Learning Methods for Spectrum Estimation

by

Joshua Pohlkamp-Hartt

A thesis submitted to the
Department of Mathematics & Statistics
in conformity with the requirements
for the degree of Doctor of Philosophy

Queen's University
Kingston, Ontario, Canada
October 2015

Abstract

Spectrum estimation is an essential technique for analyzing time series data. A leading method in the field of spectrum estimation is the multitaper method. The multitaper method has been applied to many scientific fields and has lead to the development of new methods for detection signals and modeling periodic data. Within these methods there are open problems concerning parameter selection, signal detection rates, and signal estimation. The focus of this thesis is to address these problems by using techniques from statistical learning theory. This thesis presents three theoretical contributions for improving methods related to the multitaper spectrum estimation method (1) two hypothesis testing procedures for evaluating the choice of time-bandwidth, NW, and number of tapers, K, parameters for the multitaper method, (2) a bootstrapping procedure for improving the signal detection rates for the F-test for line components, and (3) cross-validation, boosting, and bootstrapping methods for improving the performance of the inverse Fourier transform periodic data estimation method resulting from the F-test. We additionally present two applied contributions (1) a new atrial signal extraction method for electrocardiogram data, and (2) four new methods for analyzing, modeling, and reporting on hockey game play at the Major Junior level.

Acknowledgments

I would like to thank:

- My supervisors Glen and David for their support and guidance.
- My colleagues Dave, Aaron, Carly, Charlotte, Karim, Michael, and Wes for the insightful conversations and frank debates.
- My friends Natalie, Justin and Rory for their love, affection, and patience.
- My family for their never ending confidence boosts. I love you mom, dad, and Noah.

All of your support is highly appreciated and I am humbled and grateful to know such smart people.

A great thanks to the Kingston Frontenacs for their willingness to be involved this research. Additionally, thank you to the data collectors that worked on this project, without you my ideas would not have come to realization.

Finally, thank you to Kingston and Queen's for making this last decade of learning as fun as it was.

Statement of Originality

The contents of this thesis are original except where references are explicitly given. The methods and results in Chapters 5 and 7 were collaborative work done with David Riegert.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

This thesis concerns several important areas of study within statistics and, more specifically, spectrum estimation and statistical learning theory. The literature review will cover some currently accepted methods and identify performance problems where improvements can be made within these areas. We will then address several of these problems and propose methods that resolve them. Several data-driven projects will be investigated to highlight the real-world application of these new methods. We will conclude with our final remarks on these methods and how they may progress with future research.

The use of methodology from one area of study within another field has been demonstrated to be effective in solving problems that may otherwise be difficult [9, 134]. The fields of spectrum estimation and statistical learning theory have overlap and problems have previously been resolved within each by fusing ideas from both [34, 120]. With this in mind we look at the field of spectrum estimation and attempt to solve problems that arise using statistical learning tools.

Within multitaper spectral estimation [116], three current problems are parameter

selection for spectral estimation, the detection of line components within the F-test and prediction or interpolation of data points for a time-series.

The choice of time-bandwidth (NW) and number of tapers (K) in practice is a supervised decision. This decision-making by statisticians on the basis of their knowledge of the theory and assumptions of the data can lead to highly variable selections among members of the same field. In many cases these selections are reported but their effect is not discussed. While in some situations using particular values for NW and K is justified [117], we are interested in identifying an unsupervised method for finding reasonable choices. We would also like to be able to justify a choice of parameters by identifying when a spectrum is well defined by its multitaper spectral estimate.

The F-test for line components has been used as a fundamental test in the identification of signals since the test was developed [116] [83]. While a useful tool for signal detection, the F-test does have problems with the detection of line components with moderate power [120]. Missed detection can be costly in communications systems [135] as well as provide misleading evidence in scientific studies [43]. We will look to develop a new test to address the problem of missed detection and potentially improve upon the false detection rates as well.

An area of ongoing active research in time-series analysis is how to predict or interpolate values for a data set [94]. If we operate under the assumption of stationarity, we can accurately model the periodic elements of the series using a method defined by Dr. David Thomson [117]. The estimates of the series produced are affected by the significance level we use to identify periodic components in the data. We would like to have a consistent and optimal choice of significance level and we will propose

a cross-validation-based method for making this choice. As extensions to Thomson's method, we will investigate how using boosting and bootstrapping methods can help to improve the estimates we can produce.

The introduction of new methods is important, but our aim is practicality and applicability to many areas of scientific research. To test the validity of these proposed methods, we attempt to tackle a couple of problems within differing fields of study. To highlight the practical use of our improvements in parameter selection and signal detection, we will develop a method to isolate atrial signals within electrocardiogram data. In addition, we will propose a variety of statistical tools for modeling goal production and puck possession in hockey games and estimate data using our adaptations to Thomson's method.

The rest of this thesis will be presented over seven chapters. The next chapter will be our literature review covering the topics required to understand the new material presented. Chapters $3-7$ will cover a variety of contributions to statistical methods and data analysis.

The contributions are:

1. Two methods for testing the sphericity of the residuals from the F-test. These tests are used to identify the correct choices of NW and K.

2. A bootstrapping method for improving the missed and false detection performance of the F-test.

3. Three adaptations to the inverse Fourier transform periodic data modeling method developed by Thomson. The adaptations are: a cross-validation method for selecting the correct significance level, a boosting method for improving the

modeling performance, and a bootstrapping method for providing estimates of the distribution for the modeled data.

4. An improved method for isolating atrial signals within single lead electrocardiograms.

5. The development of a data analysis program for Major Junior hockey, including four new methods for player evaluation. The four methods are: (1) a new statistic for monitoring neutral zone play, (2) a bagged modeling method for identifying the optimal line combination of players, (3) a real time data reporting tool for identifying player quality in-game, and (4) a method for modeling player contributions to game-play and modeling the residual temporal patterns.

Each contribution chapter will contain data simulations or analysis along with the proposed new methods.

Lastly we have a conclusion chapter where we discuss the merits and potential pitfalls of these new methods. We will give suggestions for further advancements on these topics as well as other possible applications.

Chapter 2

Background and Literature Review

In statistical data analysis we are usually provided with temporally correlated data [11]. The investigation of the patterns and properties of this data's time dependence can lead to a greater understanding of the processes that are truly at work [1,71,103]. The fundamental question of how time is affecting our data is one that is broadly studied and has a rich history, with many useful methods [46, 55, 60, 98].

Many of the methods in time series analysis follow from the fundamental techniques of statistics, namely hypothesis testing [132] and regression [54]. In the same way that time series follows from classical methods, we expect that more recent advancements such as bootstrapping and boosting may provide a strong framework for time series to follow in the future. To understand how this framework can develop, we will explore these methods' origins.

2.1 Hypothesis Testing

The majority of time series methods rely on signal detection [99], which at its core is a hypothesis testing problem [84]. Hypothesis testing, much like many other fields of statistics, finds its roots in the study of physical processes [109, 111]. The origins stretch back to the works of Laplace [59] on birthrates of European children, in which he offers example of a null hypothesis. At the turn of the 20^{th} century, Karl Pearson developed several testing procedures, which would be added upon by his son Egon in his work with Jerzy Neyman [73] to produce the Neyman-Pearson decision theory. Independent of the work of Neyman and Pearson, Ronald Fisher developed his hypothesis testing framework, which championed the use of p-values [32]. This emphasis was in direct opposition to the rejection region and alternative hypothesis-based approach of Neyman and Pearson. These two camps would bitterly dispute the validity of each other's methods for much of Fisher's life [61]. As hypothesis testing began to be more widely accepted, a hybrid of these two approaches emerged, partially out of confusion by the general scientific community (which Fisher had predicted [32]). This hybrid method, which is first described by Lindquist [62], uses p-values in the reporting of tests as well as significance levels and alternative hypotheses. As we have progressed through the later half of the 20^{th} century, the hybrid method has grown to be the standard practice among statisticians, but across disciplines we do not find consistency in notation [75]. For ease of reading, we present the generalized framework we plan to follow within this thesis.

1. We begin by identifying a property of a population that we are not certain about and would like to come to some conclusion on. This property may be parametric, theorizing on a parameter for an assumed distribution of the data,

or non-parametric, focusing on the values of summary statistics.

2. We follow this by defining a null hypothesis, H_0, and an alternative hypothesis, H_1. These hypotheses must be mutually exclusive, the assumed model well founded and all other assumptions validated; otherwise the performance of the test will diminish.

3. Next, we define a test statistic and its distribution under the null hypothesis.

4. We now set a significance level for our test. Depending on the sample size that we expect to collect, the choice of significance level will affect the power (the probability of correctly rejecting the null hypothesis when it is untrue) of our test.

5. With a set of data samples collected from our population, we calculate the test statistic that was defined in step 3.

6. We can now determine the probability of obtaining the sample test statistic or a more extreme value under the null hypothesis. The definition of extreme is dependent on the choice of null hypothesis. This is known as the p-value for a test.

7. Lastly, we compare the p-value to the significance level selected. If the p-value is smaller, we will reject the null hypothesis in favour of the alternative.

This process can be extended to multiple populations [52], comparisons of groups within a population [30] and many other areas [3]. As an example of a hypothesis testing procedure, we will show the steps for Welch's t-test, a test we will use in Chapter 6.

2.1.1 Welch's t-Test

We have two samples, drawn from two populations. Assuming that the populations come from independent normal distributions with differing variances, we would like to test whether the mean of the populations is the same. Then, following the steps above for defining a hypothesis test, we have the following:

1. We are concerned with the mean of two populations assumed to be from normal distributions with differing variances.

2. The null hypothesis is that the means are equal for the two populations, $H_0 : \mu_1 = \mu_2$. The alternative can either be that one population is greater than the other, $H_1 : \mu_1 < \mu_2$ (one-tailed), or that they are not equal, $H_1 : \mu_1 \neq \mu_2$ (two-tailed).

3. We now identify the test statistic as

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{N_1} + \frac{S_2^2}{N_2}}}. \tag{2.1}$$

   Under the null hypothesis we have

$$T \sim t_v,\ v = \frac{(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2})^2}{\frac{(S_1^2/n_1)^2}{n_1 - 1} + \frac{(S_2^2/n_2)^2}{n_2 - 1}} - 2. \tag{2.2}$$

4. Given a sample, we can identify the p-value by finding $P(T \geq t)$ (one-tailed) or $P(\mid T \mid \geq t)$ (two-tailed), where t is the sample value of the T statistic found in formula 2.1.

5. Lastly, we check the p-value against a set significance level and from there we are able to report whether there is evidence that the means are equal.

2.1.2 Fisher's Combined Probability Test

Another example of a hypothesis test is Fisher's combined probability test [31]. This method can be used to combine the results from several independent hypothesis tests into one test. Combining the p-values from several hypothesis tests with the same null hypothesis we are able to test if all of the tests fail to reject the common null hypothesis. We will use this test in Chapter 3. The steps are:

1. We are interested in knowing whether all of the independent tests have evidence of following the common null hypothesis.

2. This method's null hypothesis, H_0, is that all of the tests follow their common null hypothesis. The alternative hypothesis, H_1, is that one or more of the hypothesis tests within the group do not follow the null hypothesis.

3. The test statistic is

$$P = -2\sum_{i=1}^{O} ln(p_i), \tag{2.3}$$

where p_i is the p-value from the i^{th} hypothesis test. Under the null hypothesis,

$$P \sim \chi^2_{2O}. \tag{2.4}$$

4. We can now determine the test statistic from the sample p-values for each hypothesis and find the overall p-value for the hypothesis across the tests, $P(P \geq \hat{P})$ where $\hat{P}$ is the sample Fisher probability statistic.

5. Last we would check this p-value against our set significance level for evidence that we should reject the hypothesis that all of the tests follow the common null hypothesis.

2.2 Time Series Analysis

The desire to evaluate temporally correlated data has long been a constant within the fields of science [136]. The main applications for this evaluation are modeling for prediction [130] or interpolation [133], identification of signals [102] or patterns [8], determining times of change [138], and exploring potentially correlated variables [47]. Many methods have been proposed to deal with each of these areas, most dealing with time series data only in the time domain.

Key methods for time series prediction and interpolation are the fitting of auto-regressive moving average (ARMA) models to the data or the use of curve fitting [18]or basis expansions [27,117]. ARMA models are the most promenent method across most fields using time series data. Autoregressive patterns in time series data were first modeled by Udny Yule [139] and Gilbert Walker [128] in the late 1920's. With the full ARMA model being introduced by Peter Whittle in his thesis from 1951 [132]. ARMA models can be considered as parametric models of the running trend and serial correlation of the time series. Many adaptations and advancements have been discovered since Whittle's thesis, including extensions to non-stationary data [127] and allowance for greater structure than initially suggested [123]. A drawback to these methods is that model selection can be difficult to employ in an unsupervised manner and that, even when supervised, there may be little evidence to help guide you to use one method over another [21]. We are also concerned with over-fitting when using higher-order models [26].

Curve fitting and basis expansions methods also rely on supervised selection, whether it is the degree of spline used to fill a gap or the basis used. We will not go into detail here about the pros and cons of each method but direct you to read

chapter 5 of [48] for a review of the commonly used methods. Throughout this thesis we will use these base techniques where applicable to aid us in dealing with low-power temporal variability that we cannot resolve with the spectrum domain techniques we propose.

2.3 Spectrum Estimation

A key technique in time series analysis is that of spectrum estimation [86]. Spectrum estimation is the process of transforming our data set from the time domain to the frequency domain. We perform this transformation in an effort to study periodic structure that exist within our data. The periodic structure itself may be of interest, as in the field of radio communications [49] [105], or we may wish to remove this structure and look for other properties found in their residuals, a useful practice in economics [112, 122].

Many methods of spectral estimation exist, with the majority of research being done under the assumptions of Gaussian noise and stationary signals. To understand the importance of these assumptions, we need to examine the spectral representation of a stationary process x_t, $t \geq 0$. The spectral representation is,

$$x_t = \int_{-\frac{1}{2}}^{\frac{1}{2}} e^{i2\pi\nu t} dX(\nu) \tag{2.5}$$

where $dX(\nu)$ is a zero-mean orthogonal increment process and we assume that x_t is harmonizable [87]. A zero-mean orthogonal increment process is one in which the correlation of the difference between any two sets of adjacent points is zero. For a process to be harmonizable, it is required that the covariance function be representable as the integral of two members of a family of complex functions over a bimeasure on

$\mathbb{R}^2$ [90],

$$r(s,t) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g_s(u)g_t(v)d\rho(u,v), \tag{2.6}$$

where $g_i : \mathbb{R} \to \mathbb{C}$ and $\rho(u,v)$ is the bimeasure. Examining the covariance function of our process helps us to understand the assumption of stationarity. The covariance function is

$$\mathbf{E}\left\{x(t_1)x^*(t_2)\right\} = \int\int e^{i2\pi(t_1f_1 - t_2f_2)}\mathbf{E}\left\{dX(f_1)dX^*(f_2)\right\} \tag{2.7}$$

where $E\{dX(f_1)dX^*(f_2)\}$ is the Loéve spectrum as defined on page 474 in [63]. If the process is stationary, the covariance function only depends upon the time difference $t_2 - t_1$. This simplifies our covariance function to Riemann integration over the frequency domain,

$$\mathbf{E}\left\{x(t_1)x^*(t_2)\right\} = \int e^{i2\pi(t_1 - t_2)f}S(f)df. \tag{2.8}$$

If we have a non-stationary process, the Loéve spectrum cannot simplify to a single-frequency representation and instead relates to $\gamma(f_1, f_2)$, the dual-frequency spectrum, an L_1 integrable complex valued function. The Loéve spectrum is instead

$$\mathbf{E}\left\{dX(f_1)dX^*(f_2)\right\} = \gamma(f_1, f_2)df_1df_2, \tag{2.9}$$

where $\gamma(f_1, f_2)$ describes the frequency correlation of each f_1 and f_2 and the covariance function is instead a double Riemann integral over the frequency domain. Without stationarity, we do not have a covariance function that is related to the spectrum of a single frequency like equation 2.8. We use the relationship in equation 2.8 to estimate the spectrum for stationary processes, but for non-stationary processes the estimation is not so simple.

Along with stationarity, there is the assumption of normality. Normality is usually assumed when developing statistical tests for use within spectrum estimation. Under

this assumption we can obtain spectral estimates that are χ^2 distributed [116] when no signal is present. In the event that a signal is present, the spectrum would be non-centralized χ^2 distributed [124]. We use this property to identify signals through power estimates and the F-test for line components [116].

With these two assumptions we adopt the following approach to estimate the spectrum. The basic idea is to take a Fourier transform of the time series multiplied by a sequence of weights (a window or taper). The choice of window is where most methods differ [85]. An example of a triangular window is shown in Figure 2.1. For most methods, depending on the choice of windows, there is a trade-off between the bias and the variance of the estimate [5]. For windows with more weight near the edges, we are including more data that is farther away from the signal of interest. This inclusion of more distant frequencies introduces bias to the estimate of the spectrum at the central frequency. The converse of this is a more narrow window, which provides you with fewer data samples and creates higher levels of variability in your spectral estimate.

2.4 Multitaper Method

This trade-off is controlled when using the multitaper spectral estimation (MTM) method [116]. The MTM was introduced by David J. Thomson [116] and allows for bias control without a significant corresponding increase in variance. The MTM is similar to other methods in that it uses windowed Fourier transforms of the data series to produce spectral estimates. However, it differs through use of an orthogonal family of windows instead of a single choice. This orthogonal family consists of a group of discrete prolate spheroidal sequences (DPSS, or Slepians) [107], several of

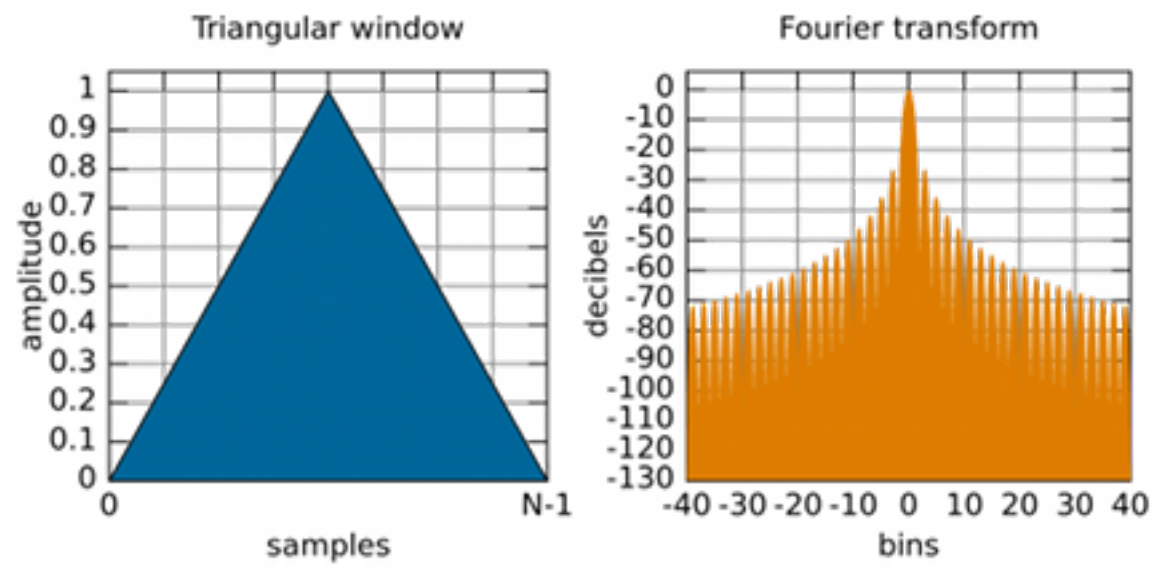

Figure 2.1: Triangular window and fast Fourier transform for 512 points of a sinusoid centred in frequency, from [76].

which are plotted in Figure 2.2. The Slepian sequences are found by solving the eigenvalue equation for a positive-definite tri-diagonal matrix [81]. The sequences are the eigenvectors from the eigendecomposition for the matrix and are ordered by the magnitude of their associated eigenvalue, from largest eigenvalue to smallest. By using an orthogonal family of functions the estimates will be computed from averaging independent sub-spectrum which lowers variance. There are many families of orthogonal functions, including the commonly used cosine functions [92]. We are motivated to use the Slepian functions because of their property that the averaged Fourier transformations of the windows maximizes the ratio of weight for in-band frequencies relative to out-of-band frequencies [107].

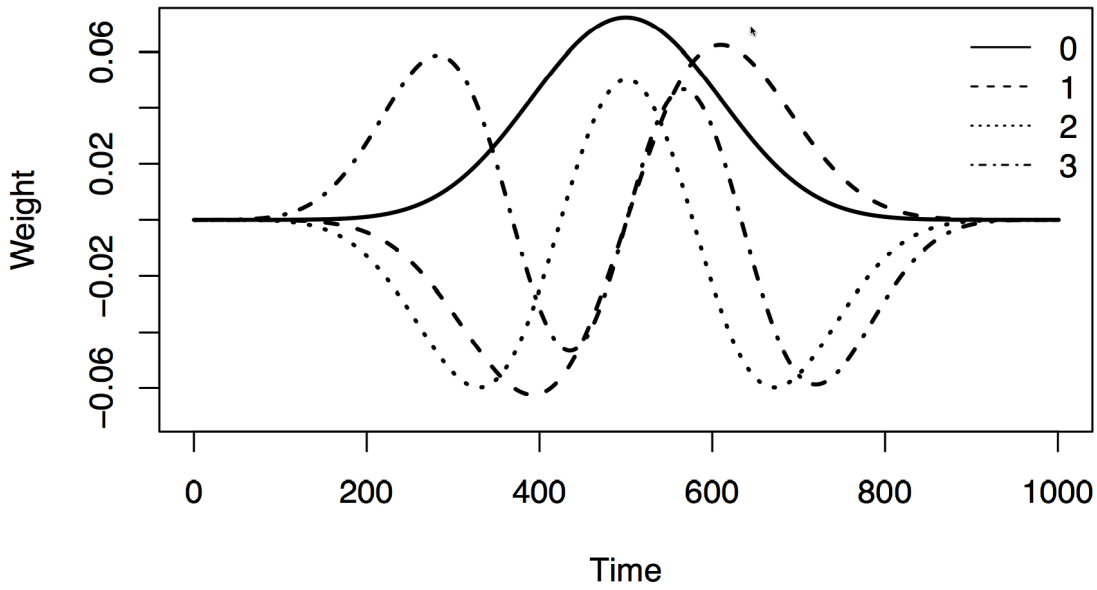

Figure 2.2: Slepian sequences of order $(0, 3)$ in the time domain with $NW = 4$, $K = 7$, $N = 1000$.

The method begins by defining a time-bandwidth product NW, with N the number of data points, and W the bandwidth parameter. Given NW, we choose to compute between NW and $2NW$ Slepian sequences of length N, $\nu_t^{(k)}$, $t = 0, ..., N-1,\ k = 0, ..., K-1$, the number of windows denoted by K. We use these as windows for K Fourier transformations, called the eigencoefficients of the data.

$$Y_k(f) = \sum_{t=0}^{N-1} \nu_t^{(k)} e^{-2i\pi ft} x_t. \tag{2.10}$$

The initial naive spectral estimate is then formed as

$$\overline{S} = \frac{1}{K} \sum_{k=0}^{K-1} |Y_k(f)|^2, \tag{2.11}$$

the average of the K eigenspectra.

The choice of NW and K are important to the shape of the spectrum and, by

extension, frequency domain−based detection methods. For smaller values of W, we get a higher-frequency resolution but increased variance in our estimates. The opposite hold for larger values. After setting W, the choice of K works as a bias-variance trade-off. For K closer to $2NW$ we get more eigenspectra, providing less variance to the estimates but higher out-of-band power, which increases the bias of the estimate. Lower values of K do no suffer as greatly with out-of-band bias but have increased variance. The choice of parameters is important to further evaluation of the data in the frequency domain. Supervised selection based on known characteristics of the data is the common practice, although this can lead to selection bias in the data analysis.

A common trick to improve the frequency granularity of a spectral estimate is a method called zero padding. In zero padding, we increase the length of our time series and, by extension, the number of frequency bins we have by adding zeros to the end of our time series. To apply this to MTM, after we have multiplied our data by the Slepian sequences, we add the desired number of zeros before taking the Fourier transforms. This will shrink the size of the frequency bins, allowing for greater resolution on the spectrum and easier detection of signals in the methods that follow.

A more advanced technique of spectral estimation using the Slepian windows is found by adaptively weighting the eigenspectra to provide superior bias control [116]. We define

$$\hat{Y}_k(f) = d_k(f)Y_k(f) \tag{2.12}$$

with

$$d_k(f) = \frac{\lambda_k^{1/2}\hat{S}(f)}{\lambda_k\hat{S}(f) + (1-\lambda_k).\sigma^2} \tag{2.13}$$

Next we solve for the spectrum estimate $\hat{S}(f)$ with

$$\hat{S}(f) = \frac{\sum_{k=0}^{K-1} |\hat{Y}_k(f)|^2}{\sum_{k=0}^{K-1} |d_k(f)|^2}, \tag{2.14}$$

where $\sigma^2 = \frac{1}{N}\sum_{n=0}^{N-1} x_n^2$ is the sample variance and λ_k is the eigenvalue associated with the kth order Slepian. We then iteratively solve for both $\hat{Y}_k(f)$ and $\hat{S}(f)$. This procedure usually converges in two or three iterations to within a few percent of the true weights when you start with the naive MTM estimate from equation 2.11 [117].

2.5 Signal Detection and the F-test

With our data now transformed into the frequency domain, we can begin to analyze the properties of the spectral estimate. The first method that is commonly applied is to graphically examine the spectral estimate for structure, whether it be spikes (peaks) or more exotic shapes such as those found in digital communications [83]. It is also possible to use our assumption of normality to test the amplitude of the spectrum for possible signals, with correction done for the noise floor and shape of the spectrum by a process called post-whitening [93]. We can use the knowledge that, in the absence of a signal, the spectrum at any frequency will follow a χ^2_{2K} distribution as a signal detection method. We know that the spectrum estimated at a given frequency will follow a χ^2_{2K} distribution because a single-windowed estimate is χ^2_2 (squaring and summing of normals), we are summing K of them in our estimate, and they are independent by orthogonality of the Slepian sequences. Examining the spectrum as realizations of a χ^2_{2K} under the null hypothesis that no signal is present allows us to determine p-values for peaks found in our estimate [124]. This is a good first test for signal detection. However, since the noise is not independent of

frequency or not easily flattened by post whitening, it may cause the data to not be easily evaluated by this type of test.

Another commonly used method of signal detection that is the F-test for line components [116]. The simplest line component is a sinusoidal signal. The F-test is computed from the eigencoefficent that are used in the multitaper method. The F-test is integral to the contributions we make in Chapters 3, 4, and 5, as well as used in the data analysis of Chapters 6 and 7.

Within the F-test we make the common assumptions of normality and stationarity as well as the important assumption that time series are made up of a collection of sinusoids in Gaussian noise. The time series is then assumed to be of the form

$$x_t = \sum_{l=1}^{M} \alpha_l \cos(\theta_l(t + \phi_t)) + \sum_{j=1}^{N} \alpha_j \sin(\theta_j(t + \phi_t)) + z_t, \tag{2.15}$$

where $\phi_t \sim U(0, 2\pi)$, represents the random phase of the signal and $z_t \sim N(0, S_N)$ is the random noise of the signal with variance S_N, the power of the background noise.

Noting that the eigenspectra of a sinusoid is an impulse (vertical rectangle in the frequency domain) of width W, our model assumption of sinusoids in noise is analogous to the eigenspectra of the time series being a sequence of impulses centred at the frequencies of the signals. Then, to determine whether there are line components within the time series, we model the set of eigenspectra at a frequency as linear functions of the Fourier-transformed Slepian sequences:

$$Y_k(f) = \mu(f)V_k(0) + r_k(f), \tag{2.16}$$

with $r_k(f) \sim CN(0, S_N)$ [117].

To detect a signal at a frequency, we test the null hypothesis $H_0 : \mu(f) = 0$. To do this we obtain estimates of the $\mu(f)$ from linear regression and then use an F-test

to determine whether there is evidence that $\mu(f)$ is non-zero. The statistic for the F-test follows an $F(2, 2K-2, p)$ distribution and is computed by:

$$F(f) = (K-1)\frac{|\hat{\mu}(f)|^2 \sum_{k=0}^{K-1} |V_k(0)|^2}{\sum_{k=0}^{K-1} |\hat{r}_k(f)|^2}, \tag{2.17}$$

$$\hat{r}_k(f) = Y_k(f) - \hat{\mu}(f)V_k(0), \tag{2.18}$$

$$\hat{\mu}(f) = \frac{\sum_{k=0}^{K-1} V_k^{\star}(0)Y_k(f)}{\sum_{k=0}^{K-1} |V_k(0)|^2}, \tag{2.19}$$

$$V_k(f) = \sum_{t=0}^{N-1} \nu_t^{(k)} e^{-2i\pi ft}, \tag{2.20}$$

where Y_k are the eigenspectra for our time series x_t, and $\nu_t^{(k)}$ are the Slepian sequences in the time domain. If the F-statistic is above $F(2, 2k-2, p)$, we will reject the null hypothesis at significance level $1-p$ and will state that a signal is present. Thomson proposed using a significance level of $1 - 1/N_f$, where N_f is the number of frequency bins used in the test. His justification for this cutoff is that under the null hypothesis we should expect to have one F-statistic at or above the cutoff across all the frequencies tested.

It is important to note that when performing the F-test, the choice in parameters NW and K can alter the resulting test statistic. These choices allow us to balance variance and bias in our spectrum estimate but can have misleading effects on the residuals from the F-test. Under correct parameter choices, the signals within our time series can be fully modeled by sinusoids and noise. We define a spectrum to be *resolved* when $\hat{r}_k(f) \sim CN(0, S_N)$ and the $\hat{r}_k(f)$ vectors are independent [117].

2.6 Bootstrap Methods

Developed in 1979 by B. Efron [24], the bootstrap method uses computational efficiency to work around problems where strict assumptions were required or where the problems were too complex for classical methods. We will use a bootstrap procedure to modify the F-test in Chapter 4.

The basic objective in the bootstrap method is to recreate the relationship between the population and the sample by considering the sample as a perfect representation of the underlying population. This is accomplished by replacing the data to generate another set of samples considered analogous to the first. The advantage gained by this is that we now work around issues of having an unknown population and make statistical inferences about the population using our samples and re-samples. As a simple example, this method can be used when we are analyzing a set of independent and identically distributed (iid) samples from an unknown joint distribution and we are interested in the mean squared error (MSE) of a location estimate obtained from the data. To have a direct estimate of the MSE, we would need to know the distribution from which our samples were taken. Even with this knowledge, the computations may not be simple or possible. To work around this, we estimate the marginal distribution of our samples. New samples are then drawn from the marginal distribution to create a new set of data. The most common method for drawing samples from the marginal distribution is to sample from the data with replacement. Other choices are possible, with another common example being a parametric bootstrap, where you suspect, the data is from a common distribution and generate new samples from it. Returning to the simple example, with new samples we can then use the multiple estimates of the location parameter to gain an understanding

of the MSE.

Another area where the framework of bootstrapping is applicable is multiple linear regression [25]. Considering the simple model: $Y(n) = X(n)B + e(n)$, where Y is an $N \times 1$ response vector, X is an $N \times p$ input matrix, e is an N length noise vector and B is an unknown vector of p parameters. These variables are not to be confused with the time series notation for $Y(f)$ and x_t although, while we can use bootstrap methods on time series data, we are only performing simple regression here. Assuming we have full rank, we follow classical regression calculations to obtain an estimate of B, $\hat{B}(n) = (X(n)X(n)^T)^{-1}X(n)Y(n)$. We now ask ourselves: how close is $\hat{B}$ to the true value of B? If we assume a model where the $X(n)$ are not random and the elements of $e(n)$ are iid with zero mean and finite variance, we can devise a bootstrapping method to obtain an estimate of how close $\hat{B}$ is to B. We estimate the $e(n)$ values, $\hat{e}(n) = Y(n) - X(n)\hat{B}$, and ensure they are centered on zero and equi-variant. If the variable being bootstrapped was not centered, we will be introducing a bias term to our parameter estimates that is dependent on the new bootstrapped values. We generally transform the $\hat{e}(n)$ terms to make them equi-variant by weighting each $\hat{e}(i)$ term by $\frac{1}{\sqrt{1-H_{ii}}}$. H_{ii} being the i^{th} diagonal element of the hat matrix ($H = X(X^TX)^{-1}X^T$). Then, drawing from the empirical distribution of $e(n)$, chosen so that we are re-sampling with replacement, we can generate new estimates for $Y(n)$, $\widetilde{Y}(n) = X(n)\hat{B} + \widetilde{e}(n)$, where $\widetilde{e}(n)$ are the re-sampled values of $e(n)$. We now compute $\widetilde{B}$ as before, $\widetilde{B} = (X(n)X(n)^T)^{-1}X(n)\widetilde{Y}(n)$. [37] showed that the distribution of $\sqrt{n}(\hat{B} - \widetilde{B})$ approximates $\sqrt{n}(B - \hat{B})$ provided n is large and variance of the inputs is well behaved. With this in mind, if we perform repeated re-samples to create a set of $\widetilde{B}$ values, we can estimate the distribution of our parameter,

B.

One issue that does present with bootstrapping in many cases is that, when testing a hypothesis, there are not theoretical distributions available under the null hypothesis. For this reason we generally have to find empirical critical region boundaries for the test statistic [25]. To do so, we repeatedly generate data under the null hypothesis and perform the bootstrapped hypothesis test. The resulting set of test statistics is an estimate of the distribution under the null hypothesis. Using the percentile associated with the significance level you expect to produce will give an estimate of the critical regions for the hypothesis test. A drawback with this test is that a large number of samples that potentially are needed to produce more extreme percentiles.

2.6.1 Bagging

Another bootstrapping method, which we will use in Chapter 3, is bagging. Bagging, was developed by Leo Brieman in 1994 to improve classification problems [12]. It is most often used on decision tree problems where we wish to avoid over-fitting and reduce variability in our estimates. It can also be used for regression problems with minor adjustments.

For bagging, we re-sample M data points from our data with replacement, much like common bootstrapping, and then analyze the M samples to make a decision. We repeat this O times, to give us a set of O decisions. Lastly we combine this set of decisions to obtain an overall decisions. The most often used method for combining decisions is to select the most common decision. This can be applied to regression in a similar fashion by modeling each subset and taking the average for the parameters from the set of O developed. Bagging gives us many of the same benefits

as other bootstrapping procedures by estimating the distribution of the decisions or parameters through the use of re-sampling [4].

2.7 Cross-validation

One the most widely used methods for determining prediction error is cross-validation [56] and one we will use in for our data synthesis method in Chapter 5. The basic premise of cross-validation is similar to that of non-parametric bootstrap methods, but instead of re-sampling the data, you are using subsets of the data to form your conclusions. For k-fold cross-validation, we randomly divide the data into k subsections. Leaving one subset out at a time, we then develop a model from the remaining $k-1$ subsets. The model we find is then used to predict the output values of the left-out portion of the data. This gives us an estimate of the prediction error for our model. We perform this by leaving out each of the k subsets and determining the prediction error of each. With this we can make decisions about parameter choices in the model development by choosing the parameter or parameters that minimize our prediction error. For example, if we were attempting to find the ideal choice of the penalization parameter in LASSO regression, T, we could use cross-validation to find the model that has the lowest estimated prediction error [125].

2.8 Gradient Boosting

Another method we will us in Chapter 5 is Gradient Boosting. Designed by Jerome Friedman in 1999 [38], gradient boosting is a model development method used most often for prediction purposes. Boosting is a general framework for reducing bias and

variance in models that are doing a poor job of describing the data. By iteratively modeling the data with increased focus on the samples the previous model have under-performed on then, adding the resulting models together, we are able to better describe the original data.

For gradient boosting, we are focused on regression problems where single models do a poor job of describing all of the relationships that exist within the data. To improve on this, Friedman proposed to iteratively model the residuals from the previous model and make a greedy (over-fitted) estimate from a linear combination of the new and old models. For a given loss function $\mathcal{L}(y, F(x))$ and set number of iterations, M, we have the following framework:

1. Fit a regression model, $F_0(x)$, minimizing for $\mathcal{L}$.
2. For each $m \in (1, M)$,
   (a) Fit a regression model, $h_m(x)$, to the residuals from the last model, $F_{m-1}(x)$, minimizing for $\mathcal{L}$.
   (b) Optimize the linear combination of the residual model and last full model that is, find $\gamma_{opt} = \underset{\gamma}{\operatorname{argmin}} \sum_{i=1}^{n} \mathcal{L}(y_i, F_{m-1}(x_i) + \gamma h_m(x_i))$.
   (c) Update the overall model, $F_m(x) = F_{m-1}(x) + \gamma_{opt} h_m(x)$.
3. The final model is F_M.

The choice of loss functions and number of iterations will alter the resulting model. It is common to used squared error. If a large number of iterations are used, we can have issues with over-fitting. Depending on the desired use for the model, we can make these choices to provide an optimal solution.

2.9 Methods For Data Analysis

2.9.1 Hierarchical Clustering

For classification problems, like we have in our atrial extraction problem in Chapter 6, clustering methods can be very useful in providing a high level of performance [40,48]. The objective of clustering methods is to group members of a sample from a population by similarities within their variables. This is often used to create classification rules for modeling a population.

A subset of these methods is hierarchical clustering methods [45]. These methods operate by either starting with all data points in one group and finding optimal ways to split the group up (called divisive or top-down) or starting with all samples as their own group and combining the groups together (called agglomerative or bottom-up). How we define distance between points for the separating/combining groups, the distance metric, and how we evaluate the distance between groups to choose the ideal set of groups, the linkage criteria, will greatly affect the results of our clustering method.

An example of a hierarchical clustering method is Ward's method [129]. With Ward's method we are performing an agglomerative hierarchical clustering by iteratively combining the two clusters with minimum squared Euclidean distance. Ward's method minimizes the within-cluster variance at each stage and the choice of stage to select, as the final clusters is not specified. Common choices for linkage criteria are minimum (single link), maximum (complete link), or the average distance between points of two clusters [45]. We would then select the stage that provided the optimum set of link criteria values across all cluster pairs (optimum is defined as giving the

minimum total or average linkage distance in most situations).

2.9.2 Quality Control Charts

Returning to our discussion of hypothesis testing from section 2.1, a direct application is found in quality control theory. When monitoring the quality of a process, we are generally concerned with the process maintaining a consistent distribution from sample to sample. First described by Walter Shewhart in 1924 while he worked for Western Electric's engineering inspection division [7], quality control charts are a major tool in controlling the quality of processes. Shewhart's control charts are essentially test statistics from sequential samples for a process plotted with significance limits. There is a temporal aspect to these charts as it is common for samples to be plotted sequentially as they are observed. Unlike most of time series statistics, the evaluation of trends within Shewhart's control charts is not commonly performed with statistically rigorous methods but rather is performed with a set of eight rules. These rules, known as Nelson Rules, were define by Lloyd Nelson in 1984 [72]. They were designed to identify changes in the distribution of a process using simple observable rules like, for example, Rule 3: A process is out of control (not following the null hypothesis) if six or more points in a row are continually increasing or decreasing. We can see in Figure 2.3 that there are three samples $(37 - 39)$ whose null hypothesis that their mean is 74.00118, at $\alpha = .01$, we would reject. These points are considered out of control in a quality sense. We also find, that under the Nelson rules, sample 40 is out of control because it is the sixth sample in a row above the target mean. Along with not being statistical in nature, a major problem with the use of Nelson rules is the increased false detection rate due to the multiple comparisons problem.

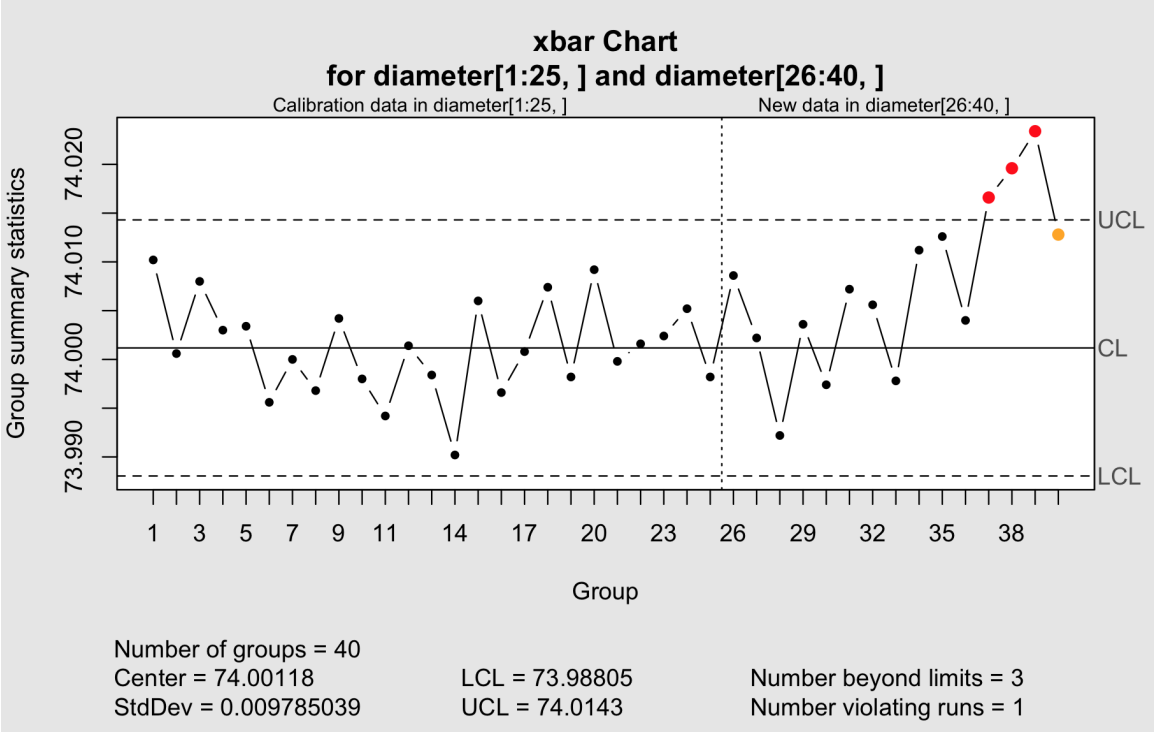

Figure 2.3: Example of a Shewhart Control Chart from the QCC package in R.

One method that has been shown to be extremely useful in practical settings, including our hockey analysis in Chapter 7, is the exponentially weighted moving average (EWMA). This model was first proposed by Roberts in 1959 [96] to help deal with small shifts in the parameters that persist over time but are not significantly different from the target or expected value. We calculate a weighted mean over all the samples collected to date with exponential weight that decreases in relation to increased distance back in time from the current sample.

We define the moving average as,

$$z_t = \lambda x_t + (1 - \lambda) z_{t-1}, 0 \leq \lambda \leq 1. \tag{2.21}$$

The null hypothesis that we will test is that each sample follows the same target mean, $\mathbb{E}(X_i) = \mu, \forall i$. Then, under the null hypothesis,

$$\mathbb{E}(z_t) = \lambda\mathbb{E}(x_t) + (1-\lambda)\mathbb{E}(z_{t-1}), \tag{2.22}$$

$$\mathbb{E}(z_t) = \lambda\sum_{i=0}^{t-1}(1-\lambda)^i\mathbb{E}(x_t) + (1-\lambda)^t\mathbb{E}(z_0), \tag{2.23}$$

with $z_0 = \mathbb{E}(x_t) = \mu$. Now subbing in for Z_0 and substituting out the series we get

$$\mathbb{E}(z_t) = (1-(1-\lambda)^t)\mu + (1-\lambda)^t\mu = \mu. \tag{2.24}$$

Similarly, we have

$$\sigma_{z_t} = \sigma\frac{\lambda}{2-\lambda}(1-(1-\lambda)^{2t}), \tag{2.25}$$

where $\sigma = VAR(x)$. From this we can easily test now whether each updated moving average follows the null hypothesis using a z-test. We can plot sequential values of z_t with the updating critical regions to monitor the process for small shifts. The variance under the null hypothesis is dependent on t giving a pinched look to the critical region near $t = 0$. As $t \to \infty$ we have $\sigma_{z_t} \to \sigma$, which gives constant bounds on the critical region for processes with a long sampling history.

Within this method we need to make decisions on the value of λ as well as the significance level for our critical region. From the context of quality control, common practical choices are $\lambda \in [.05, .2]$ and $\alpha \approx .00135$ (around 3σ). These are chosen for their ability to detect small shifts quickly while not having a large false detection rate; more discussion of this can be found in Montgomery [69]. Performing an EWMA chart with $\lambda = .2$ on the same data as in Figure 2.3, we find that in Figure 2.4 points 14 and 16 are now out of control below the limit. This was not found under the traditional Shewhart chart.

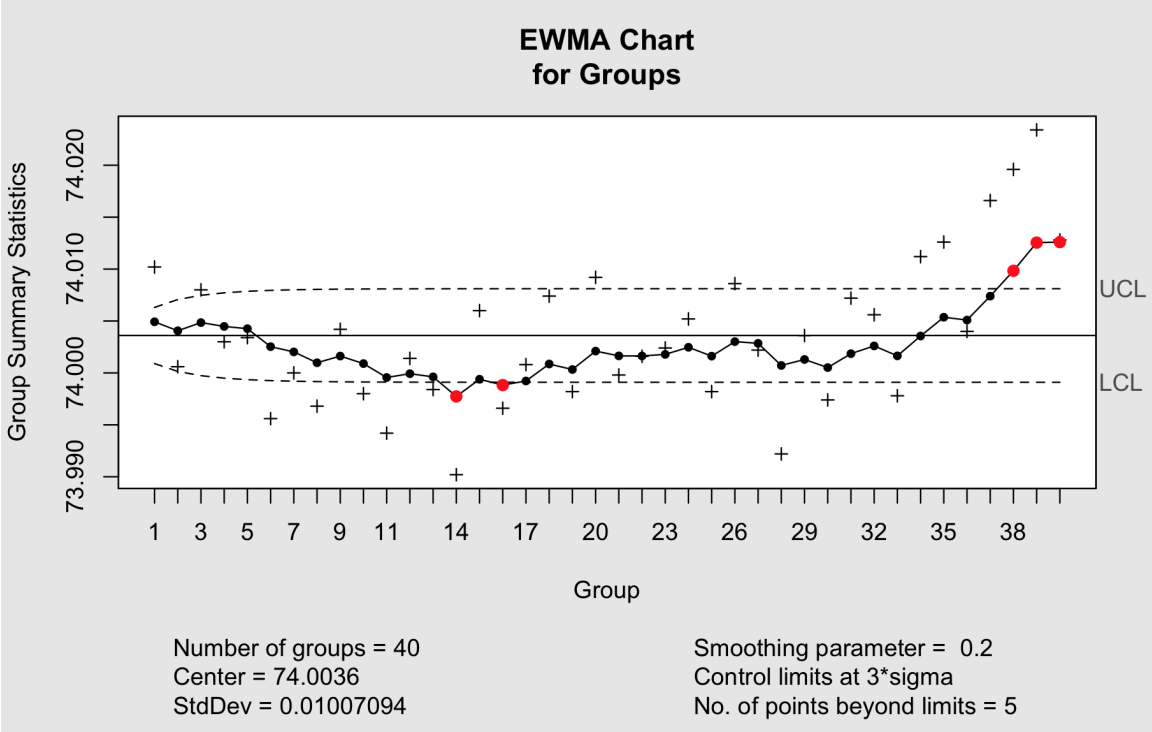

Figure 2.4: Example of an EWMA Control Chart with $\lambda = .2$ from the QCC package in R.

2.9.3 Regularized Regression

In the field of model selection for regression, there are many common methods, such as forward and backward stepwise, best subset, and forward stagewise regression, that are used regularly [68]. Each of these methods has merit and possible pitfalls. One such shortcoming is that with subset selection, your model parameters change in a discontinuous manner which often leads to high variability and the updated model does not improve upon prediction error [48]. Another issue present in many common model selection methods is that the selection is supervised, meaning that the model is chosen by the statistician under the guidance of some methodology. While

the methods are valid in the appropriate context, not all statisticians are created equal [77].

Another class of methods for use in regression model selection are known as shrinkage methods [48]. Also known as regularization, these methods provide an unsupervised and continuous choice of parameters in the regression setting. The general idea is to perform least squares regression with the L^q norm of the coefficient parameters being constrained to be less than or equal to a set maximum. The regression model is now the solution of,

$$\hat{B} = \underset{B}{\operatorname{argmin}}(\sum_{i=1}^{N}(Y_i - B_0 - \sum_{j=1}^{p} X_{n,j}B_j)), \tag{2.26}$$

subject to

$$\sum_{j=1}^{p} \mid B_j \mid^q \leq T. \tag{2.27}$$

In essence, this class of methods constrains the possible choices of parameters to a region around the origin. Several of these regions are shown in Figure 2.5 and their relation to the least squares solution is shown in Figure 2.6. The size of this region is determined by the value of T and the shape of the region is determined by q. When q is small (< 1), the region becomes concave, restricting the parameter values to being near their axis. With large q (> 1), the region of possible values is convex, allowing greater values of each parameter away from their individual axis. The two leading methods in this class are ridge ($q = 2$) and LASSO ($q = 1$) regression. Ridge regression, also known as Tikhonov-Phillips regularization, constrains the parameters to a spherical region, which is beneficial for highly correlated data, where the coefficients may be large and in opposing sign [48]. Under ridge regression, the inputs with greatest explanatory power will be given weight while allowing for considerable

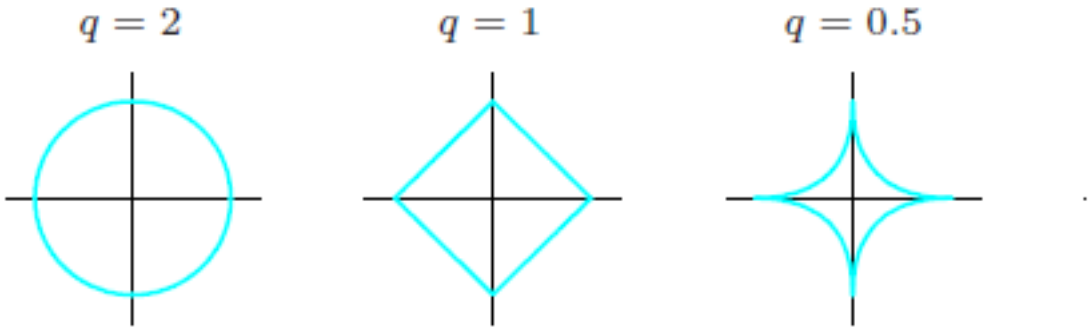

Figure 2.5: Coefficient regions for regularized regression for two dimensions, from [48].

mixture of inputs.

In contrast, the least absolute shrinkage and selection operator method (LASSO) is used not only to negate the possibility of large coefficients but also to set some to zero. The LASSO was introduced by Tibshirani in 1994 [121] and is named for the ability to shrink a parameter and remove it, there is in fact a cowboy with a lasso on the home page of Tibshirani. The region for parameters in the LASSO method is square; this allows for the possibility of the closest choice of parameters to the least squares solution to be at a corner, where one or more coefficients will be zero. This acts as soft selection, where instead of having a choice of subsets, you now have choices ranging from the full least squares solution to negligible weight for any parameter. Another beneficial aspect of these methods is that parameter selection is automatic and unsupervised for a given choice of T. It is worth noting that the choice of penalty parameter T is made by the statistician. The usual method for this choice is cross-validation and within most computer packages, this is the case. Computationally, determining the ideal value of T can be quite time-consuming for

large data sets; this can create problems for the implementation of regularization in areas like time series, where the data sets are usually quite large. We will use LASSO regression in our model development in Chapter 7.

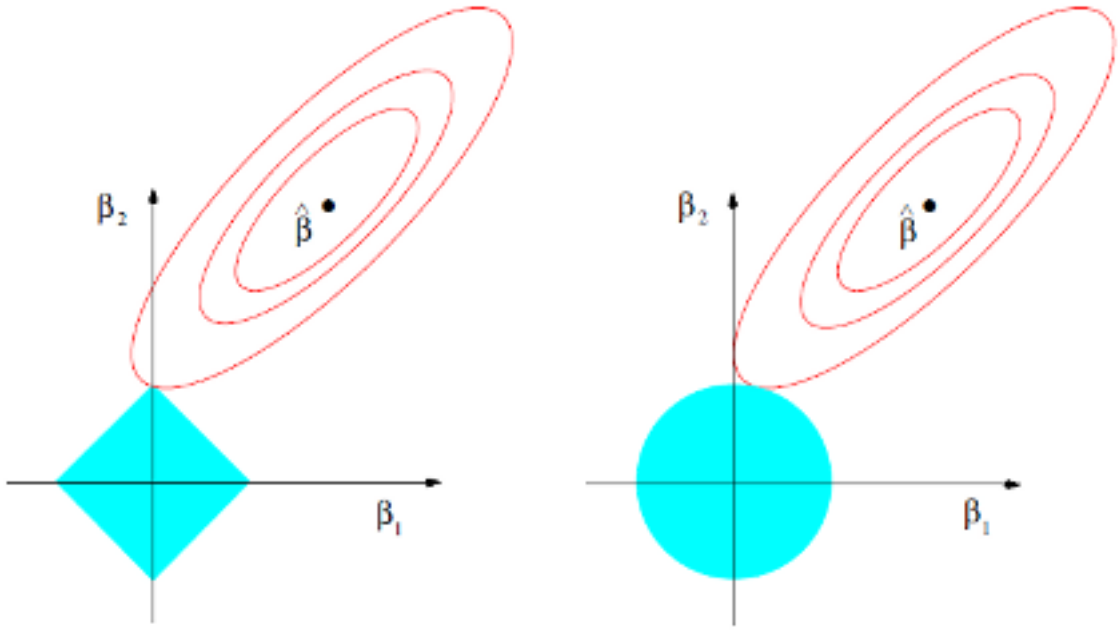

Figure 2.6: Constrained regression regions with relation to least squares estimate, from [48].

Chapter 3

Sphericity Tests for Parameter Selection

3.1 Introduction

Within the multitaper method, we are required to select parameter values for the width of the windows for the Fourier transforms known as the time-bandwidth, as well as the number of windows to be used. The choice of bandwidth (NW) and number of windows (K) is in practical settings made by a mix of iterative guessing and prior knowledge of the data set. These choices can vary drastically and the choice is not always obvious. At a recent conference session on multitaper spectral estimation[1], the number of tapers used on a range of problems varied by more than an order of magnitude. While in some situations using larger values for NW and K is justified, we are concerned with identifying from a naive mindset a sufficient

[1]The 2013 Interdisciplinary International Conference on Applied Mathematics, Modeling and Computational Science, in Waterloo, Ontario, Canada

set of parameters to start with. We would also like to be able to justify a choice of parameters by identifying when a spectrum is well resolved by its multitaper spectral estimate.

Concerning ourselves with the issue of parameter selection for MTM, the motivation is to design an unsupervised method for making an educated choice of parameters with no background knowledge of the data. Utilizing some properties of the Slepian sequences and the F-test for line components, we are able to address this issue. The use of this method relies on residuals resulting from the computation of the F-test for detection of line components.

3.2 Naive Sphericity Test

The first thing that we need is a better understanding of what the choice of NW and K are actually doing. For NW, we are in fact determining W since N (the number of samples) is set. The choice of W determines the frequency bandwidth for which the spectrum is estimated at each frequency. That is, the estimate of the spectrum at frequency f is based on the information found within the band $[f - W, f + W]$. The frequency band around f should ideally only contain a signal centred near f that is not wider than the band. There is potential when making a naive choice that we could either choose W too small and not contain all of the signal within the frequency bin or W too large and have more than one signal present within the bin. The choice of parameters is constant across the frequency range we are examining, so the choice made should be reasonable for all signals in that band.

For a set NW, the choice of K will also have to be made. The rule of thumb is "two times NW minus a couple (1 or 2)". While this choice is acceptable for many

data sets, one should not use it without thought. The choice of K determines how close to the band edge (towards $\pm W$) we want full weight within our estimation of the spectrum. At $2NW$, we have the closest estimation to a brick-wall (rectangular) filter that we can produce. While that may be ideal for some signals, as we approach $K = 2NW$, we are introducing more out-of-band bias, since the higher order Slepians have higher side-lobe (the frequencies outside of $\pm W$) power. This creates a bias-resolution trade off, where higher values of K may give a better representation of the signal, but also bias the result.

Moving back to the rule of thumb, by choosing K to be slightly less than $2NW$, we are making the choice to have full weight almost to the band edge and keep the out-of-band bias lowered. The logic behind this is that most signals are not perfectly rectangular or do not reach the band edge. Then the inclusion of the last couple tapers is only introducing bias. This assumption may be true in some cases, but we cannot be assured that this will hold for all time series we evaluate. The rule of thumb then leaves us with only a rough guide for where to start and further analysis is required.

Under correct parameter choices, we will have our line components fully described by $\hat{\mu}(f)V_k(0)$ and the residuals will follow a standard complex normal distribution. We do not expect every signal to be fully described, nor do we expect signals that are not line components or are non-stationary to be. With that in mind, we are not directly concerned with the value of the residuals at one frequency but rather the overall distribution of the residuals. We now define that the spectrum has resolved residuals if $r_k(f) \sim CN(0, \sigma^2)$. An example of well resolved residuals is shown in Figure 3.1. The test then becomes determining if the residuals follow a complex normal

distribution [117]. The test that is usually used in this case is one for sphericity. Sphericity occurs when the spread of a random variable in 2 or more dimensions is equal and uncorrelated. This occurs when the covariance matrix is diagonal. For this

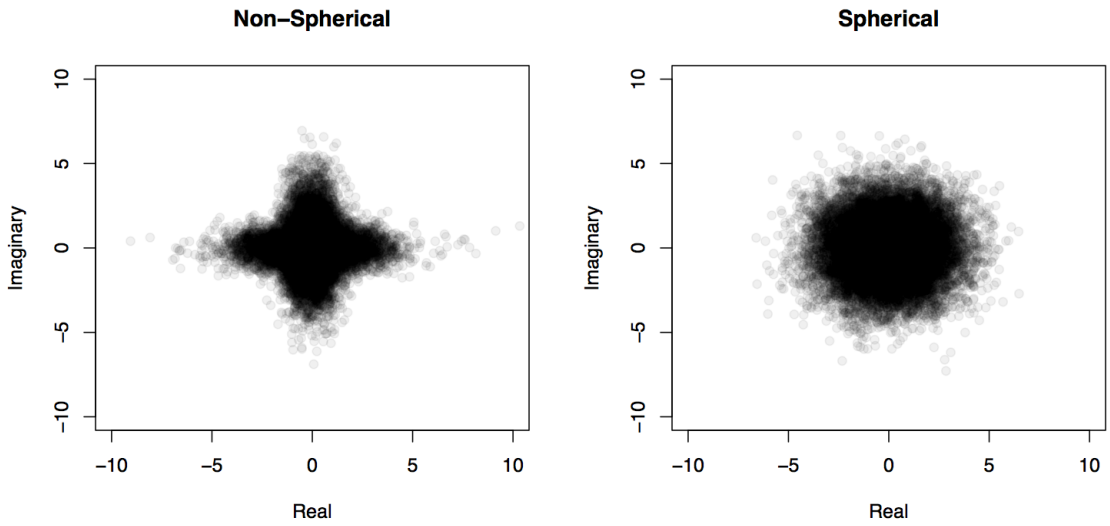

Figure 3.1: Comparison of non-spherical and spherical distributed complex-valued residuals.

test, we have K sets of residuals that are N long complex vectors. The N long vectors of frequencies cannot all be used while maintaining the assumption of independence due to the averaging in the frequency domain that is used to create our spectral estimates. Within the MTM we are averaging over the frequency range of $f \pm W$ for the spectral estimate at f. Due to this, there is a dependence between frequencies that are less than $2W$ apart from each other. We then select a subset of the N values for each complex vector of residuals that only contains frequencies that are $2W$ away from each other. We denote the length of this subset as M.

Since the hypothesis we are interested in is that all of the residuals follow a

$CN(0, \sigma^2)$ distribution, we concatenate our K subset vectors into one that is MK long and test the sphericity of all the residual terms together. We follow the test for sphericity described by S. John [53].

1. We are concerned with the covariance matrix for the real and imaginary parts of the residuals from the F-test for line components.

2. The null hypothesis is that the covariance matrix is diagonal,

$$H_0 : Cov([A, B]) = \sigma^2 \mathbb{I}. \tag{3.1}$$

We have an unknown common variance, σ^2, complex-valued sample residual from the model for the j^{th} eigenspectra at frequency f, $r_j(f) = a_j(f) + b_j(f)i$, and the real and imaginary parts in vector form,

$$A = [a_0(f_1), a_0(f_2), ..., a_0(f_M), a_1(f_1), ..., a_1(f_M), ..., a_{k-1}(f_1), ..., a_{k-1}(f_M)], \tag{3.2}$$

$$B = [b_0(f_1), b_0(f_2), ..., b_0(f_M), b_1(f_1), ..., b_1(f_M), ..., b_{k-1}(f_1), ..., b_{k-1}(f_M)]. \tag{3.3}$$

3. We identify the test statistic as

$$\Xi_N = (MK - 2)[\frac{1}{2(1 - \hat{W})^{1/2}} - 1] \tag{3.4}$$

4. Under the null hypothesis we have

$$\Xi_N \sim \mathbb{F}_{2,2MK-4} \tag{3.5}$$

with

$$\hat{W} = \frac{tr((M\hat{R})^2)}{(tr(M\hat{R}))^2}, \quad \hat{R} = Cov([A, B]). \tag{3.6}$$

5. Given a sample, we can identify the p-value by finding $P(\Xi_N \geq \xi_N)$, where ξ_N is the sample value of the Ξ_N statistic found in formula 3.5.

6. Lastly, we check the p-value against a set significance level and from there we are able to report whether there is evidence that the covariance matrix is diagonal.

In simple simulations with few spaced signals, the choice on NW and K does not greatly effect the sphericity of the residuals. This is due to the majority of frequencies only estimating noise, which will be resolved no matter the choice. This will be further explored in the simulations section to follow. That being said, with real data sets and a fair amount of computer power, one could use this test in one of two ways. First, if you have made a choice of parameters and would like to know if it is reasonable for your data, you can perform this sphericity test to see if your residuals are resolved. If we fail to reject the null hypothesis then the choice of parameters was acceptable. This test can act as a check for one's assumed choice and provide statistical evidence to support it.

The other option is that if one wants to make a naive choice of parameters for their data, they may want to run this test for a range of choices of NW and K. The literature suggests NW as small as 2 [89] and in practical settings that NW can be as large as 100 [50]. It is also noted in Slepian's original discussion of the band-limited maximized concentration properties that we should choose NW to be an integer [106]. This would allow us to select K as large as $2NW$ or the closest integer. From this, we realistically should expect $NW \in [2, 100] \cap \mathbb{Z}$. For values of K, it seems appropriate to allow a range from NW to $2NW$, possibly not allowing a range that wide as NW gets large (> 10) for computational consideration.

Performing the sphericity test for a range of NW and K then choosing the largest

p-value across all choices is a good naive method for selecting NW and K. In the event of two choices with the same p-value we selected the smaller values of NW and K. This was done to save computational costs in implementing the sphericity tests in the later chapters. This can be simplified for computational efficiency by choosing a coarse ladder of NW values and setting $K = 2NW - 1$. This method would be a good first pass before choosing a finer range to investigate more thoroughly.

This test is a good first step in choosing NW and K, while some intuition about the data and graphical examination should be used to further tune the spectral estimates. It is also important to note that if the signal is not stationary or has a considerable number of components that are not lines, this test will fail as the residuals will also deviate in distribution. Careful graphical analysis can avoid many issues with non-line components. We do not want to employ stationary spectral methods on time series with large non-stationary components, so failure in this test is welcomed in that case. We may be interested in evaluating series with minor non-stationary components and in this situation we do not expect a significant amount of deviation in the distributions of the residuals. For completeness, in time-series analysis one may want to perform a test for non-stationarity beforehand to avoid that situation all together but there are many types of non-stationarity so avoiding all of the possible pitfalls is difficult from a diagnostic standpoint.

3.3 Bagged Sphericity Test

In situations where the power (variance) of the noise in the time series, σ^2, is well known or easily estimated we can adapt the sphericity test to include this information

and provide a further refined result. We accomplish this by altering the null hypothesis to relate to a specific covariance matrix rather than the looser requirement of diagonality. In this situation, the testing procedure follows the work of Korin [57] and Anderson [3].

1. We are still concerned with the covariance matrix for the real and imaginary parts of the residuals from the F-test for line components.

2. The null hypothesis is that the covariance matrix is diagonal with a specific variance structure,

$$H_0 : Cov([A, B]) = \sigma^2 \mathbb{I}. \tag{3.7}$$

   σ^2 is the known common variance. A and B are as defined in equations 3.2 and 3.3.

3. We identify the test statistic as

$$\Xi_B = \rho(MK - 1)(\log \mid \sigma^2 \mathbb{I} \mid - \log \mid \hat{W} \mid + tr(\hat{W}(\sigma^2 \mathbb{I})^{-1}) - 2), \tag{3.8}$$

   with $\rho = 1 - \frac{15}{18(MK-1)}$.

4. Under the null hypothesis we have

$$\Xi_B \sim \chi^2_3 + \omega(\chi^2_7 - \chi^2_3) \tag{3.9}$$

   with $\omega = \frac{47}{432(MK-1)^2}$ and $\hat{W}$ defined as in the naive sphericity test.

5. Given a sample, we can identify the p-value by finding $P(\Xi_B \geq \xi_B)$, where ξ_B is the sample value of the Ξ_B statistic found in formula 3.5.

6. Lastly, we check the p-value against a set significance level and from there we are able to report whether there is evidence that the means are equal.

Unfortunately, implementing this test directly runs into two significant computational issues. First, to avoid issues with round off errors for small values when computing the determinant of the sample covariance matrix in R, we need to ensure to scale the time series so that it has an estimated background noise of $\sigma^2 > 1$. This alleviates the possibility of the computation of either determinant to be rounded to zero or negative.

The other and more fatal issue that occurs is that the test is not immune to the increasing size of available data when we increase NW and K for a fixed N. As K increases, the number of occurrences of frequencies with signals to be resolved relative to those without decreases. This will increase the sphericity of the residuals without attending to the issue of resolving the signals in the data set. In addition to this, as the amount of data used, M, increases so does the effect of the variance reduction for larger values of NW. For data sets with $M > 100$, during simulations, large values of NW had residuals that were more spherical. This effect can be attributed to averaging over more signals when widening the frequency bandwidth W used in estimating the spectrum. This is similar to over-fitting the data and does not provide accurate frequency estimates for time series prediction.

To avoid this issue of sample size and provide a choice that is geared towards use in predictive models from MTM, we use a bagging algorithm, similar to what we described in Section 2.6.1, to select constant sized and random samples of the residuals. We then perform hypothesis testing on each sample and use Fisher's method for combining p-values, as discussed in Section 2.1.2, to consider the sphericity of each sampling collectively. The procedure is adapted from equation 3.8 with the addition of a preliminary sampling stage and the use of Fisher's method at the end to provide

a single p-value. The procedure is as follows:

1) Sample the residuals with replacement O times to be used as separate data sets.

2) Test the null hypothesis, $H_0 : Cov([\hat{A}, \hat{B}]) = \sigma^2\mathbb{I}$, for all O sets of samples using the statistic from equation 3.8 and then calculate the p-value for each set.

3) Use Fisher's combined probability test to determine the overall p-value of the combination of the O sets of samples. Then, as noted in section 2.1.2, the test statistic for Fisher's Method is

$$\hat{P} = -2\sum_{i=1}^{O} ln(p_i) \sim \chi^2_{2O}. \tag{3.10}$$

From this procedure we can now obtain a statistic of how well the residuals are resolved from the choice of parameters. This test gives a good approximation of the p-value for the known noise sphericity of the residuals and by proxy how well the spectrum is resolved for a set of parameters but suffers from two potentially significant issues.

There is the potential to have runs which include only frequencies that contain no signal or only perfectly resolved signals, which will have p-values near one no matter the parameter choice. Within Fisher's method these near one p-values will cause Fisher's method to produce an insignificant result with no consideration for the other set's p-values [51]. To avoid this issue we truncating the maximum p-value for any set to $1 - 1/O$, which will alleviate this issue but will increase the false detection of the test [79].

In the situation where there is not a well informed estimate of the power of the noise in the time series the p-values given will be potentially misleading. As this testing procedure is evaluating how close the sample covariance matrix of the residuals is to $\sigma^2\mathbb{I}$, if we have an improper choice of σ^2 we are not evaluating the correct

hypothesis. This may cause us to make a decision to use a parameter set that is not in fact the optimal choice. The effects of these issues will be investigated later in this chapter.

3.4 Simulations and Comparison

In an effort to demonstrate the merits of the tests designed in this chapter, we attempted to create a situation where the signals present would have theoretical range of acceptable values and to determine if the tests could make an unsupervised choice of parameters within the range.

The objective was to determine if the two sphericity tests would choose the ideal theoretical choice of NW and K for a known time-series. To test for NW we devised a data set that is the sum of evenly spaced signals across the frequency band, $f \in (0, .5)$ in white noise. The spacing between the centres of each signal is $.013Hz$ with each signal being a combination of five sinusoids of decreasing amplitude moving away from the central frequency. This type of signal is known as a singlet in seismology.

$$X(t) = \sum_{i=1}^{38} \alpha_i \sum_{j=-2}^{2} (.3 - .1 \mid j \mid) \sin(2\pi(.013i + .002j)t) + z_t, \tag{3.11}$$

where α_i is a random amplitude for each signal that is taken from $U(.5, 1)$ and $z_t \sim N(0, 2)$. An estimate of the spectrum for three of these signals is shown in Figure 3.2. With this spacing and 1000 samples, we would ideally have a choice of NW that is $1000 \times .013/2 = 6.5$. Now a spacing larger than $NW = 7$ would cause there to be significant overlap in the signals, introducing bias. As the outer sinusoids are $.008Hz$ from each other, a choice of $NW = 1000 \times .008/2 = 4$ would be an acceptable lower bound. We choose a 5 pronged signal so that with too small a choice of NW

we would not capture the full signal. We do note that it could be possible to isolate each sinusoid in the signals by having $W < .002$. To achieve this we would need $NW < 1000 \times .002 = 2$, which past research implies is not ideal due to high variance in the spectrum [115]. We will only test for $NW \geq 2$ but do note that smaller values could provide a resolved spectrum. We computed both sphericity test for

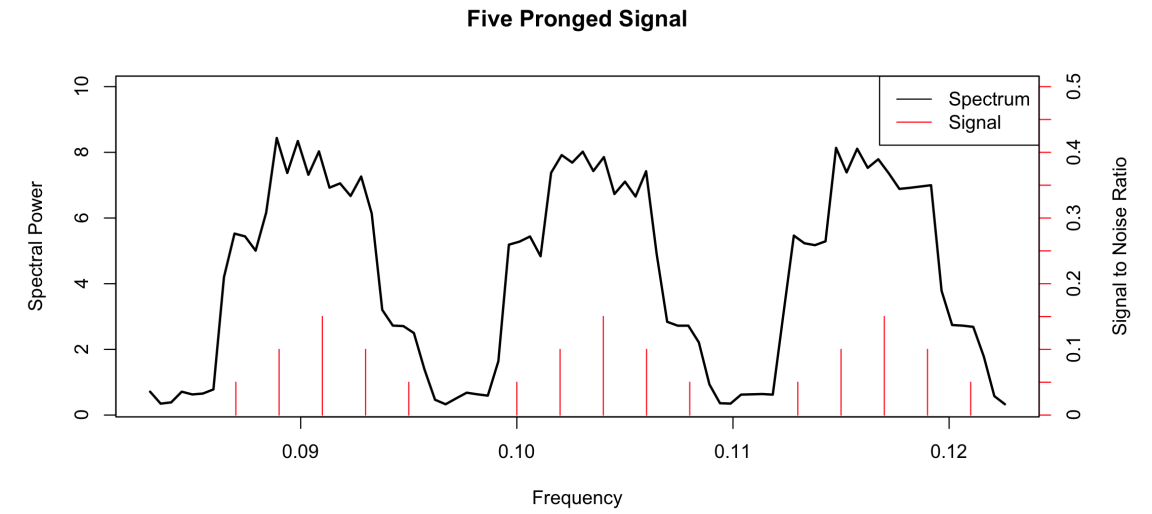

Figure 3.2: Part of the spectrum showing three test signals for the sphericity test, $NW = 4$, $K = 7$, $N = 1000$.

$NW \in [2, 10]$ and $K \in [2, 20]$ to determine the optimal parameter choices for our simulated data.

Performing the naive sphericity test on $X(t)$, we found that $NW = 6$ and $K = 10$ provided the lowest probability of the residuals not being spherical. The bagged sphericity test with $O = 50$ had a similar parameter choices of $NW = 6$ and $K = 8$. These results are within our theoretical range on the parameters and goes along with our motivation for a good uninformed first choice of parameters. We also notice

that as NW moves away from the theoretically reasonable choices, where overlap in the signals would occur and the p-value drops, this is visible in the naive test for sphericity, Figure 3.3. The probability also is lower for large K; this may be caused by the signals having little power at the band edge.

While the tests do agree on similar choices of parameters, they do observe slightly different characteristics. The naive test has considerably higher p-values, with the maximum being around .9547, compared to the bagged tests maximum p-value of .6429. Further testing did show that as O increased the p-values of bagged test decreased. Additionally the naive test had more parameter choices with p-values near the maximum, while the parameter values near the choice were similar in scale for the bagged test but all other values had p-values of 0.

To examine the performance of both methods we additionally performed 1000 simulations of five-pronged sinusoidal data with noise to determine how often a theoretically acceptable choice was given. As we can see from Figures 3.5 and 3.6, there is similar performance for both methods. We found that there was an increase in proportion found within the theoretical region by the bagged test with 73% of the parameter selections being within the range compared to 64% for the naive test. Using a two population z-test for proportions we found that the p-value for the hypothesis that the bagged proportion is not greater than the naive test was 7.38×10^{-6}. This is highly improbable and shows that we do have a significant improvement from the bagged test.

The choice of O within the bagged test had several effects on the results of the test. First, when we increase O there is convergence in the choice of parameters with the highest p-value, that is as O increases the likeliness of the parameters with the

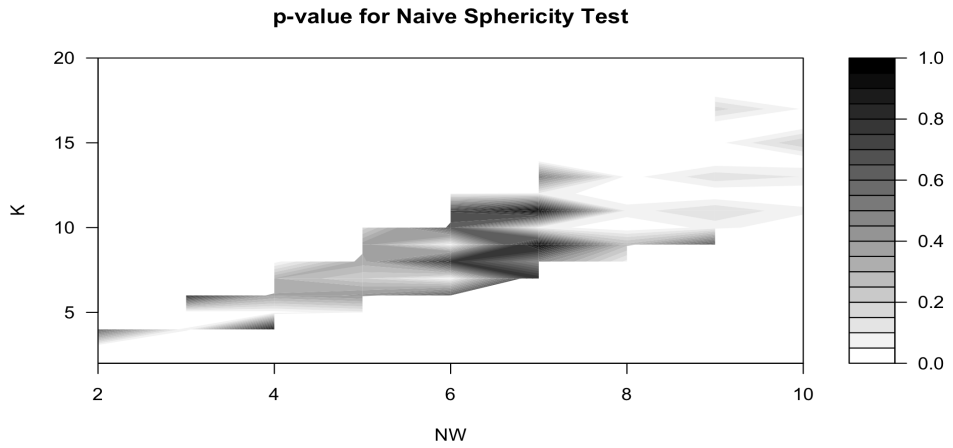

Figure 3.3: Naive sphericity test of simulated evenly spaced 5-pronged sinusoids in noise for $NW = [2, 10]$ and $K = [2, 20]$.

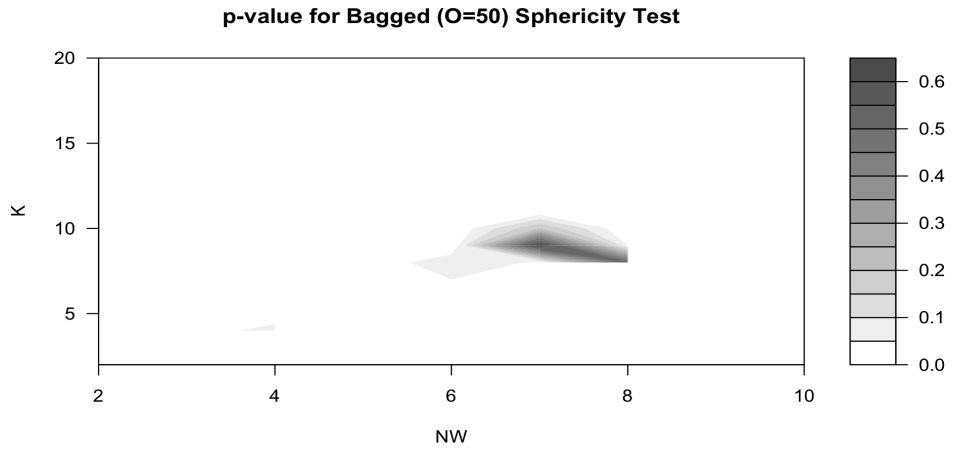

Figure 3.4: Bagged sphericity test with $O = 50$ for simulated evenly spaced 5-pronged sinusoids in noise for $NW = [2, 10]$ and $K = [2, 20]$.

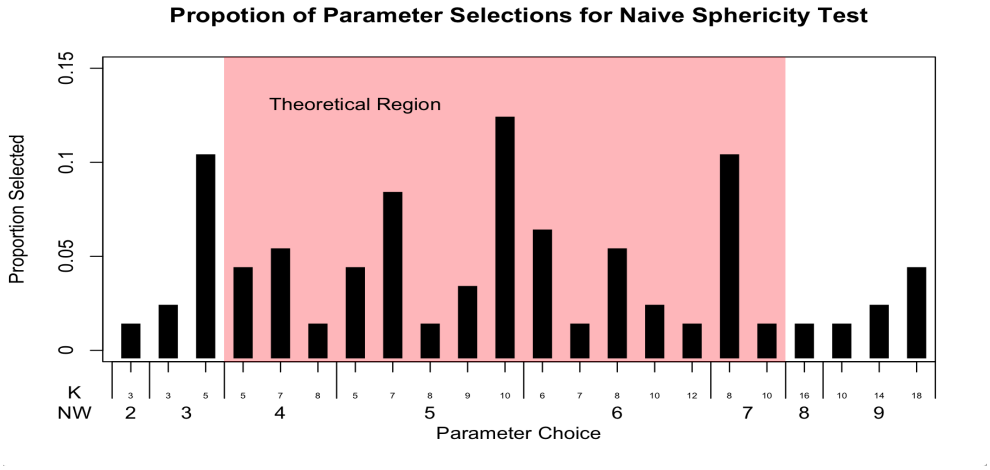

Figure 3.5: Proportion of parameter selections of the naive sphericity test for 1000 repetitions of simulated evenly spaced five-pronged sinusoids in noise for $NW = [2, 10]$ and $K = [2, 20]$. All parameter choices not listed were not selected.

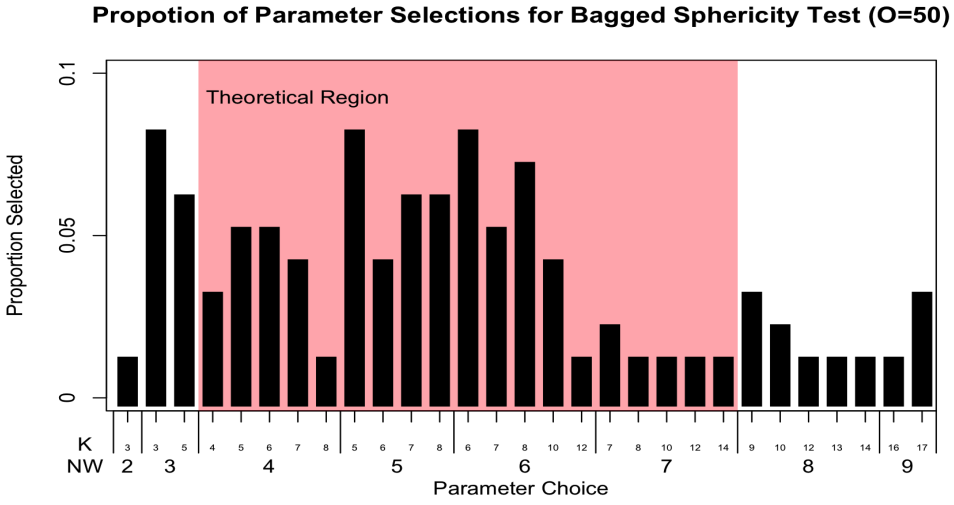

Figure 3.6: Proportion of parameter selections of the bagged sphericity test with $O = 50$ for 1000 repetitions of simulated evenly spaced five-pronged sinusoids in noise for $NW = [2, 10]$ and $K = [2, 20]$. All parameter choices not listed were not selected.

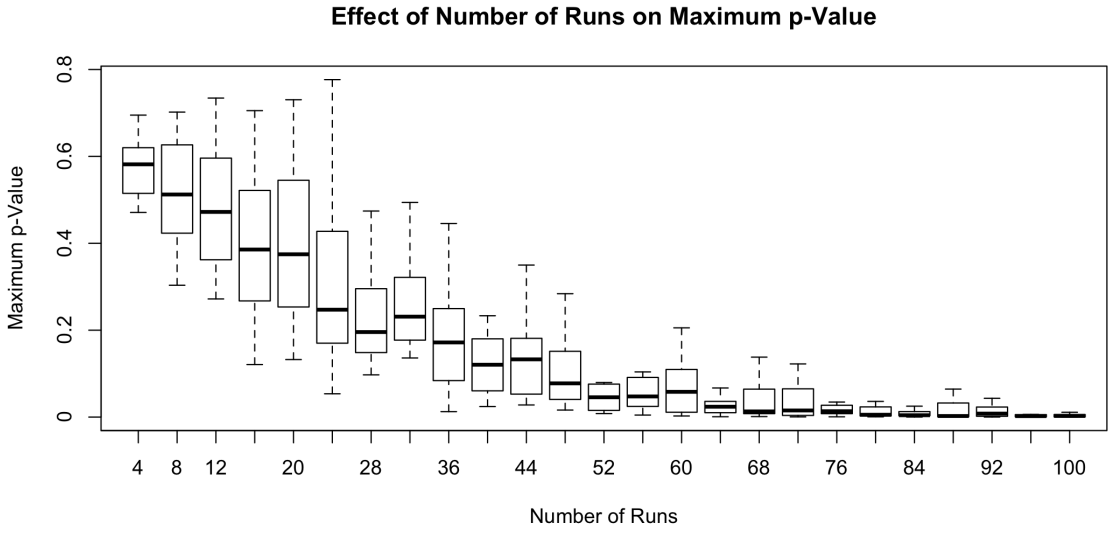

Figure 3.7: Effect of number of runs, O, on maximum p-value for the bagged sphericity test

highest p-value changing from one computation to the next decreases. This results from the sampling method used, as the total number of runs increases, the test will be sampling more combinations of the data. As well, with a larger value of O, Fisher's statistics will be summing over more p-values and the parameter choices that only work well on particular combinations will increase in p-value as more combinations are potentially used. This effect also explains the decrease in maximum p-value for Fisher's statistics as O increases, the number of combinations that contain poorly resolved residuals that are included will increase with O and their p-values near 0 will decrease the Fisher Statistic. This makes logical sense as the alternative hypothesis is that one or more of the group of sphericity tests fails. We can see this relationship in Figure 3.7.

This convergence property does come at the cost of computing time. Figure 3.8

shows a linear relationship between the computing cost and the number of runs used. The naive sphericity test for reference comes in at an average computational cost of $1.1sec$ for the same data set.[2]

For small values of O, there is considerable variance in the parameters returned as ideal. To demonstrate this property, we evaluated the bagged sphericity test 1000 times for $O = 10$ on the same data set. We then plotted the percentage of occurrences for which each parameter was the ideal choice. As you can see in Figure 3.9, there is a large set of values chosen to be the ideal parameter choices for a given sample. Fortunately, we still outperformed a random selection of parameters with a proportion of theoretically reasonable choices of .588 compared to $\frac{26}{63} = .413$ if we were to have a uniform probability of selecting any combination from the possible parameters.

While it appears that both tests are effective in this circumstance, we would like to know what effect an improper estimate of σ^2 might have on the bagged test. To examine this, we repeated this simulation but with a variety of incorrect choices of noise power from $\sigma \in [0.6, 3.7]$. The result of the average of 20 trials at each choice in the range, displayed in Figure 3.10, showed that the more you misspecify the noise variance the less likely you are to correctly select the parameters. The smaller numbered parameters being selected more often as we start to choose values farther away from the correct noise variance is a result of our code providing the smallest parameter choice that gives equal probability. As we begin to make increasingly worse choices, the p-values increase to the point that all choices are equally poor, having values of 1. We notice the near edge of reasonable choices for the noise variance we find an increased variability in choices of parameter. This effect was due to all

[2]All computer analysis in this thesis was performed in R 3.1.2 on an Apple Macbook Pro 13" running OS X 10.8.5 with a 3GHz Intel Core i7 and 8GB of RAM.

parameter choices being near 1. In addition, we show in Figure 3.11 that all p-values were 1 for all parameter choices at the extreme values we tested, which demonstrates that this test fails in the event that you do not have a reasonable estimate of the noise variance. We also notice that the bagged test appears to be more robust to over estimating the noise level rather than underestimating.

As a note on the choice of variance for the bagged test, for all tests here where we did not specify the noise level used, we used $\sigma^2 = 2$. In practice it is uncommon to have the true value of the noise variance as we do. You are more likely to estimate the sample variance from the residuals. Now noting

$$\frac{(MK-1)S^2}{\sigma^2} \sim \chi^2_{MK-1}, \tag{3.12}$$

we can identify theoretical confidence interval on the sample variance we expect to observe for the residuals. From 3.12 we get the α confidence bounds for the sample variance are

$$\frac{\chi^2_{MK-1,\alpha/2}\sigma^2}{MK-1} < S^2 < \frac{\chi^2_{MK-1,1-\alpha/2}\sigma^2}{MK-1}. \tag{3.13}$$

Noting that for our simulations $K \in [2, 20]$ and $M \in [50, 250]$, we have a $\alpha = .01$ confidence interval that could have resulted in observing sample standard deviations as small as 1.18 or as big as 1.649 when $MK = 100$. The confidence intervals for $MK > 100$ are tighter than those for $MK = 100$. From looking at Figure 3.11 we see that the test would not perform optimally but would provide reasonable results for the most extreme sample variances we expect to observe. As such, we do not expect that there would be much reduction in performance from estimating the noise variance.

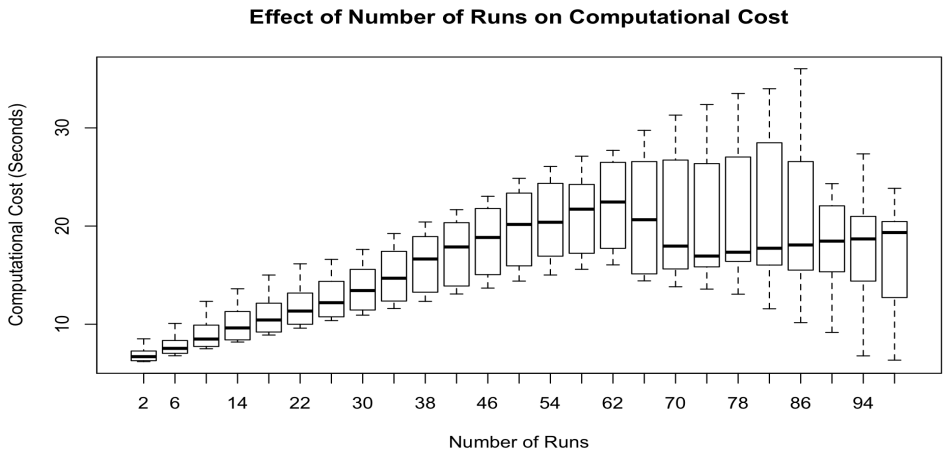

Figure 3.8: Effect of number of runs, O, on computational time for the bagged sphericity test

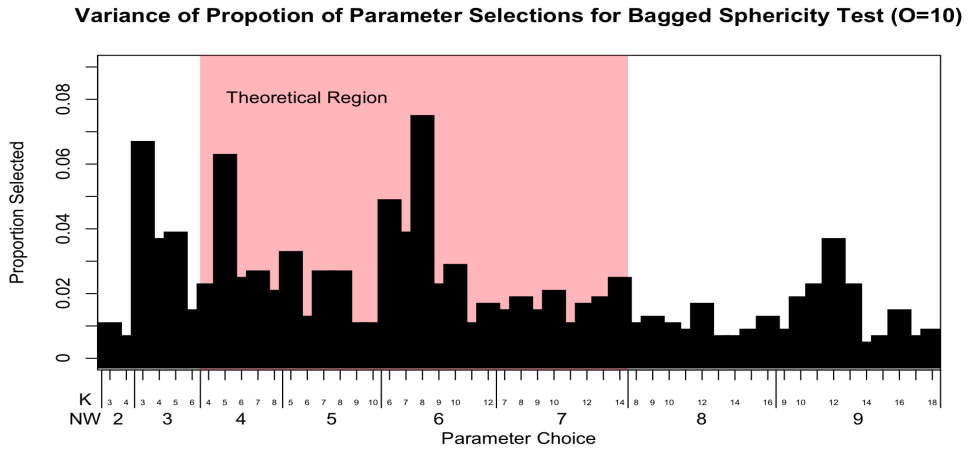

Figure 3.9: Variance in the the maximum p-value parameter choice from 1000 testings with $M = 10$. All parameter choices not listed were not selected.

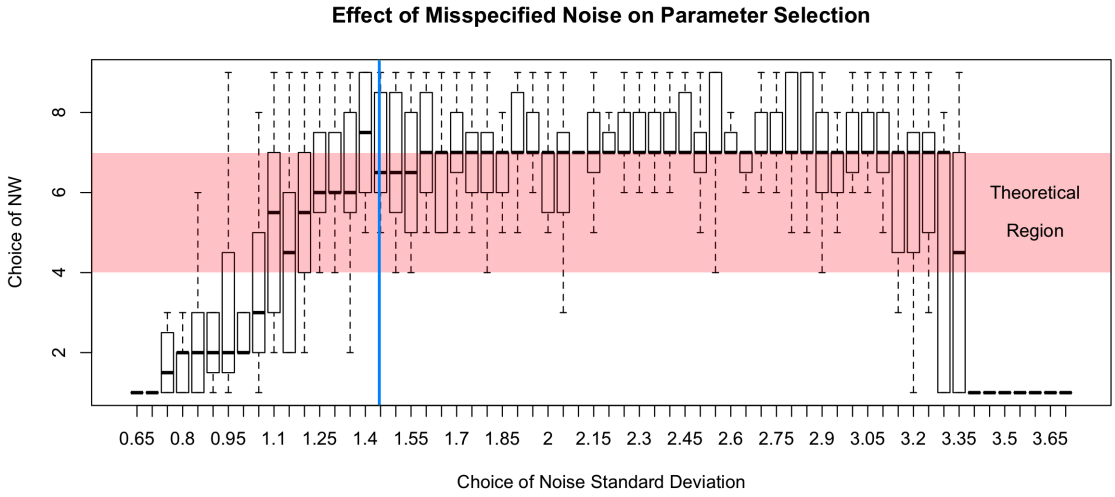

Figure 3.10: Effect on the choice of NW from wrongly specifying the noise process variance. The true variance is labeled as the blue line and the theoretically acceptable choices are highlighted by the red band.

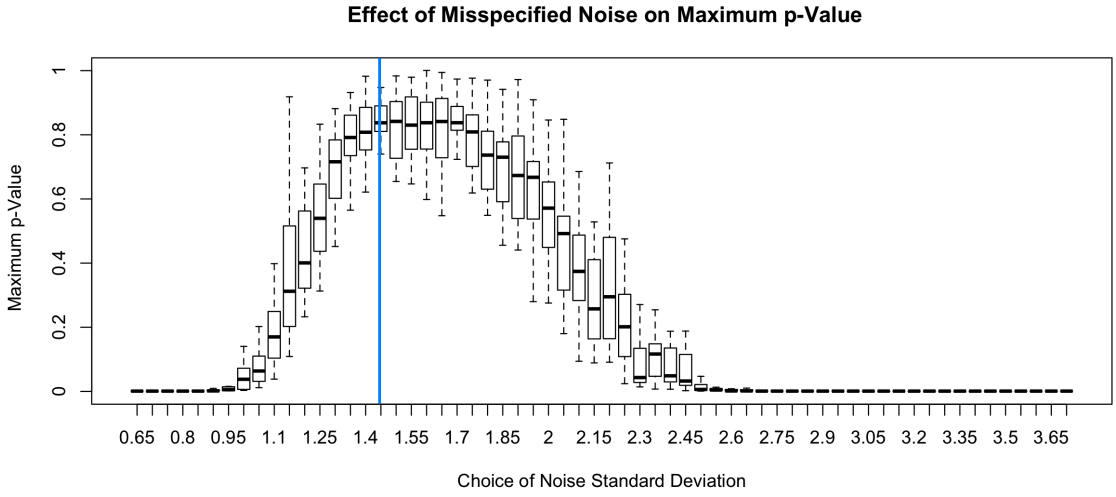

Figure 3.11: Effect on the maximum p-value for the bagged sphericity test due to wrongly specifying the noise process variance. The true variance is labeled as the blue line.

We lastly examined how both methods performed when the assumption of Gaussian noise was not upheld. To do so we examined different proportions of non-Gaussian noise sampled instead of standard Gaussian noise. We tested three alternate distributions, a centralize t-distribution with one degree of freedom, a non-centralize t-distribution with two degree of freedom and non-centrality parameter of one, and a uniform distribution with range from negative three to three. As we see in Figure 3.12, the bagged test does not perform as well on all types of non-Gaussian noise as the naive test at even the smallest proportion. We suspect this is due to the more specific hypothesis tested in the bagged test.

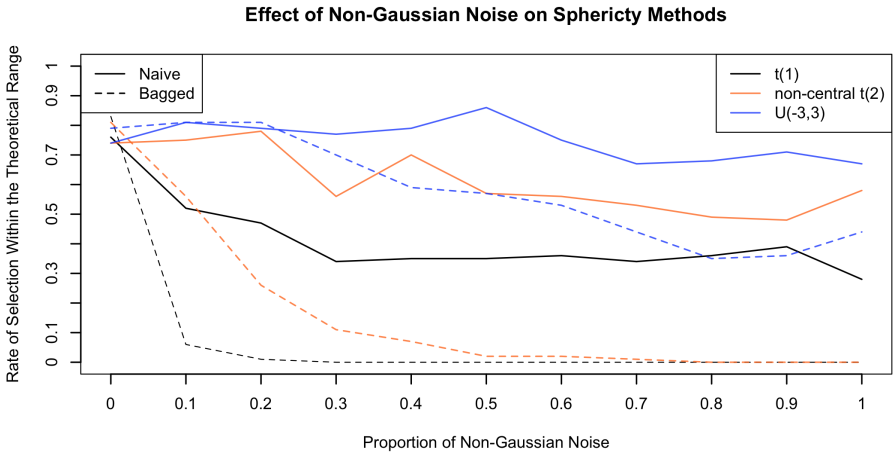

Figure 3.12: Comparison of the sphericty tests' performance under differing proportions of non-Gaussian noise. The leftmost results are calculated using standard Gaussian noise while the rightmost with the non-Gaussian distributions listed.

3.5 Conclusions on Tests

We found that both sphericity tests were able to identify when the residuals were well resolved and provided an appropriate choice in parameters. The tests also had the desirable property of decreased probability of being spherical when NW was either too small or large. While there is the concern that the naive test is not designed to differentiate between deviance in the magnitude of variance and well resolved residuals, it did identify similar parameters as the bagged test for simple simulations. The bagged test did significantly out-perform the naive test at giving reasonable results but the difference in quality of these tests was not so large that the naive test should be discarded. The naive test did perform better in conditions of non-Gaussian noise which may be a result of only testing for the diagonality of covariance matrix and non the more specific hypothesis of the bagged test.

While the bagged test may be the more thorough test, it does come with drawbacks. For reliable results the use of a reasonable ($O > 20$) number of runs is necessary but this comes at the cost of increased computer time. Conversely, the robustness of the bagged test against the misspecification of the power of the noise does not create a problem for data sets where no information about the noise process is known. The use of the naive test is recommended in situations where the noise power is not known or easily estimated, non-Gaussian noise is possible, or computational considerations need to be made. Overall both the tests did provide a reasonable choice of parameters when no theoretical parameter values are possible.

Chapter 4

Bootstrapping the F-test

4.1 Introduction

Within many scientific fields, including health care diagnostics and wireless communications, there is a need for accurate and robust signal detection [33, 141]. There are a plethora of signal detection methods available [84], with the majority of practical methods being designed for a specific data set or problem [16, 35, 49, 122]. With an aim at providing a robust method for most applications, we examined the commonly used multitaper F-test for line components (F-test). Since its development in 1982 by David J. Thomson [116], the F-test has been a commonly used test in the identification of signals within the fields of electrical engineering [83], space physics [94], neurology [36], and many others [2, 44, 70]. While an invaluable tool for signal detection in many situations, the F-test does present problems with the detection of signals with moderate to low power-to-noise ratio. As missed detection can be costly in communications systems as well as providing misleading evidence in scientific studies, this problem will need to be addressed to provide a robust signal detection test

that can be effective in multiple applications. In an effort to lessen the issue of missed detection and improve upon the false detection rate of the F-test, we have developed a bootstrapping process for use in calculating the F-test statistic.

4.2 Practical Limitations of the F-test

Within the F-test we are testing the null hypothesis that there is no sinusoidal signal present and we are examining only white noise. The test statistic is then $F_{2,2K-2}$ distributed [116]. There are several problems that can arise with the practical use of this test.

First, in the event that a signal (one or more line components) is present but the choices of NW or K used are not appropriate to resolve the signal, we will see some structure in the residuals. By structure we mean values that deviate considerably from the expected standard complex normal distribution and resemble the spectrum of the data. Since the assumption of normality in the residuals is not followed, the χ^2_{2K-2} distribution in the denominator of the F-test will not be followed. As a result, at these frequencies we will be reducing the statistic by using a larger denominator than is expected. This mechanism can result in missed detection.

Another issue is that, under the null hypothesis we can have extreme values in our residuals where the denominator of the statistic is quite small (compared to the expected value of $(2K-2)S_N(f)$). Since the numerator is independent from the denominator, the resulting statistic may be quite high even though no signal is present. This will provide false detections in our spectrum and, while they occur at the expected rates, we can deal with these unusual denominators by using re-sampling methods.

4.3 Testing Proceedure

To work around these two issues, we employ a bootstrap procedure on the residuals from the F-test that is similar to the regression bootstrapping from section 2.6. Remembering that we are performing linear regression when computing the F-test, we apply non-parametric bootstrapping to the regression problem [37]. In the new test we re-sample the residuals with replacement and then estimate both $\hat{Y}_k$ and $\hat{\mu}$. From these new values we can compute a new F-statistic. We perform a reasonable number (> 20) of re-samplings and take the average at each frequency of the F-statistics found. This average is then compared to empirical rejection regions that were found under the null hypothesis. The re-sampled residuals F-test algorithm is,

1) the F-test in the normal fashion to obtain values for $r_k(f)$, $Y_k(f)$, and $\hat{\mu}(f)$.

2) Re-sample the residuals with replacement to produce $\hat{r}_k^{(m)}(f)$.

3) Compute new values for $\hat{Y}_k^{(m)}(f)$, $\hat{\mu}^{(m)}(f)$, $\hat{r}_k^{(m)\prime}(f)$, and $\hat{F}^{(m)}(f)$:

$$\hat{Y}_k^{(m)}(f) = \mu(f)V_k(0) + \hat{r}_k^{(m)}(f), \tag{4.1}$$

$$\hat{\mu}^{(m)}(f) = \frac{\sum_{k=0}^{K-1} V_k^{\star}(0)\hat{Y}_k^{(m)}(f)}{\sum_{k=0}^{K-1} |V_k(0)|^2}, \tag{4.2}$$

$$\hat{r}_k^{(m)\prime}(f) = \hat{y}_k^{(m)}(f) - \hat{\mu}^{(m)}(f)V_k(0), \tag{4.3}$$

$$\hat{F}^{(m)}(f) = (K-1)\frac{|\hat{\mu}^{(m)}(f)|^2 \sum_{k=0}^{K-1} |V_k(0)|^2}{\sum_{k=0}^{K-1} |\hat{r}_k^{(m)\prime}(f)|^2}. \tag{4.4}$$

4) Perform steps (2) and (3) for each m, $m = 1, \ldots M$ and take the mean of the F-statistics found for each re-sampling:

$$\overline{F}(f) = \frac{1}{M}\sum_{m=1}^{M} \hat{F}^{(m)}. \tag{4.5}$$

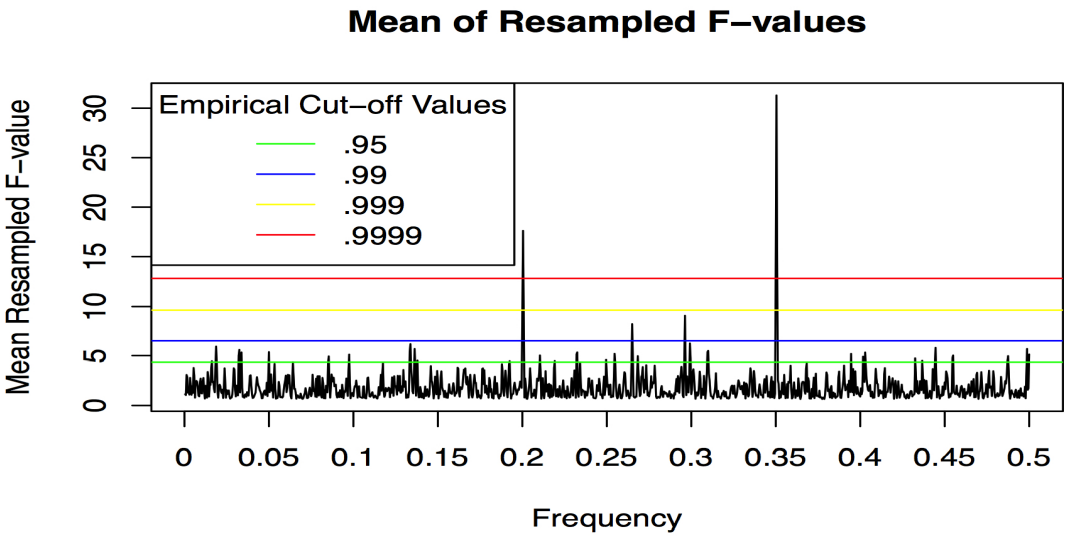

Figure 4.1: Example of the re-sampled residuals F-test of a sinusoid at $.35Hz$ in Gaussian noise, $NW = 4$, $K = 7$, $N = 1000$.

5) Lastly check whether this value exceeds our empirical rejection value to determine whether a signal is present.

Under H_0:

$$\overline{F}(f) \leq \hat{\phi}_{2,2K-2}(p), \tag{4.6}$$

where ϕ is the empirical lower bound of the rejection region for our test with significance level p. The values for $\hat{\phi}_{2,2K-2}(p)$ will be discussed in greater detail in the following section. An example of a bootstrapped test plotted with empirical cut-offs is shown in Figure 4.1.

4.4 Rejection Regions and Variance of the Bootstrapped Statistic

For use of the bootstrapped F-test as a method of signal detection, we need to identify lower bounds, $\phi_{2,2K-2}(p)$, for rejection regions for detection of a signal from the statistic computed at different significance levels. There is no simple theoretical distribution under the null hypothesis possible to produce rejection regions for our bootstrapped statistic [22]. As a result, we identified empirical rejection regions, $\hat{\phi}_{2,2K-2}(p)$, for the test statistic instead. This is accomplished by performing the bootstrapped F-test on at least $2/p$ samples of Gaussian noise with variance equal to the power of the background noise of the series. Then, after sorting the data, we are able to get an estimate of the rejection regions for our test statistic under the null hypothesis.

To compute $\hat{\phi}_{2,2K-2}(p)$ (the value of the test statistic where the empirical cumulative distribution equals $1-p\%$), we find the value at which $p\%$ of the bootstrapped F-test results from the Gaussian data are greater than p. Then for a significance level p test, we use $\hat{\phi}_{2,2K-2}(p)$ as the lower bound on the rejection region for the null hypothesis. We determined the cut-off point (lower bound) for several common parameter choices and significance levels for $S_N = 1$; these cut-off values are given in Table 4.1.

Table 4.1: Empirical cut-off values, $\hat{\phi}_{2,2K-2}(p)$, for the re-sampled F-test($S_N = 1$)

	NW/K					
Probability	3/5	3/6	4/7	4/8	5/9	5/10
.90	3.618	4.216	3.487	3.746	3.441	3.651
.95	4.582	5.487	4.377	4.813	4.582	4.627
.99	6.924	8.762	6.495	7.466	6.290	6.938
.999	10.491	14.538	9.635	11.724	9.212	10.374
.9999	14.574	22.928	12.845	16.742	12.228	14.403

Another area that is important to investigate if we plan to implement the bootstrapped F-test in practical settings is the variance of the test statistic. We require an estimate of the number of re-samplings needed to minimize the variability of the bootstrapped F-statistics to avoid unnecessary computation for our testing. Using a sinusoidal signal,

$$y_t = .3\sin(2\pi.05t) + .3\sin(2\pi.1t) + .3\sin(2\pi.135t) + .3\sin(2\pi.2t) + .3\sin(2\pi.25t) + .3\sin(2\pi.305t) + .3\sin(2\pi.35t) + N(0,1). \quad (4.7)$$

we estimated the variance of the bootstrapped F-statistic. With 20 sets of 20 replications of 1000 samples simulated from equation 4.7, we examined the variance in the bootstrapped F-statistic ($NW = 4$, $K = 7$) of each set for a given re-sampling size. The choice of $NW = 4$ and $K = 7$ was found to be a reasonable choice by the bagged sphericity test. After evaluating a range of re-sampling sizes, we found that more than 20 re-samplings were not needed to minimize the variability. This is shown

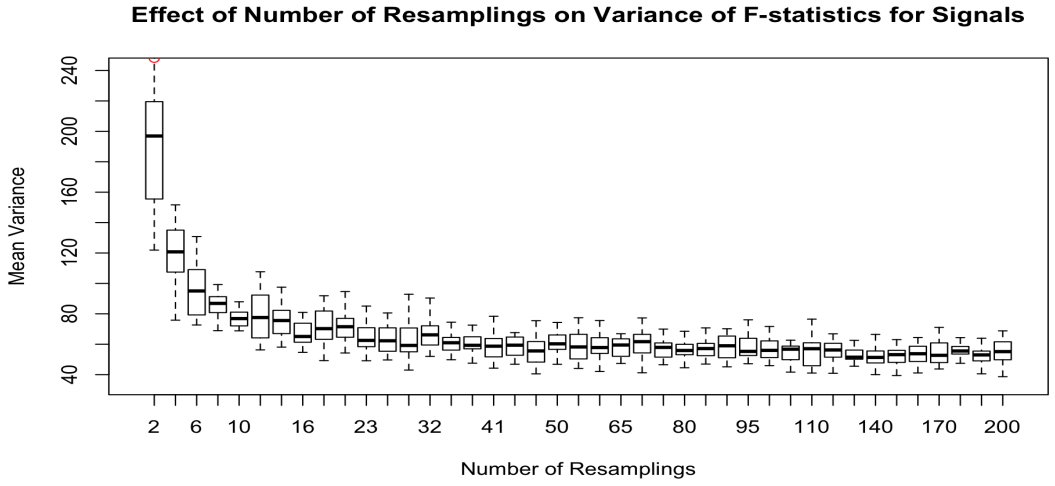

Figure 4.2: The effect of re-sampling size on the variability of the bootstrapped F-statistic for signal carrying frequencies ($NW = 4$, $K = 7$).

in Figure 4.2. We also looked at the effect the number of re-samplings had on the variability of the frequencies containing only noise. The variance of the frequencies containing no signals leveled off around 50 re-samples to the theoretical variance of the noise process, $Var(F_{2,12}) = 2.16$. This is shown in Figure 4.3.

Additionally we examined the effect that re-sampling size had on the magnitude of the bootstrap statistics for signals or noise. We found no change in the magnitude of the statistics for changes in re-sampling size in either case. From this we recommend using a minimum of 20 re-samplings to stabilize the resulting statistic. Using a larger number may be beneficial for data that is not as ideal as the simulated data presented here. The obvious drawback being extra computational costs when using more re-samplings.

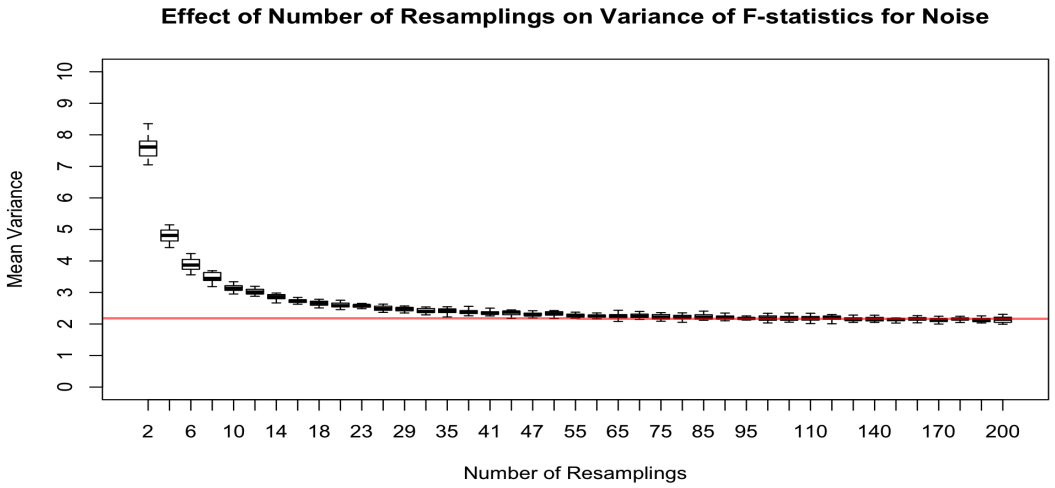

Figure 4.3: The effect of re-sampling size on the variability of the bootstrapped F-statistic for noise frequencies ($NW = 4$, $K = 7$). The red line is the theoretical variance, $Var(F_{2,12}) = 2.16$.

4.5 Comparison to the F-test

To show the advantages of re-sampling the residuals within the F-test, we examined whether this test outperforms the traditional F-test. The performance metrics that we are concerned with are the missed detection and the false detection rates. We tested this through the use of simulated line components at varying levels of power, determining how often either method detected the signals at set significance levels. For this test we used parameter values $NW = 4$ and $K = 7$. By using the bagged sphericity test from chapter 3 on the residuals of the F-test we were able to identify that the spectrum was well resolved for those parameter values.

To compare these tests, we first started by generating 1000 samples of a signal

from,

$$Y_t = \alpha \sin(2\pi(.125t)) + z_t, \tag{4.8}$$

where α is the amplitude of the signal and $z_t \sim N(0, 1)$. We then examined whether the F-test gave a statistic higher than $F(1-p, 2, 12)$ for the frequency bin centered on the sinusoid's frequency of .125Hz. p here is the probability of having a realization of that value or higher under the null hypothesis of no signal being present. We repeated the generation of a signal with noise 500 times for each value of α. We examined α values ranging from 0 to .35. We performed this test for $p = .05, .01$.

In the same vein as the testing done on the F-test, we also perform the bootstrapped test for all choices of α and p, 100 times using 30 re-samplings. We then determined whether the statistic was higher than our empirical estimates of the cut-off values at the significance level of the test.

We found, the tests performed similarly poor at very low power ($\alpha < .005$) and had near perfect rates of detection for high powered signals (($\alpha > .25$). Additionally, the performance of the tests differed as the amplitude increased, the bootstrapped test outperforming the traditional F-test. This effect is shown in Figures 4.4, 4.5, where there is an increase in the detection rates of the bootstrapped test over the traditional F-test for both the cutoff levels. We tested whether the increase in detection was significant by considering, for each signal amplitude, each attempted detection as a Bernoulli random variance and performing a comparison of proportions hypothesis test [131]. The comparison of proportions test is analogous to comparing the rates, where the sample proportions are the mean rates of detection. We defined the null hypothesis with respect to the detection rate, R_b and R_t, as $H_0 : R_b - R_t \leq 0$. The

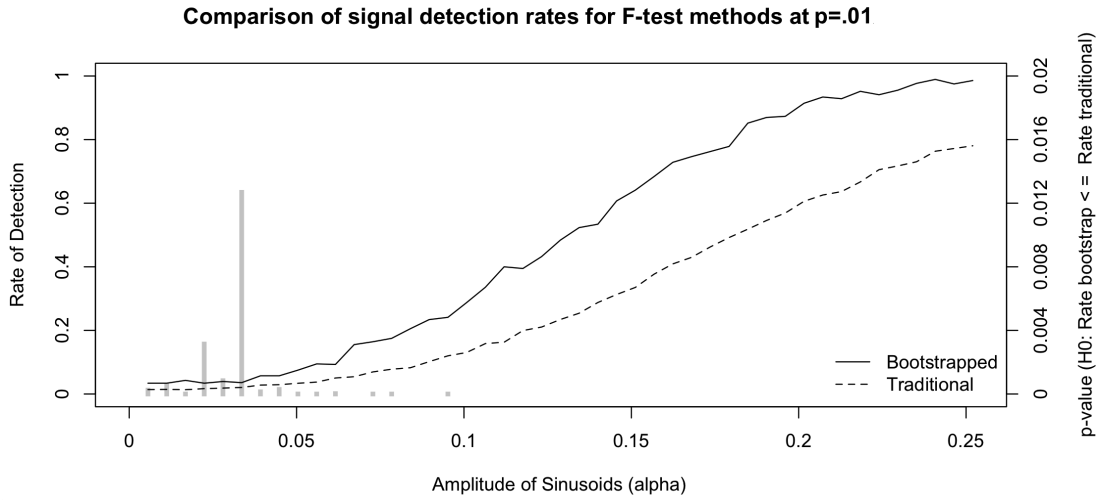

Figure 4.4: Comparison of detection rates for the F-test methods for a range of signal amplitudes with $p = .01$. p-values for $H_0 : R_{bootstrap} \leq R_{traditional}$ are provided as gray bars at each frequency. When no bar is provided the p-value is approximately zero.

test statistic being

$$z = \frac{\bar{R}_b - \bar{R}_t}{\sqrt{\bar{R}_p(1 - \bar{R}_p)(\frac{1}{N_b} + \frac{1}{N_t})}}, \tag{4.9}$$

where $\bar{R}_p = \frac{Y_b + Y_t}{N_b + N_t}$, N_b, N_t are the overall number of signals tested across all runs, and Y_b, Y_t are the total number of detection across all runs. Under the null hypothesis $z \sim N(0, 1)$. We can then calculate a p-value for each signal amplitude to determine if the difference is significant. We found that for all signal amplitude between .05 and .25 there was a significant difference in the detection rates for both methods.

Another result that came out of this simulation is that we were able to test the false detection rate of the re-sampling test when a signal is present. Using the same cut-off values, we found that for $p = .05, .01$ we had false detection rates of .0416

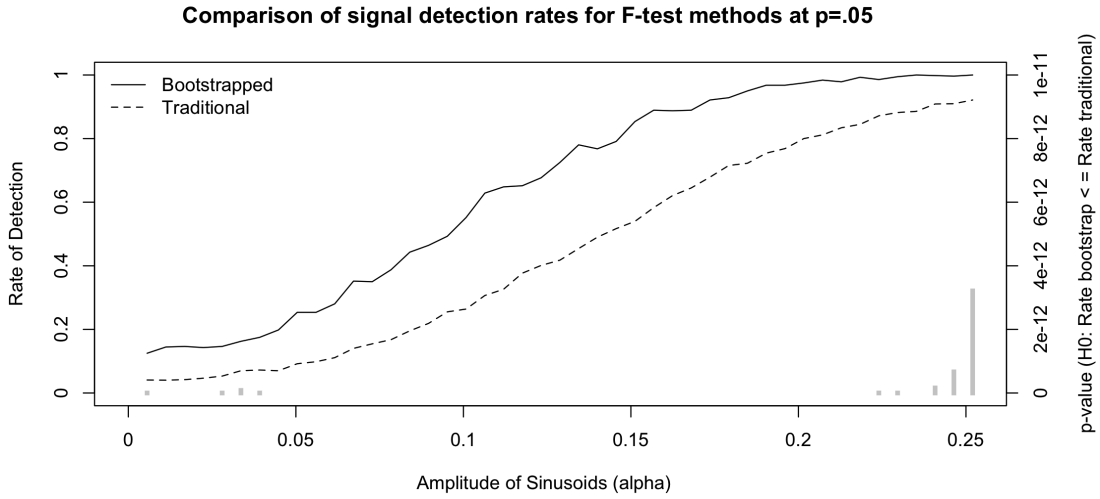

Figure 4.5: Comparison of detection rates for the F-test methods for a range of signal amplitudes with $p = .05$. p-values for $H_0 : Rate_{bootstrap} \leq Rate_{traditional}$ are provided as gray bars at each frequency. When no bar is provided the p-value is approximately zero.

and .0068. These results are lower than expected, which may be attributable to the inclusion of residuals from frequencies where a signal was present in the re-sampling distribution. These residuals are larger than most coming from the noise process so they have a slight dampening effect on the statistics.

Lastly we wanted to investigate the computational cost of the bootstrapped detection method. The computational cost is dependent on two parameters, the number of tapers used K and number of re-samplings we perform. As we increase K there is a positive correlation to computational cost, while increases in re-sampling size did not have a large effect. This makes sense as we have more residuals to re-sample when we increase K. We can see the relationship these two parameters have with computational cost in Figures 4.6, 4.7. There is a lowering in computational cost at 38 re-samplings, this may be due to R's internal optimization or the distribution of processing power during the parallelization of the re-samples. The average time for the traditional F-test was considerably less, at .034 seconds on average, with the bootstrap method at the smallest acceptable choices for the parameters still being around 1000 times larger in computational cost.

4.6 Conclusions on Simulations

We found that adding a bootstrapping procedure to the F-test for line components test has improved performance over the traditional F-test for use on simulated data. When using the bootstrapped F-test we saw improvement in both signal detection and false detection rates for sinusoidal signals. For extremely low-powered signals, we did not find any improvement. Both tests in this situation were unable to consistently detect signals. The bootstrapping method did out-performed the traditional F-test

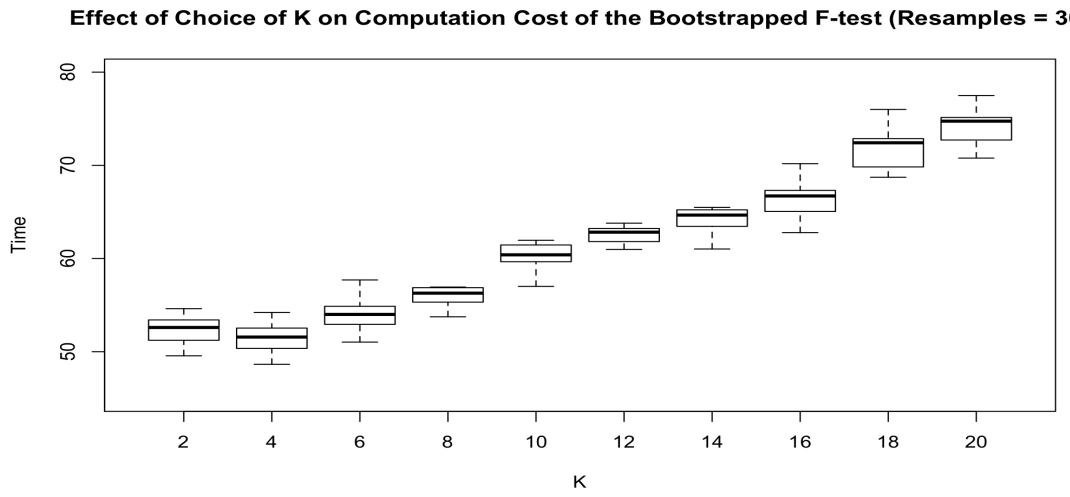

Figure 4.6: Effect of the choice of K on computational costs for the bootstrapped F-test.

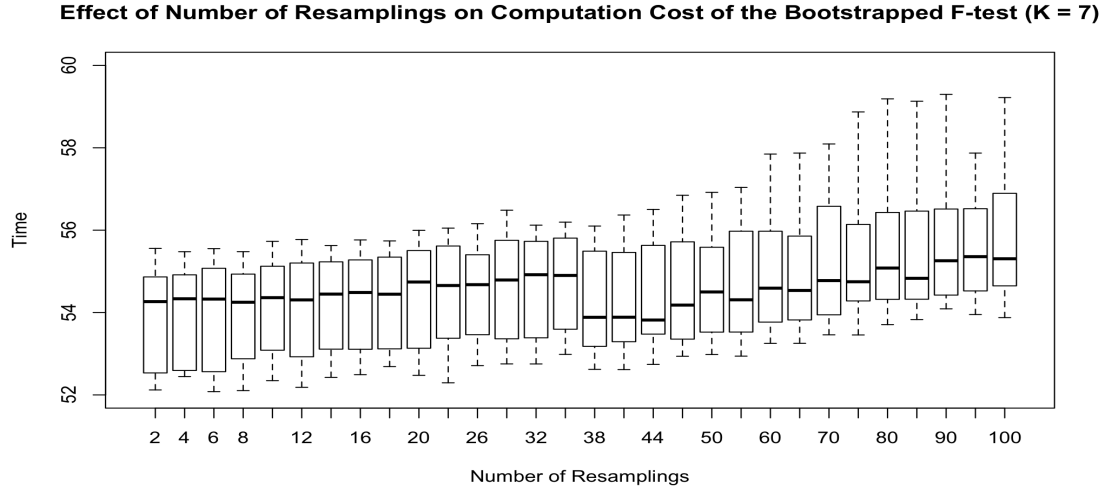

Figure 4.7: Effect of the number of re-samplings used withing the bootstrapped F-test the on computational costs of the method.

for amplitudes where the detection rate was increasing. In situations where the signal strength was strong both methods were able to identify the signals consistently.

The bootstrapped F-test provides an improvement in detection for amplitudes when the performance is not guaranteed be poor or excellent ($\alpha \in (.005, .25)$) In most real world situations we are not going to have prior knowledge of the signal amplitude for a series, and, as such, we are not likely to be able to guarantee that the signal strength will be adequate for detection consistently with the traditional F-test. In this situation, we would have equal or better performance from the use of the bootstrapped F-test. The only concern being the increased computational cost but the computational costs are still reasonably efficient. In situations where the noise power can be estimated, we do not need to produce cutoff values for the bootstrapped F-test and the computational costs are significantly reduced, making this method even more applicable.

An added bonus to the bootstrapped F-test was the lower false detection rates. This is potentially the result of having overly conservative cutoff values for our tests due to the sample sizes used to produce them. In the event that our cutoff estimates were too conservative, we would see a reduction in signal detection performance. Since the bootstrapped F-test outperforms the traditional F-test, we do not concerned with the cutoffs we have found.

The next step is to test the bootstrapping procedure on real data. We use the bootstrapped method within the study of atrial extraction for electrocardiograms and in Figure 6.1 we see that there is a significant increase in detection for this data. We do not believe that in all data situations the bootstrap method with increase the detection rate but at a minimum it will have similar detection rates to the naive

test at a marginal increase in computational cost. We recommend that in situations where the noise power can be easily estimated or computational costs are not a major concern that the the bootstrapped F-test should be used.

Chapter 5

Periodic Data Reconstruction Methods

5.1 Introduction

Often in the study of time series data we are presented with the problem of synthesizing data. This synthesis may be required to interpolate missing data [39] or to predict future values and trends [130]. In either case, there are a variety of approaches that a statistician can employ to achieve reasonable results [19, 48, 133, 139]. The focus of this chapter is to spotlight a lesser-known synthesis method that uses multitaper spectrum estimation and to provide a framework and tools for use on real data problems.

The technique we will employ is to perform an inverse Fourier transform on the complex regression coefficients that are produced when computing the F-test for line components. This method is attributed to Dr. David Thomson, who has not published a paper outlining this method but has used it in analysis [117] and has

lectured on the topic. After outlining the details of how the method works, we will provide some insight into and advancements on the basic procedure using techniques borrowed from statistical learning theory. Analysis of simulated test data is used to demonstrate how these methods work, explore their properties and evaluate their performance. We also investigate two real-world data problems to demonstrate the merit of these techniques.

The use of Thomson's synthesis method relies on the signal detection and complex mean values, $\hat{\mu}(f)$, resulting from the computation of the F-test for detection of line components, which is described in section 2.5. The choice of significance level, α, for the F-test will greatly affect the estimates returned from Thomson's method so care must be taken when making this choice. We will discuss later an unsupervised method for selecting α and the effect of the choice of α on the estimates.

5.2 Inverse Fourier Transform Signal Synthesis

Following the identification of significant frequencies within the spectrum from the F-test (the details on how to identify signals with the F-test are discussed in section 2.5), we want to synthesize the signals at these frequencies to form an estimate of the periodic elements of the time series without noise. The obvious way we may attempt to do this is by determining the phase and amplitude of the significant periodic components and modeling the time series as the sum of sinusoids with these properties. This has been shown to have been marginally successful [94] but this can become a cumbersome set of computations if the set of significant frequencies is large.

Thomson has proposed an alternative method that follows directly from the F-test. By performing an inverse Fourier transform on the regression coefficients from

the F-test, we are able to get a reasonable approximation of our original time series. That is,

$$\mathcal{F}^{-1}(\mu)(t) = \sum_{f=-f_n}^{f_n} e^{i2\pi ft} \frac{\sum_{k=0}^{K-1} V_k^\star(0) Y_k(f)}{\sum_{k=0}^{K-1} |V_k(0)|^2} \tag{5.1}$$

$$= \frac{1}{\sum_{k=0}^{K-1} |V_k(0)|^2} \sum_{f=-f_n}^{f_n} \sum_{k=0}^{K-1} e^{i2\pi ft} V_k^\star(0) Y_k(f) \tag{5.2}$$

$$= \frac{1}{\sum_{k=0}^{K-1} |V_k(0)|^2} \sum_{k=0}^{K-1} V_k^\star(0) \sum_{f=-f_n}^{f_n} e^{i2\pi ft} Y_k(f). \tag{5.3}$$

Noting that the inner summation is the discrete inverse Fourier transform of $Y_k(f)$ and $Y_k(f)$ is defined as the Fourier transform of $\nu_t^{(k)} x_t$, we can use the inversion theorem for Fourier transforms to replace the inner summation:

$$\mathcal{F}^{-1}(\mu)(t) = \frac{1}{\sum_{k=0}^{K-1} |V_k(0)|^2} \sum_{k=0}^{K-1} V_k^\star(0) \nu_t^{(k)}(N x_t) \quad = \frac{N \sum_{k=0}^{K-1} V_k^\star(0) \nu_t^{(k)}}{\sum_{k=0}^{K-1} |V_k(0)|^2} x_t. \tag{5.4}$$

Inverting this weighting we can get our time series,

$$x_t = \frac{\sum_{k=0}^{K-1} |V_k(0)|^2}{N \sum_{k=0}^{K-1} V_k^\star(0) \nu_t^{(k)}} \mathcal{F}^{-1}(\mu)(t). \tag{5.5}$$

Now evaluating the fraction we have,

$$\sum_{k=0}^{N-1} V_k^\star(0) \nu_t^{(k)} = e^{-i2\pi(0)t} \tag{5.6}$$

$$\sum_{k=0}^{N-1} V_k^\star(0) \nu_t^{(k)} = 1 \tag{5.7}$$

$$\sum_{t=0}^{N-1} \sum_{k=0}^{N-1} V_k^\star(0) \nu_t^{(k)} = \sum_{t=0}^{N-1} 1 \tag{5.8}$$

$$\sum_{k=0}^{N-1} V_k^\star(0) \sum_{t=0}^{N-1} \nu_t^{(k)} = N \tag{5.9}$$

$$\sum_{k=0}^{N-1} V_k^\star(0) \sum_{t=0}^{N-1} e^{-i2\pi(0)t} \nu_t^{(k)} = N \tag{5.10}$$

$$\sum_{k=0}^{N-1} V_k^2(0) = N \tag{5.11}$$

$$|\sum_{k=0}^{N-1} V_k^2(0)| = |N| \tag{5.12}$$

$$\sum_{k=0}^{N-1} |V_k^2(0)| = N \tag{5.13}$$

The relationship in equation 5.6 is noted in Thomson's paper on Rihaczek distributions [119]. These relationships hold approximately when summing for k from 0 to $2NW$.

Inserting the results from equations 5.7, 5.13 into equation 5.5 we get,

$$x_t \approx \frac{N(1)}{N}\mathcal{F}^{-1}(\mu)(t) \tag{5.14}$$

$$\approx \mathcal{F}^{-1}(\mu)(t). \tag{5.15}$$

for $K \geq 2NW$.

Since we do not use $K > 2NW$, we will have some truncation error and in cases with K much smaller that $2NW$, we will need to include a truncation coefficient of $\frac{N(1-\sum_{k=K}^{N-1} V_k^\star(0)\nu_t^{(k)})}{N-\sum_{k=K}^{N-1}|V_k(0)|^2}$. Several examples of truncation corrections are found in figure 5.2. We noticed that there is significant truncation bias at the edges of our intervals and an ripples that persists through the estimates. We found that the truncation bias was proportional to the length of the data. By that, we mean the truncation correction at $x\%$ into the series was the same across changes in sample size. The choice of NW and K were found to have an effect on the truncation correction. Smaller values of NW or K had more samples with considerable bias. The truncation correction values for the middle portion of the time series leveled off for most series and were slightly larger than one on average.

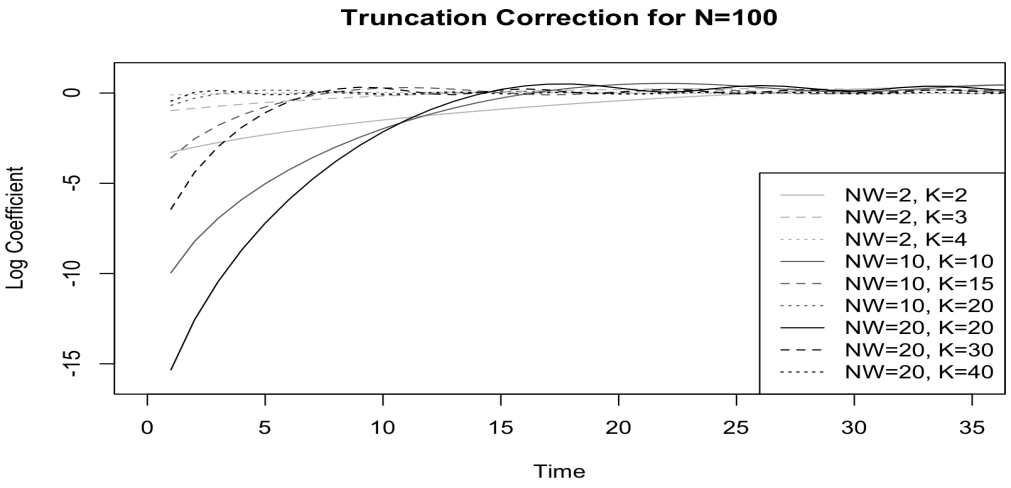

Figure 5.1: Log-scaled plot of the truncation coefficients for a variety of parameter choices with $N = 100$. The first 50 points are plotted while the latter 50 points mirror them.

For choices of NW that are near two it may be best to use the truncation correction, with NW nearer to 10 we found no truncation correction was needed. It is also important to think of the two applications we plan to use this method for. For interpolation, if we have the gap in the middle portion we are not likely to need the truncation correction for reasonable performance. Alternatively there may be issues at the boundaries of the series and we will explore their affect on performance later on in this chapter.

Now we do not have the true values for our regression coefficients, μ. We have estimates that are found from the F-test instead. Using these estimates, $\hat{\mu}$, we can get an estimate of our time series from equation 5.15:

$$\hat{x}_t \approx \mathcal{F}^{-1}(\hat{\mu})(t). \tag{5.16}$$

From this computation we can return an estimate of our time series based on the regression coefficients found within the F-test, but it is unrealistic to expect that a periodic component will be found at each frequency. Instead of using the full set of coefficients, we will use a subset that is significant under the F-test,

$$\hat{\mu}_\alpha(f) = \begin{cases} \hat{\mu}(f), & \hat{F}(f) > F_{(2NW-1,2,\alpha)}. \\ 0, & \text{otherwise.} \end{cases} \tag{5.17}$$

Now we can synthesize an estimate of the significant periodic parts within our time series by replacing $\hat{\mu}(f)$ with $\hat{\mu}_\alpha(f)$ in equation 5.16:

$$\hat{x}_{t,\alpha} \approx \mathcal{F}^{-1}(\hat{\mu}_\alpha)(t). \tag{5.18}$$

5.3 Interpolation and Prediction

The preceding method allows us to reproduce the periodic components within a time series for times where we have observations. This may be useful as a de-noising process for data clean-up, but many problems require interpolation or prediction to be made from time series data. Luckily, we can still use the same method with a small variation to solve these problems. The introduction of some proxy data in the time series will allow us to return an estimate of the periodic components operating at these unknown intervals.

For prediction of future data, we zero-pad the time series the desired number of data points. We then return the estimate of the significant periodic components. The zero-padded points within the returned estimate will now contain the periodic components, giving us a prediction estimate. The added bonus from zero-padding

is that we improve the frequency granularity, which will allow for better estimation of the significant frequencies. It is common practice to zero-pad to improve the resolution of the frequencies for analysis; therefore, in many cases there is negligible extra cost to achieve a prediction from the periodic components. To address the issue of truncation bias it is best to zero-pad the data so that the points we aim to predict are not near the end of the resulting estimated series. In most practical situations we zero-pad more than double the length of data and the resulting series has negligible truncation bias for the predicted points of interest.

To interpolate the data, we will develop some approximation of the missing data first and then perform the same method as before. The most common method is to linearly interpolate the gap in the data, as this will not introduce any new periodic trends, a problem that may occur when using a spline-based approximation [101]. A drawback of using a linear interpolation is the reduced high-frequency power that will be found in the spectrum. This will bias the interpolation by reducing the significance of the higher-frequency components. As this bias will reduce the propensity for spikes in the interpolation, this choice is reasonable for most scenarios. Another potential problem occurs when the gap edge points are extreme values of the time series process. In this situation, the interpolation will have either an incorrect central tendency or a distinct linear trend that is not found in the actual process. This problem is magnified for larger gaps, where the initial interpolation will have more effect on the resulting synthesis. In situations where the gap edge values are extreme or the gap is large, we recommend using the mean of the series instead of a linear interpolation.

5.4 Significance Level Determination (Finding α)

When synthesizing data from a time series, we need to make a choice for the significance level of the periodic components to be used. This choice can influence significantly the estimate we produce from the data. If we set the significance level too low we will accept more frequencies, producing a more turbulent estimate, which can cause undesirable spikes in the data. Likewise, if we set the significance level too high, the estimate will miss much of the periodic structure in the time series.

Finding the optimal significance level is dependent on the data as well as the problem at hand. As interpolation and prediction are two separate problems with differing methods and end goals, it is reasonable to assume that the optimal significance level will not be always the same. To identify the best choice for both cases, we will approach them separately.

For an interpolation problem, we are aiming to fill in a section of missing data with the information contained in the data on either side. It would therefore make sense to choose the optimal significance level for the data surrounding the gap. It also makes sense that the best choice would be chosen with respect to the gap size we plan to interpolate. To meet these aims we propose a cross-validation–based method, in which the data on each side of the gap is divided into bins the size of the gap and the now pseudo-missing data is interpolated.

We denote the data on either side of the gap as $X_r(t)$ & $X_l(t)$ and the size of the gap as g. For a set significance level α, we do the following:

1. Starting with $X_r(t)$, divide the data into bins of size g. In the likely event that the length of $X_r(t)$ does not evenly divide into bins of size g, truncate $X_r(t)$ by discarding the data farthest from the gap.

2. Replace one bin, $X_r((g-1)i+1)\dots X_r(gi)$, with a simple interpolation (linear or mean).

3. Compute an F-test and determine the significant frequencies.

4. Use Thomson's method to produce an estimate of the data, $\hat{X}_r(t)$

5. Calculate the mean squared error of the estimate to the removed bin, $MSE_{r,i}(\alpha) = \frac{\sum_{t=(g-1)i+1}^{gi}(X_r(t)-\hat{X}_r(t))^2}{g}$.

6. Repeat steps 2 through 5 for each bin excluding the first and last.

7. Find the mean of the mean squared errors across all interpolated bins, $MSE_r(\alpha)$.

We now repeat this process for $X_l(t)$ and take the average across both sides to obtain an overall measurement of interpolation error, $MSE(\alpha)$. This process is computed for a range of values of α with the minimum error producing level chosen, $\alpha_{opt} = \underset{\alpha}{\text{argmin}}\ MSE(\alpha)$.

If the gap in the data is larger than the size of the two adjacent portions, this method will not work. If that is the case, we recommend either using the prediction method below or, if there are multiple gaps, starting with the smallest gap and then using this interpolated data as real data for larger gaps, as this will ensure that you have the largest data series possible for the larger gaps.

For prediction problems, our main goal is to ensure the optimal set of periodic components for predicting an interval of new data. The two important parameters that will affect the significance level chosen are the length of the interval we want to predict, n_p, and the length of the data we intend to use to create this prediction, n_t. We assume $n_p < n_t$ as it is considered unwise to attempt to predict a larger time series than the sample used for modeling.

To develop a value of the prediction error associated with each significance level, we split all of the available data into overlapping bins of size $g = n_p + n_t$. The amount of overlap required is dependent on the amount of data we have; ideally, we would have a minimum of 10 bins. We now follow a process similar to that used for interpolation, with the main difference being that for each bin we replace the last n_p data points with zeros instead of using a simple interpolation.

For a data set x_t and a set significance level α, we do the following:

1. Split the data into overlapping bins of size $g = n_p + n_t$.
2. For bin i, $x_{t,i}$, replace the last n_p data points with zeros.
3. Compute an F-test and determine the significant frequencies.
4. Use Thomson's method to produce an estimate of the data, $\hat{x}_{t,i}$.
5. Calculate the mean squared error of the estimate to the predicted times, $PSE_{l,i}(\alpha) = \frac{\sum_{t=n_t+1}^{g}(x_{t,i}-\hat{x}_{t,i})^2}{n_p}$.
6. Repeat steps 2 through 4 but now replace the first n_p data points with zeros and calculate the mean squared error of the estimate to the predicted times, $PSE_{r,i}(\alpha) = \frac{\sum_{t=1}^{n_p}(x_{t,i}-\hat{x}_{t,i})^2}{n_p}$.
7. Find the mean error across both predictions for bin i, $PSE_i(\alpha) = \frac{PSE_{r,i}(\alpha)+PSE_{l,i}(\alpha)}{2}$.
8. Find the mean of the mean prediction errors across all bins, $PSE(\alpha)$.

This will provide a metric for evaluating the performance of the prediction model for a given prediction size and training interval. The assumption of stationarity is vital to the use of multiple bins, with the expectation that the prediction error of

times further away from the prediction are as useful as the closest times. If there is some concern about a gradual change in the distribution of the process, the use of a weighted mean with greater weights for bins with more current times is recommended. If there is a large shift in the distribution of the process, we can use only the most current contiguous data that is assumed to be stationary. If the intention is to use all of the data for the prediction, $n_t = n$, it is advantageous to set as large a size n_t for selecting α as possible while maintaining 10 bins with no more that 50% overlap to ensure that there is no over-training in the significance level chosen.

Additionally, we can potentially improve our reconstruction by implementing the bootstrapped F-test to identify the significant frequencies in the time series. The bootstrapped F-statistic's trait of being greater than the simple F-statistic for frequencies with true signals and smaller for frequencies containing noise can help to improve the likelihood that for the ideal cutoff we select only true signals. There are downsides to implementing the bootstrapped F-test. First, the computational costs are higher, and this, when implemented with several of the other methods, can increase the computational costs exponentially. Also, we are more likely to have spillage of the F-statistic to neighbouring bins. This issue makes it difficult to isolate key frequencies in some situations. Signal spillage within the bootstrapped F-test can be large, particularly in situations of extremely high signal-to-noise power. We will not explore the use of the bootstrapped F-test further in this section but see it as a "Cadillac" extension that could be used.

5.5 Boosting Residual Signals

In most real data situations, not all of the existing periodic components will be modeled in the estimated data. This will result in temporally correlated residuals. If we were able to model the residuals with the same method as the original model, we may be able to improve our synthesized data. To encapsulate this extra periodic information, we use a gradient boosting method. By fitting periodic components to the residuals using Thomson's method, we can identify missing periodic components of our process. We then optimize our model with the new components by finding the minimum squared error for the linear combination of the old model and residuals $M_{i-1}(t)+\gamma_i F_i(t)$, where $M_i(t) = \sum_{j=1}^{i} \gamma_j F_j(t)$ and $F_j(t)$ is the model on the $(j-1)^{\text{th}}$ residual series. $F_1(t)$ is the original model found from the data. When concerned with over-fitting, we can add a cross-validation step in here and choose the γ that minimizes the mean of the cross-validated squared errors.

We then check whether this new model is significant at a preset level, α_{boost}, compared to our old model. To do so, we perform an F-test to compare the two models. Under the null hypothesis that there is no significant improvement in the fit of the boosted model to the data, we would have our statistic,

$$F = \frac{\frac{SSE_{old} - SSEnew}{\#NewFrequencies}}{\frac{SSEnew}{n_t - \#TotalFrequencies}}, \tag{5.19}$$

which should follow an $F_{(\#NewFrequencies, n_t - \#TotalFrequencies)}$ distribution, where n_t is the number of data points used to make the interpolation or prediction.

The algorithm is as follows:

1. For a time series, y_t, we find the optimum periodic estimate, $\hat{y}_t$, as described previously.

2. We next find the residuals for these estimates, $r_t = y_t - \hat{y}_t$.

3. With these residuals we test to see whether the model is significant using the F-statistic described in equation 5.19. If the p-value of our F-statistic is smaller than α_{boost}, we accept these estimates as our non-boosted model for the time series.

4. Next, we treat the residuals, r_t, as a new time series and find the optimum periodic estimate, $\hat{r}_t$.

5. With this new estimate we now make a "greedy" model by finding $\underset{\gamma}{\operatorname{argmin}} \sum_{t \in T}(y_t - (\hat{y}_t + \gamma \hat{r}_t))^2$, where T is the set of times we are using to create the estimate with. If there is concern about over-fitting, in this step use we cross-validation to identify the optimal γ.

6. We now define our new estimates of this greedy model, $\hat{y}'_t = \hat{y}_t + \gamma \hat{r}_t$, and test for significance with an F-test like we did in step 3.

7. If the model is significant, we repeat steps $4 - 6$ with the new residuals, $r'_t = y_t - \hat{y}'_t$, as the time series.

8. We continue repeating steps $4 - 7$ with the new residuals until the model is not considered significant under the F-test in step 6. At this point we consider the model from the previous iteration as our final model.

5.6 Bootstrapped Signal Synthesis

So far in our data synthesis we have neglected the effects of randomness on our estimates. For each data point we assume that there is a set of periodic components and

a noise term. This noise term will affect our estimates of the periodic components and can cause reduced performance for interpolation or prediction. There are several ways to attempt the deal with this issue. We can use noise removal procedures by filtering the data or by using data transformations such as principal components analysis [108]. Sadly, for many types of data these procedures can be impossible to implement or detrimental to the quality of the data [126]. In an effort to produce a more meaningful estimate of the missing data, we will use a non-parametric bootstrapping method to estimate the role noise plays.

We start by assuming that our time series data is sampled from a process made of periodic components and a noise term, $x_t = \sum_{i=1}^{n} P_i(t) + \zeta(t)$. As well, we assume that the samples are uncorrelated. If the samples are correlated but no new periodic components can be removed by using boosting on the residuals, it may be best to model the data with the periodic components, an ARMA process and a noise component. To gain an estimate of the distribution the synthesized data will follow, we need to model the noise process, $\zeta(t)$. To do so, we estimate the periodic components using Thomson's method and re-sample the residuals, $r_t = x_t - \hat{x}_t$.

We now introduce noise into our initial estimates of the missing data by drawing samples from the residuals for the data points we wish to synthesize. By doing this, we emulate the noise contribution to the periodic estimates for those times. For interpolation, we would add noise samples to the simple interpolation.

This process is more tricky with prediction, as it is not possible to add the residuals directly to the missing data. If we were to add the residuals to the zero-padded section to create a scheme similar to that for interpolation, we would create a large amount of bias in our predictions. This increased bias is due to our method of trying

to model the non-existent sinusoidal components in our replacements for the zero-padding. To work around this, we propose sampling the residuals resulting from our existing reconstruction with replacement, adding them to the periodic reconstruction, producing a new data set, and then computing the new periodic estimates.

By repeating this process of drawing noise samples and obtaining periodic estimates, we produce a set of random samples for each estimated time. From these samples we can derive properties of the distribution for the synthesized data. This gives us an estimate of the mean value for each time as well as confidence intervals.

Along with the confidence intervals we get for our periodic estimates from the bootstrapping, we can also estimate the overall confidence intervals for the synthetic data by modeling the residuals by a normal distribution and adding the confidence intervals for the noise and periodic terms together. With this, we can give a distinct interval at each time that we expect the missing data to have been found.

5.7 Data Analysis and Comparison

We decided that to verify the merit of these techniques, we should first evaluate their performance on simulated data made of sinusoids in noise and then test the methods on common data sets. We are interested in knowing how well the optimized Thomson inverse estimates synthesize unknown data and under what conditions these methods perform best.

We next tested the data on two real-world applications: filling a gap in weather data and predicting the commodities market price of coffee. To determine how well these methods work on non-ideal data where all of our assumptions may not hold, we decided to check the performance of each method on a realistic problem. We

examined weather data for the city of New York and tried to interpolate a large gap that could result from a weather station going down. After that, we looked at the daily price of coffee on the commodities market and checked to see how well we were able to predict the price a year into the future.

5.7.1 Simulated Data Testing

First we tested the methods with an artificial time series containing seven sinusoids in Gaussian noise with signal amplitude varying from .15 to .3:

$$y_t = .2\sin(2\pi.2t) + .3\sin(2\pi.35t) + .25\sin(2\pi.135t) + .3\sin(2\pi.305t) + .28\sin(2\pi.25t) + .15\sin(2\pi.05t) + .27\sin(2\pi.1t) + N(0,1). \quad (5.20)$$

Wanting to see how Thomson's method and the extensions we proposed worked on a well-behaved stationary series, we attempted to interpolate and predict several realizations from the series described in equation 5.20. We began with an example of interpolation and prediction of a series of $1,200$ points from equation 5.20. For interpolation we wanted to estimate the middle 100 points of the series, while for prediction we attempted to predict the final 100 points.

The first step in both methods was to determine a reasonable choice for the parameters NW and K. For this we used the naïve sphericity test and found $NW = 5$, $K = 5$ was optimal. We use the naïve test because we plan to run a large number of simulations and the increased computational cost would introduce significant delays to this analysis. We did not expect that this choice would cause a significant issue. A discussion of the merits and pitfalls of using the naïve method, as well as the details on its procedure can be found in chapter 3.

For interpolation, we started by determining the optimal cutoff value by finding

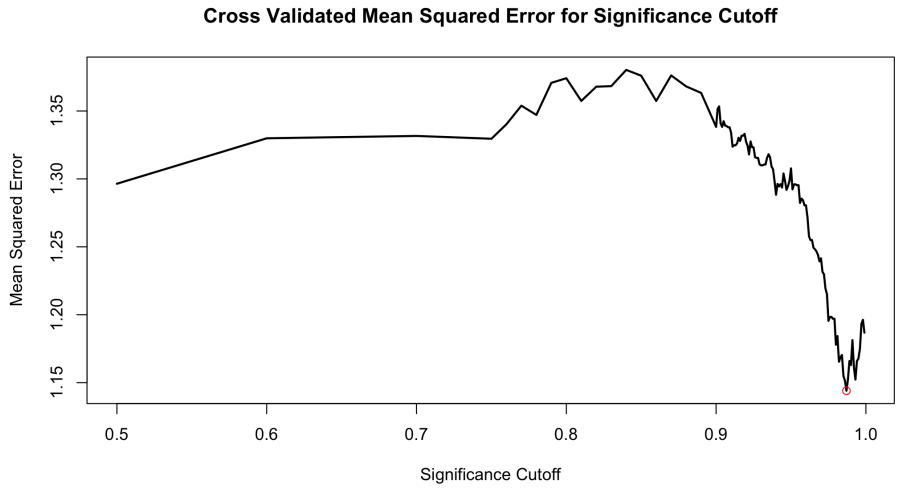

Figure 5.2: Cross-validated mean squared error of varying significance levels for interpolation of 100 points of sinusoidal data.

the minimum mean squared error across the range from .5 to 1. Shown in Figure 5.7.1, the minimum value was found to be $\alpha = .987$. We noticed that the mean squared error behaves as we expected for this series, with low cutoff values, between .5 and .9, having poor performance due to the inclusion of many falsely detected periodic components. The performance of the estimates improved as we continued to remove needless frequencies, minimizing on the optimal set, but as we continued to remove more frequencies the performance diminishes due to the removed significant periodic components.

Now with the optimal cutoff found to be $\alpha = .987$, we performed our interpolation. The interpolation, as shown in Figure 5.7.1, appeared to perform well in the gap, with less variance than the true data, as we expected with the noise levels present.

Attempting to boost the residuals, we found that there were some residual periodic

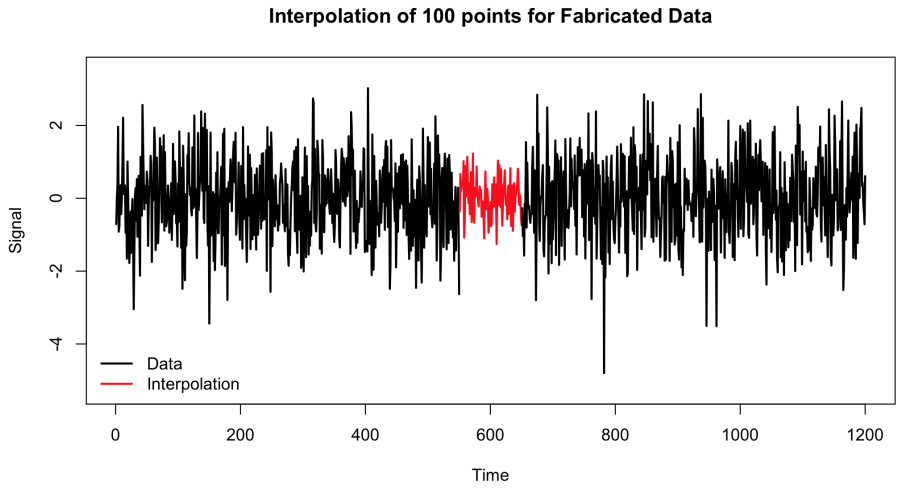

Figure 5.3: Interpolation of 100 points of sinusoidal data.

components unaccounted for. On the first iteration we added several frequencies and optimized with a greedy model with $\gamma_1 = .85$. This model, shown in Figure 5.7.1, was significant at $\alpha_{boost} = .01$. Repeating this process for the residuals from the new reconstruction, $r'_t = y_t - (\hat{y}_t + \gamma_1 \hat{r}_t)$, we found the next model not to be significant. We therefor selected the previous model, $\hat{y}'_t = \hat{y}_t + .85\hat{r}_t$, as the boosted model. This produced a slight improvement at picking up the more extreme-valued but slower periodic trend existing in the data.

Finally, sampling the residuals from the estimated series to add to the mean interpolation used gave us an estimate of the confidence intervals for the periodic error that may have existed in the interpolation. Using the residuals to model the assumed Gaussian noise for our series, we also produced overall confidence intervals for the interpolated section. For this series, the bootstrapped confidence intervals on the periodic interpolation, which we plotted in Figure 5.7.1, were extremely small,

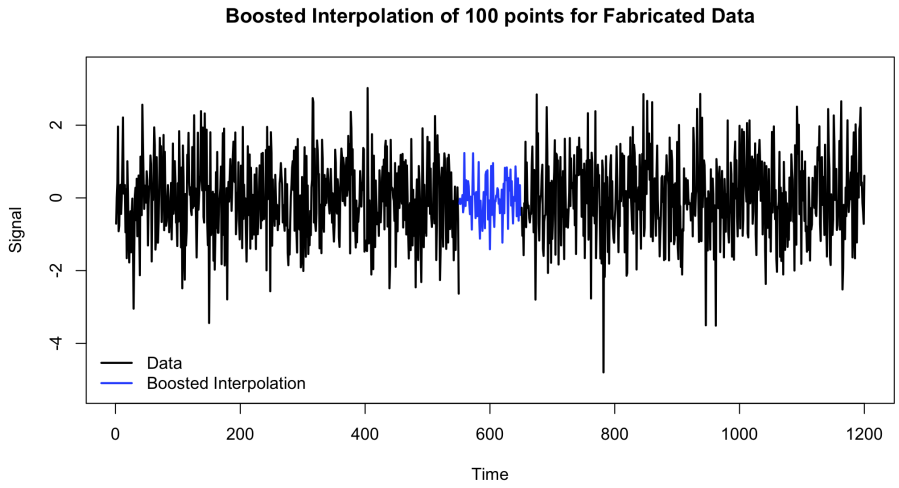

Figure 5.4: Boosted interpolation of 100 points of sinusoidal data.

as this data was strongly periodic with little residual structure or noise interference. The overall confidence bounds appeared appropriate, using $1 - \frac{1}{gapsize} = .01$ as our significance level, we saw the range match that of the surrounding data.

Comparing the confidence region of the interpolation to the data removed from the series originally, in Figure 5.7.1, we saw that no points exceeded our bounds. In addition, we saw in Figure 5.7.1 that the true periodic trends without noise tracked very closely to the confidence intervals of our periodic estimates. The periodic part of our data barely crossed the bounds on one occasion. This is as we expected, and the size of the departure was minuscule.

In Figure 5.7.1, we examined the original simple periodic reconstruction to see how well it fit within the confidence bounds that tracked the periodic components well. We saw that it performed considerably worse, rarely maintaining a similar shape. This showed that the use of boosting and bootstrapping helped to improve the estimates

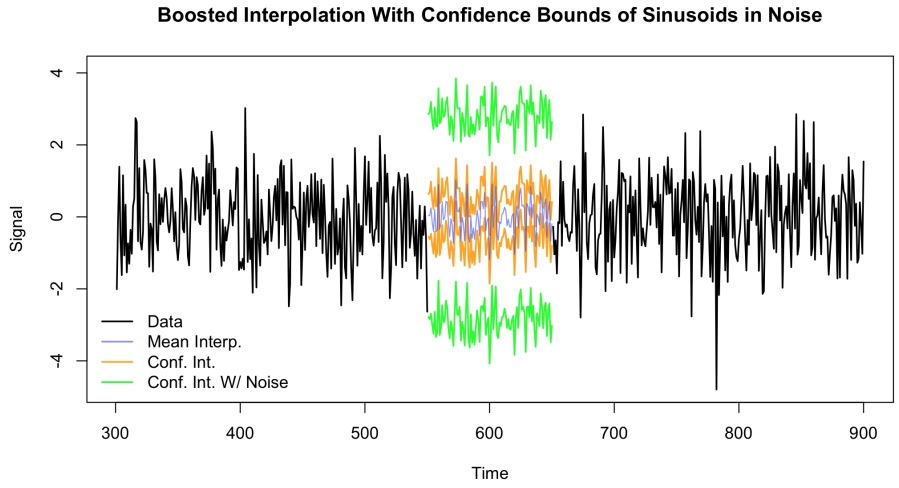

Figure 5.5: Confidence intervals for interpolation of 100 points of sinusoidal data.

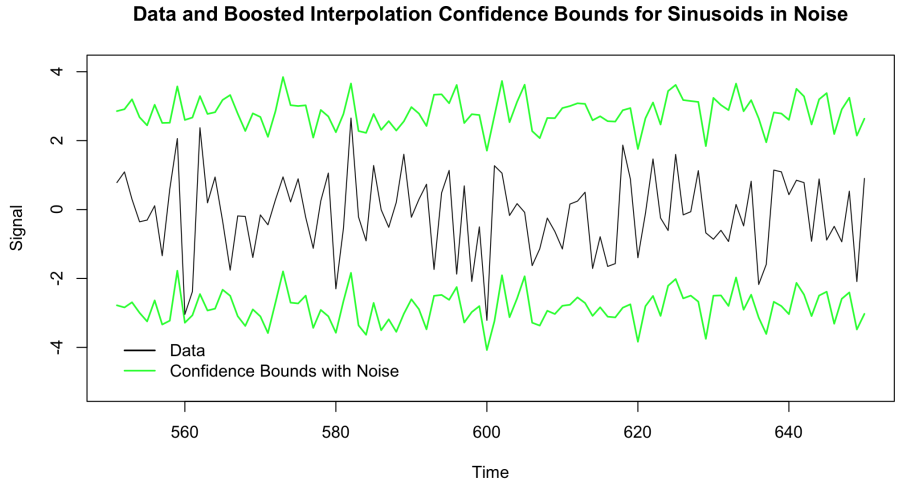

Figure 5.6: Comparison of sinusoidal data with noise to $\alpha = .01$ overall interpolation confidence intervals

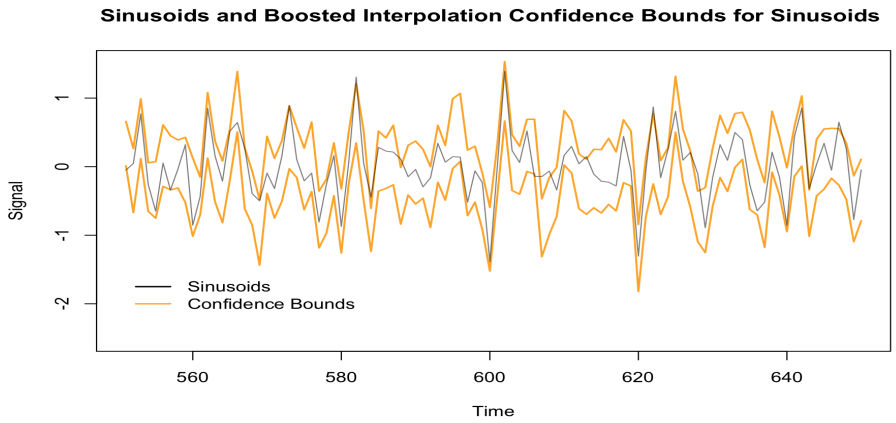

Figure 5.7: Comparison of sinusoidal data without noise to $\alpha = .01$ periodic reconstruction interpolation confidence intervals.

we provided. For this fabricated data the periodic reconstruction performed quite well and gave a useful representation of what could be found in the data gap.

Now we attempted to predict 100 data points in the future from the same set of sinusoids and noise. First we saw in Figure 5.7.1 that the optimal α is larger than before, at $\alpha_{opt} = .997$. After obtaining the optimal cutoff we made our prediction which we plotted in Figure 5.7.1. The prediction tracked well with the data and appeared to have fewer periodic terms than the interpolation. This is in line with what we expected for prediction, being more conservative than interpolation.

When modeling the residuals we found several more significant frequencies and $\gamma_1 = .94$ as the parameter for our greedy model. This model was not significant for $\alpha_{boost} = .01$. Therefore, we concluded that the original model was ideal. This may have been due to the more conservative nature of optimizing α_{boost} for prediction.

Sampling repeatedly from the residuals from our original prediction, adding them back to the reconstruction and creating periodic reconstructions of the resulting time

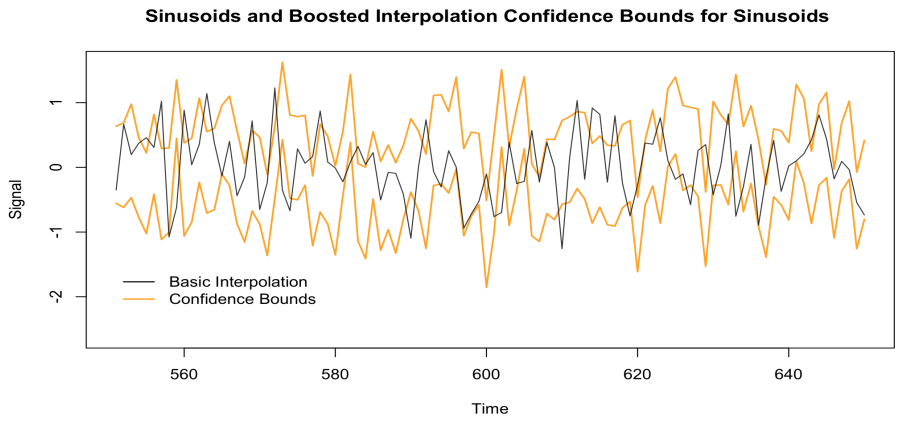

Figure 5.8: Comparison of the simple periodic reconstruction of sinusoidal data to $\alpha = .01$ periodic reconstruction interpolation confidence intervals.

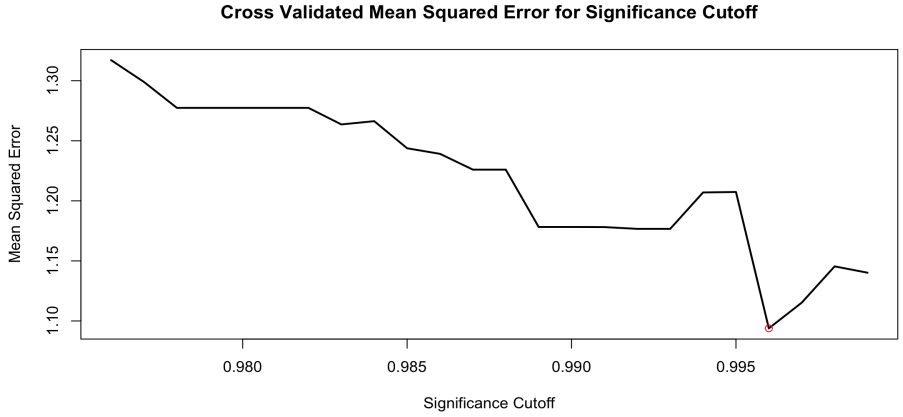

Figure 5.9: Cross-validated mean squared error of varying significance levels for predicting 100 points of sinusoidal data.

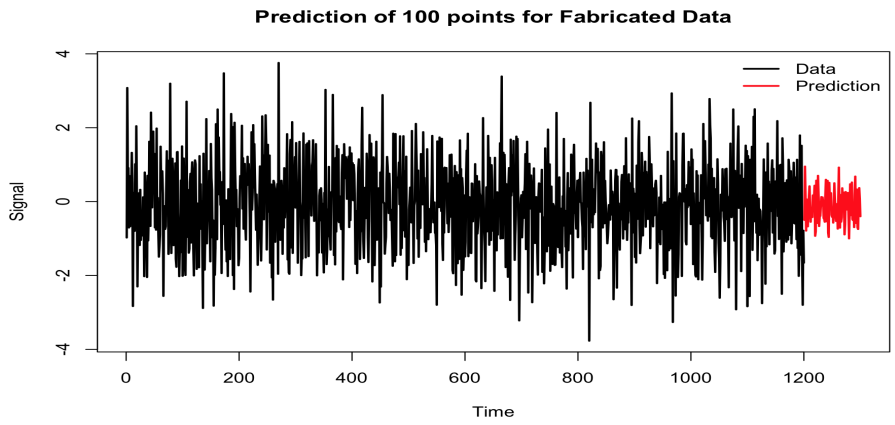

Figure 5.10: Prediction of 100 points of sinusoidal data.

series, gave us the distribution for our prediction at each time. Finding the .01 significance level confidence intervals for the periodic reconstruction with and without noise, we saw that they are slightly wider than those for interpolation and followed the data well. This is shown in 5.7.1.

In Figure 5.7.1 we looked at the confidence bounds of the prediction when compared to the true data, our prediction interval with noise contained all of the data points. As for the periodic components only, plotted in Figure 5.7.1, we see performance similar to that of our interpolation, with four occasions where the sinusoids fall outside the confidence region. Overall, the periodic estimates' bounds followed the structure of the signal well.

Following these two examples, we examined what performance gains were obtained by using the advanced methods. Using the mean squared error as the metric for performance, we performed repeated periodic reconstructions of sinusoidal data with Gaussian noise. We examined the mean squared errors of 500 repetitions of

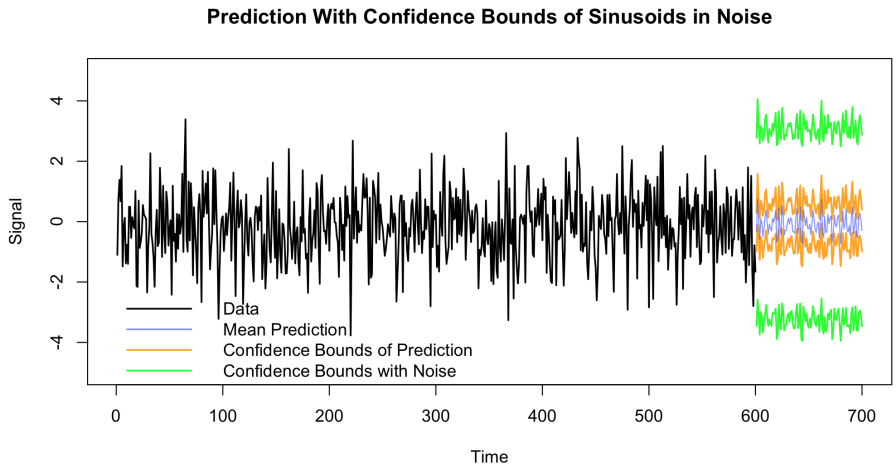

Figure 5.11: Confidence intervals for prediction of 100 points of sinusoidal data.

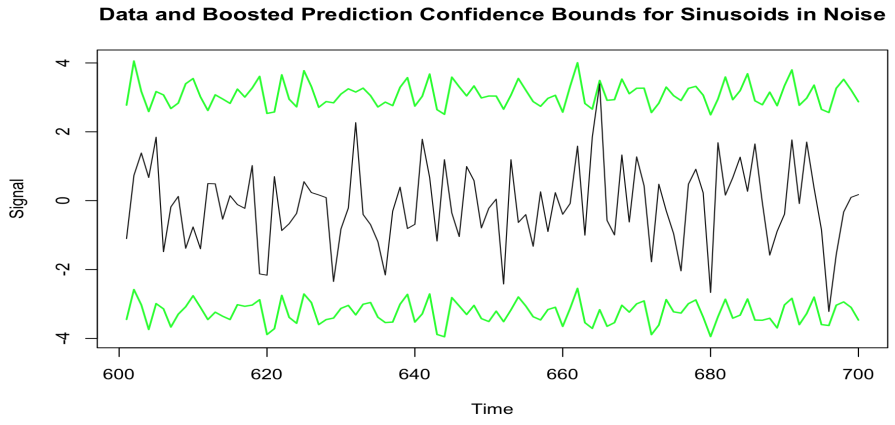

Figure 5.12: Comparison of sinusoidal data with noise to $\alpha = .01$ overall prediction confidence intervals.

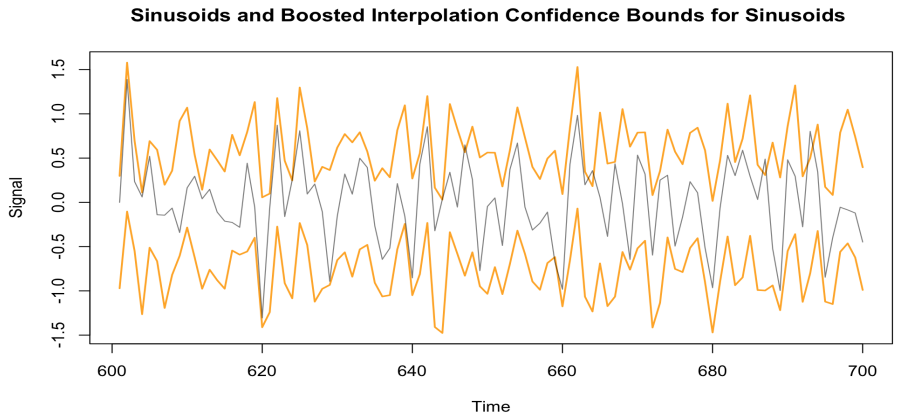

Figure 5.13: Comparison of sinusoidal data without noise to $\alpha = .01$ periodic reconstruction prediction confidence intervals.

interpolation of 100 points in the middle of $1,200$ point series for each method. The four methods used were Thomson's method with (1) optimized α, (2) optimized α and boosted residuals (significance level .01), (3) optimized α and bootstrapped modeling (60 runs), and (4) optimized α, boosted residuals (significance level .01) and bootstrapping (60 runs). We found that bootstrapping was the most significant improvement over simply optimizing for α. In Table 5.1 we saw that the boosting method offered no significant reduction in mean squared error when we tested with a one-sided t-test. This may be due to the ideal properties of the sinusoidal data. As we noted, the boosting algorithm works to model signals missed in the original reconstruction. In real-world situations, the data will not be as simple and easy to model. In this situation, the boosting method does have use.

We also examined the computational cost of each method. To do so, we measured the time to run each iteration from our performance testing to determine the average time for each method. The average times are give in Table 5.2; as we suspected, the

Table 5.1: t-test evaluating $H_0 : \mu_A > \mu_B$ for the mean squared errors of our interpolation methods.

	Method B			
Method A	Optimized α	Boosted	Bootstrap	Boosted bootstrap
Optimized α	$\hat{\mu} = 1.186223$ $n = 500$	$t = 0.0889$ p $= 0.4646$	$t = 3.6343$ p $= 0.000159$	$t = 4.0299$ p $= 3.379 \times 10^{-05}$
Boosted		$\hat{\mu} = 1.185132$ $n = 500$	$t = 3.4026$ p $= 0.0003654$	$t = 3.7766$ p $= 9.066 \times 10^{-5}$
Bootstrap			$\hat{\mu} = 1.132029$ $n = 200$	$t = 0.3009$ p$= 0.3818$
Boosted bootstrap				$\hat{\mu} = 1.126720$ $n = 200$

bootstrapping was the most costly process. When used in conjunction with boosting, the bootstrap process was considerably slower than the other methods. The boosting was not, on average, more that 15 seconds slower than simply finding the optimal α_{opt}. This can be attributed to only one iteration of residuals being reconstructed the majority of times.

Table 5.2: Average computational costs of each interpolation method.

	Optimized α	Boosted	Bootstrap	Boosted bootstrap
Time (seconds)	22.12	35.54	546.77	$1,678.77$

Prediction is a very similar problem to interpolation for each of the four methods, with the difference being that the choice of α is more conservative. We performed the same testing on each method for prediction and received similar results. The boosting method had a more significant improvement over just optimizing for α. This may be

because of the larger number of runs we were able to perform for prediction due to the lower computational costs. We still found that there was minimal improvement from use of both boosting and bootstrapping. The full analysis for each method's performance is given in Table 5.3.

In addition to modeling performance, we also examined the computational costs for the performance methods. Examining the results in Table 5.4, we still find that boosting was marginally more computationally expensive than just optimizing for α and that bootstrapping resulted in a considerable increase in cost. The boosting and bootstrapping case was still the most expensive method. We do note that the costs were two to three times less expensive than those for interpolation. This could be due to the the lower number of bins used within and the single instance of cross-validation for prediction.

Table 5.3: t-test evaluating $H_0 : \mu_A > \mu_B$ for the mean squared errors of our prediction methods.

	Method B			
Method A	Optimized α	Boosted	Bootstrap	Boosted bootstrap
Optimized α	$\hat{\mu} = 1.267972$ $n = 2000$	$t = 1.6324$ p $= 0.05157$	$t = 3.6343$ p $= 3.655 \times 10^{-05}$	$t = 1.8882$ p $= 0.04135$
Boosted		$\hat{\mu} = 1.251046$ $n = 500$	$t = 2.2646$ p $= 0.01201$	$t = 1.3733$ p $= 0.09579$
Bootstrap			$\hat{\mu} = 1.216298$ $n = 500$	$t = 0.4127$ p $= 0.3428$
Boosted bootstrap				$\hat{\mu} = 1.200918$ $n = 200$

Another pair of aspects that may affect the performance of our methods is gap size

Table 5.4: Average computational costs of each prediction method.

	Optimized α	Boosted	Bootstrap	Boosted bootstrap
Time (seconds)	6.65	15.48	288.77	586.6

and signal strength. We ran 20 simulation of interpolating and predicting sinusoidal data for varying levels of signal strength and gap sizes. The results are presented in Figures 5.14, 5.15, 5.16 and 5.17. We found that increased signal strength improved the performance of the reconstructions. This was expected, as the F-statistics are affected by signal strength. We saw that gap size had no effect. Since the data is stationary, the increased gap size will not affect the estimate. With nonstationary data or more complex signals, this may not be the case.

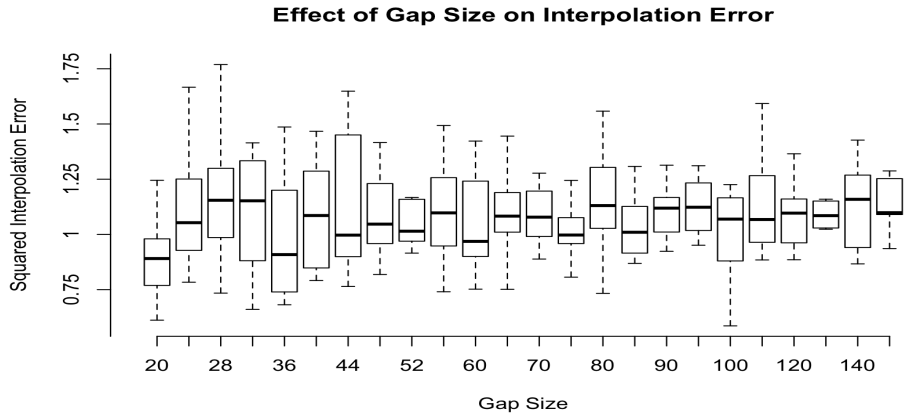

Figure 5.14: Box plots for the effect of gap size on the interpolation error of sinusoidal data with signal to noise level of .5.

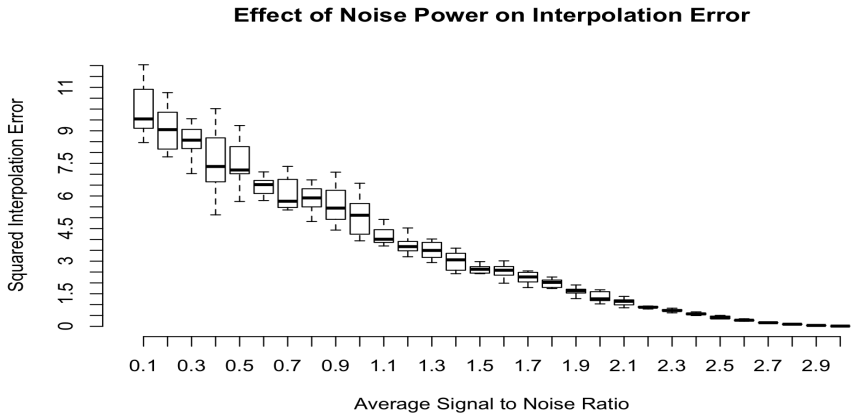

Figure 5.15: Box plots for the effect of signal strength on the interpolation error of sinusoidal data for interpolation of 100 data points.

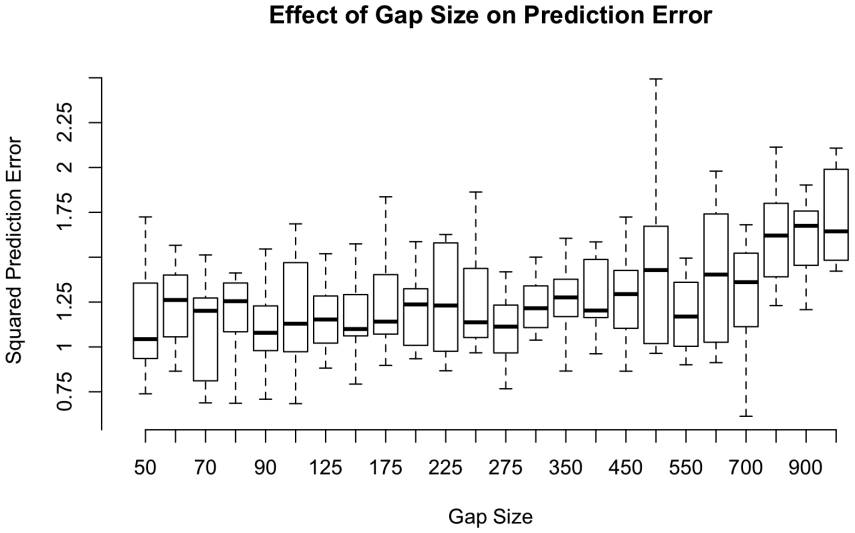

Figure 5.16: Box plots for the effect of gap size on the prediction error of sinusoidal data with signal to noise level of .5.

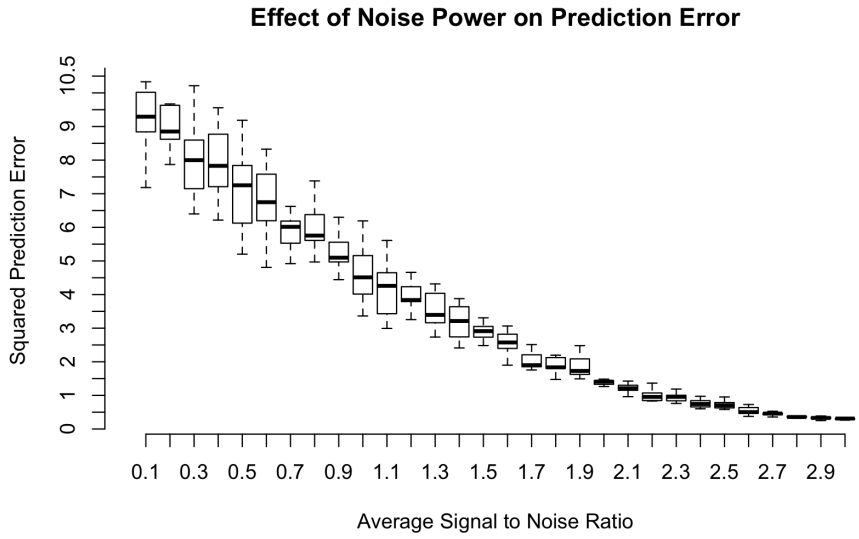

Figure 5.17: Box plots for the effect of signal strength on the prediction error of sinusoidal data for prediction of 100 data points.

5.7.2 Real-World Examples

To test these methods, we examined two real-world data sets and the potential problems that could arise. First, we examined how well we could interpolate monthly averaged New York temperature data. We removed 100 samples and attempted to interpolate them. We saw that from the boosted bootstrapped confidence intervals in Figure 5.18 there was a strong periodic trend that was maintained by our estimates. Then, looking at how well our $1 - 1/n = .01$ significance level confidence intervals performed in Figure 5.19, we saw that there was only one instance where the data falls outside the limits. This met our expectation for random chance from the uncertainty in our estimates and noise.

We next attempted to predict a year into the future for the coffee commodity

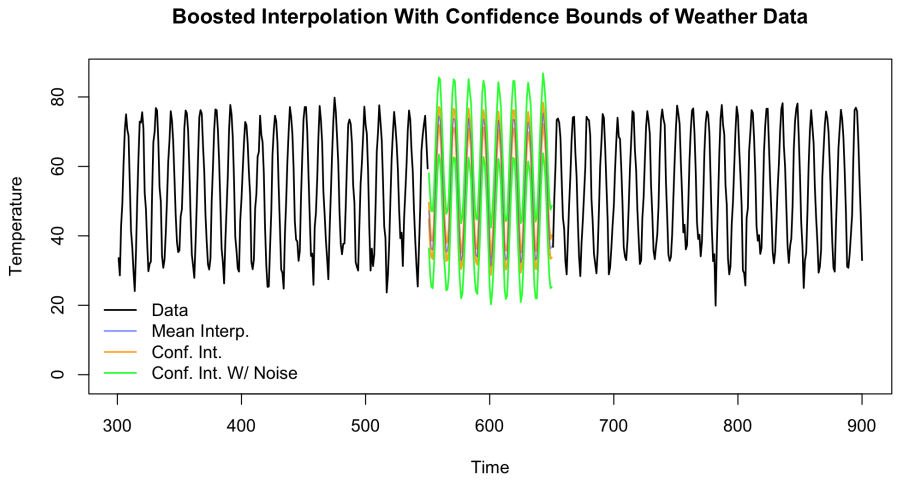

Figure 5.18: $\alpha = .01$ Confidence intervals for interpolation of 100 months of temperature data.

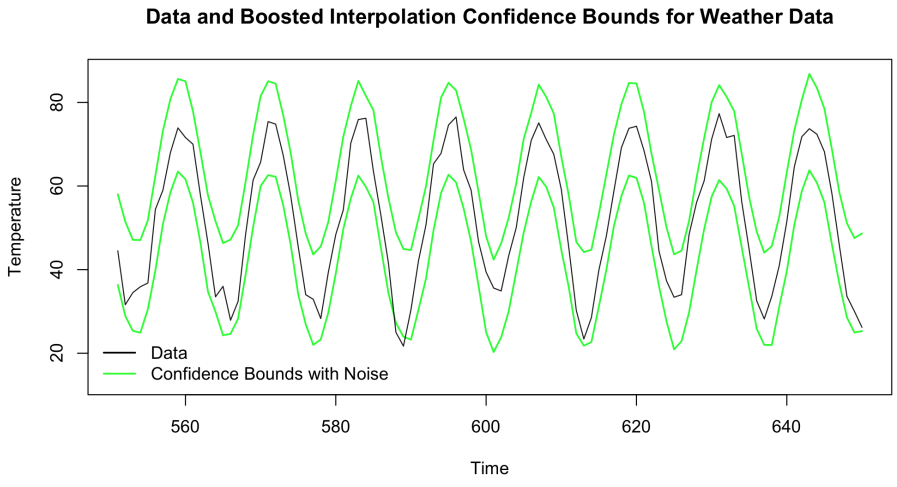

Figure 5.19: Comparison of temperature to $\alpha = .01$ periodic reconstruction interpolation confidence intervals.

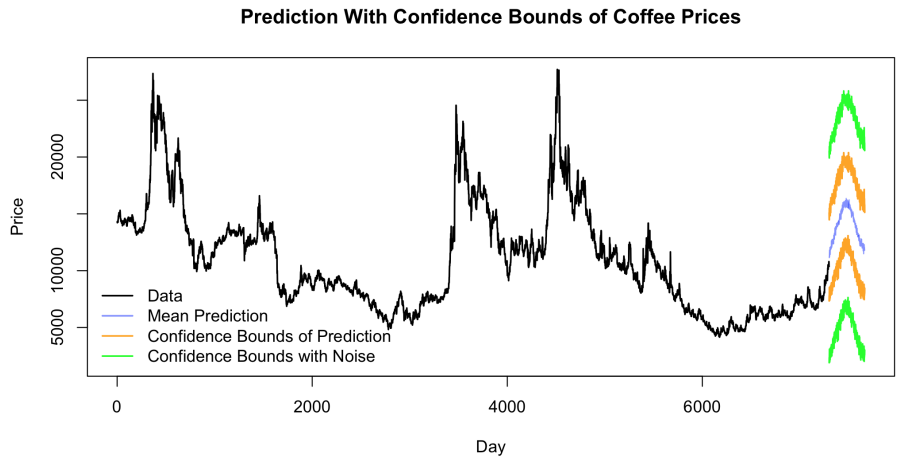

Figure 5.20: $\alpha = .01$ Confidence intervals for prediction of one year of coffee prices.

price using daily data. Examining the prediction intervals in Figure 5.20 we expected that the price of coffee would increase over the next half-year. We also noticed that the lower confidence bounds with noise are a considerably lower value than the values we had found in the data but the upper bound does not cover the range of the data. This may have been due to the residuals being from a skewed distribution. It may have been better to use a non-parametric estimate of the noise confidence intervals or to select a more appropriate distribution to model. We found in Figure 5.21 that the confidence intervals contain the actual data at all points, with the data reaching the limits on two occasions. This demonstrated that our prediction intervals were reasonable, although we may have been able to generate prediction intervals with greater performance by using another model for the residuals.

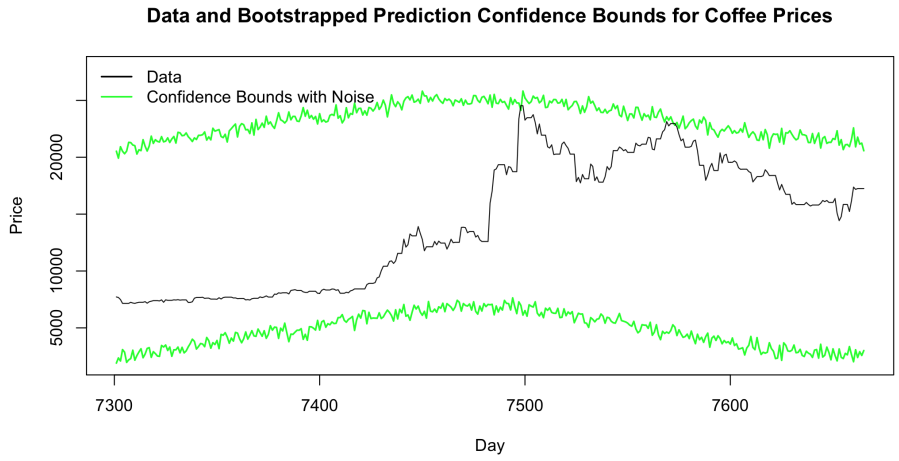

Figure 5.21: Comparison of the true coffee prices to the $\alpha = .01$ periodic reconstruction prediction confidence intervals.

5.8 Conclusions on Techniques

We have shown that these methods work well under the assumption of a time series constructed from independent noise and periodic signals. Prediction and interpolation perform similarly with slightly more conservative estimates made for prediction, as we would expect. The use of bootstrapping aids in improving the performance of Thomson's method but does so at an increased computational cost. The boosting method was not shown to create a significant improvement over simply finding the optimal α when interpolating a simple sinusoidal data series. For prediction, there was a significant improvement, but the difference in significance may be due to sample size. However, the computational cost of boosting was not nearly as large as that of bootstrapping, so boosting is the more practical method.

For use on real data projects, we believe that the development of the confidence

intervals on these reconstructions improves our ability to report on the processes governing the data. These confidence intervals followed the data well and encompassed the randomness present in the process. The mean of the bootstrapped reconstructions was also a significant improvement on simply finding the optimal α. We recommend using boosting on the original reconstructions even though few gains were demonstrated. Our justification for this is that the computational cost is minimal if you are only intending to perform a single reconstruction and is more robust at modeling the increased complexity of real data. If you were to choose one method only, we would recommend bootstrapping. With bootstrapping, you gain more information to report on your reconstruction and a significantly better estimate. The only downside is the considerable computational cost associated with bootstrapping. We would not recommend both bootstrapping and boosting to data, as, while this does give the best estimate of the series, the computational cost is far too high to justify its use.

Chapter 6

Extracating Atrial Signals from ECG Time Series

6.1 Introduction

Analysis of atrial rhythm is paramount in the identification, treatment and long-term management of atrial fibrillation [6, 108]. The main tool in the analysis of heart rhythms is the electrocardiogram (ECG), a machine that collects a set of time series measuring the electrical pulses made by the cardiac tissue observed on the skin of a patient [78]. There are commonly 12 leads observing the cardiac current simultaneously. For this reason, much of the methodology developed for evaluating ECG data is based on multiple leads being available [17].

Alternatively, in some instances where the 12 leads cannot be attached because of physical limitations or if the patient is unwilling to deal with the discomfort of the process, one or two leads are used for an ECG [91]. In most of these situations, only the first lead is reliable because of spatial changes that occur during the recording

process that can affect the signals recorded by the other leads [110]. As well, there is a larger concentration of atrial power that exists within the first lead [82]. In this chapter, we aim to produce a procedure that can perform extraction analysis similar to 12-lead methods while using only one ECG lead.

6.2 The Problem

Atrial fibrillation is a common abnormality in the heart's rhythm that can cause serious health problems, such as strokes and congestive heart failure [6]. The extraction of the atrial components of the heart's electrical pulses can help to identify atrial fibrillation and manage it as a patient undergoes treatment [95]. There are a number of common approaches to extract the atrial components from a raw ECG time series, filtering, basis expansion, and matched subtraction being some of the more common methods [15] [110].

The extraction of atrial signals can be difficult because of the ventricular pulses that also exist in the ECG data [108]. The relative amplitude of the atrial component is much smaller than that of the ventricular, and this can cause masking of the atrial processes in the time domain [82]. In addition, the noise generated by the system and methods used to perform the ECG can also be larger in amplitude than the atrial signal [82]. Another issue that arises is that there is overlap in common frequencies found in both the atrial and ventricular processes, and this overlap is not identical for all patients. The differing overlap in frequencies makes the use of simple band-pass filters unacceptable for extraction [113].

There are, however, several attributes of the atrial processes that we can use to identify and extract their signals. Our method is a version of a common basis

expansion approach, wherein we evaluate the principle components of a matrix of sample heartbeats from the ECG and determine which components are attributable to the atrial activity in the signal [108].

6.3 Advanced Principal Components Analysis

The advanced principal components analysis (APCA) method is broken down into five parts: (1) preliminary data cleanup, (2) beat extraction, (3) principal components expansion, (4) atrial component detection, and (5) component recombination. To process the data for use within our analysis, we begin by removing the mean from the data and using two six-point Butterworth filters [13] to create a band-pass filter for the region $f \in [2Hz, 20Hz]$. We now have a series in which the majority of the heartbeat power, $60 - 120$ beats per minute for a person at rest, is removed and the time series is approximately stationary, allowing for the use of spectrum estimation [113].

We next transform the time series into a matrix of heartbeats by centering equal-length sections of the series on sequential peaks of the QRS process in the heart and filling them into the rows of a matrix. The QRS process is the middle three deflections in voltage visible in an ECG for a healthy heartbeat. They are the result of the depolarization of the right and left ventricles. It is important that the beats are aligned, because if they are not, when we perform spectrum analysis on the principal components, there will be phasing problems. After creating the heartbeat matrix, we perform principal components analysis to give us an orthogonal set of components that we can evaluate for atrial signal properties.

To determine whether a principal component contains an atrial component, we use three features of the principal component: kurtosis, the eigenvalues, and the

spectrum in the range of [5Hz, 15Hz]. The first feature, kurtosis, is a strong indicator of ventricular signals. Because of the high power of the ventricle signal, the kurtosis of components that contain the ventricle process will be larger than those that lack it [41]. To utilize this feature, we use Ward's hierarchical clustering [129], which is described in section 2.9.1, to separate the components into two groupings based on Euclidean distance between kurtosis values. The cluster of components with lower-valued kurtosis will continue to be evaluated as possible atrial components.

We next evaluate the eigenvalue structure of the components. Again, from the high-powered nature of the ventricular signal, we expect that the components that contain the majority of the ventricular power will have larger eigenvalues [140]. Following this logic, it is common to classify the component with the highest eigenvalue as containing the QRS process and being produced by ventricular heart activity. It is possible though for the QRS process to split across several components and, for that reason, we use two stages of hierarchical clustering to first identify the large-valued eigenvalues that we classify as ventricular and then identify possible noise components.

In the first stage we split the components using Ward's method on the eigenvalues from the principal components analysis. From this we will have a small group with large eigenvalues, these are ventricular, and we will have a larger set of components that could be generated by noise or atrial processes. For the remaining unclassified components, those attributable to noise will have eigenvalues near zero. By using a second round of clustering on the natural logarithm of the eigenvalues, we can easily separate the noise components from the potentially atrial ones.

The final differentiation needed is to ensure that we do not misclassify moderate-powered noise components as atrial. We do this by determining the presence of signals within the $[5\text{Hz}, 15\text{Hz}]$ range. Noise components will not have any signals present. Additionally, components with signals outside this theoretical range for atrial activity will also be identified. This provides robustness against the detection of systematic artifacts with signals elsewhere in the spectrum. To test for the presence of signals in each component, we use the bootstrapped multitaper F-test. This test is chosen for its ability to identify signals well at low power and with limited sample sizes, as we showed in Chapter 4. We can see in Figure 6.1 that the bootstrapped method is able to identify signals within the components that the traditional method cannot.

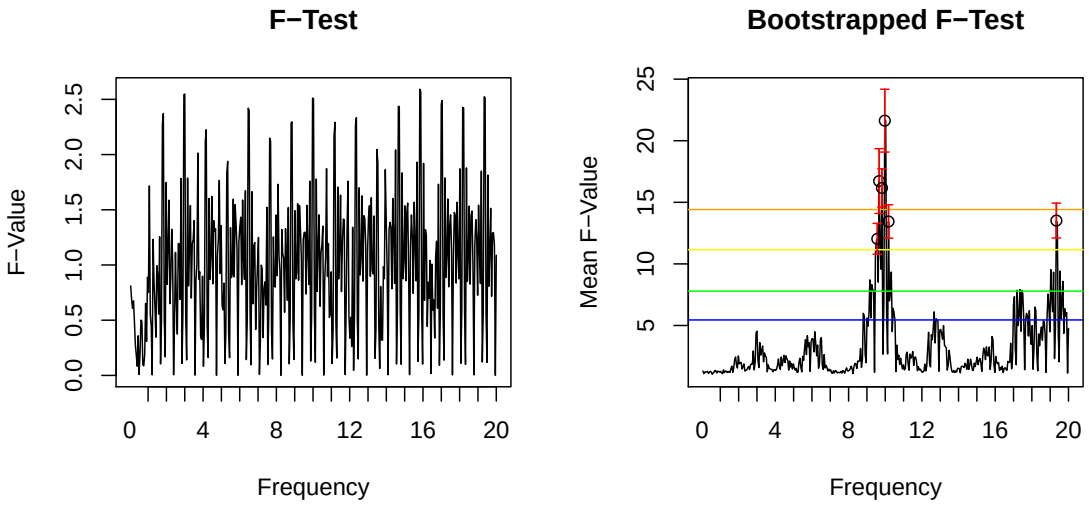

Figure 6.1: Comparison of standard and bootstrapped F-test methods on ECG extraction data.

From these three properties, we are able to classify each component. We then assume that any component that is determined to be atrial by all three properties

has been properly classified. With these classifications complete, we remove each component that is not classified as atrial, replacing it with a vector of zeros. We can now perform the inverse matrix transform to the principal components matrix to obtain an atrial process extraction from our ECG data set. We can then freely evaluate the properties of the atrial signal for signs of atrial fibrillation.

6.4 Data Study

To demonstrate the merits of this method we compare its performance with two other common methods, a more simplified eigen-value PCA-based (EPCA) method [15] and QRS cancellation from averaged beat subtraction (ABS) [110]. Evaluation of each method was based on two metrics of atrial extraction, the presence of the F-wave in the extracted signal [108] and the full removal of the QRS process [110]. We tested each method across 18 ECG signals from healthy patients. The use of healthy patients is important to control for the non-stationary effects of atrial fibrillation that can make it difficult to detect the F-wave signal [14].

The F-wave within the atrial process is a sawtooth-like wave with a frequency of $4.7 - 5$Hz [108]. We will perform a bootstrapped F-test to determine if a signal is present in that range for the atrial extractions. We will then evaluate the power of the signal for the F-wave to determine how much of the signal has been successfully extracted. The power of each signal is considered relative to the baseline of the extracted series.

To examine the removal of the QRS process, we will look at the mean of the difference between the original signal and the residual ventricular components at each QRS peak. The smaller this difference is, the lower the number of misclassifications

of ventricular components that have occurred. With these two metrics, we will be able to evaluate a proxy of both the false-detection (ventricular misclassification) and missed-detection (atrial misclassification) for each method.

The data used in this study is from the PhysioNet archive [42]. We used the MIT-BIH Normal Sinus Rhythm Database [10] from the Arrhythmia Laboratory at Boston's Beth Israel Hospital, where five men, aged 26 to 45, and 13 women, aged 20 to 50, were found to have no significant arrhythmias. For this data, there were two leads recorded and for consistency we used only lead one from each patient. Lead one was chosen for analysis because of its robustness against changes in position and orientation of the heart [17]. The data was sampled at .08Hz and 50 seconds of data was recorded for each series. We removed the first 30 seconds of the series to avoid possible non-stationarity from the person relaxing as the testing procedure commenced [108]. This left us with a 20 second long time series of each patient.

Table 6.1: Atrial extraction method comparisons

	Advanced PCA	Eigenvalue PCA	Average beat subtraction
QRS Peak	$\hat{\mu} = .473$ $\hat{\sigma} = .0075$	$\hat{\mu} = 1.836$ $\hat{\sigma} = .7536$	$\hat{\mu} = 1.391$ $\hat{\sigma} = .7817$
t-test for $H_0 : \hat{\mu}_{APCA} \geq \hat{\mu}_i$		$p_0 = 3.18 \times 10^{-7}$	$p_0 = 5.64 \times 10^{-5}$
Relative peak power $4.7 - 5$Hz	$\hat{\mu} = .0377$ $\hat{\sigma} = .0161$	$\hat{\mu} = .0097$ $\hat{\sigma} = .0062$	$\hat{\mu} = .0074$ $\hat{\sigma} = .0048$
t-test for $H_0 : \hat{\mu}_{APCA} \leq \hat{\mu}_i$		$p_0 = 8.59 \times 10^{-7}$	$p_0 = 3.08 \times 10^{-7}$
Relative run time	$\hat{\mu} = 1672.928$s $\hat{\sigma} = 619.571$s	$\hat{\mu} = .166$s $\hat{\sigma} = .020$s	$\hat{\mu} = .047$s $\hat{\sigma} = .005$s

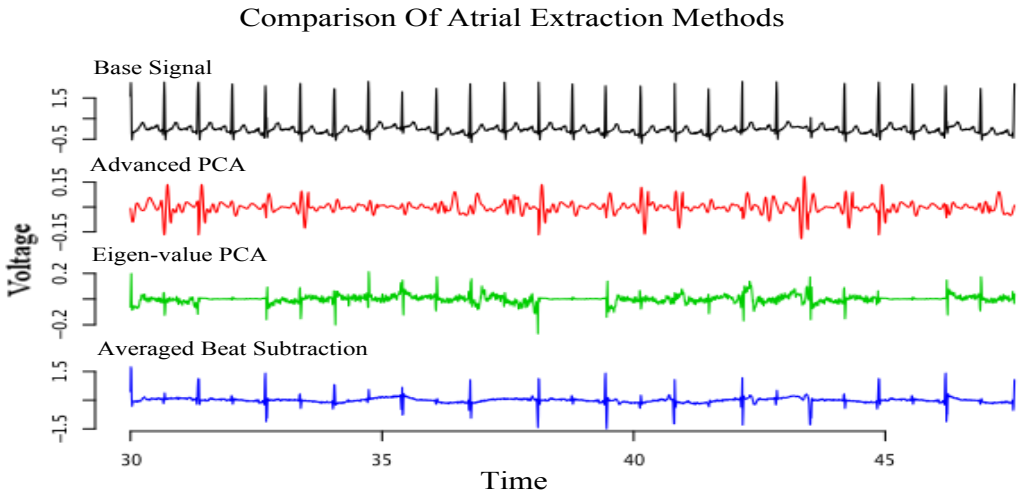

Figure 6.2: Illustration of a ECG signal and comparison of the advanced PCA, eigen-value PCA, and average beat subtraction atrial extraction methods.

Upon completion of the extractions for each series, we found that there is a significant improvement gained across both metrics when the advanced PCA approach is used. The full results of each series are given in Table 6.1. We also computed a Welch t-test for each metric, comparing our method to the reference methods, and found that the increase in performance was significant for both metrics. From graphical comparison of the extractions in Figure 6.2 and the other 17 series, we see that there is a reduced level of background noise persisting in the advanced PCA method. In addition, two other problems are found in the reference methods: after extraction, QRS peaks can persist and, there is over compensation for QRS peaks causing negative spikes.

We also wanted to see how well the new method compared to currently used methods with 12-lead data. In the 2006 paper by Langley [58] they showed that

the best 12-lead methods had a QRS peak reduction of 81.8% (.18mV remaining on average compared to .99mV originally) for the spatiotemporal QRST cancellation method. For our data we had an average QRS peak of 2.023mV. Comparing this to the QRS peak after extraction, we had a reduction of 76.6%. This is less than the best 12-lead method but much better than the other single lead methods.

Another consideration is the run time of these algorithms. The APCA method is a computationally heavy procedure, running just shy of 28 minutes on average, unlike the alternative methods that are much simpler and more efficient, with times of less than half a second. The complexity of the APCA method does require that greater care be taken for optimization; however, this was not the focus of this research. With further refinements we would expect the APCA method to come down in computer cost but not to approach the timing of the alternatives.

6.5 Conclusion

The advanced PCA method defined here for extracting atrial signals in ECGs is an improvement on the current standard methods used for single leads and is much closer in performance to 12-lead methods in performance. In the event that single leads are used in a clinical setting for the reasons of ease of use or comfort for a patient, use of this methodology is recommended for evaluating the atrial processes in the heart. The only considerations that should be noted are that this method can come at a high computational cost because of its complexity and that it would not be applicable in settings where instantaneous analysis is needed. Overall, we feel that this method is a significant improvement on standard methods and that its implementation should be considered by practitioners and researchers alike.

Chapter 7

Modeling Major Junior Hockey

7.1 Introduction

As ice hockey continues to grow in popularity [23], we are seeing a shift in the way we evaluate the quality of the players [64]. Teams, media and the public are moving away from qualitative valuation of a player's success based on personal observation and towards quantitative valuation [66]. The current surge in statistical analysis of hockey outcomes has led to improvements in data collection [65] and analysis methods [100], but as a sport there is still a lack of quantitative methods for evaluating many areas of the game [104].

Four questions in hockey that lack significant statistical analysis are of interest: How do we evaluate neutral-zone play? How can we select the optimal line combinations? Can we identify when a player is under- or over-performing during a game? And last, can we predict with any precision how a game will progress and manage the roster to optimize the potential for winning? We set out to provide solutions to these problems by (1) developing a new model for and statistics of neutral-zone play,

(2) using machine learning techniques to choose optimal lines, (3) modeling player effects on game events, and (4) applying time series methods to predict future events.

7.2 Current State of Statistics in Hockey

Like many other sports, statistics in hockey are currently in transition from simply using reports on counts of game events to using advanced statistical tools and developing exciting new metrics of in-game performance [67]. Traditionally, hockey statistics have focused on comparing the simple statistics found on the game sheet (goals, assists, etc.) of players and teams, with little focus on the effects of the other players on the ice [137].

The first advancement towards evaluating player contributions outside of these game sheet statistics was the widespread implementation of a goal differential for the time when a player is on the ice. Known as the plus-minus rating, this measurement of the difference between a player's goals for and against was first introduced at a professional level in the 1950s when Emile Francis starting using it with the Montréal Canadiens. It grew in popularity throughout the National Hockey League in the 1960s and was recorded by the NHL following the 1967 season after being introduced to the public by St. Louis Blues reporter Gary Mueller [97].

Following in the same vein as the plus-minus rating, in the early 2000s Jim Corsi, goaltending coach for the Buffalo Sabres, developed a metric reporting the differential of shot attempts by each team during a game [67]. Named the Corsi rating, the statistic has been shown to be highly correlated with puck possession [88], an area that had previously been difficult to track. This statistic is growing in popularity, with many media outlets using it as an indicator of player quality [67].

Another very similar metric proposed is the Fenwick rating, which counts shot attempts but excludes blocked shots [28]. This metric is suggested to be a closer proxy to puck possession for large amounts of data [88]. The Fenwick rating was designed by hockey blogger and engineer Matt Fenwick [28] and is demonstrative of how the majority of development in hockey statistics has come from amateur statisticians on blogs in the past few years.

Following the 2013 season, many of the top amateur statisticians in the blogging community were hired by NHL teams to expand their use of statistics [66]. As the leading innovators moved to roles behind closed doors, the role of advancing statistics in hockey fell back upon the academic world. This has been demonstrated in a 40% increase in published works on hockey statistics from 2013 to 2014 on Google Scholar. Continuing this trend, researchers from Carnegie Mellon, Brock and St. Lawrence University have continued to develop new methods, from spatial modeling of shot quality to evaluation of diversity of player nationalities on team performance [114].

One such advancement is the introduction of a logistical regression approach for evaluating player skill. Total Hockey Rating (THoR) [100], is a penalized logistical regression model for the probability of a goal being scored in the next 20 seconds that includes the majority of potential predictive variables. One of the variables used is an indicator for a player being on the ice during an event (shot, hit, face-off). After accounting for all potential effects from other variables and penalizing to minimize potential correlation of input variables, a transformation of the coefficients for player indicators is used to produce the THoR values. The transformation of $(\beta_{player} - .5) \times 80 \times 82/6$ is considered as the expected number of wins over the average player in a season, the justification being that there are roughly 80 events in

a game and 82 games in a season and it takes about six goals to ensure a victory.

Many of statistical analysis methods currently in development are not being used by NHL teams [114]. This is the result of a long-standing culture of anecdotal analysis [80]. In addition to that, the teams that are embracing quantitative methods are not providing many of the results to the academic world, in an effort to maintain their competitive advantage. In the near future, there will be real-time tracking of player movement on the ice [35] and a sharp increase in quality and availability of data for study of hockey at a professional level. The methods proposed in this chapter are designed to work within the existing framework of quantitative research while providing insights that have not been explored before in the academic domain.

7.3 Neutral Zone Play

In a recent article published in the *Toronto Sun* [104] Steve Simmons spoke of the lack of valuation of neutral zone play that is provided by the myriad of player metrics calculated today. For this reason we find it essential to introduce a way of understanding neutral zone play from a statistical vantage and to present several metrics that result from this methodology.

The neutral zone, the area between the blue-lines in the middle of the ice hockey rink, in hockey is thought of by many as the place where games are won or lost. The New Jersey Devils' famous trap system is an example of how strong neutral zone play generates success on the score sheet [29]. Outside of qualitative observations on the strength of a team, tactic, or player in the neutral zone, there is little in the way of information about the play in this area of the ice. With few recorded game events coming from the neutral zone, hits and penalties being the most common, there is

little data to employ and no identifiable event that is considered a success in the same light as a shot on net or a goal.

This leads us to evaluate what we should consider a success within neutral zone play. At its most basic, success can be defined as the ability to control which zone the puck moves to from the neutral zone. We can break this idea of success down into two independent events: either a team or player enters the neutral zone with the puck from their defensive zone and then successfully enters their offensive zone, an offensive success, or the opposing team carries the puck from its defensive zone into the neutral zone and then a team or player takes control of the puck and navigates out of the neutral zone, a defensive success.

These situations can be treated as two separate processes, and each process can be modeled as a Bernoulli trial with an estimated probability of success. Another way to think of this is as follows: If player X carries the puck into the neutral zone from their defensive zone, what is the probability of that player's team successfully carrying the puck into their offensive zone?

To model these processes we will use a maximum likelihood estimator of the probability of success in a Bernoulli trial. We consider the sum of all neutral zone successes for one situation as $Y_s \sim Binomial(n, p)$ and one game's realization as a fixed value y from Y_s. With this, we are able to estimate the probability of success, p, to be

$$\hat{p} = \frac{y}{n}, \tag{7.1}$$

where n is the number of neutral zone events that occurred during the game.

The obvious question here now is this: How do we apply this model to ice hockey over several games or even a season? For this we need only to extend this estimator

to include the larger portion of data. That is to say, for an entire season, with the games numbered $1, 2, ..., N-1, N$, the estimator becomes,

$$\hat{p}_{season} = \frac{\sum_{i=1}^{N} y_i}{\sum_{i=1}^{N} n_i}. \tag{7.2}$$

Now that we have a model to describe the events in the neutral zone, we can create a set of metrics for evaluating the play of an individual or team. The first set of metrics would be direct computation of equation 7.2 for a given type of success. This would allow us to have statistics for offensive, $\hat{p}_O$, and defensive neutral zone success, $\hat{p}_D$. These would give us an easily comparable value of each player in a league and two simple metrics for a player's success in the neutral zone.

We can take these statistics one step further to attain an overall neutral zone effectiveness score. For this we want a statistic that combines both the offensive and defensive situations and compares a player's skills to that of a league or team average. To do this, we determine the cumulative difference of the estimated probability of success over the average for the group. Thus, we define the estimated neutral zone differential (END) for player X relative to group G as

$$END(X, G) = (\hat{p}_O(X) - \hat{p}_O(G)) + (\hat{p}_D(X) - \hat{p}_D(G)). \tag{7.3}$$

This statistic gives an estimate of the success a player can be expected to have in the neutral zone. A negative value would indicate that the player's play does not positively affect neutral zone control. The opposite is true for positive scores, which would indicate that the player can be expected to control the neutral zone when on the ice.

Having these statistics is a good start to evaluating neutral zone play, but without the ability to compute them they become useless. Hockey presents several unique challenges in designing useful statistics. The free-flowing manner of play and limited

data on puck possession that is currently recorded for hockey statistics makes direct calculation of these statistics from the existing data nearly impossible.

The primary method we recommend is direct game time recording of neutral zone attempts, leading to estimates of the values for y_O, n_O, y_D, n_D. This method does pose some difficulties, as there are some neutral zone plays in hockey that may be considered non-events or may be difficult to describe as failures or successes. For this reason we define an offensive neutral zone success for a team, as a play where the puck begins in the defensive zone under their control and ends in the offensive zone under their control without the other team gaining control of the puck and leaving the neutral zone. We include in this situation the case of dump-in attempts (shooting the puck into the other team's defensive zone) that are recovered by the attacking team. In addition, non-events occur when play is stopped in the neutral zone for almost any reason. The exception is offside offensive player stoppages, which we consider failures. All other events are considered failures.

As for defensive neutral zone play, the opposite holds true; Any event where the offensive team does not have control of the puck in the offensive zone after playing through the neutral zone is a success. It does not matter which end of the ice the defending team skates into when it gains control of the puck. There may be some ambiguity when the puck becomes contested in the neutral zone, but we will maintain that this is one event until the puck leaves the neutral zone under one team's control.

Differentiating between a team's and a player's neutral zone statistic will be done by using the data for the entire game for the team's statistic and only the subset of data when a player is on the ice for the individual's statistic. This may be problematic in fast-paced leagues where the teams change players quickly on the fly, but a keen

eye and use of existing methods for tracking players can eliminate this problem.

When recording neutral zone data is not feasible, we can identify events within a game through the use of already-reported events as proxies. Depending on the league and the data available, it may possible to identify reasonable proxies of each type of event. Using the NHL as an example of the possible data available, we can use the league's online interface to obtain data, in real time, of where the puck is on the ice, who is on the ice, passing attempts, shot attempts, face off outcomes, and giveaways. Using some of this information, we can produce a decision process for determining whether an event has occurred and whether it was successful. This decision process would vary depending on the league and, as such, should only be thought of as a crude approximation for the neutral zone processes.

7.4 Optimizing Line Selection

The choices made about whom to play and what lines to use during a hockey game can have a large impact on the outcome. Historically, many of the top teams in hockey have had strong lines with a significant effect on the game [20]. Usually the problem of choosing lines is resolved using anecdotal evidence to justify a certain set of lines [74]. This practice is not particularly scientific, barely resembling qualitative analysis. This problem can be quantified by using statistics, allowing us to optimize the choice of line combinations.

No matter which metric we wish to examine over for a line combination, it is not feasible to check every potential set of lines for the team's roster. For example, if there were 16 forwards and 8 defencemen, there would be 1.7 trillion possible choices, far more than we could measure in a reasonable amount of time with standard computing

power. To make this problem more feasible, we work with a team's coaching staff to produce a set of potential line combinations.

To be able to optimize the line combinations for a team, we need to quantify success on the ice and define a metric for a particular combination's effect on this success. Success on the ice can be considered as scoring a goal on the other team's net, while we can also claim that stopping the other team from scoring is success of another type. Assuming that the actions in the current play can lead to changes in momentum that can lead to scoring within the next minute, we propose using two Boolean response variables for measuring on-ice success, goal scored by one team in the next 60 seconds and goal scored by the other team in the next 60 seconds. If we take a game and divide it into 30-second blocks, identifying which players were on the ice in that period, we can now use this data as indicator variables for a regression problem.

Using logistic regression we can identify a player's effect on the probability of goals both for and against during a game. If we also include the line combinations as two- and three-player interaction terms, we now have the full structure we expect during the game. This could also be extended to include all five-player combinations or the interactions of the two- and three-player interaction terms, but we assume that these high-level interaction terms are negligible. We are also concerned with aliasing and misleading results when introducing higher-order interaction terms due to the limited amount of data available during the season. It may be possible if teams were consistent across multiple seasons and we used smaller time intervals to obtain reasonable estimates of the higher-order interaction terms.

We can also include goaltender terms in the analysis if that data is available. The

choice of goaltender will affect the goals against that a team receives in a game and in some situations they can be involved in offensive contributions. We will include the goaltenders in the model but not in the decision making for line combinations. In situations where no clear starting goaltender is decided for a team, the inclusion of the goaltender in the decision making from this method may be beneficial.

It is worth noting that we assume that the teams are at even strength during the 30-second intervals used. In many situations where one team is not at even strength, the line combinations that are used do not match with what could be expected during regular play. Additionally, the distribution of forwards and defencemen may not be three-to-two as we are modeling here. For these reasons, we exclude the 30-second blocks where penalties occur during this analysis. These blocks are used in section 7.7.5 when attempting to predict trends within the game.

Now, for each set of line combinations, we can perform a logistic regression for both dependent variables and obtain estimates of the effects of each line through their coefficients. Then for the probability of goals for we have,

$$P_{gf} = L(\beta_0 + \sum_{i=1}^{18} \beta_i pl_i + \sum_{j=1}^{3} \gamma_j d_j + \sum_{k=1}^{4} \zeta_k f_k + + \sum_{l=1}^{2} \gamma_l g_l)), \tag{7.4}$$

where pl_i is an indicator variable for each player on the ice, d_j and f_k are the indicators for the forward and defensive lines on the ice, g_l are the indicators for the goalies and L is the logit function.

We would like a pessimistic estimate of the overall effect each possible line combination can have on the game. Assuming that the coefficients from our regression are normally distributed, we can estimate the lower α confidence limit for the coefficients on the goals-for model and the upper $1 - \alpha$ confidence limit on the goals-against model. If we were to look at the pessimistic net difference in effect each line has on

the goals-for and -against probabilities, we will have a measure of the effect on success that the potential line combination has.

$$\begin{aligned}\Delta_\alpha = (&\sum_{i=1}^{18}(\beta_{i,gf} + \Phi(\alpha)S(\beta_{i,gf})) + \sum_{j=1}^{3}(\gamma_{j,gf} + \Phi(\alpha)S(\gamma_{j,gf})) \\ &+ \sum_{k=1}^{4}(\zeta_{k,gf} + \Phi(\alpha)S(\zeta_{k,gf}))) - (\sum_{i=1}^{18}(\beta_{i,ga} + \Phi(1-\alpha)S(\beta_{i,ga})) \\ &+ \sum_{j=1}^{3}(\gamma_{j,ga} + \Phi(1-\alpha)S(\gamma_{j,ga})) + \sum_{k=1}^{4}(\zeta_{k,ga} + \Phi(1-\alpha)S(\zeta_{k,ga}))).\end{aligned} \tag{7.5}$$

Determining Δ_α for each combination of lines in our potential set, we can identify the optimal combination by the one with the largest Δ_α. Teams can adjust the significance level, α, used for the pessimistic difference depending on the situation they are in. We recommend using $\alpha = .1$ in most situations in which we would like a proven productive line combination. Line combinations with little data will have larger variation on their coefficients, causing new lines to be undervalued at $\alpha = .1$. With large changes in personnel available to a team, it may be advisable to relax the significance to greater than .1 or in situations where a team has greater prior success and little change, it may be advisable to use a smaller value.

Further analysis of the role of α can be performed. Specifically, one could create a framework or function that gives a choice or range for α depending on the factors at hand. Some of these factors are: games played in the season, strength of opponents, availability of players, days until next game, team standings, and importance of the next game. This is not the focus of this study and is considered as potential future work.

To mitigate the over-fitting bias based on line combinations that have larger amounts of time together, we propose using a bagging majority decision algorithm

to select the optimal line combination. In a bagging majority decision algorithm, we repeatedly sub-sample the available data with replacement and identify the optimal line combination for that sample. The line combination that is most often the optimal line is the one we select as the truly optimal line.

The size of the sub-samples will depend on the amount of data available. Ideally, one can use a small portion of the data while retaining the significance of the model. By using smaller portions, we can perform more repetitions without fear of large amounts of duplication. The resulting collection of optimal models for the sub-samples can offer some insight to the team by giving a clear best combination or possibly a subset of combinations that perform similarly well. From this information the coach can tweak the line combinations to best fit the personnel available to them on a given night from the pool of preferred choices.

7.5 In-game Player Monitoring

Many of the methods designed to aid in the evaluation of a player's impact on the outcome of a game are applied historically, in an offline context [65]. It is conceivable that simple count statistics, such as shots, saves, etc., could be reported as the game progresses, but these are not relative to how we expect each player to perform, or provide unsupervised continuous monitoring. In an aim to develop a straight forward and visually compelling method to provide real-time in-game player monitoring, we look to methods from traditional quality control theory.

An appealing method for providing this real-time reporting is the exponentially weighted moving average control chart (EWMA) that we described in section 2.9.2.

An EWMA chart reports the exponentially weighted cumulative standardized deviation of a player's stat values during a game to their expected value. As an example, a player's EWMA value for their Corsi rating is

$$EWMA_{Corsi}(t) = \lambda \frac{T - Corsi(t)}{S} + (1-\lambda) EWMA_{Corsi}(t-1), \tag{7.6}$$

or equivalently,

$$EWMA_{Corsi}(t) = \sum_{i=1}^{t} \lambda(1-\lambda)^{i-1} \frac{T - Corsi(i)}{T}, \tag{7.7}$$

where T is the expected or target value, λ is the weighting for the data collected at time t, usually selected to be .2 and S is the standard deviation of a player's Corsi rating.

The value of our EWMA statistics will rely on a choice of T and estimation of S. Depending on the outcome desired and the stationarity of the player's statistics throughout the year, the target value can be the mean of the last stationary section of games or an entire season if the individual's play has not changed significantly throughout. The standard deviation estimate should be made from the same set of data as the mean. Another option for the target value is to choose a value that is desired for the player to meet.

Then under the null hypothesis that $H_0 : \mu_{Corsi} = T$, we have

$$EWMA_{Corsi}(t) \sim N(0, \sqrt{\frac{\lambda}{2-\lambda}[1-(1-\lambda)^{2t}]}). \tag{7.8}$$

It is important to note that we are standardizing the EWMA metric to allow for reporting of values that are comparable across players. We can now test at any point whether their Corsi rating follows the null hypothesis or is significantly different. For this we recommend using $\alpha = .05$ significance limits, ± 1.96 (or ± 2 for simplicity in reporting to the public [69]).

7.6 Predicting Future Trends in Game Play

Another tool we can use to assist with in-game decision making, along with player monitoring, is to provide predictions of future trends within the key metrics of game play. The ebb and flow of every hockey game is different, with many unique situations presenting in each one. The objective would not be to predict these situations but to model the controllable effects and attempt to determine whether temporal correlation exists in the resulting residuals. If we can model the residuals with a predictive model, we can determine potential future values or ranges for key statistics and use the known controllable effects in the model to maintain a statistically advantageous position in the game.

After identifying the optimal line combination for the game, we can use this same pool of variables to model the variable of interest. In addition, we may want to include uncontrollable variables such as opponents, penalties and game location to further explain the outcome of the game. Once we have the selected model for the metric, which could be determined by any reasonable method (stepwise, stagewise, LASSO, etc.), we will investigate the residuals from this model. From organizing the variables into contiguous blocks of 30 seconds, we have a time series for our residuals.

To model the temporal effects in our residuals, we will use Thomson's periodic reconstruction method, for prediction as described in Chapter 5. Using the bootstrapping and boosting extensions to this method, we can obtain a distribution on the predictions at each time step and, in addition, on the potential randomness that exists in the game. Using these distributions for the predicted values, we can give with relative certainty a prediction of how the flow of the game will go outside of player choices and uncontrollable variables. We can use this to ensure that we play

the right people for the situation. For example, when the other team is expected to play poorly, we may put on strong offensive players to try to exploit their diminished play. As the choices a team makes will affect the future outcomes, it becomes important to update the predictions often.

7.7 Data Analysis: Kingston Frontenacs

To gain better insight into the use of statistics in hockey and obtain data to test our methods, we worked with the Kingston Frontenacs of the Ontario Hockey League. Acting as the team's analytics group, we were able to identify the needs of the team and areas, we could improve the quantitative methods used in decision making.

7.7.1 Data Collection

The first task we needed to address was how to acquire data from the games. The data we required was shot attempts, neutral zone events and the active players for each event. Initially we thought that this could be accomplished with hand written reports. While during our time with the team we did see rival teams' statistics groups use hard-copy reporting, we found this to be far too slow for the rate at which the games produce data. In addition, the time required to transcribe hundreds of events per game made this impossible.

Aiming to streamline our data collection, we designed a JAVA program to easily record game events. Splitting the data collection into three tasks (Shots, Neutral Zone, and Shifts) with separate collection screens, we were able to vastly increase the velocity of data we could process. The three panels used in the latest version of the

JAVA program are shown in Figures 7.1, 7.2, and 7.3.

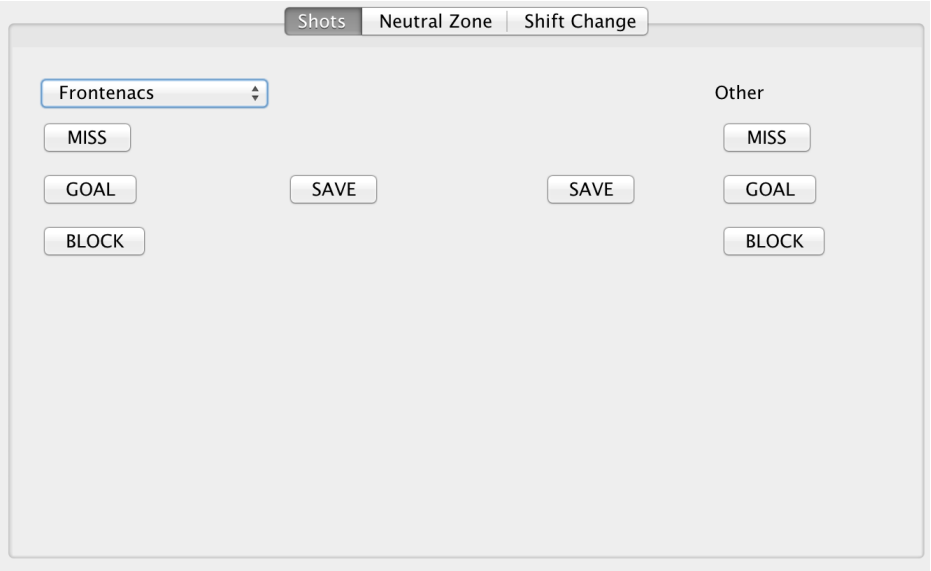

Figure 7.1: Shots panel from data collection program.

The JAVA program stores the data collected in Microsoft Access database (.db) files, which we were easily able to manipulate in *R*. To fill the three tasks required to collect data at each game, we hired undergraduate students from the Queen's University Department of Mathematics and Statistics. For home games, we would sit in the press box above the game with a clear view of all play. This allowed us the ability to follow the play and collect data effectively. The away games were a much more difficult task. After collecting the scouting film (a copy of the game recorded by the home team) from the Frontenacs, we would watch the games in classrooms. Several data quality issues did arise when we used this system. First, the video quality was highly variable, with portions of games missing, poor definition, or no audio. This took some getting used to and required that some games be watched more than once

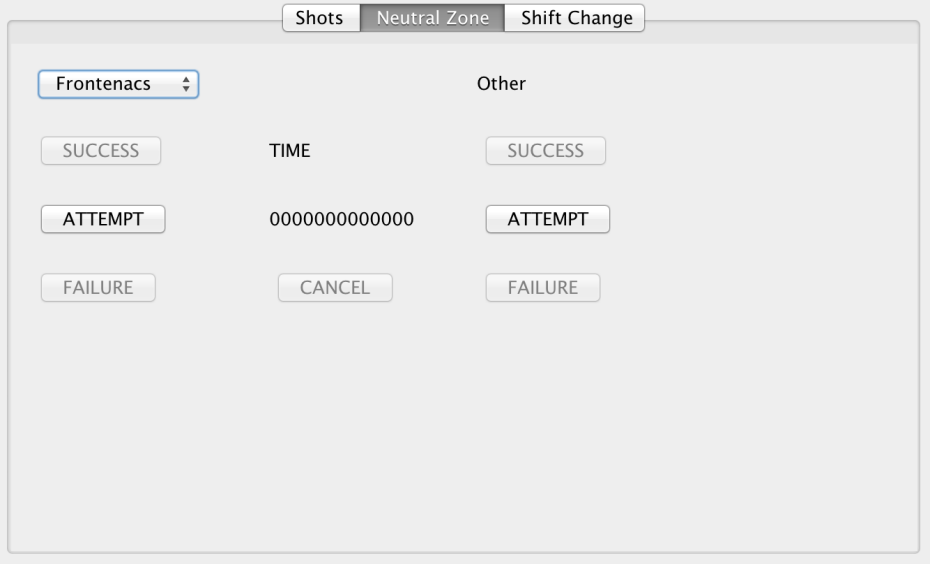

Figure 7.2: Neutral zone panel from data collection program.

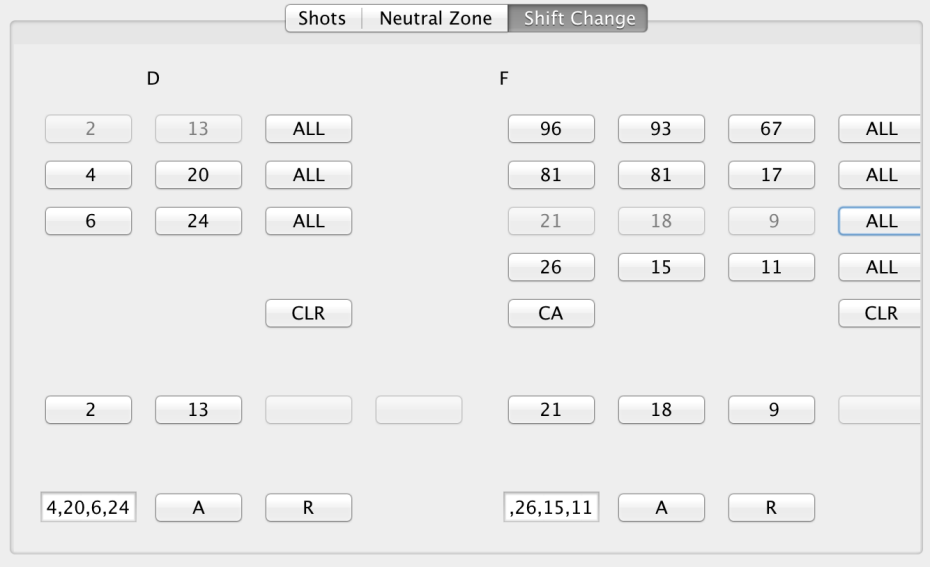

Figure 7.3: Shifts panel from data collection program.

to ensure that the data was collected correctly. The other major problem was that the film followed the movement of the puck, making it difficult to track which players were on the ice at all times. The assumption that players maintain set lines during a game is not reasonable in many cases, so missing data was an issue. This again was dealt with by watching the video multiple times, increasing training of data collectors and having supervisors help to ensure that events were not missed.

Our first task was to provide summary statistics to the Frontenacs on player Corsi ratings and other counting statistics. Writing our own code to evaluate the game data collected in R, we were able to quickly report the information of interest to the Frontenacs. As the season progressed, we continued to add more simple statistics to our reports, including shooting percentage when a player is on the ice and special teams statistics. On the basis of the feedback we received, we also provided information on recent player performance (last five and ten games) in addition to season totals.

7.7.2 Neutral Zone Statistics

After satisfying the reporting interests of the Frontenacs, we began to evaluate some of the methods we proposed in this chapter. To test these methods, we transformed the data into entries that are cumulative statistics for 30-second blocks of data. The variables included are described in Table 7.1. Using this data we were able to investigate the relationship between goal production and neutral zone statistics, player contributions, and temporal trends.

Goal production and stopping the other team are the main objectives in hockey and can be considered the most important statistics for a game. We wanted to test whether offensive and defensive neutral zone success were correlated with either

Table 7.1: Variables used in statistical modeling of hockey

Variable	Description
Goal-O	Indicator variable for the event that a goal is scored by the Frontenacs in the next minute.
Goal-D	Indicator variable for the event that a goal is scored by the opposing team in the next minute.
Corsi	Shot attempt differential during the 30-second interval.
NO	Difference between the Frontenacs' neutral zone successes and failures during the 30-second interval.
ND	Difference between the opponents' neutral zone successes and failures during the 30-second interval.
Pen	Integer variable for the number of players below even strength a team has on due to penalties being served during the 30-second interval. (it can start or end in this block and still be considered occurring).
t	Start time for the 30-second interval.
Player number	The indicator variable for each player on the ice during the 30-second interval.

team's goal production to demonstrate that teams hoping to win should consider using these statistics. To test this, we performed two multivariate logistic regressions using Goal-O and Goal-D as the dependent variables and Corsi, NO, ND, and Pen as the pool of independent variables.

Testing these two models, we found that neutral zone offensive success, Corsi rating, and penalties were significantly correlated with the Frontenacs' goal production. Neutral zone defence was the only variable for offensive goal production that was not significant. Conversely, we found that neutral zone defence was significant in the model for the opponent's goal production, with Corsi rating and penalties still being significant. We see, when we look at the regression model summaries in Table 7.2, that the coefficients for the neutral zone statistics are positive when significant. This aligns with what we expect: that more neutral zone success leads to more goals

scored.

Table 7.2: Summary of logistic regression model for goal production, including p-values for the hypothesis $H_0 : \beta = 0$

	Dependent variable	
	Goal-O	Goal-D
Corsi	$\beta = 0.372$ $p = 0$	$\beta = -0.451$ $p = 0$
NO	$\beta = 0.150$ $p = .0096$	$\beta = 0.059$ $p = .1857$
ND	$\beta = 0.098$ $p = .0570$	$\beta = 0.173$ $p = .0023$
Pen	$\beta = 0.159$ $p = .0338$	$\beta = -0.373$ $p < .0001$
Constant	$\beta = -2.889$ $p = 0$	$\beta = -2.856$ $p = 0$
Observations	$7,200$	$7,200$

We also wanted determine whether there is a relationship between Corsi rating and the neutral zone statistics. We found that the Corsi rating is significantly correlated with both neutral zone statistics. This also makes logical sense, as you need to get the puck into the zone to have a chance to get a shot on goal. Now we must ask ourselves: Do we need these extra statistics if we can monitor the Corsi rating? We believe so. By following the argument used to validate the reporting of the Corsi rating in addition to goal production when they are highly correlated, we argue that the addition of these statistics gives more information about what is going on within the game. If we were to report just the Corsi rating, we would not know what neutral

zone actions resulted in that Corsi rating. We do not go so far as recommending replacing the Corsi rating with our neutral zone statistics; we recommend only that they both be used.

The END statistic was not used in the previous modeling but we wanted to investigate its significance. To first test the significance of this statistic on the team level, we examined the correlation between the team's END statistic and the outcome of the game. Regressing the Frontenacs' END statistic on-to the outcome of the game, we found that there was not a significant relationship (p-value for the coefficient of .226). The logistic regression model did have a coefficient of 3.01, which makes sense, as for games in which a team has a positive END statistic, the team is more likely to win. The lack of significance may be attributable to the single season of data (60 games of data). This does not validate the use of the END statistic for teams and without more data or data on other teams, we are unsure of its true effect.

To check of the validity of the END statistic for players, we examined the perceived quality of the players with the highest END statistics. The logic behind this was that players who are regularly in the top two lines (six forwards and four defense) would be the highest-quality players and should perform best relative to the rest of the team.

To test this, we analyzed the hypothesis that each player's average season END statistic was greater than zero, i.e., $H_0 : E(END(player, Frontenacs)) \leq 0$, $H_1 : E(END(player, Frontenacs)) > 0$. Using a one-sided t-test and reporting the p-values of each player in Table 7.3, we found that six of the eight players with p-values below .01 were players on the top two lines. We then performed a logistic regression examining the effect of the p-value for each player on his position in the lineup. We

Table 7.3: Summary of player END hypothesis tests

Player ID	Top two lines	p-value
17	Yes	$< 10^{-19}$
22	Yes	$< 10^{-19}$
4	No	5.29×10^{-11}
19	Yes	5.26×10^{-8}
16	No	4.34×10^{-7}
11	Yes	3.83×10^{-6}
21	Yes	8.61×10^{-6}
12	Yes	2.24×10^{-3}
14	No	0.010
1	Yes	0.083
20	No	0.124
13	No	0.936
3	No	0.948
5	No	0.970
6	No	0.977
9	No	0.996
7	Yes	$> 1 - 10^{-5}$
10	No	$> 1 - 10^{-6}$
8	No	$> 1 - 10^{-8}$
15	No	$> 1 - 10^{-10}$
18	Yes	$> 1 - 10^{-14}$
2	Yes	$> 1 - 10^{-15}$
Log. regression:	$\hat{P}(\text{Top two lines}) = L(.59 - 1.59 \times \text{p-value})$	
p-value for coefficient significance:	.0947	

found that there is significant evidence (at $\alpha = .1$) that the lower the p-value, the greater the players probability of being on the top two lines. If we believe that the team has a strong understanding of player quality, we can think of a player's position in the lineup as a proxy of overall skill. From this analysis, we believe we have shown the validity of the END statistic for measuring some part of a player's skill set.

7.7.3 Line Selection

Next, we want to determine the applicability of our method of optimal line selection. This analysis was done retrospectively on the season to determine which line combination would have been a strong candidate to use during the playoffs given the players available at the end of the season. Selecting line combinations used in the final 10 games of the season and first round of the playoffs as the potential set, we were able to obtain a manageable set of 16 line combinations. We did not include indicator variables for the goaltenders in our model. This is due to the data not being readily available to us at the time of analysis and that for most of the season, including all games in the playoffs, the Frontenacs exclusively played their starting goalie.

Using the entire season as our data set, we performed 50 repetitions of 20% sub-samples. When we examine the histogram of the choices for each sub-sample in Figure 7.4, we see that line combination 14 is the optimal choice, with selection in 22 subsets out of 50. We repeated this at different levels of confidence interval to see the effect of variability on our optimal line combination. As we selected a larger value for α, we found line combination 16 to be the most often selected set of lines. At $\alpha = .45$, as is shown in Figure 7.5, we found that combination 16 was chosen 23 times, compared to nine now for combination 14.

Examining these two combinations, we see that combination 16 contained players who played considerably less frequently and had much higher variance on their regression coefficients. These players played more commonly in games against easier competition, not in the "must-win" games. We would not recommend using combination 16 in an important game as the lines are unproven. We would instead recommend combination 14.

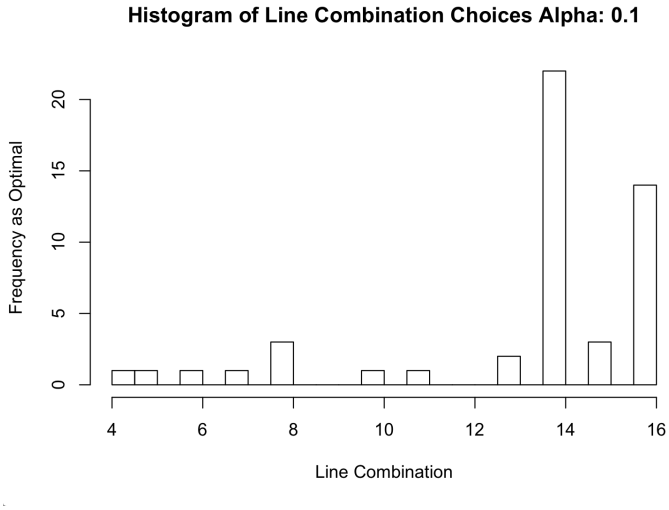

Figure 7.4: Histogram of optimal line selections for 50 20% sub-samples at $\alpha = .1$.

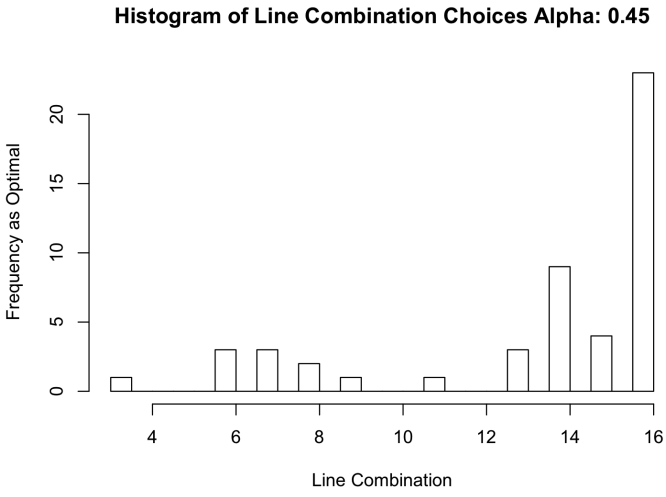

Figure 7.5: Histogram of optimal line selections for 50 20% sub-samples at $\alpha = .45$.

We wanted to know how close to optimal this line combination is. To gain some insight, we asked the Frontenacs' coaching staff to evaluate line combination 14 and

give their opinion. They stated,

> That line up you selected was close to our best line up throughout the season.

They went on to give their choice of top line for the playoffs, which was combination 8 of our possible sets. We found that the coach's choice was selected three times at $\alpha = .1$ and once for $\alpha = .45$. It is worth noting that both the coach's and our analysis selected the same optimal defensive pairings.

There were two problems with the coach's forward line selections within this analysis. First, because of injuries, the skaters selected for these lines played very few games together throughout the season. This created very high variances on the forward line interaction terms. Second, the majority of games played together by the top two lines were during the very tough playoff series against North Bay. The poor results in these games led to a large bias in the coefficients for these lines.

The difference in line selections demonstrates a flaw in the context of this analysis. We want to provide a proven (shown to be significant) set of lines that provides the best chance of winning. Many combinations may be untested and can be overlooked by this method. Another major problem is the difference in the amount of data available for all line combinations and the potential bias of the results from games in which uncommon line choices were used. Line combination 14 was identified using the entire season of data, which makes it much closer to the best line combination of the Frontenac's season, as is reflected in the quote from the coaching staff.

We stand by our method for situations where a team wants the optimal low-risk line choice. Adjusting α, we could get a riskier choice, offering higher variation but also possibly higher gains. Due to the affect of the choice of α, we do think

that improvements could be gained by developing a decision making framework or function for identifying the correct α value for a team's situation and coach's mindset. Additionally, more variables could be introduced to improve the accuracy of the model to a specific upcoming game. Some variables that could be included are indicators for home games, opponents, opposing goaltenders, game significance, and time since last home game. Lastly, to combat the issue of small samples of combinations we can take a subset of the data that would better represent the play of the line members together. However, this would introduce bias in our choice of optimal combination.

7.7.4 Player Monitoring

To demonstrate the utility of EWMA chart monitoring that we discussed in section 7.5, we will show two ways to employ this method for easy reporting to teams. First, we will show a retrospective view of a player across a game so that we can see how their play progresses and observe areas where we could make in-game decisions to improve our potential outcome. This can be used as a teaching tool after games to help players identify areas in their game where they can improve. This can also be used by coaches to help identify player performance trends at certain times during a game. In fact, a more simplified control chart system was used during our reporting of player performance throughout the season. We will also demonstrate a method for in-game monitoring of players. This can be used for on-the-fly decision making by coaches and can provide feedback to players for quick correction of their play.

Using the data collection methodology described earlier, we can update the EWMA statistics for each new time period throughout a game. With this data we can plot entire games, using time as the x-axis and the EWMA statistic as the y-axis. Putting

in the cut-off limits for 2σ, we can determine when a player is performing significantly differently from his regular play. The estimates of the mean and standard deviations for the players can be updated between games to reflect their most recent play.

In the example illustrated in Figures 7.6 and 7.7, we use the entire season of data to produce the parameter estimates of the player's Corsi rating values. We can see in Figure 7.6 that this player significantly out-performed his season average on multiple occasions during this game and approximately half-way through the third period he is playing exceptionally. We also note that he has a significant decrease in Corsi rating in the final minutes of the game. These results could be used by the coaching staff to identify that this player could be having a stamina problem late in the game and that his play late in the game should be monitored in future games.

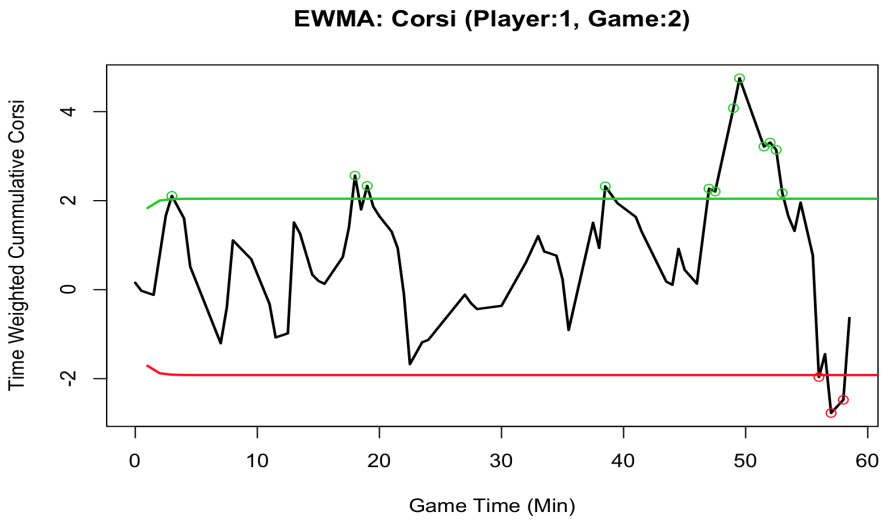

Figure 7.6: Example of a player's EWMA_{corsi} across a game with 2σ limits.

To monitor the game in an ongoing fashion, we propose using a simple box plot

approach. We will give the current EWMA statistic and limits for all player, at the current game time on one plot. To use this in a game situation, we would need to design a network system to allow us to run the code online while we update the database files. If we write the data to and run the analysis on a server and load the results to tablets or similar technology on the player bench, we could provide real-time charts in the games. These are not unreasonable expectations for teams at the junior or professional level, where there is already a large amount of technological infrastructure used.

In Figure 7.7 we show an example of how these results could be displayed. It is apparent that players numbered 5 and 11 are significantly under-performing at this point in the game and that most of the team is playing below average. The coach could use this information to inquire as to possible problems with players 5 and 11 or to rest them until the end of the period, when they can make a better assessment of the player's performance.

Overall, we have shown two easy-to-implement methods for monitoring player statistics from a game. Neither of these methods is computationally heavy, taking less than 30-seconds to run. In addition, the plots resulting from these methods are easy to understand, with clear results for use by teams. We believe that with the growth of data analysis and the integration of technology into hockey, these methods would be beneficial tools for improving team performance.

7.7.5 Predicting Game Trends

We can use Thomson's periodic reconstructions to produce estimates of several key statistics. Within our analysis we have already found the optimal set of lines that

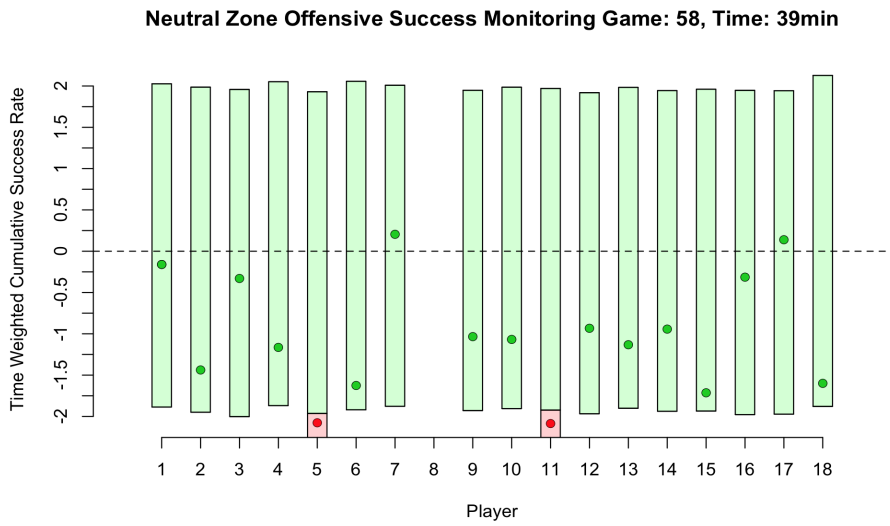

Figure 7.7: Example of an in-game EWMA$_{NO}$ with 2σ limits.

we can use as our independent variables. In addition, we include penalties as an independent variable. The statistics that we can model are probability of goals, neutral zone success, and Corsi rating. All will follow a similar framework for making a prediction. Because of the ever-changing nature of games, when a game seems disjointed as a result of the significant differences in play between periods or after major events such as injuries or fights, we recommend using only the data from after that point. This will ensure that the closest approximation to stationarity possible is maintained for this game.

As an example, we have modeled the Corsi rating for the final five minutes of the 11$^{\text{th}}$ game of the season. We used cross-validated LASSO regression to select the optimal set of variables for our model. The LASSO model estimating the Corsi rating

differential for a 30-second interval for the Frontenacs was

$$\begin{aligned}\hat{Corsi} = &-.0085 + .0644P_1 + .0442P_4 - .1266P_6 - .0287P_{11} - .0015P_{12} \\ &-.0222P_{13} + .02P_{14} + .0611P_{16} + .0525P_{17} + .1717P_{18} + .0441P_{22} \\ &-.0367P_2P_{10} + .0252P_{14}P_{15}P_{16} + .2285PEN.\end{aligned} \tag{7.9}$$

We can now take the residuals from the model and try to predict the final five minutes. We have decided to only use the data from the 2nd and 3rd periods, as we were concerned about changes in the flow of the game following the first period. This appears to be a logical change point as we witnessed the Frontenacs play more consistently in the 1st period of the first ten games of the season. We tested the equality of variances in the residuals from the first period and the later two, finding that they were significantly different (p-value = 0.002979).

We can see from the resulting prediction in Figure 7.8 that we should expect to have slightly worse than even puck possession immediately and around the 57 minute mark we should expect the other team to increase the pace of play. We expect that the Frontenacs' play will improve in the final two minutes. From this we can decide to select which players or lines to go on, to ensure that we have a positive predicted Corsi rating. For example, we may avoid putting on the defensive pairing of player 2 and 10 right now if all lines are equally rested.

This data does not appear to be strongly periodic, and the resulting uncertainty in the periodic terms is quite large. We also notice that the confidence interval with noise is very wide. This is most likely to be a result of the small amount of periodic structure in the data.

The performance of this method is not as great as we would like for use in real game situations. We believe that an increase in available co-variate data for the

LASSO model may improve the quality of the residuals, decreasing the amount of noise in the time series. In many situations these residuals may be non-stationary and the use of models from that field may be more applicable.

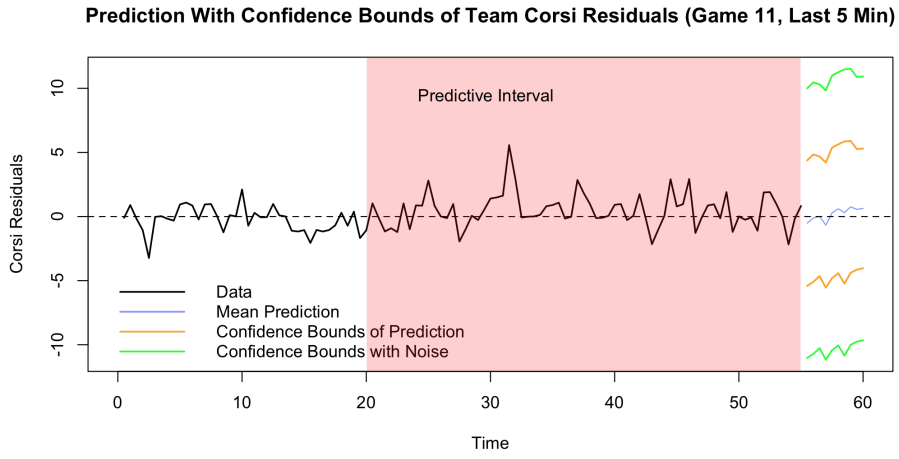

Figure 7.8: Example of the prediction of the final five minutes of game play.

7.8 Conclusions and Discussion

As we have shown, there are many ways to improve the decision making on players and teams in hockey through the use of statistics. Through our analysis, we have provided a tangible framework for monitoring base player statistics including using the new neutral zone metrics, selecting optimal lines from a potential set, monitoring player in-game quality, and providing predictions for future events. With all of our examples performed on real data, we have demonstrated that creating a bridge between advanced statistical methods and hockey is not as difficult as popular opinion would suggest.

We feel that in the next few years, as the sport and its fans gain a greater understanding of and appreciation for the merits of statistics, we will see a rise in many new techniques for analyzing statistics on the sport. The techniques described here show promise and we believe they will be useful additions to any team's statistical program. In particular, we believe that the END statistics and EWMA monitoring method are methods that would have large impacts on team performance. The END statistic is a simple and accessible metric that can be used to describe the quality of neutral zone play, an area that is currently underdeveloped. The EWMA method and other online reporting methods are not currently used in high levels of hockey. The EWMA method provides easy to understand charts and can offer a great level of in-game feedback to teams. We will recommend that the Kingston Frontenacs use the methods described here as they move forward and hope that, with time, more teams will follow suit.

Chapter 8

Concluding Remarks

The ability to make links between scientific fields is a powerful tool. We believe that this thesis demonstrates this sentiment, albeit on a smaller scale, merging areas within statistics. The problems we presented within spectrum analysis were resolved through the use of statistical learning techniques. These new hybrid methods we proposed were then shown to have applications to real-world problems.

Addressing several open problems within spectrum analysis, we demonstrated the potential for statistical learning theory to aid in other areas of statistics. The two focuses were removing supervised decisions from the spectrum estimation process and giving statisticians the tools to make informed and unbiased decisions. This is highlighted in the sphericity tests presented in Chapter 3 and the cross-validation method for cutoff selection within the inverse Fourier transform synthesis procedure discussed in Chapter 5. With the sphericity tests we were able to provide reasonable and well-founded parameter choices for the multitaper spectrum estimation method. For Thomson's method of time series synthesis, we developed a method to select a problem-specific optimal significance level for selecting periodic components. These

methods demonstrated how the implementation of simple techniques from another area of statistics can help us to avoid potentially significant mistakes in our analysis.

Along with trying to help minimize statistician bias in spectrum estimation, we worked to reduce the bias after decisions had been made. Our bootstrap-based methods for signal detection in Chapter 4 were able to provide a significant level of improvement over traditional methods, including in situations where user error is present. We also used a boosting method in Chapter 5 to identify additional signals that may have been missed by the naive optimization method or from incorrect user-defined cutoffs. The addition of statistical learning procedures following statistical decision-making can help to improve the performance of the methods. This is a useful property for real-world data where there are large amounts of uncertainty and where correct decisions may be difficult to identify.

Finally, we felt it important to tackle the most practical of time series methods, data synthesis, in Chapter 5. Building on the rich history of data modeling and estimation, we were able to introduce improved methods for dealing with data under ideal conditions. Applying our methods to real-world data, we were able to examine their performance in less-than-ideal situations. We found that the methods still performed well on real-world data, successfully interpolating New York temperature data, predicting coffee commodities market prices and, in Chapter 7, predicting hockey statistics.

In an effort to justify and validate the methods we proposed, we studied two different yet challenging data projects. Firstly, with the atrial signal extraction project in Chapter 6, we were able to demonstrate how the use of advanced methodologies can improve on the standard practices. Identifying atrial components is essentially a

signal detection problem with a large amount of complexity. The low signal-to-noise ratio made this problem ideal for demonstrating the benefits of the bootstrapped F-test. We were able to show a considerable and significant improvement on the currently accepted methods (average beat subtraction and eigenvalue-based principal components analysis) for both of the metrics we tested. Our one concern with the new advanced principal components analysis method is the increased computational cost. Because we spent little time on optimizing the code for this project, we believe that our timing estimates are a gross overestimation of the amount of time it would take to run the advanced principal components analysis method in an optimized setting. Additionally, most atrial signal examinations are performed retrospectively by physicians so the speed of extraction is not a vital concern.

Our final data project, evaluation of player and team performance in hockey, being a passion of ours, was not primarily focused on demonstrating how our proposed theoretical methods could be used. Instead, we aimed to use statistical methods to improve upon the evaluation of the game. We began by showing how, with only a moderate computer science background, it is possible to develop an adequate data collection program. While designing and implementing a data collection and analysis project was a significant undertaking, we felt that many improvements in how hockey can be evaluated were also possible. All of the methods we proposed in this project showed promise and utility for hockey at the major junior level. We provided justification for the inclusion of our proposed neutral zone statistics in evaluating player and team performance. We showed how quality control methods and, by extension, hypothesis testing can be useful tools in monitoring player performance. Our method for selecting line combinations demonstrated how regression methods can be useful

within hockey analysis, the primary concern being the proper identification of variables that describe success in play. Ideally, through data collection, we have data on goal production and can use this as our dependent variable. We believe that it is integral to have accurately timed shot event data to analyze player effects properly. This should be the major factor in designing a data collection method. Finally, we were able to also show that the periodic data synthesis methods from Chapter 5 have potential for use in predicting hockey trends. This method was not as successful as we had hoped and we believe that the major issue was missing variables. We felt that, overall, we were able to provide some innovations to the game that has captivated us for many years. We are excited to see how these methods will be employed by the Kingston Frontenacs this coming season in the OHL.

To summarize, we believe that the use of statistical learning tools can help in many scientific areas. Spectrum analysis, being a poorly understood and less-often-used field, was an ideal area for the introduction of these tools. We feel that considerable advancement can still be made within the field of spectrum estimation. We strongly encourage students and researchers to examine this area. Additionally, many more research areas have application for the use of advanced spectrum analysis methods. By partnering with leading researchers in these fields, we can increase the exposure of spectrum analysis within the overall scientific community and further advance the spectrum analysis methods to deal with the unique challenges of the data.

From a theoretical perspective, here are four areas in which we think that advancement can be made and that are potential examples of how to apply statistical learning theory to spectrum estimation:

1. Through the use of regularization on the basis expansions in quadratic inverse

theory [118], we could produce noise-reduced time-frequency estimates.

2. By clustering the frequencies within the F-test, we could produce a non-supervised signal detection within the multitaper framework. This is similar to the optimal α method described in section 5.4.

3. Cross-validation and bootstrapping could be used to produce optimized confidence intervals on multitaper spectrum estimates where sub-sampling is available.

4. Many of the methods used within spectrum and time series analysis involve modeling and produce residuals. There is limited analysis of the residuals in these fields, an example being the sphericity tests we designed in Chapter 3. By introducing more methods for analyzing and using the residuals from time series and spectrum analysis methods, we should be able to improve on existing modeling techniques. One such way would be by treating the residual K series from the F-test as eigenspectra and calculating a MTM estimate, we can look at the unexplained power from our model, examine the normality assumption, identify outliers, and tune the model.

These are only a few of the possibilities that are not direct extensions of the work in this thesis. We are hopeful that more research will be done on the methods proposed within this thesis.

Bibliography

[1] H. Abarbanel, T. Frison, and L. Tsimring. Obtaining order in a world of chaos. *Signal Processing Magazine of the IEEE*, 15(3):49–65, 1998.

[2] A. Almasri. Testing the periodicity on the Swedish varve data. *Centre for Labour Market Policy Research, Linnaeus University, Technical report*, (14), 2009.

[3] T. Anderson. An introduction to multivariate statistical analysis. *Wiley*, 1958.

[4] J. Aslam, R. Popa, and R. Rivest. On estimating the size and confidence of a statistical audit. In *Proceedings of the USENIX/ACCURATE Electronic Voting Technology Workshop*, 2007.

[5] M. Bayram and R. Baraniuk. Multiple window time-frequency analysis. In *Proceedings of the IEEE-SP International Symposium on Time-Frequency and Time-Scale Analysis*, pages 173–176, 1996.

[6] E. Benjamin, P. Wolf, R. DAgostino, H. Silbershatz, W. Kannel, and D. Levy. Impact of atrial fibrillation on the risk of death the Framingham heart study. *Circulation*, 98(10):946–952, 1998.

[7] M. Best and D. Neuhauser. Walter A Shewhart, 1924, and the Hawthorne factory. *Quality and Safety in Health Care*, 15(2):142–143, 2006.

[8] W. Beveridge. Wheat prices and rainfall in western Europe. *Journal of the Royal Statistical Society*, pages 412–475, 1922.

[9] M. Bordons, F. Morillo, and I. Gómez. Analysis of cross-disciplinary research through bibliometric tools. In *Handbook of Quantitative Science and Technology Research*, pages 437–456. 2005.

[10] R. Bousseljot, D. Kreiseler, and A. Schnabel. Nutzung der EKG-Signaldatenbank CARDIODAT der PTB über das internet. *Biomedical Engineering*, 40(s1):317–318, 1995.

[11] G. Box, G. Jenkins, and G. Reinsel. Time series analysis: forecasting and control. *Wiley*, 2011.

[12] L. Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.

[13] S. Butterworth. On the theory of filter amplifiers. *Wireless Engineer*, 7(6):536–541, 1930.

[14] C. Cammarota, G. Guarini, E. Rogora, and M. Ambrosini. Non stationary model of the heartbeat time in atrial fibrillation. In *Proceedings of Fifth ESMTB Conference on Mathematical Modeling and Computing in Biology and Medicine*, 2002.

[15] F. Castells, P. Laguna, L. Sörnmo, A. Bollmann, and J. Millet. Principal component analysis in ECG signal processing. *EURASIP Journal on Applied Signal Processing*, (1):98–98, 2007.

[16] F. Castells, C. Mora, J. Rieta, D. Moratal-Pérez, and J. Millet. Estimation of atrial fibrillatory wave from single-lead atrial fibrillation electrocardiograms using principal component analysis concepts. *Medical and Biological Engineering and Computing*, 43(5):557–560, 2005.

[17] J. Catalano. Guide to ECG analysis. *Lippincott Williams & Wilkins*, 2002.

[18] C. Chatfield. The analysis of time series: an introduction. *CRC Press*, 2013.

[19] L. Cohen. Time-frequency analysis. *Prentice Hall*, 1995.

[20] F. Conner. Hockey's most wanted: The top 10 book of wicked slapshots, bruising goons and ice oddities. *Potomac Books*, 2002.

[21] N. Crato and B. Ray. Some problems in the overspecification of ARMA and processes using arfima models. In *Proceedings of the Third Congress of the Portuguese Statistical Society*, pages 527–539, 1996.

[22] A. Davison and D. Hinkley. Bootstrap methods and their application. *Cambridge University Press*, 1997.

[23] A. DiMento. Six reasons why the NHL is increasing in popularity in the USA. *Upper Deck*, 2014. http://upperdeckblog.com/2014/01/six-reasons-why-the-nhl-is-increasing-in-popularity-in-the-usa/.

[24] B. Efron. Bootstrap methods: another look at the jackknife. *The Annals of Statistics*, 7(1):1–26, 1979.

[25] B. Efron and R. Tibshirani. An introduction to the bootstrap. *CRC Press*, 1994.

[26] B. Everitt and A. Skrondal. The Cambridge dictionary of statistics. *Cambridge University Press*, 2006.

[27] J. Fan and Q. Yao. Nonlinear time series: nonparametric and parametric methods. *Springer*, 2003.

[28] M. Fenwick. Footnotes. *Battle of Alberta*, February 2015. http://battleofalberta.blogspot.ca/2015/02/footnotes.html.

[29] J. Fischer. The neutral zone trap and the New Jersey Devils. 2011. http://www.inlouwetrust.com/2011/9/6/2408894/the-neutral-zone-trap-the-new-jersey-devils.

[30] R. Fisher. On the "probable error" of a coefficient of correlation deduced from a small sample. *Metron*, 1:3–32, 1921.

[31] R. Fisher. Statistical methods for research workers. *Genesis Publishing*, 1925.

[32] R. Fisher. Statistical methods and scientific induction. *Journal of the Royal Statistical Society, Series B*, 17(1):69–78, 1955.

[33] G. Foschini and M. Gans. On limits of wireless communications in a fading environment when using multiple antennas. *Wireless Personal Communications*, 6(3):311–335, 1998.

[34] J. Franke and W. Hardle. On bootstrapping kernel spectral estimates. *The Annals of Statistics*, 20(1):121–145, 1992.

[35] A. Franks, A. Miller, L. Bornn, and K. Goldsberry. Characterizing the spatial structure of defensive skill in professional basketball. *The Annals of Applied Statistics*, 9(1):94–121, 2015.

[36] M. Frantseva, J. Cui, F. Farzan, L. Chinta, J. Velazquez, and Z. Daskalakis. Disrupted cortical conductivity in schizophrenia: TMS–EEG study. *Cerebral Cortex*, 24(1):211–221, 2014.

[37] D. Freedman. Bootstrapping regression models. *The Annals of Statistics*, 9(6):1218–1228, 1981.

[38] J. Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 29(5):1189–1232, 2001.

[39] M. Friedman. The interpolation of time series by related series. *Journal of the American Statistical Association*, 57(300):729–757, 1962.

[40] G. Gan, C. Ma, and J. Wu. Data clustering: theory, algorithms, and applications. *Siam*, 2007.

[41] A. Ghaffari, M. Homaeinezhad, M. Khazraee, and M. Daevaeiha. Segmentation of holter ECG waves via analysis of a discrete wavelet-derived multiple skewness–kurtosis based metric. *Annals of Biomedical Engineering*, 38(4):1497–1510, 2010.

[42] A. Goldberger, L. Amaral, L. Glass, J. Hausdorff, P. Ivanov, R. Mark, J. Mietus, G. Moody, C. Peng, and H Stanley. Physiobank, Physiotoolkit, and Physionet components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220, 2000.

[43] L. Goldman, R. Sayson, S. Robbins, L. Cohn, M. Bettmann, and M. Weisberg. The value of the autopsy in three medical eras. *New England Journal of Medicine*, 308(17):1000–1005, 1983.

[44] D. Groll and D. Thomson. Incidence of influenza in Ontario following the universal influenza immunization campaign. *Vaccine*, 24(24):5245–5250, 2006.

[45] H. Halff. Graphical evaluation of hierarchical clustering schemes. *Center for the Study of Reading, University of Illinois, Technical Report*, 1975.

[46] S. Halli and K. Rao. Advanced techniques of population analysis. *Springer*, 2013.

[47] E. Hannan. Multiple time series. *Wiley*, 2009.

[48] T. Hastie, R. Tibshirani, and J. Friedman. The elements of statistical learning. *Springer*, 2001.

[49] S. Haykin, D. Thomson, and J. Reed. Spectrum sensing for cognitive radio. *Proceedings of the IEEE*, 97(5):849–877, 2009.

[50] L. Hinnov and S. Meyers. Paleoclimate time scale estimation using multitaper spectral methods. In *International Conference: Applied Mathematics, Modeling and Computational Science*, 2013.

[51] F. Hong and R. Breitling. A comparison of meta-analysis methods for detecting differentially expressed genes in microarray experiments. *Bioinformatics*, 24(3):374–382, 2008.

[52] C. Huberty and S. Olejnik. Applied MANOVA and discriminant analysis. *Wiley*, 2006.

[53] S. John. The distribution of a statistic used for testing sphericity of normal distributions. *Biometrika*, 59(1):169–173, 1972.

[54] B. Kedem and K. Fokianos. Regression models for time series analysis. *Wiley*, 2005.

[55] J. Klein. Statistical visions in time: a history of time series analysis, 1662-1938. *Cambridge University Press*, 1997.

[56] R. Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *International Joint Conference on Artificial Intelligence*, pages 1137–1145, 1995.

[57] B. Korin. On the distribution of a statistic used for testing a covariance matrix. *Biometrika*, 55(1):171–178, 1968.

[58] P. Langley, J. Rieta, M. Stridh, Jo. Millet, L. Sörnmo, and A. Murray. Comparison of atrial signal extraction algorithms in 12-lead ECGs with atrial fibrillation. *IEEE Transactions on Biomedical Engineering*, 53(2):343–346, 2006.

[59] P. Laplace. Mémoire sur les probabilités. *Mémoires de lAcadémie Royale des Sciences de Paris*, 1778:227–332, 1781.

[60] C. Lawson and R. Hanson. Solving least squares problems. *SIAM*, 1974.

[61] E. Lehmann. The Fisher, Neyman-Pearson theories of testing hypotheses: One theory or two? *Journal of the American Statistical Association*, 88(424):1242–1249, 1993.

[62] E. Lindquist. Statistical analysis in educational research. *Houghton Mifflin*, 1940.

[63] M. Loeve. Probability theory. *Springer*, 1963.

[64] M. Masisak. Super 16: Growth of analytics a boon for hockey fans. *NHL.com*, 2015. http://www.nhl.com/ice/news.htm?id=762471.

[65] J. Matisz. NHL.com getting fancy with inclusion of 'enhanced' stats. *Toronto Sun*, 2015. http://www.torontosun.com/2015/02/19/nhlcom-getting-fancy-with-inclusion-of-enhanced-stats.

[66] S. McIndoe. The NHL's analytics awakening. *Grantland*, 2014. http://grantland.com/the-triangle/the-nhls-analytics-awakening/.

[67] B McKenzie. The real story of how Corsi got its name. 2014. http://www.tsn.ca/mckenzie-the-real-story-of-how-corsi-got-its-name-1.100011.

[68] A. Miller. Subset selection in regression. *CRC Press*, 2002.

[69] D. Montgomery. Introduction to statistical quality control. *Wiley*, 2007.

[70] I. Moore and E. Tompa. Understanding changes over time in workers' compensation claim rates using time series analytical techniques. *Occupational and Environmental Medicine*, 68(11):837–841, 2011.

[71] F. Mormann, R. Andrzejak, C. Elger, and K. Lehnertz. Seizure prediction: the long and winding road. *Brain*, 130(2):314–333, 2007.

[72] L. Nelson. Column: Technical aids: The Shewhart control chart–tests for special causes. *Journal of Quality Technology*, 16(4), 1984.

[73] J. Neyman and E. Pearson. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 231(694–706):289–337, 1933.

[74] NHL.com. Stars coach calls new lineup options 'tremendous'. *NHL.com*, 2015. http://www.nhl.com/ice/news.htm?id=775142.

[75] R. Nickerson. Null hypothesis significance testing: a review of an old and continuing controversy. *Psychological Methods*, 5(2):241, 2000.

[76] O. Niemitalo. Window function and frequency response - triangular. 2013. http://en.wikipedia.org/wiki/Window_function.

[77] S. Nieuwenhuis, B. Forstmann, and E. Wagenmakers. Erroneous analyses of interactions in neuroscience: a problem of significance. *Nature neuroscience*, 14(9):1105–1107, 2011.

[78] J. O'Keefe Jr, S. Hammill, M. Freed, and S. Pogwizd. The complete guide to ECGs. *Jones & Bartlett*, 2010.

[79] A. Owen. Pearsons test in a large scale multiple meta-analysis. *Stanford University, Technical Report*, 2007.

[80] A. Parnass. Analytics, not statistics, driving NHL evolution. *NHL.com*, 2015. http://www.nhl.com/ice/news.htm?id=754099.

[81] D. Percival and A. Walden. Spectral analysis for physical applications. *Cambridge University Press*, 1993.

[82] S. Petrutiu, J. Ng, G. Nijm, H. Al-Angari, S. Swiryn, and A. Sahakian. Atrial fibrillation and waveform characterization. *Engineering in Medicine and Biology Magazine of the IEEE*, 25(6):24–30, 2006.

[83] J. Pohlkamp-Hartt. The development and practical study of a grey space detector for cognitive radio. Master's thesis, Queen's University, Kingston, Ontario, Canada, 2010.

[84] H. Poor. An introduction to signal detection and estimation. *Springer*, 2013.

[85] K. Prabhu. Window functions and their applications in signal processing. *CRC Press*, 2013.

[86] M. Priestley. Spectral analysis and time series. *Academic Press*, 1981.

[87] G. Prieto, F. Vernon, G. Masters, and D. Thomson. Multitaper Wigner-Ville spectrum for detecting dispersive signals from earthquake records. In *Conference Record of the Thirty-Ninth Asilomar Conference on Signals, Systems and Computers*, pages 938–941. IEEE, 2005.

[88] T Purdy. Shots, Fenwick and Corsi. 2011. http://objectivenhl.blogspot.ca/2011/02/shots-fenwick-and-corsi.html.

[89] K. Rao, A.and Hamed and H. Chen. Nonstationarities in hydrologic and environmental time series. *Springer*, 2003.

[90] M. Rao. Representation and estimation for harmonizable type processes. In *Conference Record of the Thirty-Sixth Asilomar Conference on Signals, Systems and Computers*, pages 1559–1564. IEEE, 2002.

[91] A. Ratna. Development of the heart rate monitor by using finger detector. Bachelor thesis, Universiti Teknikal Malaysia Melaka, Hang Tuah Jaya, Melaka, Malaysia, 2008.

[92] K. Riedel and A. Sidorenko. Minimum bias multiple taper spectral estimation. *IEEE Transactions on Signal Processing*, 43(1):188–195, 1995.

[93] D. Riegert. Post-whitening spectra for comparison and determining significant peaks. 2013. http://driegert.wordpress.com/2013/04/06/post-whitening-spectra-for-comparison-and-determining-significant-peaks/.

[94] D. Riegert, A. Springford, and D. Thomson. Forecasting the likelihood of solar flares using an inferred solar stress index. In *Statistical Society of Canada Annual Meetings*, 2014.

[95] J. Rieta, F. Castells, C. Sánchez, V. Zarzoso, and J. Millet. Atrial activity extraction for atrial fibrillation analysis using blind source separation. *IEEE Transactions on Biomedical Engineering*, 51(7):1176–1186, 2004.

[96] S. Roberts. Control chart tests based on geometric moving averages. *Technometrics*, 1(3):239–250, 1959.

[97] J. Rutherford. 100 things Blues fans should know and do before they die. *Triumph Books*, 2014.

[98] H. Sakoe and S. Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 26(1):43–49, 1978.

[99] L. Scharf. Statistical signal processing. *Addison-Wesley*, 1991.

[100] M. Schuckers and J. Curro. Total hockey rating (THoR): A comprehensive statistical rating of National Hockey League forwards and defensemen based upon all on-ice events. In *Proceedings of the MIT Sloan Sports Analytics Conference*, 2013.

[101] M. Schulz and K. Stattegger. SPECTRUM: spectral analysis of unevenly spaced paleoclimatic time series. *Computers & Geosciences*, 23(9):929–945, 1997.

[102] D. Shiau. *Signal identification and forecasting in nonstationary time series data*. PhD thesis, University of Florida, Gainesville, Florida, USA, 2001.

[103] R. Shumway and D. Stoffer. Time series analysis and its applications. *Springer*, 2013.

[104] S. Simmons. Why hockey's trendy advanced stats are a numbers game. *Toronto Sun*, 2014. http://www.torontosun.com/2014/05/20/why-hockeys-trendy-advanced-stats-are-a-numbers-game.

[105] K. Sithamparanathan and A. Giorgetti. Cognitive radio techniques: spectrum sensing, interference mitigation, and localization. *Artech House*, 2012.

[106] D. Slepian. Some asymptotic expansions for prolate spheroidal wave functions. *Journal of Mathematical Physics*, 44(2):99–140, 1965.

[107] D. Slepian and H. Pollack. Prolate spheroidal wave functions, Fourier analysis and uncertainty - I. *Bell System Technical Journal*, 40(1):43–64, 1961.

[108] L. Sörnmo, M. Stridh, and J. Rieta. Atrial activity extraction from the ECG. *Understanding Atrial Fibrillation: The Signal Processing Contribution*, pages 53–80, 2008.

[109] S. Stigler. The history of statistics: The measurement of uncertainty before 1900. *Harvard University Press*, 1986.

[110] M. Stridh and L. Sörnmo. Spatiotemporal QRST cancellation techniques for analysis of atrial fibrillation. *IEEE Transactions on Biomedical Engineering*, 48(1):105–111, 2001.

[111] Student. The probable error of a mean. *Biometrika*, 6(1):1–25, 1908.

[112] G. Teyssière and A. Kirman. Long memory in economics. *Springer*, 2006.

[113] N. Thakor, J. Webster, and W. Tompkins. Estimation of QRS complex power spectra for design of a QRS filter. *IEEE Transactions on Biomedical Engineering*, 31(11):702–706, 1984.

[114] A. Thomas, M. Schuckers, B. Macdonald, S. Ventura, and K. Mongeon. Statistics on ice: Advances in methods for the analysis of ice hockey. In *Joint Statistical Meetings*, 2014.

[115] D. Thomson. Spectrum estimation techniques for characterization and development of wt4 waveguidei. *Bell System Technical Journal*, 56(9):1769–1815, 1977.

[116] D. Thomson. Spectrum estimation and harmonic analysis. *Proceedings of the IEEE*, 70(9):1055–1096, 1982.

[117] D. Thomson. Quadratic-inverse spectrum estimates: applications to paleoclimatology. *Philosophical Transactions: Physical Sciences and Engineering*, 332(1627):536–597, 1990.

[118] D. Thomson. An overview of multiple-window and quadratic-inverse spectrum estimation methods. *Proceedings In Acoustics, Speech, and Signal Processing*, 6(1):185–94, 1994.

[119] D. Thomson. Quadratic-inverse expansion of the Rihaczek distribution. In *Conference Record of the Thirty-Ninth Asilomar Conference on Signals, Systems and Computers*, pages 912–915. IEEE, 2005.

[120] D. Thomson. Jackknifing multitaper spectrum estimates. *Signal Processing Magazine of the IEEE*, 24(4):20–30, 2007.

[121] R. Tibshirani. Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society, Series B*, 58(1):267–288, 1996.

[122] R. Tsay. Analysis of financial time series. *Wiley*, 2005.

[123] F. Tseng and G. Tzeng. A fuzzy seasonal ARIMA model for forecasting. *Fuzzy Sets and Systems*, 126(3):367–376, 2002.

[124] H. Urkowitz. Energy detection of unknown deterministic signals. *Proceedings of the IEEE*, 55(4):523–531, 1967.

[125] M. Usai, M. Goddard, and B. Hayes. LASSO with cross-validation for genomic selection. *Genetics Research*, 91(6):427–436, 2009.

[126] H. van Lanen and S. Demuth. FRIEND 2002: Regional hydrology: Bridging the gap between research and practice. *International Association of Hydrological Sciences*, (274), 2002.

[127] K. Vu. The ARIMA and VARIMA time series: Their modelings. *AuLac Technologies*, 2007.

[128] G. Walker. On periodicity in series of related terms. *Proceedings of the Royal Society of London, Series A*, 59(7):518–532, 1931.

[129] J. Ward Jr. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301):236–244, 1963.

[130] A. Weigend. Time series prediction: forecasting the future and understanding the past. *Santa Fe Institute Studies in the Sciences of Complexity*, 1994.

[131] N. Weiss and C. Weiss. Introductory statistics. *Pearson Education*, 2012.

[132] P. Whittle. Hypothesis testing in time series analysis. *Almqvist & Wiksells*, 1951.

[133] N. Wiener. Extrapolation, interpolation, and smoothing of stationary time series. *The MIT press*, 1949.

[134] A. Wiles. Modular elliptic curves and Fermat's last theorem. *Annals of Mathematics*, 141:443–551, 1995.

[135] R. Wohlstetter. Pearl Harbor: warning and decision. *Stanford University Press*, 1962.

[136] H. Wold. Bibliography on time series and stochastic processes. *The MIT Press*, 1965.

[137] M. Wright. The National Hockey League, 1917-1967: A year-by-year statistical history. *McFarland*, 2010.

[138] W. Yaghi. Detecting autocovariance change in time series. *ProQuest*, 2007.

[139] U. Yule. On a method of investigating periodicities in disturbed series, with special reference to Wolfer's sunspot numbers. *Philosophical Transactions of the Royal Society of London, Series A*, 226(636–646):267–298, 1927.

[140] H. Zhang and L. Zhang. ECG analysis based on PCA and support vector machines. In *International Conference on Neural Networks and Brain*, volume 2, pages 743–747. IEEE, 2005.

[141] W. Zong, G. Moody, and D. Jiang. A robust open-source algorithm to detect onset and duration of qrs complexes. In *Computers in Cardiology*, pages 737–740. IEEE, 2003.