Project 4
Spectral Analysis and Signal Processing

Any periodic function $X(t)$ with period T may be written as a sum of sinusoidal waves:

$$\hat{X}_k = \frac{1}{T}\int_0^T X(t)\exp\bigl(-2\pi\mathrm{i}kt/T\bigr)\,\mathrm{d}t, \tag{1}$$

$$X(t) = \sum_{k=-\infty}^{\infty} \hat{X}_k \exp\bigl(2\pi\mathrm{i}kt/T\bigr). \tag{2}$$

The coefficients $\hat{X}_k$ are complex numbers and they define the amplitude and phase of a wave with frequency $f(k) = k/T$. Equations (1) and (2) give the forward and backward Fourier transforms on the interval $[0, T)$, respectively. When the signal $X(t)$ is sampled using N regularly spaced points ($X_n = X(n\Delta t)$, $\Delta t = T/N$) we have to use the discrete version of the Fourier transform. The most common definitions of the forward and backward *discrete Fourier transform* (DFT) are the following:

$$\hat{X}_k = \mathcal{F}\bigl\{X_n\bigr\}_k = \sum_{n=0}^{N-1} X_n \exp\bigl(-2\pi\mathrm{i}kn/N\bigr), \qquad k = 0, 1, 2, \ldots, N-1 \tag{3}$$

$$X_n = \mathcal{F}^{-1}\bigl\{\hat{X}_k\bigr\}_n = \frac{1}{N}\sum_{k=0}^{N-1} \hat{X}_k \exp\bigl(2\pi\mathrm{i}kn/N\bigr), \qquad n = 0, 1, 2, \ldots, N-1. \tag{4}$$

The complex coefficients $\hat{X}_k$ can be now determined only over a finite frequency range from $f_{\min} = -(N/2-1)/T$ to $f_{\max} = (N/2)/T$. Any modes with higher frequencies are mapped back to the range $f \in [f_{\min}, f_{\max}]$. For consistency with the Fourier series definition (2) the index k should run from $-N/2+1$ to $N/2$. However, because the Fourier transformation implicitly assumes all signals (in time and frequency domain) are periodic with period N, we can let k run from 0 to $N-1$ and map k to the correct frequency $f(k)$ according to the rule

$$f(k) = \frac{1}{T}\cdot\begin{cases} k, & 0 \le k \le N/2, \\ k-N, & N/2 < k < N \end{cases}. \tag{5}$$

This means that the lower half of the array $\hat{X}_k$ stores modes with positive frequencies, whereas the upper half stores modes with negative frequencies. For real input signals X_n the Fourier coefficients with negative frequencies are just complex conjugates of the coefficients with positive frequencies:

$$\hat{X}_{N-k} = \hat{X}_k^*. \tag{6}$$

The data stored in the upper half of the complex array $\hat{X}_k$ is therefore redundant when the input signal is real.

Quite often we are interested in the distribution of wave amplitudes of the signal across the frequency range. A common measure for the distribution of wave amplitudes is the power spectral density (or power spectrum):

$$P_k = \frac{T^2}{N^2}|\hat{X}_k|^2. \tag{7}$$

The spectrum P_k can reveal interesting features of a signal which are sometimes not obvious when looking in the time domain. The peaks in the spectrum represent the dominant modes which are the most essential for describing the signal. Flat spectra with no apparent peaks correspond to random (white) noise signals.

Another example where the Fourier transform comes in handy is the correlation function for signals $X(t)$ and $Y(t)$:

$$C_{X,Y}(\tau) = \lim_{T\to\infty} \frac{1}{T} \int_0^T X(t+\tau)Y(t)^* \,\mathrm{d}t. \tag{8}$$

The correlation function tells us how well do the signals match when shifted by a time lag τ. The correlation function when computed with $X = Y$ is known as the autocorrelation function.

For discrete signals of length N we have to take into account that the number of data points in the overlapping region changes when we vary the discrete time lag l. A first step in the definition of the discrete correlation function requires that we extend the X and Y arrays with N zeros:

$$\tilde{X}_n = (X_0, X_1, \ldots, X_{N-1}, 0, 0, \ldots, 0), \qquad \tilde{Y}_n = (Y_0, Y_1, \ldots, Y_{N-1}, 0, 0, \ldots, 0). \tag{9}$$

The size of the extended arrays is now $2N$. Using the new arrays, we may write the discrete correlation as

$$C_{X,Y}(l) = \frac{1}{N-|l|} \sum_{n=0}^{2N-1} \tilde{X}_{n+l}\tilde{Y}_n^*. \tag{10}$$

The maximal range for the lag l is from $-N+1$ to $N-1$ though the quality of the estimates degrates for large lags because there are fewer points in the region where the signals overlap. The above definition assumes a circular correlation (i.e. periodic data). That is, whenever the array index $n + l$ for $\tilde{X}$ exceeds the array bounds it is "wrapped around" to become in-bounds using the rule $\tilde{X}_{n\pm 2N} = \tilde{X}_n$.

Using the definitions (3) and (4) for the forward and backward DFT it can be shown that the correlation function can be conveniently written as

$$C_{X,Y}\big(l(n)\big) = \frac{1}{N-|l(n)|} \mathcal{F}^{-1}\left\{\mathcal{F}\{\tilde{X}_n\}_k \cdot \mathcal{F}\{\tilde{Y}_n\}_k^*\right\}_n. \tag{11}$$

$C_{X,Y}$ is an array of size 2N and the lag l can be written as a function of the array index $n = 0, 1, 2, \ldots, 2N-1$ as

$$l(n) = \begin{cases} n, & 0 \le n \le N, \\ n-2N, & N < n < 2N \end{cases}. \tag{12}$$

The lower half of the inverse DFT of $\hat{\hat{X}}_k \hat{\hat{Y}}_k^*$ corresponds to correlations for positive lags, whereas the upper half of the array corresponds to negative time lags. A division by zero occurs for the element with index N in $C_{X,Y}$ because the normalization factor in equation (11) is zero. This particular element can be safely discarded in any analysis because it corresponds to a situation where the size of the overlapping region for the X and Y arrays is zero. Instead of computing the correlations from the raw data X_n, Y_n it is common practice to normalize the signals using

$$X_n \to (X_n - \overline{X})/\sigma_X, \qquad Y_n \to (Y_n - \overline{Y})/\sigma_Y, \tag{13}$$

where $\overline{X}$, $\overline{Y}$ are the time averages of X_n, Y_n, and σ_X, σ_Y are the standard deviations of X_n, Y_n, respectively. When the signals are normalized as above the correlation function will be bounded between −1 and 1. Values close to 1 indicate very high correalations, wheareas values close to -1 indicate high anticorrelations.

1 Assignments

For this project, you will perform Fourier analysis on two different data sets: monthly mean sunspot numbers and cosmic ray counts over a period of $T = 50$ years from beginning of 1966 to the end of 2015. The sunspot numbers are provided in the file `sunspots_monthly.txt` and information about the data is given in `sunspots_info.txt`. The cosmic ray counts are provided in the file `rays_monthly.txt` and information about the data is given in `rays_info.txt`. Both files contain $N = 600$ data points. The provided data contains several columns. For the analysis you should use:

- 2nd column of `sunspots_monthly.txt` (monthly mean total sunspot number),
- 3rd column of `rays_monthy.txt` (corrected cosmic ray count rates).

Sunspots are dark spots on the surface of the Sun with very intense magnetic fields. They typically last for several days, although very large ones may last for weeks. Sunspot measurements have been performed on a regular basis for over 150 years. The number of sunspots is used as a measure of solar (magnetic) activity with a cycle of approximately 11 years.

Cosmic rays are high energy charged particles mainly originating from outside the solar system. They can effect cloud formation through ionization of atmospheric molecules, and they constitute a fraction of the total annual radiation exposure on Earth. During periods of increased solar activity the interplanetary magnetic field strengthens making it more difficult for the cosmic rays to reach the Earth. The cosmic ray flux is therefore anticorrelated with solar activity.

The assignments are the following:

1. Compute power spectra of the signals `sunspots_monthly.txt` and `rays_monthly.txt` and estimate the time period (inverse frequency) of the dominant nonzero frequency mode in the spectrum. The dominant mode should lie in the low-frequency range.

2. Compute the normalized autocorrelation functions of the signals using the formula (11). Do not forget to normalize the data using (13) and to pad the normalized arrays with zeros before applying the Fourier transforms. Estimate the length of the solar cycle and cosmic ray cycle from the time periods of their autocorrelation functions.

3. Compute the normalized cross-correlation function between the sunspot numbers and cosmic ray flux. What is the value of the correlation coefficient $C_{X,Y}(l = 0)$? Where is the (global) minimum of the cross-correlation function? Is it exactly at $l = 0$ or is it perhaps slightly shifted?

4. The signals `sunspots_monthly.txt` and `rays_monthly.txt` are quite noisy. Smooth the signals using a sharp low-pass filter with a cutoff frequency f_0. To implement the filter, transform the original signal to frequency space, set all modes with frequency $|f| \geq f_0$ to zero, and transform the modified signal back to the time domain. Compare the signals filtered with different cutoff frequencies on the same graph.

Below we give for reference the online sources where the data for the assignments was obtained:

- Sunspot numbers: `http://www.sidc.be/silso/datafiles`, WDC-SILSO, Royal Observatory of Belgium, Brussels.
- Cosmic ray counts: `http://cosmicrays.oulu.fi`, Oulu Cosmic Ray Station, Sodankyla Geophysical Observatory, Finland.

A Supplemental Information

A naive computation of the discrete Fourier transforms (3) and (4) requires order of N^2 operations. Thanks to the fast Fourier transform (FFT) algorithm the sums can be recursively broken down into smaller parts using a "divide and conquer" strategy which requires only a total of $N \log N$ operations. The algorithm works best if N is a power of 2 though modern implementations usually work well for any size N that can be written as a product of small prime numbers. The FFT is one of the most widely known and celebrated numerical algorithms of the 20th century. Implementations of FFT are available in every major numerical package or computational environment:

- Python: functions `rfft()` and `irfft()` or `fft()` and `ifft()` from the NumPy library,