Spectral Analysis of Time Series Data

- Rebecca Warner

Book Summary

This document briefly summarizes the main points of the book, "Spectral Analysis of Time Series Data" by Rebecca M.Warner.

1 Research Questions for Time Series analysis & Spectral Analysis studies

The core idea of the book is *variance partitioning.* The variance in a time series can be decomposed in to many components. First it can be attributed to *linear* or a *curvilinear* trend. Second there can be a *cyclical* component that can give rise to variance. Every signal of length N can be obtained by adding a set of cyclical components with the following periods or cycle lengths : $N/1, N/2, N/3, \ldots, 2$. One can attribute the variance to each of these cyclical components and then shortlist only those components that contribute significantly to the variance. Each of these cyclical components has a sum of squares associated with it that has 2 degrees of freedom. Essentially this is a form of ANOVA, in which total sum of squares for the time series is divided in to sum of squares components that account for each of the $N/2$ periodic components.

Here is some basic terminology of spectral analysis :

- Amplitude
- Period
- Frequency : It is convenient to talk about "Which frequencies contribute to maximum variance"? than "Which time periods contribute to maximum variance"?
- Phase
- Harmonic analysis : This entails fitting a single sinusoid to the data and estimating the amplitude, phase, period and the mean for the same.
- Periodogram / Spectral analysis : This entails fitting a complete set of $N/2$ sinusoids to the signal.
- Coherence

Most of the econometricians deseasonalize data and then build models. This is the right approach as fitting a model with out deseasonalizing can throw up features that are sometimes completely at odds with the real world. The classic case is fitting a GARCH model with out deseasonalizing intraday volatility. The volatility persistence effect is attenuated by the misspecification. In any case, fitting sinusoidal is not the only approach. In fact the wavelet approach is more applicable to series that show non stationary behavior. In any case, this book is about fitting sinusoids and doing a ANOVA type analysis across various components of a time series.

2 Issues in Time-Series Research Design, Data Collection and Data Entry

From a spectral analysis perspective, there are two issues that the author stresses on. The first issue is that length of time series should be at least 5 to 10 times the cycle length that the researcher is interested in. The second issue is the aspect of sampling frequency. If there is no adequate sampling of the signal, a high frequency signal might appear like a low frequency signal. Hence one needs to take steps to avoid "aliasing". Thirdly if the researcher wants to detect a lagged response between two variables, then the lag value will influence the sampling frequency.

3 Preliminary Examination of Time-Series data

One sometimes tends to downplay the preprocessing stage of analysis, and this leads to incorrect analysis and inference. This chapter provides the reader a set of basic questions that needs to be explored in the preprocessing stage :

- Are there outliers ? Few outliers regularly spaced can make you falsely conclude that the frequency corresponding to those outliers is a pattern in the time series
- Is there a sufficient variance in the data so that spectral analysis makes sense ? If not, the amplitudes of sinusoids might be so small that there might not be anything meaningful.
- Does the time series show predictable patterns ?
- How stationary is the series ?

The standard tools available to check for sources of pattern are autocorrelation function, Box-Ljung Q test, Durbin Watson test for lag 1 correlation. This chapter introduces an example and shows the steps one needs to follow before doing a spectral analysis. An OLS regression is used to remove the trend or curvilinear trend. The residuals of the model are then checked for the checked for the violation of "white noise" hypothesis. ACF and Box-Ljung tests are used to reject the white noise hypothesis. Subsequently ANOVA test is carried out on the residuals. One of the requirements for the spectral analysis is the homogenity of variance. In the example shown, Levene's test is used to check the homogenity of variance. In the violation of this assumption, typically two approaches are taken: one, is where this fact is sidelined and the usual spectral analysis is carried out. In the second approach, complex modulation is used to account for this varying amplitude.
The takeaways from the chapter are

- Periodogram and spectral analysis describe the average amplitude of the cycle. These analysis by themselves obscure the fact that the amplitude is changing over time.
- If the trend is not removed, it produces artificial patterns in the periodogram
- It is better to use OLS than differencing methods to remove the trend. The reason is that differencing methods tend to over-correct for the trend.

4 Harmonic analysis

This is fancy name for fitting a sinusoid to the data with a known period. This involves estimating the amplitude, the phase and mean of the sinusoid via OLS. Typically one never knows the frequency of the signal right away. One tries to plot a periodogram or do a spectral analysis to zero in one the key frequencies. Once these frequencies are selected , a Harmonic analysis is done to estimate the parameters of the sinusoids representing those frequencies.

The basic equation involved in harmonic analysis is

$$y_t = \mu + A\cos(\omega t) + B\sin(\omega t) + \epsilon_t = \mu + A\cos(2\pi t/\tau) + B\sin(2\pi t/\tau) + \epsilon_t$$

For a given frequency or a time period , the sine and cosine function are orthogonal and they form a basis for representing any signal with time period τ. In fact, the Discrete Fourier Transform that one comes across in the study of fourier analysis is nothing but a representation of the signal via complex exponentials where each complex exponential is a component of the basis. The estimation of coefficients is a piece of cake as the coefficients can be easily evaluated based on the inner product between the signal and the respective cos and sin terms. The amplitude of the sinusoid is $\sqrt{A^2 + B^2}$. This amplitude gives a measure of how much the time series fluctuates above and below the mean. This chapter uses a simulated time series with time period (τ) of 7 days and then uses OLS to estimate amplitude, mean and phase. Once this is done, the residuals are tested for white noise hypothesis. Obviously there are issues to be considered beyond harmonic analysis such as

- If the series has a linear or curvilinear trend, then it must be handled before harmonic analysis.
- If the time period of the signal is not known, one should resort to periodogram or spectrum analysis before doing harmonic analysis.
- Outliers need to treated.
- If there are more that two cyclic components, a complex description of the process is needed to perform harmonic analysis.
- If the data is not stationary, then methods such as complex demodulation might be needed.

5 Periodogram analysis

In any signal of length N, one can fit $N/2$ sinusoids and hence there needs to be a way to select the appropriate true sinusoids of the signal. The tool used is called periodogram where a plot between the frequency and the intensity or power of the signal at various frequencies is plotted $(2/N \cdot (A^2 + B^2))$. Note that the energy of a periodic signal is infinite and hence in such cases, one tends to plot the power of the signal. In any signal that is driven by certain periodicities, the periodogram shows the peaks at precisely those points. However one needs to be careful about spurious patterns. Hence one needs some statistical test to check the null that the signal is white noise. Fisher's g test can be used for hypothesis testing. The key idea behind g test is that it checks for the proportion of intensity accounted for a specific frequency and then decides whether the observed peak at a specific frequency is random or not. The test has critical values that vary based on the rank of the frequencies being tested. This theory is followed up with some simulated data and analysis.

The chapter mentions an important point about *leakage*. This arises when the sample length is small and cannot capture the true frequency of the signal. In this case, the periodogram shows spikes at the nearby frequencies. The basic point here is that one should not blindly follow the pattern from the periodogram. A grid search around the peak should be done to check for any leakage of frequencies. The takeaway is that periodogram is plagued with sampling error. The problem with periodogram can be addressed by power spectrum analysis. This involves smoothing of the periodogram using statistical analysis. The spectrum analysis gives a far better picture about the distribution of power over a set of frequencies than a periodogram. However the downside with the spectrum analysis is that you cannot pinpoint the exact frequency and do a harmonic analysis. Well, not all is rosy about sticking to periodogram only. One needs to deal with *leakage* problem. Also one needs to do grid search and repeat harmonic analysis till one finds the best fitting periods.One thing I like about this book is that each chapter ends with a brief review of the concepts discussed and also provides a teaser to the next chapter. This kind of approach keeps the reader interested in the whole experience of going over the book.

6 Spectral analysis

Given the periodogram is discrete and has problems associated with it such as *leakage*, it is better to smoothen the discrete estimate. This is done by choosing different window functions and window widths. This is similar to Kernel density estimation methods where one chooses a kernel density function along with a specific width to smoothen the histogram. Obviously there are methods to choose the bandwidth of the kernel smoother. In fact in the "density estimation" literature, one learns that it does not matter which Kernel function is used as long as an appropriate bandwidth is chosen. This choice of bandwidth is usually driven by some cross validation kind of exercise.
The author mentions two aspects of smoothing, one being the *width* and second being the *shape*. Various windows can be used such as Hamming, Bartlett, Parzenm Tukey, Daniell. The decision to choose the width is clearly a bias-variance trade off situation. The chapter also mentions Bartlett window where FFT is done on ACF of a specific lag length M. Since power spectrum in itself is an estimate, one needs to have an handle on the confidence bands. For computing the confidence bands, one needs to know the effective degrees of freedom. Computing this for Danielle's window is easy than Tukey-Hamming window, which typically includes some numerical computation. The logical step after getting the confidence bands for the power spectrum is to test the null hypothesis of white noise.

An estimated power spectrum is a smoothed periodogram. How should one choose the frequencies ? Should they be based on periodogram or power spectrum. Here are some guidelines provided by the author :

- If you have multiple time series, you can choose periodogram as it helps you quickly validate the observed frequencies across a range of datasets. However if you have just one time series, it is better to go for power spectrum analysis as it reduces the sampling error of the periodogram
- Statistical testing of various frequencies in periodogram is done via Fisher's g test which is conservative in nature
- Sometimes independent sources of information can be used to decide between periodogram result and power spectrum result
- A straightforward way is to use the respective frequencies of the output and try to check how good the fit is to the actual data. Reconstructing the sinusoids and doing a harmonic analysis would be a good way to choose between the frequencies suggested by power spectrum and periodogram.

7 Summary of issues for univariate time-series data

This chapter gives some guidelines for reporting spectral analysis output. It also addresses the topic of analyzing multiple time series. There are two ways to aggregate the time series. At each time interval, you can aggregate data or aggregate periodogram across subjects. The choice between these approaches is context specific and subjective in nature. There are three methods mentioned in the chapter that are useful for assessing the changes in cycle parameters across time

1. Analysis of Segments of Time Series : This method comprises doing analysis over various segments and comparing relevant statistics across various time segments
2. Complex demodulation : This involves plotting a relevant parameter over time
3. Band pass filtering : This involves applying a band pass filter to a specific segment of signal and then analyzing the bandpass signal for different segments of signal.

8 Assessing Relationships between Two Time Series

This chapter presents some of the basic tools to assess the relationship between any two time series. These fall under the context of time domain analysis plus a combination of univariate spectral analysis. One of the common tools to assess the relationship between two time series is the unlagged Pearson correlation coefficient. The problems with using correlation as a statistic to estimate dependency is :

1. Within timseries observations may not be independent. One has to prewhiten the series to remove this dependency
2. Serial dependence with in the time series can manifest as a spurious correlation between the time series. If two time series are positively trending with time, then there is a positive correlation. If the two time series are negatively trending with time, then there is a negative correlation. It is not about causality at all. In fact if causuality has to be explained, then the residuals of the two time series need to be examined. Any type of serial dependence such as trends and cycles produce spurious correlation. If you take any set of time series of economic indicators, you will find some correlation. This correlation could be spurious arising out of common shared trends rather than anything else.
3. Both time series can be correlated at some other lag : Between two time series, one might be interested in computing CCF. However it is suggested that computing CCF should aways be on prewhitened series so that there is no bias in the reported CCF estimates

The author lists three common findings amongst the massive literature on behavioral time series

- Most of the time series are nonrandom,
- Most of the models have been fitted using ARIMA / categorical time series etc.
- Most of the analysis involves prewhitening the series before doing any analysis

This chapter gives a set of of guidelines that one needs to follow for reporting a bivariate time series analysis. They are

- Conduct a univariate analysis of both time series.
- Report the shared trends between the two time series, if there are reasons to believe that these are not merely artifactual.
- Report the presence of coordinated cycles between the two time series , if there are reasons to believe that these are not merely artifactual.
- Report whether the residuals are correlated between the two time series.

The takeaway from this chapter is standard method of removing trend and cycles before analyzing the two time series might not be the only approach. In many instances, the analyst is interested in the overall dependency on cycles or trends and in all such cases, it is better to perform analysis on raw series than the prewhitened series. Another takeaway is that even if you do prewhitening and analyze the series, it is possible that the residuals are driven by a third factor and there is an exogenous variable creating spurious correlation between the time series.

9 Cross -Spectral Analysis

Cross spectrum analysis is a kind of exploratory analysis tool. If the data analyst does not know a priori what frequency bands account for most of the variance in each time series and/or does not know the time lag between the two time series, cross spectral analysis provides estimates of these across all the $N/2$ frequencies.The components of the cross-spectrum analysis are coherence and phase.
The following are the three questions of interest :

1. What proportion of the variance in each of the two individual time series is accounted for by this particular frequency band ? This question can be answered via periodogram or spectrum analysis for each of the univariate time series
2. Within this frequency band, how highly correlated are the pair of time series ? This is estimated via Coherence.
3. Within each of frequency bands, what is the phase relationship of time lag between the series ? This is estimated via Phase.

The way one can verbalize the whole aspect of cross spectral analysis is as follows :

> Let's say you have two time series X and Y . You have removed the trend component from each of the time series. Now let's say you draw a periodogram for each of the time series. There are two estimates that cross spectral analysis gives, one is the coherence and the second is the phase. Coherence talks about the percentage of variation in X accounted by variation in Y for a specific frequency band. Phase talks about the lead lag relationship between X and Y for a specific frequency band. These two estimates for, let's say series X, are useful and interpretable only for those frequencies that form a significant portion of total variance of the series X. So, the first filter is the set of frequencies that contribute significantly to the total variance. Out of these frequencies, the coherence gives a measure of dependency amongst the two series. If the coherence is large, only then it makes sense to look at the phase estimate to get an idea of the lead-lag relationship.

How does on compute Cross spectrum ?
Coherence :The idea is to compute the cross covariance function, take the fourier transform of it , smoothen it via one of the windows. The squared coherence between the time series X and Y at frequency ω is computed from the cross-spectra and the individual spectral of X and Y as follows :

$$s_{x,y}(\omega)^2 = \frac{g_{x,y}(\omega)^2}{g_{x,x}(\omega)^2 g_{y,y}(\omega)^2}$$

where $g_{x,y}$ is the cross-spectrum and $g_{x,x}, g_{y,y}$ are the spectrum for the individual series.
Phase :The phase of cross spectrum is computed as

$$\phi_{x,y}(\omega) = \arctan\left\{\frac{\mathrm{Re}\, g_{x,y}(\omega)}{\mathrm{Im}\, g_{x,y}(\omega)}\right\}$$

10 Applications of Bivariate Time-Series and Cross-Spectral Analyses

This chapter gives two examples that combines all the concepts from the previous two chapters, i.e cross correlation function, coherence and phase. The detailed commentary for each of the examples allows any reader to thoroughly understand the application of principles described in the book. The fact that the book also gives the data in the appendix, for the two examples, means that one can actually independently analyze the datasets and verify the results presented in the chapter.

11 Pitfalls for the Unwary : Examples of Common Sources of Artifact

This chapter is priceless as it shows the many examples where a periodogram or a spectrum analysis can be spurious in nature. One way to assess whether the peak in the frequency domain is a spurious one or not, is to basically plot the signal and the time series together and check for a fit. Datasets are simulated and commentary is provided for the observed periodogram. The datasets for which the periodograms are analyzed are :

- White noise
- White noise + Linear trend
- White noise + 2 outliers
- Effect of including baseline periods before and after a treatment. The data is typically in the shape of boxcar
- Truncated dataset where the number of observations in the time series is not a integer multiple of the cycle length