! To all of my dear friends whom I have made over these years of study, thank you for making this journey more enjoyable than it certainly could have been. Special thanks go to Carla Wiles, for many things, including her support and her valued opinions on all the aesthetic aspects of this thesis.

I want to thank my family for their unwavering support in my choice to study something as academic as the distant stars. I thank my mother Patricia, my father Paul, my sister Bobbie Lee, her partner Johnny, my brother Sean, my sister-in-law Valentina, my niece Nia, my nephews Rashay and Darius, and my step-parents Ron and Nina.

Last, and certainly not least, I dedicate this thesis to my mentor, Prof. Dr. Shashi M. Kanbur, who has continuously and actively encouraged me over the past decade to pursue my "academic dreams." Thank you, Shashi, for always being there for me, and for showing me the light of variable stars.

Curriculum vitae

Earl Patrick Bellinger

Education

Ph.D. Candidate, Institute of Computer Science, University of Göttingen
International Max Planck Research School for Solar System Science
Fellow of the National Physical Science Consortium

M.Sc. Computer Science, Indiana University Bloomington, USA 2014
Fellow of the National Physical Science Consortium
GPA: 3.95/4.0

B.Sc. Applied Mathematics, SUNY Oswego, NY, USA 2012
B.Sc. Computer Science, *ibid.* 2012
Presidential Scholar
Honors Thesis: *Multiphase Relations of Magellanic Cloud Cepheids*
GPA: 3.81/4.0 (*summa cum laude*, ranked #1 in Computer Science)

Research Positions

Max Planck Institute for Solar System Research *(Germany)* 2015 – 2018
Doctoral Candidate, Stellar Ages & Galactic Evolution Group

Yale University *(USA)* 2016 – 2017
Visiting Assistant in Research, Department of Astronomy

Indiana University *(USA)* 2013 – 2015
Research Assistant, School of Informatics & Computing

NIST Information Technology Laboratory *(USA)* 2013 – 2014
Guest Researcher, Scientific Applications and Visualization Group

National Center of Sciences *(Japan)* 2013
Research Student, National Institute of Informatics

NASA Jet Propulsion Laboratory *(USA)* 2012
SURF Fellow, Cassini Mission to Saturn

Federal University of Alagoas *(Brazil)* 2011
REU Student, Institute of Physics

Federal University of Santa Catarina *(Brazil)* 2010
REU Student, Department of Physics

Teaching Positions

Yale University Spring 2017
Teaching Assistant, Department of Astronomy

University of Göttingen Summer 2016
Assistant, Institute for Astrophysics

Indiana University Fall 2012
Associate Instructor, School of Informatics & Computing

SUNY Oswego Fall 2010
Seminar Leader, Honors Department

Selected Talks

⋆*invited*

⋆**Stellar Astrophysics Centre Seminar** (Aarhus, Denmark) 2018
"*Determining stellar structure with asteroseismology using novel techniques*"

TESS/Kepler Asteroseismic Science Consortium (Aarhus, Denmark) 2018
"*Testing stellar physics with asteroseismic inversions of solar-type stars*"

⋆**Madison Seminar** (University of Wisconsin–Madison, USA) 2017
"*From Starlight to Stellar Ages with Asteroseismology*"

Rocks & Stars II (Max Planck Institute, Göttingen, Germany) 2017
"*The Seismic Structures of Solar-Type Stars*"

ERES-III (Yale University, New Haven, CT, USA) 2017
"*Fundamental Parameters of Exoplanet Host Stars with Asteroseismology*"

⋆**Science Today** (Public talk at SUNY Oswego, NY, USA) 2017
"*A Look Inside the Private Lives of Stars*"

⋆**Red Giant Modeling Workshop** (Göttingen, Germany) 2016
"*Fundamental Stellar Parameters in an Instant with Machine Learning*"

RR Lyrae (Visegrád, Hungary) 2015
"*Resolving Combination Frequency Amplitudes of Multimode Pulsators*"

American Astronomical Society (Seattle, WA, USA) 2015
"*Optimal Model Discovery of Periodic Variable Stars*"

⋆**Delhi Workshop on Variable Stars** (Delhi, India) 2015
"*Calibrating the Cepheid Distances to the Magellanic Clouds*"

⋆**Kerala Workshop on Stellar Astrophysics** (Kerala, India) 2014
"*Automated Supervised Classification of Variable Stars*"

Honors & Awards

Stellar Astrophysics Centre Postdoctoral Fellowship 2018 – 2021
National Physical Science Consortium Graduate Fellowship 2012 – 2017
SUNY Oswego Presidential Scholarship 2008 – 2012
Oebele Van Dyk Outstanding Computer Science Senior Award 2012
SUNY Chancellor's Award 2012
SUNY Oswego Student/Faculty Collaborative Challenge Grant 2011