

UNIVERSITY OF CALGARY

LAD Method for Detecting Variable Stars in Large-Scale Telescopic Survey

by

Charmaine Adelle Thomsen

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES
IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE
DEGREE OF MASTER OF SCIENCE

GRADUATE PROGRAM IN MATHEMATICS AND STATISTICS

CALGARY, ALBERTA

January, 2025

© Charmaine Adelle Thomsen 2025

Abstract

The huge technological leaps in the last few decades in the field of telescopic surveys have necessitated the use of faster and more accurate computational methods. Stellar lightcurves which demonstrate variability are often used to determine inter-galactic distances, predict the age of stellar clusters, and locate dark objects. In this thesis, we propose to use a novel approach to identify these variable-type stars, i.e. applying a least absolute deviations (LAD) regression to a wavelet-based model of the lightcurve data. When comparing our new method to an established method – maximum a posteriori (MAP) weighted regression - I find that the LAD approach is much more conservative and misclassifies many of the variable stars. However, I also find that the MAP method is overly permissive. Thus to take advantage of the strengths of each method, I propose and construct a two-tier voting classifier based on both. This final classifier correctly predicts 70% of the 5412 variable-type test stars while simultaneously classifying 82% of the non-variable-type test stars correctly.

Preface

This thesis is an original work by the author. No part of this thesis has been previously published.

Acknowledgements

I have been incredibly blessed to be surrounded by a whole community of people who have been encouraging me, helping, feeding me, and just generally keeping me going. This has been an unexpectedly crazy long journey, and I am so grateful to each and every one of you.

Dr. Jingjing Wu, thank you, thank you, thank you. When you accepted me as your student neither one of us could have imagined the many years ahead. As my health has waxed and waned, you have seen me at my worst and yet you have continued to support and encourage me. I will be forever grateful for your patience, thoughtfulness, kind words, and understanding. It has been an honour to watch your career grow over the years I've been your student. You work incredibly hard and you deserve every accolade you receive. I continue to be inspired by you as both an academic and as a woman. I will only make it across the grad school finish line because of you.

Sean, husband and love of my life, I can't even list the numerous ways you have supported me to help me get this far. Thanks for playing this two-player game with me and I'm excited to see what this next chapter of our life has in store.

Deb and Royce, you gave me a home when I needed it badly, the support I needed when my health collapsed, made me part of your family, and taught me so many valuable lessons about life. There is no way for me to fully express my gratitude for everything you have done for me. Thank you for all the love and memories and I look forward to making many more with you.

Taylor, Yue and Yilan, it has been an honour doing grad school with you. You are all such amazing, smart women and it's been so fun getting to know you. Thank you for all the vent sessions, the homework help, the shopping adventures, and everything in between.

Thank you to my first family, Momsers and Daddo, Kathleen, Justin, Ian and Chris for your support and encouragement and for providing a safe landing place when I couldn't take care of myself. Thank you Leanne for everything. Thanks also to the rest of the Navis and Boonstra clans. I love you all so much.

Thank you to my new family, Gaylene and Robert, Ryan and Natasha, and Michelle, for all the meals, the words of encouragement, and for supporting me even when you weren't quite sure what a thesis was.

Irene and all the other lovely ladies of ASCHA, thank you for teaching me how the world works outside of school, for all the "mother-hen" love, and for releasing me to follow my dreams even when it was inconvenient for you. Jenny, thank you for showing me how to be persistent to get my foot in the door.

Richard, Troy, Deena, and the rest of the team I met through CAP, thank you for encouraging me to finish, for giving me time and space to keep plugging away at my thesis and for all your words of encouragement and advice. I have been inspired and learned so much from each of you.

Thank you to my professors and everyone else I got to know or work with at the University of Calgary. I learned so much from all of you and I am forever grateful for the experience. Additionally, thank you to my committee members Dr. Lu and Dr. Stil for their time and input.

Finally, last but certainly not least, thank you to Akira for being the best goober a woman could want. Thanks for keeping me company while I worked and for all the doggie snuggles.

Table of Contents

Abstract	ii
Preface	iii
Acknowledgements	iv
Table of Contents	vi
List of Figures and Illustrations	viii
List of Tables	ix
List of Symbols, Abbreviations and Nomenclature	x
1 Introduction	1
1.1 Stellar Classification	1
1.1.1 The Main Sequence	2
1.1.2 Variable stars - Cepheids and RR Lyrae	3
1.1.3 Eclipsing binary systems	5
1.1.4 Microlensing	7
1.2 Challenges in Stellar Data Mining	9
1.3 Objectives and Organization of the Thesis	11
2 Large-Scale Telescopic Surveys	12
2.1 MACHO Data	13
2.2 Machine Learning and Wavelet Transform based Classifications	14
3 Methodologies	18
3.1 The Wavelet Model	19
3.2 Review of Methodologies	25
3.2.1 Maximum a posteriori (MAP)	25
3.2.2 Least Absolute Deviations (LAD) Regression	27
3.2.3 EM algorithm	28
3.3 Proposed Methods	29
3.3.1 MAP with EM	29
3.3.2 LAD with EM	31

4	Analysis Results	33
4.1	Description of the Testing Dataset	33
4.2	Comparing MAP and LAD Results	35
4.3	Voting Classifier	36
4.4	Extension to MACHO Full Dataset	40
5	Conclusions	42
	Bibliography	44
A	Wavelets	51
B	Data Tables	54

List of Figures and Illustrations

1.1	Hertzprung-Russell diagram	3
1.2	Four Eclipsing Binary Lightcurves from M31 Galaxy	7
1.3	A geometric representation of gravitational lensing	8
2.1	Example of a Null, Variable and Event Object	15
3.1	A subset of the Symlet-4 wavelet bases	22
3.2	A subset of the Symlet-4 wavelet bases (con.)	23

List of Tables

4.1	Breakdown of the testing data	34
4.2	Results of MAP and LAD classifiers	36
4.3	Results of MAP-LAD classifier	37
4.4	Results of Weighted MAP-LAD classifier	38
4.5	Results of Varying Weighted MAP-LAD classifier	41
B.1	A subset of the variable star data and results	54
B.2	A subset of the non-variable star data and results	58

List of Symbols, Abbreviations and Nomenclature

Term/Symbol	Definition
Cepheid	A group of variable stars who pulsate with a stable period and frequency; often used to determine Galactic distances.
Eclipsing Binary Star System	Two stars orbiting around each other.
Electromagnetic Spectrum, The	The full range of electromagnetic radiation, organized by longest to shortest wavelength corresponding to the lowest to highest frequencies. From longest to shortest wavelength the generally accepted categories are as follows: radio, infrared, visible, ultra-violet, X-ray and Gamma.
EM	Expectation Maximization. An iterative method to find statistical parameters.
Flux	A measure of brightness; specifically the energy received per unit of time per unit area.
Frequency	The number of complete waveforms over a given unit of time; as frequency increases the corresponding wavelength gets shorter. The mathematical inverse of period.
Galactic Bulge	A tightly packed group of stars found at the center of our Galaxy.
iid	independent and identically distributed.
LAD	Least Absolute Deviations. A statistical optimization technique comprised of minimizing the sum of the absolute values of the residuals.

Lightcurve	The graph formed by multiple observations of the same star over time.
LMC	Large Magellanic Cloud.
LRT	Likelihood Ratio Test.
LSST	Legacy Survey of Space and Time.
Luminosity	The absolute amount of radiated electromagnetic energy, or light, radiated per time unit by an astronomical body.
MACHO	MASSive Compact Halo Object. A stellar body that emits little to no radiation and is virtually undetectable by conventional methods.
Magnitude	Also known as radiant flux, this is a measure of brightness; which for historical reasons is measured on a reverse-logarithmic scale.
Main Sequence	The stellar classification to which most stars belong; when the luminosity of all stars are plotted against their colour index (i.e. temperature) these stars will form a strong diagonal band from the upper left to the lower right (see section 1.1.1 and Figure 1.1).
MAP	Maximum A Posteriori. A Bayesian optimization technique which augments the Maximum Likelihood Estimator with a prior distribution.
Mass	The quantity of matter in a physical body, determined by the strength of its gravitational attraction to other bodies.
Microlensing	The brightening effect created when light from a distant object is gravitationally bent around a nearer object similar to the effect of an Einstein Ring, although no complete ring is formed.
MLE	Maximum Likelihood Estimation.
OLS	Ordinary Least Squares.
Period	A length of time over which wavelengths are measured. The inverse of frequency.
RR Lyrae	A class of old stars which pulsate fairly rapidly.

SMC	Small Magellanic Cloud.
Spectra	The pattern of absorption lines in the light emitted by a star. These measurements correlate to the chemical structure and temperature of a body.
Symlet-4	The Least Asymmetric Daubechies 4 wavelet.
Uninformative Prior	A prior, or prior distribution, is an assumed probability density function of a random variable before any data is considered. An uninformative, or diffuse prior, expresses only vague or general information.
Visible Light Spectrum	See the Electromagnetic Spectrum (above).
Wavelet	In essence, a vector of discrete numbers, referred to as wavelet coefficients, that represent some detail of a more complex waveform.
A^T	transpose of vector or matrix A.
$\lfloor a \rfloor$	Floor, which gives the greatest integer less than or equal to a .
$N(\mu, \sigma^2)$	normal distribution with mean μ and standard deviation σ^2 .
t_ν	t -distribution with ν degrees of freedom.
χ_ν^2	chi-squared distribution with ν degrees of freedom.
ϕ	The wavelet basis defined in Sec. 3.1.

Chapter 1

Introduction

Since Galileo first pointed his telescope at the sky in 1609, astronomical insights powered by ever-increasingly powerful telescopes have grown by leaps and bounds. However, none have surpassed the data deluge that astronomers now face with the advent of CCD cameras, first invented in the 1980s. As our data storage and computing power have grown over the past forty years, so have the size of the databases of images now captured by a single telescope. One of the largest such projects is the Vera C. Rubin Observatory, set to conduct the 10-year Legacy Survey of Space and Time (LSST). This telescope is currently under construction on the Cerro Pachon ridge in northern Chile, and is set to begin its scientific observations sometime in 2025. The camera contains over three billion pixels and is expected to generate 20 terabytes of observational data *every night* (Observatory, 2024). This scale of data collection requires faster and more efficient methods of analysis. But what are astronomers looking for? What does this data look like? How does statistics play a role in this type of big-data problem?

1.1 Stellar Classification

For as long as humankind has been observing the heavens, they have created ways to summarize their knowledge through classification systems. Today's astrophysical classifica-

tion landscape is a hodgepodge of different systems that together can be used to describe an individual stellar body. A star's apparent brightness taken together with measurements of its spectra¹ provides the framework for determining a stellar body's temperature, chemical composition, luminosity², distance from the Sun, and information about what is between the Earth and the object.

In this thesis, we use magnitude data collected specifically in the red and blue visible light spectrum to perform classification of a database of stars. More details about this specific dataset utilized for this project will be covered in Chapter 2. But first, a discussion of the specific types of stars this project will be attempting to classify is given below.

1.1.1 The Main Sequence

Somewhere around 90% of the known stars in the universe fall on the Main Sequence (NASA, 2024). These stars generate their energy through the nuclear fusion of hydrogen into helium, and their colour and brightness is determined at least in part by the amount of hydrogen present within the star to burn. Because these stars exist in hydrostatic equilibrium³ they remain stable for millions or even billions of years. Our Sun is an example of a Main Sequence star.

Plotting the surface temperature⁴ of many stars against their luminosity shows why these stars are known as the Main Sequence. This is shown by the diagonal line in Figure 1.1. Main Sequence stars remain on the Main Sequence until they burn all of the hydrogen in their central region and either expand into a Giant due to the overwhelming force of the thermal energy from the core or collapse inwards due to gravity and become a Dwarf. These

¹The pattern of absorption lines in the light emitted by a star. These measurements correlate to the temperature and chemical structure of a body.

²The absolute amount of radiated electromagnetic energy, or light, radiated per time unit by an astronomical object. Although this is never able to be directly measured since a star radiates energy across the entire electromagnetic spectrum and in every direction from a sphere, its magnitude or radiant flux (a measure of brightness; measured on a reverse-logarithmic scale) is observable and is related to luminosity through the inverse square law for light.

³The force of gravity is balanced by the outward pressure of thermal energy from the star's core.

⁴Temperature is directly correlated with spectra.

stars are then considered to have left the Main Sequence, and are in the process of slowly dying.

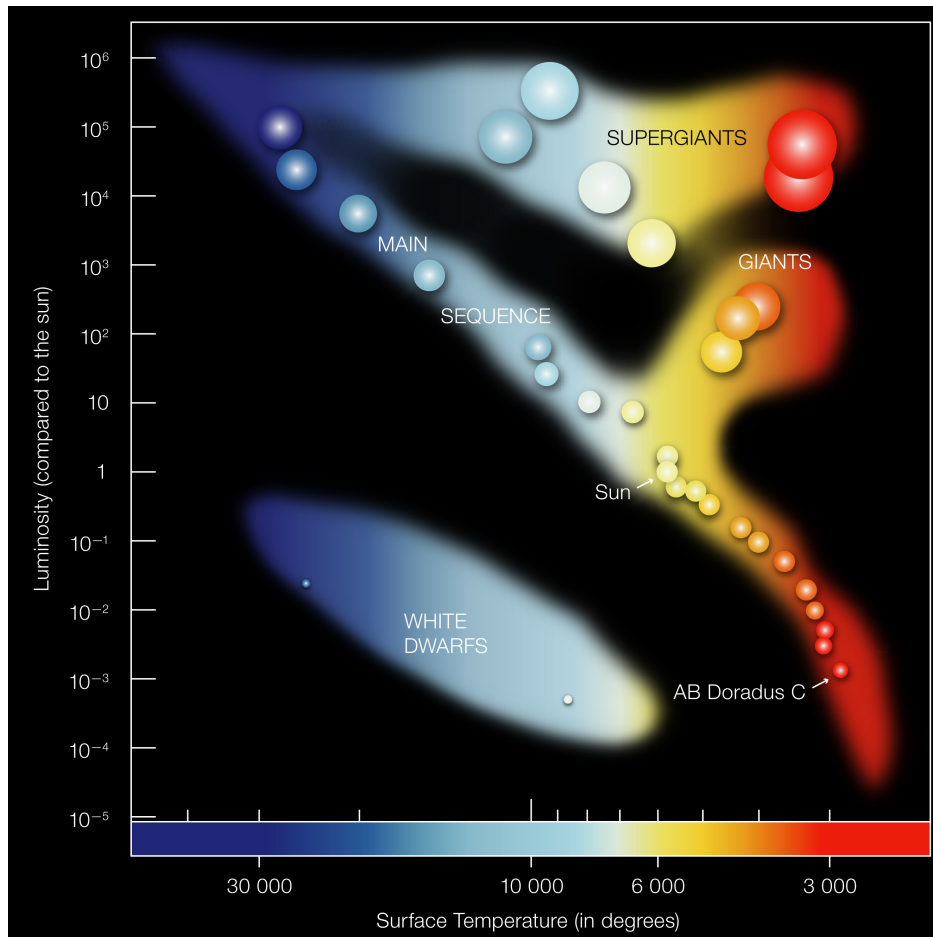


Figure 1.1: **Hertzsprung-Russell diagram.** When surface temperature is plotted against luminosity, it can be seen that all stars which burn hydrogen into helium lie along a diagonal branch called the Main Sequence. This type of graph is called a Hertzsprung-Russell diagram. This figure is taken from ESO (2024). Credit: ESO.

1.1.2 Variable stars - Cepheids and RR Lyrae

In contrast to the constancy of Main Sequence stars, variable stars' flux⁵ as observed from Earth changes over time. Two groups of stars who vary due to external factors will be discussed in the sections below. Here we discuss two groups of stars whose intrinsic luminosity varies periodically over time.

⁵See definition of magnitude.

It is estimated that out of several hundred billion stars, only several million are variable (also known as pulsating) stars (Carroll and Ostlie, 2009). The properties of pulsating stars can be used in the same way that geologists use seismic waves to study the interior of the Earth. Pulsating stars are similar in many ways to Main Sequence stars, except they are not in hydrostatic equilibrium. Instead, an inner layer of the star becomes slowly compressed, in essence damming up the thermal energy pushing to the surface of the star. This in turn pushes the star's surface layers upwards. Eventually the expanding layer becomes less dense and more transparent, and the trapped heat escapes, causing the layer to fall back down to resume the cycle. This process takes the form of a standing sound wave with multiple overtones.

The most well known of the pulsating stars are the Classical Cepheids, referred to here simply as cepheids⁶. These stars are incredibly valuable as tools to determine the sizes and distances in both our own galaxy and the nearby galaxies. This is because there is a direct and well-established relationship between their period of pulsation and their luminosity, allowing for accurate measurement of the distance to the cepheid through the inverse square law for light. Generally cepheids have a period between 1 and 100 days and tend to be very bright and very numerous, making their variability easy to observe by even an amateur astronomer with a telescope (Percy, 2007). Their amplitudes tend to vary from differences of just a few hundredths of a magnitude to over a full magnitude between minima and maxima⁷.

Similar to cepheids, RR Lyrae stars - named after the brightest of their type, RR Lyrae in the Lyra constellation - are useful for determining distances within our Galaxy. They are primarily found in tight groups of stars known as Globular Clusters. These groups are convenient to study because the stars within all have generally a similar age, chemical composition and reddening⁸, and often don't require movement of a telescope to take in the

⁶In fact Classical Cepheids are just the most common type of Cepheid, there are three other known types who appear similar in their lightcurves but differ in their Fourier decompositions or consist of more complex overtones.

⁷As mentioned earlier, magnitudes are measured on a reverse logarithmic scale, making this difference much larger than it sounds

⁸Reddening will be discussed in Section 1.2.

whole cluster. Additionally, RR Lyrae can be found ‘in the field’, that is, the areas of our Galaxy above or below the galactic plane⁹ not distinguished by any other object, whereas cepheids tend to be found primarily on the galactic plane. But other than geographical differences, what distinguishes cepheids from RR Lyrae? They both tend to have stable pulses, although RR Lyrae tend to pulsate much faster than cepheids with a rate of a whole cycle completed in just a day or two. However, cepheids are young, metal-rich stars whereas RR Lyrae are old and nearing the end of their lifespan. They have exhausted the hydrogen in their core and are now burning helium. They are actually so old that most of the heavy metals in the universe had not yet been produced when they were formed. As a group, RR Lyrae tend to be extremely homogenous in temperature, colour and composition. This can be useful when trying to determine distances or reddening¹⁰. RR Lyrae would perhaps claim the title of most useful variable stars, except that they tend to also be very dim when compared to cepheids (although they are generally 100 times more luminous than the Sun) and therefore more difficult to detect. While they have been extensively studied within the Milky Way Galaxy, they have only been detected with great difficulty in a few other near galaxies (Percy, 2007).

1.1.3 Eclipsing binary systems

In contrast to the variable stars described above, the brightening and dimming of the lightcurve of an eclipsing binary system is an artificial construct created by the motion of two or more stars. Although we observe only points in the sky, it turns out that more than half of all stars are actually systems of two or more stars orbiting around a shared center of mass¹¹.

⁹The galactic plane is similar to Earth’s equator - an artificial line cutting horizontally through the center of the Galaxy

¹⁰A phenomena caused by the extinction of light emitted by a star by dust or other materials in the interstellar medium. Light at shorter wavelengths tends to scatter more than those at longer wavelengths, so when blue light is compared to red light more blue than red will be lost to this effect, thus the term reddening.

¹¹Recently a six-star system was discovered, where three pairs of stars orbit each other, and the pairs also orbit the other pairs in a complex geometric dance. (NASA, 2024; Carroll and Ostlie, 2009)

Consider briefly our Solar System. The Sun lies at the center and the planets orbit around it. The center of each planet's orbit lies at the point in space where the Sun is located. But what if one of the planets had a size and mass¹² similar to the Sun? We would then expect that the Sun and this huge planet would rotate *around each other*, and the center of their orbits would not lie at the Sun or at the planet, but at some point in space between them. If they were the exact same mass, following Newton's gravitational laws they would orbit each other with the center of mass, i.e. the center of their orbit, equidistant from each.

This is the situation with a binary system. Two stars orbit each other, with the center of mass of the system somewhere in the space between them. If an observer on Earth was very close to these stars or if the stars were sufficiently far apart, they could watch as one passed in front of the other, similar to how we can occasionally observe planetary transits of the sun or solar eclipses. There are in fact some systems where this is possible - we can observe both stars and watch them pass in front and then behind each other. This is called a visual binary system. The brightest star in the sky, Sirius A, and its companion, Sirius B, fall in this group. However, in general this is not the case. Most stars are simply too far away, too dim, too small, or there can be a host of other reasons that make it too difficult to directly observe the binary system.

For these systems, the lightcurve of the "star" we observe must be examined to determine if it has a companion. When the binary pair's orbital plane is oriented along the line of sight of an observer on Earth, it is possible to observe the effect of the stars eclipsing despite no ability to resolve the viewed object into its members. Consider a solar eclipse. As the moon passes in front of the Sun, the light from the Sun is dimmed for a few minutes as the moon covers part or all of the Sun. The exact same phenomena can be observed in the light curve of an eclipsing binary. Four examples of eclipsing binary light curves are shown in Figure 1.2. Each of the dips in the curve represent a star passing in front of the other. In general, the stars of a binary system are not the same size or mass, and so when the smaller star passes

¹²The quantity of matter in a physical body, determined by the strength of its gravitational attraction to other bodies.

in front of the larger, the amount of light that is cut out is smaller than the amount which is reduced when the larger passes in front of the smaller. After the eclipse ends the stars are viewed side-by-side and the maximum amount of flux will be received by the observer.

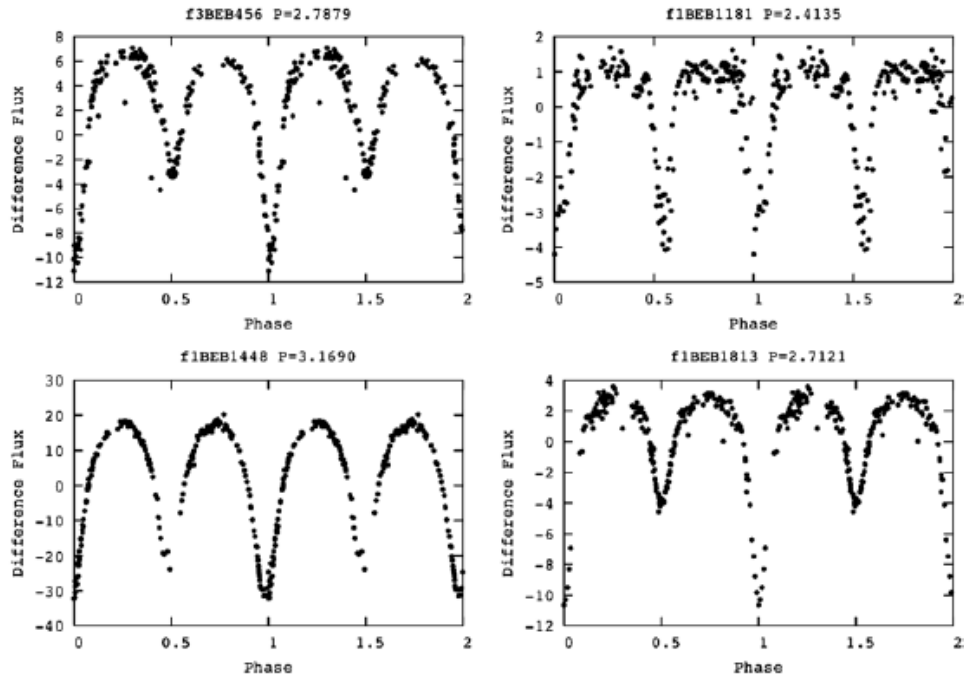


Figure 1.2: **Four Eclipsing Binary Lightcurves from M31 Galaxy.** The larger dips are the larger star passing in front of the smaller, and the smaller dips are the smaller star passing in front of the larger. As the stars separate and the eclipse ends, the system returns to its peak brightness. This figure is taken from Todd et al. (2006).

1.1.4 Microlensing

Perhaps the strangest artificial change to a lightcurve occurs during a microlensing event. The concept of gravitational lensing was first posited as early as 1784, but it was not proven until Einstein published his work on general relativity. A version of this lensing effect was famously used to prove Einstein's theories during the transit of Venus in 1919.

When observable, gravitational lensing is important as it can help astronomers measure the mass of distant objects and statistically constrain the number density of mass concentrations which can in turn give a sense of the number of dark objects present in areas like

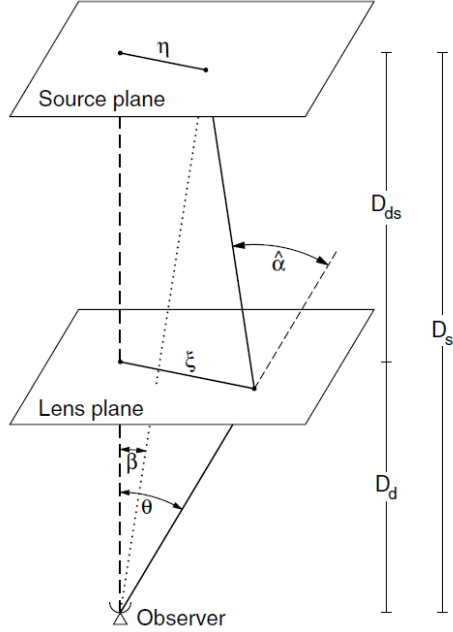


Figure 1.3: **A geometric representation of gravitational lensing.** A simple geometric representation of gravitational lensing. Taken from (Schneider et al., 2006).

the Galactic Bulge at the center of the Milky Way. Finally, gravitational lensing can act as a natural telescope for objects that would otherwise be too faint to detect. Microlensing specifically is often used in the search for planets orbiting distant stars (Schneider et al., 2006).

The idea of gravitational lensing is quite simple - that an object with sufficiently large mass can exert a gravitational effect on a photon of light. A simple representation of gravitational lensing can be seen in Figure 1.3. Here D_s represents the distance from the observer to the source, D_{ds} represents the distance from the source to the lens, and D_d represents the distance from the observer to the lens. Further, η represents the size of the source, which could be anything from a point object like a star¹³ to a full galaxy. Then ξ is the displacement of the source on the plane of the lens and the angle $\hat{\alpha}$ is known as the deflection angle.

When a lensing event occurs, the observer in Figure 1.3 will see a displaced image of

¹³Stars are generally considered to be point objects despite their massive size due to the even more massive distances involved in astrophysical measurements.

the source on the lens plane as the objects move into the correct alignment. Without going into the mathematical details, there comes a moment when the source and lens are perfectly aligned so that the observer will view a ring of images of the source around the lens, similar to the effect of an Einstein ring¹⁴. However, because stellar objects tend to be in motion (albeit slowly from our perspective), this effect will generally last at most for weeks before it vanishes. Additionally, since the movement of these objects with respect to a viewer on Earth is more or less random, perfect alignment may never occur.

There are three generally accepted forms of gravitational lensing - strong, weak and micro. Strong and weak lensing effects are visually observable because of the size of the source, something perhaps on the scale of a galaxy or very close like Venus transiting the Sun. Microlensing occurs when the source is essentially a point object such as a star and the image created is not visually observable due to the very small angles involved. Instead, the observer will see the light from the source appear to brighten as the light bends around the lens. At the moment closest to forming an Einstein ring the object will appear at its brightest. Then it will gradually dim again until it reaches its regular magnitude. This contrasts with the variable stars or eclipsing binaries described above, as this increase in magnitude would only be expected to occur irregularly. If the lens was simply a dark star passing in front of another star, it may happen only once.

1.2 Challenges in Stellar Data Mining

Space is far from empty. The interstellar medium, “space” as it is colloquially known, is full of dust and gas that is affected by explosive events, moved on stellar winds, and shaped by galactic magnetic fields, among other more complex effects. These clouds create a phenomena known as interstellar extinction, where some or all of the wavelengths of light traveling from a star are absorbed or scattered by the cloud. In general, the amount of

¹⁴For more information on this phenomena, P. Schneider’s book (Schneider et al., 2006) provides an excellent description along with a derivation of the equations.

extinction is wavelength dependent. Longer wavelengths will not be scattered as strongly as shorter wavelengths. In the visible range, this means that red light will travel better through a dust cloud than blue light, causing an effect called interstellar reddening (Carroll and Ostlie, 2009). When making observations with a telescope from Earth, this reddening must be checked and corrected for whenever possible.

In addition to the reddening issue, stars that exist in a tight group, such as within a globular cluster or in the Galactic Bulge¹⁵, can be hard to distinguish individually. Air currents in the atmosphere can cause distortions to the light received by a ground based telescope. These issues can insert large amount of noise into a collected dataset.

Whether observations are taken from a ground-based or space-based telescope, the data will necessarily have gaps due to the rotation of the earth, the cycles of the moon, the path of the Earth around the Sun, and so on. Additionally, telescopes can be in high demand and the differing priorities for telescope time can artificially add gaps in the the data. Ground based telescopes face additional issues with bad weather or sky visibility concerns. Since the surveys required to obtain a lightcurve with any kind of meaning must often be collected over a number of years (the MACHO study discussed below studied the same parts of the sky for almost a decade), gaps can even be introduced into the data from telescope maintenance issues or repositioning times.

Finally, the data can be difficult to study with supervised classification methods as developing a training sample can be difficult. Catalogues are large, consisting of millions or even billions of observations. And yet, because they either didn't study the same area of the sky, or because they were taken at different times, combining catalogues can leave a researcher with a much smaller dataset than would be expected based on the size of the original catalogues. Beyond this difficulty, as mentioned several times in the last section, the astrophysical phenomena of interest generally happens very few times in comparison to the number of stars contained within a catalogue.

¹⁵The Galactic Bulge refers to a tightly packed large group of stars at the center of our galaxy.

All of these difficulties taken together suggest that statistical algorithms must be carefully chosen and tuned to obtain useful results. They must be both robust against outliers and gaps in the data, and able to deal with non-Gaussian and heteroscedastic errors. However, they must also be fast, as new telescope projects such as the LSST create new data deluges every night (Huijse et al., 2014).

1.3 Objectives and Organization of the Thesis

In this thesis, I will demonstrate a novel method for analyzing the data of a large-scale telescopic survey with the goal of separating out stars whose light varies over time from ordinary unchanging Main Sequence stars. This method involves utilizing wavelets to analyze the full lightcurve of each star at the end of several years of observations. Pairing this wavelet method with the power of least absolute deviations (LAD) regression forms a new way to computationally examine these lightcurves and quickly suggest which stars are worthy of further study by astrophysicists interested in finding variable stars, eclipsing binaries, or microlensing events.

The rest of the thesis is organized as follows. In the next chapter I provide an introduction to the particular dataset utilized within this thesis. I also provide a brief overview of other research performed in this area. Chapter 3 contains a discussion of the paper which motivated this study, followed by a description of the mathematical model that I used. I present my results in Chapter 4, including a replication of the old method in literature, results from my new method, and a quick comparison between the two. Furthermore, I propose a two-tiered classifier which combines the two methods, and present the results from applying my final method to the complete dataset. Finally in Chapter 5, concluding remarks are provided.

Chapter 2

Large-Scale Telescopic Surveys

The first large-scale telescopic surveys were conducted in the nineties. At that time astrophysicists had calculated the total mass of multiple galaxies, including our own Milky Way. However, the true mass of the galaxy was many times larger than any kind of approximate mass calculated by summing the total mass of all visible objects within the galaxy, giving rise to the concept of dark matter¹. A theory at the time held that a significant fraction of the dark matter in these galaxies was composed of MACHOs, or MAssive Compact Halo Objects, like brown dwarf stars or planets. These objects are best identified by detecting microlensing events, discussed in the previous chapter. But, as mentioned previously, these events are so rare that it became necessary to study many stars over a period of years to obtain a useful detection rate (Welch, 2012). Thus the large-scale survey was born².

Today these surveys are used to determine everything from how galaxies are formed to viewing the deep universe to building a more accurate map of our own Milky Way Galaxy. The LSST described in the previous chapter is the newest iteration of this technique. Most such surveys make their data available to the scientific community to answer questions we have now about the universe, and questions in the future that we haven't even thought to

¹Matter whose presence is discerned by the force of its gravitational attraction rather than its luminosity. Thought to account for about 85% of the matter in the universe.

²These first surveys in fact disproved the idea that MACHOs could make up a large part of the galaxy's dark matter. Instead, the MACHO survey team found that these objects could make up at most 20% of galactic dark matter (Alcock et al., 1996), so the search continues.

ask.

In this chapter I will introduce the MACHO survey, whose data will be used to test my new methodology³. Following that I will review in brief a few of the ways statistics and data mining techniques are currently being used to study large-scale telescopic surveys.

2.1 MACHO Data

The primary dataset used in this thesis was collected by a group called the MACHO Collaboration⁴. The observations were conducted between 1992 and 1999 on a 50-inch telescope mounted with eight CCDs⁵ located at the Mount Stromlo Observatory, Australia. The telescope was fully dedicated to their project and they viewed the sky on every possible clear night, reducing some of the missing data issues outlined in the previous chapter (Alcock et al., 2000).

The dataset consists of observations for more than 70 million stars, split between the Large and Small Magellanic Clouds⁶ (LMC and SMC respectively) and the Galactic Bulge. About 55% of the observations were of the LMC. Data was collected in two visible light colour bands simultaneously, blue and red (Alcock et al., 2001). To collect the data they divided the sky into fields, each 42 by 42 *arcmin*², and then each field was observed approximately once each night, with higher priority fields observed twice ⁷ (Welch, 2012). In total there were about 80 fields in the LMC, 6 fields in the SMC, and 100 fields in the Bulge. I obtained the data from the MACHO database through the Australian National Computing Infrastructure's data repository (Australian National University, 2020).

³Information about this can be found in Chapters 3 and 4.

⁴This paper utilizes public domain data obtained by the MACHO Project, jointly funded by the US Department of Energy through the University of California, Lawrence Livermore National Laboratory under contract No. W-7405-Eng-48, by the National Science Foundation through the Center for Particle Astrophysics of the University of California under cooperative agreement AST-8809616, and by the Mount Stromlo and Siding Spring Observatory, part of the Australian National University.

⁵A charge-coupled device, or CCD, converts photons to electrons with a light-sensitive integrated circuit.

⁶The Large Magellanic Cloud (LMC) and Small Magellanic Cloud (SMC) are two of the closest satellite galaxies - smaller companion galaxies which orbit a more massive host galaxy similar to the way a moon orbits a planet - to the Milky Way.

⁷A minute of arc, or *arcmin*, is a unit used to measure angular separation in the sky. $1^\circ = 60 \text{ arcmin}$.

Examples of three of the primary types of classes found in the MACHO dataset are shown in Figure 2.1. The first is a lightcurve of a non-variable star, the second is a periodic lightcurve for a variable star, while the third contains a microlensing event. Note that both the blue and red light curves are shown since behaviours in one colour band may not match exactly behaviours in the other band. Finally, as mentioned in a footnote in Chapter 1, astrophysicists measure brightness on a reverse logarithmic scale, so the downwards spike in the event object actually corresponds to an increase in brightness.

2.2 Machine Learning and Wavelet Transform based Classifications

With first light of the LSST rapidly approaching, fast and automated approaches to classification of stellar objects is a regular theme in the astrophysical literature. Machine learning approaches are being extensively explored as researchers attempt to grapple with the huge amount of data each new iteration of ground or space based telescope provides.

Support vector machine is a common approach, likely due to its flexibility and relatively straightforward implementation. It has been used to classify stars by spectral type (Zhongbao, 2016; Du et al., 2017), to differentiate stars from quasi-stellar objects (Wang et al., 2022), and - similar to the objectives of this thesis - to identify variable stars (Johnston and Oluseyi, 2017). Random forest is another popular algorithm due to its ability to define clear classes of objects. To classify variable stars, Kim and Bailer-Jones (2016) used a pure random forest method, while Armstrong et al. (2015) combined a random forest with Kohonen Self-Organizing Maps and Zhang et al. (2023) created a voting classifier that combines a Self-Paced Ensemble method with a random forest. Huijse et al. (2015) similarly used a random forest to classify variable stars, but their process is differentiated from the earlier examples in that they use the image files directly. They then extracted features through non-negative matrix factorization, before applying the random forest.

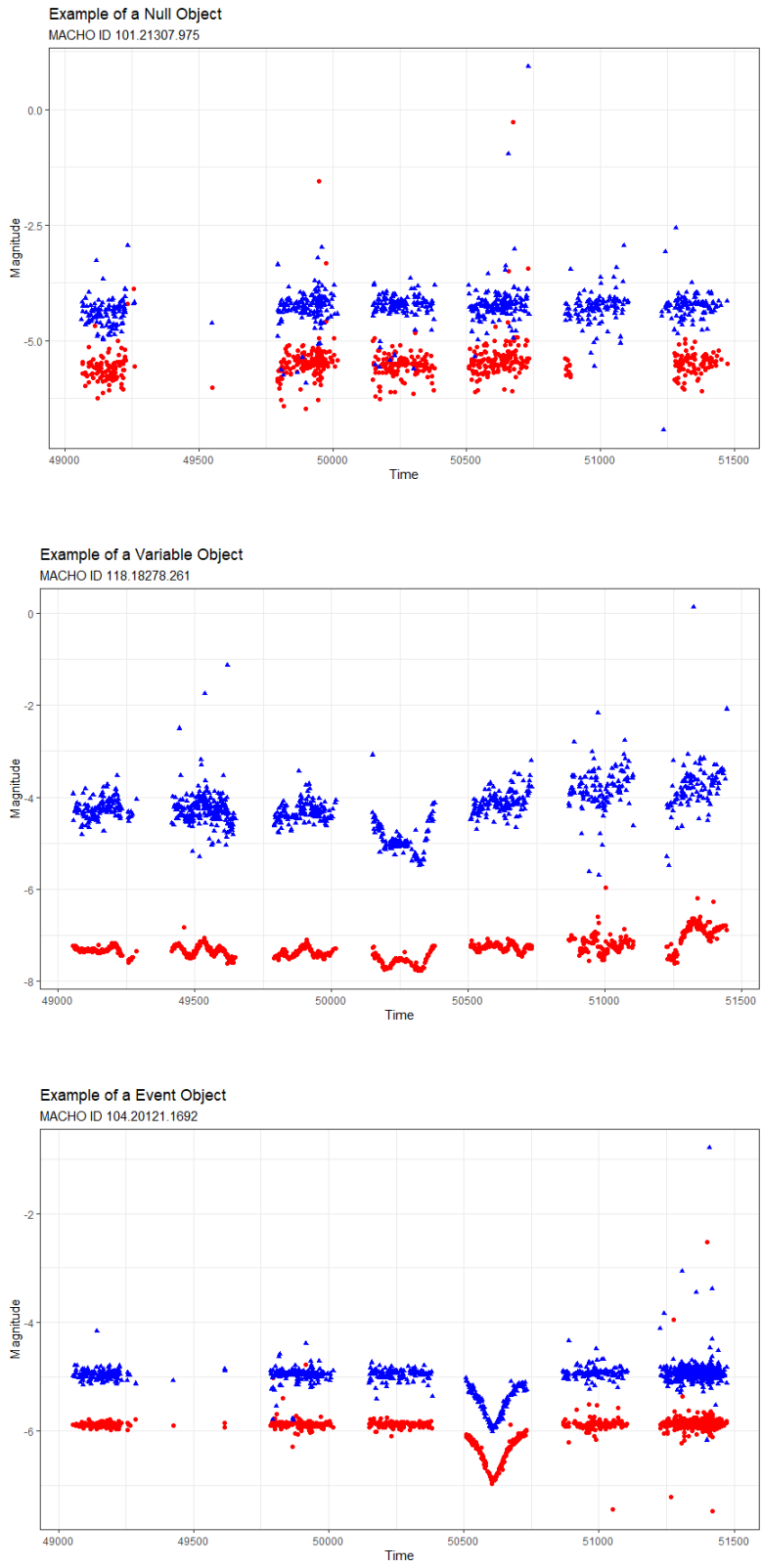


Figure 2.1: Example of a Null, Variable and Event Object

In contrast, Paegert et al. (2014) described a neural net lightcurve classifier that is specific for eclipsing binaries, and Carrasco-Davis et al. (2019) classified more generally variable stars with a recurrent convolutional neural network on sequences of images instead of lightcurves. Convolutional neural networks have also been used to distinguish stars from galaxies at low signal to noise ratios, once again straight from the telescope image files (Stoppa et al., 2023).

Finally, in a bit of a unique example of using machine learning to classify specifically periodic variable stars, Kügler et al. (2015) treated the time series of the lightcurves as a density model and then performed classification on the obtained distance matrix of a mixture of Gaussians which takes into account the measurement uncertainty, also called photometric error. Similarly, Blomme et al. (2011) utilized a multistage tree algorithm where each of the nodes consists of a Gaussian mixture classifier to detect subclasses of variable stars, such as eclipsing binaries or non-radial pulsators.

For a more thorough, albeit already a bit dated, review of the challenges in automated classification and the variety of statistical and machine learning tools that have been applied to these problems, see for example Graham et al. (2017), Babu and Mahabal (2015) and Huijse et al. (2014).

With the exception of the article that motivated the work in this thesis, described in the next chapter, no previous research was found that utilized wavelets to perform classification of variable stars. However, in the astrophysical literature wavelets are a regularly utilized methodology. Edes-Huyal et al. (2021) used wavelet decomposition to address noisy features in the set of multiband simulated lightcurves from the Photometric LSST Astronomical Time-series Classification Challenge (PLAsTiCC). They were then able to follow up with a neural network to perform feature ranking. Wavelet decomposition has also been used to detect gravitational wave events, as seen in Li et al. (2020) where the derived wavelet coefficients are used in a convolutional neural network to detect the existence of damped sinusoid gravitational wave signals that contain white Gaussian noises.

Wavelet transforms have been used recently as part of classification pipelines to reduce the

effect of redshift on supernovae light curves (de Oliveira et al., 2022) and improve detection of active galactic nuclei and classify blazar candidates (Cao et al., 2024). Velicheti et al. (2023) used a wavelet transform to remove instrument noise in order to improve detection of weak lensing events for the soon-to-launch Nancy Grace Roman Space Telescope. Finally, Zaritsky et al. (2021) used wavelet filtering to detect large ultra diffuse galaxies in the Sloan Digital Sky Survey images. These are just a few of the most recent applications of wavelets to the field of astrophysical classification. Wavelet-based classification is not a new concept. For example, Zappalà et al. (1995) utilized a wavelet analysis alongside hierarchical clustering to find asteroid families within our Solar System.

Chapter 3

Methodologies

As mentioned in Chapter 1, astrophysical data presents a unique data analysis problem, which makes the detection of specific phenomena within a large database more complex than many other similar machine learning problems. In addition, due to the necessarily limited length of telescopic observations, an observer can never have the complete waveform. This adds an additional challenge of determining whether the observed event is really an event or simply a globally variable light curve where the variability was only observed once. In this project I will not distinguish between objects who are candidates for a microlensing event vs. some sort of variable star phenomena, and leave any further discernment between curves to the astrophysical experts and future observations.

In this chapter, I will first in Section 3.1 describe generally the approach that motivated my new novel solution to this large data problem. In the following Section 3.2, I will show the details of the methodologies. Finally, in Section 3.3, since the differences between the methods are important but subtle I will compare them through the lens of the Expectation Maximization (EM) algorithm, which both methods employ.

3.1 The Wavelet Model

The specific motivation for my work detailed in the rest of this thesis came from Blocker and Protopapas (2012), herein referred to as B&P. Their work aimed to create an automated methodology to sort through the massive dataset created by a large-scale telescopic survey and determine which objects were candidates for a microlensing event.

B&P suggested that we can differentiate between lightcurves that experience an event and those that are simply periodic by analyzing small groups of observations that differ substantially from those around them. However, as explored in Chapter 1, within a database of stellar observations there will be objects that are non-periodic as well as objects that pulsate. Additionally, while not discussed at any length within this thesis, there will be objects which experience an anomalous event such as a supernova. Thus, any proposed algorithm must be robust enough to remove the stars that are non-variable and those that are anomalies, the first of which will likely constitute the majority of the objects in the database.

To differentiate between classes, the time-scale over which an event occurs is examined. To reduce the difficulty of the problem and increase algorithmic efficiency, B&P decomposed the classification problem into a two-step process using one of the fundamental probability theorems as follows. Let V be the set of all light curves that have variation at the time scale we are interested in and S be the subset of V that includes the lightcurves with a microlensing event, then

$$P(Y_i \in S) = P(Y_i \in V \cap S) = P(Y_i \in V)P(Y_i \in S | Y_i \in V).$$

In B&P, the first step $P(Y_i \in V)$ is calculated using a parametric¹ wavelet model, while the second step $P(Y_i \in S | Y_i \in V)$ is calculated using a regularized logistic regression

¹B&P described their statistical model here as ‘semi-parametric’. However, after review it is clear that all parameters within the model are fully parameterized with clearly stated distributions and there are no nonparametric components.

classification scheme. The logistic regression training model is calculated using a combination of labelled data from the MACHO survey and simulated microlensing event data. The model itself is created from two features they engineered, with the equations and reasons for using them provided in their paper. However, since neither the data they used to train their model or a pre-trained model was available to this author, this second step is not examined in this thesis and I only focus on the classification problem in the first step, i.e. to identify light curves that have variation at the time scale.

To address the irregular sampling issues consistently present in astrophysical observations (discussed in Chapter 2), B&P constructed their model using wavelets. In general, wavelets are a useful tool for analyzing data with discontinuities or sharp change (Edes-Huyal et al., 2021). A wavelet is basically a vector of discrete numbers, referred to as wavelet coefficients, that represent some detail of a more complex waveform. If one had infinite wavelets that were compared in some way to an original waveform, theoretically you could recreate the original wave no matter how complex. A lengthier discussion on wavelets and wavelet construction can be found in Appendix A for interested readers. Here it is sufficient to state that B&P chose to use the Least Asymmetric Daubechies 4 wavelet, also called Symlet-4.

Instead of processing each star’s lightcurve directly through a Symlet-4 filter, a standard basis of discretized wavelet coefficients was generated. B&P included code in their GitHub repository (discussed further in Chapter 4) which I utilized to ensure consistency when comparing methods to generate the basis. The first column of the basis is an added intercept column equal everywhere to $2^{-11/2}$, and each successive column contains the wavelet coefficients for successive Symlet-4 wavelets $\phi(t) = (\phi_0(t), \phi_1(t), \dots, \phi_{2047}(t))^T$. To create the wavelet basis, B&P performed the Symlet-4 discrete wavelet transform on a vector of zeroes, and after some rearrangement created a matrix of size $2^{11} \times 2^{11}$, i.e. 2048×2048 . Any size basis could have been constructed as long as it was of a size 2^X . However, the dataset used by both B&P and by this thesis contains up to 1200 or more observations per star, and thus using a basis of size $2^{11} = 2048$ provides the closest size to the data.

The complete basis can be thought of as a collection of wavelets which, when applied to a general waveform, will represent increasingly finer detail of the original wave. To help show how the basis looks, the first eighteen wavelets (that is, columns 2-19 of the basis) are depicted in Figure 3.1. The wavelet coefficients are shown on the y -axis and graphed against the X -axis i , $i = 1, \dots, 2048$. It can be seen from Figure 3.1 that the area where $y_i \neq 0$ is shrinking over each successive wavelet as that wavelet represents finer details of a general waveform.

A basic physical principle is the inverse relationship between the period T and frequency f of a waveform. Here period represents the length of time it takes a waveform to complete one cycle or oscillation and inversely frequency is the number of oscillations per time unit. This can be expressed simply as

$$f = \frac{1}{T} = \frac{\omega}{2\pi},$$

where ω is the angular frequency, measured in radians per second. Therefore, when I discuss a microlensing event - a star whose lightcurve shows a singular spike or bump - this translates to an object with a low frequency (e.g. it only happened once during the whole waveform) which further translates to a large period. Conversely, a pulsating star will have a low periodicity since its frequency will be high (i.e. the waveform happened many times during the time the star was observed).

For each lightcurve, let $y(t)$ be the observed magnitude of the curve at time t , then a linear model can be fitted with $y(t)$ regressed on the wavelet basis vector

$\phi(t) = (\phi_0(t), \phi_1(t), \dots, \phi_M(t))^\top$, i.e.

$$y(t) = \beta_0\phi_0(t) + \sum_{i=1}^k \beta_i\phi_i(t) + \sum_{j=k+1}^M \beta_j\phi_j(t) + \epsilon(t), \quad (3.1)$$

where $\beta = (\beta_0, \beta_1, \dots, \beta_M)^\top$ is the unknown regression coefficient vector and $\epsilon(t)$ is the error/noise term possibly depending on time t . Here M is an integer much less than 2047 in order to have a much simpler whereas efficient enough linear model to fit $y(t)$. In addition,

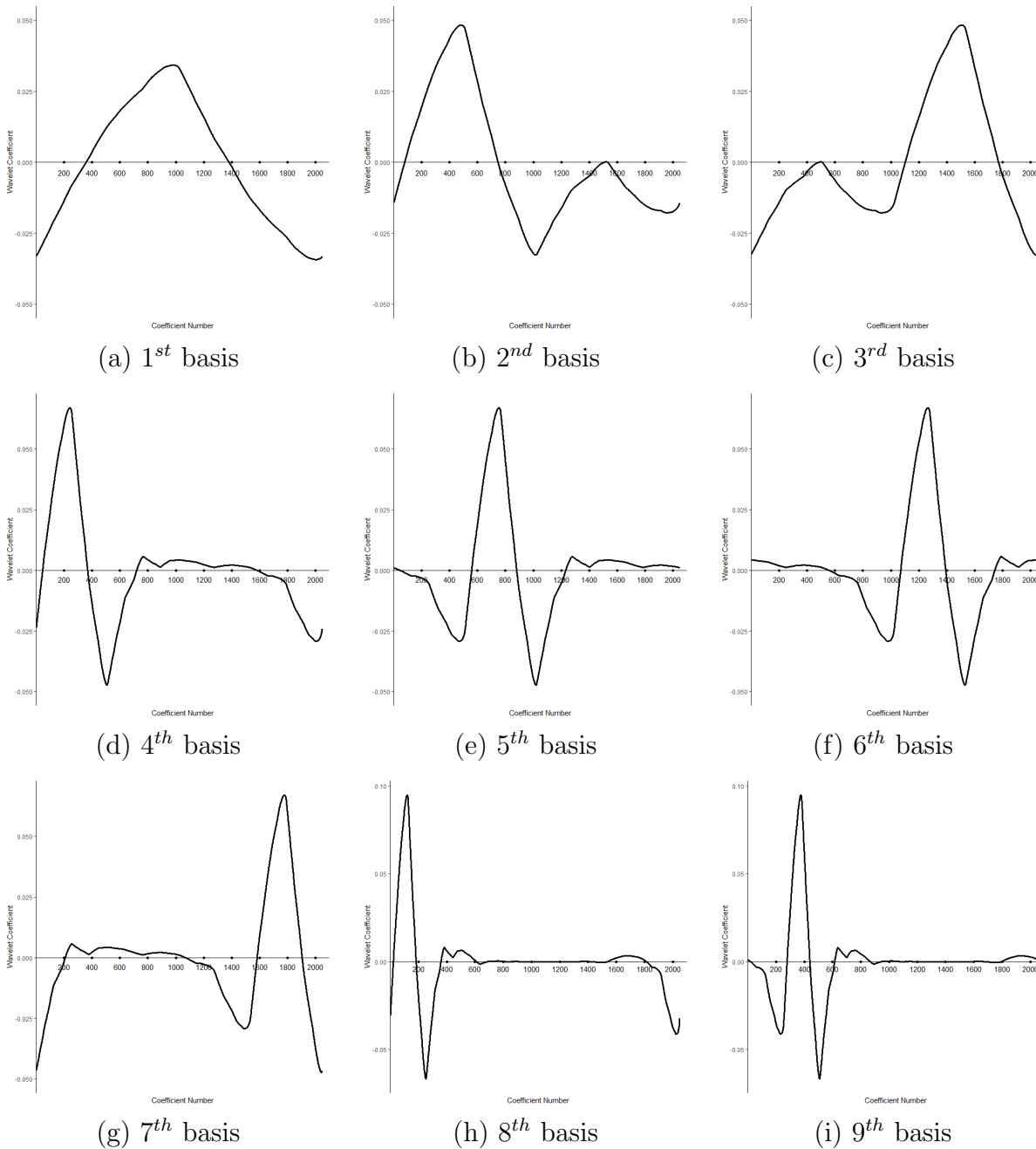


Figure 3.1: A subset of the Symlet-4 wavelet bases

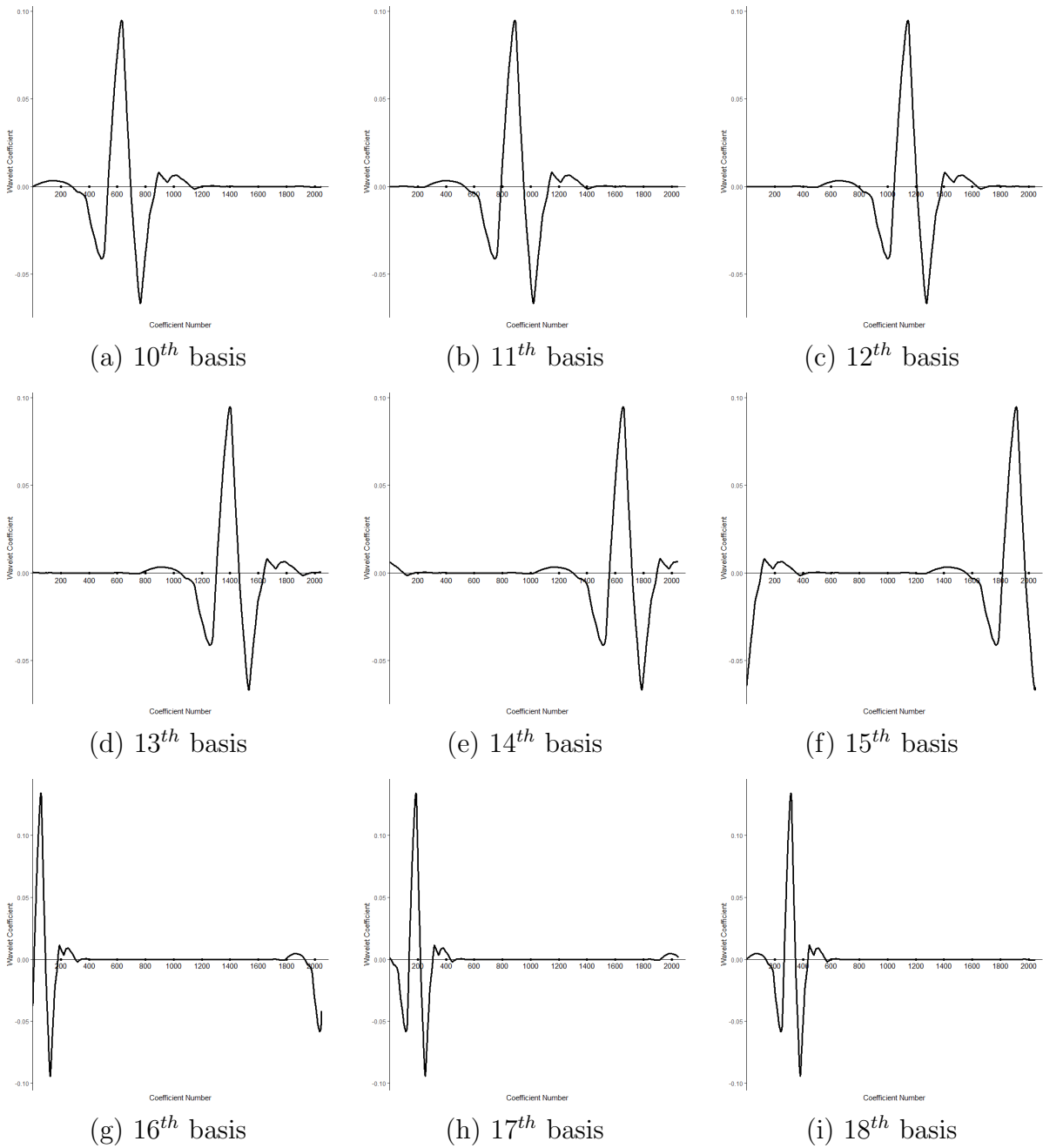


Figure 3.2: A subset of the Symlet-4 wavelet bases (con.)

k is used to separate event-like variation from trends - a trend being a lightcurve with rapid and regular changes in period. Recall that high frequency corresponds to low periodicity, and vice versa. Thus $(\phi_1, \dots, \phi_k)^\top$ model the structure due to trends, while the higher frequency components $(\phi_{k+1}, \dots, \phi_M)^\top$ model the structure due to events. Following B&P, I will always use $M = 128$ and $k = 8$ in the analysis shown in this thesis. B&P noted that any variation found beyond $M = 128$ likely corresponds to an isolated outlier or some other anomalous phenomena not of interest to this study. B&P chose $k = 8$ in order to provide the best results for finding events. However, if future research were to use this method to search for a different type of phenomena, it may be interesting to modify these values.

Consider now a lightcurve with n observations $y = (y_1, \dots, y_n)^\top$ at n different time points $t = (t_1, \dots, t_n)^\top$, i.e. we observe n pairs $(t_1, y_1), (t_2, y_2), \dots, (t_n, y_n)$. To calculate $\phi(t)$, I first transform each time value t_i to

$$t'_i = \left(\frac{t_i - t_1}{t_n - t_1} \times 2048 \right) + 0.5, \quad i = 1, \dots, n.$$

I then calculate $t_i^* = \lfloor t'_i \rfloor$, the floor of t'_i ,² which is used as an index to select rows of the wavelet basis matrix ϕ . This selection process is not one to one, as some elements of the wavelet basis will be selected multiple times, whereas others will be skipped entirely. I use $\phi_{n \times 129}^* = (\phi_0^*, \dots, \phi_{128}^*)$ to denote the finally selected matrix from ϕ . Let $\beta = (\beta_0, \dots, \beta_{128})^\top$. Now I regress y against ϕ^* as

$$y = \beta_0 \phi_0^* + \sum_{i=1}^8 \beta_i \phi_i^* + \sum_{i=9}^{128} \beta_i \phi_i^* + \epsilon = \phi^* \beta + \epsilon. \quad (3.2)$$

This linear wavelet model will be the starting point for which the two methods, i.e. the MAP used in literature and the LAD that I propose, will be applied and compared in this thesis.

The details of these two estimation methods will be described over the next two sections.

²Floor gives the greatest integer less than or equal to t'_i . B&P used this awkward method because their code is written in \mathcal{C} language which doesn't have a simple-to-use math method built in the way that Python-based and other similar languages do.

Let $\hat{\beta}_{MAP}$ and $\hat{\beta}_{LAD}$ denote the estimate obtained by using MAP and LAD respectively. Based on previous discussion, in order to test the event-like variation, I can just test

$$\begin{aligned}
 H_0 : \beta_9 = \beta_{10} = \dots = \beta_{128} = 0 \\
 \text{vs.} \\
 H_a : \beta_j \neq 0 \text{ for some } j = 9, \dots, 128,
 \end{aligned}
 \tag{3.3}$$

for which likelihood ratio test (LRT) or others can be applied. In other words, (3.2) is the full model under H_a and the reduced model under H_0 is

$$y = \beta_0 \phi_0^* + \sum_{i=1}^8 \beta_i \phi_i^* + \epsilon.
 \tag{3.4}$$

Details on testing (3.3) will be shown in Section 3.3.

3.2 Review of Methodologies

In this section, I first describe in Section 3.2.1 the MAP method that B&P used and then in Section 3.2.2 introduce the LAD estimator. Each of the two methods can be reduced to a statistical optimization problem, and as a result I will employ the EM algorithm introduced in Section 3.2.3. Both methods are meant to improve upon the performance of classical methods such as maximum likelihood estimation (MLE) or ordinary least squares (OLS) when data concerns such as large outliers or missing data are present. Later in Chapter 4, I will show that each method has its strengths and weaknesses when tested on a labelled dataset, which motivated us to combine the two methods to have improved results.

3.2.1 Maximum a posteriori (MAP)

To address the missing data issue and take advantage of their proposed pre-knowledge in regards to the parameters, a similar procedure to MLE was used by B&P, called MAP

estimation.

First comes the general definition of MAP estimation. Let $y = (y_1, \dots, y_n)$ be an i.i.d. sample from a population with density function $f(\cdot|\theta)$ which depends on the parameter θ . Suppose that θ has some known prior density $g(\theta)$. Then by Bayes theorem³ the posterior density of θ is

$$g(\theta|y) = \frac{\prod_{i=1}^n f(y_i|\theta) \cdot g(\theta)}{\int \prod_{i=1}^n f(y_i|\theta)g(\theta)d\theta} \propto \prod_{i=1}^n f(y_i|\theta) \cdot g(\theta).$$

Then the general MAP estimator is given as

$$\hat{\theta}_{MAP} = \arg \max_{\theta} g(\theta|y) = \arg \max_{\theta} \prod_{i=1}^n f(y_i|\theta)g(\theta). \quad (3.5)$$

For the regression model (3.2), the MAP estimator of β is given as

$$\hat{\beta}_{MAP} = \arg \max_{\beta} \prod_{i=1}^n f(y_i|\phi_i^*, \beta)g(\beta),$$

where $y_i \sim f(\cdot|\phi_i^*, \beta)$.

B&P define their prior density to be $g(\beta) := N(0, \sigma^2/\tau)$ suggesting that this will stabilize their inferences⁴. They set $\tau = \frac{1}{100}$ to create an uninformative prior that they claim is sufficient to regularize estimates where gaps are present in the data. Additionally they treat their residuals as $\epsilon(t) \sim t_{\nu}(0, \sigma^2)$ with $\nu = 5$.⁵ The wide tails of the t -distribution help account for the large data outliers seen in astronomical survey datasets, as discussed in Chapter 2. However, the likelihood function of the t -distribution has no general closed form solution. EM can be applied directly from here, but B&P instead followed the data augmentation

³ $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$

⁴B&P published two copies of their paper. In the original (Blocker and Protopapas, 2012) they incorrectly define their prior to be $g(\beta) := N(0, \sigma^2\tau)$ but correct it in their (Blocker and Protopapas, 2013) publication to the version stated here.

⁵ ν can be treated as a random variable as well, but B&P found that the increase in computational time did not justify the increase in predictive accuracy.

scheme presented in Meng and Van Dyk (1997) due to the substantial reduction in number of iterations required. This method defines $z(t) \sim N(0, 1)$ and $q(t) \sim \text{InvGamma}(\frac{\nu}{2}, \frac{\nu}{2})$, where $z(t)$ and $q(t)$ are independent of one another. If $q(t)$ is considered to represent the missing data, then the residuals can be defined as

$$\epsilon(t) \sim t_5(0, \sigma^2) \sim z(t) \cdot \sqrt{q(t)}.$$

This has the additional benefit that the prior $g(\beta)$ is conditionally conjugate to this augmented form of the model. Then, as shown in Meng and Van Dyk (1997), calculating the MAP estimator for β becomes a straightforward weighted least-squares procedure where

$$\hat{\beta}_{MAP} = ([\phi^*]^\top \sqrt{w} [\phi^*])^{-1} [\phi^*]^\top \sqrt{w} y \quad (3.6)$$

with

$$w = E[q(t) | y, \phi^*, \beta, \sigma] = \frac{\nu + 1}{\nu + [\epsilon(t)]^2 \tau^{-1}} = \frac{6}{5 + [\epsilon(t)]^2 \tau^{-1}} \quad (3.7)$$

3.2.2 Least Absolute Deviations (LAD) Regression

LAD regression is similar to Ordinary Least Squares (OLS) but with the L_2 -norm replaced by L_1 -norm. In addition, LAD is considerably more robust than OLS as it does not emphasize as strongly outlier points within the data. Since astrophysical data generally contains many outliers, as shown in Chapter 2, this LAD method is suitable to find a robust solution while being simpler to understand than the method shown in the previous section.

For the regression model (3.2), the LAD estimator of β is defined as

$$\hat{\beta}_{LAD} = \min_{\beta} \sum_{i=1}^n |y - \phi^* \beta|_i$$

where A_i denotes the i^{th} element of the column vector A . While there is no closed-form solution to this optimization problem, Phillips (2002) proposed an EM algorithm to calculate

$\hat{\beta}_{LAD}$ iteratively, via a model reformulization as follows. The regression model (3.2) can be rewritten as

$$y_i = \phi_i^* \beta + \sigma(\sqrt{2\nu_i})^{-1} \epsilon_i,$$

where $\sigma > 0$, ϵ_i 's are assumed i.i.d. standard normal errors, and the ν_i 's are i.i.d. positive random variables independent of ϵ_i 's. For convenience, assume that $(\sqrt{2\nu_i})^{-1}$ has an exponential distribution, then the conditional density $f(\cdot|\phi_i^*, \nu_i, \beta, \sigma)$ of y_i is normal. Finally, the marginal density of y_i given x_i is a double exponential density; see Phillips (2002) for the proofs.

3.2.3 EM algorithm

The Expectation Maximization (EM) procedure consists of two steps which are performed iteratively until convergence, e.g. the difference between the likelihood value of the new iteration and the previous iteration reaches some arbitrarily small pre-determined cut-off value. The general procedure is as follows. Given the unknown parameter vector θ , let y denote the observed sample of size n with likelihood $L(\theta) = f_Y(y|\theta)$, and z denote the unobserved data with likelihood function $f_Z(z|\theta)$. Denote the full likelihood as $L_c(\theta) = f_c(y, z|\theta)$. Now let $\hat{\theta}^{(k)}$ denote the estimate of θ from the k^{th} iteration. Then for the current $(k+1)^{th}$ iteration, the Expectation step (E-step) and the Maximization step (M-step) of the EM algorithm are given below.

E-step: Calculate

$$Q(\theta; \theta^{(k)}) := E_{z \sim f_Z(z|y, \theta^{(k)})} [L_c(\theta) | y],$$

i.e. the expected value of the likelihood (or log-likelihood) of θ with respect to the current conditional distribution of z given y and the current estimate $\theta^{(k)}$.

M-step: Calculate the updated predictor $\hat{\theta}^{(k+1)}$ by

$$\hat{\theta}^{(k+1)} = \arg \max_{\theta} Q(\theta; \theta^{(k)}).$$

Iterations are performed over these two steps until convergence is reached, e.g. $L(\theta^{(k+1)}) - L(\theta^{(k)}) \approx 0$.

The well-known work Dempster et al. (1977) first described the EM algorithm and showed that

$$L(\theta^{(k+1)}) \geq L(\theta^{(k)}).$$

That is, the observed data likelihood function $L(\theta)$ does not decrease after an EM iteration. Said another way, convergence will always be reached as long as the likelihood function has an upper bound.

3.3 Proposed Methods

3.3.1 MAP with EM

To iteratively calculate $\hat{\beta}_{MAP}$ I follow B&P and use the EM method with an augmented data scheme, as shown in McLachlan (2008) and Meng and Van Dyk (1997). Before I can begin iterations, an initial value for w must be chosen; B&P chose the parameter τ in the distribution of the prior $g(\beta)$ to be $\frac{1}{100}$. Then an initial value of $\hat{\beta}_{MAP}$ can be found by calculating (3.6). The residuals $\epsilon(t)$ and an updated value for τ are calculated as

$$\epsilon = y - \phi^* \hat{\beta},$$

$$\tau = \frac{\sum_{i=1}^n \epsilon_i w_i}{\sum_{i=1}^n |w_i|}.$$

Finally, an initial value for the log-prior and log-posterior are found. Once these initial values are calculated, the EM method is applied as follows.

E-step: Calculate the weights $w^{(k+1)} = E \left[q(t) \mid y, \phi^*, \hat{\beta}^{(k)}, \hat{\epsilon}(t)^{(k)}, \hat{\tau}^{(k)} \right]$ by (3.7), i.e.

$$w^{(k+1)} = \frac{\nu + 1}{\nu + \frac{[\hat{\epsilon}(t)^{(k)}]^2}{\hat{\tau}^{(k)}}} = \frac{6}{5 + \frac{[\hat{\epsilon}(t)^{(k)}]^2}{\hat{\tau}^{(k)}}}.$$

M-step: Update the values $\hat{\beta}^{(k+1)}$, $[\hat{\epsilon}(t)]^{(k+1)}$ and $\hat{\tau}^{(k+1)}$ by

$$\begin{aligned} \hat{\beta}^{(k+1)} &= \left(\phi^{*T} \sqrt{w^{(k+1)}} \phi^* \right)^{-1} \phi^{*T} \sqrt{w^{(k+1)}} y, \\ [\hat{\epsilon}(t)]^{(k+1)} &= y - f(\hat{\beta}^{(k+1)}, \phi^*) = y - \phi^* \hat{\beta}^{(k+1)}, \\ \hat{\tau}^{(k+1)} &= \frac{\sum_{i=1}^n [\hat{\epsilon}(t)_i^{(k+1)}]^2 + \sum_{j=1}^m (\hat{\beta}_j^{(k+1)})^2 [g(\beta)]_j}{n + \sum_{j=1}^m [g(\beta)]_j}, \end{aligned}$$

where $g(\beta)$ is the vector of priors and m is the number of regressors currently being modelled.

After each iteration the convergence condition $\ln L(\hat{\beta}^{(k+1)}) - \ln L(\hat{\beta}^{(k)}) \approx 0$ is checked, where

$$\begin{aligned} \ln L(\hat{\beta}^{(k+1)}) &= \sum_i \ln f(y_i | \phi_i^*, \hat{\beta}^{(k+1)}) \\ &= \sum_{i=1}^n \left(\frac{(\nu + 1)}{2} \ln \left(1 + \frac{[\hat{\epsilon}(t)_j]^2}{\nu \hat{\tau}} \right) - \ln(\sqrt{\hat{\tau}}) \right) - \sum_{j=1}^m \left(\frac{[\hat{\beta}_j^{(k+1)}]^2}{2 \sqrt{\frac{\hat{\tau}}{g(\beta)_{(1)}}}} + \ln \left(\sqrt{\frac{\hat{\tau}}{[g(\beta)]_1}} \right) \right) \end{aligned} \quad (3.8)$$

represents the log posterior of the current iteration. Note that $[g(\beta)]_1$ is the first value of the prior vector (although any value would have been sufficient due to the fact that I used a uniform prior).

To calculate the p -value or find the rejection region, B&P employed a χ^2 approximation

with a modified Benjamini-Hochberg False Discovery Rate (FDR) procedure, introduced by Benjamini and Yekutieli (2001)⁶. This enables them to conduct simultaneous hypothesis testing while controlling the FDR, despite the positive dependency of the test statistics (Benjamini and Yekutieli, 2001). Then the test statistic that B&P employed is $(\hat{\ell}_1 - \hat{\ell}_0)$, the approximate log-ratio of the marginal posterior probabilities, where $\hat{\ell}_1$ and $\hat{\ell}_0$ represent the log-likelihoods of the full and restricted models shown in (3.2) and (3.4) respectively and calculated through (3.8).

3.3.2 LAD with EM

To calculate my LAD test statistics, I also use an EM procedure, which has subtle but important differences from the EM procedure with MAP estimation. However, I utilize the Likelihood Ratio Test (LRT) statistics directly instead of with the Benjamini-Hochberg FDR procedure.

Unlike the former section, I treat y_1, \dots, y_n as the incomplete data, and consider $(y_1, \nu_1), \dots, (y_n, \nu_n)$ to be the complete data. Recall that $(\sqrt{2}\nu_i)^{-2}$ has an exponential distribution.

Then, for the $(k + 1)^{th}$ iteration, $k = 1, 2, \dots$, the EM algorithm is given below.

E-step: As shown in Phillips (2002), if I drop the terms that do not depend on β or σ I can derive the proxy log-likelihood

$$-\frac{n}{2} \ln(\sigma^2) - \frac{1}{\sigma^2} \sum_{i=1}^n E[\nu_i^2] (y_i - \phi_i^* \beta)^2, \quad (3.9)$$

with

$$E[\nu_i^2] := E[\nu_i^2 \mid y_i, \phi_i^*, \beta, \sigma] = \frac{\sigma}{\sqrt{2} |y_i - \phi_i^* \beta|} \quad (3.10)$$

⁶This procedure replaces the more basic LRT, although B&P also provided within their code a calculation for the LRT in case a future user might prefer this statistical test. The Benjamini-Hochberg FDR procedure replaces the more traditional family-wise error rate used to control the probability of rendering type I error with a Bonferroni-type calculation of the proportion of errors among the rejected hypotheses.

for all $|y_i - \phi_i^* \beta| > 0$. Then weights can be defined as $W^{(k+1)} = \text{diag}(E[\nu_i^2])$.

M-step: From the previous step $\hat{\beta}_{LAD}$ can be reduced to

$$\hat{\beta}^{(k+1)} = ([\phi^*]^\top W^{(k+1)} [\phi^*])^{-1} [\phi^*]^\top W^{(k+1)} y.$$

After each iteration the convergence condition $L(\beta^{(k+1)}) - L(\beta^{(k)}) \approx 0$ is checked. Note that because the weights do not depend on σ , and therefore the calculated values of $\hat{\beta}$ are independent of any calculated values of $\hat{\sigma}$, $\hat{\sigma}$ can be calculated in a single step by

$$\hat{\sigma} = \frac{\sqrt{2}}{n} \sum_{i=1}^n |y_i - \phi_i^* \hat{\beta}|$$

after the EM algorithm reaches convergence.

Once I obtain estimates of β and σ , I perform the hypothesis test as shown in (3.3). However, because the LAD statistic is already robust, as mentioned earlier, I opt to use the LRT directly. Then my test statistic is $-2(\hat{\ell}_0 - \hat{\ell}_1) \sim \chi^2$, with degrees of freedom equal to 120, where $\hat{\ell}_0$ and $\hat{\ell}_1$ represent the log-likelihoods calculated from (3.9) of the null and alternative models respectively.

Chapter 4

Analysis Results

In this chapter, I first describe the pre-classified MACHO data subset that I apply the MAP and LAD classifiers to. Then I provide a comparison of the two methods and propose a two-tier voting classifier utilizing both the MAP and the LAD methods to provide the best classification results.

All computations were performed on the University of Calgary’s Advanced Research Computing (ARC) cluster high performance computing system¹. The system is built for University of Calgary research projects and consists of over 16,000 cores. For this research project I was allocated a working space of up to 15 TBs, with an individual job runtime of up to 7 days.

4.1 Description of the Testing Dataset

To obtain a clear comparison between methods it was first necessary for me to obtain my own labelled dataset. There is no predefined or commonly used group of test data, so to achieve this aim I found representative examples of each of the stellar types outlined

¹The author would like to thank the Research Computing Services group at the University of Calgary for access to the ARC cluster which was vital to this research project, and the staff for their invaluable expertise and assistance.

in Chapter 1 by searching the VizieR Astronomical Catalogue² (Ochsenbein, 2024). This online tool provides access to a huge library of published astronomical catalogues. Experts and researchers from around the world voluntarily provide VizieR access to their research results when they have confidently classified any stellar object from any catalogue. It also scans literature and retrieves data from publications. The tool then provides tables of combined research with the aggregated results and various confidence levels for any previously researched stellar object, and references to the original papers or catalogues. Using this tool I found 5480 labeled stars from the MACHO survey which form my testing dataset. An immediate limitation of the data is clear here, that it is only as accurate as the accuracy of the pre-classified data. I attempted to use stars which had several separate references validating their type, but there is definitely room for error. The breakdown of this testing data is shown in Table 4.1.

Table 4.1: Breakdown of the testing data

Type	Number of Stars
Non-variable	68
Cepheid	1868
RR Lyrae	108
Eclipsing Binary	3031
Microlensing	405
Total	5480

As can be seen from Table 4.1, I was only able to find 68 stars that are definitely from the non-variable Main Sequence class. Unfortunately it proved extremely difficult to find verified non-variable stars, despite these making up a large part of the MACHO survey data. This provides a major limitation to the results shown in the rest of this thesis. In the future it would certainly be interesting to obtain a larger dataset of non-variable-type stars, perhaps by collaborating with an expert who could identify stars via visual inspection, or

²This research has made use of the VizieR catalogue access tool, CDS, Strasbourg, France (DOI : 10.26093/cds/vizieR). The original description of the VizieR service was published in 2000, A&AS 143, 23.

who had greater familiarity with the VizieR tool. This also creates the additional limitation of possibly optimizing the classifier to the data. This could also be mitigated by testing using additional data but was outside of the scope of this thesis, although it would certainly be interesting to try in the future.

As can be seen from Table 4.1, I obtained four different types of variable stars to test whether the proposed methods are robust enough to identify the variability despite of the type of the variability. However, I do not make any further distinction in the rest of this thesis than between variable or non-variable stars. A further method would be required to make more fine-grained distinctions³. A few of the papers discussed in Section 2.2 performed such classifications, for the interested readers.

4.2 Comparing MAP and LAD Results

Blocker and Protopapas (2012) tested their MAP model on a simulated dataset⁴ as well as the full MACHO dataset. They claimed they were able to obtain an approximate test power of 80%. Unfortunately they did not provide any way to replicate their simulated data or test, so this claim was unable to be verified here. However, the authors did provide their code in a GitHub repository (Blocker, 2013). It includes code to explicitly define the wavelet basis which enables clear comparisons between the two methods.

Recall that each star has data in two wavelengths, red and blue. Each method was ran on both wavelengths, producing a pair of p -values for each method. Significance level $\alpha = 0.1$ is used for each individual test. I consider a method to have identified a variable star if it finds variability in at least one of the two wavelength colours. It follows that a star is not classified as variable if both tests on red and blue wavelengths produce p -values bigger than 0.1.

³Blocker and Protopapas (2012) did discuss such a method, but since their results were not easily replicable, this work is left out of the scope of this thesis.

⁴This dataset consists of simulated 50,000 event curves and 50,000 null curves based on the MACHO dataset.

A summary of the results from both MAP and LAD is presented in Table 4.2. From Table 4.2 I observe that the MAP method consistently outperforms the LAD method in terms of the proportion of correctly identified variable stars. However, MAP does this at the cost of being overly permissive, i.e. it struggles to identify the non-variable stars. In contrast, while the LAD method is very conservative and does not come close to the MAP method in correctly identifying variable stars, it does a much better job at identifying the non-variable stars. This is likely because it had such a low likelihood of predicting a star to be variable. The results of a subset of individual stars are given in Appendix B.

Table 4.2: Results of MAP and LAD classifiers

Truth		MAP		LAD	
Star Type	Total Number	Non-variable	Variable	Non-variable	Variable
Non-variable	68	18 (26.5%)	50 (73.5%)	45 (66.2%)	23 (33.8%)
Variable	5412	1029 (19.0%)	4383 (81.0%)	5259 (97.2%)	153 (2.8%)

In terms of timing, the MAP method ran quite a bit faster than LAD. On average, it takes 1.19 seconds to perform a MAP procedure, while in contrast the LAD method takes about 9.65 seconds. For each star the test was conducted on the same computing node with the same computing resources available, ensuring that any difference is not due to computer hardware but the methods themselves.

4.3 Voting Classifier

If my goal was simply to identify as many stars to be variable as possible, I would conclude that the MAP method was satisfactory and conclude this thesis here. However, recall that the actual goal of many classification problems in this field is to identify a subset of the complete survey data which warrants researchers' further study and investigation. If a larger subset with false discoveries is identified, there will be waste of time and resources put on those false discoveries in followup studies. For the problem under our consideration,

the time and effort that researchers spend on false variable stars should be minimized, if it can not be avoided completely. Thus it is imperative that an efficient classifier not be overly permissive of predicting variability when the star is in fact non-variable. I therefore have the goal to integrate the high power of MAP and the low false discovery rate of LAD.

As a starting point to combine MAP and LAD, for each star I simply consider all the four tests from both MAP and LAD at once, instead of two for each method separately. More specifically, for each star I calculate four p -values: $p_{MAP,r}$ for MAP on the red filter band, $p_{MAP,b}$ for MAP on the blue filter band, $p_{LAD,r}$ for LAD on the red filter band, and $p_{LAD,b}$ for LAD on the blue filter band. I conclude a star to be variable if any of the four gives a $p \leq 0.1$, and to be non-variable if all the four p -values are bigger than 0.1 (i.e. none of them predicts variability). I simply call this method MAP-LAD. The results for this MAP-LAD classifier are presented in Table 4.3. From Table 4.3, I observe that the number of correctly identified variable stars is higher than that of both MAP and LAD, and the number of correctly identified non-variable stars is lower than that of both MAP and LAD. Thus as expected, this MAP-LAD classifier has higher power while at the same time has higher type I error than the LAD classifier alone.

Table 4.3: Results of MAP-LAD classifier

Truth		MAP-LAD	
Star Type	Total Number	Non-variable	Variable
Non-variable	68	13 (19.1%)	55 (80.9%)
Variable	5412	4391 (81.1%)	1021 (18.9%)

In the MAP-LAD method, I essentially use

$$p = \min\{p_{MAP,r}, p_{MAP,b}, p_{LAD,r}, p_{LAD,b}\}$$

as the p -value to make our decision: if $p \leq 0.1$ then the star is classified as variable; otherwise as non-variable.

As an extension of such a p , I consider the linear combination

$$p_w = w_1 p_{MAP,r} + w_2 p_{MAP,b} + w_3 p_{LAD,r} + w_4 p_{LAD,b}$$

as the p -value for decision making, where $w = (w_1, w_2, w_3, w_4)^T$ is the weight vector such that $w_i \in [0, 1]$ for $i = 1, \dots, 4$ and $\sum_{i=1}^4 w_i = 1$. I simply call such a classifier as Weighted MAP-LAD. Then the problem is reduced to how to choose an appropriate weight vector w .

As an exploration, I try various weight vectors w and the more interesting results are shown in Table 4.4. As can be seen there, if our goal is to balance the proportion of correctly identified variable stars and that of correctly identified non-variable stars, then the best choice of w among those considered is $w = (.45, .45, .05, .05)$. I also observe from Table 4.4 that when the weights for MAP increases, the number of correctly identified variable stars increases while that of correctly identified non-variable stars decreases, indicating that the behavior is getting closer to that of MAP when the associated weights increase. Table 4.4 only considers the cases when $w_1 = w_2$ and $w_3 = w_4$. I did test weighting the colours differently, but did not find any substantial improvement to the results.

Table 4.4: Results of Weighted MAP-LAD classifier

w	Correct # Variable	Correct # Non-Variable
(0, 0, .5, .5)	68 (1.26%)	67 (98.53%)
(.2, .2, .3, .3)	63 (1.16%)	67 (98.53%)
(.25, .25, .25, .25)	68 (1.26%)	67 (98.53%)
(.4, .4, .1, .1)	99 (1.83%)	63 (92.65%)
(.425, .425, .075, .075)	153 (2.83%)	59 (86.76%)
(.45, .45, .05, .05)	3882 (71.73%)	50 (73.53%)
(.475, .475, .025, .025)	3924 (72.51%)	45 (66.18%)
(.5, .5, 0, 0)	3924 (72.51%)	45 (66.18%)
Truth	5412	68

For each consideration of weight vector in Table 4.4, all the stars use the same w . Intu-

itively, I may choose different w for different stars. I simply call such a classifier as Varying Weighted MAP-LAD. Following this direction, I start with the easiest extension from constant weight vector to two different weight vectors for different stars based on some easily computable criteria.

For example, depending on whether MAP identifies a star as variable or not, I may use different set of weights to calculate the p -value p_w . Several dozen of such different p_w are examined and below are the selected eight of them, with classification results presented in Table 4.5.

Variation 1:

$$p_w = \begin{cases} 0.4(p_{MAP,r} + p_{MAP,b}) + 0.1(p_{LAD,r} + p_{LAD,b}), & \text{if } \min\{p_{MAP,r}, p_{MAP,b}\} \leq 0.1, \\ 0.1(p_{MAP,r} + p_{MAP,b}) + 0.4(p_{LAD,r} + p_{LAD,b}), & \text{otherwise.} \end{cases}$$

Variation 2:

$$p_w = \begin{cases} 0.45(p_{MAP,r} + p_{MAP,b}) + 0.05(p_{LAD,r} + p_{LAD,b}), & \text{if } \min\{p_{MAP,r}, p_{MAP,b}\} \leq 0.1, \\ 0.05(p_{MAP,r} + p_{MAP,b}) + 0.45(p_{LAD,r} + p_{LAD,b}), & \text{otherwise.} \end{cases}$$

Variation 3:

$$p_w = \begin{cases} 0.45(p_{MAP,r} + p_{MAP,b}) + 0.05(p_{LAD,r} + p_{LAD,b}), & \text{if } \min\{p_{MAP,r}, p_{MAP,b}\} \leq 0.1, \\ 0.25(p_{MAP,r} + p_{MAP,b}) + 0.25(p_{LAD,r} + p_{LAD,b}), & \text{otherwise.} \end{cases}$$

Variation 4:

$$p_w = \begin{cases} 0.45(p_{MAP,r} + p_{MAP,b}) + 0.05(p_{LAD,r} + p_{LAD,b}), & \text{if } \min\{p_{MAP,r}, p_{MAP,b}\} \leq 0.9, \\ 0.05(p_{MAP,r} + p_{MAP,b}) + 0.45(p_{LAD,r} + p_{LAD,b}), & \text{otherwise.} \end{cases}$$

Variation 5:

$$p_w = \begin{cases} 0.45(p_{MAP,r} + p_{MAP,b}) + 0.05(p_{LAD,r} + p_{LAD,b}), & \text{if } \min\{p_{LAD,r}, p_{LAD,b}\} \leq 0.1, \\ 0.05(p_{MAP,r} + p_{MAP,b}) + 0.45(p_{LAD,r} + p_{LAD,b}), & \text{otherwise.} \end{cases}$$

Variation 6:

$$p_w = \begin{cases} 0.25(p_{MAP,r} + p_{MAP,b}) + 0.25(p_{LAD,r} + p_{LAD,b}), & \text{if } \min\{p_{LAD,r}, p_{LAD,b}\} \leq 0.1, \\ 0.45(p_{MAP,r} + p_{MAP,b}) + 0.05(p_{LAD,r} + p_{LAD,b}), & \text{otherwise.} \end{cases}$$

Variation 7:

$$p_w = \begin{cases} 0.25(p_{MAP,r} + p_{MAP,b}) + 0.25(p_{LAD,r} + p_{LAD,b}), & \text{if } \min\{p_{LAD,r}, p_{LAD,b}\} \leq 0.1, \\ 0.475(p_{MAP,r} + p_{MAP,b}) + 0.025(p_{LAD,r} + p_{LAD,b}), & \text{otherwise.} \end{cases}$$

Variation 8:

$$p_w = \begin{cases} 0.35(p_{MAP,r} + p_{MAP,b}) + 0.15(p_{LAD,r} + p_{LAD,b}), & \text{if } \min\{p_{LAD,r}, p_{LAD,b}\} \leq 0.1, \\ 0.45(p_{MAP,r} + p_{MAP,b}) + 0.05(p_{LAD,r} + p_{LAD,b}), & \text{otherwise.} \end{cases}$$

From Table 4.5 I find that Variation 8 produces relatively better results than other variations considered. I also observe from, say, Variations 2 and 3, that small changes in the weights can produce a large difference in the classification results, as also observed in Table 4.4.

4.4 Extension to MACHO Full Dataset

A subset of 38,051,094 stars from the MACHO survey database is analyzed with my proposed Varying Weighted MAP-LAD classifier. There are an average of 458 observations

Table 4.5: Results of Varying Weighted MAP-LAD classifier

Variation	Correct # Variable	Correct # Non-Variable
1	95 (1.76%)	63 (92.65%)
2	3875 (71.60%)	50 (73.53%)
3	3875 (71.60%)	50 (73.53%)
4	3875 (71.60%)	50 (73.53%)
5	130 (2.40%)	61 (89.71%)
6	3812 (70.44%)	56 (82.35%)
7	3861 (71.34%)	51 (75.00%)
8	3817 (70.53%)	56 (82.35%)
Truth	5412	68

per star, with range from 0 to 1200+. A pre-processing step is performed to remove all stars that have less than 10 observations, as this is considered too small to obtain a meaningful statistical conclusion. In addition, only a small number of stars consist entirely of missing data. For most of the stars, observations are missing in only one of the two color bands so that classification could still be carried out based on the other color band. Finally 32,100 stars are removed from the analysis because there are too few observations in either colour band of those stars.

Using the Varying Weighted MAP-LAD Variation 8 classifier, I identify a total of 15,183,670 stars to be of a variable type. Such a high proportion seems unlikely at first, but when it is considered that eclipsing binary systems were included in the testing dataset it makes a little bit more sense. Additionally, I consider that this number may be biased due to the still relatively high probability of making a type I error with this classifier. If I treat only 70.53% of my results as correct identifications of a variable type star, I find 10,709,042 lightcurves that have a higher chance of being predicted as variable-type stars.

Chapter 5

Conclusions

In this thesis, I proposed novel classification methods to identify variable-type stars within a large-scale telescopic survey. Using the data from the MACHO sky, I compared the MAP classifier introduced by Blocker and Protopapas (2012) to my proposed LAD classifier under the regression setting that utilizes wavelets to decompose a lightcurve into components that may demonstrate variability. Specifically, the Daubechies 4 wavelet was employed as a basis to construct the linear model used to fit the raw lightcurve data.

A testing dataset was obtained using the VizieR online catalogue tool. The labelled data were used to compare the two classifiers. My results showed that the MAP classifier outperforms the LAD classifier in terms of both the number of identified variable stars and the computation speed. Nevertheless, the MAP was found to be overly permissive in the sense of classifying many non-variable stars as variable.

To address this issue, the MAP and LAD were combined in three proposed ways. Firstly, for each star, I proposed the MAP-LAD classifier which uses the minimum of the four individual p -values as the p -value to make classification. This method produced similar and even more extreme results than MAP. Secondly, I proposed the Weighted MAP-LAD classifier which uses constant p_w , the weighted linear combination of the four individual p -values with fixed weight vector w , to make classification. This showed an encouraging improvement in

identifying non-variable stars, as seen in Table 4.4. Thirdly, I proposed Varying Weighted MAP-LAD classifier which uses varying p_w to make classification. This method provided more flexibility in the choice of weights and further improved the classification results by increasing the number of identified non-variable stars without sacrificing much in identifying variable stars.

A few limitations have already been discussed, including the limited number of non-variable stars included in the test dataset and a concern around optimizing the solution to the dataset instead of to some general measure. An additional limitation should be noted, that the data consists of a time series and is by definition dependent on previous observations. However, the mathematical models applied here treat the data as independent.

Considering the limitations of the data used in this thesis, it would be interesting to examine more classified stars, particularly more non-variable stars. One way this could be achieved would be collaborating with an expert to expand the non-variable star group with lightcurves classified by visual inspection. Another would be to create simulated data. Choosing a different set of reference objects might provide additional insight into strengths of the classifier. Especially for the proposed Varying Weighted MAP-LAD classifier, there are many different ways to choose varying weights and the various ones studied in this thesis are just a few of them. I only did a very exploratory study of such a classifier, the promising results warrant further investigation of this class of classifiers in order to improve classification results. It would also be interesting to work with only the brightest 10% of the data for each lightcurve, as is standard practice in astrophysical studies, and see how this changes or improves the results. Additionally, here I only require one $\hat{\beta}_i$ to be non-zero, but it would be interesting to study the case of requiring a group to be non-zero to better simulate a variable lightcurve.

Bibliography

Alcock, C., Allsman, R. A., Alves, D., Axelrod, T. S., Becker, A. C., Bennett, D. P., Cook, K. H., Freeman, K. C., Griest, K., Guern, J., Lehner, M. J., Marshall, S. L., Peterson, B. A., Pratt, M. R., Quinn, P. J., Rodgers, A. W., Stubbs, C. W., and Sutherland, W. (1996). The MACHO project: Limits on planetary mass dark matter in the galactic halo from gravitational microlensing. *The Astrophysical Journal*, 471(2):774–782.

Alcock, C., Allsman, R. A., Alves, D. R., Axelrod, T. S., Becker, A. C., Bennett, D. P., Cook, K. H., Dalal, N., Drake, A. J., Freeman, K. C., Geha, M., Griest, K., Lehner, M. J., Marshall, S. L., Minniti, D., Nelson, C. A., Peterson, B. A., Popowski, P., Pratt, M. R., Quinn, P. J., Stubbs, C. W., Sutherland, W., Tomaney, A. B., Vandehei, T., and Welch, D. (2000). The MACHO project: Microlensing results from 5.7 years of large magellanic cloud observations. *The Astrophysical Journal*, 542(1):281–307.

Alcock, C., Allsman, R. A., Alves, D. R., Axelrod, T. S., Becker, A. C., Bennett, D. P., Cook, K. H., Drake, A. J., Freeman, K. C., Geha, M., Griest, K., Lehner, M. J., Marshall, S. L., Minniti, D., Nelson, C. A., Peterson, B. A., Popowski, P., Pratt, M. R., Quinn, P. J., Stubbs, C. W., Sutherland, W., Tomaney, A. B., Vandehei, T., and and, D. W. (2001). The MACHO project: Microlensing detection efficiency. *The Astrophysical Journal Supplement Series*, 136(2):439–462.

Armstrong, D. J., Kirk, J., Lam, K. W. F., McCormac, J., Osborn, H. P., Spake, J., Walker, S., Brown, D. J. A., Kristiansen, M. H., Pollacco, D., West, R., and Wheatley, P. J.

- (2015). K2 variable catalogue – ii. machine learning classification of variable stars and eclipsing binaries in k2 fields 0–4. *Monthly Notices of the Royal Astronomical Society*, 456(2):2260–2272.
- Australian National University (2020). Macho data collection. Retrieved on 10/10/2019 from https://geonetwork.nci.org.au/geonetwork/srv/eng/catalog.search#/metadata/f6028_2676_1478_3956.
- Babu, G. J. and Mahabal, A. (2015). Skysurveys, light curves and statistical challenges. *International Statistical Review*, 84(3):506–527.
- Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29(4):1165–1188.
- Blocker, A. W. (2013). Robust (semi-parametric) wavelet-based event detection. GitHub Repository. Retrieved on 13/09/2017 from <https://github.com/awblocker/rowavedt>.
- Blocker, A. W. and Protopapas, P. (2012). *Statistical Challenges In Modern Astronomy V*, chapter Semi-Parametric Robust Event Detection for Massive Time Domain Databases, pages 177–187. Lecture Notes in Statistics. Springer.
- Blocker, A. W. and Protopapas, P. (2013). Semi-parametric robust event detection for massive time-domain databases.
- Blomme, J., Sarro, L. M., O’Donovan, F. T., Debosscher, J., Brown, T., Lopez, M., Dubath, P., Rimoldini, L., Charbonneau, D., Dunham, E., Mandushev, G., Ciardi, D. R., De Ridder, J., and Aerts, C. (2011). Improved methodology for the automated classification of periodic variable stars: Automated classification of periodic variable stars. *Monthly Notices of the Royal Astronomical Society*, 418(1):96–106.
- Cao, H., Xiao, H., Luo, Z., Zeng, X., and Fan, J. (2024). Identification of 4fgl uncertain

- sources at higher resolutions with inverse discrete wavelet transform. *The Astrophysical Journal*, 961(1):91.
- Carrasco-Davis, R., Cabrera-Vives, G., Förster, F., Estévez, P. A., Huijse, P., Protopapas, P., Reyes, I., Martínez-Palomera, J., and Donoso, C. (2019). Deep learning for image sequence classification of astronomical events. *Publications of the Astronomical Society of the Pacific*, 131(1004):108006.
- Carroll, B. W. and Ostlie, D. A. (2009). *An introduction to modern astrophysics*. Pearson/Addison-Wesley, San Francisco, Calif.
- Daubechies, I. (1992). Ten lectures on wavelets. CBMS-NSF regional conference series in applied mathematics ; 61, Philadelphia. Society for Industrial and Applied Mathematics.
- de Oliveira, F. M. F., dos Santos, M. V., and Reis, R. R. R. (2022). Data-driven photometric redshift estimation from type ia supernovae light curves. *Monthly Notices of the Royal Astronomical Society*, 518(2):2385–2397.
- Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38.
- Du, C., Luo, A., and Yang, H. (2017). Adaptive stellar spectral subclass classification based on bayesian svms. *New Astronomy*, 51:51–58.
- Edes-Huyal, F. K., Cataltepe, Z., and Kahya, E. O. (2021). On the classification and feature relevance of multiband light curves. *The Astronomical Journal*, 161(4):168.
- ESO (2024). Hertzsprung-russell diagram. Retrieved on 23/01/2024 from <https://www.eso.org/public/images/eso0728c/>.
- Graham, M., Drake, A., Djorgovski, S., Mahabal, A., and Donalek, C. (2017). Challenges in

- the automated classification of variable stars in large databases. *EPJ Web of Conferences*, 152:03001.
- Graps, A. (1995). An introduction to wavelets. *IEEE Computational Science and Engineering*, 2(2).
- Huijse, P., Estevez, P. A., Protopapas, P., Principe, J. C., and Zegers, P. (2014). Computational intelligence challenges and applications on large-scale astronomical time series databases. *IEEE Computational Intelligence Magazine*, 9(3):27–39.
- Huijse, P., Estévez, P. A., Förster, F., and Berrocal, E. (2015). Discriminating variable star candidates in large image databases from the hits survey using nmf. *Procedia Computer Science*, 53:29–38.
- Johnston, K. and Oluseyi, H. (2017). Generation of a supervised classification algorithm for time-series variable stars with an application to the linear dataset. *New Astronomy*, 52:35–47.
- Kim, D.-W. and Bailer-Jones, C. A. L. (2016). A package for the automated classification of periodic variable stars. *Astronomy & Astrophysics*, 587:A18.
- Kügler, S. D., Gianniotis, N., and Polsterer, K. L. (2015). Featureless classification of light curves. *Monthly Notices of the Royal Astronomical Society*, 451(4):3385–3392.
- Li, X.-R., Yu, W.-L., Fan, X.-L., and Babu, G. J. (2020). Some optimizations on detecting gravitational wave using convolutional neural network. *Frontiers of Physics*, 15(5).
- Mallat, S. G. (2009). *A wavelet tour of signal processing*. Elsevier/Academic Press.
- McLachlan, G. J. (2008). *The EM algorithm and extensions*. Wiley series in probability and statistics. Wiley-Interscience, Hoboken, N.J, 2nd ed. edition.

- Meng, X.-L. and Van Dyk, D. (1997). The em algorithm—an old folk-song sung to a fast new tune. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 59(3):511–567.
- NASA (2024). Types of stars. website. Retrieved on 23/01/2024 from <https://science.nasa.gov/universe/stars/types/>.
- Observatory, V. C. R. (2024). Explore - rubin numbers. Retrieved on 17/01/2024 from <https://rubinobservatory.org/explore/numbers>.
- Ochsenbein, F. (2024). The Vizier database of astronomical catalogues .
- Paegert, M., Stassun, K. G., and Burger, D. M. (2014). The eb factory project. i. a fast, neural-net-based, general purpose light curve classifier optimized for eclipsing binaries. *The Astronomical Journal*, 148(2):31.
- Percy, J. R. (2007). *Understanding variable stars*. Cambridge University Press, Cambridge. Includes bibliographical references (pages 332-341) and index.
- Phillips, R. F. (2002). Least absolute deviations estimation via the em algorithm. *Statistics and Computing*, 12(3):281–285.
- Schneider, P., Kochanek, C., and Wambsganss, J. (2006). *Gravitational lensing: strong, weak and micro: Saas-Fee advanced course 33*, volume 33. Springer Science & Business Media.
- Stoppa, F., Bhattacharyya, S., Ruiz de Austri, R., Vreeswijk, P., Caron, S., Zaharijas, G., Bloemen, S., Principe, G., Malyshev, D., Vodeb, V., Groot, P. J., Cator, E., and Nelemans, G. (2023). Autosourceid-classifier: Star-galaxy classification using a convolutional neural network with spatial information. *Astronomy & Astrophysics*, 680:A109.
- Todd, I., Pollacco, D., Skillen, I., Bramich, D. M., Bell, S., and Augusteijn, T. (2006).

- Eclipsing binary stars in nearby galaxies. *Astrophysics and Space Science*, 304(1–4):223–225.
- Tran, L. M. (2006). From fourier transforms to wavelet analysis: Mathematical concepts and examples.
- Velicheti, P. D., Wu, J. F., and Petric, A. (2023). Quantifying roman wfi dark images with the wavelet scattering transform. *Publications of the Astronomical Society of the Pacific*, 135(1050):084502.
- Wang, C., Bai, Y., López-Sanjuan, C., Yuan, H., Wang, S., Liu, J., Sobral, D., Baqui, P. O., Martín, E. L., Andres Galarza, C., Alcaniz, J., Angulo, R. E., Cenarro, A. J., Cristóbal-Hornillos, D., Dupke, R. A., Ederoclite, A., Hernández-Monteagudo, C., Marín-Franch, A., Moles, M., Sodr e, L., Vázquez Rami o, H., and Varela, J. (2022). J-plus: Support vector machine applied to star-galaxy-qso classification. *Astronomy & Astrophysics*, 659:A144.
- Welch, D. (2012). The MACHO project overview. Retrieved on 06/24/2022 from <https://www.macho.anu.edu.au/Project/Overview/status.html>.
- Zappal a, V., Bendjoya, P., Cellino, A., Farinella, P., and Froeschl e, C. (1995). Asteroid families: Search of a 12,487-asteroid sample using two different clustering techniques. *Icarus*, 116(2):291–314.
- Zaritsky, D., Donnerstein, R., Karunakaran, A., Barbosa, C. E., Dey, A., Kadowaki, J., Spekkens, K., and Zhang, H. (2021). Systematically measuring ultra-diffuse galaxies (smudges). ii. expanded survey description and the stripe 82 catalog. *The Astrophysical Journal Supplement Series*, 257(2):60.
- Zhang, J., Zhang, Y., Kang, Z., Li, C., Tao, Y., Zhao, Y., and Wu, X.-B. (2023). A novel approach for variable star classification based on imbalanced learning. *Publications of the Astronomical Society of Australia*, 40.

Zhong-bao, L. (2016). Stellar spectral classification with locality preserving projections and support vector machine. *Journal of Astrophysics and Astronomy*, 37(2).

Appendix A

Wavelets

Wavelets derive their basis from the same principle that Joseph Fourier presented in 1807, namely that any periodic function can be represented by a linear combination of simple sinusoidal waves. This principle is reflected through the well-known inverse Fourier transform shown below in $L^1(\mathbb{R})$. Let $f(t)$ be some waveform. Then f can be reconstructed from a set of Fourier transforms \hat{f} as

$$f(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \hat{f}(\omega) e^{i\omega t} d\omega,$$

where ω is the frequency and $e^{i\omega t}$ is the complex representation of a wave as defined by Euler's formula $e^{i\omega t} = \cos(\omega t) + i \sin(\omega t)$.

Expressing a waveform through its constituent frequencies has wide-ranging applications across the mathematical and physical disciplines. However, the Fourier transform is only expressed on the interval $[0, 2\pi]$, only uses sines and cosines, and it does not handle well large datasets that contain some kind of complicated pattern, for example, discontinuities or sharp spikes (Tran, 2006). Stellar lightcurves often have both of the mentioned complications. For data in such a area wavelets provide a better solution.

Consider a generic continuous waveform, such as a complete lightcurve. Then a wavelet transform can be used to create a mathematically reversible time-frequency representation of that signal. There are many families of wavelets that all have a different structure and basis

but follow the same basic principles. They can be discrete or continuous. Examples include the Haar, Daubechies, Meyer, and Mexican Hat wavelets. Selection of the best wavelet family depends on the scientific application. Additionally, within each wavelet family there are subclasses classified by the number of coefficients and the iteration level. Here, I will focus specifically on the Daubechies 4 class of wavelets, as this is the class used in Eq. 3.1 for my model, and the class used by Blocker and Protopapas (2012).

A general discrete orthogonal wavelet is formed in the following way. The ‘father wavelet’ $\Phi(x)$ is defined as

$$\Phi(x) = \sum_{k=-\infty}^{\infty} c_k \Phi(2x - k) \quad (\text{A.1})$$

and the ‘mother wavelet’ is defined implicitly as

$$\phi(x) = \sqrt{2} \sum_{k=-\infty}^{\infty} (-1)^k c_{(1-k)} \phi(2x - k), \quad (\text{A.2})$$

where c_k ’s are the wavelet coefficients. To ensure orthogonality, the coefficients must satisfy

$$\sum_{k=0}^{N-2} c_k = 2, \quad \sum_{k=0}^{N-1} c_k c_{k+2l} = 2\delta_{l,0},$$

where δ is the Dirac delta function and l is the wavelet location index (Graps (1995)).

Ingrid Daubechies, in her *Ten Lectures on Wavelets* (1992), showed that to represent a polynomial of order p under these conditions, (A.2) will have at least $2p$ nonzero coefficients. Thus a linear Daubechies wavelet transform (i.e. $p = 2$) is known as the Daubechies 4, and the four coefficients are

$$c_0 = \frac{1 + \sqrt{3}}{4\sqrt{2}}, \quad c_1 = \frac{3 + \sqrt{3}}{4\sqrt{2}}, \quad c_2 = \frac{3 - \sqrt{3}}{4\sqrt{2}}, \quad c_3 = \frac{1 - \sqrt{3}}{4\sqrt{2}}. \quad (\text{A.3})$$

For details on the derivation and proofs of the wavelet, see Mallat (2009).

The wavelet does not have a closed form solution, but methods to calculate it exist within most mathematical software. For this thesis project, I used the implementation written by Blocker (2013) in a python-based language to ensure comparability between the MAP method

by B&P and our proposed methods.

Appendix B

Data Tables

Table B.1: **A subset of the variable star data and results.** The first column shows the ⟨field⟩-⟨tile⟩-⟨sequence⟩ identifier from the MACHO survey. The first identifier, field, can be used to determine where the star is located: Fields 1-99 are the LMC, fields 201-299 are the SMC, and the other fields, that is 101-199, 301-399 and 401-499 are the Galactic Bulge. The second column shows what the true type of the star is. As discussed in Chapter 1, the four types used in these tests are Cepheids, eclipsing binaries, microlensing events, and RR Lyrae stars. Finally, the last four columns show the calculated p -values used to identify variable stars with significance level $\alpha = 0.1$.

MACHO ID	Type	$p_{MAP,r}$	$p_{MAP,b}$	$p_{LAD,r}$	$p_{LAD,b}$
9-4391-25	cepheid	0	0	2.2×10^{-16}	2.2×10^{-16}
1-4657-198	cepheid	0	0	0.5628	0.004138
1-3569-71	cepheid	0	0	0.0948	1
1-3567-1010	cepheid	0	0	1	1
1-3570-29	cepheid	0	0	1	1
9-4396-51	cepheid	0	0	1	1
9-5121-17	cepheid	0	0	1	1
9-5608-11	cepheid	0	0	1	1
11-9232-496	cepheid	2.2511×10^{-83}	0	1	1
12-11169-24	cepheid	0	0	1	1
13-6933-3421	cepheid	0	7.8440×10^{-95}	0.9962	1

Table B.1: (continued)

MACHO ID	Type	$PMAP,r$	$PMAP,b$	$PLAD,r$	$PLAD,b$
14-9104-32	cepheid	0	0	0.9999	1
81-8393-12	cepheid	0	0	1	1
82-8045-64	cepheid	2.0599×10^{-90}	0	1	1
19-4792-10	cepheid	0	0	1	1
79-5144-146	cepheid	1	6.7882×10^{-10}	1	1
55-3858-11	cepheid	5.2561×10^{-18}	1	1	1
77-8036-20	cepheid	0	1	1	1
77-8036-23	cepheid	0	1	1	1
77-7911-27	cepheid	1	1	1	1
77-8151-20	cepheid	1	1	1	1
77-8152-14	cepheid	1	1	1	1
78-6099-68	cepheid	1	1	0.4603	1
22-5106-45	eclipsbin	2.3070×10^{-6}	1.2533×10^{-8}	0.001376	0.07832
78-6466-241	eclipsbin	3.0619×10^{-5}	2.5179×10^{-12}	0.01667	3.455×10^{-6}
80-7558-52	eclipsbin	0	0	2.2×10^{-16}	2.2×10^{-16}
82-9009-55	eclipsbin	0	0	2.2×10^{-16}	2.2×10^{-16}
11-8751-629	eclipsbin	0	0	1.004×10^{-11}	0.8752
82-9011-7	eclipsbin	0	0	2.2×10^{-16}	1
7-7180-37	eclipsbin	0	0	1	0.05355
18-2121-19	eclipsbin	0	5.1072×10^{-48}	1	0.04698
23-3902-21	eclipsbin	0.000247	5.4641×10^{-24}	0.0003929	1
78-6221-90	eclipsbin	1	4.6677×10^{-7}	0.05014	1
24-3706-28	eclipsbin	0	1	0.0005632	1
2-4902-4632	eclipsbin	0	0	1	1
3-6481-38	eclipsbin	0	0	1	1
15-10072-13	eclipsbin	0	0	1	0.1365

Table B.1: (continued)

MACHO ID	Type	$PMAP,r$	$PMAP,b$	$PLAD,r$	$PLAD,b$
18-2840-2731	eclipsbin	2.0493×10^{-47}	0	1	1
19-3818-32	eclipsbin	0	0	1	0.3807
78-5979-122	eclipsbin	0	0	1	1
79-4776-42	eclipsbin	0	0	1	1
79-5739-5807	eclipsbin	2.2503×10^{-31}	1.1342×10^{-30}	1	1
80-6473-124	eclipsbin	0	0	1	1
81-9000-164	eclipsbin	0	0	1	1
82-8402-183	eclipsbin	0	0	1	1
24-3582-1515	eclipsbin	1	1.1945×10^{-4}	1	1
78-5980-1008	eclipsbin	1	5.0958×10^{-16}	0.9985	1
22-4262-303	eclipsbin	0	1	1	1
79-5745-95	eclipsbin	0	1	1	1
22-4873-3686	eclipsbin	1	1	1	1
23-3418-140	eclipsbin	1	1	1	1
57-5308-32	eclipsbin	1	1	1	1
77-7672-521	eclipsbin	1	1	0.9999	1
109-19850-956	microlens	1.1208×10^{-96}	1.8337×10^{-62}	2.2×10^{-16}	2.2×10^{-16}
113-18414-1512	microlens	0	0	2.986×10^{-5}	6.762×10^{-7}
179-21844-84	microlens	2.2197×10^{-48}	1.0445×10^{-74}	2.2×10^{-16}	2.2×10^{-16}
108-18951-1221	microlens	0	0	2.2×10^{-16}	1
110-22454-1042	microlens	0	0	8.088×10^{-7}	1
304-35564-26	microlens	1	2.0348×10^{-11}	2.2×10^{-16}	0.001934
402-48103-1719	microlens	0	1	1	2.2×10^{-16}
172-32358-313	microlens	1	1	1	0.00362
118-18402-495	microlens	4.0818×10^{-12}	9.6229×10^{-25}	0.5409	0.2419
128-22057-2384	microlens	1.6207×10^{-24}	8.3518×10^{-16}	1	1

Table B.1: (continued)

MACHO ID	Type	$P_{MAP,r}$	$P_{MAP,b}$	$P_{LAD,r}$	$P_{LAD,b}$
167-23910-3170	microlens	8.3738×10^{-68}	9.9373×10^{-35}	0.3854	0.5184
403-47907-3004	microlens	0	0	1	1
104-21421-131	microlens	1	3.8185×10^{-7}	0.1611	1
109-20379-2012	microlens	0	1	0.3091	0.996
109-20635-2193	microlens	0	1	1	1
176-19214-2902	microlens	8.0336×10^{-7}	1	1	1
104-21421-273	microlens	1	1	1	1
9-5483-1857	rrlyr	0	0	1	1
9-5487-736	rrlyr	0	0	1	1
11-8750-1694	rrlyr	0	0	1	1
14-9703-450	rrlyr	0	0	1	1
80-6590-1844	rrlyr	0	0	1	1
6-5732-3906	rrlyr	3.5276×10^{-35}	1	1	1
6-5853-3986	rrlyr	1	1	1	1

Table B.2: **A subset of the non-variable star data and results.** As before, the first column shows the ⟨field⟩-⟨tile⟩-⟨sequence⟩ identifier from the MACHO survey. All stars in this set, including the ones not shown, are located in the Galactic Bulge. The second column shows what the true type of the star is, i.e. all are non-variable. Finally, the last four columns show the calculated p -value used to identify variable stars with significance level $\alpha = 0.1$.

MACHO ID	Type	$p_{MAP,r}$	$p_{MAP,b}$	$p_{LAD,r}$	$p_{LAD,b}$
119-20090-3749	nonvar	1	1	0.9999	1
119-20091-3882	nonvar	1	1	0.7638	0.9997
119-20091-3891	nonvar	1	1	1	1
119-20091-3857	nonvar	3.7987×10^{-13}	1	0.8544	1
119-20091-3881	nonvar	1.3996×10^{-7}	1	0.8094	1
119-20090-3744	nonvar	1	2.4662×10^{-5}	1	1
119-20091-3861	nonvar	1	1.7392×10^{-40}	0.1415	1
119-20220-74	nonvar	1	1.8921×10^{-6}	1	0.2619
119-20220-31	nonvar	1.4579×10^{-62}	1.2460×10^{-7}	1	1
119-20221-138	nonvar	2.4481×10^{-6}	1.2171×10^{-7}	1	1
119-20221-173	nonvar	0.00031	7.8707×10^{-6}	1	1
119-20220-151	nonvar	1	1	1.36×10^{-11}	1
119-20222-2548	nonvar	1	1	0.01131	1
119-20091-3964	nonvar	3.7890×10^{-5}	1	2.2×10^{-16}	0.9998
119-20221-298	nonvar	1.6877×10^{-13}	1	1.495×10^{-9}	1
119-20091-3867	nonvar	1	1.1133×10^{-45}	1.393×10^{-10}	0.9999
119-20091-3858	nonvar	2.7089×10^{-13}	4.4582×10^{-39}	0.0003515	0.6198
119-20220-63	nonvar	1.9562×10^{-49}	3.450×10^{-144}	2.2×10^{-16}	2.2×10^{-16}