STAT 153 - Introduction to Time Series Lecture Six

Fall 2022, UC Berkeley

Aditya Guntuboyina

September 14, 2022

1 Recap

In the last class, we discussed the problem of detecting sinusoidal oscillations in observed time series data. The standard approach for doing this involves calculating the Discrete Fourier Transform of the data and then the Periodogram. We assume here that time is denoted by $0, 1, \dots, n-1$ (where n is the length of the observed time series). Fourier frequencies are given by $0, \frac{1}{n}, \frac{2}{n}, \dots, \frac{n-1}{n}$. Sinusoidal vectors of complex exponentials with Fourier frequencies sampled at the times $0, \dots, n-1$ are denoted by $u^0, \dots, u^{n-1}$. These are defined as:

$$u^j = (1, \exp(2\pi ij/n), \exp(2\pi i2j/n), \dots, \exp(2\pi i(n-1)j/n))^T.$$

We proved in the last class that these vectors are orthogonal i.e.,

$$\left\langle u^j, u^k \right\rangle = 0 \qquad \text{whenever } 0 \le j \ne k \le n-1, \tag{1}$$

and also that $\langle u^j, u^j \rangle = n$. Here the inner product between two complex valued vectors $w = (w_0, \dots, w_{n-1})^T$ and $v = (v_0, \dots, v_{n-1})^T$ is given by

$$\langle w, v \rangle = \sum_{j=0}^{n-1} w_j \bar{v}_j.$$

As a result of this orthogonality, $u^0, \dots, u^{n-1}$ form a basis for the space of all vectors (with complex entries) of length n. This means that every dataset $y_0, \dots, y_{n-1}$ can be written as a linear combination of $u^0, \dots, u^{n-1}$ and, further, the coefficients of the linear combination are easy to find. This motivates the definition of the DFT. More precisely, the DFT of the dataset $y_0, \dots, y_{n-1}$ is defined by

$$b_j = \langle y, u^j \rangle = \sum_{t=0}^{n-1} y_t \exp\left(-\frac{2\pi ijt}{n}\right) \qquad \text{for } j = 0, 1, \dots, n-1. \tag{2}$$

It is easy to check that

1. $b_0 = y_0 + \dots + y_{n-1}$
2. $b_{n-j} = \overline{b_j}$ when $y_0, \dots, y_{n-1}$ are all real numbers.

One can recover the original dataset $y_0, \dots, y_{n-1}$ from its DFT $b_0, \dots, b_{n-1}$ using the formula:

$$y_t = \frac{1}{n} \sum_{j=0}^{n-1} b_j \exp\left(\frac{2\pi i j t}{n}\right) \qquad \text{for } t = 0, 1, \dots, n-1. \tag{3}$$

Note that the two formulae (2) and (3) are quite similar; the differences being in the sign of the exponent in the complex exponential and the presence of the factor $1/n$ in (3). (2) is known as the DFT formula while (3) is known as the Inverse DFT formula.

Formula (3) becomes

$$y = \frac{1}{n} \sum_{j=0}^{n-1} b_j u^j \tag{4}$$

in vector form (note $y = (y_0, y_1, \dots, y_{n-1})^T$) and this encapsulates the idea that y can be written as a linear combination of the basis sinusoids $u^0, u^1, \dots, u^{n-1}$.

2 Periodogram

The Periodogram is a tool for visualizing the DFT (note that the DFT coefficients b_j can be complex). The periodogram is defined by

$$I\left(\frac{j}{n}\right) := \frac{|b_j|^2}{n} \qquad \text{for } 0 < \frac{j}{n} \leq \frac{1}{2}.$$

One visualizes the size of the DFT terms by plotting the periodogram. Note that $j = 0$ is not plotted as b_0 is simply the sum of the data values and does not provide any information on the sinusoidal components present in the data. Further, the periodogram $I(j/n)$ is only plotted for $(j/n) \leq 0.5$ as b_j's for larger values of j are determined by earlier b_j's by complex conjugacy.

Because

$$b_j = \sum_{t=0}^{n-1} y_t \exp\left(-\frac{2\pi i j t}{n}\right) = \sum_{t=0}^{n-1} y_t \cos\frac{2\pi j t}{n} - i \sum_{t=0}^{n-1} y_t \sin\frac{2\pi j t}{n},$$

we can write the periodogram as:

$$I\left(\frac{j}{n}\right) = \frac{1}{n}\left[\left(\sum_{t=0}^{n-1} y_t \cos\frac{2\pi j t}{n}\right)^2 + \left(\sum_{t=0}^{n-1} y_t \sin\frac{2\pi j t}{n}\right)^2\right] \qquad \text{for } 0 < \frac{j}{n} \leq \frac{1}{2}. \tag{5}$$

We can view the periodogram ordinates as a decomposition of the sample variance. More precisely, because of the following fact, we can view $2I(j/n)$ as the portion of sample variance that is explained by sines and cosines at frequency j/n.

Fact 2.1. *The periodogram ordinates $I(j/n)$ for the dataset $y_0, \dots, y_{n-1}$ satisfy:*

$$\sum_{t=0}^{n-1} (y_t - \bar{y})^2 = 2 \sum_{j=1}^{(n/2)-1} I\left(\frac{j}{n}\right) + I\left(\frac{1}{2}\right) \qquad \textit{if } n \textit{ is even},$$

and

$$\sum_{t=0}^{n-1} (y_t - \bar{y})^2 = 2 \sum_{j=1}^{(n-1)/2} I\left(\frac{j}{n}\right) \qquad \textit{if } n \textit{ is odd}.$$

Proof of Fact 2.1. Because of (4), we can write

$$\begin{aligned}
\sum_{t=0}^{n-1} y_t^2 = \|y\|^2 &= \left\| \frac{1}{n}\sum_{j=0}^{n-1} b_j u^j \right\|^2 \\
&= \left\langle \frac{1}{n}\sum_{j=0}^{n-1} b_j u^j, \frac{1}{n}\sum_{k=0}^{n-1} b_k u^k \right\rangle \\
&= \frac{1}{n^2}\sum_{j=0}^{n-1}\sum_{k=0}^{n-1} b_j \overline{b_k} \left\langle u^j, u^k \right\rangle \\
&= \frac{1}{n^2}\sum_{j=0}^{n-1} |b_j|^2 \left\langle u^j, u^j \right\rangle \quad \text{(by orthogonality)} \\
&= \frac{1}{n^2}\sum_{j=0}^{n-1} |b_j|^2 \times n = \frac{1}{n}\sum_{j=0}^{n-1} |b_j|^2.
\end{aligned}$$

Now because $b_0 = y_0 + \cdots + y_{n-1} = n\bar{y}$, we can move b_0^2 to the left hand side to get

$$\sum_{t=0}^{n-1} (y_t - \bar{y})^2 = \sum_{t=0}^{n-1} y_t^2 - n\bar{y}^2 = \frac{1}{n}\sum_{j=1}^{n-1} |b_j|^2.$$

Now we use complex conjugacy $b_j = \overline{b_{n-j}}$ to write the sum above in terms of $1 \le j \le n/2$. When n is even, we get

$$\begin{aligned}
\sum_{t=0}^{n-1} (y_t - \bar{y})^2 &= \frac{1}{n}\sum_{j=1}^{n-1} |b_j|^2 \\
&= \frac{1}{n}\sum_{j=1}^{(n/2)-1} |b_j|^2 + \frac{1}{n}|b_{n/2}|^2 + \frac{1}{n}\sum_{j=(n/2)+1}^{n-1} |b_j|^2 \\
&= \frac{1}{n}\sum_{j=1}^{(n/2)-1} |b_j|^2 + \frac{1}{n}|b_{n/2}|^2 + \frac{1}{n}\sum_{j=(n/2)+1}^{n-1} |\overline{b_{n-j}}|^2 \\
&= \frac{1}{n}\sum_{j=1}^{(n/2)-1} |b_j|^2 + \frac{1}{n}|b_{n/2}|^2 + \frac{1}{n}\sum_{j=(n/2)+1}^{n-1} |b_{n-j}|^2 \\
&= \frac{1}{n}\sum_{j=1}^{(n/2)-1} |b_j|^2 + \frac{1}{n}|b_{n/2}|^2 + \frac{1}{n}\sum_{j=1}^{(n/2)-1} |b_j|^2 \\
&= \frac{2}{n}\sum_{j=1}^{(n/2)-1} |b_j|^2 + \frac{1}{n}|b_{n/2}|^2 = 2\sum_{j=1}^{(n/2)-1} I\left(\frac{j}{n}\right) + I\left(\frac{1}{2}\right).
\end{aligned}$$

The proof for n odd is very similar and is left as exercise. □

The definition (5) of the periodogram is for Fourier Frequencies. For more general frequencies $f \in (0, 1/2]$, we can define the periodogram ordinate as:

$$I(f) := \frac{1}{n}\left[\left(\sum_{t=0}^{n-1} y_t \cos(2\pi f t)\right)^2 + \left(\sum_{t=0}^{n-1} y_t \sin(2\pi f t)\right)^2\right] \qquad \text{for } 0 < f \le \frac{1}{2}.$$

Note also that throughout, we assumed that the time points are equally spaced between $0, 1, \ldots, n-1$. In more general situations, when the observed time series values $y_1, \ldots, y_n$ are sampled at arbitrary times $t_1, \ldots, t_n$, we can define the Periodogram as

$$I(f) := \frac{1}{n}\left[\left(\sum_{i=1}^{n} y_i \cos(2\pi f t_i)\right)^2 + \left(\sum_{i=1}^{n} y_i \sin(2\pi f t_i)\right)^2\right] \qquad \text{for } -\infty < f < \infty.$$

Note that now f is ranging over all real numbers whereas in the case where $t = 0, 1, \ldots, n-1$, we argued in the last class that there is no point in considering frequencies outside the range $[0, 1/2]$.

3 DFT for Sinusoids

To understand the DFT, let us calculate the DFT of the cosine wave $x_t = R\cos(2\pi f_0 t + \phi), t = 0, 1, \ldots, n-1$. In the last class, we saw that we can without loss of generality take $0 \leq f_0 \leq 1/2$.

3.1 Case 1: When f_0 is a Fourier frequency

Suppose f_0 is of the form k/n for some k where $0 \leq k/n \leq 1/2$. Then the DFT is given by

$$\begin{aligned} b_j &= \sum_{t=0}^{n-1} R\cos(2\pi(k/n)t + \Phi)\exp\left(-2\pi i(j/n)t\right) \\ &= \frac{Re^{i\Phi}}{2}\sum_{t=0}^{n-1}\exp\left(2\pi i t\frac{k-j}{n}\right) + \frac{Re^{-i\Phi}}{2}\sum_{t=0}^{n-1}\exp\left(-2\pi i t\frac{j+k}{n}\right). \end{aligned}$$

Note that we do not need to consider the DFT b_j for $j/n > 1/2$. So we assume that $0 \leq j/n \leq 1/2$. Because the original cosine wave was assumed to have frequency in the range $[0, 1/2]$, we have $0 \leq k/n \leq 1/2$.

We need to consider the case $j = k$ and $j \neq k$ here. When $j \neq k$, both terms above are zero (because of the geometric sum formula).

When $j = k \neq n/2$, the second term is zero while the first term equals $Re^{i\Phi}n/2$. When $j = k = n/2$ (this can only happen if n is even), we get $b_k = nR(e^{i\Phi} + e^{-i\Phi})/2 = nR\cos\Phi$. Thus

$$b_j = \begin{cases} nRe^{i\Phi}/2 & \text{if } j = k \neq n/2 \\ nR\cos\Phi & \text{if } j = k = n/2 \\ 0 & \text{for all other } 1 \leq j \leq n/2 \end{cases}$$

The periodogram thus equals

$$I\left(\frac{j}{n}\right) = \begin{cases} \frac{nR^2}{4} & \text{if } j = k \neq n/2 \\ nR^2\cos^2\Phi & \text{if } j = k = n/2 \\ 0 & \text{for all other } 1 \leq j \leq n/2 \end{cases}$$

3.2 Multiple Fourier Frequencies

Now consider data that is linear combination of multiple Fourier frequencies:

$$x_t = \sum_{l=1}^{m} R_l \cos(2\pi t(k_l/n) + \Phi_l) \tag{6}$$

where each k_l is an integer satisfying $0 \leq k_l/n \leq 1/2$. Because the definition of the DFT is linear in the data $\{x_t\}$, it follows that the DFT of (6) is given by

$$b_j = \begin{cases} nR_l e^{i\Phi_l}/2 & \text{if } j = k_l \neq n/2 \\ nR_l \cos\Phi_l & \text{if } j = k_l = n/2 \\ 0 & \text{otherwise} \end{cases}$$

for $0 \leq j/n \leq 1/2$. Similarly

$$I(j/n) = \begin{cases} nR_l^2/4 & \text{if } j = k_l \neq n/2 \\ nR_l^2 \cos^2\Phi_l & \text{if } j = k_l = n/2 \\ 0 & \text{otherwise} \end{cases}$$

This shows that the DFT picks out the frequencies present in the data. The strength (absolute value) of the DFT at a frequency is proportional to the amplitude (R_l) of the cosine wave at that frequency.

3.3 DFT of a sinusoid at a non-Fourier frequency

The DFT of a sinusoid at a non-Fourier frequency is calculated in the following way: Consider the signal $x_t = e^{2\pi f_0 t}$ where $f_0 \in [0, 1/2]$ is not necessarily of the form k/n for any k. Its DFT is given by

$$b_j := \sum_{t=0}^{n-1} x_t e^{-2\pi i t(j/n)} = \sum_{t=0}^{n-1} e^{2\pi i(f_0-(j/n))t}.$$

If we denote the function

$$S_n(g) := \sum_{t=0}^{n-1} e^{2\pi i g t} \tag{7}$$

then we can write

$$b_j = S_n(f_0 - (j/n)).$$

The function $S_n(g)$ can clearly be evaluated using the geometric series formula to be

$$S_n(g) = \frac{e^{2\pi i g n} - 1}{e^{2\pi i g} - 1}$$

Because

$$e^{i\theta} - 1 = \cos\theta + i\sin\theta - 1 = 2e^{i\theta/2}\sin\theta/2,$$

we get

$$S_n(g) = \frac{\sin \pi n g}{\sin \pi g} e^{i\pi g(n-1)}$$

Thus the absolute value of the DFT of $y_t = e^{2\pi i f_0 t}$ is given by

$$|b_j| = |S_n(f_0 - (j/n))| = \left| \frac{\sin \pi n(f_0 - (j/n))}{\sin \pi(f_0 - (j/n)))} \right|$$

This expression becomes meaningless when $f_0 = j/n$ but since we are assuming that f_0 is not a Fourier frequency, we do not need to worry about this.

The behavior of this DFT can be best understood by plotting the function $g \mapsto (\sin \pi n g)/(\sin \pi g)$.

When f_0 is not of the form k/n for any k, the term $S_n(f_0 - j/n)$ is non-zero for all j. This situation when one observes a non-zero DFT term $_j$ because of the presence of a sinusoid at a frequency f_0 different from j/n is referred to as **Leakage**. Leakage due to a sinusoid with frequency f_0 not of the form k/n is present in all DFT terms b_j but the magnitude of the presence decays as j/n gets far from f_0 (this is due to the form of the function $S_n(f_0 - j/n)$).

The behavior of the DFT of the cosine wave $x_t = R\cos(2\pi f_0 t + \Phi)$ can be understood from the DFT of $x_t = e^{2\pi f_0 t}$.

4 Interpreting the Periodogram

The DFT writes the given data in terms of sinusoids with frequencies of the form k/n. Frequencies of the form k/n are called Fourier frequencies.

Suppose that we are given a dataset $x_0, \dots, x_{n-1}$. We have calculated its DFT: $b_0, b_1, \dots, b_{n-1}$ and we have plotted $|b_j|^2/n$ for $j = 1, \dots, (n-1)/2$ for odd n and for $j = 1, \dots, n/2$ for even n.

If we see a single spike in this plot, say at b_k, we are sure that the data is a sinusoid with frequency k/n.

If we get two spikes, say at b_{k_1} and b_{k_2}, then the data is slightly more complicated: it is a linear combination of two sinusoids at frequencies k_1/n and k_2/n with the strengths of these sinusoids depending on the size of the spikes.

Multiple spikes indicate that the data is made up of many sinusoids at Fourier frequencies and, in general, this means that the data is more complicated.

However, sometimes one can see multiple spikes in the periodogram even when the structure of the data is not very complicated. A typical example is leakage due to the presence of a sinusoid at a non-Fourier frequency.

5 Periodogram vs Model-based Methods

The periodogram is an exploratory data analysis tool and is not tied to any formal model. It gives an idea of the frequencies of sinusoidal cycles present in the data. However, for formal inference of frequencies, it is better to use a model-based approach. For example, if we are hunting for the sinusoid which best fits the data, instead of looking at the periodogram which can have all sorts of spurious spikes, it is better to consider the model:

$$Y_i = \beta_0 + \beta_1 \cos(2\pi f t_i) + \beta_2 \sin(2\pi f t_i) + Z_t \qquad \text{with } Z_i \overset{\text{i.i.d}}{\sim} N(0, \sigma^2). \tag{8}$$

This model has 5 unknown parameters: $\beta_0, \beta_1, \beta_2, \sigma, f$ of which our main interest lies in f. We can write the model (8) in vector form as:

$$Y = X(f)\beta + Z \qquad \text{where } Z \sim N_n(0, \sigma^2 I_n). \tag{9}$$

Here Y is the $n \times 1$ vector consisting of the observed time series values $y_1, \dots, y_n$, $X(f)$ is the $n \times p$ (with $p = 3$) matrix given by

$$X(f) = \begin{pmatrix} 1 & \cos(2\pi f t_1) & \sin(2\pi f t_1) \\ 1 & \cos(2\pi f t_2) & \sin(2\pi f t_2) \\ . & . & . \\ . & . & . \\ . & . & . \\ 1 & \cos(2\pi f t_n) & \sin(2\pi f t_n) \end{pmatrix}$$

Also $Z \sim N_n(0, \sigma^2 I_n)$ means that Z has the n-dimensional multivariate normal distribution with mean vector consisting of all zeros and covariance matrix σ^2 multiplied by the identity matrix (this is same as saying that $Z_1, \dots, Z_n$ are i.i.d $N(0, \sigma^2)$). We performed Bayesian inference under this model in Lecture Four under a simple uniform prior and calculated the posterior density for f as:

$$\text{posterior density for } f \propto \|Y - X(f)\hat{\beta}(f)\|^{-(n-p)} \left|X(f)^T X(f)\right|^{-1/2}. \tag{10}$$

where $\hat{\beta}(f)$ is the least squares estimator for the regression of Y on $X(f)$:

$$\hat{\beta}(f) := \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \|Y - X(f)\beta\|^2.$$

In the formula (10), the factor $\left|X(f)^T X(f)\right|^{-1/2}$ is usually not important and the posterior density is mainly driven by the first term:

$$\|Y - X(f)\hat{\beta}(f)\|^{-(n-p)}$$

This term is usually highly peaked because

$$\|Y - X(f)\hat{\beta}(f)\|^{-(n-p)} \propto \left(\frac{\|Y - X(\hat{f})\hat{\beta}(\hat{f})\|^2}{\|Y - X(f)\hat{\beta}(f)\|^2}\right)^{(n-p)/2} \tag{11}$$

where $\hat{f}$ is the value of f which minimizes $\|Y - X(f)\hat{\beta}(f)\|$ over all f (this is the frequency which gives the best fit to the observed data). From (11), it is clear that only those frequencies f for which $\|Y - X(f)\hat{\beta}(f)\|$ is very close to $\|Y - X(\hat{f})\hat{\beta}(\hat{f})\|$ get high posterior weight. Generally this will only be the case for f which are very close to $\hat{f}$. To summarize, unlike the periodogram which may show multiple spurious peaks, the Bayesian posterior (10) generally has a single sharp peak around $\hat{f}$.

It is helpful to note that the posterior density formula (10) holds more generally for any model that is of the form (9). This allows us to perform inference in many other interesting models. For example, suppose we want to fit a two-component sinusoidal model to the data. Then we would consider the model:

$$Y_i = \beta_0 + \beta_1 \cos(2\pi f_1 t_i) + \beta_2 \sin(2\pi f_1 t_i) + \beta_3 \cos(2\pi f_2 t_i) + \beta_4 \sin(2\pi f_2 t_i) + Z_i \tag{12}$$

with, as before, $Z_i \overset{\text{i.i.d}}{\sim} N(0, \sigma^2)$. This model is again of the form (9) with $f = (f_1, f_2)$ and $X(f)$ given by

$$X(f_1, f_2) = \begin{pmatrix} 1 & \cos(2\pi f_1 t_1) & \sin(2\pi f_1 t_1) & \cos(2\pi f_2 t_1) & \sin(2\pi f_2 t_1) \\ 1 & \cos(2\pi f_1 t_2) & \sin(2\pi f_1 t_2) & \cos(2\pi f_2 t_2) & \sin(2\pi f_2 t_2) \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 1 & \cos(2\pi f_1 t_n) & \sin(2\pi f_1 t_n) & \cos(2\pi f_2 t_n) & \sin(2\pi f_2 t_n) \end{pmatrix}$$

The number of columns of $X(f_1, f_2)$ is $p = 5$. Posterior inference for $f = (f_1, f_2)$ in the model (12) can therefore be also carried out by the formula (10). Note now though that this is a bivariate posterior density.

More generally, inference on the frequency parameters for a k-component sinusoidal model can be done, in principle, by considering $X(f)$ to be the $n \times (1 + 2k)$ matrix which columns given by $1, \cos(2\pi f_1 t), \sin(2\pi f_1 t), \cos(2\pi f_2 t), \sin(2\pi f_2 t), \dots, \cos(2\pi f_k t), \sin(2\pi f_k t)$ as t varies over $t_1, \dots, t_n$. This will give us a k-dimensional posterior density for $f = (f_1, \dots, f_k)$.

One can also use this methodology for other models with applications such as change-point inference. For example, for change-point inference, it is natural to consider the model:

$$Y_i = \beta_0 + \beta_1 I\{t_i > f_1\} + \beta_2 I\{t_i > f_2\} + \dots \beta_k I\{t_i > f_k\} + Z_i \tag{13}$$

with $Z_i \overset{\text{i.i.d}}{\sim} N(0, \sigma^2)$. Here $I\{t_i > f_j\}$ denotes the indicator function which equals the value 1 when $t_i > f_j$ and 0 otherwise. The function:

$$t \mapsto \beta_0 + \beta_1 I\{t > f_1\} + \beta_2 I\{t > f_2\} + \dots \beta_k I\{t > f_k\}$$

is piecewise constant with k jumps at the locations $f_1, \dots, f_k$. These $f_1, \dots, f_k$ are also known as change-points. The model (13) can be written in the general form (9) with $X(f_1, \dots, f_k)$ being the $n \times (k+1)$ matrix with columns $1, I\{t > f_1\}, \dots, I\{t > f_k\}$ for $t = t_1, \dots, t_n$. Thus posterior inference for the change-points $f_1, \dots, f_k$ can be carried out using (10).

6 Recommended List of Readings for Today

All the books mentioned below are freely available online (let me know if you are not able to find them).

1. The DFT and periodogram are standard topics in time series analysis and can be found in many books. For example, Section 4.3 of the book "Time Series Analysis and its Applications (4th edition)" by Shumway and Stoffer. Note that Shumway and Stoffer define the DFT with a additional multiplicative factor of $n^{-1/2}$. A book focussing on the DFT for time series analysis is "Fourier Analysis of Time Series" by Bloomfield.

2. The Bayesian approach for sinusoidal frequency inference was developed in the book "Bayesian spectrum analysis and parameter estimation" by Bretthorst. Chapter 2 of this book deals with the single sinusoid model. Application to the sunspots dataset is in Section 2.6.