MATH 7339 - Machine Learning and Statistical Learning Theory 2

Section- Discrete Fourier transform and Wavelet transform

1. Spectral estimation
2. Sample autocovariance
3. Discrete Fourier transform
4. Periodogram
5. Wavelet Transform introduction

Spectral Estimation:

So far, we've looked at the spectral density, which gives an alternative view of stationary time series. It is a population quantity.

We'll next consider the sample version: the periodogram.

Given a realization $x_1, \ldots, x_n$ of a time series, how can we estimate the spectral density?

- One approach: replace $\gamma(h)$ in the definition of spectral density

$$f(\omega) = \sum_{-\infty}^{\infty} \gamma(h) e^{-2\pi i \omega h}$$

with sample autocovariance $\hat{\gamma}(h)$.

- Another approach, called the **periodogram**: compute $I(\omega)$, the squared modulus of the discrete Fourier transform (at frequencies $\omega = k/n$).

Remarks:

- These two approaches are identical at the Fourier frequencies $\omega = k/n$.
- The asymptotic expectation of the periodogram $I(\omega)$ is $f(\omega)$. We can derive some asymptotic properties, and hence do hypothesis testing.
- Unfortunately, the asymptotic variance of $I(\omega)$ is constant. It is not a consistent estimator of $f(\omega)$.
- We can reduce the variance by smoothing the periodogram—averaging over adjacent frequencies. If we average over a narrower range as $n \to \infty$, we can obtain a consistent estimator of the spectral density.

Estimating the spectrum: Sample autocovariance

Idea: Use sample autocovariance $\hat{\gamma}(h)$ replace $\gamma(h)$

$$\hat{\gamma}(h) = \frac{1}{n} \sum_{t=1}^{n-|h|} \left(x_{t+|h|} - \bar{x}\right)(x_t - \bar{x})$$

in the definition of spectral density

$$f(\omega) = \sum_{h=-\infty}^{\infty} \gamma(h) e^{-2\pi i \omega h}$$

That is, for $-\frac{1}{2} \leq \omega \leq \frac{1}{2}$, estimate $f(\omega)$ by

$$\hat{f}(\omega) = \sum_{h=-n+1}^{n-1} \hat{\gamma}(h) e^{-2\pi i \omega h}$$

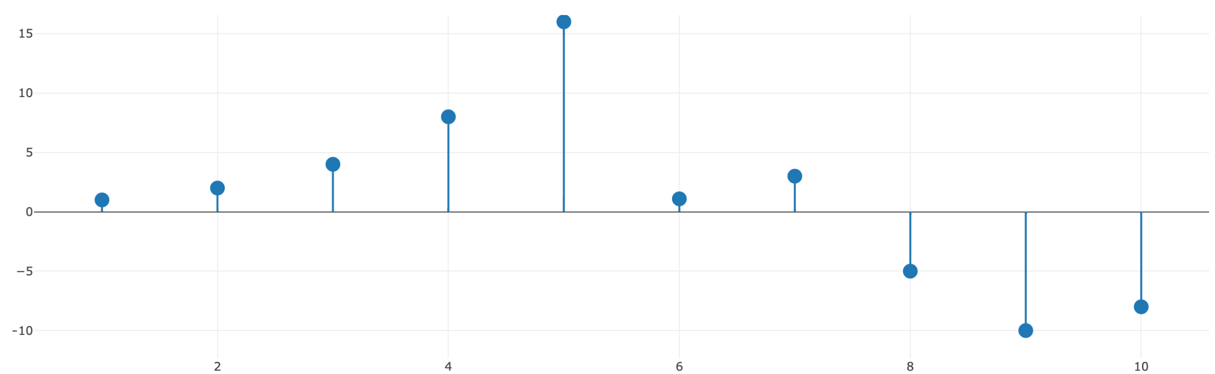
Estimating the spectrum: Periodogram

Another approach to estimating the spectrum is called the periodogram. It was proposed in 1897 by Arthur Schuster (at Owens College, which later became part of the University of Manchester), who used it to investigate periodicity in the occurrence of earthquakes, and in sunspot activity.

Arthur Schuster, "On Lunar and Solar Periodicities of Earthquakes," Proceedings of the Royal Society of London, Vol. 61 (1897), pp. 455–465.

To define the periodogram, we need to introduce the discrete Fourier transform (DFT) of a finite sequence $x_1, \ldots, x_n$.

$$\vec{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \in \mathbb{C}^n$$



❑ Discrete Fourier Transform

In some cases, a spectrum-domain representation is more convenient in describing a process. To transform a time-domain representation to a spectrum-domain representation, we use the Fourier transform.

For a times series sequence $\vec{x} = (x_1, \ldots, x_n)$, define the **discrete Fourier transform (DFT)** as $(X(\omega_0), X(\omega_1), \ldots, X(\omega_{n-1}))$, where

$$X(\omega_k) = \frac{1}{\sqrt{n}} \sum_{t=1}^{n} x_t e^{-2\pi i \omega_k t}$$

and $\omega_k = \frac{k}{n}$, for $k = 0,1, \ldots, n-1$, are called the **Fourier frequencies.** (Think of $\{\omega_k \mid k = 0, \ldots, n-1\}$ as the discrete version of the frequency range $\omega \in [0, 1]$.)

First, let's show that we can view the DFT as a representation of x in a different basis, the **Fourier basis**.

Consider the space $\mathbb{C}^n$ of vectors of n complex numbers, with inner product $\langle \vec{a}, \vec{b} \rangle = \vec{a}^* \vec{b}$, where $\vec{a}^*$ is the complex conjugate transpose of the vector $\vec{a} \in \mathbb{C}^n$.

Suppose that a set $\{\vec{\phi}_j \mid j = 0, 1, \ldots, n-1\}$ of n vectors in $\mathbb{C}^n$ are orthonormal:

$$\langle \vec{\phi}_j, \vec{\phi}_k \rangle = \begin{cases} 1 & if\ j = k \\ 0 & otherwise \end{cases}$$

Then these $\{\vec{\phi}_j\}$ span the vector space $\mathbb{C}^n$, and so for any vector $\vec{x} \in \mathbb{C}^n$, we can write $\vec{x}$ in terms of this new orthonormal basis,

$$\vec{x} = \sum_{j=0}^{n-1} \langle \vec{\phi}_j, \vec{x} \rangle \vec{\phi}_j$$

An alternative way to represent the DFT is by separately considering the real and imaginary parts,

$$X(\omega_j) = \langle E_j, \vec{x} \rangle = \frac{1}{\sqrt{n}} \sum_{t=1}^{n} e^{-2\pi i \omega_j t} \, x_t$$

$$= \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \cos(2\pi t \omega_j) \, x_t - i \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \sin(2\pi t \omega_j) \, x_t$$

$$= X_c(\omega_j) - i X_s(\omega_j)$$

where this defines the sine and cosine transforms, $X_s \; and \; X_c, of \; \vec{x}$.

Fast Fourier Transform

Fast Fourier Transform (FFT) is used to compute the Discrete Fourier Transform (DFT) to make the computations more efficient.

❑ Periodogram

Recall we want to estimate $f(\omega) = \sum_{-\infty}^{\infty} \gamma(h) e^{-2\pi i \omega h}$

The **periodogram** is defined as

$$I(\omega_j) = |X(\omega_j)|^2 = \frac{1}{n}\left|\sum_{t=1}^{n} e^{-2\pi i \omega_j t}\, x_t\right|^2 = X_c^2(\omega_j) + X_s^2(\omega_j)$$

$$X_c^2(\omega_j) = \frac{1}{\sqrt{n}}\sum_{t=1}^{n} \cos(2\pi t \omega_j)\, x_t$$

$$X_s^2(\omega_j) = \frac{1}{\sqrt{n}}\sum_{t=1}^{n} \sin(2\pi t \omega_j)\, x_t$$

The periodogram $I(\omega_j) = |X(\omega_j)|^2$ for one of the Fourier frequencies $\omega_j = \frac{j}{n}$ (for $j = 0, 1, \ldots, n-1$).
The orthonormality of the E_j implies that we can write

$$\vec{x}^*\vec{x} = \left(\sum_{j=0}^{n-1} X(\omega_j)E_j\right)^* \left(\sum_{j=0}^{n-1} X(\omega_j)E_j\right)$$

$$= \sum_{j=0}^{n-1} |X(\omega_j)|^2 = \sum_{j=0}^{n-1} I(\omega_j)$$

For $\overline{\vec{x}} = 0$, we can write this as

$$\hat{\sigma}_x{}^2 = \frac{1}{n}\sum_{t=1}^{n} x_t^2 = \frac{1}{n}\sum_{j=0}^{n-1} I(\omega_j)$$

This is the discrete analog of the identity

$$\hat{\sigma}_x{}^2 = \gamma_x(0) = \int_{-\frac{1}{2}}^{\frac{1}{2}} f_x(\omega) d\omega$$

Think of $I(\omega_j)$ as the discrete version of $f(\omega)$ at the frequency $\omega_j = j/n$, and think of $\left(\frac{1}{n}\right)\sum_{\omega_j}$. as the discrete version of $\int \quad d\omega$

Estimating the spectrum: Periodogram

Why is the periodogram at a Fourier frequency (that is, $\omega = \omega_j$) the same as computing $f(\omega)$ from the sample autocovariance?

Almost the same. They are not the same at $\omega_0 = 0$ when $\vec{\bar{x}} \neq 0$.

For $j = 0$, we have

$$I(0) = n\bar{x}^2$$

But, for $j \in \{1, \ldots, n-1\}$,

But if either $\vec{\bar{x}} \neq 0$, or we consider a Fourier frequency ω_j with $j \in \{1, \ldots, n-1\}$,

$$I(\omega_j) = |X(\omega_j)|^2 = \frac{1}{n}\left|\sum_{t=1}^{n} e^{-2\pi i\omega_j t}\, x_t\right|^2 = \frac{1}{n}\left|\sum_{t=1}^{n} e^{-2\pi i\omega_j t}\,(x_t - \bar{x})\right|^2$$

$$= \frac{1}{n}\left(\sum_{t=1}^{n} e^{-2\pi i\omega_j t}\,(x_t - \bar{x})\right)\left(\sum_{t=1}^{n} e^{-2\pi i\omega_j t}\,(x_t - \bar{x})\right)$$

$$= \frac{1}{n}\left(\sum_{s,t=1}^{n} e^{-2\pi i\omega_j (s-t)}\,(x_s - \bar{x})(x_t - \bar{x})\right) = \sum_{h=-n-1}^{n-1} \hat{\gamma}(h) e^{-2\pi i\omega_j h}$$

where the fact that $\omega_j \neq 0$ implies $\sum_{t=1}^{n} e^{-2\pi i\omega_j t} = 0$ (we showed this when we were verifying the orthonormality of the Fourier basis) has allowed us to subtract the sample mean in that case.

Interpreting the Periodogram

We can think of the inverse Fourier transform as a regression of x_t on sines and cosines with the coefficients equal to $\frac{2}{\sqrt{h}}$ times the sine part and the cosine part of the Fourier transforms respectively.

Therefore, $X_c(\omega_j)$ and $X_s(\omega_j)$ measure the contribution the frequency ω_j has in explaining the variation in the time series. The bigger $X_c(\omega_j)$ and $X_s(\omega_j)$, the greater the contribution from the frequency ω_j.

One can show that

$$\sum_{t=1}^{n}(x_t - \bar{x})^2 = 2\sum_{j=1}^{m}\left[X_c^2(\omega_j) + X_s^2(\omega_j)\right] = 2\sum_{j=1}^{m} I(\omega_j)$$

The sum of squares can be decomposed into 2 times the sum of the periodograms over frequencies ω_j for $1 \leq j \leq m$. In other words, the variation in the series x_t is distributed over frequencies ω_j, where the amount of variation explained by frequency ω_j is $2I(\omega_j)$.

Thus, we can interpret the periodogram as the amount of variation at a certain frequency. This is how we also interpret the spectral density. The periodogram is the sample version of the spectral density, which is a population quantity.

Asymptotic properties of the periodogram

We want to understand the asymptotic behavior of the periodogram $I(\omega)$ at a particular frequency ω, as n increases. We'll see that its expectation converges to $f(\omega)$.

Let $\omega_{j:n}$ denote a frequency of the form $\frac{j_n}{n}$, where $\{j_n\}$ is a sequence of integers so that $j_n \to \infty$ and $n \to \infty$. It turns out that

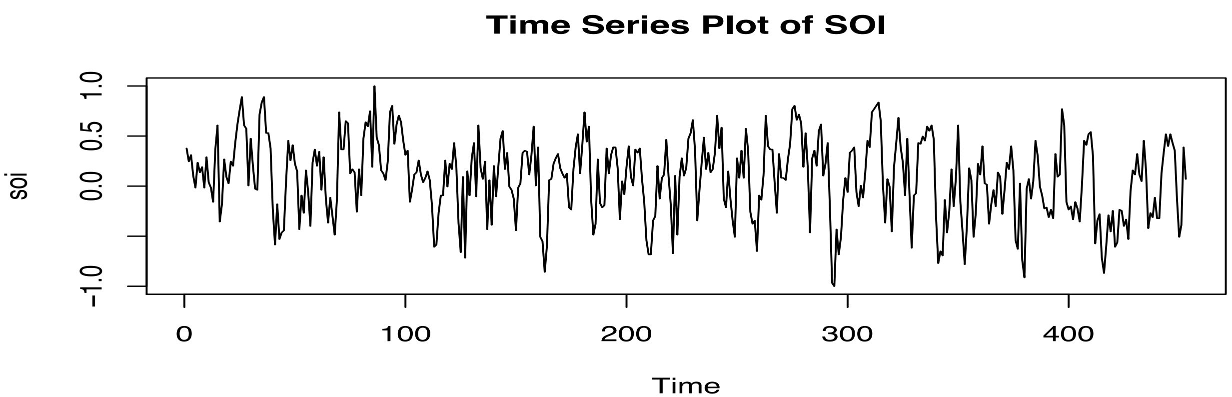
$$E\left[I\left(\omega_{j:n}\right)\right] \to f(\omega) = \sum_{-\infty}^{\infty} \gamma(h) e^{-2\pi i \omega h}$$

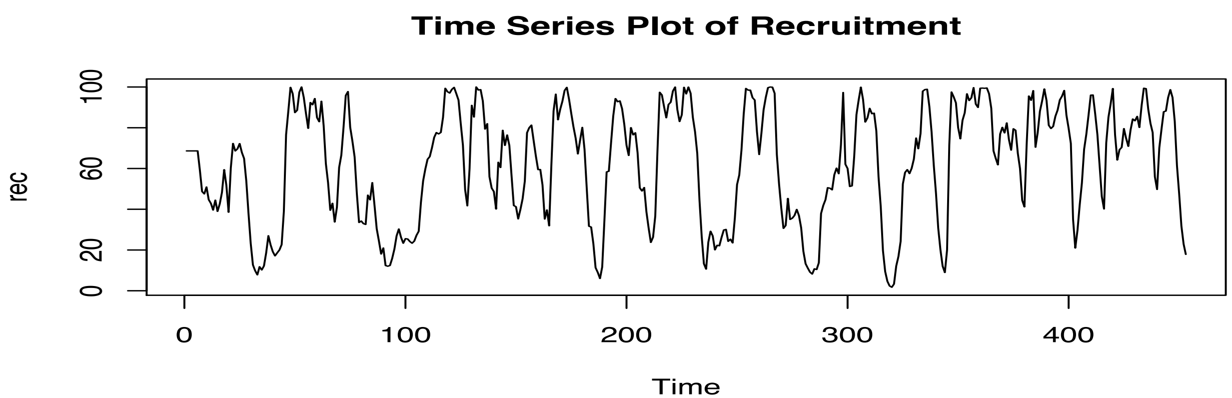
The spectral density is the long run average of the periodogram.

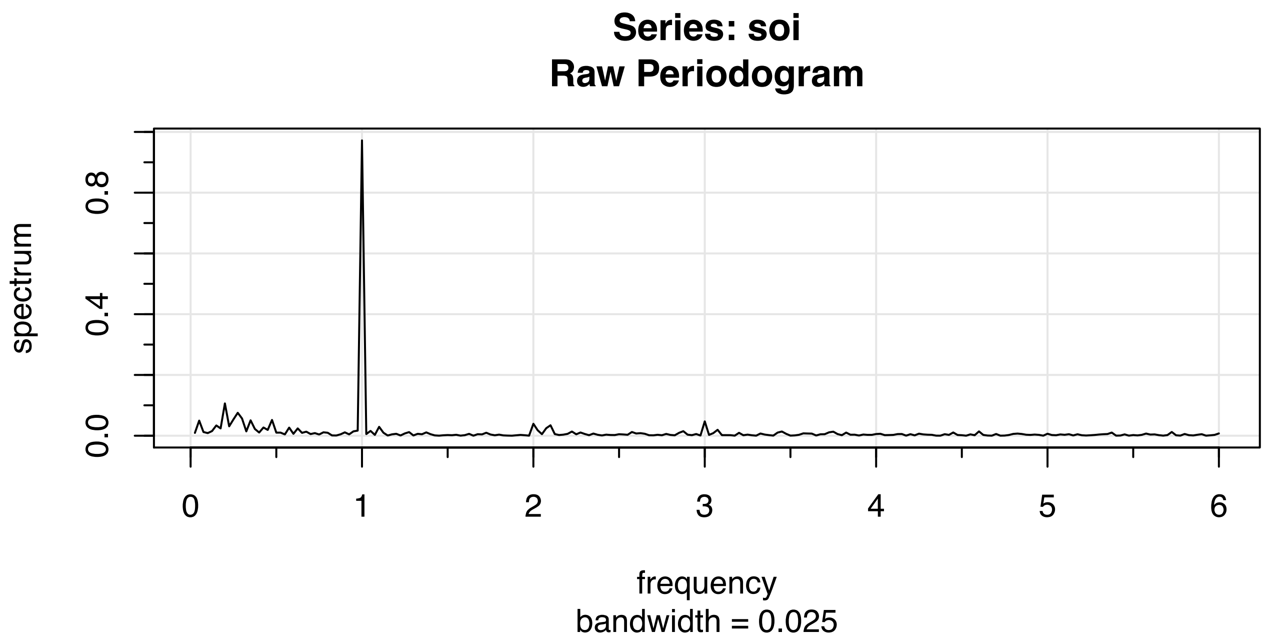
Further more, $\frac{2I\left(\omega_{j:n}\right)}{f(\omega_j)}$ follows *chi-squared distribution* and confidence interval can be calculated.

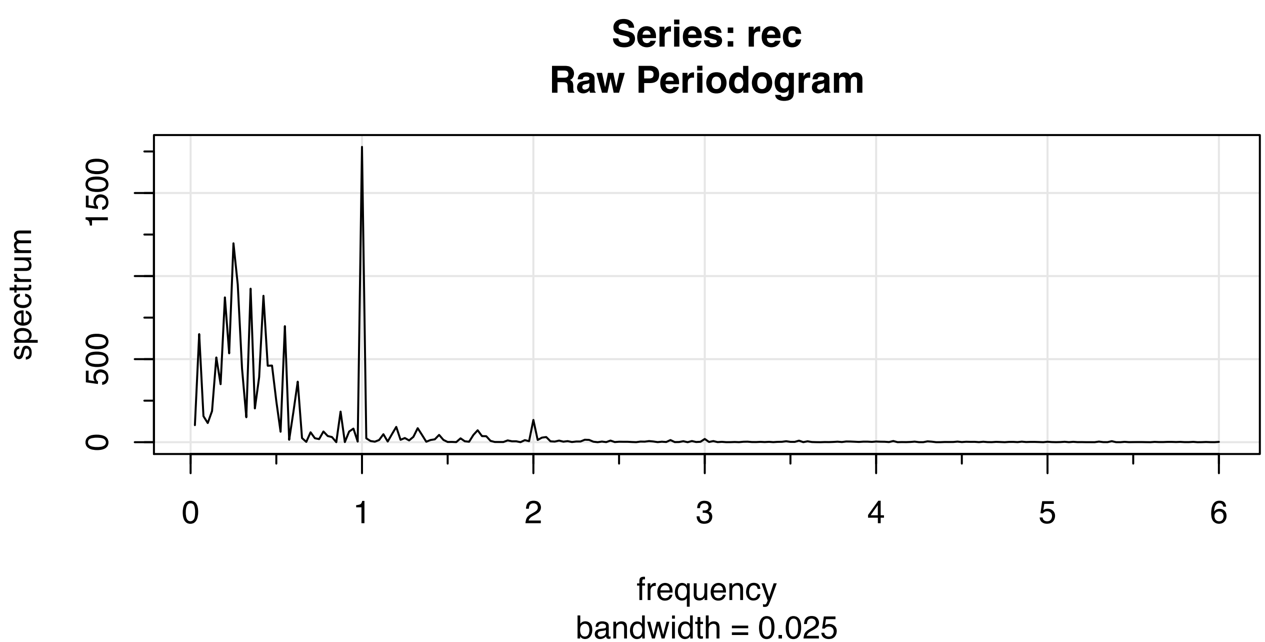
Example:

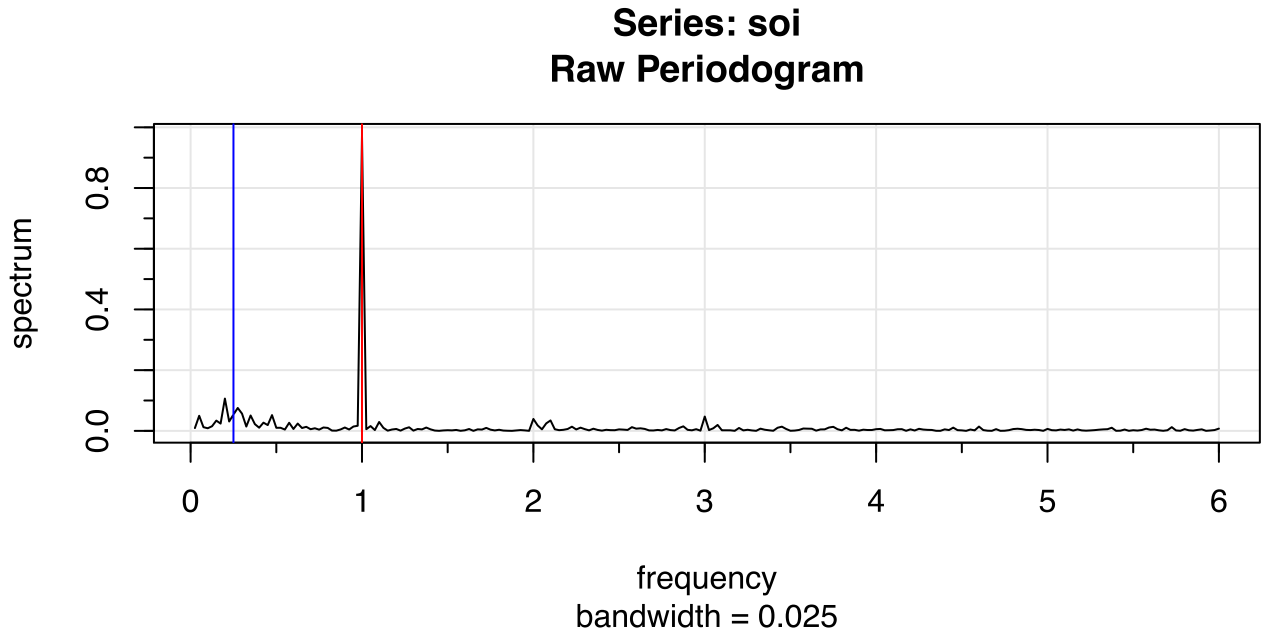
In this example, we will look at the Southern Oscillation Index (SOI) and recruitment datasets, which contain monthly data on the changes in air pressure and estimated number of new fish in the central Pacific Ocean from 1950 to 1987. The central Pacific Ocean warms approximately every three to seven years due to El Nino.

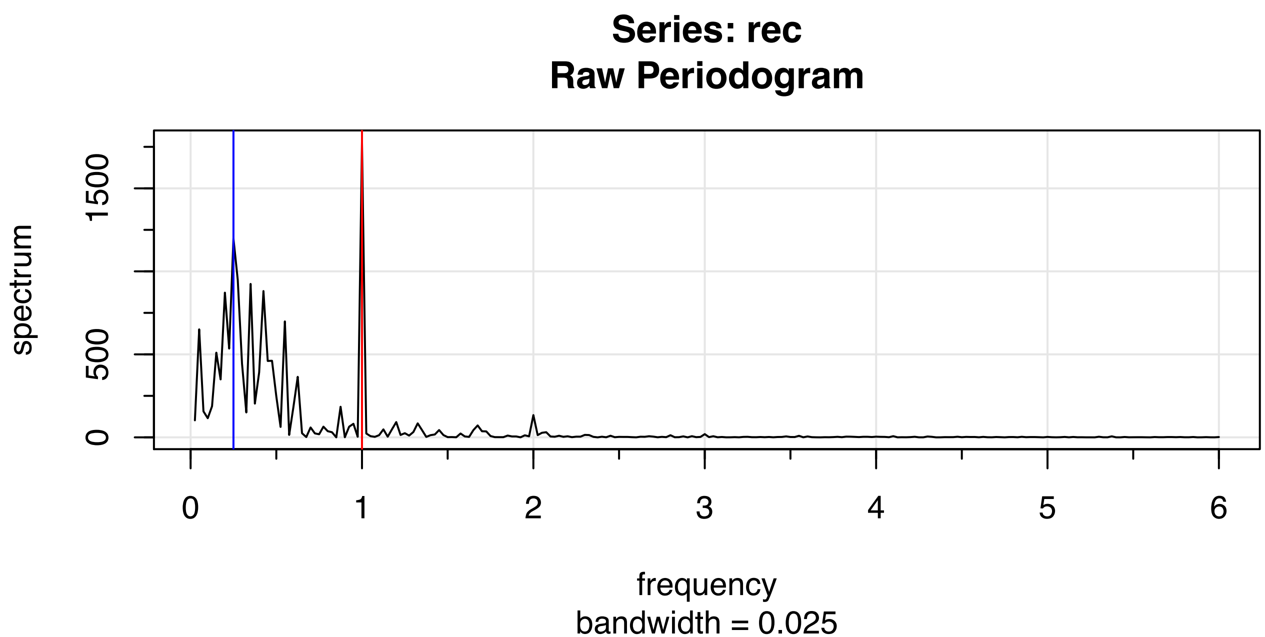





Series: soi
Raw Periodogram
spectrum
0.0
0.4
0.8
0
1
2
3
4
5
6
frequency
bandwidth = 0.025


Series: rec
Raw Periodogram
spectrum
0
500
1500
0
1
2
3
4
5
6
frequency
bandwidth = 0.025


Series: soi
Raw Periodogram
spectrum
0.0
0.4
0.8
0
1
2
3
4
5
6
frequency
bandwidth = 0.025


Series: rec
Raw Periodogram
spectrum
0
500
1500
0
1
2
3
4
5
6
frequency
bandwidth = 0.025

From the periodograms:

- obvious peak at $\omega = 1/12$ for yearly cycle.
- some peaks at around $\omega = 1/48$ for El Nino cycle. The wide band of activity suggests that this cycle is **not very regular**.

From the SOI data, the value of the periodogram at $\omega = \frac{1}{12}$ is $I(\frac{1}{12}) = 0.9722$. Since $\chi_2^2(0.025) = 0.0506$ and $\chi_2^2(0.975) = 7.3778$, an approximate 95% confidence interval for the spectrum $f(\frac{1}{12})$ is

$$\left(\frac{2(0.9722)}{7.3778}, \frac{2(0.9722)}{0.0506}\right) = (0.2636, 38.4011).$$

At $\omega = \frac{1}{48}$, $I(\frac{1}{48}) = 0.0537$, therefore an approximate 95% confidence interval for the spectrum $f(\frac{1}{48})$ is

$$\left(\frac{2(0.0537)}{7.3778}, \frac{2(0.0537)}{0.0506}\right) = (0.0146, 2.1222).$$

- "Time Series Analysis and Its Applications", 4th ed. 2017, by Shumway and Stoffer.

  Sections 4.4

- I. Daubechies (Ten Lectures on Wavelets; Orthonormal Bases of Compactly Supported Wavelets) https://epubs.siam.org/doi/book/10.1137/1.9781611970104

- Mark Kon, lecture notes: http://math.bu.edu/people/mkon/Wavelets.pdf