9-15-2016

Synergistic Effects of Phase Folding and Wavelet Denoising with Applications in Light Curve Analysis

Andrew M. Armstrong

SYNERGISTIC EFFECTS OF PHASE FOLDING AND WAVELET DENOISING WITH APPLICATIONS IN LIGHT CURVE ANALYSIS

DISSERTATION

Andrew M. Armstrong, Captain, USAF

AFIT-ENC-DS-16-S-001

DEPARTMENT OF THE AIR FORCE
AIR UNIVERSITY

AIR FORCE INSTITUTE OF TECHNOLOGY

Wright-Patterson Air Force Base, Ohio

AFIT-ENC-DS-16-S-001

SYNERGISTIC EFFECTS OF PHASE FOLDING AND WAVELET DENOISING WITH APPLICATIONS IN LIGHT CURVE ANALYSIS

DISSERTATION

Presented to the Faculty

Graduate School of Engineering and Management

Air Force Institute of Technology

Air University

Air Education and Training Command

in Partial Fulfillment of the Requirements for the

Degree of Doctorate of Philosophy in Applied Mathematics

Andrew M. Armstrong, MS

Captain, USAF

September 2016

AFIT-ENC-DS-16-S-001

SYNERGISTIC EFFECTS OF PHASE FOLDING AND WAVELET DENOISING WITH APPLICATIONS IN LIGHT CURVE ANALYSIS

DISSERTATION

Andrew M. Armstrong, MS
Captain, USAF

Committee Membership:

Dr. Christine Schubert Kabban
Chair

Lt Col Ryan Kappedal, PhD
Member

Maj Charlton David Lewis, PhD
Member

Dr. Mark Oxley
Member

ADEDEJI B. BADIRU, PhD
Dean, Graduate School of Engineering
and Management

Abstract

The growing size of cosmological data sets is causing the current human-centric approach to cosmology to become impractical. Autonomous data analysis techniques need to be developed in order to advance the field of cosmology. This research examines the benefits of combining two signal analysis techniques, namely phase folding and wavelet denoising, into a newly-developed suite of autonomous light curve analysis tools which includes aspects of component extraction and period detection. The improvements these tools provide, with respect to autonomy and signal quality, are demonstrated using both simulated and real-world light curve data. Although applied to light curve data, the suite of tools developed in this dissertation are advantageous to the processing, modeling, or extractions to any periodic signal analysis.

Table of Contents

Page

Page

List of Figures

SYNERGISTIC EFFECTS OF PHASE FOLDING AND WAVELET DENOISING WITH APPLICATIONS IN LIGHT CURVE ANALYSIS

I. Introduction

Moore's law, and other exponential growth patterns in technology, has given rise to a quantity of data so large that standard data analysis techniques have proven inadequate. The need to develop methodologies capable of processing these vast data stores has lead to the development of a new discipline called Big Data. Scientists at the forefront of Big Data research attempt to merge statistical theory with the raw power of machine learning in order to develop tools suitable to the task. Many big data scientists specialize in a more narrow area of study, creating subdisciplines. One such subdiscipline, astrostatistics, attempts to apply big data tools to astronomical data in an effort to gain insight into the foundations of the universe.

Though a large number of astronomical data sources exist, many astrostatisticians focus on a specific data type called a light curve. A light curve is a graph of light intensity of an astronomical object, or set of objects, over time. Using light curves, it is possible to not only classify a large number of astronomical objects, but also to calculate some of their inherent properties such as size, location, and even composition. The simple nature of light curve data, coupled with its potential for deep understanding of underlying phenomenon, has lead to the creation of massive light curve repositories. These light curve repositories are often open to the public and easily accessible for analysis.

One of the most common techniques of light curve analysis is phase folding. Phase folding takes advantage of the periodic nature of many light curves by overlaying the

successive periods of a signal on top of each other. This process creates a data dense signal composed of a single period, which helps to aid in further analysis. Once folded, light curves can be studied using a large number of techniques.

A promising, but seldomly employed, light curve analysis tool is called wavelet analysis. Wavelets are mathematical tools which can be used to implement a variety of different transforms. Wavelet transforms are used to process data for the purposes of analysis, denoising, and compression in a wide range of fields. Unfortunately, much of light curve analysis is still completed by hand, albeit aided at times with mathematical tools. With the rise of big data sets, the current human centric approach is no longer efficient or effective [24].

The goal of this research is to utilize synergistic properties of phase folding and wavelet analysis, specifically wavelet denoising, with respect to light curve analysis in order to create an autonomous suite of tools to accomplish light curve analysis. Although applied to light curve data, the suite of tools developed in this dissertation are advantageous to the processing, modeling and extraction of any periodic signal. The goal is broken down into four primary objectives.

The first objective is to establish the mathematical foundations of phase folding and to prove mathematically and demonstrate empirically the benefits of combining phase folding with wavelet denoising. Objective two is to develop an automated method of signal decomposition by isolating components with phase folding and extracting these components using wavelet techniques. Objective three is to create an automated period detection algorithm using wavelets and phase folding when component periods are unknown. Lastly, objective four is to demonstrate the viability of these tools and processes in the analysis of real-world light curve data.

This dissertation has 9 chapters. Chapter 2 provides a general background on astrostatistics and an in-depth look at modern light curve analysis techniques. In

Chapter 3, the mathematical theory underlying wavelets is discussed as well as the wavelet denoising process. Chapter 4 focuses on the properties of phase folding along with the current methods of period determination. The process and results of combining phase folding and wavelet denoising are discussed in Chapter 5. Component extraction and period detection are covered in Chapters 6 and 7, respectively. Chapter 8 will apply the newly developed tools to real-world light curve data and Chapter 9 will provide a summary of the work and suggest some areas of future research.

II. Astrostatistics

Historically, astronomy has been a data-driven science. Larger and more precise data sets have led to the development and refinement of more accurate cosmological models of the universe. These data sets continue to grow in size, and have been traditionally managed by astronomers. With the advent of modern data gathering methods such as the Sloan Digital Sky Survey (SDSS), the Kepler satellite, and the forthcoming Large Synoptic Survey Telescope (LSST), the human-centric approach to astronomy is becoming strained [13, 24, 25, 63]. More than ever, astronomers have been forced to look to statisticians and machine-learning experts for more efficient data analysis methods. This merging of specialties has given rise to a new scientific discipline called Astrostatistics.

The roll of astrostatistics is to to test and refine cosmological theories using the raw data gathered by astronomers [50]. When presented with raw data, astrostatisticians first attempt to detect and classify known astronomical objects as well as flag unknown phenomenon for further investigation. Once objects are identified, object specific information such as various orbital parameters are calculated. This information is then combined with that of similar objects to develop canonical parameters, which are parameters devoted to summarizing a set of data (similar to the concept of a sufficient statistic). These canonical parameters can then be used to refine and test various cosmological theories.

One of the most public examples of astrostatistics is the recent rapid identification of exoplanets, that is, planets orbiting distant stars. The discovery of these new exoplanets is due largely to the analysis of data provided by the Kepler spacecraft, which was launched in 2009 [25]. Kepler was designed to record the brightness of more than 145,000 stars over an extended period of time. The plot of this brightness over time, called a light curve, is then analyzed in order to detect periodic dimming

of the star. Such dimming could indicate a potential planet. Through this transit method of detection, scientists using Kepler data have been able to identify over 1,000 confirmed exoplanets [38].

This chapter will present the relevant background and motivation for an improvement in light curve classification methodology. The roll of astrostatistics in modern cosmology will be discussed first. Next will be an overview of light curves and their impact on astrostatistics. Concluding the chapter will be a discussion of the shortcomings of light curve analysis and an investigation into how such analysis can be improved.

2.1 Cosmological Foundations

The primary motivation for improved automated classification methods for astronomical phenomenon is their use in astrostatistics for testing different cosmological models [50]. In order to understand the impact improved methods would have on cosmology, a basic understanding of cosmological models is important. Once a cosmological foundation has been established, the ability of astrostatistics to both refine and test these models will be discussed individually.

Modern Cosmological Models.

In 1916 Albert Einstein published his seminal paper, "The Foundations of the General Theory of Relativity" which fundamentally changed how cosmologist viewed the universe [11]. Einstein was able to develop what are now known as Einstein's Field Equations, which describe the geometric effects that matter and radiation have on spacetime and how the curvature of spacetime effects matter and radiation. These field equations are nonlinear and extremely difficult to solve, but have provided the mathematical foundations to some of the most well know cosmological phenomenon.

For example, the first nontrivial solution to the field equations was derived by Karl Schwarzschild in 1916 [52]. This solution was the first step in our modern understanding of black holes.

Einstein went on to apply his field equations to the universe as a whole, where he first described his cosmological constant term denoted by Λ [12]. This term was included in the new iteration of his equations as a result of his assumption of a static universe. Five years later, Alexander Friedmann would go on to publish another exact solution to Einstein's original field equations, without the use of the cosmological constant, which described an expanding universe [16, 17]. In 1929, Edwin Hubble would publish his work indicating that the universe is, in fact, expanding, giving credit to Friedmann's solution [22].

Friedmann's field equation solution would not gain notoriety until it was independently derived by Georges Lematre in 1927, two years before Hubble's paper [31]. Though Friedmann was the first to derive the solution allowing for an expanding universe, Lematre was the first to propose that it was actually occurring, in what is now known as the Big Bang Theory. This model of the universe would continue to be used until the late 90s when it was discovered that the expansion of the universe is actually accelerating [48]. This discovery, along with the discovery of the cosmic microwave background radiation, necessitated the re-inclusion of the cosmological constant into Friedmann and Lematre's equation [3, 43, 44].

The Friedmann-Lematre equation with the cosmological constant is now widely accepted as the most accurate cosmological model to date, and has been dubbed the Λ Cold Dark Matter (ΛCDM) model [50]. The cold dark matter portion of the name comes from the hypothetical matter which is believed to be responsible for the formation of galaxies in a sparse universe [4]. The cosmological constant was redefined to refer to the dark energy believed to exist in all of space [42, 48]. There are several

versions of the ΛCDM model. One of the most commonly referenced models is the seven parameter model, detailed in Table 1. In this model, the underlying structure of the universe can be described using only seven values.

Table 1. Seven Parameters of the ΛCDM Model [50].

Parameter	Description	MLE	68% Interval
Ω_b	Baryonic matter density	0.0490	0.0490±0.0073
Ω_m	Total matter density	0.3175	0.314±0.020
Ω_Λ	Dark energy density	0.6825	0.686±0.020
H_0	Rate of expansion (km/s/Mpc)	67.11	67.4±1.4
τ	The optical depth	0.0925	0.097±0.038
A_s	Amplitude of initial spectrum ($*10^9$)	2.215	2.23±0.16
n_s	Spectral index of initial spectrum	0.9624	0.9616±0.0094

The parameters provided in Table 1 do not constitute the entire set of parameters for all versions of the ΛCDM model. One possible modification to the ΛCDM model was proposed in 1981 by Alan Guth [19]. In an attempt to resolve several long standing problems associated with the standard model, Guth suggested that the universe underwent a period of exponential growth shortly after the big bang. The exponential growth of the universe can be accounted for in the model by modifying the assumptions on entropy in the early universe, causing a small change in the ΛCDM model.

Although the ΛCDM model and its many variations appear to be the most promising, there are a wide variety of other cosmological models that have been proposed [15]. Alternative solutions to Einstein's field equations have been developed into models, along with variations on Newtonian gravity called Modified Newtonian Dynamics (MOND). Some modern models even reject the concept of a big bang, such as the

steady state model [21]. Each model must be tested and refined in order to determine which model best represents reality. Thus, there are two primary objectives of astrostatistics. The first objective is to test different models of the universe. The second objective of astrostatistics is to refine the parameters of the various cosmological models, such as those in Table 1, to improve the model's accuracy.

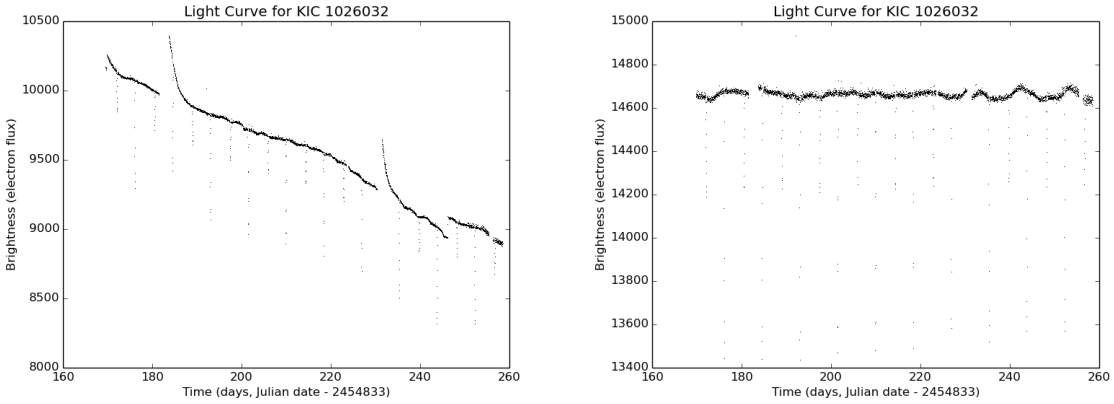
Refinement of Cosmological Parameters.

One of the primary objectives of astrostatistics is to refine the estimates of the cosmological parameters that shape the universe [50]. In order to accomplish this lofty goal, the raw data from individual objects are used to derive their object specific parameters. Once the object specific parameters for a large enough sample are collected, they are summarized by various canonical parameters. Finally, the canonical parameters are incorporated into a cosmological model in order to provide an estimate for the cosmological parameters. The framework described in [50] uses statistical analysis to create a direct link between the raw data and the testing of cosmological models.

The raw observables for a given astronomical phenomenon are generally recordings of the intensity of different wavelengths of the electromagnetic spectrum at a given time or over a set period. These intensity recordings can be broken down into spectroscopic and photometric data. The recording of spectroscopic data is time consuming and cost prohibitive, making it not as common as photometric data, even though spectroscopic data can be richer in information [13]. The photometric data is often recorded over a set spectral range for a period of time and can be used to create light curve data. Both types of data may be used in the estimation process of parameters either individually or jointly [26].

In the conversion process from raw observables to object specific parameters, first, data must be preprocessed. In the case of a spectroscopic measurement, preprocessing

the data requires removal of the red shift and other noise producing variables that may be included in the spectrum. For light curves, some detrending process is usually necessary. The detrending process is used to remove the gradual changes over time of the detection system's position relative to the object of interest. The effects of detrending on the light curve can be seen in the notional (Figure 1a) and the detrended light curve (Figure 1b).



(a) Original Light Curve **(b) Detrended Light Curve**

Figure 1. Detrending Example.

The derivation of object specific parameters includes a detection and classification component. For example, a light curve with a periodic changing of intensity could be any number of complex astronomical systems including a star with a single planet, a pulsar, or an EB system [5, 9, 65]. Without accurate classification, the derived object parameters will be meaningless and even potentially harmful to future analysis if they are assumed accurate. Once the object's specific classification has been determined, its parameters can be derived using a variety of different techniques such as the Wilson-Devinney approach discussed later [66].

When the object specific parameters for a sufficient number of related objects have been determined, they can be summarized into canonical parameters. This summary

is similar in concept to a sufficient statistic in that all of the gathered information on a parameter is condensed into the smallest set of data points. A key example of this summary process is the relationship between absolute magnitude of quasars to their redshift, as seen in Figure 2.

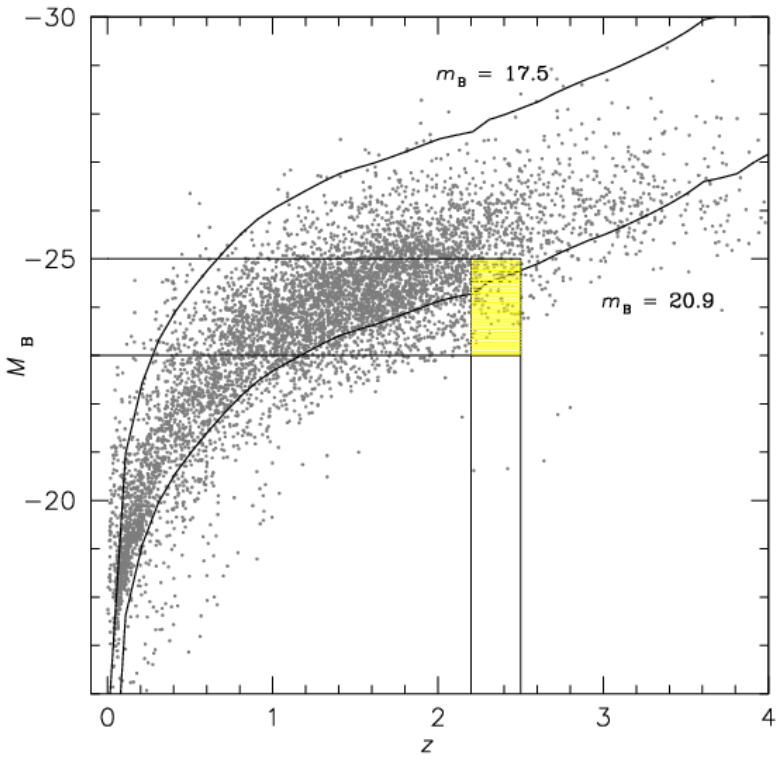


Figure 2. Distribution of Quasars vs Red Shift [62].

This information can be condensed into a distribution modeling the relationship between quasar brightness and redshift which can provide insight into the expansion rate of the universe.

To create better estimates for the cosmological parameters, the relationship between the various canonical parameters is described using a cosmological model. Estimates of the cosmological parameters can be produced through detailed analysis of the canonical parameters. This hierarchical process of converting raw data into

cosmological parameters has given rise to the use of Bayesian statistical methods [50]. Bayesian methodology places probability distributions on each cosmological parameter, as opposed to defining each parameter with a single value. The cosmological parameters are then refined using information derived from the canonical and object specific parameters; information which may be expressed either as specific values or distributions themselves.

Testing of Cosmological Models.

The second primary objective of astrostatistics is to test different cosmological models. As discussed previously, a wide variety of models describing the physical universe exists in the literature, all of which need to be compared to determine which most accurately describes the universe. Cosmological models are commonly tested by generating various predictions based on the mathematical foundations of each model [64]. These predictions are then compared to the observable universe and evaluated based on their similarities.

One of the most basic examples of this testing procedure is in the evaluation of Einstein's original model of the universe. Einstein's original model assumed that the universe existed in a steady state, that is, the universe had always been the way it is now [42]. When Hubble observed that the universe was expanding, Einstein's model no longer accurately represented the universe and a revised model gained popularity. Revolutionary discoveries, such as Hubble's discovery of an expanding universe, are not sufficient to test all possible cosmological models due to the scarcity of such groundbreaking discoveries and plethora of potential models. Therefore, more refined methods are necessary for testing models with similar predictions.

High performance computers provide another way to test cosmological models that was unavailable before such computers gained prevalence [23, 64]. Cosmological

models can now be simulated at various levels and complexities to determine more specific predictions which can then be analyzed. A common approach is to simulate the universe's evolution based on a given model on a large scale from its beginning to its current state. Based on the sophistication of the simulation, any number of different astronomical predictions can be tested. One such example is the testing of inflation, which is the leading candidate to explain why the universe is expanding.

Another test compares the predicted distribution of galaxies to what is observed in the cosmos. For example, the ΛCDM model predicts that galaxies formed in the early universe because dark matter was prevalent along the edges of large gas clouds [54]. These large gas clouds would then combine leaving a halo of dark matter around the edges. This dark matter would then cause the gas clouds to rotate at an accelerated pace causing the formation of thin disk like galaxies. Cosmologists can then use this theory to analyze the universe for the prevalence of such galaxy structures as a test for accuracy of the ΛCDM model.

Analyzing the composition of different galaxies (such as star density, black hole density, and luminous matter) is another test commonly applied to cosmological models. One of the most commonly used phenomenon are eclipsing binary (EB) stars [26]. EBs make for excellent testing systems because of the wide variety of testable attributes which can be accurately determined including age, luminosity, composition, distance, and even their prevalence in a galaxy [56, 40].

The fundamental attributes of individual stars described above serve as the foundational elements for calculating galactic properties, such as the distance between galaxies. Galactic properties can then be used to test various predictions of stellar evolutionary models. Our sun is the only star for which methods exist to determine its various parameters directly [56]. While methods have been developed to determine other parameters of stars from light curves and other data sources, no means of

determining a star's mass to a sufficiently small confidence level has been found for systems containing only a single star. Usually, the presence of a second observable body with other derivable parameters, such as its composition, allows for smaller confidence intervals to be placed on each star's mass [26]. The determination of a star's mass serves a critical role in the determination of stellar properties such as its distance, and can therefore be used to discern galactic distances. As such, EBs fill an important roll in the testing of stellar evolutionary models.

2.2 Light Curve Analysis

A light curve is the recording of a cosmological object's, or group of objects', brightness over time. Using light curves a large variety of cosmological objects, such as EBs, can be detected, classified, and studied. This section discusses the role of light curves in astrostatistics and provides a brief overview of their analysis process.

Role of Light Curves.

Light curves are a powerful tool in astrostatistics due largely to their abundance and far reaching applications. The abundance of light curves is due to the ease in which they are gathered and the low costs associated with thier attainment. The American Association of Variable Star Observers (AAVSO), a non-profit association of both professional and amateur astronomers who create and study light curves, have people gathering light curve data with tools as simple as a watch and binoculars [53]. More advance and precise systems also exist for capturing light curve data, the most well known is the Kepler satellite [25].

Kepler was launched on March 6, 2009 and began gathering data on May 13th of that same year with the primary mission of detecting earth-sized planets in the habitable zone of other stars. To accomplish this goal, Kepler gathers brightness

measurements of $\approx 156,000$ stars every 29.4 minuets over three month intervals. These data are then used to create light curves which can be used to detect exoplanets.

Scientists search for light curves which show a periodic dip in brightness over time. In the case of an exoplanet, this dip in brightness is caused by the planet passing in front of, or eclipsing, its host star relative to Kepler. Once detected, the planet's orbit and size can be determined from the frequency and intensity of these dips when combined with orther information on the system. To date, Kepler data has been used to detected and confirm the existence of over $2,300$ planets.

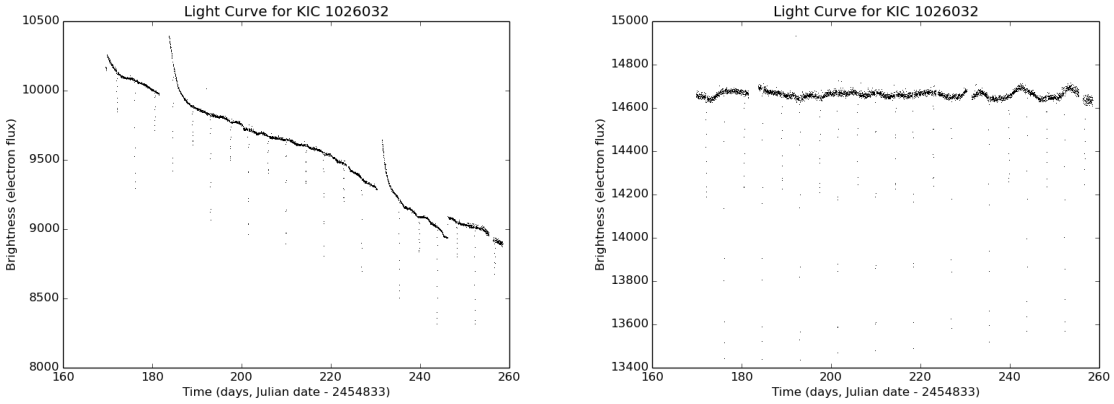
The detection and study of exoplanets is not the lone use of light curves. In [58] the light curves of 32 types of variable stars are discussed, broken down into 6 different classes. The classes of variable stars include cataclysmic, or exploding, stars such as supernove, pulsating stars which periodically expand and contract, and EB systems which are composed of two stars which orbit each other. The stars/systems with potentially the largest impact on modern cosmology are supernove and EB systems.

Supernove and eclipsing binary systems can act, as what astronomers refer to, as standard candles. This means that the actual brightness of the systems can be empirically determined. With a known brightness, the inverse square law can be then used to determine the distance to the system in question. This information can then be used to derive canonical parameters such as the distribution of matter in the universe which can then be used to test and refine cosmological models.

In order to be of use, light curves typically go through a rigorous preprocessing. Light curve preprocessing generally consists of three key steps: detrending, filtering, and smoothing. Once preprocessed, light curves are then classified using a wide variety of techniques. The remainder of this section will discuss each of these steps.

Detrending.

Detrending is the process used by cosmologists to remove the error caused by velocity aberration [61]. For ground based systems, such as SDSS or the Catalina Sky Survey (CSS), this error is a combination of the earth's rotation on its axis as well as the earth's rotation around the sun. For space based systems, such as Kepler, the error is related to Kepler's specific orbit. When looking at a light curve, this error is readily apparent (see Figure 1a) as a gradual change in the overall intensity, or flux, over time. Figure 1b shows the same light curve after detrending, where the effects of velocity aberration have been removed. Once this error is removed, classification algorithms may be more effective because the algorithms are operating on a more accurate representation of the phenomenon.



(a) Original Light Curve **(b) Detrended Light Curve**

Figure 1. Detrending Example.

The actual detrending process varies according to the gathering mechanism, the object of interest, and the cosmologist performing the detrending. One of the easiest methods in which to remove the error trend is through the subtraction of the best fit linear regression model [7]. In this instance, the experimenter will perform a basic linear regression analysis to find the slope (b) and the intercept (a) of the best fitting

line for the data as a whole, using flux as the dependent variable ($\mathbb{Y}$) and time as the independent variable ($\mathbb{X}$). The fitted function, $\hat{\mathbb{Y}} = a + b\mathbb{X}$, is then subtracted from the original data. Although this method can be effective at removing the error trend, it is not commonly used because the linear model of the error is a poor representation of the phenomenon causing such error.

An alternative approach to detrending employed by the National Aeronautics and Space Administration's (NASA) team working on the Kepler data is the systematic removal of Cotrending Basis Vectors (CBVs) [28]. The CBVs are the 16 best-fit vectors which represent the most common features in a large set of targets over a given period of time. A Bayesian method called Maximum A Posteriori (MAP) is used to calculate the CBVs in lieu of a least squares approach in order to prevent overfitting. The CBVs are ranked on their overall error contribution, and subsequently subtracted from the target light curves.

Unfortunately, as discussed in [35] , the detrending process used by NASA's Kepler team may remove pertinent data. Specifically, in the search for variable astronomical phenomenon, where an object's flux changes over time, NASA's detrended data may not include periodic signals whose thresholds are not met, and are hence excluded. Therefore, in the search for EBs which are considered variable phenomenon, another detrending process may be necessary. An alternative detrending process commonly used in the search for EBs is the sigma-clipping algorithm derived [55].

[55]'s sigma-clipping algorithm uses the light curve and associated uncertainty for the time-series flux data points in order to fit a Legendre polynomial of order l to the flux using the generating function of:

$$P_l(x) = \frac{1}{2^l l!} \frac{d^l}{dx^l} \left[\left(x^2 - 1 \right)^l \right]. \tag{1}$$

Data points which reside outside one standard deviation of the fit are iteratively

removed until no more remain. The function then returns a normalized flux value with a new uncertainty level for each input point. This approach to detrending allows for fitting a wide variation of functions by changing the order of the Legendre polynomial (l) and by setting the threshold of the clipping function (σ). For example, when searching for EBs, [40] selected a 10^{th} order Legendre polynomial, seen in Equation 2, with a sigma-clipping threshold of $(-1\sigma, 3\sigma)$.

$$P_{10}(x) = \frac{1}{256}\left(46,189x^{10} - 109,395x^{8} + 90,090x^{6} - 30,030x^{4} - 3,465x^{2} - 63\right). \quad (2)$$

Preliminary Filter.

Once the detrending process has been completed, the light curves are often filtered to remove errors or objects which do not appear significant. Depending on the application and its implementation, filtering can significantly reduce the overall computation time of the system. Although there are a wide variety of filtering methods, only a few of the most common will be discussed after describing uses of the filtering process.

One use of the filtering process is to remove artifacts from the data set. In the context of time domain astronomy, an artifact is an error in the data provided by the collection mechanism. There are a large variety of possible instrument errors, many of which are difficult to identify. The causes of these errors can include such things as cosmic rays, dust accumulation, or simply the movement of the device [61]. For the Kepler satellite, confidence levels for each object over each collection interval are provided. This information aids in the removal of artifacts, but may not be sufficient to remove all artifacts [28].

Another common use of the filtering process is the removal of known erroneous

entities from a data set. For example, this can include variations in light curves caused by natural and unnatural satellites passing in front of the collection apparatus during its collection period. These data points can be removed based on information provided by NASA, which catalogs the various objects in the solar system [39]. Another example of filtering objects may occur during the search for supernovae, where an astronomer may remove potential supernovae candidates which do not appear to be originating from a known galaxy [41].

Filtering entire light curves from the database strictly due to the level of noise in their signal is also a common practice. When searching for slight variations in flux to indicate the presence of an exoplanet, an astronomer may not consider light curves with excessive noise since the small variations of interest may be unrecognizable. The most commonly used method of filtering noisy light curves is by setting a signal to noise ratio (SNR) threshold. Light curves with SNRs below a given value are discarded [1].

Finally, specific filters may be used for specific applications. In [40], a filter is used to identify systems conforming to certain periodic characteristics. The periodic signals are detected by a folding function devised in [51]. The periodic filter eliminates light curves which do not exhibit significant non-random periodicity, the periodic characteristic of interest.

Smoothers and Splines.

Smoothers and data reduction techniques are common filtering methods used to describe patterns in phenomenon and reduce computation time in many classification systems. In statistics, smoothers are functions which may be used to remove excess noise from a signal in order to get a cleaner representation of the underlying function or pattern. One of the simplest implementations of a smoother is a moving average.

The most basic moving average smoother uses the average of an equal number of values on either side of a data point in order to calculate a new value. A moving average will lower the impact of high frequency variation over the entire data set.

Curve fitting is similar to smoothing in that it is an attempt to describe the underlying function for data, however, it may or may not make assumptions as to the underlying structure of the data. For instance, curve fitting may refer to fitting a simple regression model or it may model the data as complex polynomial functions in order to account for a certain level of variability. As with all curve fitting, it is assumed that the data is easily represented by the fitted function, which might not always be the case.

To overcome some of the weaknesses in curve fitting, many practitioners incorporate splines. Splines are used to model a data set as a piecewise sequence of polynomials connected by knots [46, 40]. These knots are the transition points from one polynomial to another. Determining the knot points can itself be a difficult problem and has inherent issues, such as the function not being differentiable where the polynomials connect.

These smoothing methods can be used independently or in conjunction with each other, and with various data reduction techniques. Data reduction can be a very important process for data sets used in most astronomical classification systems because it can significantly impact the computational time and accuracy of the classification function. Although many techniques exist for reducing data, one of the simplest is called binning. One implementation of binning is to represent every n data points as a single number, such as their average, in order to produce an n-fold reduction in the data size. The shortfalls of binning are similar to those for the functional form methods or smoothers over a large range in that a significant and potentially important amount of data may be lost in the process.

Classification Procedure.

One of the greatest challenges in light curve analysis is accurate classification of the represented system. This difficulty, combined with the growing number of light curves, led to the increasing adoption of autonomous classification systems. These systems are often concerned with classifying light curves in a very small category. For example, [40] is concerned only with classifying three types of EB systems and is not concerned with any other type of variable star.

In [34] light curves were phase folded, which means that data from each period are overlayed to create a single period with denser sampling, and fitted using a program called `polyfit`. `polyfit` creates m piece wise polynomial segments of order n which best fit the data. Once fit, the models are then sampled at 1000 equidistant points, and analyzed using Locally Linear Embedding (LLE). LLE is similar to Principal Component Analysis (PCA) in that LLE is a dimensionaly reduction technique, however it determines relationships between points as opposed to global properties. In effect, this technique creates a lower dimensional space in which to perform a nearest neighbor classification. Though able to accurately classify the majority of the light curves, this approach requires the pruning of over a quarter of its dataset due to the classification mechanism, mainly the inability of `polyfit` to accurately fit the light curve.

A more far-reaching scheme was developed in [2] . Armstrong et al. attempted to classify $68,910$ light curves into 7 different classes of variable stars to include δ Scuti, λ Doradus, RR Lyrae, two types of EBs, noise, and other periodic variables. To accomplish this, the light curves were each phase folded, then several features were derived to include period, amplitude, and standard deviation. Each feature set was then added to a self organizing map as a means of dimensionality reduction, thereby creating hybrid features. The light curves were then classified using these hybrid

features using a random forest classification process. The random forest algorithm then returned a probability of class membership for each light curve from a set of decision trees. Overall the system had a successful classification rate of 92%, however one class was only accurately classified 76% of the time.

III. Wavelets

In linear algebra, a basis for a linear space can be used to represent any vector in the space using an unique linear combination of these basis vectors. The selection of such a basis can have a significant impact on the tools and techniques available when analyzing a given signal in the linear space of signals. Representing a signal using a specific set of basis vectors can provide greater insight into the signal's behavior and composition. One of the most frequently used set of basis vectors are those employed by Fourier analysis, that is, the sine and cosine. Using these basis vectors as a means of transformation, a function in the time domain is converted into a function in the frequency domain.

The Fourier basis, while very powerful, is only capable of analyzing a signal in the frequency domain. This means that time specific information of a signal, when represented using the Fourier basis, is essentially lost. Wavelets seek to fill the gap between a purely time-based and frequency-based analysis by representing a signal using a basis which incorporates both time and scale, which can be related to frequency, information simultaneously. This combined representation incurs the same loss as that of the Fourier basis, but distributed across both the time and frequency domain. These losses can be tuned in order to optimize the signal representation for a given application.

This hybrid domain grants the wavelet transform many advantages over that of the different Fourier transforms. One of the most important advantages is the ability to locate signal features in both time and frequency. This ability allows the analysis of non-stationary signals, signals which change their component frequencies over time. Coupled with the ability to perfectly reconstruct a signal from its wavelet coefficients, the wavelet transform is a powerful tool in both compression and denoising applications.

This chapter will cover the basic foundations of wavelet theory and the wavelet tools used throughout the rest of this dissertation. The Fourier series and transforms, as well as their relevant strengths and weaknesses, will be discussed first. Following this will be an in depth look at wavelet theory and the advantages wavelet analysis has over Fourier analysis. Concluding this chapter will be an overview of several wavelet tools, with a special emphasis on the denoising applications of wavelets.

3.1 Wavelet Foundations

Wavelet were first created in 1910 by Haar and popularized in 1992 by Daubechies with roots in the Fourier transform [6, 20]. Wavelets can be used to create a basis for signals which can be used in applications as diverse as classification, compression, and denoising, yet have rarely, if ever, been applied to the study of light curves. This section discusses the underlying theory of wavelets with special emphasis placed on the foundations needed to understand wavelet denoising.

Fourier Transform.

The Fourier transform converts signals from the time domain into the frequency domain. This conversion is accomplished by breaking a signal down into its component frequencies and their respective amplitudes. The new signal representation allows the detection of frequency changes that would be difficult to detect using the original domain. Given a signal defined over all time, $t \in \mathbb{R}$, that has finite energy, $f(t)$, the Fourier transform produces the frequency representation of the signal, $\hat{f}(\omega)$ by

$$\hat{f}(\omega) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(t) \mathrm{e}^{-i\omega t} dt. \tag{3}$$

The Continuous Fourier Transform (CFT) is used for continuous signals, however real-world signal analysis is more often concerned with discrete time signals. This

process can then be easily reversed using the inverse Fourier transform given as

$$f(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \hat{f}(\omega) \mathrm{e}^{i\omega t} d\omega. \tag{4}$$

To perform a Fourier transform on a discrete signal requires the use of a discrete version of the Fourier transform, called the Discrete Fourier Transform (DFT). The formula for calculating the DFT for a signal f of length N, where n is the sample and k is the current frequency under consideration (0 hertz up to $N-1$ hertz), is given by

$$\hat{f}[k] = \sum_{n=0}^{N-1} f[n] \mathrm{e}^{\frac{-i2\pi kn}{N}}, \tag{5}$$

and its inverse transform is given by

$$f[n] = \frac{1}{N} \sum_{k=0}^{N-1} \hat{f}[k] \mathrm{e}^{\frac{i2\pi kn}{N}}, \tag{6}$$

where $f[n]$ denotes $f(t_n)$.

The Fourier transform provides useful frequency information about the signal, but at the cost of losing all time domain information. This trade off prevents the detection of frequency changes in the signal since the information returned from the Fourier transform is simply an average over the whole signal. This compromise is actually a result of the Heisenberg uncertainty principle as applied to signal processing, which is referred to as the Gabor limit [18].

The Gabor limit states that signals cannot have arbitrarily small precision in both time and frequency simultaneously. A useful tool for describing and visualizing this relationship is a Heisenberg rectangle which is of size $\sigma_t \times \sigma_\omega$ and has an area of $\frac{1}{4\pi}$ (called the Gabor limit). This means that when analyzing a signal, the product of the standard deviation in miliseconds, σ_t, and the standard deviation in hertz, σ_ω,

must be greater than or equal to $\frac{1}{4\pi}$. An example of this relationship can be seen in Figure 3.

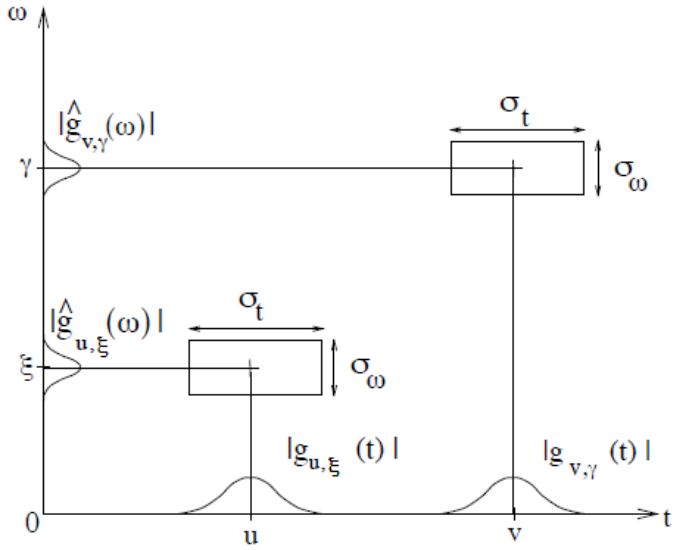


Figure 3. Heisenberg Rectangles [32].

The Windowed Fourier Transform (WFT) was developed to incorporate time and frequency information at the Gabor limit. In effect, a time and frequency width are chosen which conforms to the Gabor limit. These values are then used to create a grid of Heisenberg rectangles which span the time-frequency plane. The Fourier transform of each rectangle is then taken and the result stored in a matrix. Since only the relationship between σ_t and σ_ω is fixed, and not the actual values, the Heisenberg rectangles can be modified based on application.

The main downside to this approach of combining time and frequency information into a single transform is that of resolution. Once set, the dimensions for Heisenberg rectangles remain constant for the WFT, however, it is often desirable to vary them throughout the transform. A wide window (in time) provides better frequency resolution while a narrow window provides better time resolution. Therefore it is often more desirable have a wide rectangle at lower frequencies, to provide more accurate

time information, which narrows as the frequency increases to improve frequency resolution. This ability to change Heisenberg rectangle dimensions is one of the biggest reasons for the popularity of wavelet transforms.

Continuous Wavelet Transform.

A more adaptable alternative to the Fourier transform is the Wavelet transform. Given a mother wavelet, $\psi \in L^2(\mathbb{R})$, one first creates a family of normalized wavelets through dilations and translations of the mother wavelet with scale value $a > 0$ and translation value $b \in \mathbb{R}$ that is:

$$\psi^{a,b}(x) = |a|^{\frac{-1}{2}} \psi \left(\frac{x-b}{a} \right) \tag{7}$$

where

$$||\psi^{a,b}|| = ||\psi|| = 1 \ , \tag{8}$$

and $|| \cdot ||$ is the L^2-norm. The set of $\{\psi^{a,b} : a > 0, b \in \mathbb{R}\}$ is called a wavelet basis for $L^2(\mathbb{R})$. Given a mother wavelet function ψ, one can define a continuous wavelet transform (CWT), denoted by $\mathcal{W}_\psi$, of a function $f \in L^2(\mathbb{R})$, by

$$\mathcal{W}_\psi f(a,b) = \langle f, \psi^{a,b} \rangle = \int_{\mathbb{R}} f(x) |a|^{\frac{-1}{2}} \overline{\psi \left(\frac{x-b}{a} \right)} dx \ . \tag{9}$$

If certain restrictions are imposed on the choice of wavelets, namely the admissibility criteria given by

$$C_\psi = \int_{\mathbb{R}} \frac{|\hat{\psi}(\omega)|^2}{|\omega|} d\omega < \infty. \tag{10}$$

Then the inverse of the CWT can be calculated using

$$f = \frac{1}{C_\psi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \mathcal{W}_\psi f(a,b) \psi^{a,b} \frac{da db}{a^2}. \tag{11}$$

The CWT can be useful in some applications, such as signal analysis, however, the discrete wavelet transform (DWT) is more commonly employed for discrete data.

Multiresolution Analysis.

The DWT uses wavelets to create an orthonormal basis for the linear space of $\ell^2(\mathbb{Z})$ sequences which is used to decompose a discrete set of data into the coefficients. In order to construct such wavelets multiresolution analysis (MRA) is used [6]. The foundation of MRA is a sequence of nested subspaces, $\{V_j : j \in \mathbb{Z}\} \in L^2(\mathbb{R})$, which satisfy the following four conditions.

1.
$$\cdots \subset V_2 \subset V_1 \subset V_0 \subset V_{-1} \subset V_{-2} \cdots \subset L^2(\mathbb{R}) \tag{12}$$

2.
$$\overline{\bigcup_{j\in\mathbb{Z}} V_j} = L^2(\mathbb{R}) \tag{13}$$

3.
$$\bigcap_{j\in\mathbb{Z}} V_j = \{0\} \tag{14}$$

4.
$$f \in V_j \Leftrightarrow f(2^j \cdot) \in V_0 \tag{15}$$

Though many subspace sequences satisfy Equations 12 - 14, the final condition in Equation 15 requires each of the subspaces to be scaled versions of the central subspace, V_0. This means that for every function $f_j \in V_j$ that there exists $f_0 \in V_0$ such that f_j and f_0 are scaled versions of each other (same functional form but compressed or expanded).

In order to create a wavelet basis, two additional properties are required. The first is that V_0 is invariant under integer translations, which is formally expressed as

$$f \in V_0 \Rightarrow f(\cdot - n) \in V_0 \quad \forall n \in \mathbb{Z}. \tag{16}$$

The final assumption is that there exists $\phi \in V_0$ such that

$$\{\phi_{0,n} : n \in \mathbb{Z}\} \text{ is an orthonormal basis for } V_0, \tag{17}$$

for all $j, n \in \mathbb{Z}$, and

$$\phi_{j,n}(x) = 2^{\frac{-j}{2}} \phi(2^{-j}x - n). \tag{18}$$

Equations 15, 17, and 18 imply that $\{\phi_{j,n}; n \in \mathbb{Z}\}$ is an orthonormal basis for V_j $\forall j \in \mathbb{Z}$. In this case, ϕ is referred to as the scaling function. When the properties described in Equations 12-18 hold, then there exists a wavelet basis $\{\psi_{j,k} : j, k \in \mathbb{Z}\}$ of $L^2(\mathbb{R})$ where $\psi_{j,k} = 2^{\frac{-j}{2}} \psi(2^{-j}x - k)$ such that for all $f \in L^2(\mathbb{R})$

$$P_{j-1}f = P_j f + \sum_{k \in \mathbb{Z}} \langle f, \psi_{j,k} \rangle \psi_{j,k} \tag{19}$$

where P_j is the orthogonal projector onto V_j. .

To derive the wavelet ψ from the scaling function, first define the orthogonal complement of V_j in V_{j-1} as W_j. Therefore

$$V_{j-1} = V_j \oplus W_j, \tag{20}$$

where

$$W_j \perp W_k \text{ if } j \neq k. \tag{21}$$

It then follows that for $j < J$,

$$V_j = V_J \oplus \bigoplus_{k=0}^{J-j-1} W_{J-k}, \tag{22}$$

which means

$$L^2(\mathbb{R}) = \bigoplus_{j \in \mathbb{Z}} W_j. \tag{23}$$

This then creates a way to decompose $L^2(\mathbb{R})$ into mutually orthogonal subspaces. These properties can now be used to derive ψ from the scaling function ϕ. Given that $\phi \in V_0 \subset V_{-1}$ and $\{\phi_{-1,n} : n \in \mathbb{Z}\}$ constitutes an orthonormal basis in V_{-1}, then

$$\phi(x) = \sum_{n \in \mathbb{Z}} h_n \phi_{-1,n}(x) = \sum_{n \in \mathbb{Z}} h_n \sqrt{2} \phi(2x - n), \tag{24}$$

almost everywhere $x \in \mathbb{R}$ where

$$h_n = \langle \phi, \phi_{-1,n} \rangle \tag{25}$$

and

$$\sum_n |h_n|^2 = 1. \tag{26}$$

Define $g(x)$ such that

$$g(x) = \sum_n g_n \phi_{-1,n}(x) \tag{27}$$

where

$$g_n = (-1)^n h_{-n+1}. \tag{28}$$

So it follows that $g \in V_{-1}$, however, $g \notin V_0$, therefore $g \in W_0$. The function g is then called the mother wavelet, making

$$\psi = g = \sum_n (-1)^n h_{-n+1} \phi_{-1,n} \tag{29}$$

and

$$\psi(x) = \sqrt{2} \sum_n (-1)^n h_{-n+1} \phi(2x - n). \tag{30}$$

As an example of a MRA, select ϕ such that

$$\phi(x) = \begin{cases} 1 \text{ if } 0 \leq x < 1 \\ 0 \text{ otherwise.} \end{cases} \tag{31}$$

Then

$$h_k = \langle \phi, \phi_{1,k} \rangle = \sqrt{2} \int_{\mathbb{R}} \phi(x) \overline{\phi(2x - k)} dx = \begin{cases} \frac{1}{\sqrt{2}} & \text{if } k = 0, 1 \\ 0 & \text{otherwise.} \end{cases} \tag{32}$$

Therefore,

$$g_0 = (-1)^0 h_{-0+1} = h_1 = \frac{1}{\sqrt{2}} \tag{33}$$

and

$$g_1 = (-1)^1 h_{-1+1} = -h_0 = \frac{-1}{\sqrt{2}} \tag{34}$$

which makes

$$\psi = \frac{1}{\sqrt{2}} \phi_{-1,0} - \frac{1}{\sqrt{2}} \phi_{-1,1}, \tag{35}$$

or more simply

$$\psi(x) = \begin{cases} 1 & \text{if } 0 \leq x < \frac{1}{2} \\ -1 & \text{if } \frac{1}{2} \leq x < 1 \\ 0 & \text{otherwise.} \end{cases} \tag{36}$$

Discrete Wavelet Transform.

Performing a DWT can be thought of as passing a function through a series of filters. These filters are the byproduct of the MRA process discussed above for a specific scaling function and mother wavelet pair. In the process of calculating the mother wavelet from a given scaling function, the original space V_0 is broken down into two complementary subspaces V_1 and W_1. Subspace V_1 can then be broken down into complementary subspaces V_2 and W_2, a process which can be continued ad infinitum in a cascading fashion.

By definition, V_0 can be defined by its basis vectors $\phi_{0,n}$. A function $f \in V_0$ can then be represented as coeffients for the basis vectors of V_0, by

$$f = \sum_n c_n \phi_{0,n} \tag{37}$$

where

$$c_n = \langle f, \phi_{0,n} \rangle \tag{38}$$

are the coefficients. These coefficients can then be used to calculate coefficients for the basis vectors composing the complimentary subspaces V_1 and W_1 using ϕ and ψ, respectively. Given that

$$\psi = \sum_n g_n \phi_{-1,n} = \sum_n \langle \psi, \phi_{-1,n} \rangle \phi_{-1,n} = \sum_n (-1)^n h_{-n+1} \phi_{-1,n} \tag{39}$$

and

$$\psi_{j,k}(x) = 2^{-j/2}\psi(2^{-j}x - k) \tag{40}$$

$$= 2^{-j/2}\sum_n g_n 2^{1/2}\phi(2^{1-j}x - 2k - n) \tag{41}$$

$$= \sum_n g_{n-2k}\phi_{j-1,n}(x), \tag{42}$$

the coefficients for the basis vectors in $\psi_{1,k}$ can be calculated by

$$d_{1,k} = \langle f, \psi_{1,k}\rangle = \sum_n \overline{g_{n-2k}}\langle f, \phi_{0,n}\rangle. \tag{43}$$

Equation 43 can then be generalized for $\psi_{j,k}$ by

$$d_{j,k} = \langle f, \psi_{j,k}\rangle = \sum_n \overline{g_{n-2k}}\langle f, \phi_{j-1,n}\rangle. \tag{44}$$

Similarly the coefficients for the basis vectors of V_1 can be calculated. Starting with

$$\phi_{j,k} = 2^{-j/2}\phi(2^{-j}x - k) \tag{45}$$

$$= \sum_n h_{n-2k}\phi_{j-1,n}(x), \tag{46}$$

the basis coefficients can then be determined by

$$c_{j,k} = \langle f, \phi_{j,k}\rangle = \sum_n \overline{h_{n-2k}}\langle f, \phi_{j-1,n}\rangle. \tag{47}$$

This process can be continued by finding the coefficients for the basis vectors of V_2 and W_2 from the coefficients in V_1, however, the number of coefficients remains constant. Therefore, the number of coefficients in V_2 is half the number of coefficients in V_1, and equal to the number of coefficients in W_2. This is accomplished by a process

called downsampling, which means that only every other coefficient value is retained. The splitting of the coefficients into two complimentary sets can be thought of as breaking the function down into detail coefficients, $d_{j,k}$, and coarse or approximation coefficients, $c_{j,k}$.

Due to the way in which these subspaces were initially defined, h and g above constitute what is known as a quadrature mirror filter (QMF) pair. One of the most important byproducts of this QMF relationship is the potential for perfect reconstruction of the original function from its coefficients. This ability for perfect reconstruction holds regardless of the depth of the decomposition and allows for, among many things, wavelet denoising.

3.2 Wavelet Denoising

Wavelet denoising is powerful regression like technique which can accurately estimate an unknown function from noisy data. This section will detail this estimation process by borrowing heavily from [8] , one of the foundational works on the topic.

Spatially Adaptive Methods.

Define a signal $\mathbf{y} = (y_1, y_2, \ldots, y_n) \in \mathbb{R}^n$ such that

$$y_i = f(t_i) + e_i, \qquad i = 1, \ldots, n \tag{48}$$

where f is an unknown function to be estimated, $t_i = \frac{i}{n}$ and $\mathbf{e} = (e_1, e_2, \ldots, e_n)$ is independent and identically distributed normally as $N(0, \sigma^2)$. The estimate of f is denoted $\hat{f}$ such that $\mathbf{f} = (f(t_1), f(t_2), \ldots, f(t_n))$ and $\hat{\mathbf{f}} = (\hat{f}(t_1), \hat{f}(t_2), \ldots, \hat{f}(t_n))$. The performance of $\hat{\mathbf{f}}$ may be measured using the risk function $R(\hat{\mathbf{f}}, \mathbf{f})$ where E is the

expectation operator of the random variables and $||\cdot||$ denotes the ℓ^2 norm on $\mathbb{R}^n$ by

$$R(\hat{\mathbf{f}}, \mathbf{f}) = \frac{E||\hat{\mathbf{f}} - \mathbf{f}||^2}{n}. \tag{49}$$

The best estimate of $\mathbf{f}$ is the $\hat{\mathbf{f}}$ which has the smallest risk value, $\min_{\hat{\mathbf{f}}} R(\hat{\mathbf{f}}, \mathbf{f})$.

In order to construct $\hat{\mathbf{f}}$ the reconstruction formula $T(\mathbf{y}, \delta)$ is used, which is used to represent any possible estimate function to include ordinary least squares and wavelet denoising. The reconstruction formula takes as input the noisy data $\mathbf{y}$ and a spatial smoothing parameter δ. Let $d(\mathbf{y})$ denote the data-adaptive choice for δ. Then, the function for $\hat{\mathbf{f}}$ is given by:

$$\hat{\mathbf{f}} = T(\mathbf{y}, d(\mathbf{y})). \tag{50}$$

As an example, consider the piecewise polynomial (PP) reconstruction, $T_{PP(D)}(\mathbf{y}, \delta)$, which uses polynomials of degree D to estimate $\mathbf{f}$. In this instance, δ is a finite list of L real numbers which defines the partitions of $\mathbf{f}$. Let

$$1_{I_l} = \begin{cases} 1 & \text{if } t_i \in I_l \\ 0 & \text{if } t_i \notin I_l. \end{cases} \tag{51}$$

Then $\hat{\mathbf{f}}$ is estimated by

$$T_{PP(D)}(\mathbf{y}, \delta)(t_i) = \sum_{l=1}^{L} \hat{p}_l(t_i) 1_{I_l}(t_i) \tag{52}$$

where $\hat{p}_l$ is the least squares polynomial estimate for the interval l.

Given the risk function in Equation 49, define the ideal risk given a reconstruction

formula as

$$\mathcal{R}(T, \mathbf{f}) = \inf_{\delta > 0} R(T(\mathbf{y}, \delta), \mathbf{f}). \tag{53}$$

This is the lowest possible risk, which in effect, is the selection of the best possible δ, denoted δ^*.

Wavelet Reconstruction.

Given $\mathbf{y} = \{y_i\}_{i=1}^n$ where y_i is defined as in Equation 48 and $n = 2^{J+1}$, $J \in \mathbb{Z}_+$. Construct an $n \times n$ orthogonal wavelet matrix $\mathcal{W}$ with M vanishing moments, a support width of S, and the low-resolution cut-off (the lowest level of deconstruction) as j_0. The wavelet coefficients for $\mathbf{y}$, denoted by $\mathbf{w}$, are derived as

$$\mathbf{w} = \mathcal{W}\mathbf{y}, \tag{54}$$

which can be inverted to yield

$$\mathbf{y} = \mathcal{W}^T\mathbf{w}. \tag{55}$$

There are a total of $n = 2^{J+1}$ wavelet coefficients which are indexed dyadically by

$$w_{j,k} \qquad (j = 0, \ldots, J; k = 0, \ldots, 2^j - 1) \tag{56}$$

with the remaining element labeled as $w_{-1,0}$. These coefficients correspond to row basis vectors of $\mathcal{W}$, denoted $W_{j,k}$. This makes the inversion formula in Equation 55 equivalent to

$$y_i = \sum_j \sum_k w_{j,k} W_{j,k}(i). \tag{57}$$

The wavelet basis vectors, $W_{j,k}$ are then the same vectors described in Equation 40.

Given the above derivation of the wavelet basis vectors, two important properties are known:

1. $W_{j,k}$ has vanishing moments up to order M as long as $j \geq j_0$, and

2. Given that $j \geq j_0$, $W_{j,k}$ is supported in $[2^{J-j}(k-S), 2^{J-j}(k+S)]$.

Therefore, in order to determine if there is a significant change, as defined in [8], in f near time t, one need only look at the wavelet coefficients at levels $j = j_0, \ldots, J$ and spacial indicies k where $k2^{-j} \approx t$. If the absolute value of these coefficients is large, a significant change is said to have occurred at t.

Creating a reconstruction function using the above wavelet definition, define δ to be a set of indicies (j,k), which makes

$$\hat{\mathbf{f}} = T_{\text{Wave}}(\mathbf{y}, \delta) = \sum_{(j,k)\in\delta} w_{j,k} W_{j,k} \tag{58}$$

Ideal Risk.

To determine the ideal risk of the wavelet reconstruction function in Equation 58, first consider the derivation of the wavelet coefficients described in Equation 54. Since $\mathbf{y}$ is composed of the function f and normally distributed error, the expanded derivation becomes

$$\begin{aligned} \mathbf{w} &= \mathcal{W}\mathbf{y} && (59)\\ &= \mathcal{W}(\mathbf{f} + \mathbf{e}) && (60)\\ &= \mathcal{W}\mathbf{f} + \mathcal{W}\mathbf{e} && (61)\\ &= \theta + \mathbf{z}, && (62) \end{aligned}$$

where $\theta = \mathcal{W}\mathbf{f}$ and $\mathbf{z} = \mathcal{W}\mathbf{e}$. Therefore,

$$w_{j,k} = \theta_{j,k} + z_{j,k}. \tag{63}$$

This means that the "white noise" from $\mathbf{e}$ affects all the wavelet coefficients equally, while the effect of the function's coefficients are still limited to a small subset of coefficients.

Using Equation 63, define the δ which achieves the ideal risk, δ^*, as the indicies (j,k) where the wavelet coefficients for the function are nonzero, $\theta_{j,k} \neq 0$. In effect, this removes all of the coefficients for $\mathbf{y}$ which would only contribute noise. Then, Properties 1 and 2 in Subsection 3.2 put a limit on the size of δ^* since the coefficients $\theta_{j,k}$ of $\mathbf{f}$ will all be equal to 0 except:

1. The coefficients at coarse levels, that is, where $0 \leq j < j_0$, and;

2. The coefficients at levels $j_0 \leq j \leq J$ where a breakpoint of $\mathbf{f}$ is contained in the associated interval $[2^{-j}(k-S), 2^{-j}(k+S)]$.

Since there are only 2^{j_0} coefficients which could satisfy 1, and at most (# of breakpoints)$\times$ $(2S+1)$ coefficients which could satisfy 2, then

$$|\{(j,k) : \theta_{j,k} \neq 0\}| \leq 2^{j_0} + (J+1-j_0)(2S+1)L, \tag{64}$$

where L is the number of partitions and $|\cdot|$ is the cardinality.

Due to the orthogonality of the (W_j, k), the ideal risk for the wavelet reconstruction of $\mathbf{f}$ is

$$R(T(\mathbf{y}, \delta^*), \mathbf{f}) = \frac{|\delta^*|\sigma^2}{n} \tag{65}$$

since $\hat{\mathbf{f}} = \sum_{(j,k)\in\delta^*} w_{j,k} W_{j,k}$.

Practical Application.

In practical applications, achieving the ideal risk using wavelet denoising is impossible since there is no way to discern what wavelet coefficients contribute to the signal

and which only contribute error. Therefore, when wavelet denoising is implemented it requires some other means of determining δ. In most applications the data adaptive estimate for δ, $d(\mathbf{y})$, is found through the use of a thresholding function.

Define a hard threshold where w is a wavelet coefficient, λ is a given threshold, and $|\cdot|$ is the absolute value by

$$\eta_H(w, \lambda) = wI_{|w|>\lambda} \tag{66}$$

and a soft threshold by

$$\eta_S(w, \lambda) = \text{sgn}(w)(|w| - \lambda)I_{|w|>\lambda}. \tag{67}$$

The hard threshold then sets any w where $|w| < \lambda$ to 0 while leaving the rest of the coefficients untouched. Soft thresholding sets any w where $|w| < \lambda$ to 0 as well as subtracts λ from all the remaining positive coefficients and adds λ to all the remaining negative coefficients. In effect, soft thresholding shrinks the magnitude of all the wavelet coefficents by λ or sets them to zero if they are to small.

The use of a thresholding function is motivated by the knowledge that only a small number of the wavelet coefficients, θ, for $\mathbf{f}$, are nonzero. Therefore if the threshold value, λ, is chosen such that it is larger than the magnitude of most of the noise coefficients, $z_{j,k}$, while being smaller than most signal coefficients, $\theta_{j,k}$, then thresholding can result in near-ideal risk. Much of the research in wavelet denoising is then focused on the fine tuning of λ.

One of the most popular choices for λ is the minimax threshold which seeks to minimize the variation from the ideal risk. This minimax threshold is often estimated

using

$$\lambda_{\text{minimax}} = 0.3936 + 0.1829 \times \frac{\log(N)}{\log(2)} \tag{68}$$

where $n > 32$ is the number of samples [27]. The estimate for $\mathbf{f}$, denoted $\hat{\mathbf{f}}$, is then found by

$$\hat{\mathbf{f}} = \mathcal{W}^T \hat{\theta} \tag{69}$$

where

$$\hat{\theta} = \eta_S(w, \lambda_{\text{minimax}}). \tag{70}$$

IV. Phase Folding

This chapter details the theory behind the periodic signal decomposition methods to be employed in the following chapters. The approach is based on the work of Stobie and Hawarden in [59] and [60] in which the effects of two periodic signals are isolated in two beat period cepheids, a type of variable star. In [59] and [60] the authors attempt to identify the first periodic component of a signal, isolate it, and remove its impact from the original signal. This enables them to analyze the second periodic component with minimal interference from the first component.

In [60], a light curve is first created from the available data. The light curve is then folded based on the previously calculated period as a means of isolating its first component. Next, a mean curve corresponding to the primary signal is determined. From the mean curve, the residuals for each observation are found. The residuals are then searched for the second periodic component using a technique from [30]. Finally, a mean curve for the second component is found using the same method as that for the first component.

The remainder of this chapter is broken down into three sections which will expand upon and improve the process described by Stobie and Hawarden as well as characterize its effects. Section 4.1 will discuss the need for and the effects of phase folding. Section two will explore the sources of noise for phase folded signals. Finally, the three most popular period determination methods will be characterized.

4.1 Phase Folding

Phase folding is a technique commonly employed in light curve analysis which overlays information from successive periods of a signal onto the plot of a single period. This plot is then a function of phase, or the fraction of a period, as opposed

to the original light curve which was a function of time. Overlaying light curve data by periods creates a more data dense plot where the periodic behavior of the signal can be more easily identified. This section will describe in more detail the impact of this technique and if such a technique is mathematically justified for more complex signals.

Motivation for Phase Folding.

The primary benefit to phase folding, other than its potential ability to extract components, is the creation of a more data dense representation of the signal over one full period. A cursory analysis of this data dense representation would lead to the belief that such a technique would lead to a more accurate estimate of the underlying function since there are more points contained in the folded curve. Though this belief is well founded, it has yet to be proven. In order to prove that a more data dense representation of a signal improves its functional form fit, first consider the notation used in the wavelet denoising section above.

Consider a signal $\mathbf{y} \in \mathbb{R}^n$ such that

$$y_i = f(t_i) + \epsilon_i, \qquad i = 1, ..., n \tag{48}$$

$$t_i = \frac{i}{n} \tag{71}$$

$$\epsilon_i \sim N(0, \sigma^2). \tag{72}$$

The ideal risk (error) when denoising a signal using wavelets is a function of two components:

1. The coefficients at coarse levels, that is, where $0 \leq j < j_0$, and

2. The coefficients at levels $j_0 \leq j \leq J$ where a breakpoint of $\mathbf{f}$ is contained in the associated interval $[2^{-j}(k - S), 2^{-j}(k + S)]$.

There are at most 2^{j_0} coefficients which satisfy the first condition and at most (# of breakpoints) $\times$ $(2S+1)$ coefficients which satisfy the second. Therefore, the ideal risk for the wavelet reconstruction of $\mathbf{f}$ is

$$R(T(\mathbf{y},\delta^*),\mathbf{f}) = \frac{|\delta^*|\sigma^2}{n} \tag{73}$$

where

$$|\delta^*| \leq 2^{j_0} + (J+1-j_0)(2S+1)L \tag{74}$$

and L is the number of partitions of the function.

Therefore, in order to prove that a more data dense function signal representation would lower the error it must be shown that

$$\frac{|\delta_n^*|\sigma^2}{n} \geq \frac{|\delta_m^*|\sigma^2}{m}, \tag{75}$$

when $m > n$.

Theorem 1 *Let the signal* $\mathbf{y}$ *be defined as in Equation 48 with the ideal risk given by Equation 53. If over the same time interval* y_i *is sampled* $m > n$ *times, where* $m = 2^{K+1}$ *and* $n = 2^{J+1}$ *where* $K, J \in \mathbb{Z}_+$ *and* $K > J$*, then the ideal risk for the* m *sample signal will be less than or equal to the ideal risk for the* n *sampled signal.*

Proof If the ideal risk for the signal with m samples is less than the signal with n samples, then

$$R(T(\mathbf{y},\delta^*),\mathbf{f})_n \geq R(T(\mathbf{y},\delta^*),\mathbf{f})_m \tag{76}$$

$$\Rightarrow \quad \frac{|\delta_n^*|\sigma^2}{n} \geq \frac{|\delta_m^*|\sigma^2}{m} \tag{77}$$

since $n = 2^{J+1}$ and $m = 2^{K+1}$ for some $K > J$ such that $K = J + C_K$ where $C_K \in \mathbb{Z}^+$. This makes $m = 2^{C_K} n$, making the inequality which needs to proven to be

$$\frac{|\delta_n^*|\sigma^2}{n} > \frac{|\delta_m^*|\sigma^2}{2^{C_K} n}. \tag{78}$$

In order to prove this, assume the opposite inequality and demonstrate a contradiction for both cases.

Case 1 $L > 0$, Assume

$$\frac{|\delta_n^*|\sigma^2}{n} \leq \frac{|\delta_m^*|\sigma^2}{2^{C_K} n}. \tag{79}$$

Multiplying both sides by n dividing both sides by σ^2 gives

$$|\delta_n^*| \leq \frac{|\delta_m^*|}{2^{C_K}}. \tag{80}$$

Replacing $|\delta^*|$ with the function for the count produces

$$2^{j_0} + (J + 1 - j_0)(2S + 1)L \leq \frac{2^{j_0} + (J_m + 1 - j_0)(2S + 1)L}{2^{C_K}}. \tag{81}$$

Multipyling both sides by 2^{C_K} results in

$$2^{C_K}\left(2^{j_0} + (J + 1 - j_0)(2S + 1)L\right) \leq 2^{j_0} + (J_m + 1 - j_0)(2S + 1)L. \tag{82}$$

Replacing J_m with $J + C_K$ makes the equation

$$2^{C_K}\left(2^{j_0} + (J + 1 - j_0)(2S + 1)L\right) \leq 2^{j_0} + (J + C_K + 1 - j_0)(2S + 1)L. \tag{83}$$

Rearranging the right side of the equation gives

$$2^{C_K}\left(2^{j_0}+(J+1-j_0)(2S+1)L\right) \leq 2^{j_0}+(C_K+(J+1-j_0))(2S+1)L. \quad (84)$$

Substituting $(J+1-j_0)$ with A, and $(2S-1)$ with B results in

$$2^{C_K}\left(2^{j_0}+ABL\right) \leq 2^{j_0}+(C_K+A)BL. \quad (85)$$

$$= 2^{j_0}+C_KBL+ABL \quad (86)$$

$$= (2^{j_0}+ABL)+C_KBL \quad (87)$$

Subtracting $(2^{j_0}+ABL)$ from both sides produces

$$2^{C_K}\left(2^{j_0}+ABL\right)-(2^{j_0}+ABL) \leq C_KBL. \quad (88)$$

$$(2^{C_K}-1)\left(2^{j_0}+ABL\right) \leq C_KBL \quad (89)$$

Dividing both sides by $BL(2^{C_K}-1)$ yields

$$\frac{2^{j_0}+ABL}{BL} \leq \frac{C_K}{2^{C_K}-1}. \quad (90)$$

$$\frac{2^{j_0}}{BL}+\frac{ABL}{BL} \leq \frac{C_K}{2^{C_K}-1} \quad (91)$$

Which simplifies to

$$\frac{2^{j_0}}{BL}+A \leq \frac{C_K}{2^{C_K}-1}. \quad (92)$$

Replacing the temporary variables gives

$$\frac{2^{j_0}}{(2S-1)L} + (J+1-j_0) \leq \frac{C_K}{2^{C_K}-1}. \tag{93}$$

Since $J \in \mathbb{Z}_+$ and $J > j_0$, it is known that

$$\frac{2^{j_0}}{(2S-1)L} + (J+1-j_0) > 1 \tag{94}$$

and that

$$1 \geq \frac{C_K}{2^{C_K}-1}. \tag{95}$$

Since $C_K \in \mathbb{Z}_+$, this implies that

$$1 < \frac{2^{j_0}}{(2S-1)L} + (J+1-j_0) \leq \frac{C_K}{2^{C_K}-1} \leq 1 \tag{96}$$

which is a contradiction. Therefore the original assumption is wrong and

$$\frac{|\delta_n^*|\sigma^2}{n} > \frac{|\delta_m^*|\sigma^2}{2^{C_K} n}. \tag{97}$$

Case 2 $L = 0$. Assume

$$\frac{|\delta_n^*|\sigma^2}{n} \leq \frac{|\delta_m^*|\sigma^2}{2^{C_K} n}. \tag{98}$$

Multiplying both sides by n and dividing both sides by σ^2 gives

$$|\delta_n^*| \leq \frac{|\delta_m^*|}{2^{C_K}}. \tag{99}$$

Replacing $|\delta^*|$ with the function for the count makes the equation

$$2^{j_0} + (J + 1 - j_0)(2S + 1)L \leq \frac{2^{j_0} + (J_m + 1 - j_0)(2S + 1)L}{2^{C_K}}. \tag{100}$$

Multipyling both sides by 2^{C_K} and replacing L with 0 results in

$$2^{C_K}\left(2^{j_0} + (J + 1 - j_0)(2S + 1)0\right) \leq 2^{j_0} + (J_m + 1 - j_0)(2S + 1)0. \tag{101}$$

$$2^{C_K}\left(2^{j_0}\right) \leq 2^{j_0} \tag{102}$$

Dividing by 2^{j_0} produces

$$2^{C_K} \leq 1. \tag{103}$$

Since $C_K \in \mathbb{Z}_+$, therefore $2^{C_K} > 1$, resulting in

$$1 < 2^{C_K} \leq 1. \tag{104}$$

Which is a contradiction. Therefore

$$\frac{|\delta_n^*|\sigma^2}{n} > \frac{|\delta_m^*|\sigma^2}{2^{C_K} n}. \tag{105}$$

□

Therefore, the ideal risk for the same function with more samples is less than that for fewer samples. Since the function in question is periodic, and the folded function has a lower error for $\mathbf{f}$, the improvement in signal quality extends to any number of desired periods.

Formal Definition of Phase Folding.

Given $\tau > 0$ and a function g where $g(t) = g(t+\tau)$ for all $t \in \mathbb{R}$, then $g \in P(\mathbb{R}, \mathbb{R})$ where $P(\mathbb{R}, \mathbb{R})$ is the set of all periodic functions from $\mathbb{R}$ to $\mathbb{R}$. If $g(t) = g(t+\tau)$ for all $t \in \mathbb{R}$, then $g(t) = g(\text{mod}_\tau(t))$ for all $t \in \mathbb{R}$ where

$$\text{mod}_\tau(t) = \begin{cases} \dots \\ t - n\tau \quad \text{for} (n-1)\tau \leq t < n\tau \\ \dots \end{cases} \tag{106}$$

. Since the reverse also holds, $g(t) = g(\text{mod}_\tau(t))$ for all $t \in \mathbb{R}$ can be used as an alternative definition for a periodic function. Define the fundamental function f for g as

$$f(t) = g(t)1_{[0,\tau)}, \tag{107}$$

where f is one full period of g starting at $t = 0$, ending at $t = \tau$ and

$$g(t) = f \circ \text{mod}_\tau(t) \tag{108}$$

$$= f(\text{mod}_\tau(t)) \tag{109}$$

for all $t \in \mathbb{R}$.

Define the interval $I = [0, T]$ where $T > \tau$, then g is in the periodic functions on the interval I if $g(t) = g(\text{mod}_\tau(t)) \forall t \in \mathbb{R}$. Since g is periodic on the whole real line it is periodic an any closed interval of $\mathbb{R}$. Therefore, $g(t) = g(\text{mod}_\tau(t)) \forall t \in I \subset \mathbb{R}$ so $g \in P(I, \mathbb{R})$. Define $\mathbf{t} = (t_1, t_2, \dots, t_n)$ where $t_i = \frac{i-1}{n}T$ and y such that

$$y(t_i) = g(t_i) + \epsilon_i \tag{110}$$

for all $t \in \mathbf{t}$ where $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$.

Given $\mathbf{t}$ and $y(t_i)$ for all $t_i \in \mathbf{t}$, then the expected value for g at $s_j \in \overline{\mathbf{t} \cap I}$ where $\text{mod}_\tau(t_i) = \text{mod}_\tau(s_j)$ is

$$\widehat{g(s_j)} = E(y(s_j)) \tag{111}$$

$$= E(g(s_j) + \epsilon_j) \tag{112}$$

$$= E(g(s_j)) + E(\epsilon_j) \tag{113}$$

$$= E(g(\text{mod}_\tau(s_j)) + 0 \tag{114}$$

$$= E(f(\text{mod}_\tau(s_j)) \tag{115}$$

$$= E(f(\text{mod}_\tau(t_i)) \tag{116}$$

$$= y(t_i). \tag{117}$$

$y(t_i)$ is then the phase folded estimate for g at s_j since $g \in P(I, \mathbb{R})$.

Define y phase folded over τ as the estimate for g at each $s_j \in [0, \tau)$ where there exists a $t_i \in \mathbf{t}$ such that $\text{mod}_\tau(s_j) = \text{mod}_\tau(t_i)$, or $\forall t_i \in \mathbf{t}$

$$y_{[0,\tau)}(\text{mod}_\tau(t_i)) = E(y(s_j)) \tag{118}$$

$$= y(t_i). \tag{119}$$

In effect, this phase folding creates a sample dense representation of the fundamental function of g, f, on $[0, \tau)$ which incorporates all the information from the original signal.

4.2 Noise Characteristics

This section will cover the noise characteristics of properly and improperly folded signals. The result of folding a solitary component signal will be covered first. Next

the effects of improperly folding a signal, folding it not over its period, will be discussed. Finally, the impact multiple components have on the residual values will be investigated.

Single Component.

Let a signal, y be composed of a single periodic function g and independent and identically distributed (iid) noise ϵ, where $g \in P(I, \mathbb{R})$ with period τ and fundamental function f.

$$y(t_i) = g(t_i) + \epsilon_i \tag{120}$$

$$t_i = \left(\frac{i-1}{n}\right) T \qquad T > \tau \tag{121}$$

$$i = 1, \ldots, n \tag{122}$$

Phase folding y over θ effectively condences the signal down to the interval $[0, \theta)$ while still maintaining all the information from the original signal. Define a proper fold as the folding of a function over the period of the function being estimated, τ. If the function undergoes a proper fold, meaning that $\theta = \tau$, the fold becomes

$$y_{[0,\tau)}(\text{mod}_\tau(t_i)) = E(y(s_j)) \tag{123}$$

$$= y(t_i) \tag{124}$$

for all $t_i \in \mathbf{t}$. After a proper fold the iid error is not modified in anyway, only moved around. This means that the error sample is still iid.

Improperly Folded Signals.

In order to determine the effects of additional components on a folded signal, the impact of an improperly folded signal must fist be analyzed. Consider the sine function in Figure 4, with a period of 1 second, and showing 3 periods over 3 seconds. The figure has 50 samples taken every $\frac{6}{25} = .06$ seconds, averaging 16.7 samples per period. These samples are indicated by the different color circles corresponding to the period in which they were taken. When properly folded at $\theta = 1$s, this results in the function graphed as the magenta line in Figure 5. Other than the slight truncation, this new signal matches that of the true signal almost perfectly. It should also be noted that each pair of successive points from the first period contain a point from each of the successive periods between them.

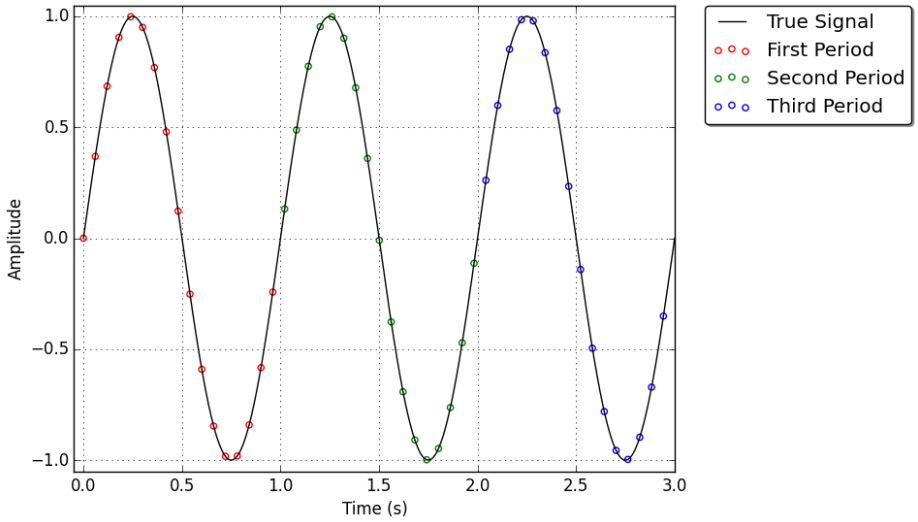


Figure 4. Original Signal.

A slight error in the fold can have a noticeable effect on the resultant folded signal. Folds of the signal in Figure 4 over slightly larger periods (by 1% and 5%) can be seen in Figure 6a and Figure 6b respectively. This fold error creates a saw tooth like effect on the resultant signal caused by sampling from each slightly shifted version of

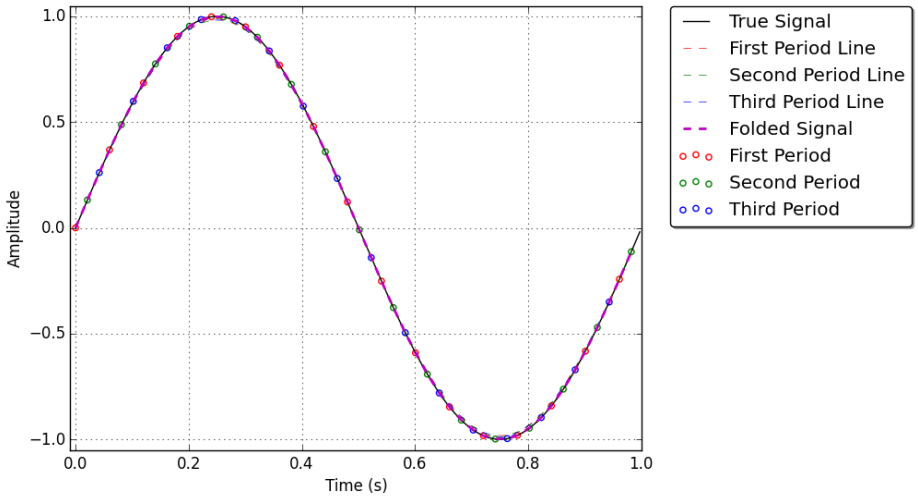
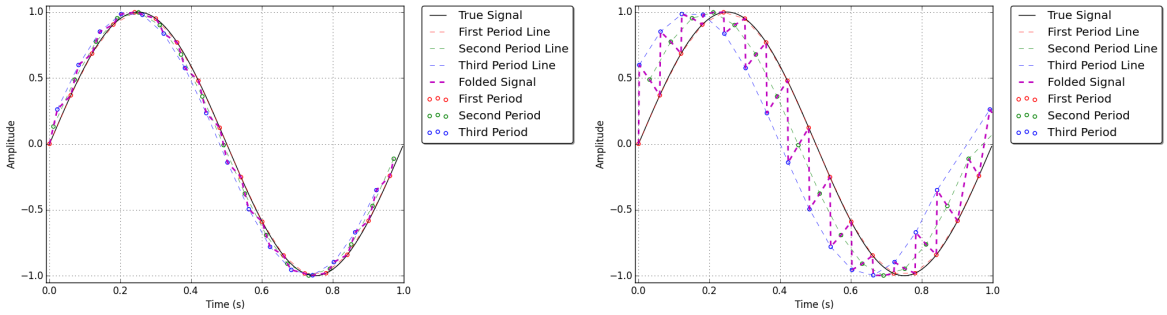


Figure 5. Properly Folded Signal.

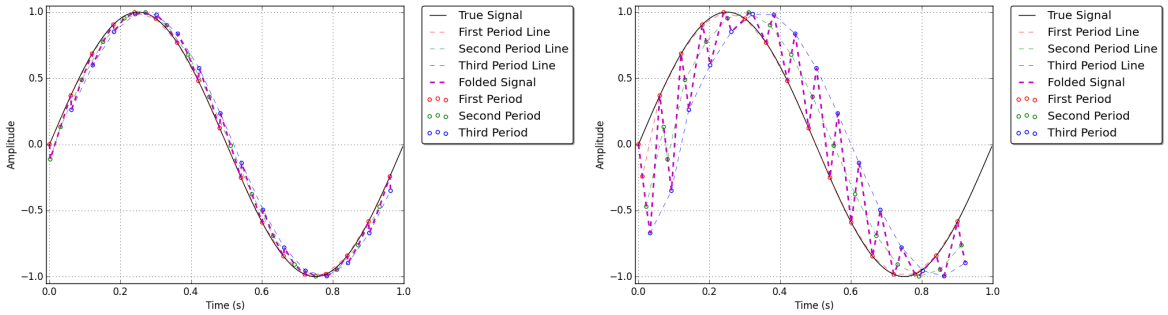
the true signal for every sample from the first period. A similar effect occurs when the fold is too small, as can be seen in Figure 7, where the sawtooth effect is reflected over the true signal. Depending on the shape of the signal being folded, errors from improperly folded signals generally increase the further the fold is from an integer multiple of the true period.



(a) 1% **Error** **(b)** 5% **Error**

Figure 6. Improperly Folded Signal - Large Fold.

When a signal has a sufficiently large number of periods folded over each other, even a small variation in the fold can make the signal indistinguishable from noise. However, this apparent noise will be distributed in the same way as that of a properly



(a) 1% **Error** **(b)** 5% **Error**

Figure 7. Improperly Folded Signal - Small Fold.

folded signal as can be seen in Figures 8 and 9 where the left plot shows the phase folded signal and the right plot show the histogram of the samples. This should come as no surprise since the distribution of the function will not changed no matter what the fold.

Multiple Component Signals.

Now that the effects of an improperly folded signal are understood, consider the two component signal $y(t_i) = g_1(t_i) + g_2(t_i) + \epsilon_i$ where g_1 has the period τ_1 and g_2 has the period $\tau_2 \neq \tau_1$. Since $\tau \neq \theta$, regardless of the fold chosen, at least one of the component functions will be improperly folded. This means that there will always be some signal error in addition to that resulting from ϵ_i. This error will be highly correlated based on the functional form of the improperly folded signal as well as the number of overlain folds.

Figure 10 contains the unfolded plot of a complete two component signal as well as plots for both individual components. Component one is the wavelet test signal called Blip, with a period of three ($\tau_1 = 3$), and component two is a sine wave with a period of five ($\tau_2 = 5$). All three signals folded over τ_1 are shown in Figure 11 along with their wavelet denoised signal. If component two was not included, the denoised

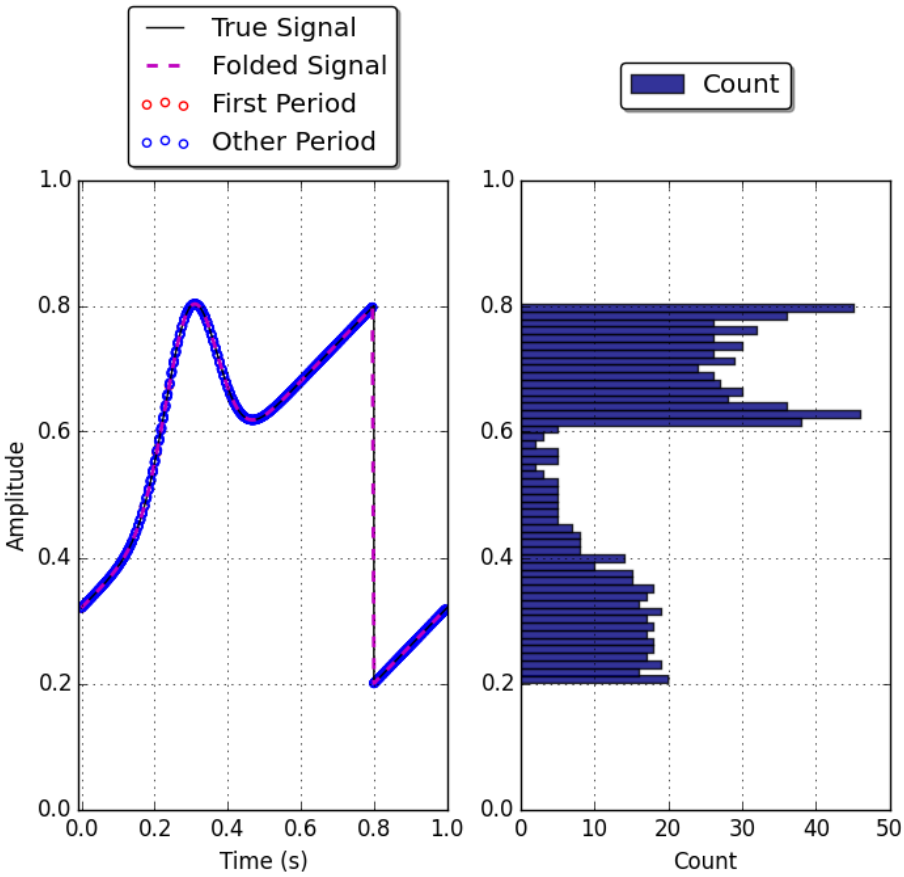


Figure 8. Properly Folded Blip Signal. The left plot shows the phase folded signal and the right plot shows its histogram.

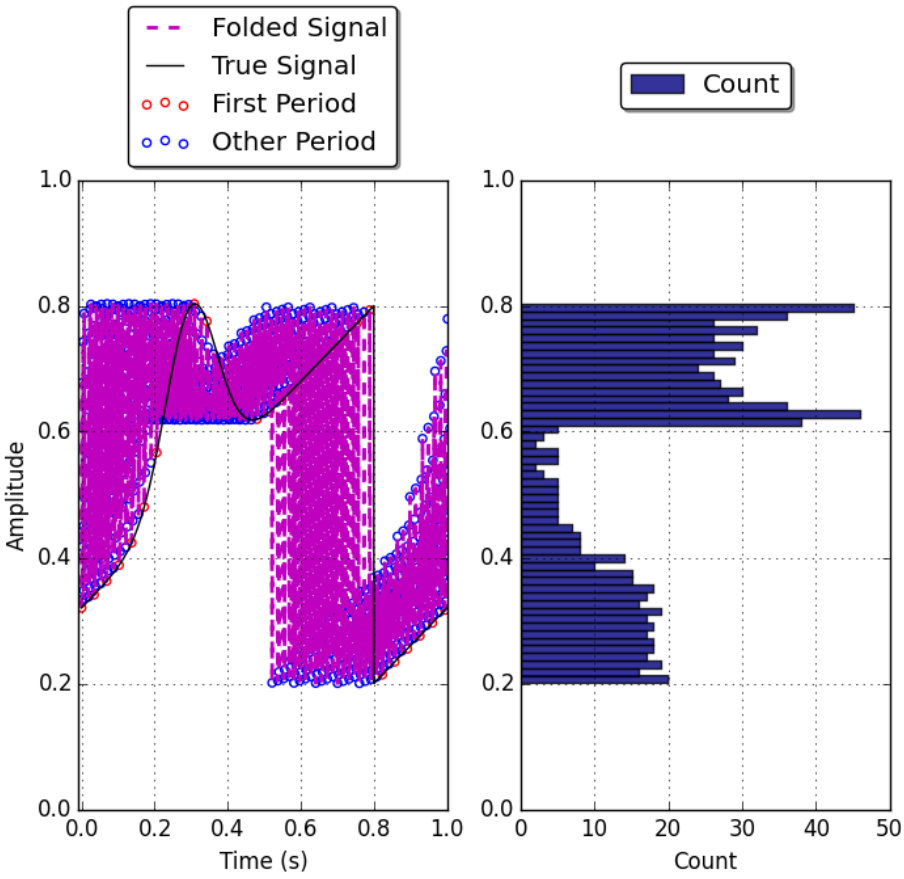


Figure 9. Improperly Folded Blip Signal - 1% Error. The left plot shows the phase folded signal and the right plot shows its histogram.

plots for the complete signal and component one would be identical. However, due to the impact of the second component, the complete signal folded over the first component's period has some error as can be seen in Figure 12.

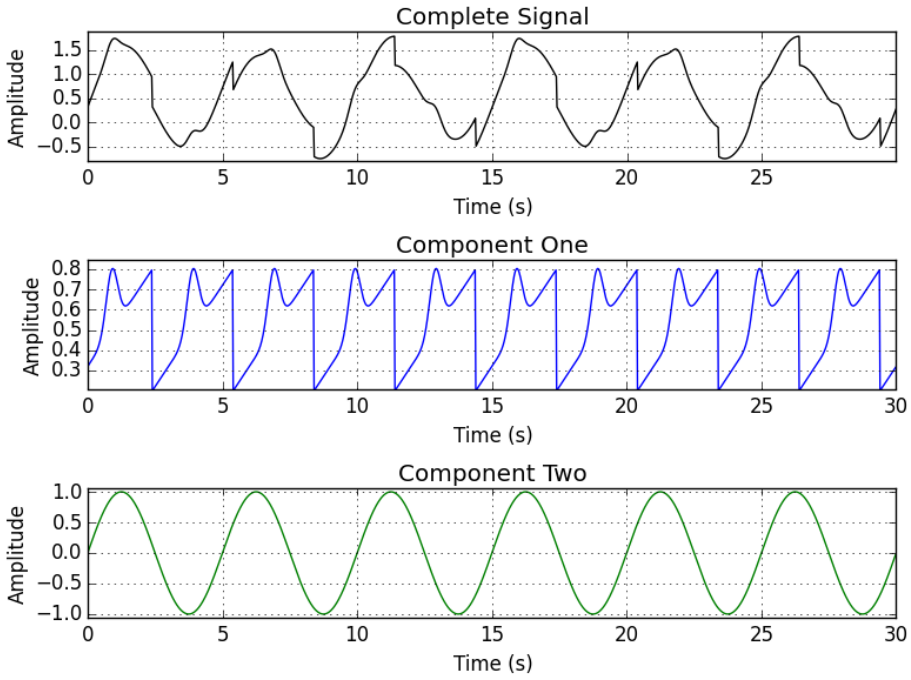


Figure 10. Two Component Signal - Original.

A similar, though less impactful effect can be seen when the signals are folded over the period for component two. Figure 13 shows the three signals folded over τ_2 along with their wavelet denoised signals. Since the amplitude of the signal causing the noise, component one, is smaller than that of component two the impact on the complete signal is much smaller. This effect extends to the denoised signal which can be seen in more detail in Figure 14.

This section serves as an introduction to the effects of phase folding on various types of signals and provides a mathematical framework in which to discuss the folding process. These topics will be address in greater detail with special emphasis placed on component extraction, denoising, and period detection in later chapters.

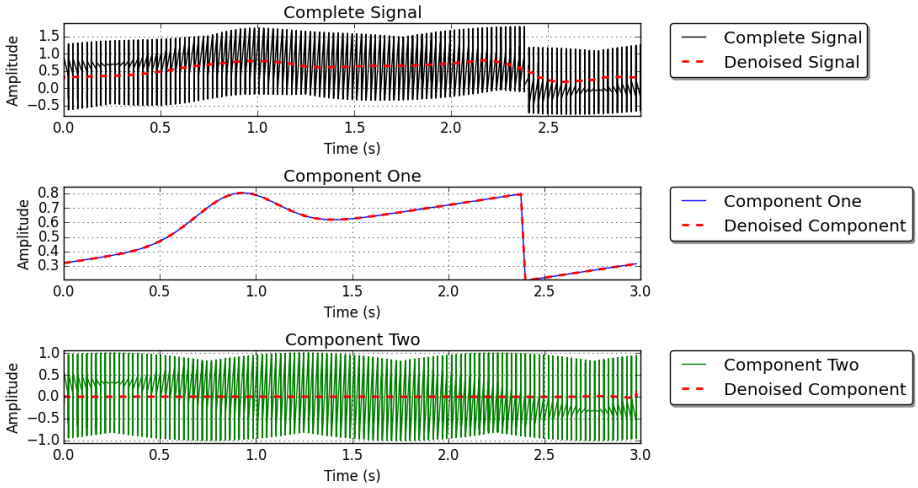


Figure 11. Two Component Signal - Fold A.

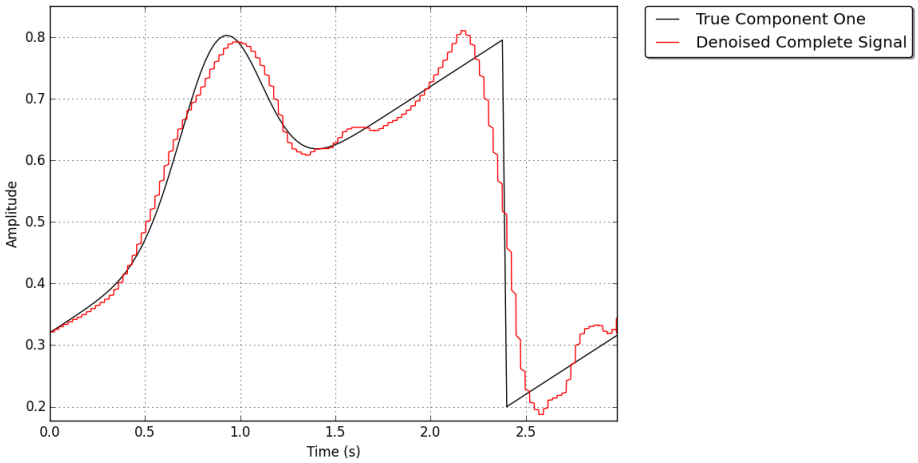


Figure 12. Two Component Signal - Fold A Denoised.

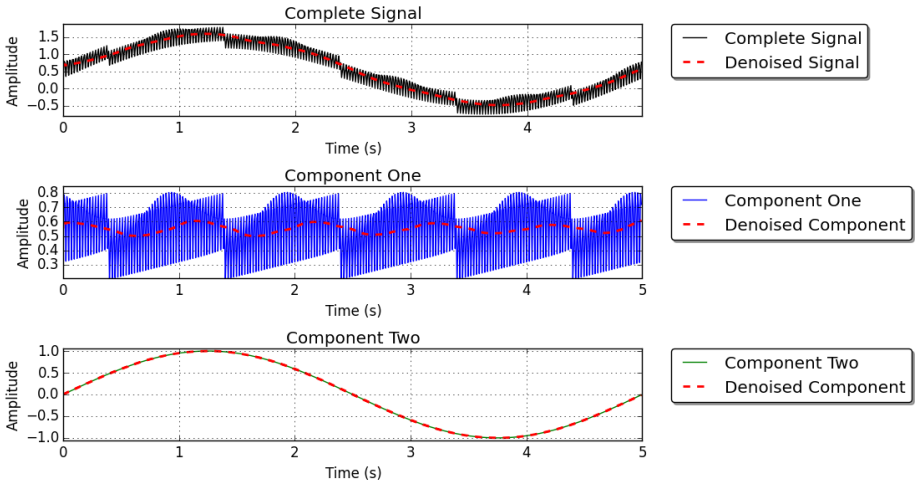


Figure 13. Two Component Signal - Fold B.

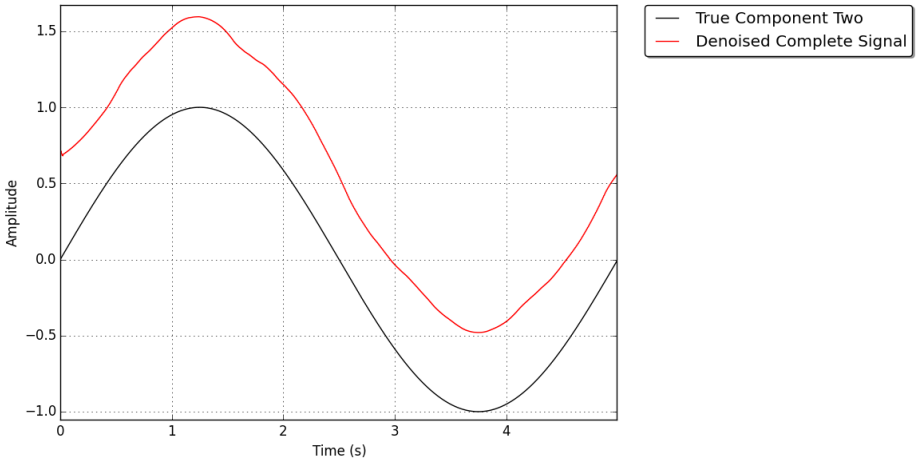


Figure 14. Two Component Signal - Fold B Denoised.

4.3 Period Determination

The NASA Exoplanet Archive is an online resource aimed at collecting and analyzing astronomical data, specifically those associated with exoplanets. In addition to the raw data normally collected, such as light curves and radial velocity curves, derived parameters such as positions and temperatures are also available. One service offered by the NASA Exoplanet Archive is the creation of periodograms, a plot demonstrating the relative impact of a range of periods [37]. This service uses three different algorithms for period determination, which are then used to create the periodogram. These algorithms and their variations are widely used across the astronomical community described here.

Lomb-Scargle.

The so called classic periodogram is based on the the Discrete Fourier Transform (DFT). Consider the function

$$\begin{gathered} y(t_i) = f(t_i) + \epsilon_i \text{ where} \\ \epsilon_i \sim N(0, \sigma^2) \text{ and} \\ i = 1, \ldots, n. \end{gathered}$$

The DFT of $\mathbf{y}$, denoted $FT_y(\omega)$, is then

$$FT_y(\omega) = \sum_{j=1}^{n} y(t_j) e^{-i\omega t_j}. \tag{125}$$

Using the DFT of the original signal, the classic periodogram is found by

$$P_y(\omega) = \frac{|FT_y(\omega)|^2}{n}. \tag{126}$$

Since there are an infinite number of frequencies at which the above function could be evaluated, common practice dictates that the evaluations be limited to $\frac{n}{2}$ evenly spaced frequencies. The first problem with the classic periodogram is the characteristics of its inherent noise. Even with relatively low noise signals, the classic periodogram results in a noisy output that can not be improved even with more samples [49]. The second problem is that of specctral leakage, where the power of a given frequency leaks over into other frequencies. These leaks can effect close frequencies and distant frequencies due to a variety of reasons such as aliasing.

The first algorithm used by the NASA Exoplanet Archive is called the Lomb-Scargle (L-S) algorithm , and it is a modified version of the classic periodogram [49]. Instead of calculating P_y using the DFT, the L-S algorithm uses

$$P_y(\omega) = \frac{1}{2}\left\{ \frac{\left[\sum_j y_j \cos\omega(t_j-\tau)\right]^2}{\sum_j \cos^2\omega(t_j-\tau)} + \frac{\left[\sum_j y_j \sin\omega(t_j-\tau)\right]^2}{\sum_j \sin^2\omega(t_j-\tau)} \right\} \tag{127}$$

where τ is defined using

$$\tan(2\omega\tau) = \frac{\sum_j \sin 2\omega t_j}{\sum_j cos2\omega t_j}. \tag{128}$$

This formulation constitutes a simple and well understood statistical behavior, and is equivalent to the original formulation with evenly sampled data, as well as being shift invariant.

The well known statistical behavior of the L-S periodogram is what provides superior utility to the original. Due to the method in which the L-S function is constructed, the power of the periodogram at a given frequency is exponentially distributed [49]. This means that the probability of random error creating any given spike in the periodogram can be easily determined. Therefore, a researcher need only put limits on

their allowable error in order to automate the L-S as a means of period detection.

Box-fitting Least Squares.

The L-S algorithm is a powerful method of finding periodic behavior in noisy signals, however, due to its reliance on trigonometric functions it is ill suited to the detection of more "blocky' periodic signals. This means the the L-S algorithm would be very capable at detecting periodic behavior of pulsating variable stars, but would perform much more poorly in the detection of periodic behavior of transiting exoplanets. The second available method in the NASA Exoplanet Archive is the Box-fitting Least Squares (BLS) algorithm, which has been designed to detect these more "blocky" periodic signals [29].

The BLS algorithm assumes that a given periodic signal has two states, H and L, which stand for high and low respectively. Since the algorithm was designed to find transiting exoplanets, L should only be present for a small fraction of the signal at periodic intervals. The low state is therefore in effect when the exoplanet is transiting in front of the star, and high at all other times. For a period of length P_0, the low state should only be in effect for qP_0 where q is assumed to be small (approximately $0.01-0.05$). Since $H = \frac{-Lq}{1-q}$, the BLS algorithm is focused on finding the best estimate for only four parameters: P_0, q, L, and t_0 (the epoc of the transit).

Given a signal as in Equation 48, each $y(t_i)$ is assigned a weight $w_i = \sigma_i^{-2}\left[\sum_{j=1}^{n}\sigma_j^{-2}\right]^{-1}$. This causes the arithmetic mean of $y_i w_i$ to be zero. The weighted signal is then folded over each of the trial periods, the periods to be tested, and the new series is then indexed using $\tilde{y}_i$ and weights $\tilde{w}_i$. A step function is then fit to the folded signal with the low amplitude dip, $\hat{L}$, on the interval $[i_1, i_2]$ and the high amplitude baseline, $\hat{H}$, on the intervals $[1, i_1)$ and $(i_2, n]$. The time spent in $\hat{L}$ is then characterized by the sum of the weights of data points in the interval $[i_1, i_2]$, denoted r.

To find the period of the signal, the minimum value of $\mathcal{D}$ is found for each period, where for a given (i_1, i_2) pair and a particular fold,

$$\mathcal{D} = \sum_{i=1}^{i_1-1} \tilde{w}_i(\tilde{y}_i - \hat{H})^2 + \sum_{i=i_2+1}^{n} \tilde{w}_i(\tilde{y}_i - \hat{H})^2 + \sum_{i=i_1}^{i_2} \tilde{w}_i(\tilde{y}_i - \hat{L})^2 \tag{129}$$

$$= \sum_{i=1}^{n} \tilde{w}_i \tilde{y}_i{}^2 - \frac{s^2}{r(1-r)} \tag{130}$$

where

$$s = \sum_{i=i_1}^{i_2} \tilde{w}_i \tilde{y}_i. \tag{131}$$

Since $\sum_{i=1}^{n} \tilde{w}_i \tilde{y}_i{}^2$ doesn't change, the value of the periodogram from any particular fold choice is

$$SR = \max \left\{ \left[\frac{s^2(i_1, i_2)}{r(i_1, i_2)\left[1 - r(i_1, i_2)\right]} \right]^{\frac{1}{2}} \right\}. \tag{132}$$

The proper period is then the point on the periodogram where $SR = (H-L)\sqrt{r(1-r)}$. In practical applications the algorithm generally splits the work into bins, the number of which depends on q. Though computationally intensive, BLS is a powerful algorithm for detecting transiting exoplanet-like periodic behavior.

Plavchan.

When attempting to determine the period, both L-S and BLS make assumptions as to the nature of the signal. L-S assumes that the signal is the sum of sinusoidal functions, and BLS assumes there there is only a single dip in the signal which is both narrow and shallow. If a truly autonomous period detection method is desired, neither of these assumptions are justified and a more general approach would be

necessary. One of the most popular near-agnostic period detection methods is called Phase Dispersion Minimization (PDM) [57].

Similar to BLS, PDM folds the signal in question over a set of trial periods. Once folded, the signal variance, σ^2 is calculated for the whole signal. The signal is also broken up into a set of bins, which may or may not be disjoint. The sample variance, s_j^2, for each bin is then calculated, where s^2 is the weighted sum of all variances over all j. The value $\Theta = \frac{s^2}{\sigma^2}$ is then used as the statistic to determine the best choice of a period. If the tested period is not the true period then s^2 will be approximately equal to σ^2, making $\Theta \approx 1$. When testing the actual period of the signal, Θ will reach reach a local minimum, potentially near zero.

This third period detection method provided by the NASA Exoplanet Archive is a binless variation on the standard PDM method, where BLS is a binned variation [45]. Instead of using Θ as the measure of a correct period, Plavchan uses the χ^2 difference between the original light curve and the smoothed light curve (using a boxcar method) for the worst 25 data points. The periods which minimize this χ^2_{25} value are then considered the best choice for the signals period.

Areas for Improvement.

When attempting to decompose a complex signal into its periodic component pieces, it was found that even slight variations in the proposed period can have a significant impact on the result. Since in the general case, no assumptions can be made as to the nature of the components, both the L-S and the BLS algorithms are alone poor choices for period detections algorithms. Variations on PDM, such as Plavchan, attempt to overcome these problems by not making any assumptions on the signal, however Plavchan uses a very naive approch to signal smoothing which could be greatly improved. Means of improving these methods, through the incorporation

of wavelet smoothing techniqes and knowledge about the resultant noise of additional components, are proposed and compared to these standard period detection methods in Chapter VII.

V. Improved Signal Quality

Phase folding's primary purpose is to improve the quality of an extracted periodic signal. To accomplish this, successive periods of the signal of interest are overlayed to increase the sample density of one full period. The phase folded signal is then smoothed, or denoised, using one of any number of possible techniques to determine the best estimate for the signal. The basic mechanics of this process are well understood and have been used for years by scientists, especially astronomers and cosmologists, in order to study a variety of periodic signals.

Though phase folding is widely used and accepted as a viable signal analysis technique, very little effort has been put forth to understand the effects of different types of variability on the extracted signal. This variability may originate from different folds, sampling rates, and additional components. This chapter is concerned with exploring the practical aspects of the theoretical work discussed previously, with respect to the effects from different types of variability, and the synergistic benefits of combining phase folding and wavelet denoising to mitigate these effects. The following sections will examine the effects of different folds on signal quality, how sampling variations can effect the resultant signal, and if the benefits of phase folding extend to signals with multiple components.

5.1 Fold Effects

In this section the effects of different folds on a signal will be investigated. The first topic to be covered will be even folds, where the complete signal is composed of a single component with an integer number of periods. Next, the effects of uneven, or noninteger, periods on the fold will be examined. These two topics will cover the extent of period variability which may be encountered when using real world data.

Integer Folds.

In the most basic example of phase folding, an integer number of periods are overlain in order to increase the number of samples for a single period. This results in a repetitive sampling pattern in the folded signal, which holds for the whole signal length. Recall, the periodic signal $y(t_i)$, where

$$y(t_i) = g(t_i) + \epsilon_i \tag{120}$$

$$t_i = \left(\frac{i-1}{n}\right) T \tag{121}$$

$$i = 1, \ldots, n \tag{122}$$

Take two successive samples, t_i and t_{i+1}, from the unfolded signal where $\left\lfloor \frac{t_i}{\tau} \right\rfloor = \left\lfloor \frac{t_{i+1}}{\tau} \right\rfloor$, meaning they were taken from the same period. Once the signal is folded over τ, there will be exactly $m-1$ samples between t_i and t_{i+1}, where $m = \frac{T}{\tau}$. These m samples (counting one of the end point samples) can result in three different possible sampling patterns.

If the sampling rate for the original signal is such that $t_i + k * \tau = t_j$ for some $k \in \mathbf{Z}_+$, t_i, and t_j, causes t_i and t_j to map to the same location. Since $y_\tau(t_i)$ may not equal $y_\tau(t_j)$, and wavelets are not designed to deal with repeated measures, it is recommended that the average value of the repeatedly sampled points be substituted before any wavelet denoising techniques are applied. The second possibility is that the number of samples per period and the number of periods in the signal are so perfectly related that the folded signal has evenly spaced samples. In this case, a wavelet denoising technique is perfectly suited to denoising the folded signal.

The finally possibility, which is more likely than the other two, is that the folded signal will result in a sampling rate with a repeated pattern every m samples. This means that $t_{i+1} - t_i = t_{i+m+1} - t_{i+m}$ for all $t_i < t_{n-m-1}$, however $t_{i+1} - t_i$ may not

be equal to $t_{i+2} - t_{i+1}$ depending on the sampling pattern caused by the folds. There are a wide variety of possibilities for this pattern that are all based on the number of samples per period and the number of periods in the signal. Due to the vast number of possible sampling patterns, it is difficult to determine their effects on the wavelet denoising process. However, since the sampling patterns are all functions of the number of samples per period and the number of periods in the signal, the effects of varying these two parameters can be tested in their stead.

To test the effect that the number of samples per period and number of periods have on the denoised signal, quality test signals must first be chosen. In wavelet denoising research, the ten test signals shown in Figure 15 are used most often [33]. These signals were chosen to test the versatility of wavelet denoising techniques by incorporating as many difficult-to-denoise signal characteristics as possible. Therefore, these signals will be used for testing purposes throughout the remainder of this dissertation.

Each of the test signals was sampled at three different levels (2^7, 2^9 and 2^{11} samples), using three different levels of noise (SNRs of 5 dB, 15 dB, and 25 db), and with one to 100 periods. This resulted in a total of $3 \times 3 \times 100 = 900$ different simulations, each of which was repeated 100 times and the mean mean squared error (MSE) was recorded. It was expected that the MSE using the standard approach (denoising without phase folding) would result in more error with more periods and that the folded signal would remain relatively constant with respect to the number of periods. Higher noise (lower SNR) was expected to raise the error of both approaches, however, cause a greater variation between the two denoising techniques as the number of periods increased. Sample size was anticipated to lower the error of both approaches and potentially mitigate the effects of more periods.

The shape of each test signal was presumed to have some effect on the MSE as

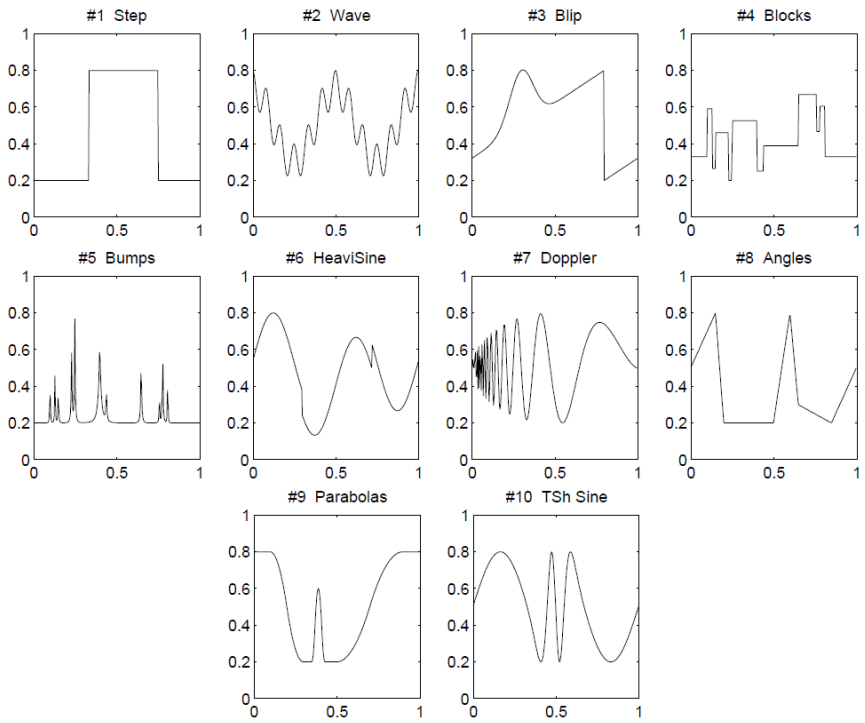
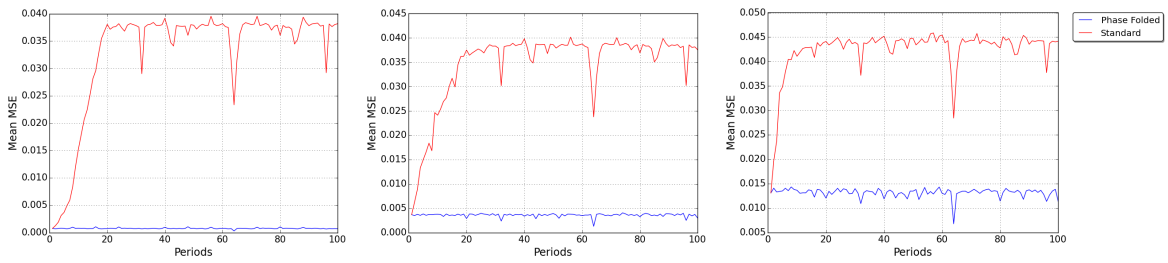


Figure 15. Test Signals [33].

well. Shapes with rapid variations, such as the Doppler and Bumps, were thought to cause a more rapid increase in difference between the standard approach and the phase folded approach as the number of periods increased. Whereas smoother signals such as HeaviSine, Parabolas, and Time Shifted Sine were predicted to somewhat mitigate the effects of increased periods.

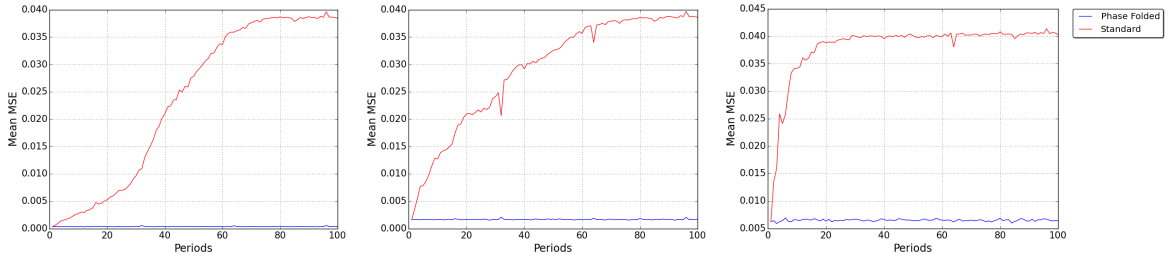
As can be seen in Figures 16, 17, and 18, a lower SNR (more noise) increased the baseline level of error for all three sample levels as was expected. There was also an increase in the difference in noise between the standard approach and the phase folded approach as the number of periods increased, while the phase folded approach remained relatively constant. This effect seemed to plateau, however, once a certain noise level was reached, a level that appears to be constant with respect to the test signal, regardless of the number of samples.

An increase in the number of samples appears to increase the number of periods



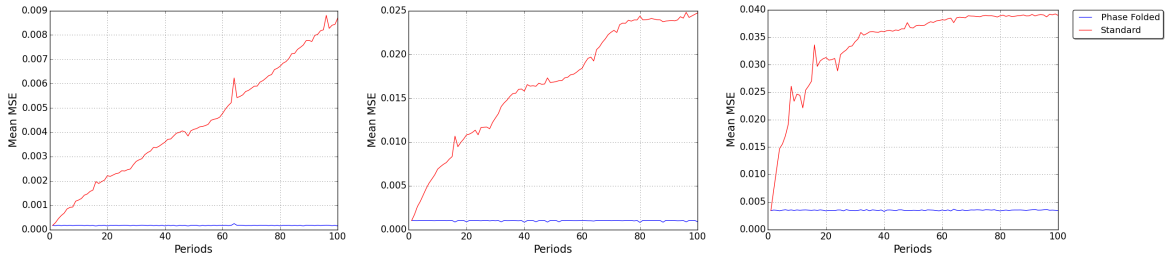
(a) SNR of 25 **(b) SNR of 15** **(c) SNR of 5**

Figure 16. Mean Squared Error of Blip Integer Fold Simulation Results - 2^7 Points.



(a) SNR of 25 **(b) SNR of 15** **(c) SNR of 5**

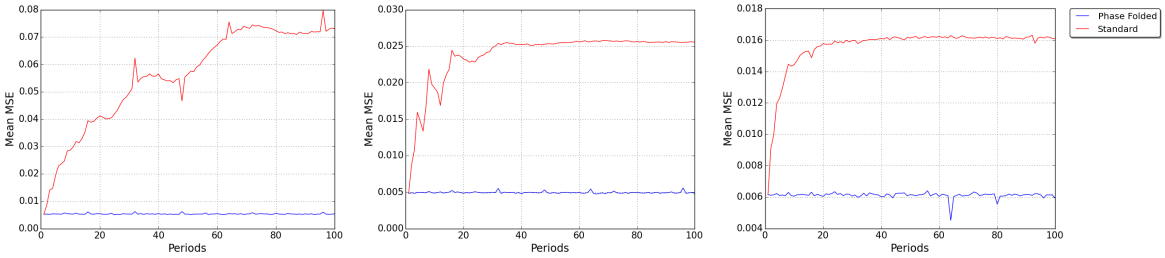
Figure 17. Mean Squared Error of Blip Integer Fold Simulation Results - 2^9 Points.



(a) SNR of 25 **(b) SNR of 15** **(c) SNR of 5**

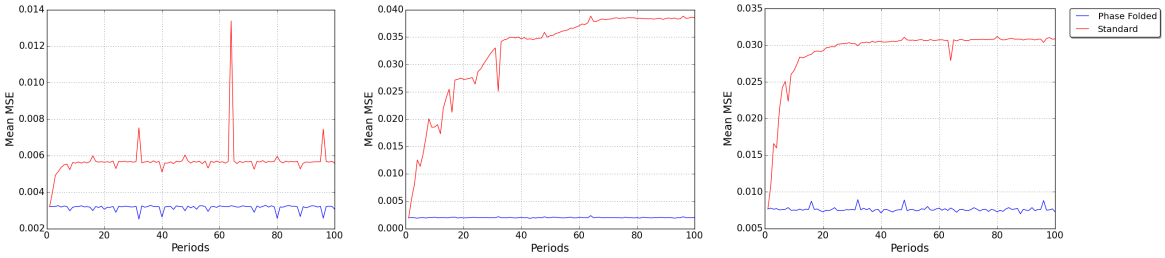
Figure 18. Mean Squared Error of Blip Integer Fold Simulation Results - 2^{11} Points.

necessary to reach the error plateau. In Figure 16a, the noise plateau is reached at approximately the 20 period point, whereas it seems to take to about 70 periods for an SNR of 25 for the Blip folded signal with 2^9 samples (Figure 17a). Figures 18a and 18b do not reach the error plateau, and appear to still be increasing in error with more periods, especially for larger SNRs. The plateau is finally reached for a SNR of 5 (Figure 18c) at around the 60 period point. Therefore, there appears to be a maximum MSE for the Blip signal which is constant with respect to SNR and sample size.



(a) Step **(b) Wave** **(c) Blocks**

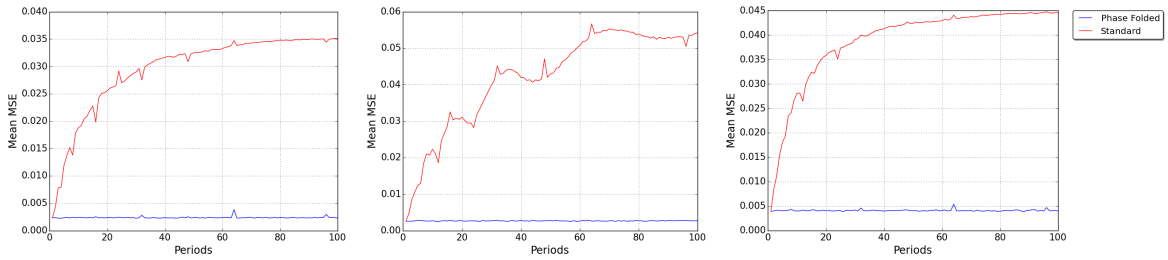
Figure 19. Mean Squared Error for 2^{11} Points with an SNR of 5 dB.



(a) Bumps **(b) HeaviSine** **(c) Doppler**

Figure 20. Mean Squared Error for 2^{11} Points with an SNR of 5 dB.

Figures 19, 20, and 21 show the MSE for the standard and phase folded denoising techniques for samples of size 2^{11} and an SNR of 5 dB for each of the remaining test signals. These figures demonstrate a similar pattern to that of the Blip test signal shown in Figure 18c with 2^{11} points and an SNR of 5 dB. Each signal appears to



(a) Angles **(b) Parabolas** **(c) Time Shifted Sine**

Figure 21. Mean Squared Error for 2^{11} Points with an SNR of 5 dB.

have a error plateau for the standard approach, and a lower, nearly constant, error for the phase folded method. The error plateaus vary by signal and have rates for obtainment which are independent of the plateau error level.

The highest error plateaus occurred with the Time Shifted Sine followed by HeaviSine, Angles, Doppler and Wave. All of these signals, with the exception of Angles, have a sinusoidal component, as does, to a lesser extent, Blip. Those signals with the lowest MSE plateau are Parabolas, Bumps, and Step. Blocks falls somewhere in the middle for standard error plateaus. The rapid changes of Doppler and Wave appear to create higher error levels, however, this is obviously not the case for Bumps.

The shapes in Bumps and the shape of Parabolas are very similar in form to that of *db*4, the wavelet used to denoise all the signals. It is believed that this similarity helps to lower the error plateau. It is also theorized that the piecewise constant nature of Step and Blocks helps to lower the peak error. Changing the wavelet used to denoise the signals is expected to have a significant impact on the plateau error. The effects of different wavelet choices on signal denoising will be explored in a later section.

A final point of interest is the large spikes in error for the Bumps signal at periodic intervals. The first anomaly appears at the 8 period point where the error level dips by a small fraction. Similar effects occur for every period which is an integer multiple

of 8, higher error for even multiples, lower error for odd, and with the largest effect being a more than doubling of the MSE for 64 periods. Similar effects were observed for the Bumps signal at different levels of sample size and error. It is believed that these jumps in MSE are the result of the interaction between the test signal, the number of periods, and the sampling rate and that they will disappear with small modifications to the rate or period.

Non-Integer Folds.

In real-world data, it is highly unlikely that the gathered signal will consist of an exact integer number of periods. Therefore, the effects of non-integer folds must be examined. In order to accomplish this, a simulation of each of the 10 wavelet test signals for sample sizes of 2^7, 2^9, and 2^{11} and SNR levels of 25 dB, 15 dB, and 5 dB were folded over 500 different periods between 2 and 15.5 periods.

In addition to the effects discussed in the previous subsection, it was expected that lower non-integer folds would result in higher error due to signal distortion from the perspective of the wavelet, and that these effects would diminish for more periods. For example, a signal consisting of 2.5 periods, once folded, would result in 1.5 as many samples over a given interval in the first half of the signal as compared to the same interval in the last half of the signal. Since the DWT treats the samples as evenly spaced, it was thought that this may increase error. With larger numbers of periods, such as 15.5, the difference between the first half of the signal and last half of the signal drops to a sample multiplier of $1.0\bar{6}$, meaning that the impact should be smaller than with a smaller number of periods.

After conducting the simulation, it was found that the initial assumptions were correct. The gradual changes as in the previous subsection were still observed for the standard method. Also, the increase in error at non-integer folds was also observed

which decreased in impact as more full periods were added. An excellent example of this effect can be seen in Figure 22. The phase folded MSE has a diminishing sinusoidal type shape, which is essentially smooth at the 12 periods point.

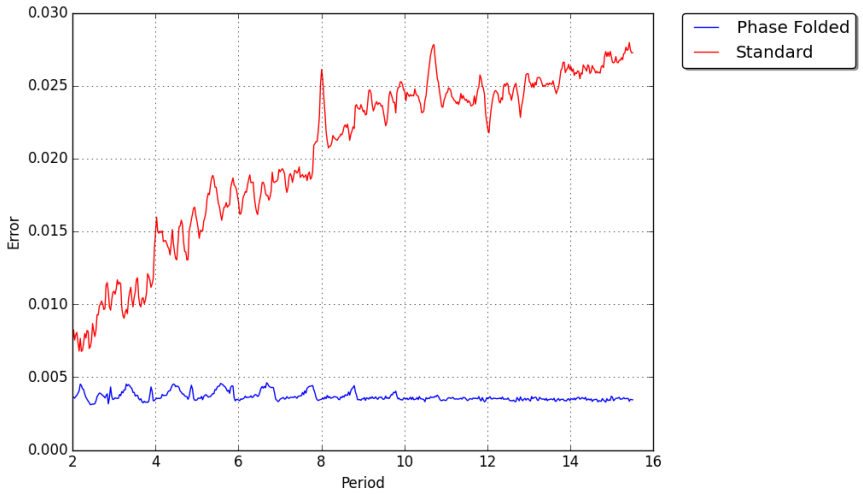


Figure 22. Blips - 2^{11} Samples and SNR of 5 dB.

5.2 Sampling Effects

In the previous section, two types of period variability were discovered. In a similar fashion, this chapter aims at investigating the effects of various types of sampling on the resulting error. The first topic to be examined is the effect of evenly spaced samples which contain no variation in the sampling rate. Unevenly sampled data will be examined next, in which there is random variation in the time between samples. Finally, the impact of sample sizes other than powers of two will be explored.

Even Samples.

Even sampling is the most basic form of sampling used in denoising applications. In evenly sampled data, each sample has the same distance between its neighbor(s) as any other sample. This type of sampling is ideal for the wavelet which does not take

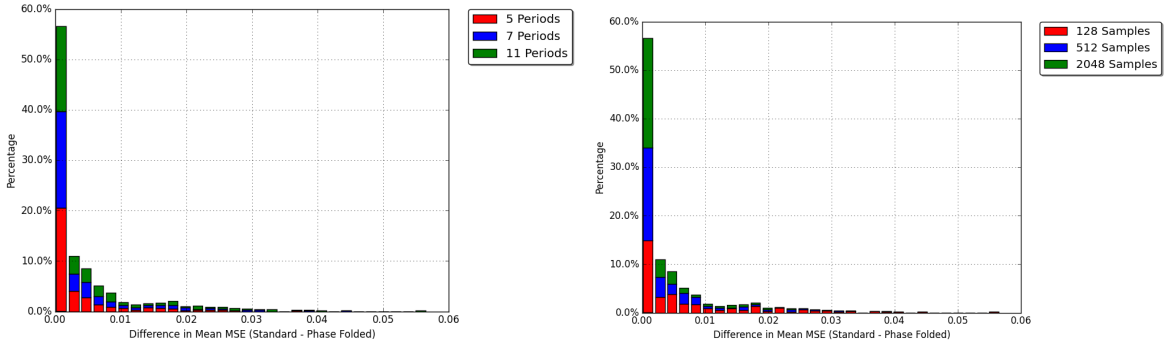
Table 2. Even Sample Simulation Factors and Levels.

Samples	2^7, 2^9, 2^{11}
Periods	5, 7, 11
Amplitude Multiplier	.1, .5, 1
SNR	25, 15, 5

into account the time variable. To test the effects of even sampling on MSE, each of the 10 wavelet test signals were tested using a full factorial design encompasing all of the 81 possible combinations of factor levels found in Table 2. Each of the 810 simulations was run 100 times, and their average used as the best estimate of the true performance.

While the phase folded technique was believed to be superior for all combinations, it was expected that more periods would result in a greater difference between the standard and phase folded resulting MSEs. Larger sample sizes were also anticipated to lower the error, however since sample size should effect both the standard method and the phase folded method, these effects should be comparatively minimal. Higher levels of noise (lower SNR) was thought to have a bigger impact on the standard approach, resulting in better performance for the phase folded technique. The amplitude multiplier was assumed to have no effect on the performance since the added noise is simply a function of the signal power. Finally, it was expected that the different signals would result in small variations in MSE differences in a similar way as those seen in the fold analysis section previously.

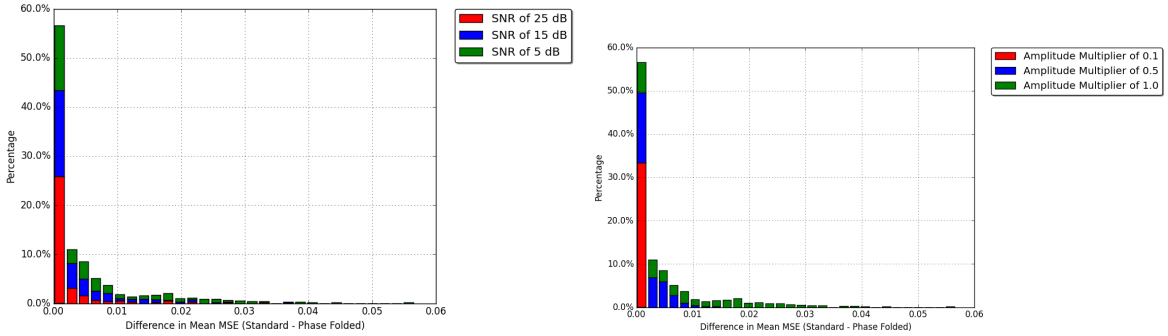
As can be seen in Figure 23a, lower periods result in a smaller error difference than higher periods. However, since the number of periods do not differ greatly, this difference is relatively small. It is to be expected that this difference would be more noticeable if there were a wider range of periods. More samples actually appear to lower the difference in mean MSE as can be seen in Figure 23b. This is once again



(a) Periods **(b) Samples**

Figure 23. One Signal Even Samples Mean MSE Difference Comparison.

likely a result of the periods, and a higher number of periods will likely shrink these differences.



(a) Noise **(b) Amplitude Multiplier**

Figure 24. One Signal Even Samples Mean MSE Difference Comparison.

Figure 24a shows the effect that noise has on the differences in error. Lower noise levels results in smaller mean MSE, while higher noise level increases these differences. The amplitude multiplier, surprisingly, has a noticeable effect on the difference in means of the MSEs. After further investigation it was found that all of the amplitude multipliers have the same distribution of error (just on different scales). The apparent lack of various levels of mean MSE for the .1 amplitude multiplier is because all of the variation fits in one of the histogram bins.

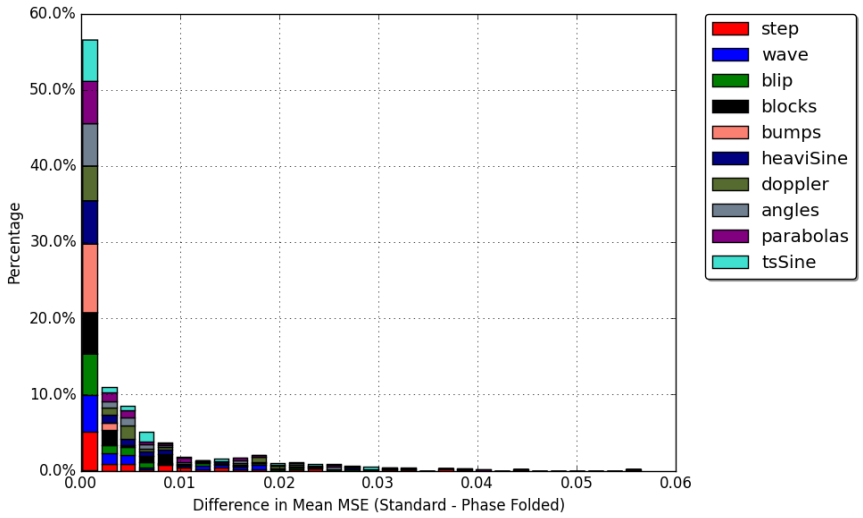


Figure 25. One Signal Even Samples Mean MSE Difference Signal Comparison.

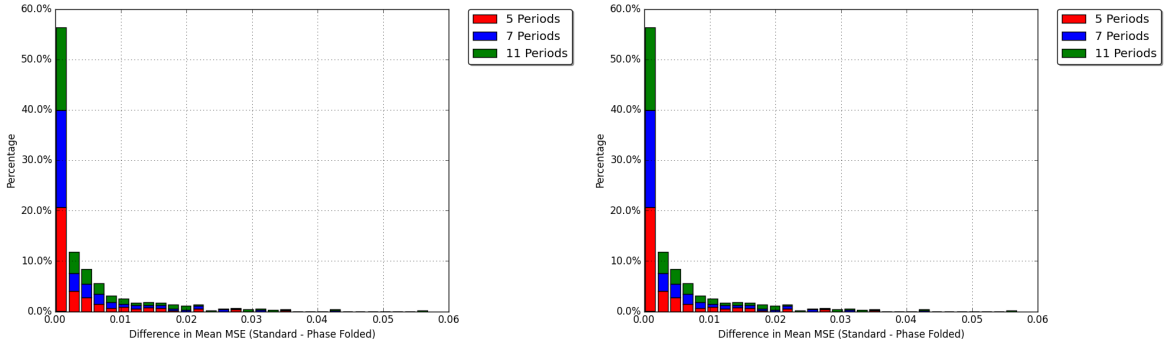
Finally, Figure 25 shows the effects of the various test signals on the difference in mean MSEs. Bumps appears to have very little difference between the standard and phase folded MSE. This again is likely due to the small numbers of periods. More periods would result in a larger difference. Other signals, such as Blocks, had a much larger range of mean MSE values.

Uneven Samples.

To test the effects of uneven samples, the same factors and levels found in Table 2 were used. The only difference in experimental setup between this simulation and that for the even samples was in how the time values were created. For the uneven samples simulation, the appropriate number of samples were taken from a uniform distribution. This means that there were no consistent differences, spacing-wise, between samples like those used in the even sample experiment.

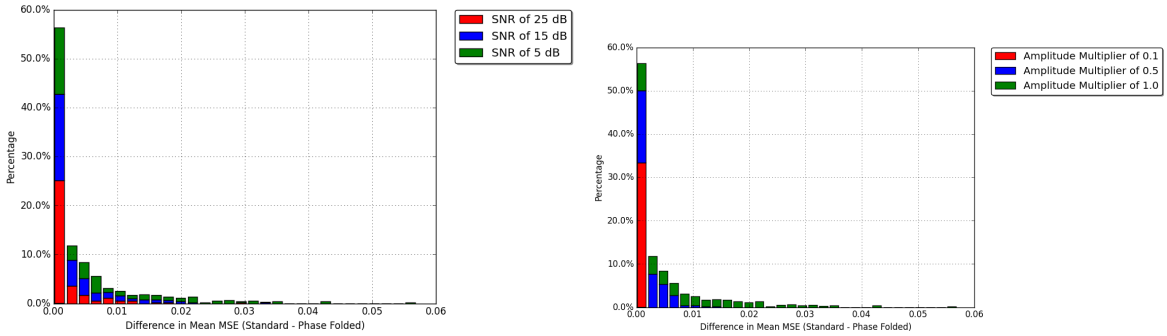
It was assumed that the same effects as those demonstrated by the even samples would be observed, however, with a slightly larger baseline error. Since this increased baseline error would be experienced by both the standard and phase folded

approaches, the resultant distributions for error were expected to be nearly identical to those found for the even sample simulation. As can be seen in Figures 26-28 these assumptions were well founded, and results follow that as discussed for even samples.



(a) Periods **(b) Samples**

Figure 26. One Signal Uneven Samples Mean MSE Difference Comparison.



(a) Noise **(b) Amplitude Multiplier**

Figure 27. One Signal Uneven Samples Mean MSE Difference Comparison.

To test the baseline increase in error, the same experiments used for the integer fold analysis (Section 5.1) were ran on the unevenly sampled test signals. This means that for each test signal, all $2,700$ possible choices for factor levels from Table 3 were simulated 100 times each, the means of which were recorded. The results can be seen in Figures 29-32. Note that the difference in errors between the two types of sampling are nearly identical. The strange effects due to specific periods for the Bumps signal

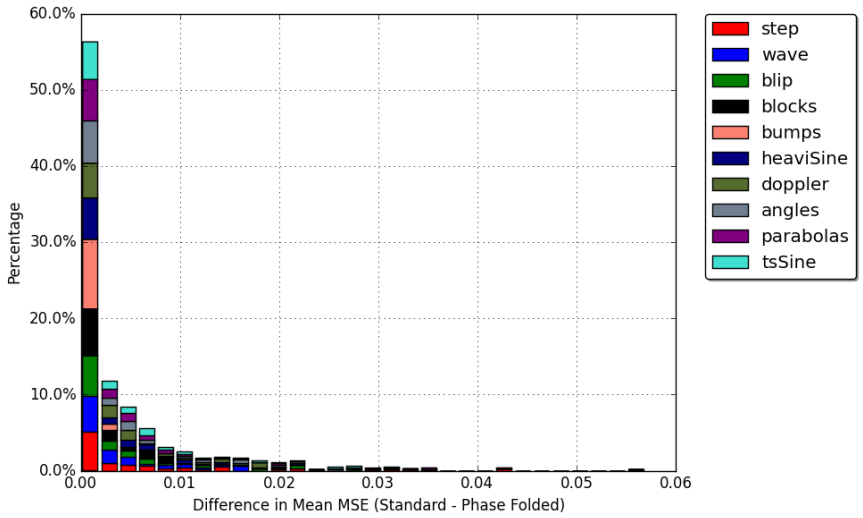


Figure 28. One Signal Une Samples Mean MSE Difference Signal Comparison.

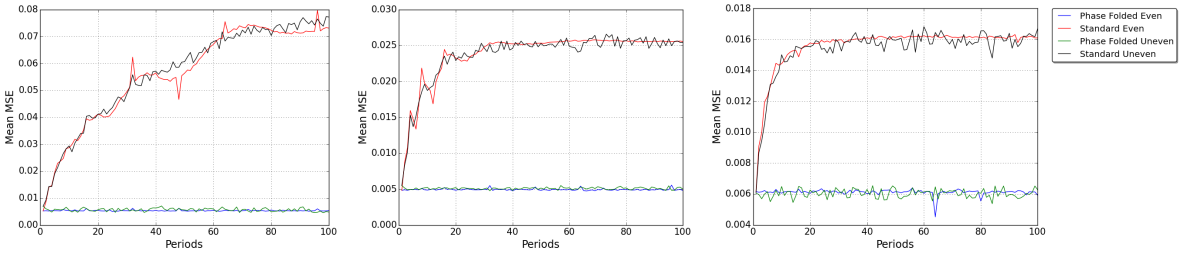
Table 3. Uneven Sample Simulation Factors and Levels for Comparison.

Samples	2^7, 2^9, 2^{11}
Periods	1, 2, . . . , 100
Amplitude Multiplier	.1, .5, 1
SNR	25, 15, 5

also seem to have disappeared when the samples became uneven (Figure 32). This means that the anomalous error was an artifact due to an interaction between the sampling rate and the signal shape.

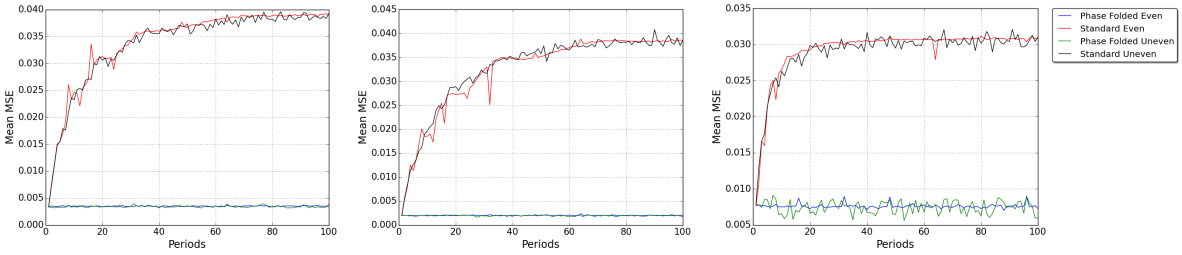
Non-Powers of Two.

Due to the way in which wavelets are constructed, wavelet denoising programs require sample sizes to occur in powers of two. In real world data this is rarely the case, therefore, the data must either be discarded or extended in some way. Since it is inadvisable to discard already gathered data, most scientist apply some extension process to data sets that are not of standard lengths. This extension process is usually



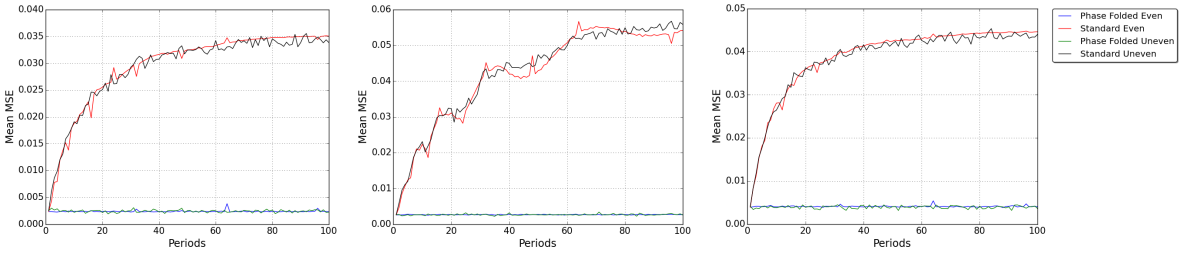
(a) Step (b) Wave (c) Blocks

Figure 29. Sample Type Comparison for Mean Squared Error for 2^{11} Points with an SNR of 5 dB.



(a) Blip (b) HeaviSine (c) Doppler

Figure 30. Sample Type Comparison for Mean Squared Error for 2^{11} Points with an SNR of 5 dB.



(a) Angles (b) Parabolas (c) Time Shifted Sine

Figure 31. Sample Type Comparison for Mean Squared Error for 2^{11} Points with an SNR of 5 dB.

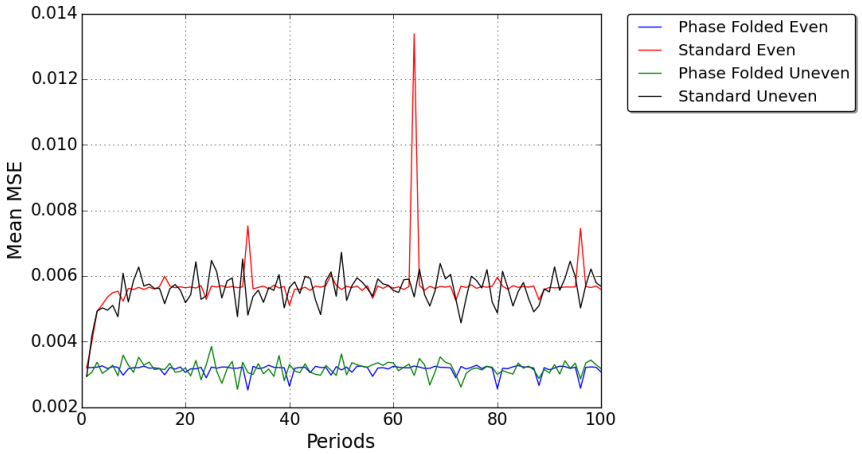


Figure 32. Sample Type Comparison for Mean Squared Error for 2^{11} Points with an SNR of 5 dB - Bumps.

conducted in one of four ways.

The first way scientists extend their data sets is by simply adding zeros before and after the gathered data until the desired length is reached, called zero padding (zpd). The second approach, called constant padding (cpd), repeats the first and last values out to accomplish the same goal. Since these approaches are very naive in nature they are rarely used. A third and more commonly used method is the symmetric padding method (sym), where the signal is lengthened by reflections at the end points. The final method is called periodic padding (ppd), where the signal is considered to be periodic and is repeated appropriately before and after the true signal.

Before testing these methods, it was conjectured that the worst performing method would be the zpd since none of the signals start or end at 0. This would simply cause a sharp drop off at the end of the signal that the wavelet would need to model, most likely having a detrimental effect on the error. It was also thought that the cpd would have a similar negative effect on the MSE, though not to the extent of the zpd approach. The sym and ppd approaches were believed to be the best choices depending on the signal shape.

Table 4. Sample Variation Factors and Levels.

Samples	128, 129, . . . , 512
Periods	5, 7, 11
Amplitude Multiplier	1
SNR	25, 15, 5

Using the factors and levels shown in Table 4, each of the nine combinations were simulated 100 times using the number of samples as the independent variable, the average of these runs was used used to represent them. Shown in Figure 33 is a typical resultant graph. It was found that zpd produced the highest MSE for all cases and that the ppd was the best performing approach, with slightly better mean MSE than sym in almost every instance. Since Figure 33 contains sample sizes of three different powers of two without any noticeable effect on the errors, it appears that having too few samples, if extended appropriately, will have little to no effect on the observed error when using the phase folding method of denoising.

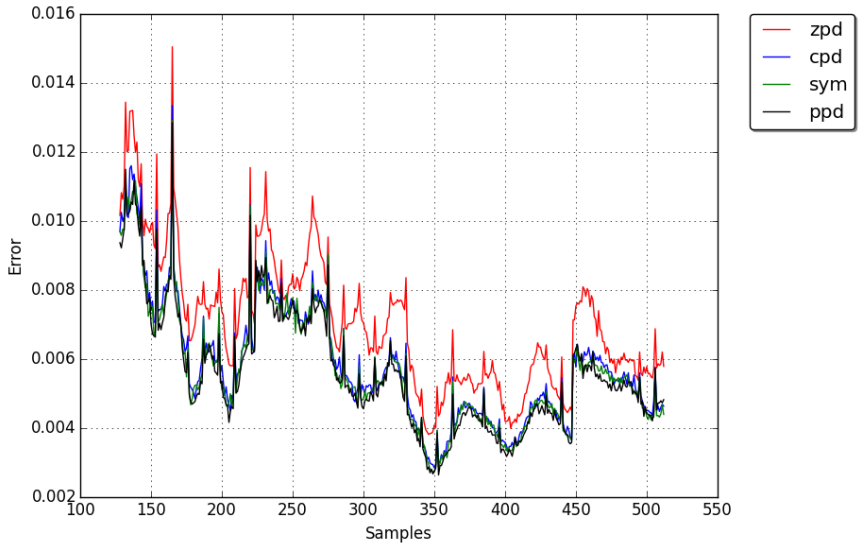


Figure 33. Sample Size Comparison for Mean Squared Error for 2^{11} Points with an SNR of 5 dB - Angels.

5.3 Multiple Components

The final section of this chapter is concerned with improving the extracted signal quality when the signal consists of two or more components. To accomplish this the signal is folded over the period of each of its components in turn. Once folded over a given period, the signal is denoised using the wavelet method discussed throughout this chapter. The denoised component is then subtracted from the original signal and the process continues until all components have been extracted. Once all components have been removed, they are added together to get the best estimate for the original signal.

To test the validity of this approach, two component signals will be explored first, followed by three component signals. This section is only concerned with the extraction of components using phase folding for the purpose of improving the signal quality of the original signal. Improving the quality of the extracted components while phase folding will be covered extensively in the next chapter.

Two Components.

To test the ability of phase folding to improve signal quality over the standard approach, signals composed of two components were tested first. These signals were created by first producing each periodic component separately over the same time interval and then adding them together. These components consisted of each pair of different wavelet test signals, as provided in Figure 15. Once the complete signal was created, the appropriate level of error was calculated based on the signal power and desired SNR, then added to the signal. The phase folding extraction method was then used on the signal to get the best possible estimate of the original (before noise was added) signal.

The simulation used to test this process consisted of testing each of the $10 \times 9 = 90$

Table 5. Improved Signal Quality - Two Components Factors and Levels.

Signals	Step, Wave, Blip, Blocks, Bumps, HeaviSine, Doppler, Angles, Parabolas, Time Shifted Sine
Samples	2^7, 2^9, 2^{11}
Periods	5, 7, 11
Amplitude Multiplier	.1, .5, 1
SNR	25, 15, 5

possible combinations of wavelet test signals (meaning that the same signal shape was not used for both components simultaneously). Each possible signal combination was tested at three different sample sizes, 2^7, 2^9, and 2^{11}, and three different levels of noise, SNR of 25 dB, 15 dB, and 5 dB. The possible periods for each of the components were 5, 7, and 11. Once again, only one of the components would have any of the given periods at a time making $3 \times 2 = 6$ total possible combinations of periods.

Finally, each component was also multiplied by a constant value to simulate the interaction of two signals of different amplitudes. These amplitude multipliers were .1, .5, and 1. These modifiers were allowed to be repeated in a single signal, making the total possible combinations of amplitude multipliers equal to $3 \times 3 = 9$. This makes the grand total of $90 \times 3 \times 3 \times 6 \times 9 = 43,740$ simulations, each of which were ran 100 times and the average MSE recorded. A table summarizing the possible factors and levels can be found in Table 5.

The phase folding method was anticipated to perform better than the standard approach in nearly all instances, except perhaps in a small number of cases where random variations may give some small edge to the standard approach. To compare

the two approaches, the differences in mean MSE were computed between the standard and the phase folding approaches such that positive values reflect a lower mean MSE from the phase folding method. The simulation results shown in Figure 34 fit the assumed distribution nearly perfectly, with negative values (where the mean MSE from the standard approach was less than that from the phase folding approach) consisting of only 2.63% of the total simulations. These negative results were almost exclusively limited to cases with low noise levels and high sample sizes. This caused the difference in mean MSE between the two methods to be so small that random variation caused the standard approach to outperform the phase folding approach in a limited number of cases.

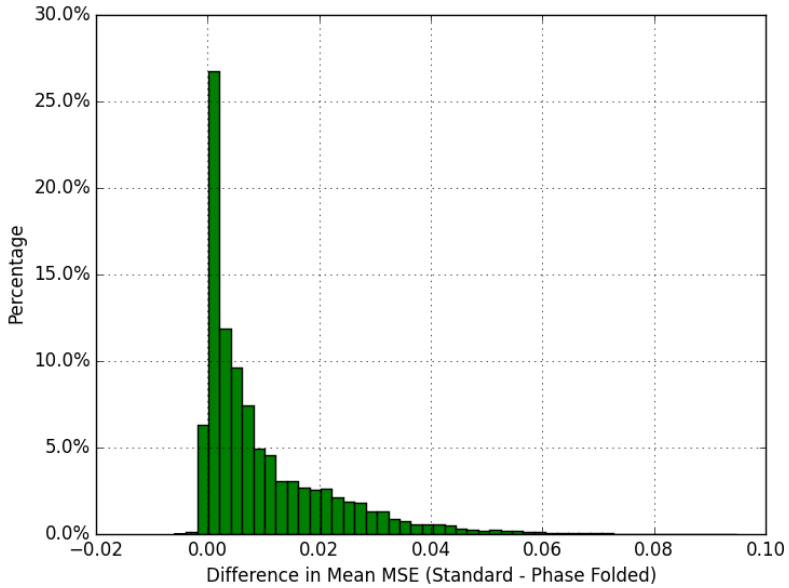


Figure 34. Two Signals Mean MSE Difference.

It was seen in previous simulations that small numbers of periods resulted in a smaller gap between the resulting phase folded and standard MSEs. When distorted components are used as a base for the denoised signal, these small gaps can be eroded almost completely. Random variations then have the potential to give a slight edge

to the standard approach in a limited number of situations. Figures 35 and 36 show the effect of noise levels and sample sizes on the results. Higher noise levels (lower SNR) improved the performance of the phase folding method relative to the standard approach, while lower sample sizes appeared the have the same effect.

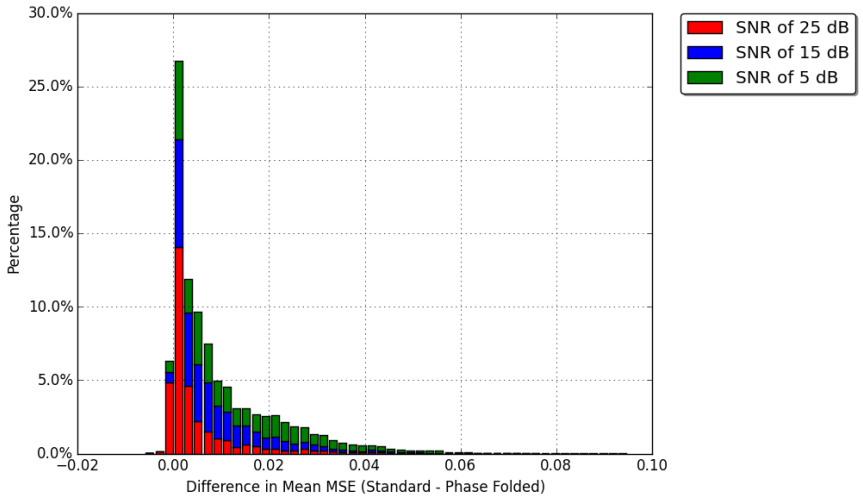


Figure 35. Two Signals Mean MSE Difference - Noise.

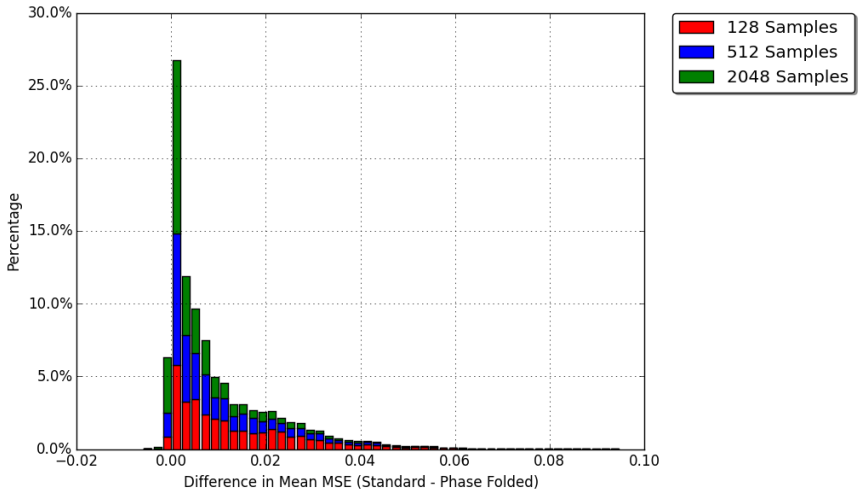


Figure 36. Two Signals Mean MSE Difference - Points.

To test the effects of the various factors on the phase folding mean MSE, a simple linear regression was ran. Using the parameters in Table 6, and a significance level

Table 6. Two Component - Regression Parameter Estimates.

Factor	Estimate	P-value
Intercept	-4.639	< 0.001
Period_1	0.003	0.163
Amplitude_1	0.223	< 0.001
Period_2	-0.004	0.028
Amplitude_2	0.237	< 0.001
Points	0.448	< 0.001
Noise	-0.103	< 0.001
Order	0.007	0.301

of .05, it was found that the period for component one was not significant(P-value = 0.163), while the period for component two was (P-value = 0.028). The mean MSE increased on average as amplitude for component 1 (P-value < 0.001), amplitude for component 2 (P-value < 0.001), the number of samples (P-value < 0.001), and the SNR (P-value < 0.001) increased and as the number of periods for component 2 decreased (P-value < 0.001). Additionally, it was determined that the order in which components were extracted had no significant effect on the MSE (P-value = 0.301). These results are given in Table 6. The model resulted in an adjusted coefficient of determination of 0.902, which was only increased to 0.922 with the addition of the test signals to the model. When added to the model, all test signals were found to be significant, however blip was found to be of only moderate significance when used as either signal and parabolas was only moderately significant when used for the second signal. Effects of the other parameters remaind significant (or not) as for the model which didn't include the test signals as predictor variables.

Table 7. Improved Signal Quality - Three Components Factors and Levels.

Signals	Step, Wave, Blip, Blocks, Bumps, HeaviSine, Doppler, Angles, Parabolas, Time Shifted Sine
Samples	2^7
Periods	5, 7, 11
Amplitude Multiplier	.1, .5, 1
SNR	25, 15, 5

Three Components.

The final test for the effectiveness of the phase folding denoising approach was to denoise signals consisting of three components. The simulation combined all of the $10 \times 9 \times 8 = 720$ possible combinations of three test signals at each of the three noise levels. Each of these combinations was then tested using the $3 \times 3 = 27$ possible choices for amplitude multipliers and $3 \times 2 \times 1 = 6$ combinations of periods to bring the number of simulations up to $720 \times 3 \times 6 = 349,920$. Due to this vast number of combinations only signals with 2^7 samples were used. These factors and corresponding levels are summarized in Table 7.

Using the information from the previous simulation, it was expected that there would be a slightly higher fraction of instances where the standard approach produced a lower error than the phase folding method. This was expected due to the the added variation caused by the additional components coupled with the use of lower periods for each component. However, it was believed that these variations would be counteracted by the use of only 2^7 samples for each of the simulations, which would

cause an improved result in the previous simulation.

The results shown in Figure 37 demonstrate an increased level of negative values (standard method produced lower mean MSE) constituting 12.9% of the simulations. A further investigation into the underlying causes of these negative values found that nearly all of them were the result of the highest levels of noise. The stacked histogram shown in Figure 38 shows that the variation in performance at the SNR of 5dB (closely) approximates a normal distribution. This result implies that it is difficult to discern a difference between the performance of the phase folding method and the standard approach, although overall, in the majority of scenarios (81.1%), the phase folding method produced lower mean MSE.

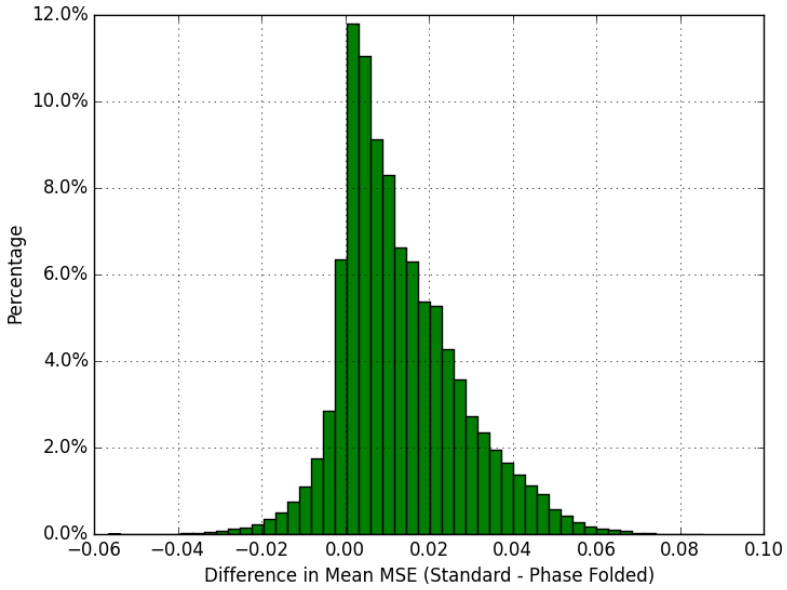


Figure 37. Three Signals Mean MSE Difference.

These results were opposite of those seen in the two signal simulations, where the higher noise improved the performance of the phase folding approach over that of the standard approach. A decrease in the sample size had a similar effect, where small sample sizes cased the phase folding approach to outperform (lower MSE) the

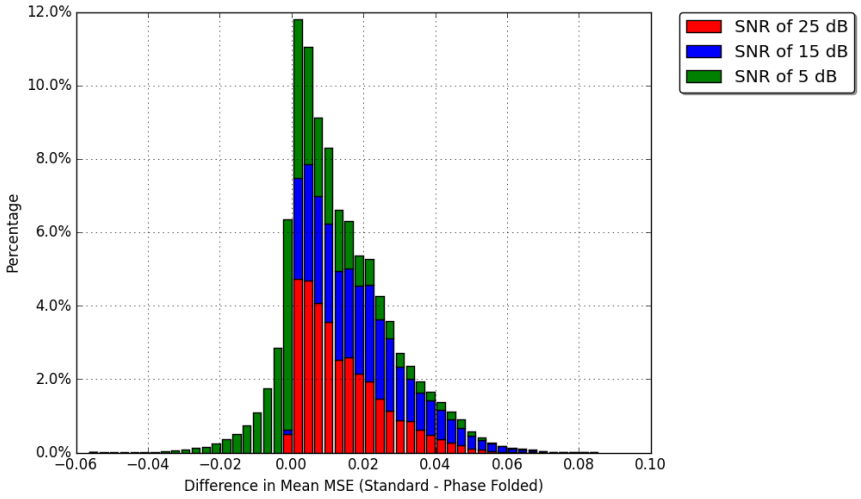


Figure 38. Three Signals Mean MSE Difference - Noise.

standard approach. It is believed that, when combined with such high levels of error, that there were not a sufficient number of samples to counteract the effects of the increased components, which caused the extracted components to have higher levels of noise. To test this hypothesis, a simulation with 2^9 samples was ran with a SNR of 5dB. The results of of this simulation, shown in Figure 39, clearly show that the phase folding approach outperforms the standard approach with a higher sample size and a SNR of 5db.

This chapter demonstrated the utility of combining phase folding and wavelet denoising to improve signal estimation. The effect of different periods on the process were explored and found to have little to no effect on the overall signal quality, while sample size and error levels had an impact. Different sampling rates and counts were also investigated and found to cause little to no variation in overall signal quality. Finally, it was discovered that for signals composed of multiple periodic components, removing each component in turn using the phase folding and denoising approach improved the quality of the extracted signal in nearly every instance when compared to basic wavelet denoising. Further, overall signal mean MSE was not affected by the

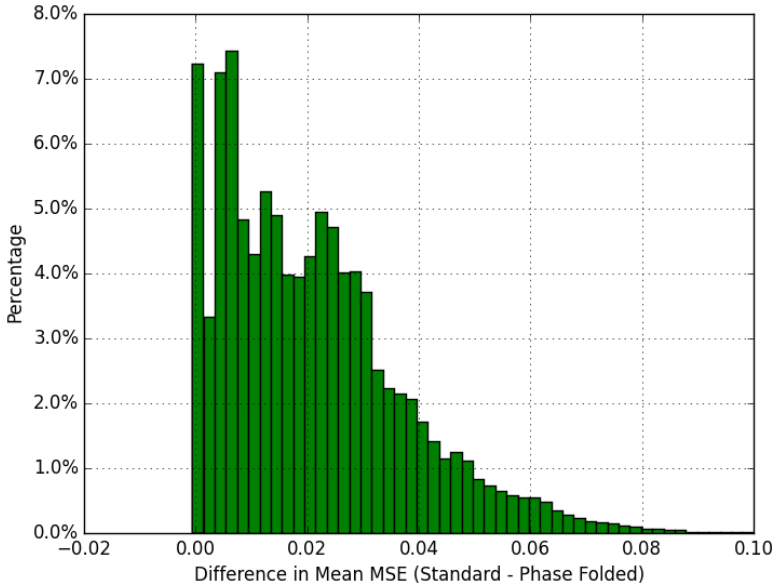


Figure 39. Three Signals Mean MSE Difference - 2^9 Samples - SNR of 5dB.

order in which the components were extracted. The process of extraction individual components and their effects on the the overall signal MSE will be covered in the next chapter.

VI. Component Extraction

In the previous chapter, the phase folding approach outperformed the standard method in nearly every situation. However, these improvements appeared to diminish with an increase in the number of components used to create the signal. This chapter focuses on improving the quality of the extracted components which is desirable in its own right, and is expected to improve the quality of the denoised signal as a whole.

This investigation consists of two main sections, the first of which looks at the basic component extraction technique, as well as various points of interest in component extraction such as order of extracted components, the interaction of the periodic components, and the effects of wavelet selection. Finally, the second section proposes a new method of component extraction aimed at lowering the MSE of each extracted component.

6.1 Iterative Denoising

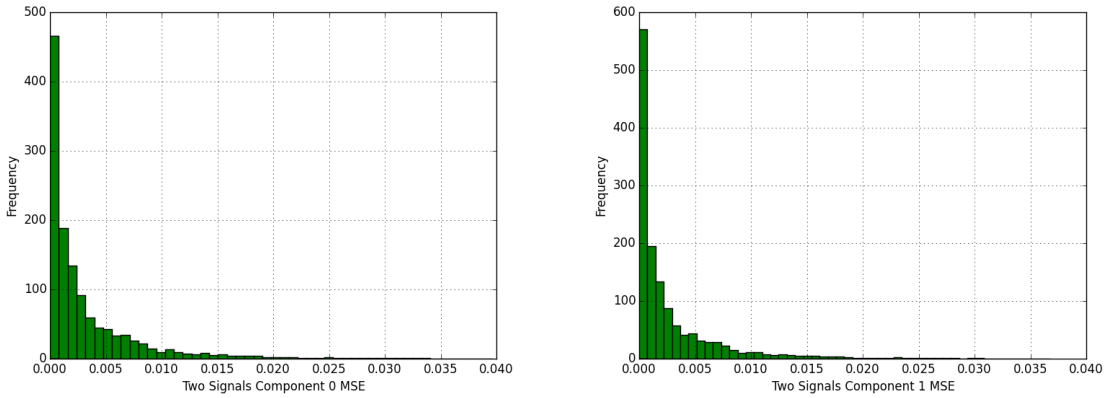
For each of the component parts, the iterative denoising method involves folding the signal over the component's period, denoising the signal using the wavelet method, and subtracting the denoised component from the signal. Once each component has been removed, the best estimate of the complete signal consists of the sum of all the components. This section will take a close look at this process for signals composed of two and three components.

Two Components.

In order to investigate the performance of the iterative denoising method's effectiveness in component extraction, a set of test signals was required. For this simulation, the same test signals described in Table 5 were used, constituting a total of

43,750 different simulations each of which was ran 100 times and the results averaged to remove the effects of random variation. It was expected that the errors for each of the components would be approximately equal to 50% of the error for the signal as a whole, that is, each component contributes equally to the error rate for the signal.

Figure 40 shows the histograms for the MSEs of the first and second components extracted using the iterative denoising method. These distributions look nearly identical with the exception of a significantly large percentage of the total simulations residing in the first bin for Component 2 than that for Component 1. This means that the error for the second extracted component was lowered by removing the first component.



(a) Component 1 **(b) Component 2**

Figure 40. Two Component Signal - Component MSEs.

Other effects seen in previous simulations were also replicated in this experiment. The first of which was the impact of noise, lower noise levels resulted in lower MSEs for both components. More samples also helped to lower the error for each component. The period counts for each signal had little to no effect, while the amplitude multiplier lowered error with lower multiples. The effect of amplitude multiplication is to be expected since lower signal amplitude results in lower noise. Holding all other factors

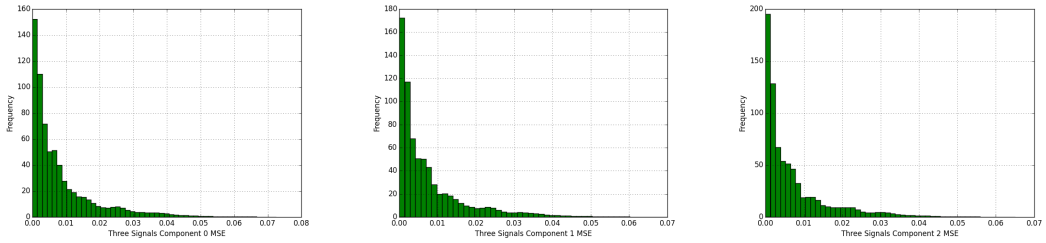
and levels equal, this results in the signals with smaller amplitude always having lower mean MSE values.

In order to determine the effect of individual component's MSE on the overall signal MSE a regression model using the error of each component as predictors for the complete signal error was constructed. This model resulted in a coefficient of determination (adjusted) of .9127, indicating the errors of each component are almost exclusively able to predict the error of the complete signal. The regression model produced component parameter estimates of $\approx .90$ for both of the components error, and an intercept of ≈ 0. This implies that while some error may counteract or exacerbate each other, that lower component error equates to lower signal error. Therefore, if the component extraction technique were to be improved even further, it would most likely improve the MSE of the complete signal estimate. This topic will be explored further later in the chapter.

Three Components.

To test the extraction of three signal components using the iterative denoising method, the same simulation and levels were used as in the previous three component signal test which are summarized in Table 7. It was assumed that the same trend seen in the two component extraction experiment above would continue in the three component case. The only expected change was for an increase in the errors for each component since the addition of more components increases the noise.

Figure 41 shows the resulting histograms for the $349,920$ simulations for each of the extracted components. Once again the most noticeable trend is the compression of the histogram for each successive component. Component 3, therefore, has significantly less error than Component 1, so much so that both axes scales were changed to account for it.



(a) Component 1 **(b) Component 2** **(c) Component 3**

Figure 41. Three Component Signal Component MSEs.

Once again, running a regression model was constructed using the component MSEs as predictors for the complete signal MSE. It was found that the component errors had a significant ability to predict the overall signal error with a coefficient of determination (adjusted) value of .9699. The parameter estimates for component 1, component 2, and component 3 were 1.077, 1.126, and 0.869 respectively. These results once again demonstrate a relationship between the errors for individual components and the complete signal error. This relationship will be exploited in Section 6.2 in order to improve the overall signal quality.

Harmonic Periods.

One large potential drawback to the iterative denoising approach to component extraction is the prospect of components with harmonic frequencies. When extracting components with harmonic frequencies, the iterative denoising approach may not be able to completely remove the effects of the other components. This is due to the fact that the same pieces of other components occur at the same points when folded over another component's period, compounding the effect.

To test the impact of these harmonic frequencies, a simulation similar to the two component extraction experiment was ran using periods of 5, 10, and 15. The factors and levels of the experiment are summarized in Table 8. It was anticipated that

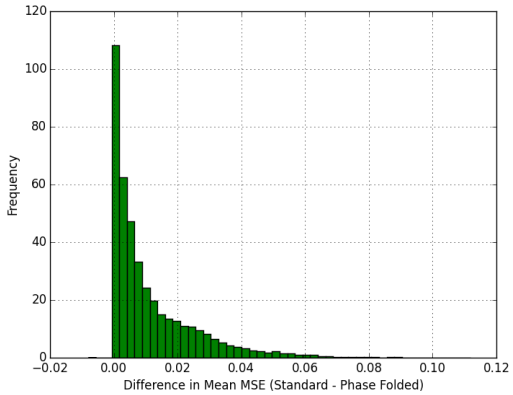
Table 8. Harmonic Effects Experiment - Two Components Factors and Levels.

Signals	Step, Wave, Blip, Blocks, Bumps, HeaviSine, Doppler, Angles, Parabolas, Time Shifted Sine
Samples	2^7, 2^9, 2^{11}
Periods	5, 10, 15
Amplitude Multiplier	.1, .5, 1
SNR	25, 15, 5

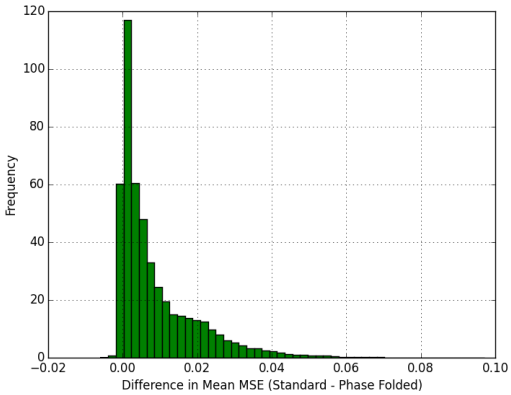
there would be nearly identical differences between the phase folding and standard approaches for both choices of period. However, the component errors were expected to be significantly higher for the harmonic frequencies due to the compounding error effect previously discussed. Other factors such as SNR and amplitude multipliers were believed to have the same effect as seen in previous experiments.

Figure 42 shows the histograms of the standard MSE minus the phase folded MSE for both the harmonic and prime periods. Surprisingly, the iterative denoising approach performed better with harmonic periods than prime periods. In retrospect, this should have been obvious since both components were essentially denoised twice, each time improving the picture of the whole signal more. This effect would, therefore, have an obviously negative impact on the MSEs of both components, which can be seen in Figures 43 and 44 where errors were essentially doubled.

When the effect of the individual factors were investigated, the only noticeable difference was the effect of period order on component extraction. In the prime period case, period order had little to no effect on the MSE of the component. For the harmonic periods the higher the frequency (more periods) of the first component

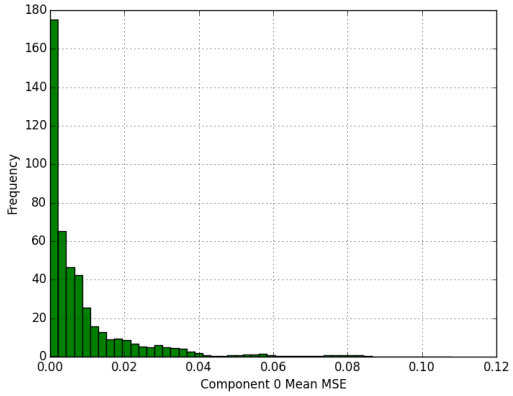


(a) Harmonic Periods

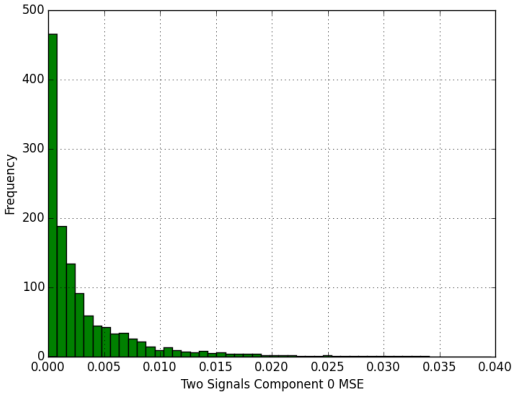


(b) Prime Periods

Figure 42. Two Component Signal Harmonic and Prime Period Comparison.

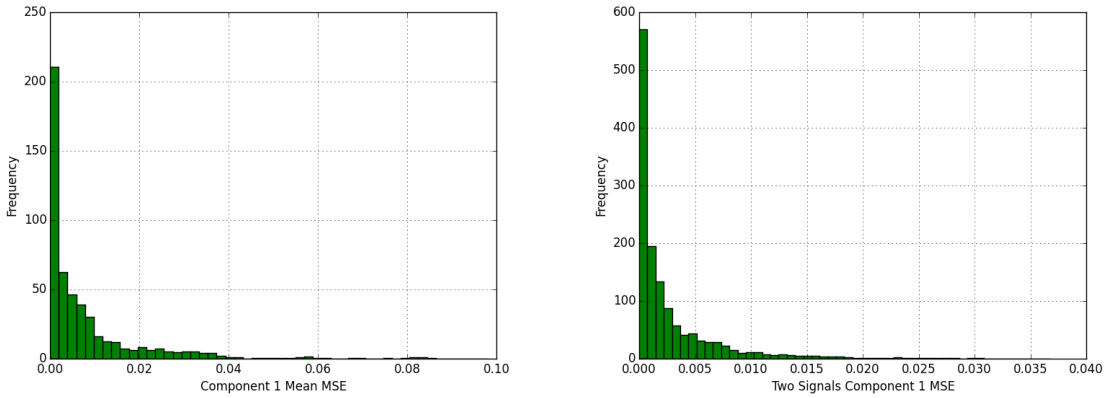


(a) Harmonic Periods



(b) Prime Periods

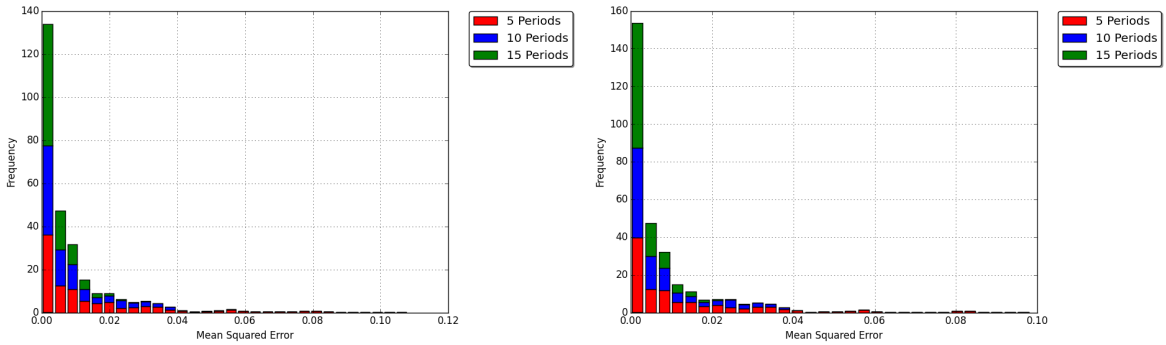
Figure 43. Two Component Signal - Component 1 - Harmonic and Prime Period Comparison.



(a) Harmonic Periods **(b) Prime Periods**

Figure 44. Two Component Signal - Component 2 - Harmonic and Prime Period Comparison.

extracted, the lower the MSE of both components. Figure 45 shows the effect that the period of component 1 has on the MSE of both components.



(a) Component 1 **(b) Component** 2

Figure 45. Two Harmonic Components - Effect of Period Order on Component MSE.

Wavelet Selection.

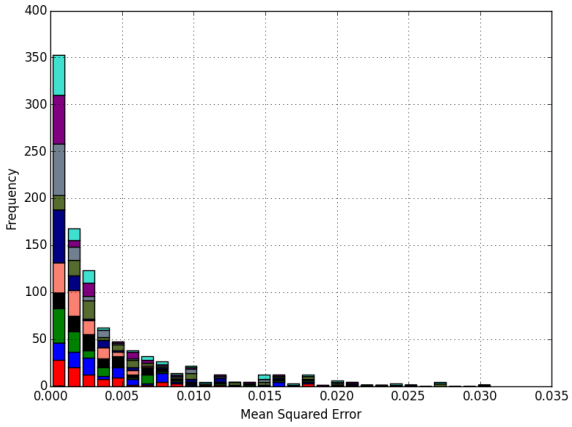
The choice of the wavelet used to denoise a signal can have a large effect on the resultant MSE. To test the effect of seven different wavelets on the test signals, an experiment using the factors and levels summarized in Table 9 was ran. This

Table 9. Wavelet Effect on MSE.

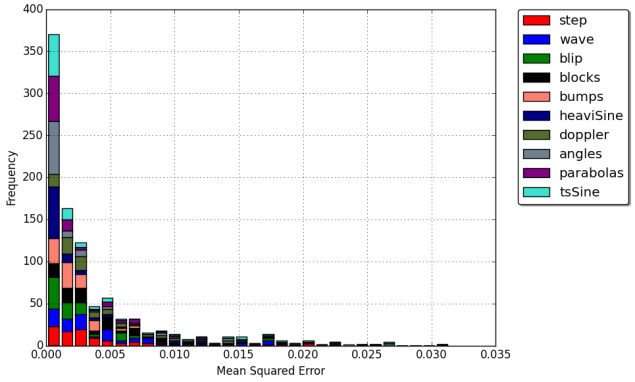
Signals	Step, Wave, Blip, Blocks, Bumps, HeaviSine, Doppler, Angles, Parabolas, Time Shifted Sine
Wavelets	db4, db5, sym5, sym7 coif4, coif5, haar
Samples	2^7, 2^9, 2^{11}, 2^{13}
Periods	3, 5, 7, 9
Amplitude Multiplier	1
SNR	20, 15, 10, 5

experiment consisted of extracting the sole component of the signal for 610 simulations ran for each of the 7 different wavelets. It is was assumed that the more similar a wavelet is to the signal, the lower the MSE. This means that shapes such as Step and Blocks were assumed to have lower MSEs when denoised with the haar wavelet, while more sinusoidal signals such as HeaviSine, Parabolas, and Time Shifted Sine would have lower MSEs with higher order Daubechies wavelets.

Figures 46-49 show the distribution of the phase folded MSE for each of the wavelets by test signal. Overall, the db4 wavelet appears to have the lowest MSE of the 7 wavelets tested, while the haar by far has the worst. The various signals appear to have a similar distribution across all the wavelets tested, leading to the conclusion that any effort to optimize component extraction by matching wavelets to signal shape would not be fruitful.

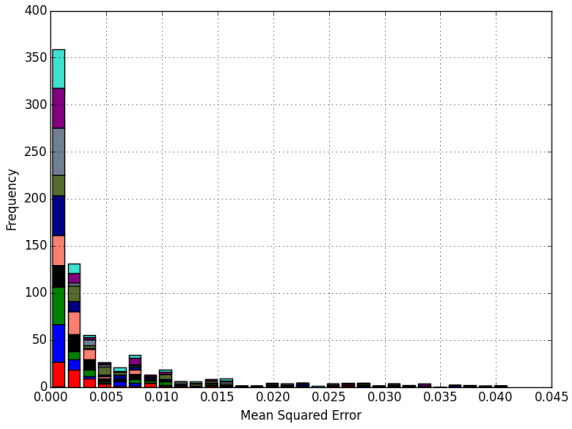


(a) sym5

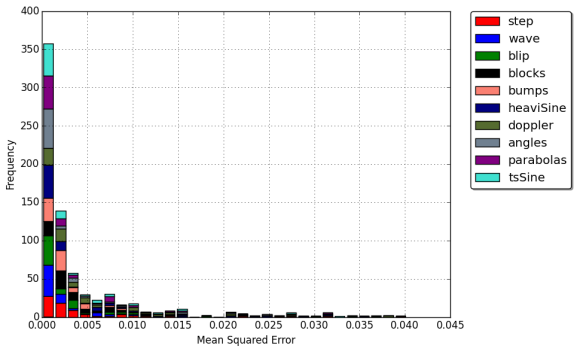


(b) sym7

Figure 46. Wavelet Effect on Phase Folded MSE.

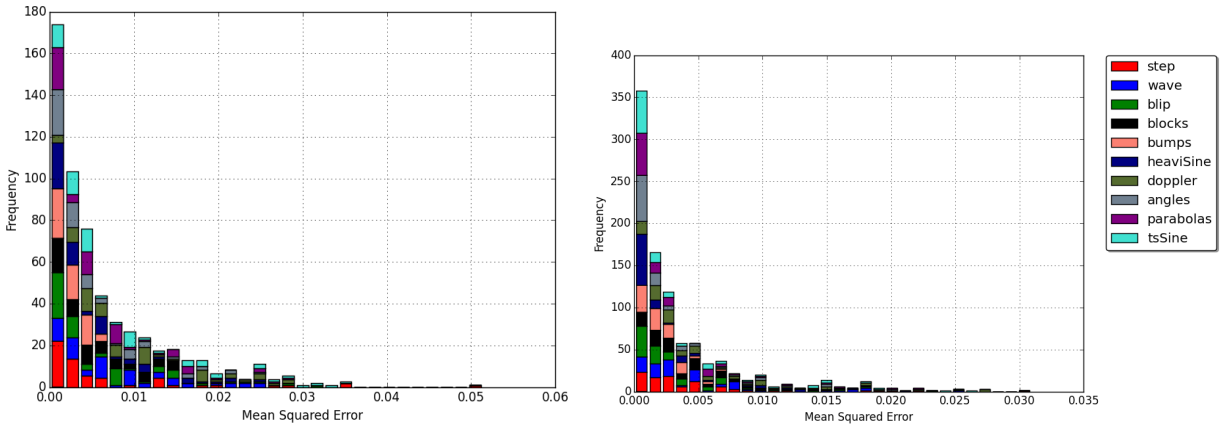


(a) coif4



(b) coif5

Figure 47. Wavelet Effect on Phase Folded MSE.



(a) haar **(b) db5**

Figure 48. Wavelet Effect on Phase Folded MSE.

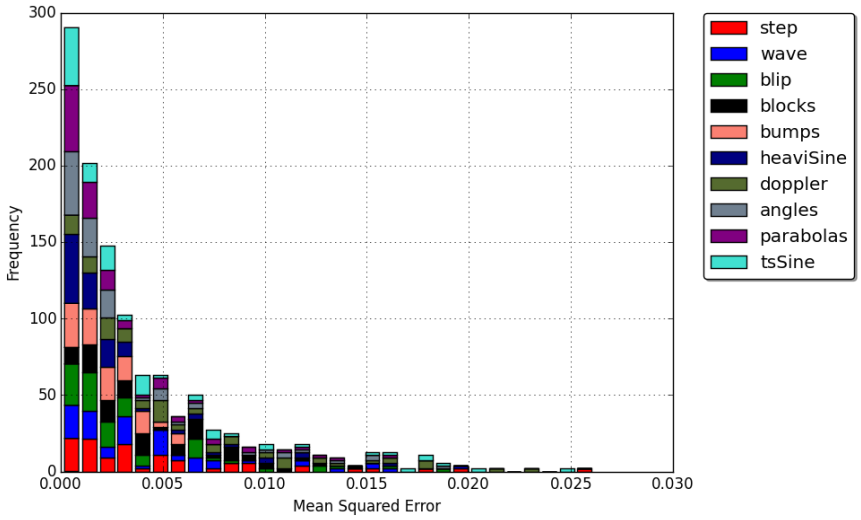


Figure 49. Wavelet Effect on Phase Folded MSE - db4.

Effects of Variation.

In the previous chapter it was found that order did not affect the total error, however, in this chapter it was found to effect the quality of the extracted components. After a component was extracted, subsequent component extractions appeared to have lower mean MSEs. The expected cause of this phenomenon was that the extraction of earlier components removed variation, which resulted in a cleaner signal for subsequent component extractions. When this analysis was expanded to encompass all of the $43,740$ two component simulations, it was found that in the majority of the cases (61.2%) component 1's mean MSE was higher than component 2's mean MSE. These results show there there is an order effect on the component mean MSEs, however, that this effect does not extend to all combinations of factor levels.

One potential cause for this discrepancy among the simulations was the effect of components inherent variation. It was theorized that for the 38.8% of the cases where the second component had a higher mean MSE than the first component would have a smaller overall effect on the signal than the second component. It was found that in 96.9% of cases where the amplitude multiplier for the first component extracted was higher than that for the second component, the first component had a higher mean MSE than the second component. In a similar effect, it was discovered that in 79.8% of the cases where the amplitude multiplier for the second component was greater, that it also had the higher mean MSE.

6.2 Improved Component Extraction

Throughout this chapter two insights suggested a means in which to improve the signal quality even further than through a simple iterative denoising approach. The first of which was the verification that the more accurately individual components could be extracted, the lower the error in the overall denoised signal. The second

insight was that the order of extraction, and the inherent variation of the components extracted, effected the mean MSE of the components. The aim of this section is to improve extracted component quality, which will improve the overall signal quality, by utilizing this knowledge.

To accomplish this goal, three methods of improved component extraction will be tested. The first method extracts signal components in several permutations and only saves a component when it is the last extracted (ordered extraction). Method two (amplitude extraction) extracts the components with the highest effect on variation first in an effort to lower the mean MSE of subsequent components. The third method (recursive extraction) combines these two approaches by first extracting components in order of effect on variability (amplitude extraction), which is followed by a recursive component update process similar to the process used in method one. Finally, all three methods will be compared and the areas in which they excel will be discussed.

Ordered Extraction.

In Section 6.1 it was discovered that while order had little to no effect on the overall mean MSE of the denoised signal, it did influence the mean MSE of the different components. It was found in 61.2% of cases that the earlier in the iterative phase folding and denoising process a component was removed, the higher the component's mean MSE. Leveraging this phenomenon for a signal with n components, the ordered extraction technique performs n complete iterative denoising operations, extracting a different component last for each of the n runs, saving the last component extracted. In effect, this results in each component being removed last by the iterative denoising operation.

To test this extraction technique, the factors and levels summarized in Table 10 were used. This resulted in a total of $14,580$ different combinations of factors and

Table 10. Improved Component Extraction - Two Components Factors and Levels.

Signals	Step, Wave, Blip, Blocks, Bumps, HeaviSine, Doppler, Angles, Parabolas, Time Shifted Sine
Samples	2^7
Periods	5, 7, 11
Amplitude Multiplier	.1, .5, 1
SNR	25, 15, 5

levels where the mean of 100 simulations was used as the representative value. It was expected that in roughly 60% of cases, when the mean MSE for component 1 was greater than that for component 2 using the iterative denoising method, that the mean MSE for component 1 would be less using ordered extraction than that observed for the basic iterative denoising method, since component 1 was, in effect, removed later in the process. Since component 2 was removed in the exact same fashion using both algorithms, no difference between its mean MSE was expected for the two algorithms. It was also believed that the overall signal mean MSE would decrease since the mean MSE of one of its components was expected to decrease while the other stayed constant.

It was found that in 82.5% of the cases, the mean MSE for component 1 was lower using the ordered extraction method then when using the basic iterative approach. Therefore, in 89.5% of the cases where the mean MSE for component 1 was less than that for component 2 using the basic iterative method, the ordered method improved (lowered) the mean MSE of component 1. This result far exceed the expected improvement of only 60% of cases, and demonstrates that order is in fact an

important factor in component extraction. Though the ordered extraction technique did not change the mean MSE of component 2, since both methods extracted the second component in the exact same way, it ultimately resulted in a lower mean MSE for the complete signal in 74.2% of cases by an average of 10.1%. These results are summarized in Table 11.

Amplitude Extraction.

The amplitude extraction technique attempts to lower the mean MSE of the components by extracting the components with the largest effect on the signal variability first. This process is conducted by extracting each component from the noisy signal individually while noting the variability of the resultant residuals. Once completed, the component which, when extracted, lowered the residual variability the most was removed and the process started over using these residuals as the signal. This process proceeds until all of the components have been extracted.

It was believed that the amplitude extraction technique would improve the mean MSE for the components and complete signal in approximately 50% of the $14,580$ simulations, the same as those summarized in Table 10. The 50% estimate was proposed to incorporate all of the instances in which the amplitude of component 2 was higher than component 1 ($33.\bar{3}$% of the instances), and half of the instances in which the amplitude multipliers for the signals were equal ($16.\bar{6}$% of the instances).

The amplitude extraction technique result, summarized in Table 11, show an improvement in mean MSE for component 1 in 60.0% of the instances and no change in 28.1% of the instances. This result is in stark contrast to the results for component 2, where in only 13.9% of the instances was the amplitude method mean MSE less than that of the basic iterative denoising approach.These changes in component mean MSE values caused a near equal number of instances where the amplitude extraction

technique outperformed the basic iterative denoise technique for full system error (36.7%), and where the amplitude technique performed worse (35.23%).

Recursive Extraction.

The recursive extraction technique attempts to take advantage of the benefits of both the order extraction technique and the amplitude extraction technique. First the components are extracted using the amplitude extraction technique, where the component which lowers the residual variation the most is extracted first. The components are then updated using an order extraction type approach, where the last component extracted using the amplitude technique is considered a clean component. The clean component is subtracted from the original noisy signal, the remaining components are then extracted from this updated noisy signal in the same order used in the amplitude extraction phase. The component extracted last is then considered to be clean.

This process continues, adding a new component to the clean component list until no more components remain to be extracted. These new, clean, components are now an updated version of the original extracted components. This process can then be repeated, flipping the order of component extraction each iteration, as many times as desired. It was believed that this hybridized method would improve the component mean MSE values as well as the complete signal mean MSE for all of the signals at each iteration up to a point at which there would be diminishing returns for further updates.

With a single update round the recursive extraction technique resulted in an improvement over the basic iterative denoise approach in 83.3% and 23.8% of cases, for components 1 and 2 respectively. Overall, with only a single update round, the recursive extraction technique improved over the base iterative denoise method in

89.6% of the instances while improving in at least one component in only 84.1% of cases. This means that, for a limited number of instances, that both component errors increased while the complete system error decreased.

Using two update rounds, the recursive method improved over the basic iterative method in 76.1% of cases for component 1, and 34.8% cases for component 2. Resulting in improvement over the basic iterative technique for the whole signal in 80.2% of the instances. Once again, in a small number of case (1.2%), the error for both components increase while the error for the complete signal decreased. The results for the recursive extraction technique, for both one and two update rounds, as summarized in Table 11.

Comparison of Improved Extraction Techniques.

Table 11 shows a summary of the different extraction methods as well as the percentage of instances in which they outperfomed the oridinal iterative denoising algorithm. For overall system error the best performing algorithm was the recursive extraction technique with a single update round. The recursive extraction with a single update outperformed every other algorithm in the majority of the various factor and level combinations and most of the time with the largest margin. However, the algorithm with the lowest component error was the order extraction technique. This was a surprising result because the order technique could only possibly improve upon component 1's mean MSE, since it used the same method as the basic algorithm to extract component 2.

Table 11. Percentage of Instances in which Mean MSE Improved by Extraction Method.

Extraction Method	Updates	Improved Mean MSE		
		Comp 1	Comp 2	Total
Ordered	–	82.5%	0.00%	74.2%
Amplitude	–	60.0%	13.9%	36.7%
Recursive	1	83.3%	23.8%	89.6%
Recursive	2	76.1%	34.8%	80.2%

VII. Period Detection

Without the correct period it becomes very difficult, if not impossible, to isolate the periodic components of a signal. Therefore, an accurate period detection method is one of the most important tools needed for a successful automated periodic component extraction system. Section one discusses the difficulties of creating such a system as well as a potential tool for an improved implementation. The second section explores the three most predominately used algorithms and their effectiveness in various situations. The final section proposes a new method of period detection based on the iterative phase folding technique, and compares its performance to the industry standards.

7.1 Component Effects

The addition of more components to a system increases the difficulty of period detection in two key ways. Part one covers the difficulty which arises from the addition of error as a result of each new component. The second section discusses the exponential growth of the solution space which accompanies the linear growth of components.

Inherent Error.

In Chapter V the effects of various factors on the quality of denoised signals was tested. When the signals summarized in Table 2, which consisted of only a single signal with even samples and prime periods, the largest mean MSE found was 0.0174 while the average mean MSE was 0.0014. With the addition of another component (signals summarized by Table 5), the largest mean MSE and average mean MSE jumped up to 0.0761 and 0.0056 respectively. In both the largest and average cases,

there was approximately a four fold increase in the level of mean MSE observed. Therefore, the addition of more components to a signal greatly increase the observed error, and consequently, lowers the chance of an accurate period detection.

In addition to the increased error as a result of more components, there is the compounding effect of improper folds for component extraction. Since the increase in components raises the error levels, it becomes more difficult to accurately determine the period of a component. If the period of a component is not exactly determined, it creates additional error for subsequent components (discussed in Section 4.2), making the detection of their period even more unlikely. There is then a cascading effect where the estimate for each successive period is worse than the last.

Solution Space.

The addition of more components to a signal increases the signal error and causes the accurate detection of component periods to become more difficult. These additional components also bring with them an exponential growth in the solution space. This exponential growth comes from two effects, the first is the need to check for more periodic components. If a researcher wishes to check n periods for a one component signal, they would then need to check n^k periods for a k component signal. The second cause for exponential growth of the solution space is the additional error discussed previously. Since more components increase the error, which concurrently increases the difficulty of accurate period detection, in order to maintain a level of accuracy, more periods must be tested for each component.

The shape of the solution space can also cause a researcher to increase the number of tested periods. Consider the solution space, created using the Phase Dispersion Minimization (PDM) detection algorithm, of the two component signal described in Table 12. The solution space, shown in Figure 50, indicates the best estimate for the

two periods where the Θ statistic (where Θ is the bin variance divided by the signal variance as used by the Phase Dispersion Minimization algorithm) reaches its global minimum. As can be seen in the green and blue lines crisscrossing in Figure 50, there are a large number of candidates which may achieve the lowest Θ. These green and blue lines represent a deep, yet narrow, dips in Θ that occurs suddenly, with little or no gradual change indicating a local minimal value for Θ. These characteristics of the solution space mean that many of the most popular methods of determining a global minimum value will be ineffective.

Table 12. Two Component Signal.

Property	Component 1	Component 2
Signal	Blocks	Blip
Amplitude Multiplier	1.0	1.0
Periods	17	31
Samples	2^{11}	
Noise	5dB	

7.2 Current Approaches

Section 4.3 discusses the theory behind three of the most commonly used period detection algorithms in light curve analysis. Two of the algorithms, Lomb-Scargle (L-S) and Phase Dispersion Minimization (PDM), show great promise as signal agnostic algorithms, whereas Box-fitting Least Squares (BLS) has a more limited application. The limitations of BLS, and its resultant unsuitability to a general automated system, will be discussed first. Following that will be an in depth look at both the L-S and PDM methods, specifically at their ability to ascertain the period of a series of signals.

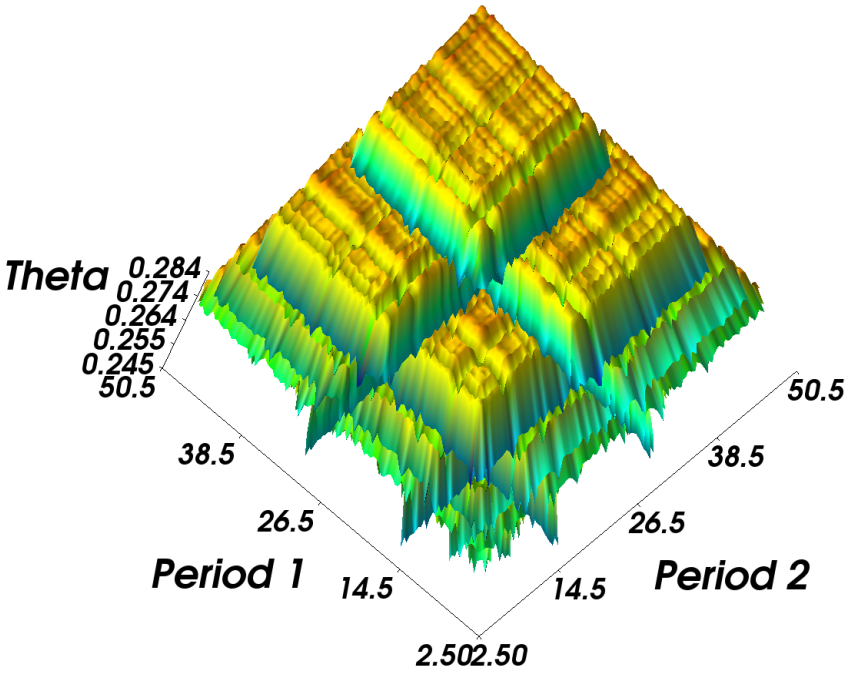


Figure 50. Example Period Detection Solution Space.

Box-fitting Least Squares.

As discussed in Section 4.3, BLS assumes that a signal operates in only two states, a high and a low state. It is also assumed that the signal is in the low state only a small percentage of the time, on the order of 1%. To find the period of a signal, BLS uses these assumptions to fit the block-like signal with a near constant amplitude and a single short drop to every test period. The fold with the corresponding best fitting block signal is then considered the period.

Since the BLS algorithm puts such stringent requirements on the signal in order to be effective, it is a poor choice for a signal agnostic algorithm. Due to this shortcoming, and the shapes of the test signals which do not follow this predefined shape, the BLS algorithm was not explored further than a simple cursory test. In that test, the BLS algorithm was unable to determine the period of a single test function even at the lowest noise levels.

Table 13. Period Detection Test - Factors and Levels.

Signals	Step, Wave, Blip, Blocks, Bumps, HeaviSine, Doppler, Angles, Parabolas, Time Shifted Sine
Samples	2^7, 2^9, 2^{11}
Periods	5, 17, 29, 43
Amplitude Multiplier	0.1
SNR	25, 15, 5

Lomb-Scargle.

The L-S algorithm is an offshoot of the Discrete Fourier Transform (DFT), which in essence attempts to fit a series of sinusoidal waves to the signal under investigation. The fit of these waves can be evaluated in such a way so that they follow a well known statistical distribution, the exponential distribution. Due to this relatively unique feature, a researcher can set strict limits on their desired error levels.

Though the ability to set these error limits hints at a more autonomously-friendly period detection algorithm, L-S still makes the assumption that the signal is a mix of sinusoidal waves. Though this assumption can be a limiting factor, it is not so strict as to make the algorithm inadmissible like the BLS algorithm. In order to test the effectiveness of the L-S algorithm, a period detection simulation was conducted, the factors and levels of which are summarized in Table 13.

Each of the 10 signals was simulated with 4 different periods at the same amplitude multiplier of 0.1. They were then sampled at 3 different rates and at 3 different noise levels, resulting in $10 \times 4 \times 3 \times 3 = 360$ simulations. Each simulation was then ran

25 times, and the returned periods recorded for each iteration. It was assumed that the L-S algorithm would perform best for the sinusoidal signals such as wave and HeaviSine, while performing poorly for the more blocky signals and jumpy signals such as Step and Bumps respectively.

It was determined that the best means of evaluating the accuracy of the L-S algorithm was by noting the percent correct for each run of 25. A correct period was defined as one in which the estimated period for the algorithm was ± 1 from the true period. Surprisingly, it was found that L-S performed the worst when detecting the period for the wave signal with only a 1% accuracy for 900 total simulations, while scoring a perfect 100% for the step function. L-S also performed well with the parabolas function and the blip function with 100% and 99.6% accuracy respectively, while the remainder of the signals had between 30% and 50% accuracy.

The L-S algorithm also performed better with higher periods, scoring in the 70% range for 29 and 43 periods and only about 30% for 5 and 17 periods. More samples also appeared to help the L-S algorithm by going from 39.5% accuracy for 2^7 samples up to 61.5% accuracy at 2^{11} samples. Interestingly, the level of noise had nearly no impact with accuracy in the 50% range for all three noise levels. Overall, the L-S algorithm averaged 53.3% accuracy for all simulations and runs. These results are all summarized in Tables 14-17.

Phase Dispersion Minimization.

PDM and its offshoots are probably the most widely used period detection algorithms in light curve analysis. The basic algorithm has been around since 1978 and was first developed by Stellingwerf, and has had many variations developed since [57, 45]. PDM uses a similar approach to BLS in that it folds the signal over every test period. Whereas BLS attempts to fit a function to the data for every fold, PDM

takes a more signal agnostic approach and simply compares the variation in a series of bins to the variation of the signal as a whole. The period of the function is considered found when the inter-bin variation reaches it lowest fraction of the overall signal variance.

In order to test the effectiveness of PDM to accurately determine the period of various signals, the same signals used to test the L-S algorithm were employed. Each of these signals, summarized in Table 13, were subjected to the PDM algorithm 25 times and their results recorded. Once again, a period was deemed correct if it was ± 1 of the true period.

Due to the signal agnostic nature, it was predicted that the PDM method would perform equally well for nearly every signal with the exception of Bumps, Doppler, and Wave. It was thought that these three signals have such rapid variation that they may vary too much in the different bins to get an accurate estimate. It was also believed that the more samples and the less noise the better the results, while the period count was believed to be of little to no consequence.

It was found the PDM had between 80% and 95% accuracy for 8 of the 10 signals, with the exception of Bumps (61.1%) and Wave (37.2%). These results lined up well the predicted results, with the exception of Doppler which outperformed expectations, reaching 91.2% accuracy. It was also determined that the number of periods had little to no effect, with the accuracy in the range of $82\% \pm 5\%$. The sample size appeared to have an effect on accuracy, increasing with more samples from 64.3% with 2^7 samples to 93.4% with 2^{11} samples. Finally, only the highest error level, SNR of 5dB, lowered the accuracy of PDM down to 66.7% from the 87.3% and 88.8% at the 15dB and 25dB levels respectively. The PDM method averaged an accuracy level of 80.98% overall.

7.3 Revised Approach

Throughout this dissertation the synergistic benefits of combining phase folding and wavelet denoising have been expounded. Due to the previous results it was believed that these two tools could be combined to create an improved period detection algorithm. This section looks at one possible method of creating a phase folding/wavelet denoising period detection algorithm and compares its performance to that of the L-S and PDM methods discussed previously.

Wavelet Phase Dispersion Minimization.

One of the key properties of BLS, L-S, and PDM is that each of them attempt to, in their own way, fit some sort of function to the data in order to ascertain its period. For BLS this consisted of fitting a block signal with a single dip to every test fold. L-S attempted to fit a series of sinusoidal waves to the signal to find the period. Finally, PDM created a series of bins in which variation for the bin mean was recorded and used to detect periodicity, essentially attempting to fit a piecewise constant function to the signal. Each algorithm is therefore making some inherent assumptions as to the underlying nature of the signal.

One of the greatest benefits of wavelet denoising, especially when combined with different choices of wavelets, is the nearly complete signal agnostic way in which a best fit line is drawn. In order to use this powerful tool to create a period detection algorithm, one key concept from the original PDM algorithm was used. To test for the periodic signal, the PDM algorithm creates a Θ statistic where $\Theta = \frac{s^2}{\sigma^2}$ for each test period. The Θ value is in effect comparing the residuals of a piecewise constant functional fit to the variation of the whole signal.

The Wavelet Phase Dispersion Minimization ($\mathcal{W}$-PDM) algorithm uses this same concept to create its own Θ value. The $\mathcal{W}$-PDM algorithm folds the signal over

each of the test periods. After the signal has been folded, a best fit line is drawn using wavelet denoising. The $\mathcal{W}$-PDM then calculates its Θ statistic by dividing the variation of the residuals from the best fit line by the variation of the original signal. The test folded signal with the lowest Θ value is then considered to be folded over the correct period.

Performance Comparison.

To test the performance of the $\mathcal{W}$-PDM algorithm, the same simulated signals used to test the L-S and PDM algorithms were used, which are summarized in Table 13. Each simulation consisted of 25 runs in which the $\mathcal{W}$-PDM algorithm determined its best estimate of the period. For each run, the $\mathcal{W}$-PDM algorithm was considered to have accurately determined the period if the returned period was ± 1 from the true period.

Tables 14-17 show the results of all the simulations summarized by the different factors. When broken out by signals, $\mathcal{W}$-PDM had the best performance for 3 of the 10 signals, while coming in second for the remaining 7. When broken up by periods, samples, and SNR, $\mathcal{W}$-PDM had the best overall performance with the highest scores in 3 of 4, 2 of 3, and 3 of 3 instances respectively. Overall, the $\mathcal{W}$-PDM algorithm outperformed the other two, averaging 82.2% accuracy over all simulations compared to 81.0% for PDM and 53.3% for L-S.

Table 14. Percentage of Correct Period Detection by Signal.

Signal	L-S	PDM	$\mathcal{W}$-PDM
angles	35.0%	92.9%	91.1%
blip	99.6%	81.1%	89.6%
blocks	32.9%	86.4%	85.4%
bumps	36.6%	61.1%	63.9%
doppler	40.0%	91.2%	88.7%
heaviSine	39.3%	87.2%	89.1%
parabolas	100.0%	85.7%	89.6%
step	100.0%	93.0%	94.1%
tsSine	48.9%	93.9%	90.3%
wave	1.0%	37.2%	40.6%

Table 15. Percentage of Correct Period Detection by Number of Periods.

Periods	L-S	PDM	$\mathcal{W}$-PDM
5	30.9%	80.8%	80.8%
17	34.4%	77.5%	78.5%
29	75.4%	78.1%	84.8%
43	72.5%	87.5%	84.8%

Table 16. Percentage of Correct Period Detection by Number of Samples.

Samples	L-S	PDM	$\mathcal{W}$-PDM
2^7	39.5%	64.3%	60.5%
2^9	59.0%	85.2%	91.2%
2^{11}	61.5%	93.4%	95.0%

Table 17. Percentage of Correct Period Detection by SNR.

SNR	L-S	PDM	$\mathcal{W}$-PDM
5dB	51.4%	66.7%	67.8%
15dB	54.2%	87.4%	87.4%
25dB	54.4%	88.8%	91.5%

VIII. Light Curve Application

Throughout this dissertation, a suite of tools have been developed to aid in the automated analysis of periodic time series data. Up until this point, these tools have only been applied to simulated data. In this chapter these newly developed tools are applied to real light curve data gathered by the Kepler satellite.

The light curves analyzed in this chapter, which were taken directly from the Kepler database, are almost completely unprocessed [36]. Before classification, light curves are typically subjected to a three part data preprocessing pipeline consisting of detrending, filtering, and smoothing. Of these three preprocessing steps, the light curves analyzed in this chapter have only undergone detrending and a fraction of the filtering process.

The detrending process removes the error caused by velocity aberration, which, when dealing with Kepler data, consists of the error introduced by the change in relative position of Kepler to the object of interest. The detrending process undergone by these light curves consists of removing the Cotrending Basis Vectors (CBVs) which contribute the the most error to the signal [28]. Though widely used, the CBV detrending algorithm employed by NASA to flatten out the light curves has the potential to remove periodic signals which don't meet specific thresholds [35]. Therefore, periodic detection algorithms which are able to find patterns in the NASA provided Kepler data, will most likely be able to identify periodic behavior in data detrended using other methods, such as the sigma-clipping algorithm [55].

Before NASA attempts to classify the various light curves provided by the Kepler satellite, the light curves are submitted to a rigorous filtering process. This filtering process can be broken down into four filtering steps [1]. The first filtering step consists of correcting discontinuities which are usually caused by the impact of cosmic rays, or similar particles, on the flux detection system. The second filtering step, the

correction of systematic errors, removes error correlated with an ancillary data. Step three is the removal of excess flux caused by crowding in the flux sensor, usually caused by velocity aberration. The final filtering step is the identification and removal of outlier. These ourliers are identified using a sliding window mean and standard deviation calculation.

Of these four filtering steps, the data provided by NASA to the public has only undergone the first three. This means that the light curves analyzed throughout this chapter likely contain large outliers. However, since there is no justification for the removal of these data points, they have been retained, though they may ultimately have an impact on the results.

The remainder of this chapter is devoted to the analysis of three Kepler light curves, refereed to by their Kepler Input Catalog (KIC) number. Sections one (KIC 2831632) and two (KIC 2835289) will apply the newly developed suite of period signal analysis tools to a light curves containing the flux pattern of eclipsing binary star systems. Sections three (KIC 10358759) applies these same tools to a system containing a single star hosting two planets.

8.1 KIC 2831632

KIC 2831632 is an eclipsing binary (EB) star first identified in 2011 from the first Kepler data release [47]. KIC 2831632 is located at Right Ascension (RA) 285.6901, Declination (DEC) 38.0761 and has an effective temperature of $7045K$. A summary of KIC 2831632 other properties can be found in Table 18.

Period Detection.

The light curve for KIC 2831632 was taken from the 7^{th} quarter release of Kepler data, which was stored in the MAST database for public access [36]. The 7^{th} quarter

Table 18. KIC 2831632 - Summary of Properties [10].

Property	Primary Eclipse	Secondary Eclipse
Period (Days)	2.5731	
Eclipse Depth (Normalized)	0.0092	0.0078
Eclipse Width (Fraction of phase)	0.2342	0.2722
Eclipse Separation (Secondary - Primary)	0.5145	

light curve was chosen primarily due to its small number of discontinuities, which allowed for a clearer result after processing. The plot of the normalized light curve for KIC 2831632 is shown in Figure 51. From the plot it appears as though there is a sinusoidal like shape that occurs periodically.

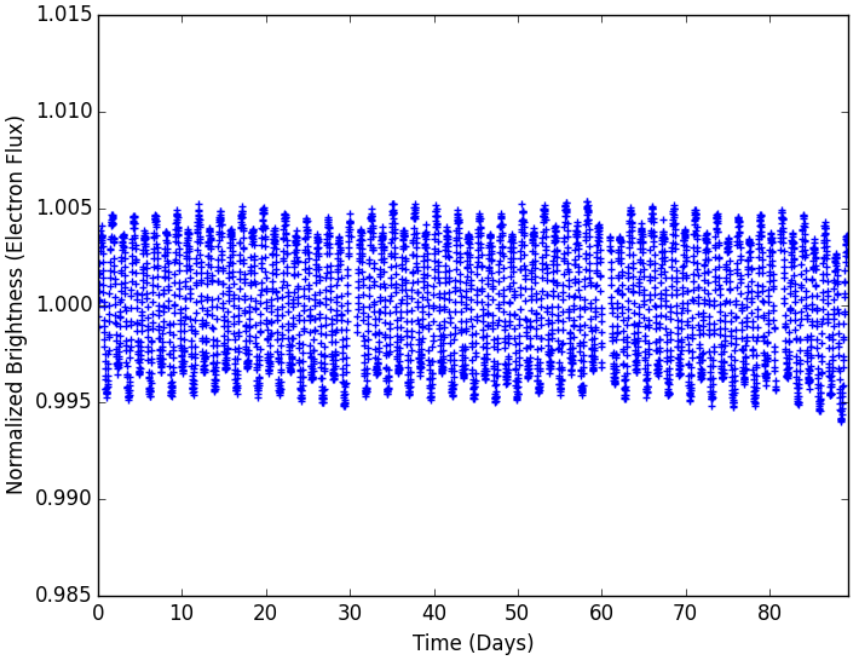


Figure 51. KIC 2831632 - Original Signal.

In order to isolate this periodic component, the period of the phenomenon must first be identified. The period of the signal was found using the Wavelet Phase Dispersion Minimization algorithm ($\mathcal{W}$-PDM) developed in Section 7.3. To find the

best estimate of the signal period the $\mathcal{W}$-PDM algorithm searched between 2 and 75 period counts at intervals of .01. The resulting Θ plot, shown in Figure 52, achieved its global minimum at 34.96 periods. Dividing the maximum time (89.352) by this period count resulted in a calculated period of 2.556 days, whereas the true period found by NASA was determined to be 2.573, making the $\mathcal{W}$-PDM estimate quite accurate.

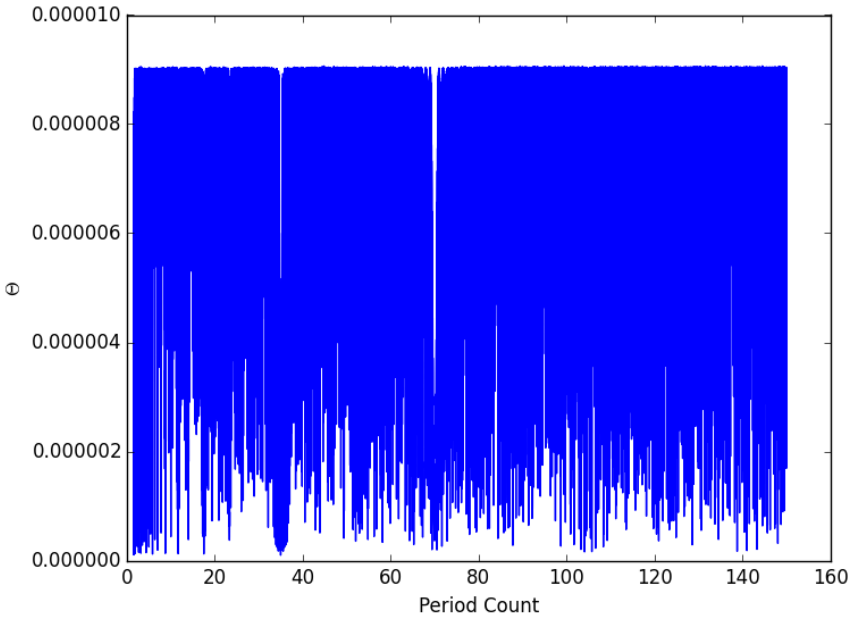


Figure 52. KIC 2831632 - Component One Period Detection.

Component Extraction.

Once the signal period of 2.556 was identified, the periodic component could be isolated and denoised. This process was completed using the basic phase folding and wavelet denoising process developed in Chapter V. More advanced component extraction algorithms, such as those developed in Chapter VI, were not used because they require > 1 components, and KIC 2831632 has only one component.

The periodic component was isolated through phase folding over the period of

2.556, and denoised using the *db*4 wavelet. The resultant can be seen in Figure 53, which is very similar to the NASA generated plot shown in Figure 54. Unfolding the denoised signal results in the clean light curve shown in Figure 55. The newly cleaned figure exhibits a much more regular pattern with very little variation.

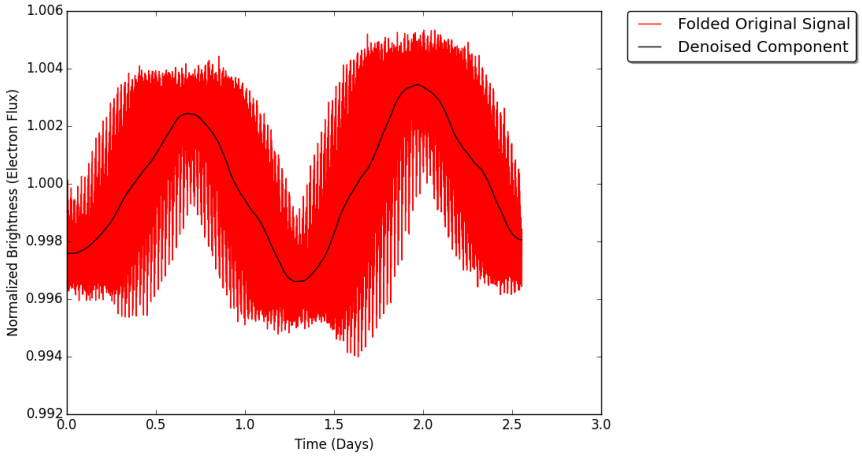


Figure 53. KIC 2831632 - Component One Isolated and Denoised.

The results in this section were found through a completely automated process requiring the user to provide only the period search range and the component counts. Using a light curve that had not been completely filtered, the period found by the $\mathcal{W}$-PDM method was only off by 0.6%. The resulting denoised component had a near perfect match to the isolated component provided by NASA. These results indicate that the suite of developed tools are not suited to only simulated data, but have the potential for real world application.

8.2 KIC 2835289

KIC 2835289 is one of the $1,879$ eclipsing binary stars identified in 2011 from the same Kepler data release as KIC 2831632 [47]. KIC 2835289 is located at Right Ascension (RA) 286.8593, Declination (DEC) 38.0275 and has an effective temperature of $6228K$. A summary of KIC 2835289 other properties can be found in Table 19.

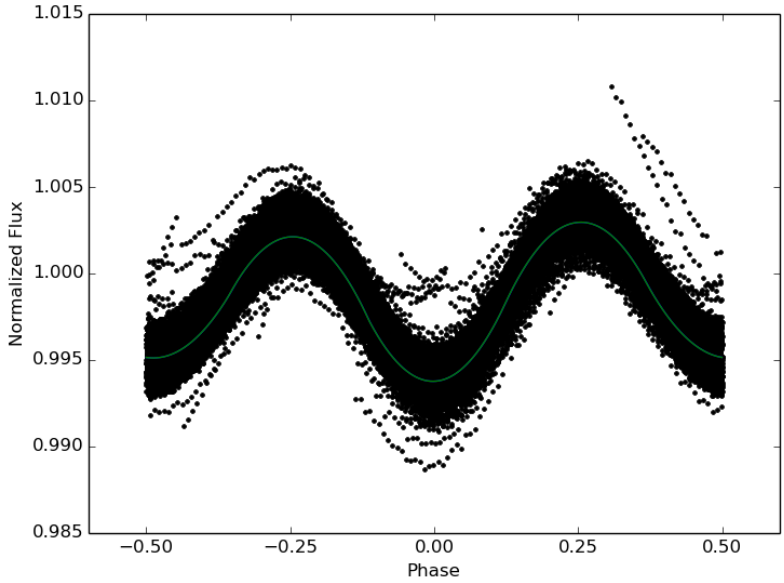


Figure 54. KIC 2831632 - NASA Generated Phase Plot [10].

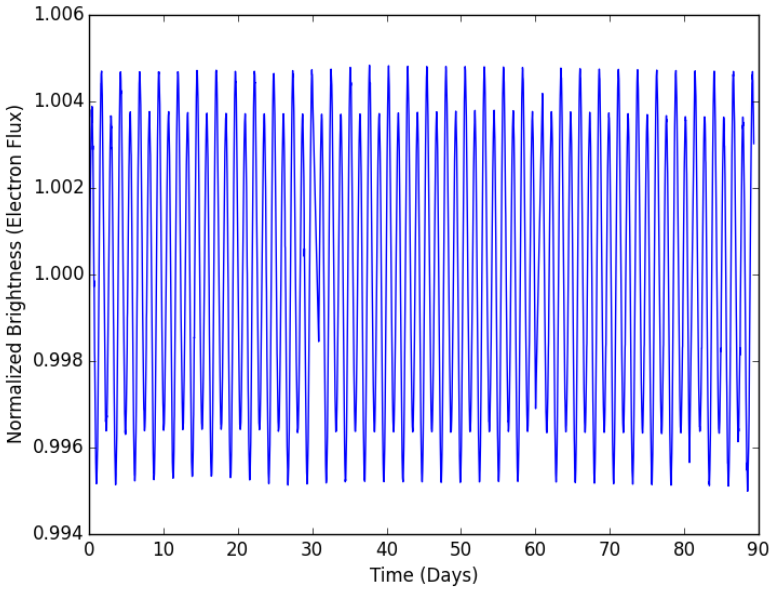


Figure 55. KIC 2831632 - Clean Signal.

Table 19. KIC 2835289 - Summary of Properties [10].

Property	Primary Eclipse	Secondary Eclipse
Period (Days)	0.8577	
Eclipse Depth (Normalized)	0.0226	0.0183
Eclipse Width (Fraction of phase)	0.2371	0.2726
Eclipse Separation (Secondary - Primary)	0.5077	

Period Detection.

Like KIC 2831632, the light curve for KIC 2835289 was taken from the 7th quarter Kepler data release due to its low level of discontinuities. The normalized plot for this light curve can be seen in Figure 56. Figure 56 has a noticeable periodic behavior which appears to be much more rapid than that demonstrated in Figure 51. This observation aligns with the respective periods for the two systems.

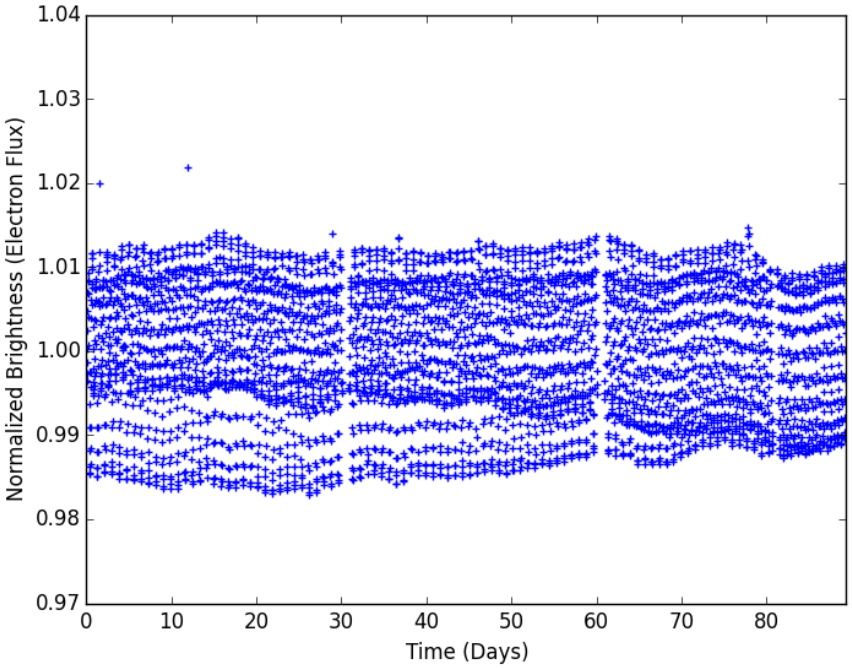


Figure 56. KIC 2835289 - Original Signal.

In order to isolate the periodic component of KIC 2835289 the period must first be found using the $\mathcal{W}$-PDM algorithm derived previously. Provided with a period count search range of $2 - 75$ periods and an interval of .01 resulted in the Θ plot shown in Figure 57. The global minimum for the Θ plot was reached at 104.88, indicating a period of $89.352/104.88 = 0.8519$ days for the solitary component. The true period provided by NASA was 0.8578 days, making the period detection error of the $\mathcal{W}$-PDM only 0.79%

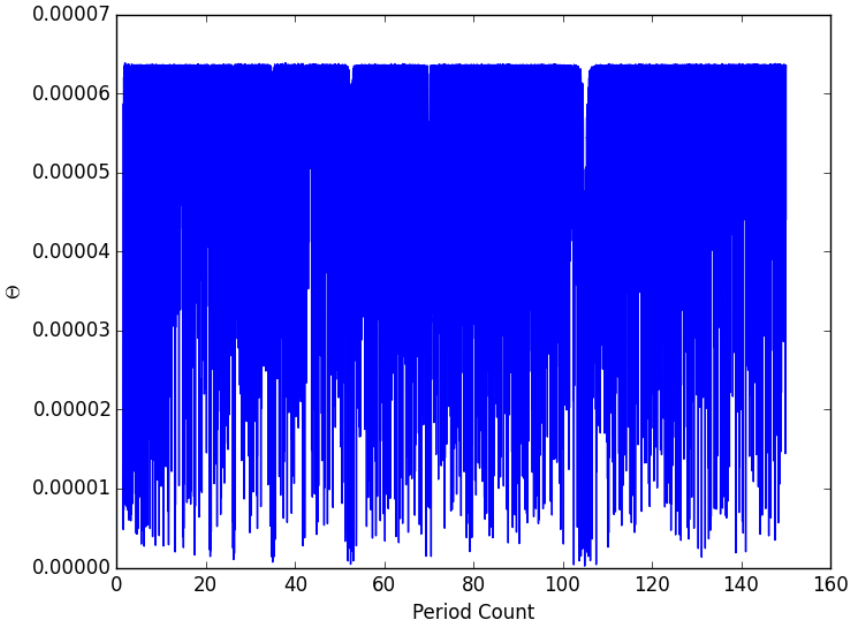


Figure 57. KIC 2835289 - Component One Period Detection.

Component Extraction.

Using the period of 0.8519 days determined in the previous subsection, the solitary component was isolated using phase folding. Once isolated, the component was denoised using wavelet denoising with the *db*4 wavelet. The results, shown in Figure 58, show a clear sinusoidal like pattern to the periodic component. The denoised component in Figure 58 has a very similar shape to NASA isolated compoenent shown in

Figure 59.

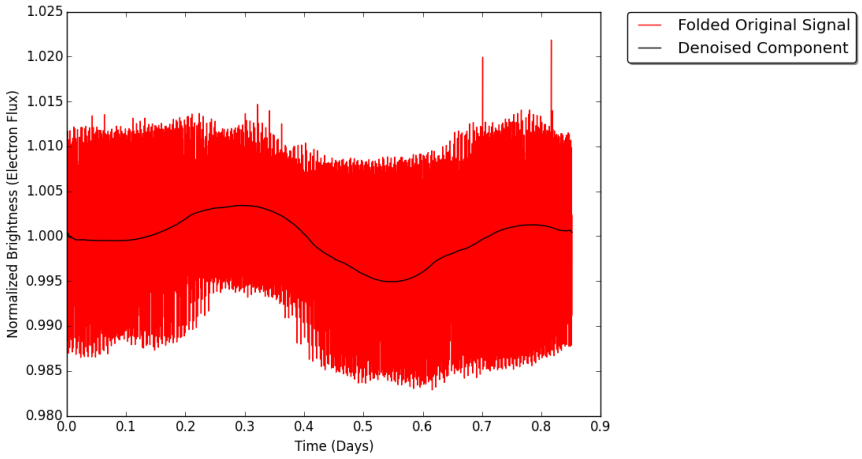


Figure 58. KIC 2835289 - Component One Isolated and Denoised.

Unfolding the isolated component resulted in the clean signal in Figure 60, which has a much smoother form than Figure 56. It should be noted that the seemingly gradual increase in the smaller dip's depth is the result of interaction between the sampling rate and the period of the component. Once again, these results indicate the the newly developed suite of tools can be applied to noisy real world with comparable results to those determined by NASA.

8.3 KIC 10358759

The final real world light curve to be analyzed by the newly developed autonomous tool suite is taken from KIC 10358759, also known as Kepler-29. The actual light curve used in the analysis is a combination of the light curves from the Kepler data release quarters $2 - 6$, which span a total of 459.53 days. The reason for this larger data set is that the periods of the two components, two separate exoplanets, are longer than the periods for the two previous EB systems. Since the flux changes for exoplanets are, in general, much smaller than those for EBs, it was believed that more periods would aid in the detection of the components.

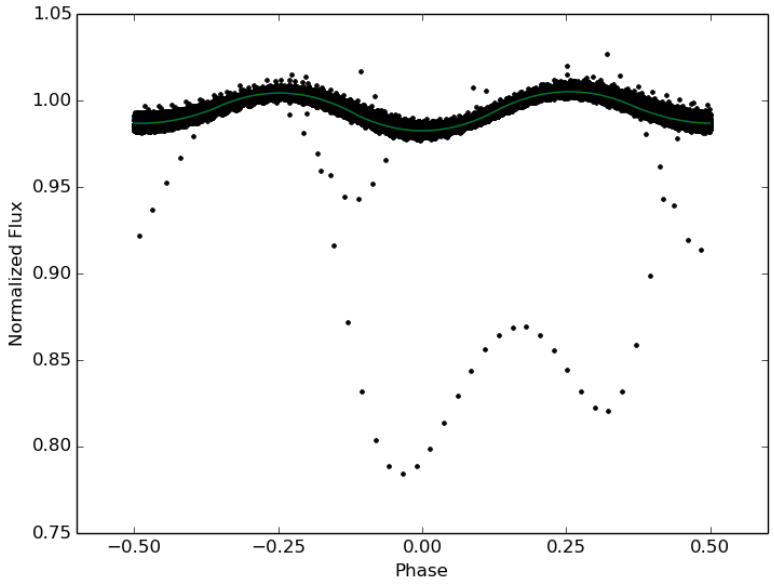


Figure 59. KIC 2835289 - NASA Generated Phase Plot [10].

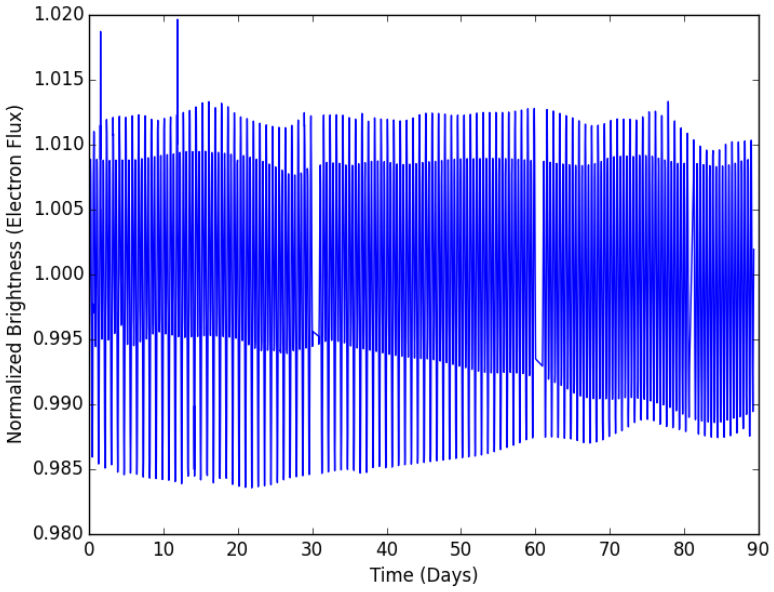


Figure 60. KIC 2835289 - Clean Signal.

Table 20. KIC 2835289 - Summary of Properties [14].

Property	Exoplanet 1	Exoplanet 2
Designation	Kepler-29*b*	Kepler-29*c*
Period (Days)	10.3376	13.2907
Mass (Jupiter Mass)	0.4	0.3

Kepler-29, which is located at RA 19^{h} 53^{m} 23.60^{s}, DEC 47^{h} 29^{m} 28.4^{s}, is a relatively unique system. At the time of its discovery, in 2012, it was 1 of only 80 identified systems with 2 or more planets in orbit [14]. Due to its more complicated nature, having two periodic components, a more detailed and complicated analysis is necessary. A summary of the relevant properties of the Kepler-29 system can be found in Table 20, and the original light curve in Figure 61.

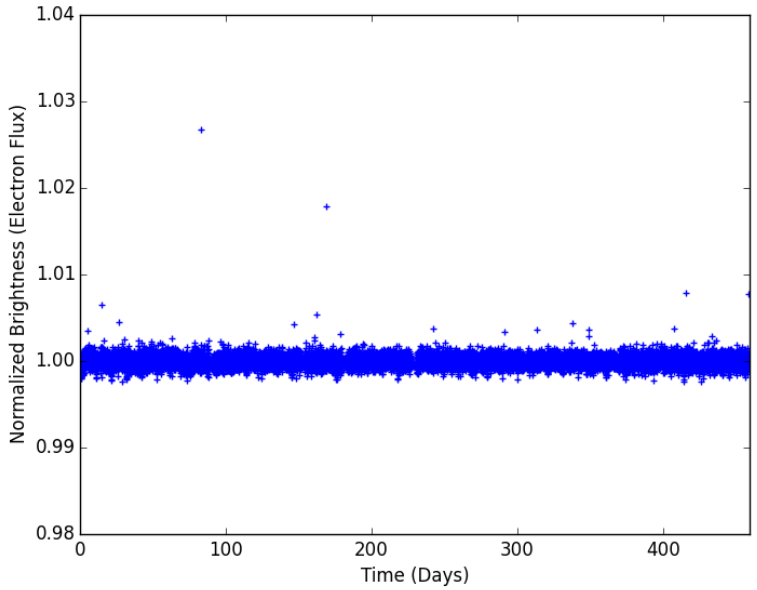


Figure 61. KIC 10358759 - Original Signal.

Component Isolation.

Since there may be some interaction between the two periodic components of KIC 10358759, the first component must be identified and removed before component two can be subsequently isolated and extracted. To find the period of the first component of Kepler-29 the light curve was passed to the $\mathcal{W}$-PDM algorithm with a search range of $2 - 75$ periods and an interval of .01. The resultant Θ plot, which can be seen in Figure 62, reaches its global minimum at 44.50 periods, indicating a period of $459.532/44.50 = 10.327$ days for the first component.

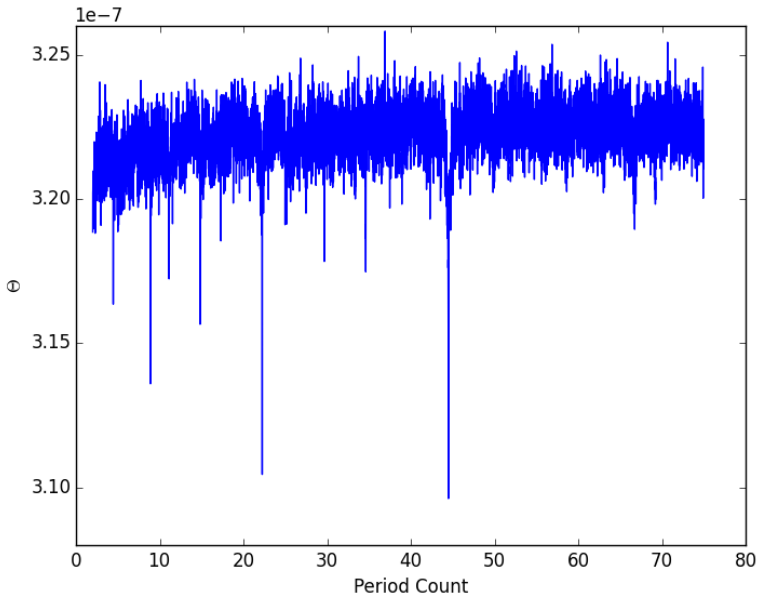


Figure 62. KIC 10358759 - Component One Period Detection.

Phase folding the original signal over the period of 10.327 and denoising using the *db*4 wavelet resulted in the plot shown in Figure 63. The newly denoised component was then subtracted from the original signal and unfolded. This revised flux signal was then passed back to the $\mathcal{W}$-PDM algorithm with the same search range $(2 - 75)$ and interval (0.01) used for component one. The Θ plot, shown in Figure 64, reached it global minimum at 34.62 periods. Therefore the second component had a period

Table 21. KIC 10358759 - Period Summary.

Exoplanet	$\mathcal{W}$-PDM Period	NASA Period	Percent Error
Kepler-29b	10.327	10.337	0.10%
Kepler-29c	13.247	13.291	0.33%

of $459.532/34.62 = 13.274$ days.

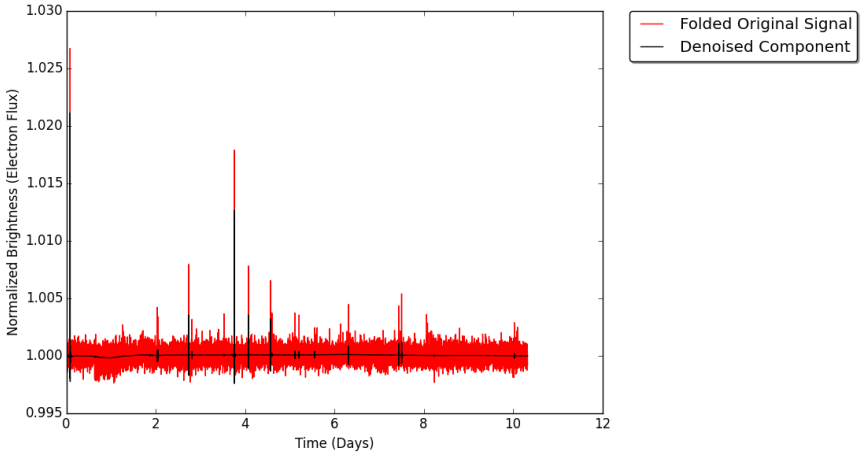


Figure 63. KIC 10358759 - Component One Isolated and Denoised.

Folding the revised flux plot over the calculated period for component two and denoising resulted in the plot shown in Figure 65. It is difficult to determine where in Figures 63 or 65 where the dips indicating a transit occur. This is due to the influence of the added noise in the signal which would have, most likely, been removed by NASA's fourth filtering stage. It should be noted, however, that the calculated periods for the two components, Summarized in Table 21, were very close to those release by NASA.

The NASA released plots for each component, shown in Figure 66, show flux dips cased by the two planets to last only $3 - 4$ hours and have a resultant drop of only ≈ 0.001 from the normalized flux. This makes the detection of the periodic components with the $\mathcal{W}$-PDM algorithm all the more impressive, especially considering the

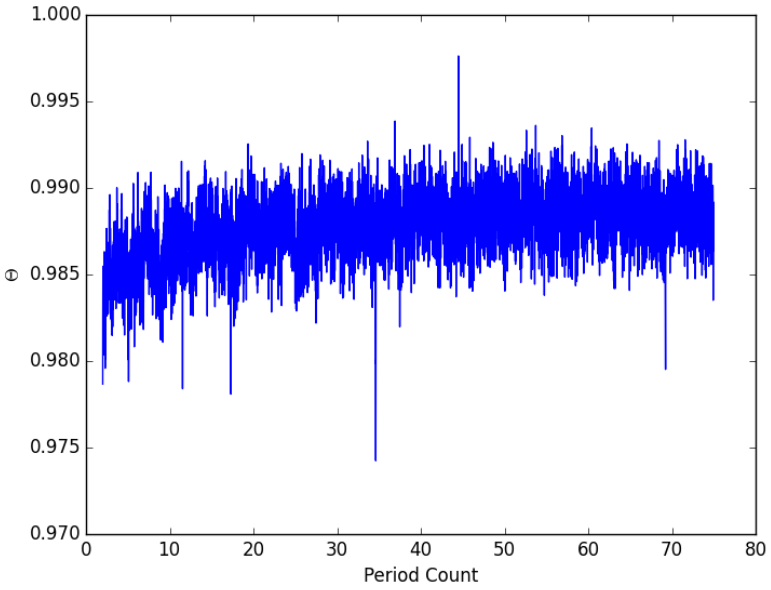


Figure 64. KIC 10358759 - Component Two Period Detection.

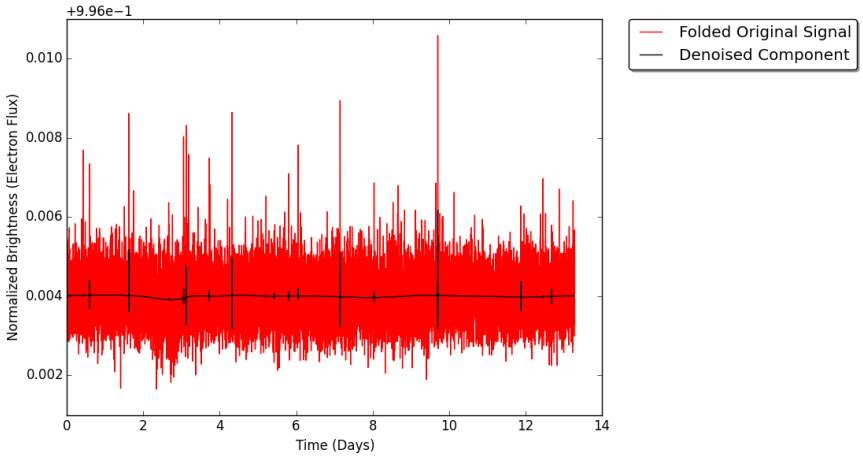
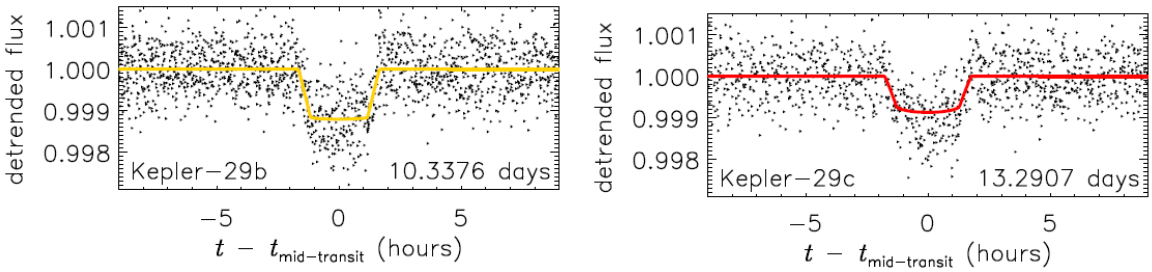


Figure 65. KIC 10358759 - Component Two Isolated and Denoised.

potential outlier events, reaching as high as 1.025, that are scattered throughout the signal.



(a) Component 1 **(b) Component 2**

Figure 66. KIC 10358759 - Published Components [14].

Improved Component Extraction.

In an attempt to improve the component and complete signals the Order and Recursive (with one update round) techniques from Chapter VI were applied to the light curve. Using the periods found with the $\mathcal{W}$-PDM of 10.327 and 13.247 the order extraction method produced the components shown in Figures 67 and 68. Once combined, the denoised complete signal using the Order extraction method can be found in Figure 69.

Running the same analysis with the Recursive method resulted in components 1 and 2 shown in Figures 70 and 71 respectively. The plot for the complete signal resulting from the Recursive extraction method is shown in Figure 72. Due the the results in Chapter VI, the Ordered technique is expected to provide the best results for the individual components, while the best clean signal should be provided by the Recursive technique. Due to the nature of the data, however, it is impossible to determine which estimate are in fact the best since there is no truth information.

Though the components extracted using the three different methods in this section were such that the dips caused by the transiting exoplanets were difficult to detect,

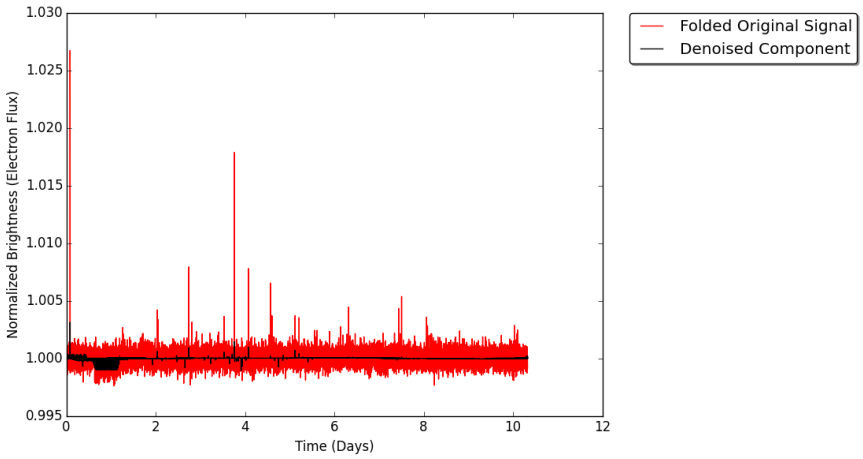


Figure 67. KIC 10358759 - Component One Isolated and Denoised - Order Method.

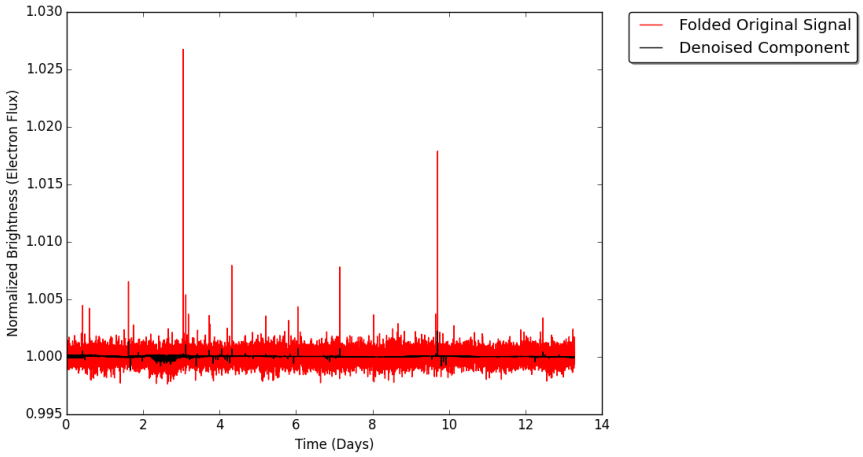


Figure 68. KIC 10358759 - Component Two Isolated and Denoised - Order Method.

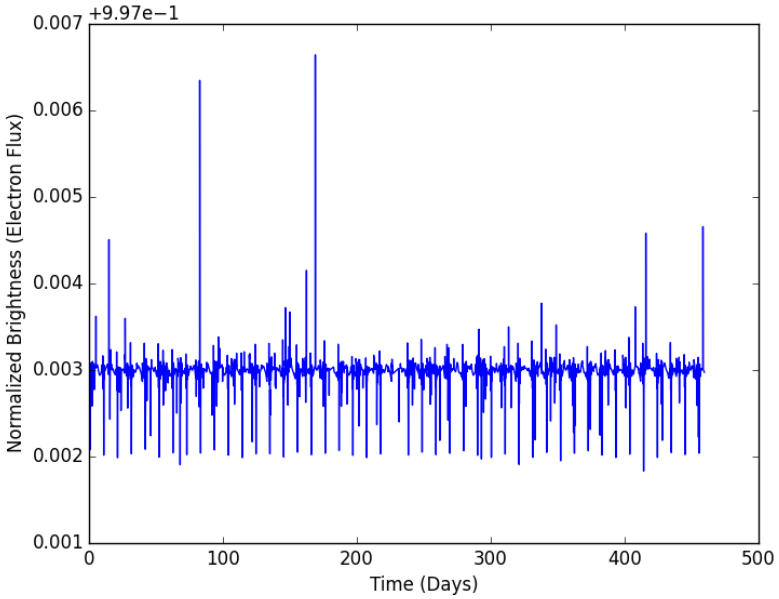


Figure 69. KIC 10358759 - Clean Signal - Order Method.

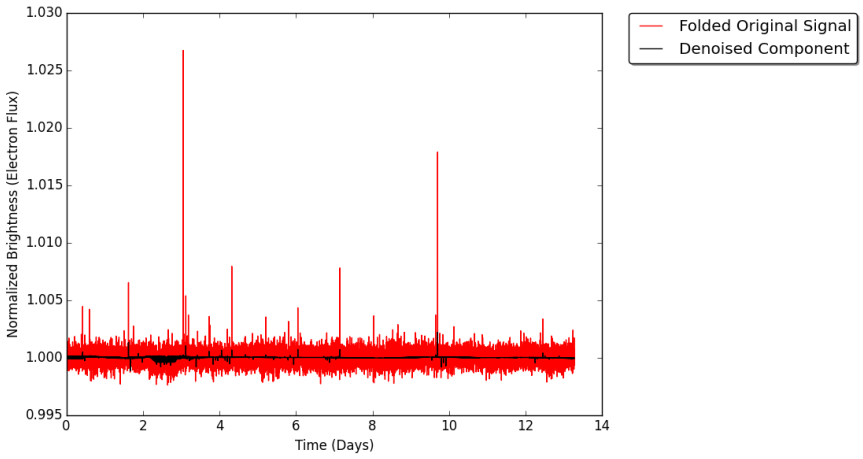


Figure 70. KIC 10358759 - Component One Isolated and Denoised - Recursive Method.

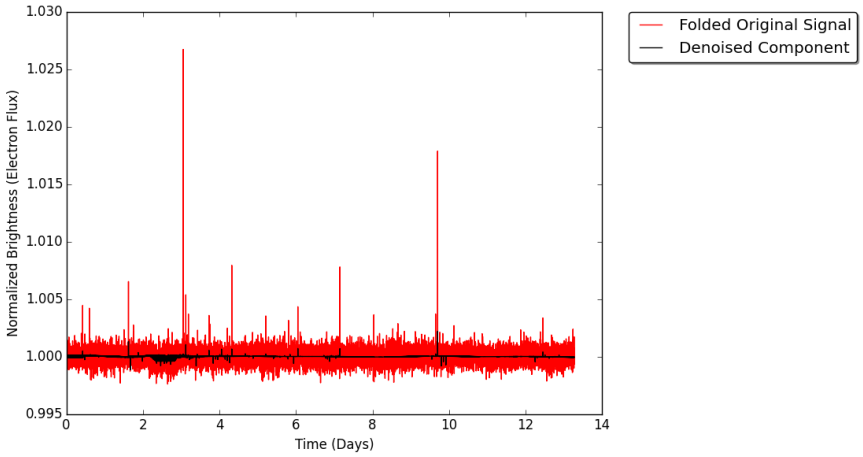


Figure 71. KIC 10358759 - Component Two Isolated and Denoised - Recursive Method.

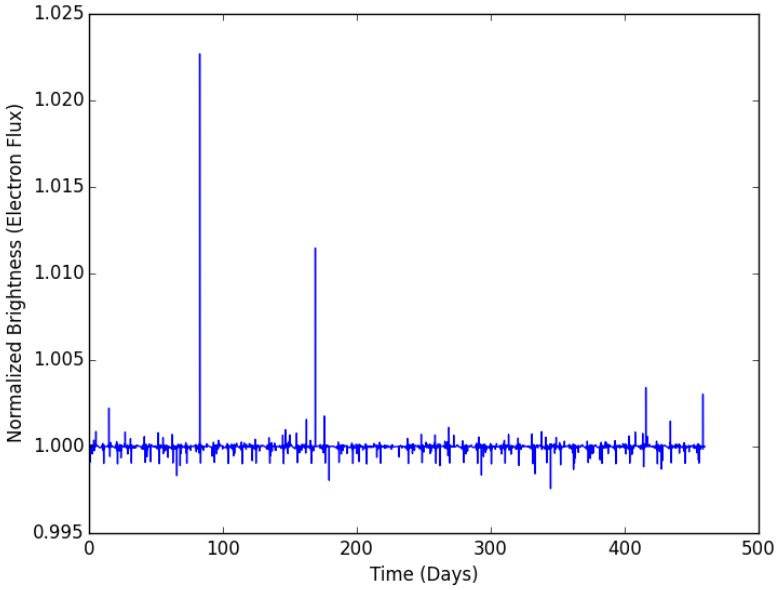


Figure 72. KIC 10358759 - Clean Signal - Recursive Method.

this section still shows the viability of the various approaches to the analysis of real light curves. The $\mathcal{W}$-PDM once again performed a near flawless detection of the component periods, even while overcoming the large spikes in flux that have a high chance of being erroneous. With the use of a more filtered light curve, it is expected that the $\mathcal{W}$-PDM method and the two improved extraction techniques would be even more effective than demonstrated here.

IX. Conclusion

The goal of this research was to combine two powerful analysis techniques, phase folding and wavelet denoising, in order to develop a suite a tools for the automated processing of vast stores of light curve data. To accomplish this aim, the research was broken down into four primary objectives. The first objective consisted of developing a mathematical framework in which to discuss the phase folding process and to use this new framework to demonstrate, both mathematically and empirically, the benefits of combining phase folding and wavelet denoising. To accomplish this objective, a mathematical proof was presented demonstrating the benefits of these two techniques operating synergistically. This proof was accompanied by a rigorous exploration of various factors which may have had an adverse effect on the proposed approach. Where it was found that some factors, such as sample counts not being powers of two, could have an adverse effect, that the combination of phase folding and wavelet denoising were still able to outperform the standard approach.

Objective two was concerned with the development of an automated method of signal decomposition. Four different methods were proposed to accomplish this objective. The first method, iterative denoising, consisted of folding a signal over each of its component periods in order to extract them using a wavelet denoising technique. This method proved to be extremely effective and was even able to improve upon overall signal quality. While developing and testing the iterative denoising method two factors, the order of extraction and the component variabilities, were identified as potential tools for an improved technique. Taking advantage of these factors resulted in the development three additional component extraction techniques, the Ordered, Amplitude, and Recursive techniques. Two of these techniques, Ordered and Recursive, proved to be even better at component extraction than the already impressive iterative denoising method.

The development of these component extraction techniques was based on the assumption of accurate period estimates for the different components. While several techniques existed for the detection of such periods, they were all based on some assumption as to the nature of the signal shape. Objective three, therefore, became the development of a completely shape agnostic approach to period detection. Based heavily on the Phase Dispersion Minimization (PDM) technique, the Wavelet PDM ($\mathcal{W}$-PDM) algorithm was able to more accurately determine the periods of component signals than the industry standards.

The final objective of this research was to utilize the complete suite of tools to analyze real world light curve data. Three different light curves were selected, two of which represented eclipsing binary (EB) systems and the third a solitary star with two transiting exoplanets. In all cases the $\mathcal{W}$-PDM algorithm was able to determine the periods of the components to within 1% of the values released by NASA. For the two EB systems, the extracted components matched almost exactly with those provided by NASA. For the two exoplanet system, error, unfortunately, was able to mask the component shapes.

These results demonstrate conclusively the synergistic effects of combining phase folding and wavelet denoising for the purposes of periodic signal analysis, especially in the case of light curves. However, the field is still ripe for further research, specifically in two areas. The first is in the development of similar suite of tools which take advantage of the developments in second generation wavelet which are built for unevenly sampled data and promise improvements in computational time. The second area for future work is in a more refined method of exploring the dynamic solution space of the $\mathcal{W}$-PDM algorithm, potentially through the utilization of an evolutionary algorithm. Automated tools, such as those developed in this paper (and their successors), will prove to be of the utmost importance for the future of signal

processing. This holds especially true in the field of astrostatistics, where projects such as the Large Synoptic Survey Telescope promise to overwhelm modern methods of light curve analysis.

Bibliography

1. "Kepler Data Processing Handbook". *NASA Ames Research Center*, (April), 2011.

2. Armstrong, D. J., J. Kirk, K. W. F. Lam, J. McCormac, H. P. Osborn, J. Spake, S. Walker, D. J. A. Brown, M. H. Kristiansen, D. Pollacco, R. West, and P. J. Wheatley. "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, 2015. ISSN 13652966.

3. Baker, Joanne C., Keith Grainge, M. P. Hobson, Michael E. Jones, R. Kneissl, A. N. Lasenby, C. M. M. O'Sullivan, Guy Pooley, G. Rocha, Richard Saunders, P. F. Scott, and E. M. Waldram. "Detection of Cosmic Microwave Background Structure in a Second Field with the Cosmic Anisotropy Telescope". *Monthly Notices of the Royal Astonomical Society*, 308:1173–1178, 1999. ISSN 0035-8711. URL `http://arxiv.org/abs/astro-ph/9904415`.

4. Blumenthal, George R., S. M. Faber, Joel R. Primack, and Martin J. Rees. "Formation of Galaxies and Large-scale Structure with Cold Dark Matter". *Nature*, 311(5986):517–525, 1984. ISSN 0028-0836.

5. Borucki, William J. and The Kepler Team. "Characteristics of Kepler Planetary Candidates Based on the First Data Set: The Majority are Found to be Neptune-Size and Smaller". *The Astrophysical Journal*, 728:117–136, 2010. ISSN 0004-637X. URL `http://arxiv.org/abs/1006.2799`.

6. Daubechies, Ingrid. *Ten Lectures on Wavelets*. Society for Industrial and Applied Mathematics, Jan 1992. ISBN 978-0-89871-274-2. URL `http://epubs.siam.org/doi/book/10.1137/1.9781611970104`.

7. Debosscher, J., L. M. Sarro, C. Aerts, J. Cuypers, B. Vandenbussche, R. Garridio, and E. Solano. "Automated Supervised Classification of Variable Stars I. Methodology". *Astronomy & Astrophysics*, volume 7638. 2014.

8. Donoho, David L. and Iain M. Johnstone. "Ideal Spatial Adaption by Wavelet Shrinkage". *Biometrika*, 81(3):425–455, 1994. ISSN 0006-3444.

9. Doyle, Laurance R, Joshua A. Carter, Daniel C. Fabrycky, Robert W. Slawson, Steve B. Howell, Joshua N. Winn, Jerome A. Orosz, Andrej Prsa, William F. Welsh, Samuel N. Quinn, David Latham, Guillermo Torres, Lars A. Buchhave, Geoffrey W. Marcy, Jonathan J. Fortney, Avi Shporer, Eric B. Ford, Jack J. Lissauer, Darin Ragozzine, Michael Rucker, Natalie Batalha, Jon M. Jenkins, William J. Borucki, David Koch, Christopher K. Middour, Jennifer R. Hall, Sean McCauliff, Michael N. Fanelli, Elisa V. Quintana, Matthew J. Holman, Douglas A. Caldwell, Martin Still, Robert P. Stefanik, Warren R. Brown, Gilbert A.

Esquerdo, Sumin Tang, Gabor Furesz, John C. Geary, Perry Berlind, Michael L. Calkins, Donald R. Short, Jason H. Steffen, Dimitar Sasselov, Edward W. Dunham, William D. Cochran, Alan Boss, Michael R. Haas, Derek Buzasi, and Debra Fischer. "Kepler-16: A Transiting Circumbinary Planet." *Science (New York, N.Y.)*, 333(6049):1602–1606, 2011. ISSN 0036-8075.

10. EB Working Group. "Kepler Eclipsing Binary Catalog - Third Revision", 2016. URL `http://keplerebs.villanova.edu/`.

11. Einstein, Albert. "Die grundlage der allgemeinen relativitatstheorie". *Annalen der Physik*, 49(4):769–822, 1916. ISSN 00033804. URL `http://www3.interscience.wiley.com/journal/112483878/abstract`.

12. Einstein, Albert. "Cosmological Considerations in the General Theory of Relativity". *Sitzungsber.Preuss.Akad.Wiss.Berlin (Math.Phys.)*, (142-152), 1917.

13. Eisenstein, Daniel J., David H. Weinberg, Eric Agol, Hiroaki Aihara, Carlos Allende Prieto, Scott F. Anderson, James A. Arns, Eric Aubourg, Stephen Bailey, Eduardo Balbinot, Robert Barkhouser, Timothy C. Beers, Andreas A. Berlind, Steven J. Bickerton, Dmitry Bizyaev, Michael R. Blanton, John J. Bochanski, Adam S. Bolton, Casey T. Bosman, Jo Bovy, Howard J. Brewington, W. N. Brandt, Ben Breslauer, J. Brinkmann, Peter J. Brown, Joel R. Brownstein, Dan Burger, Nicolas G. Busca, Heather Campbell, Phillip A. Cargile, William C. Carithers, Joleen K. Carlberg, Michael A. Carr, Yanmei Chen, Cristina Chiappini, Johan Comparat, Natalia Connolly, Marina Cortes, Rupert A. C. Croft, Luiz N. da Costa, Katia Cunha, James R. A. Davenport, Kyle Dawson, Nathan De Lee, de Gustavo F. Porto Mello, Fernando de Simoni, Janice Dean, Saurav Dhital, Anne Ealet, Garrett L. Ebelke, Edward M. Edmondson, Jacob M. Eiting, Stephanie Escoffier, Massimiliano Esposito, Michael L. Evans, Xiaohui Fan, Bruno Femenia Castella, Leticia Dutra Ferreira, Greg Fitzgerald, Scott W. Fleming, Andreu Font-Ribera, Eric B. Ford, Peter M. Frinchaboy, Ana Elia Garcia Perez, B. Scott Gaudi, Jian Ge, Luan Ghezzi, Bruce A. Gillespie, G. Gilmore, Leo Girardi, J. Richard Gott, Andrew Gould, Eva K. Grebel, James E. Gunn, Jean-Christophe Hamilton, Paul Harding, David W. Harris, Suzanne L. Hawley, Frederick R. Hearty, Jonay I. Gonzalez Hernandez, Shirley Ho, David W. Hogg, Jon A. Holtzman, Klaus Honscheid, Naohisa Inada, Inese I. Ivans, Linhua Jiang, Peng Jiang, Jennifer A. Johnson, Cathy Jordan, Wendell P. Jordan, Guinevere Kauffmann, Eyal Kazin, David Kirkby, Mark A. Klaene, Jean-Paul Kneib, G. R. Knapp, C. S. Kochanek, Lars Koesterke, Juna A. Kollmeier, Richard G. Kron, Dustin Lang, James E. Lawler, Jean-Marc Le Goff, Brian L. Lee, Young Sun Lee, Jarron M. Leisenring, Yen-Ting Lin, Jian Liu, Daniel C. Long, Craig P. Loomis, Sara Lucatello, Britt Lundgren, Robert H. Lupton, Bo Ma, Zhibo Ma, Nicholas MacDonald, Claude Mack, Suvrath Mahadevan, Marcio A. G. Maia, Elena Malanushenko, Viktor Malanushenko, Steven R. Majewski, Martin Makler, Rachel Mandelbaum, Claudia Maraston, Daniel Margala,

Paul Maseman, Karen L. Masters, Cameron K. McBride, Patrick McDonald, Ian D. McGreer, Richard G. McMahon, Olga Mena Requejo, Brice Menard, Jordi Miralda-Escude, Heather L. Morrison, Fergal Mullally, Demitri Muna, Hitoshi Murayama, Adam D. Myers, Tracy Naugle, Angelo Fausti Neto, Duy Cuong Nguyen, Robert C. Nichol, David L. Nidever, Robert W. O'Connell, Ricardo L. C. Ogando, Matthew D. Olmstead, Daniel J. Oravetz, Nikhil Padmanabhan, Martin Paegert, Nathalie Palanque-Delabrouille, Kaike Pan, Parul Pandey, John K. Parejko, Isabelle Paris, Paulo Pellegrini, Joshua Pepper, Will J. Percival, Patrick Petitjean, Robert Pfaffenberger, Janine Pforr, Stefanie Phleps, Christophe Pichon, Matthew M. Pieri, Francisco Prada, Adrian M. Price-Whelan, M. Jordan Raddick, Beatriz H. F. Ramos, Celine Ryle, I. Neill Reid, James Rich, Gordon T. Richards, George H. Rieke, Marcia J. Rieke, Hans-Walter Rix, Annie C. Robin, Helio J. Rocha-Pinto, Constance M. Rockosi, Natalie A. Roe, Emmanuel Rollinde, Ashley J. Ross, Nicholas P. Ross, Bruno Rossetto, Ariel G. Sanchez, Basilio Santiago, Conor Sayres, Ricardo Schiavon, David J. Schlegel, Katharine J. Schlesinger, Sarah J. Schmidt, Donald P. Schneider, Kris Sellgren, Alaina Shelden, Erin Sheldon, Matthew Shetrone, Yiping Shu, John D. Silverman, Jennifer Simmerer, Audrey E. Simmons, Thirupathi Sivarani, M. F. Skrutskie, Anze Slosar, Stephen Smee, Verne V. Smith, Stephanie A. Snedden, Keivan G. Stassun, Oliver Steele, Matthias Steinmetz, Mark H. Stockett, Todd Stollberg, Michael A. Strauss, Masayuki Tanaka, Aniruddha R. Thakar, Daniel Thomas, Jeremy L. Tinker, Benjamin M. Tofflemire, Rita Tojeiro, Christy A. Tremonti, Mariana Vargas Magana, Licia Verde, Nicole P. Vogt, David A. Wake, Xiaoke Wan, Ji Wang, Benjamin A. Weaver, Martin White, Simon D. M. White, John C. Wilson, John P. Wisniewski, W. Michael Wood-Vasey, Brian Yanny, Naoki Yasuda, Christophe Yeche, Donald G. York, Erick Young, Gail Zasowski, Idit Zehavi, and Bo Zhao. "SDSS-III: Massive Spectroscopic Surveys of the Distant Universe, the Milky Way Galaxy, and Extra-Solar Planetary Systems". *National Science Foundation*, 1–67, 2011. ISSN 0004-6256. URL `http://arxiv.org/abs/1101.1529`.

14. Fabrycky, Daniel C., Eric B. Ford, Jason H. Steffen, Jason F. Rowe, Joshua A. Carter, Althea V. Moorhead, Natalie M. Batalha, William J. Borucki, Steve Bryson, Lars A. Buchhave, Jessie L. Christiansen, David R. Ciardi, William D. Cochran, Michael Endl, Michael N. Fanelli, Debra Fischer, Francois Fressin, John Geary, Michael R. Haas, Jennifer R. Hall, Matthew J. Holman, Jon M. Jenkins, David G. Koch, David W. Latham, Jie Li, Jack J. Lissauer, Philip Lucas, Geoffrey W. Marcy, Tsevi Mazeh, Sean McCauliff, Samuel Quinn, Darin Ragozzine, Dimitar Sasselov, and Avi Shporer. "Transit Timing Observations from Kepler. IV. Confirmation of Four Multiple-Planet Systems by Simple Physical Models". *The Astrophysical Journal*, 750(2):114, 2012. ISSN 0004-637X.

15. Famaey, Benoit and Stacy McGaugh. "Challenges for Lambda-CDM and MOND". *High Energy Physics - Phenomenology*, 437(1):14, 2013. URL `http://arxiv.org/abs/1301.0623`.

16. Friedmann, A. "On the Curvature of Space". *General Relativity and Gravitation*, 31(12):1991–2000, 1999. ISSN 1434-6001. URL `http://adsabs.harvard.edu/abs/1922ZPhy...10..377F`.

17. Friedmann, A. "On the Possibility of a World with Constan t Negative Curvature of Space". *General Relativity and Gravitation*, 31(12):2001–2008, 1999.

18. Gabor, Dennis. "Theory of Communication". *Journal of the Institution of Electrical Engineers*, 93(26):429–441, 1946. ISSN 09252312. URL `citeulike-article-id:4452465`.

19. Guth, Alan H. "Inflationary universe: A possible solution to the horizon and flatness problems". *Physical Review D*, 23(2):347–356, 1981. ISSN 05562821.

20. Haar, Alfred. "Zur Theorie der orthogonalen Funktionensysteme". *Mathematische Annalen*, 71(1):38–53, 1911. ISSN 00255831.

21. Hoyle, F., G. Burbidge, and J. V. Narlikar. "A quasi-steady state cosmological model with creation of matter". *Astrophysical Journal*, 410(2):437–457, 1993. ISSN 0004-637X.

22. Hubble, E. "a Relation Between Distance and Radial Velocity Among Extra-Galactic Nebulae." *Proceedings of the National Academy of Sciences of the United States of America*, 15(3):168–173, 1929. ISSN 0027-8424.

23. Irwin, John B. "Tables Facilitating the Least-Squares Solution of an Eclipsing Binary Light-Curve". *The Astrophysical Journal*, 106:380, 1947. ISSN 0004-637X.

24. Ivezic, Z., J. A. Tyson, E. Acosta, R. Allsman, S. F. Anderson, J. Andrew, R. Angel, T. Axelrod, J. D. Barr, A. C. Becker, J. Becla, C. Beldica, R. D. Blandford, J. S. Bloom, K. Borne, W. N. Brandt, M. E. Brown, J. S. Bullock, D. L. Burke, S. Chandrasekharan, S. Chesley, C. F. Claver, A. Connolly, K. H. Cook, A. Cooray, K. R. Covey, C. Cribbs, R. Cutri, G. Daues, F. Delgado, H. Ferguson, E. Gawiser, J. C. Geary, P. Gee, M. Geha, R. R. Gibson, D. K. Gilmore, W. J. Gressler, C. Hogan, M. E. Huffer, S. H. Jacoby, B. Jain, J. G. Jernigan, R. L. Jones, M. Juric, S. M. Kahn, J. S. Kalirai, J. P. Kantor, R. Kessler, D. Kirkby, L. Knox, V. L. Krabbendam, S. Krughoff, S. Kulkarni, R. Lambert, D. Levine, M. Liang, K-T. Lim, R. H. Lupton, P. Marshall, S. Marshall, M. May, M. Miller, D. J. Mills, D. G. Monet, D. R. Neill, M. Nordby, P. O'Connor, J. Oliver, S. S. Olivier, K. Olsen, R. E. Owen, J. R. Peterson, C. E. Petry, F. Pierfederici, S. Pietrowicz, R. Pike, P. A. Pinto, R. Plante, V. Radeka, A. Rasmussen, S. T. Ridgway, W. Rosing, A. Saha, T. L. Schalk, R. H. Schindler, D. P. Schneider, G. Schumacher, J. Sebag, L. G. Seppala, I. Shipsey, N. Silvestri, J. A. Smith, R. C. Smith, M. A. Strauss, C. W. Stubbs, D. Sweeney, A. Szalay, J. J. Thaler, D. Vanden Berk, L. Walkowicz, M. Warner, B. Willman, D. Wittman,

S. C. Wolff, W. M. Wood-Vasey, P. Yoachim, H. Zhan, and the Lsst Collaboration. "LSST: from Science Drivers to Reference Design and Anticipated Data Products". *Bulletin of the American Astronomical Society*, volume 41, 34. 2008. ISBN 9781479980017. URL http://arxiv.org/abs/0805.2366.

25. Jenkins, Jon M., Douglas A. Caldwell, Hema Chandrasekaran, Joseph D. Twicken, Stephen T. Bryson, Elisa V. Quintana, Bruce D. Clarke, Jie Li, Christopher Allen, Peter Tenenbaum, Hayley Wu, Todd C. Klaus, Christopher K. Middour, Miles T. Cote, Sean McCauliff, Forrest R. Girouard, Jay P. Gunter, Bill Wohler, Jeneen Sommers, Jennifer R. Hall, Kamal Uddin, Michael S. Wu, Paresh A. Bhavsar, Jeffrey Van Cleve, David L. Pletcher, Jessie A. Dotson, Michael R. Haas, Ronald L. Gilliland, David G. Koch, and William J.Borucki. "Overview of the Kepler Science Processing Pipeline". *The Astrophysical Journal Letters*, 713(2):8, 2010. ISSN 2041-8205. URL http://arxiv.org/abs/1001.0258.

26. Kallrath, J. and E. F. Milone. *Eclipsing Binary Stars: Modeling and Analysis*. Springer Science and Business Media, 2009. ISBN 9781441906991. URL http://books.google.com.au/books?id=CrXBnZFdjXgC.

27. Karthikeyan, P., M. Murugappan, and S. Yaacob. "ECG signal denoising using wavelet thresholding techniques in human stress assessment". *International Journal on Electrical Engineering and Informatics*, 4(2):306–319, 2012. ISSN 20856830.

28. Kinemuchi, K., M. Fanelli, J. Pepper, M. Still, and Steve B. Howell. "Demystifying Kepler Data: A Primer for Systematic Artifact Mitigation". *Publications of the Astronomical Society of the Pacific*, 124(919):963–984, 2012. ISSN 00046280.

29. Kovács, G., S. Zucker, and T. Mazeh. "A box-fitting algorithm in the search for periodic transits". *Astronomy & Astrophysics*, 391(1):369–377, aug 2002. ISSN 0004-6361. URL http://www.aanda.org/10.1051/0004-6361:20020802.

30. Lafler, J. and T. D. Kinman. "An RR Lyrae Star Survey with Ihe Lick 20-INCH Astrograph II. The Calculation of RR Lyrae Periods by Electronic Computer." *The Astrophysical Journal Supplement Series*, 11(173):216, Jun 1965. ISSN 0067-0049. URL http://adsabs.harvard.edu/doi/10.1086/190116.

31. Lematre, A. G. "A Homogeneous Universe of Constant Mass and Increasing Radius accounting for the Radial Velocity of Extra-galactic Nebulae". *Monthly Notices of the Royal Astronomical Society*, 91(5):483–490, 1931. ISSN 0035-8711. URL http://mnras.oxfordjournals.org/cgi/doi/10.1093/mnras/91.5.483.

32. Mallat, S.A. *Wavelet Tour of Signal Processing*. Academic Press, 3 edition, 2008. ISBN 9788578110796.

33. Marron, J. S., S. Adak, I. M. Johnstone, M. H. Neumann, and P. Patil. "Exact Risk Analysis of Wavelet Regression". *Journal of Computational and Graphical Statistics*, 7(3):278, 1998. ISSN 10618600. URL http://www.jstor.org/stable/1390705?origin=crossref.

34. Matijevic, Gal, Andrej Prsa, Jerome A. Orosz, William F. Welsh, Steven Bloemen, and Thomas Barclay. "Kepler Eclipsing Binary Stars. III. Classification of Kepler Eclipsing Binary Light Curves with Locally Linear Embedding". *The Astronomical Journal*, 143(5), 2012. URL http://arxiv.org/abs/1204.2113http://dx.doi.org/10.1088/0004-6256/143/5/123.

35. Mcquillan, A., S. Aigrain, and S. Roberts. "Astrophysics Statistics of stellar variability from Kepler I . Revisiting Quarter 1 with an astrophysically robust systematics correction". *Astronomy & Astrophysics*, 137(2010):1–13, 2012.

36. Mikulski Archive for Space Telescopes. "Kepler Eclipsing Binaries (Revision 3)", 2015. URL https://archive.stsci.edu/kepler/eclipsing{_}binaries.html.

37. National Aeronautics and Space Administration. "NASA Exoplanet Archive Algorithm Documentation". URL http://exoplanetarchive.ipac.caltech.edu/applications/Periodogram/docs/Algorithms.html.

38. National Aeronautics and Space Administration. "Kepler Discoveries", 2015. URL http://kepler.nasa.gov/Mission/discoveries/.

39. National Aeronautics and Space Administration and Jet Propulsion Laboratory. "JPL Small-Body Database Search Engine", 2015. URL http://ssd.jpl.nasa.gov/sbdb{_}query.cgi.

40. Parvizi, Mahmoud, Martin Paegert, and Keivan G. Stassun. "The EB Factory Project. II. Validation with the Kepler Field in Preparation for K2 and TESS". *The Astronomical Journal*, 148(6):26, 2014.

41. Paturel, G., C. Petit, Ph. Prugniel, G. Theureau, J. Rousseau, M. Brouty, P. Dubois, and L. Cambrosy. "HYPERLEDA". *Astronomy and Astrophysics*, volume 412, 45–55. 2003. ISSN 0004-6361.

42. Peebles, P. J. E., P. J. E. Peebles, Bharat Ratra, and Bharat Ratra. "The Cosmological Constant and Dark Energy". *Reviews of Modern Physics*, 75(2):559–606, 2002. ISSN 0034-6861. URL http://link.aps.org/doi/10.1103/RevModPhys.75.559.

43. Penzias, A. A. and R. W. Wilson. "A Measurement of Excess Antenna Temperature at 4080 Mc/s." *The Astrophysical Journal*, 142:419, 1965. ISSN 0004-637X.

44. Penzias, A. A. and R. W. Wilson. "Measurement of the Flux Density of CAS a at 4080 Mc/s." *The Astrophysical Journal*, 142:1149, 1965. ISSN 0004-637X.

45. Plavchan, Peter, M. Jura, J. Davy Kirkpatrick, Roc M. Cutri, and S. C. Gallagher. "Near-Infrared Variability in the 2MASS Calibration Fields: A Search for Planetary Transit Candidates". *The Astrophysical Journal Supplement Series*, 175(2006):191–228, 2008. ISSN 0004-637X.

46. Prsa, A., E. F. Guinan, E. J. Devinney, M. DeGeorge, D. H. Bradstreet, J. M. Giammarco, C. R. Alcock, and S. G. Engle. "Artificial Intelligence Approach to the Determination of Physical Properties of Eclipsing Binaries. I. The EBAI Project". *The Astrophysical Journal*, 687(1):52, 2008. URL `http://arxiv.org/abs/0807.1724`.

47. Prsa, Andrej, Natalie Batalha, Robert W. Slawson, Laurance R. Doyle, William F. Welsh, Jerome A. Orosz, Sara Seager, Michael Rucker, Kimberly Mjaseth, Scott G. Engle, Kyle Conroy, Jon Jenkins, Douglas Caldwell, David Koch, and William Borucki. "Kepler Eclipsing Binary Stars. I. Catalog and Principal Characterization of 1879 Eclipsing Binaries in the First Data Release". *The Astronomical Journal*, 141(3):83, Mar 2011. ISSN 0004-6256. URL `http://stacks.iop.org/1538-3881/141/i=3/a=83?key=crossref.235ce1102f588c963700e60a41d300fa`.

48. Riess, Adam G., Alexei V. Filippenko, Peter Challis, Alejandro Clocchiattia, Alan Diercks, Peter M. Garnavich, Ron L. Gilliland, Craig J. Hogan, Saurabh Jha, Robert P. Kirshner, B. Leibundgut, M. M. Phillips, David Reiss, Brian P. Schmidt, Robert A. Schommer, R. Chris Smith, J. Spyromilio, Christopher Stubbs, Nicholas B. Suntzeff, and John Tonry. "Observational Evidence from Supernovae for an Accelerating Universe and a Cosmological Constant". *The Astronomical Journal*, 116(3):36, 1998. ISSN 00046256. URL `http://arxiv.org/abs/astro-ph/9805201`.

49. Scargle, J. D. "Studies in astronomical time series analysis. II - Statistical aspects of spectral analysis of unevenly spaced data". *The Astrophysical Journal*, 263:835–853, 1982. ISSN 0004-637X.

50. Schafer, Chad M. "A Framework for Statistical Inference in Astrophysics". *Annual Review of Statistics and Its Application*, 2(1):141–162, 2015. ISSN 2326-8298. URL `http://www.annualreviews.org/doi/abs/10.1146/annurev-statistics-022513-115538`.

51. Schwarzenberg-Czerny, A. "On the advantage of using analysis of variance for period search". *Mnras*, 241:153–165, 1989.

52. Schwarzschild, K. "Über das Gravitationsfeld eines Massenpunktes nach der Einsteinschen Theorie". *Sitzungsberichte der Königlich Preussischen Akademie der Wissenschaften*, 7:189–196, 1916. URL `http://arxiv.org/abs/gr-qc/0406090$\delimiter"026E30F$nhttp://adsabs.harvard.edu/abs/1916SPAW.....189S`.

53. Scovil, Charles. *Manual for Visual Observing*. The American Association of Variable Star Observers, 2013.

54. Skibba, Ramin A., Alison L. Coil, Alexander J. Mendez, Michael R. Blanton, Aaron D. Bray, Richard J. Cool, Daniel J. Eisenstein, Hong Guo, Takamitsu Miyaji, John Moustakas, and Guangtun Zhu. "Dark Matter Halo Models of Stellar Mass-Dependent Galaxy Clustering in PRIMUS+DEEP2 at 0.2<z<1.2". *The Astrophysical Jouran*, 802(2):16, 2015. URL `http://arxiv.org/abs/1503.00731`.

55. Slawson, Robert W., Andrej Prsa, William F. Welsh, Jerome A. Orosz, Michael Rucker, Natalie M. Batalha, Laurance R. Doyle, Scott G. Engle, Kyle Conroy, Jared Coughlin, Trevor Ames Gregg, Tara Fetherolf, Donald R. Short, Gur Windmiller, Daniel C. Fabrycky, Steve B. Howell, Jon M. Jenkins, Kamal Uddin, Fergal Mullally, Shawn E. Seader, Susan E. Thompson, Dwight T. Sanderfer, William Borucki, and David Koch. "Kepler Eclipsing Binary Stars. II. 2165 Eclipsing Binaries in the Second Data Release". *The Astronomical Journal*, 142(5):29, 2011. ISSN 0004-6256. URL `http://arxiv.org/abs/1103.1659`.

56. Stassun, Keivan G., Leslie Hebb, Mercedes Lopez-Morales, and Andrej Prsa. "Eclipsing Binary Stars as Tests of Stellar Evolutionary Models and Stellar Ages". *The Ages of Stars*, 258:161–170, 2009. ISSN 1743-9213. URL `http://arxiv.org/abs/0902.2548`.

57. Stellingwerf, R. F. "Period determination using phase dispersion minimization". *The Astrophysical Journal*, 224:953, 1978. ISSN 0004-637X.

58. Sterken, C. and C. Jaschek (editors). *Light Curves of Variable Stars*. Cambridge University Press, Cambridge, 1996. ISBN 9780511564796. URL `http://ebooks.cambridge.org/ref/id/CBO9780511564796`.

59. Stobie, R. S. "The Beat Period of Y Carinae". *Monthly Notices of the Royal Astronomical Society*, 157(2):167–170, Apr 1972. ISSN 0035-8711. URL `http://mnras.oxfordjournals.org/cgi/doi/10.1093/mnras/157.2.167`.

60. Stobie, R. S. and T. Hawarden. "The Beat Period of AX Velorum". *Monthly Notices of the Royal Astronomical Society*, 157(2):157–165, apr 1972. ISSN 0035-8711. URL `http://mnras.oxfordjournals.org/cgi/doi/10.1093/mnras/157.2.157`.

61. Stumpe, Martin C., Jeffrey C. Smith, Jeffrey E. Van Cleve, Joseph D. Twicken, Thomas S. Barclay, Michael N. Fanelli, Forrest R. Girouard, Jon M. Jenkins, Jeffery J. Kolodziejczak, Sean D. McCauliff, and Robert L.Morris. "Kepler Presearch Data Conditioning IArchitecture and Algorithms for Error Correction in Kepler Light Curves". *Publications of the Astronomical Society of the Pacific*, 124(919):985–999, 2012. ISSN 00046280.

62. Sulentic, Jack W., Ascensión del Olmo, and Paola Marziani. "Exploring Low Luminosity Quasar Diversity at z ˜ 2.5 with the Gran Telescopio Canarias". *Advances in Space Research*, 54(7):1401–1405, 2014. ISSN 02731177. URL `http://linkinghub.elsevier.com/retrieve/pii/S0273117713006091`.

63. The LSST Science Collaborations and LSST Project. *The LSST Science Book*. 2009. URL `http://arxiv.org/abs/0912.0201`.

64. Wambsganss, Joachim, Renyue Cen, and Jeremiah P. Ostriker. "Testing Cosmological Models by Gravitational Lensing: I. Method and First Applications". *The Astrophysical Journal*, 494(1):25, 1996. ISSN 0004-637X. URL `http://arxiv.org/abs/astro-ph/9610096`.

65. Watters, Kyle P., Roger W. Romani, Patrick Weltevrede, and Simon Johnston. "An Atlas For Interpreting Gamma-Ray Pulsar Light Curves". *The Astrophysical Journal*, 695(2):13, 2008. ISSN 0004-637X. URL `http://arxiv.org/abs/0812.3931`.

66. Wilson, Robert E. and Edward J. Devinney. "Realization of Accurate Close-Binary Light Curves: Application to MR Cygni". *The Astrophysical Journal*, 166:605–619, 1971. ISSN 0004-637X. URL `http://adsabs.harvard.edu/abs/1971ApJ...166..605W`.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704–0188

1. REPORT DATE *(DD–MM–YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From — To)*
15–09–2016	Doctoral Dissertation	September 2013 — September 2016

4. TITLE AND SUBTITLE

Synergistic Effects of Phase Folding and Wavelet Denoising with Applications in Light Curve analysis

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Armstrong, Andrew M., Capt, USAF

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Air Force Institute of Technology
Graduate School of Engineering and Management (AFIT/EN)
2950 Hobson Way
WPAFB OH 45433-7765

8. PERFORMING ORGANIZATION REPORT NUMBER

AFIT-ENC-DS-16-S-001

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

intentionally left blank

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT

DISTRIBUTION STATEMENT A:
APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

The growing size of cosmological data sets is causing the current human-centric approach to cosmology to become impractical. Autonomous data analysis techniques need to be developed in order to advance the field of cosmology. This research examines the benefits of combining two signal analysis techniques, namely phase folding and wavelet denoising, into a newly-developed suite of autonomous light curve analysis tools which includes aspects of component extraction and period detection. The improvements these tools provide, with respect to autonomy and signal quality, are demonstrated using both simulated and real-world light curve data. Although applied to light curve data, the suite of tools developed in this dissertation are advantageous to the processing, modeling, or extractions to any periodic signal analysis.

15. SUBJECT TERMS

Light Curves, Wavelet, Phase Folding

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Dr. Christine M. Schubert Kabban, AFIT/ENC
U	U	U	UU	161	19b. TELEPHONE NUMBER *(include area code)* (937) 255-3636, x4555; christine.schubertkabban@afit.edu