DATA, MODELS & UNCERTAINTY IN THE NATURAL SCIENCES

Problem Set 6: Due on Dean's Date

We discussed how the **power spectral density** relates to the expected value of the squared magnitude of the Fourier coefficients of a signal $f(x)$ when these are treated as random variables:

$$S(k) = E\{|\tilde{f}(k)|^2\}. \tag{1}$$

We discussed how, in real life, we will always work with a *finite*, *sampled*, realization of the truth, and how this truth is (therefore) always *windowed* to fit within the observation domain. If "no" window is applied, we really *are* applying a window, too — the *boxcar* window, a *mask* that weights (with values 1) the samples inside of our observation domain, and weights (with values 0) those samples that are outside of what we observe.

Power spectral estimation thus consists of finding a data window that makes the sum of the squared Fourier coefficients of what we observe, weighted by the window, a **power spectral estimate** that is a "good" approximation to the true power spectral density. As usual "good" involves notions of *bias* and *variance*, which we want as small as possible.

One such estimate might be given by

$$\hat{S}(k) = |\tilde{g}(k)|^2, \tag{2}$$

with $\tilde{g}(k)$ the discrete Fourier transform of the sampled and windowed signal $g(x) = a(x)f(x)$, for some window function $a(x)$. Matlab has a function `periodogram` that allows us to try a variety of windows and see the results right away. If the window is a boxcar, we call the result the **periodogram**, if the window is another function, we call the result the **modified periodogram**. The periodogram has some awful statistical properties that the modified periodogram with better windows alleviates — not that those will be immediately obvious from the examples below.

But remember that the outcome of the *windowing-Fourier-squaring* operation, in expectation, is the *convolution* of the true power spectral density with the periodogram of the window itself. If $\hat{S}(k)$ is a modified periodogram as in eq. (2), and thus an estimate of the true power spectral density $S(k)$ in eq. (1), and you used a windowing function $a(x)$ to pre-multiply the data $f(x)$, then

$$E\{\hat{S}(k)\} = \int_{-\frac{1}{2\Delta x}}^{\frac{1}{2\Delta x}} S(k')\,|\tilde{a}(k-k')|^2\,dk'. \tag{3}$$

In what follows the *periodogram of the data window*, $|\tilde{a}(k)|^2$ will thus be as important to understand the results as the *modified periodogram of the windowed function*, $|\tilde{g}(k)|^2$, and you will want to look at both.

If you see any peaks and clear structure in the power spectral density estimates that you will compute using the built-in function `periodogram`, you should be comfortable to try your hand at going a little more bare-bones, and use Matlab's `fft` to obtain similar results. Doing `abs(fft(gofx)).^2` should get you close! But remember that you need to make your own frequency axis to do the plotting (as we discussed in class, $k_n = n/(N\Delta x)$, and remember that you should focus on only half of the set of frequencies for real signals. Look into using `fftshift` to get plots that look like those that `periodogram` generates automatically.

1. If a bandlimited function is not sampled faster than the *Nyquist rate*, it will be subject to *aliasing*. The Nyquist rate is twice the largest frequency contained in the signal, $1/(2\Delta x)$. Illustrate the breakdown due to aliasing by generating a synthetic signal (invent something interesting!) and sampling it using various values of Δx: above, at, and below the Nyquist rate. Make plots of the time-domain signal and of the *power spectral density*, for which you can rely on Matlab's `periodogram` function and its default parameters (i.e. a boxcar window).

   Note that this question basically amounts to you running the Matlab example from inside `periodogram` *for a made-up signal with a series of made-up sampling rates, and interpreting the results.*

2. Study the power spectral density of the mysterious signal of *yearly* observations contained in the data vector `YEARLY.PLT`. Note that the first column in this data file is the calendar year and the second column the measurement.

3. Do the same for the mystery signal of *decadal* observations (of the same process as above) contained in the vector `DECADAL.PLT`. Note that the first column in this data file is the number of years *before calendar year 1950*, the second the measurement, and the third a measure of uncertainty.

4. Explore the effect of *windowing* on the shape of the power spectral density estimated from the data. Experiment with a variety of windowing functions using `periodogram`. Make plots of the input signal, the window functions and their periodograms, and the modified periodograms providing the spectral estimate. Popular data windows are `BARTLETT`, `BLACKMAN`, `CHEBWIN`, `HAMMING`, `HANN`, `KAISER`. See Harris [1978] for a comprehensive survey; I've put this paper on the e-Reserves.

5. What can you say about either/both data sets? Which periods, if any, are prominent? Google and see if you can find out what data sets these might actually be.

6. You've only used `periodogram` up to this point, but now make your own plot, "from scratch", of one of your favorite power spectral density estimates above, using `fft` and your own frequency axis, as described above. Or use my `fftaxis` which you can find on my "software" page. Google my name with "software". You will find other routines of interest there.

7. For your favorite power spectral density estimate above, including of the synthetic signals, forget the Fourier transform altogether. Target a set of *specific* frequencies and fit sines and cosines of those target frequencies in the signal using the tools of inverse theory that we discussed and you have used extensively in class. To capture a frequency, you will get one sine and one cosine coefficient, so quote the sums of the squares of the expansion coefficients for the particular frequencies under consideration.

   Note that this question basically amounts to redoing a golfball-style fit to the data but now with sines and cosines of a particular *frequency with unknown expansion coefficients.*

8. Discuss the quality of those sine/cosine fits in terms of the ratio of the variance of the residuals (the data minus the fits) to the variance of the original signal. You want this ratio to be small, of course. See if you can refine the (modified) periodogram by finding, more specifically than on the grid of Fourier frequencies that `periodogram` or your implementation of `fft` give you, which particular frequencies give you the best **variance reduction** in this sense.

Note that this question is to make you realize that Fourier series coefficients "solve" the inverse problem of fitting a time series by a Fourier series in a least-squares sense, and that if you have a good idea of which frequency range to look for, you can build up an appreciation for which frequencies contribute most to the signal under consideration. See Gubbins [2004], Chapter 10.

References

Gubbins, D., 2004. *Time Series Analysis and Inverse Theory for Geophysicists*, Cambridge Univ. Press, Cambridge, UK.

Harris, F. J., 1978. On the use of windows for harmonic analysis with the discrete Fourier transform, *Proc. IEEE*, **66**(1), 51–83.