The Sphericity Test for Parameter Selection for Multitaper Spectral Estimation

J.W.W. Pohlkamp-Hartt

October 1, 2014

1 Introduction

Spectrum Estimation is a commonly used tool for research on temporally correlated data. Within the field of Spectrum Estimation, there are many commonly used methods. The Multitaper Method is a leading method due to it's desirable bias-variance properties. 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 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 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.

In statistical data analysis, we are often provided with temporally correlated data. 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. 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 [5]. A key technique in time series analysis is that of spectrum estimation [7].

2 Spectrum Estimation

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 exists within our data. The periodic structure itself may be of interest, as in the field of radio communications[3], or we may wish to remove this structure and look for other properties found in their residuals, a useful practice in Economics [10]. Many methods of spectral estimation exist with the majority of research being done under the assumptions of Gaussian noise and stationary signals.

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. For most methods, depending on the choice of windows, there is a trade off between the bias and the variance of the estimate[2].

3 Multitaper Method

This trade off is controlled when using the Multitaper Spectral Estimation method (MTM) [9]. The MTM was introduced by David J. Thomson 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 its 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) [8]. By using any orthogonal family of functions they will have maximal energy within a given frequency band. An attractive property of the Slepians is that the Fourier transformations of the windows minimize the weight given to out-of-band frequencies.

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

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

length N, $\nu_t^{(k)}$, where $t = 0, ..., N-1,\ k = 0, ..., K-1$, the number of windows denoted by K. We use these as windows for K Fourier transformations, called the eigenspectra of the data.

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

The initial naive spectral estimate is then formed as

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

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 not suffer as poorly 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 your research.

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 values resulting from the computation of the F-test for detection of line components. We think of the F-test as a regression problem, where we are regressing the frequency domain Slepian sequences taken at the base frequency, $V_k(0)$, onto the eigenspectra for each frequency, $Y_k(f)$.

$$Y_k(f) = \hat{\mu}(f) V_k(0) + e(f), \tag{3}$$

where $e(f) \sim CN(0, \sigma^2)$ is Complex Gaussian distributed with variance equal to the background noise of the environment, $\sigma^2 = S_N(f)$.

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 if 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:

$$F(f) = (K-1) \frac{|\hat{\mu}(f)|^2 \sum_{k=0}^{K-1} |V_k(0)|^2}{\sum_{k=0}^{K-1} |\hat{r}_k(f)|^2}, \tag{4}$$

$$\hat{r}_k(f) = Y_k(f) - \hat{\mu}(f) V_k(0), \tag{5}$$

$$\hat{\mu}(f) = \frac{\sum_{k=0}^{K-1} V_k^{\star}(0) Y_k(f)}{\sum_{k=0}^{K-1} |V_k(0)|^2}, \tag{6}$$

$$V_k(f) = \sum_{t=0}^{N-1} \nu_t^{(k)} e^{-2i\pi ft}, \tag{7}$$

where Y_k are the eigenspectra for our time series x_t, and $\nu_t^{(k)}$ are the Slepian sequences in the time domain.

We are also able to get an estimate for the residuals, $\hat{r}_k(f) = Y_k(f) - \hat{\mu}(f) V_k(0)$. These residual are used to evaluate the validity of the parameter choices used in the MTM.

4 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 window 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 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 1. The test then becomes determining if the residuals follow a complex normal distribution. The test that is usually used in this case is one for sphericity. Sphericity is had when the spread of a random variable in 2 or more dimensions is equal. This occurs when the covariance matrix is diagonal. For this test, we have K sets of residuals that are N long complex vectors. Since the concern is that

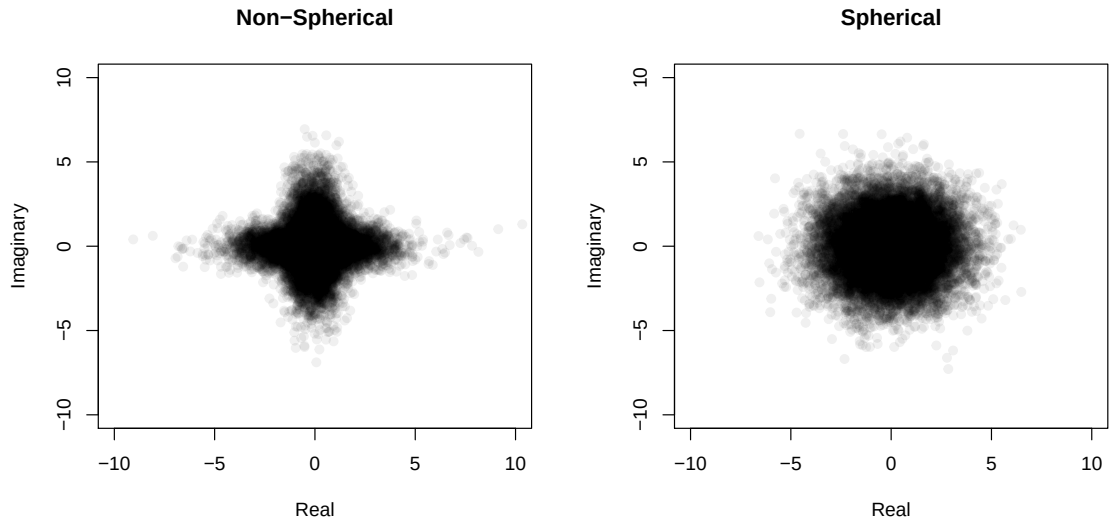

Figure 1: Comparison of spherical and non-spherical distributed complex-valued residuals.

all of the residuals follow a $CN(0, \sigma^2)$ distribution, we concatenate our K vectors into one that is NK long and test the sphericity of all the residual terms together. We follow the test for sphericity described by S. John[4].

$$R = [r_0(f_1), r_0(f_2), ..., r_0(f_N), r_1(f_1), ..., r_1(f_N), ..., r_{k-1}(f_1), ..., r_{k-1}(f_N)] \tag{8}$$

$$H_0 : Cov([Re(R), Im(R)]) = \sigma^2 \mathbb{I}, \tag{9}$$

where σ^2 is unknown and not defined.Under H_0,

$$(N-2)[\frac{1}{2(1-W)^{1/2}} - 1] \sim \mathbb{F}_{2,2N-4} \tag{10}$$

with

$$W = \frac{tr((NV)^2)}{(tr(NV))^2}, \;\; V = Cov([Re(R), Im(R)]) \tag{11}$$

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 being 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. 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 and in practical settings that NW can be as large as 100. Also due to the fact that the first $2NW$ Slepian sequences maximize the in-band, $f \in (-W, W)$, to total, $f \in (-1/2, 1/2)$, spectral concentration, we should choose NW to be an integer. This would allow us to select K as large as $2NW$ and not 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 minimal probability of non-sphericity across all choices is a good naive method for selecting NW and K. 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 non-stationary time series, so failure in this test is welcomed in that case. 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.

5 Bagged Sphericity Test

In situations where the power 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 mirrors the work of Korin [6] and Anderson [1].

$$H_0 : Cov([Re(R), Im(R)]) = \sigma^2 \mathbb{I}, \tag{12}$$

σ^2 is known.Under H_0,

$$\rho(N-1)(\log \mid \sigma^2 \mathbb{I} \mid - \log \mid V \mid + tr(V(\sigma^2 \mathbb{I})^{-1}) - 2) \sim \chi^2_3 + \omega(\chi^2_7 - \chi^2_3) \tag{13}$$

with

$$\omega = \frac{47}{432(N-1)^2},\ \rho = 1 - \frac{15}{18(N-1)}. \tag{14}$$

Unfortunately, implementing this test directly runs into two significant computational issues. First, to avoid issues with round off for small values we need to ensure to scale the time series has an estimated background noise of σ^2 1. This alleviates the possibility of the computation of either determinant to be rounded to zero.

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. As K increases, the amount 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, N, increases so does the effect of the variance reduction for larger values of NW. For data sets with $N > 100$, 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 to select constant sized and random sample of the residuals. We then perform hypothesis testing on each sample and use Fisher's method for meta-analysis to consider the sphericity of each sampling collectively. The procedure is adapted from equation 13 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 now is as follows:
1) Sample the residuals with replacement M times to be used as separate data sets.
2) Test the null hypothesis, $H_0 : Cov([Re(R), Im(R)]) = \sigma^2 \mathbb{I}$, for all M sets of samples using the statistic from equation 13 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 M sets of samples. Fisher's method states that $-2 \sum_{i=1}^{M} ln(p_i) \sim \chi^2_{2M}$
.

From this procedure we can now obtain a measure of the 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 3 potentially significant issues.

First, due to the random re-sampling that is used, for each evaluation of this test the p-value returned is not consistent for a give set of parameters. This is still true for large numbers of runs, $M > 1000$, and the potential for a consistent result due to an increase in run size may be difficult to achieve with reasonable computational efficiency.

Next, there is the potential to have runs with no signals in them, which will have p-values near zero no matter the parameter choice. Within Fisher's method these near zero p-values will cause the summation to approach negative infinity with no consideration for the other set's p-values. To avoid this issue we use the common approach of setting the minimum p-value for any set to $1/M$ which will alleviate this issue but will increase the false detection of the test.

Lastly, 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 in the simulations portion of this paper.

6 Simulations and Comparison

In an effort to demonstrate the merits of the tests designed in this report, 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 a naive 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 5 sinusoids of lessening amplitude as moving away from the central frequency.

$$X(t) = \sum_{i=1}^{38} \alpha_i \sum_{j=-2}^{2} (.3 - .1 \mid j \mid) \sin(2\pi(.013i + .002j)t) + z_t, \tag{15}$$

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 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 that 6 would cause there to be 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 ran both tests sphericity test for $NW \in [2, 10]$ 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 $K = 6$ provided the lowest probability of the residuals not being spherical. The bagged sphericity test with $M = 4000$ had the same parameter choices of $NW = 6$ and $K = 6$. These results are within our theoretical range on the parameters and goes along with our motivation for a good naive 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 probability of non-sphericity rises, this is visible in the naive test for sphericity, Figure 3. The probability also rises with K; this may be caused by the signals having little power at the band edge.

While the tests do agree on the same choice of parameters, they do observe slightly different characteristics. The naive test has considerably lower p-values, with the minimum being around .006, compared to the bagged tests minimum p-value of .997. Further testing did show that as M increased the p-values of bagged test also increased. Additionally the naive test had more parameter choices with p-values near the minimum, while the parameter values near the choice were similar in scale for the bagged test but all other values had p-values of 1.

The choice of M within the bagged test had several effects on the results of the test. First as would be expected when we increase M there is convergence in the choice of parameters with the lowest p-value, that is as M increases the likeliness of the parameters with the lowest p-value changing from one computation to the next decreases. This is an obvious result 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 M 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 increase in minimum p-value for Fisher's statistics as M increases, the number of combinations that are perversely non-spherical that are included will increase with M and their p-values near 1 will increase the Fisher Statistic. This relationship is shown in 5.

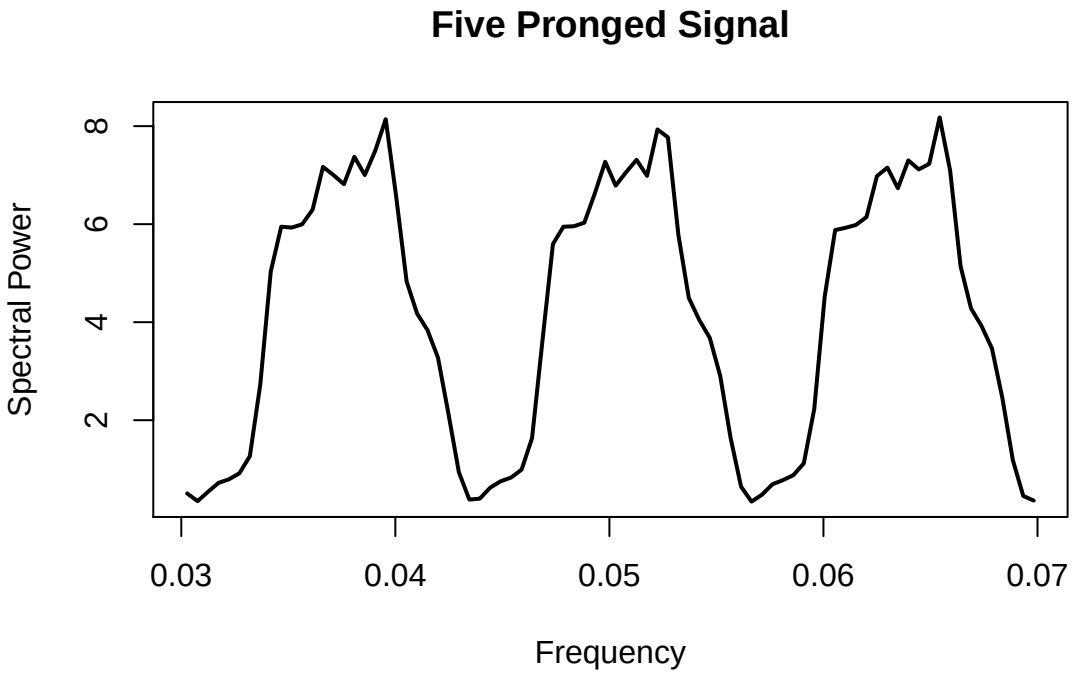

Figure 2: Part of the spectrum showing three test signals for the Sphericity Test, $NW = 4$, $K = 7$, $N = 1000$.

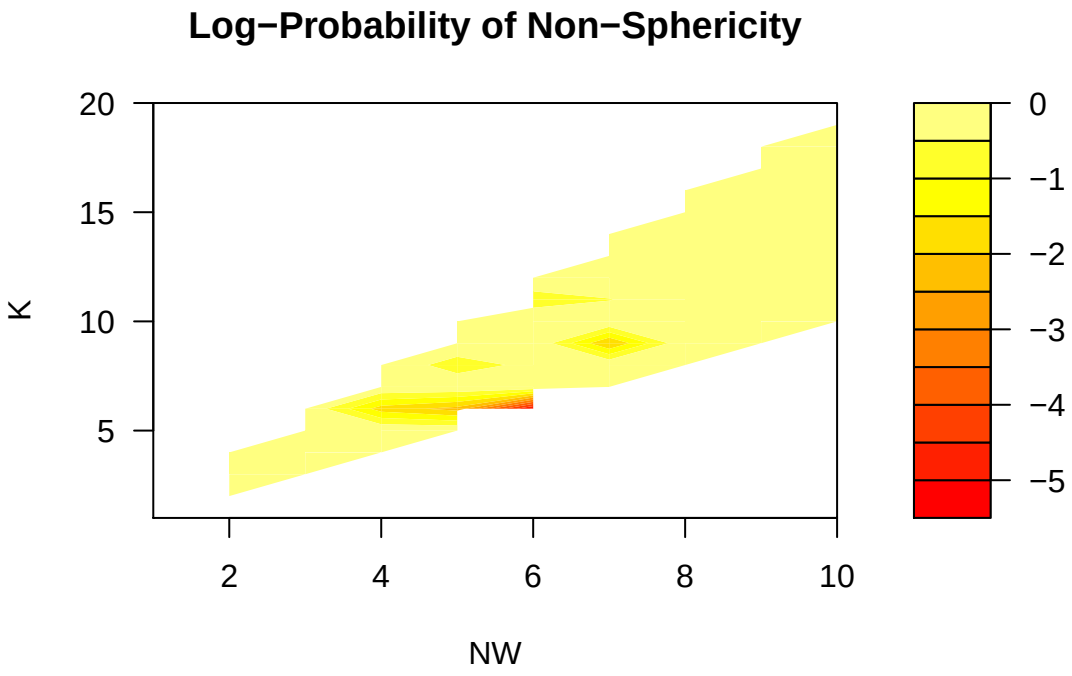

Figure 3: Naive Sphericity Test of simulated evenly spaced 5-pronged sinusoids in noise for $NW = [2, 9]$ and $K = [2, 18]$.

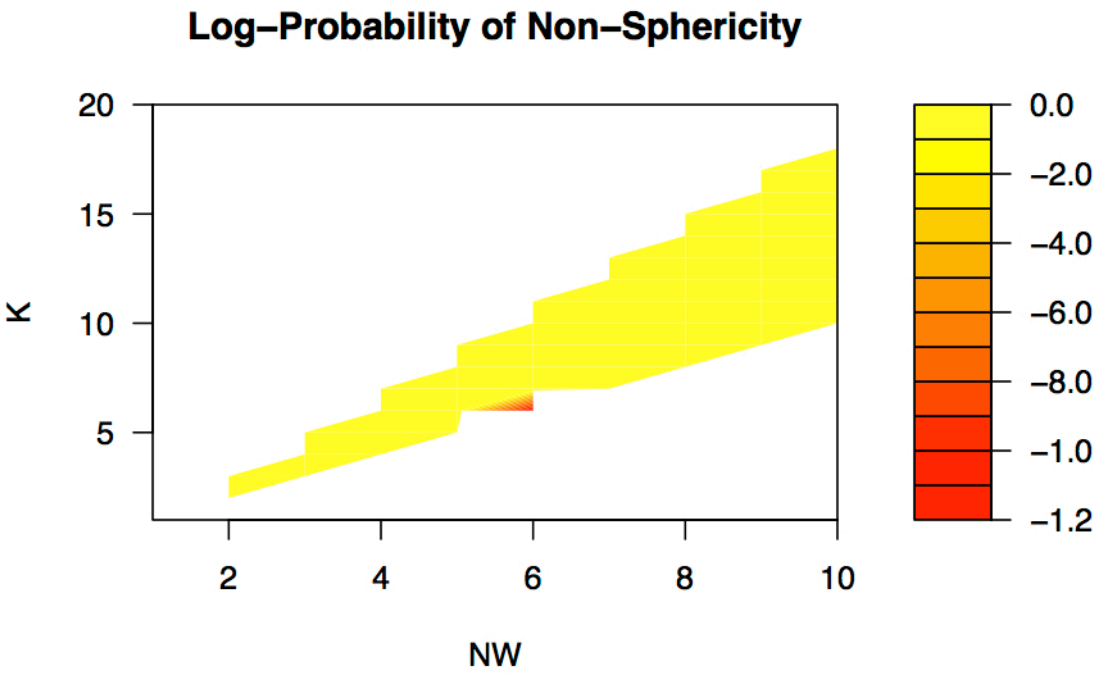

Figure 4: Bagged Sphericity Test of simulated evenly spaced 5-pronged sinusoids in noise for $NW = [2, 9]$ and $K = [2, 18]$.

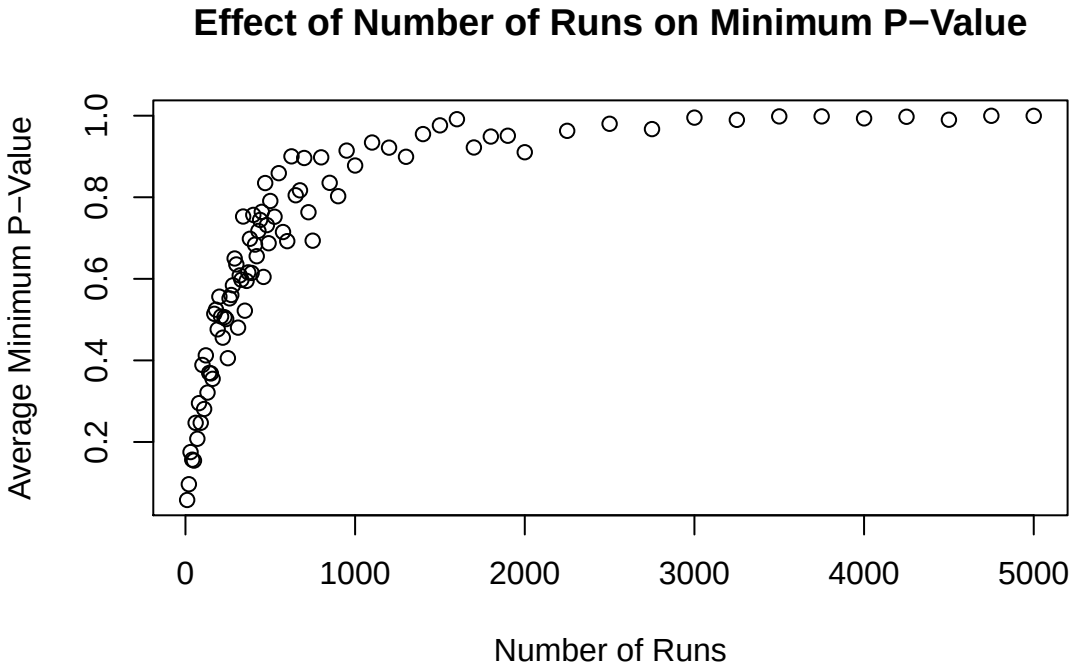

Figure 5: Effect of Number of Runs, M, on Minimum P-Value for The Bagged Sphericity Test

This convergence property does come at the cost of computing time. Figure 6 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 $1.3sec$ for the same data set, this is similar in cost to $M \sim 50$.

For small values of M, there is considerable variance in the parameters returned as ideal. To demonstrate this property, we evaluated the bagged sphericity test 1000 times for $M = 50$. We then plotted the percentage of occurrences for which each parameter was the ideal choice. As you can see in Figure 7, there is a large set of values chosen to be the ideal parameter choices for a given try. Fortunately, we still have the choice of $NW = 6$ and $K = 6$ as the most common result.

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 $\sigma^2 \in [1.4, 3.4]$. The result of the average of 10 trials at each choice in the range, displayed in Figure 8, showed that as you under-estimate the noise level, the lower-valued parameters begin to have lower p-values. The opposite hold for over-estimated values, where the p-values of the larger valued parameters were lower. As we reached the upper limit on the range tested, we found the variance in the choice selected was increased. This effect was due to all parameter choices being near 1. In addition, outside of this range, we found that all p-values were 1 for all parameter choices which demonstrates that this test fails in the event that you do not have a reasonable estimate of the noise level.

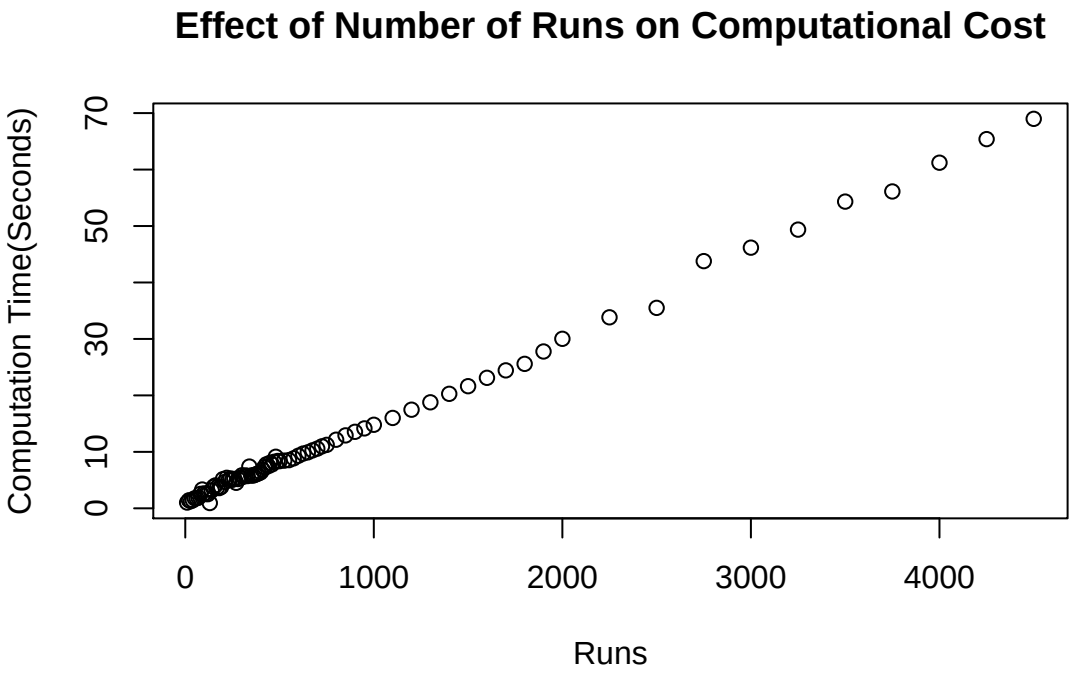

Figure 6: Effect of Number of Runs, M, on Computational Time for The Bagged Sphericity Test

7 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 increased probability of being not 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 the same parameters as the bagged test for simple simulations.

While the bagged test may be the more thorough test, it does come with drawbacks. For reliable results the use of a large number of runs is necessary but this comes at the cost of significant computer time. In addition the lack of robustness of the test against the misspecification of the power of the noise creates a significant 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. Overall both the tests did provide a reasonable choice of parameters when no theoretical parameter values are possible.

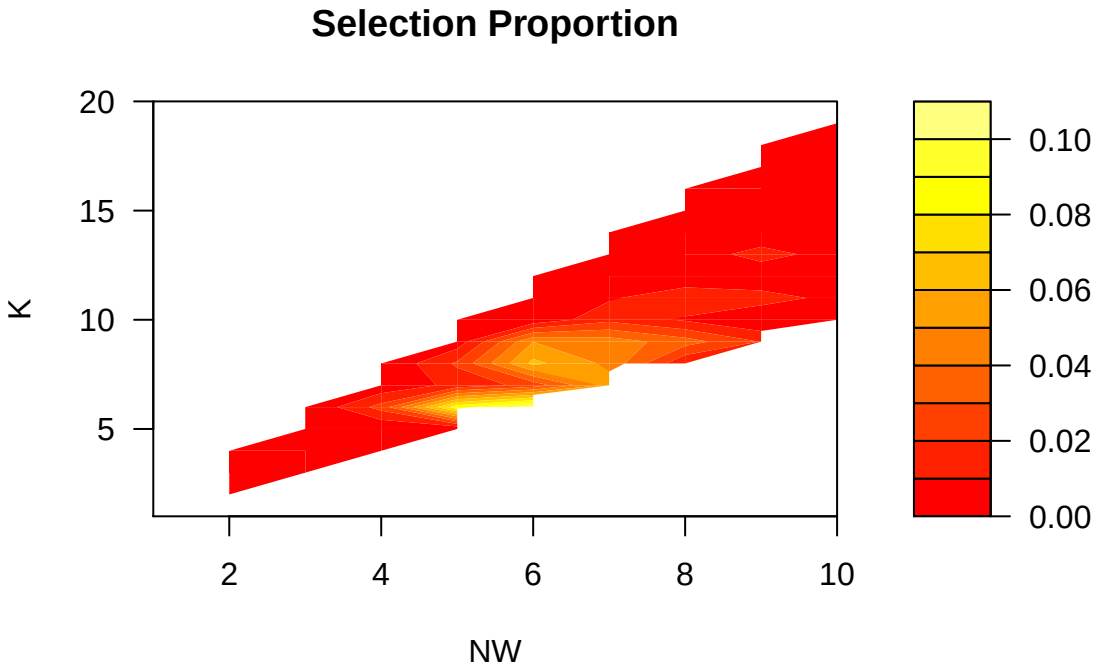

Figure 7: Variance in the the Minimum P-Value Parameter Choice from 1000 testings with $M = 50$.

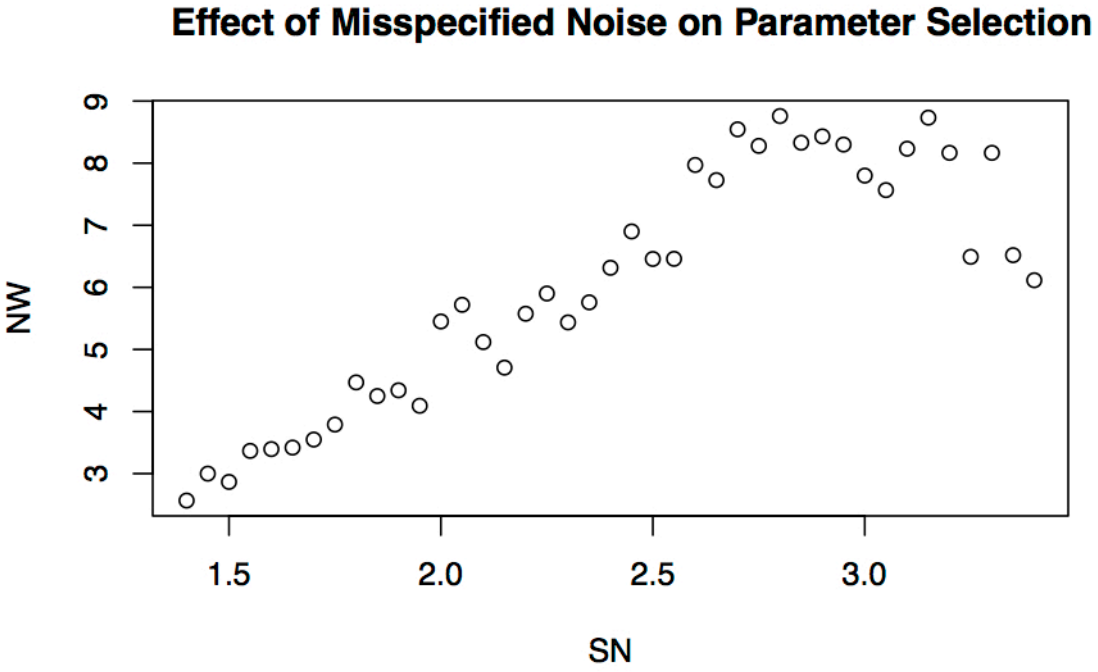

Figure 8: Effect on the Minimum P-Value Parameter from Wrongly Specifying the Noise Process Variance.

References

[1] T.W. Anderson. *An introduction to multivariate statistical analysis*, volume 2. Wiley, 1958.

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

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

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

[5] J. Klein. *Statistical visions in time: a history of time series analysis, 1662-1938*, volume 19. Cambridge University Press, 1997.

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

[7] M.B. Priestley. *Spectral analysis and time series*, volume 1. Academic press, 1981.

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

[9] D.J. Thomson. Spectrum estimation and harmonic analysis. *Proceedings of the IEEE*, 70(09):1055–1096, 1982.

[10] R. Tsay. *Analysis of financial time series*, volume 543. Wiley, 2005.