

Harmonic Regression

Overview

1. Example: variable star data
2. Discrete Fourier transform (periodogram)
3. Examples of the DFT

Example: Periodic Data

Magnitude of variable star This integer time series is reported to be the magnitude of a variable star observed on 600 successive nights (Whittaker and Robinson, 1924, `varstar.dat`).

Periodic model Classical example of “signal in noise,” with a hidden (though not very hidden in this case) periodic signal. Model for periodic data with a single sinusoid is (Example 2.7)

$$x_t = R \cos(2\pi\omega t + \phi) + w_t$$

Watch out for w_t (white noise) versus ω (frequency, the reciprocal of the period in t units). Note that SS write A rather than R for the amplitude.

Caution: Aliasing and identifiability The frequency must be typically restricted to the range $-1/2 < \omega \leq 1/2$ in order for the model to be identified. For discrete equally spaced data, frequencies outside of this range are *aliased* into this range. For example, suppose that $\lambda = 1 - \delta$, $0 < \delta < 1/2$, then for $t = 1, 2, \dots$

$$\begin{aligned} \cos(2\pi\lambda t) &= \cos(2\pi(1 - \delta)t + \phi) = \cos(2\pi t + (-2\pi\delta t + \phi)) \\ &= \cos(2\pi t) \cos(2\pi\delta t - \phi) + \sin(2\pi t) \sin(2\pi\delta t + \phi) \\ &= \cos(2\pi\delta t - \phi); . \end{aligned}$$

Thus, a sampled sinusoid with frequency $\omega > 1/2$ appears as a sinusoid with frequency in the interval $(-1/2, 1/2]$. If the data is sampled at a spacing of Δ time units, then the highest frequency observable in

discrete data is $2/\Delta$, which is known as the *folding frequency* or *Nyquist frequency*.

Notation alert! Many books (including the ones I learned from) write the sinusoids using the angular frequency notation that embeds the 2π into the frequency as $\lambda = 2\pi\omega$. Thus, the harmonic regression would look like

$$x_t = R \cos(\lambda t + \phi) + w_t$$

Question How to estimate unknown parameters?
the amplitude R , the frequency ω and the phase ϕ .

Estimation by regression Convert the expression of the model into one more suited to OLS estimation by using the formula $\cos(a + b) = \cos a \cos b - \sin a \sin b$ which allows us to write it as the linear regression (sans intercept)

$$x_t = \underbrace{R \cos(\phi)}_{\beta_1} \cos(2\pi\omega t) + \underbrace{-R \sin(\phi)}_{\beta_2} \sin(2\pi\omega t) + w_t$$

Notice that $\beta_1^2 + \beta_2^2 = R^2$. The sum of the squared regression coefficients is the squared amplitude of the hidden sinusoid.

Frequency The estimation is linear if ω is known. What's a good estimator for ω ?

Hint: count cycles. The estimation of ω is non-standard: we can get an estimator within $o_p(1/n)$ rather than the usual root- n rate.

Regression at the Fourier Frequencies

Idea Rather than expect to count peaks (as in the variable star), fit regressions at many frequencies to find hidden sinusoids (simulated data). Use the estimated amplitude at these frequencies to locate hidden periodic components. It turns out that it is particularly easy to estimate the amplitude at a grid of evenly spaced frequencies from 0 to $1/2$.

Fourier frequencies determine the grid of frequencies. A frequency $-1/2 < \omega \leq 1/2$ is known as a Fourier frequency if the associated sinusoid

completes an integer number of cycles in the observed length of data

$$\omega_j = \frac{j}{n}, \quad j = 0, 1, 2, \dots, n/2,$$

assuming n is an even integer.

Because of aliasing, the set of j 's stop at $n/2$. Because cosine is an even function and sine is odd, we don't need to consider negative frequencies.

Orthogonality Cosines and sines at the Fourier frequencies generate an orthogonal (though not orthonormal) set of regressors. Consider the $n \times n$ matrix (assuming n is even) given by

$$X = \left(\begin{array}{c|cc|ccc|c} \vdots & \vdots & \vdots & \dots & \vdots & \vdots & \vdots \\ \cos(2\pi\omega_0 t) & \cos(2\pi\omega_1 t) & \sin(2\pi\omega_1 t) & \dots & \cos(2\pi\omega_{n/2-1} t) & \sin(2\pi\omega_{n/2-1} t) & \cos(2\pi\omega_{n/2} t) \\ \vdots & \vdots & \vdots & \dots & \vdots & \vdots & \vdots \end{array} \right)$$

Note that the first column is all 1's ($\omega_0 = 0$) and the last column alternates ± 1 ($\omega_{n/2} = \frac{1}{2}$). Then it follows that

$$X'X = \begin{pmatrix} n & 0 & 0 & \dots & 0 \\ 0 & n/2 & 0 & \dots & 0 \\ 0 & 0 & \ddots & 0 & 0 \\ 0 & \dots & 0 & n/2 & 0 \\ 0 & \dots & 0 & 0 & n \end{pmatrix}$$

The orthogonality results from the fact that

$$\sum_{t=1}^n \cos(2\pi\omega_k t) = \sum_{t=1}^n \sin(2\pi\omega_k t) = 0$$

at the Fourier frequencies. The cross-products reduce to these since, for instance,

$$\begin{aligned} \sum_{t=1}^n \cos(2\pi\omega_k t) \sin(2\pi\omega_j t) &= \frac{1}{2} \sum_{t=1}^n \sin(2\pi\omega_{k+j} t) - \sin(2\pi\omega_{k-j} t) \\ &= 0, \end{aligned}$$

For diagonal terms, notice that $\cos(a) \cos(b) = \frac{1}{2}(\cos(a+b) + \cos(a-b))$ and $\sin(a) \sin(b) = \frac{1}{2}(\cos(a-b) - \cos(a+b))$ handle the sums of squares

$$\sum_{t=1}^n \cos(2\pi\omega_j t)^2 = \frac{1}{2} \sum_{t=1}^n \cos(4\pi\omega_j t) + 1 = n/2, \quad j \neq 0, n/2.$$

Harmonic regression If we use Fourier frequencies in our harmonic regression, the regression coefficients are easily found since “ $X'X$ ” is almost diagonal. Specifically, (with A’s on the cosine terms and B’s on the sines)

$$\begin{aligned} A_0 &= \sum_t x_t/n & A_{n/2} &= \sum_t x_t(-1)^t/n \\ A_j &= \frac{2}{n} \sum_t x_t \cos 2\pi\omega_j t & B_j &= \frac{2}{n} \sum_t x_t \sin 2\pi\omega_j t. \end{aligned}$$

Note that $B_0 = B_{n/2} = 0$; there is no sine component for these terms.

Estimated amplitude The sum of squares captured by a specific sine/cosine pair is ($j \neq 0, n/2$)

$$\frac{n}{2}(A_j^2 + B_j^2) = \frac{n}{2} R_j^2.$$

That is, the amplitude of the fitted sinusoid determines the variance explained by this term in a regression model.

Since we have fit a regression of n parameters to n observations x_1, \dots, x_n , the model fits perfectly. Thus the variation in the fitted values is exactly that of the original data, and we obtain the following decomposition of the variance:

$$\sum_t X_t^2 = n(R_0^2 + R_{n/2}^2) + \frac{n}{2} \sum_j R_j^2,$$

where the weights on R_0 and $R_{n/2}$ differ since there is no sine term at these frequencies.

Vector space The data $\mathbf{X} = (x_1, \dots, x_n)'$ is a vector in n -dimensional space. The usual basis for this space is the set of vectors

$$e_j = (0 \dots 0 \ 1_j \ 0 \ 0 \dots 0)'.$$

Thus we can write $\mathbf{X} = \sum_t X_t e_t$. The harmonic model uses a different orthogonal basis, namely the sines and cosines associated with the Fourier frequencies. The “saturated” harmonic regression

$$X_t = A_0 + A_{n/2}(-1)^t + \sum_{j=1}^{n/2-1} (A_j \cos(2\pi\omega_j t) + B_j \sin(2\pi\omega_j t))$$

represents the vector \mathbf{X} in this new basis. The coordinates of \mathbf{X} in this basis are the coefficients A_j and B_j .

Discrete Fourier transform

Time Origin When dealing with the Fourier transform, it is more simple to index time from $t = 0$ to $t = n - 1$.

Definition Changing to complex variables leads to the discrete Fourier transform. The discrete Fourier transform (DFT) of the real-valued n sequence y_0, \dots, y_{n-1} is defined as

$$d(\omega_j) = D_j = \frac{1}{n} \sum_{t=0}^{n-1} y_t \exp(-i2\pi\omega_j t), \quad j = 0, 1, 2, \dots, n - 1.$$

The DFT is the set of harmonic regression coefficients, written using complex variables. For $j = 0, 1, \dots, \frac{n}{2}$,

$$\begin{aligned} D_j &= \frac{1}{n} \sum_{t=0}^{n-1} y_t \exp(-i2\pi\omega_j t) \\ &= \frac{1}{n} \sum_{t=0}^{n-1} y_t (\cos 2\pi\omega_j t - i \sin 2\pi\omega_j t) \\ &= \frac{1}{2}(A_j - iB_j) \end{aligned}$$

Matrix form The DFT amounts to a change of basis transformation. Define the $n \times n$ matrix $F_{jk} = \exp(-i2\pi jk/n)$. The inner product of columns (or rows) is (remember to conjugate the second)

$$\begin{aligned} F_j' \bar{F}_k &= \sum_{t=0}^{n-1} e^{-i2\pi jt/n} e^{i2\pi kt/n} \\ &= \sum_{t=0}^{n-1} e^{i2\pi(k-j)t/n} \\ &= \begin{cases} n & j = k \\ 0 & j \neq k \end{cases} \end{aligned}$$

The zero follows from summing the geometric series (for $h = 1, 2, \dots$)

$$\sum_{t=0}^{n-1} e^{i2\pi ht/n} = \sum_{t=0}^{n-1} \left(e^{i2\pi h/n} \right)^t$$

$$\begin{aligned}
 &= \frac{1 - e^{i2\pi h}}{1 - e^{i2\pi h/n}} \\
 &= \frac{e^{i2\pi h/2} \sin 2\pi h}{e^{i2\pi h/(2n)} \sin 2\pi h/n} = 0
 \end{aligned}$$

Hence, $F^*F = nI_n$ (* denotes the conjugate of the transpose); the matrix $n^{-1/2}F$ is *orthonormal*. Notice how much more simple — if you are okay with complex variables, that is — it is to see that the columns are orthogonal than with sines and cosines (no need for the double-angle formulas).

Normalizing There are many ways to write the transform, depending on how n appears. For example, to get something close to the $\{A_j, B_j\}$ representation, express the transform as

$$D = \frac{1}{n} F^* Y .$$

The fast Fourier transform (FFT) is a method for evaluating this matrix multiplication (which appears to be of order n^2) in order $n \log n$ steps by a clever recursion. (Wavelets, another orthogonal basis, allow calculation in order n operations.)

Symmetry & Redundancy Since we begin with n real-valued observations y_t , but obtain n complex values D_j , the DFT has a redundancy (or symmetry):

$$\begin{aligned}
 \overline{D_{n-j}} &= \frac{1}{n} \sum_{t=0}^{n-1} y_t \exp(i2\pi\omega_{n-j}t) \\
 &= \frac{1}{n} \sum_{t=0}^{n-1} y_t \exp(-i2\pi\omega_j t) \\
 &= D_j .
 \end{aligned}$$

One can exploit this symmetry to obtain the transform of two real-valued series at once from one application of the FFT.

Inversion We can recover the original data from the inverse transform,

$$\sum_j D_j \exp(i2\pi\omega_j t) = \frac{1}{n} \sum_j \sum_s y_s \exp(i2\pi\omega_j t - i2\pi\omega_j s)$$

$$\begin{aligned}
 &= \frac{1}{n} \sum_s y_s \sum_j \exp(i2\pi\omega_j(t-s)) \\
 &= y_t
 \end{aligned}$$

since $\sum_j \exp(i2\pi\omega_j(t-s)) = 0$ for $s \neq t$, and otherwise is n . Using the matrix form, multiplying both sides by F gives $\mathbf{Y} = F\mathbf{D}$ immediately.

Variance decomposition As in harmonic regression, we can associate a variance with D_j . In particular,

$$\begin{aligned}
 \sum_t y_t^2 &= \sum_t |y_t|^2 = \sum_t \left| \sum_j D_j \exp(i2\pi\omega_j t) \right|^2 \\
 &= n \sum_j D_j \bar{D}_j = n \sum_{j=0}^{n-1} |D_j|^2,
 \end{aligned}$$

which is a much "neater" formula than that offered in the real-valued harmonic regression model. In matrix form, this is easier still:

$$\sum_t y_t^2 = \mathbf{Y}^* \mathbf{Y} = (F\mathbf{D})^* (F\mathbf{D}) = \mathbf{D}^* (F^* F) \mathbf{D} = n \sum_j |D_j|^2$$

Periodogram The variance at the Fourier frequency $2\pi\omega_j$ is $\frac{n}{2}(A_j^2 + B_j^2) = 2nD_j\bar{D}_j$ for $j \neq 0, n/2$. The periodogram gives the same decomposition of variance, defined as

$$I_n(\omega_j) = \frac{n}{2\pi} D_j \bar{D}_j = \frac{n}{2\pi} |D_j|^2 = \frac{1}{2\pi n} \left| \sum_t x_t \exp(-i2\pi\omega_j t) \right|^2$$

Examples of the DFT (for future reference)

Linearity Since its just a linear transformation (change of basis), the DFT is a *linear* operator. Hence, *e.g.*, the DFT of a sum is the sum of the DFT's:

$$D_{x+y,j} = \frac{1}{n} \sum_t (x_t + y_t) \exp(-i2\pi\omega_j t) = D_{x,j} + D_{y,j}.$$

Thus, once we understand how the DFT behaves for some simple series, we can understand it for any others that are sums of these simple cases.

Convolutions If the input data are a product, $x_t = y_t z_t$, the DFT has again a very special form. Using the inverse transform we find that the transform of the product is the *convolution* of the transforms,

$$\begin{aligned} D_{x,j} &= \frac{1}{n} \sum_t y_t z_t \exp(-i2\pi\omega_j t) \\ &= \frac{1}{n} \sum_t y_t \left(\sum_k D_{z,k} \exp(i2\pi\omega_k t) \right) \exp(-i2\pi\omega_j t) \\ &= \sum_k D_{z,k} \left(\frac{1}{n} \sum_t y_t \exp(-i2\pi\omega_{j-k} t) \right) \\ &= \sum_{k=0}^{n-1} D_{z,k} D_{y,j-k} \end{aligned}$$

Recall the comparable property of r.v.'s: the MGF of a sum of two ind. r.v.'s is the product of the MGF's and the distribution of the sum is the convolution.

Some Special Cases Several useful special cases are:

Constant. If the series $y_t = k$ for all t , then

$$D_j = \frac{1}{n} \sum_t y_t \exp(-i\omega_j t) = \frac{k}{n} \sum_t \exp(-i\omega_j t)$$

which is zero unless $j = 0$, in which case $D_0 = k$. Hence a constant input generates a single “spike” in the output.

Spike. If the input is zero except for a single non-zero value k at index s , then $D_j = \frac{k}{n} \exp(-i2\pi\omega_j s)$. The amplitude of the DFT is constant, with the phase a linear function of the location of the single spike.

Sinusoid. If $y_t = k \exp(i2\pi\lambda t)$, then we obtain a multiple of the *Dirichlet kernel*,

$$D_j = \frac{1}{n} \sum_t \exp(i2\pi t(\lambda - \omega_j)) = \exp\left(i \frac{(n-1)(\lambda - \omega_j)}{2}\right) D_n(\lambda - \omega_j),$$

where the version of Dirichlet kernel used here is

$$D_n(\lambda) = \frac{\sin(n\lambda/2)}{n \sin(\lambda/2)} \approx \frac{\sin(n\lambda/2)}{n\lambda/2} \text{ if } \lambda \approx 0.$$

If $\lambda = k/n$ is a Fourier frequency, only D_k is non-zero.

Boxcar. If the input is the step function (or “boxcar”),

$$y_t = 1, t = 0, 1, \dots, m - 1, \quad y_t = 0, t = m, m + 1, \dots, n - 1,$$

$$\text{then } |D_j| = \frac{m}{n} D_m(2\pi\omega_j) .$$

Periodic function. Suppose that the input data y_t is composed of K repetitions of the sequence of H points x_t so that $n = KH$. Then the DFT of y_t is

$$\begin{aligned} D_{y,j} &= \frac{1}{n} \sum_t y_t e^{-i2\pi\omega_j t} \\ &= \frac{1}{KH} \sum_{k=0}^{K-1} \sum_{h=0}^{H-1} x_h e^{-i\frac{2\pi j}{KH}(h+kH)} \\ &= \frac{1}{K} \sum_k e^{-i\frac{2\pi j k}{K}} \frac{1}{H} \sum_h x_h e^{-i\frac{2\pi j h}{KH}} \\ &= D_K(2\pi j/K) \frac{1}{H} \sum_h x_h e^{-i\frac{2\pi j h}{KH}} \\ &= \begin{cases} 0 & \text{for } j \neq 0, K, 2K, \dots, (H-1)K. \\ D_{x,\ell} & \text{for } j = \ell K. \end{cases} \end{aligned}$$

The transform of the y 's is zero except at multiples of $2\pi K/n$, which is known as the *fundamental frequency*.