Data Analysis and Compression: How Low Can We Go?

Some very important datasets involve the relationships between observations taken at different times or at different spatial locations. In this homework, we experiment with some transform tools (discrete cosine transform, discrete sine transform, and wavelet transforms) to extract patterns that might be hidden in the raw data.

In our first challenge, we use transforms for data analysis, looking for patterns in stock prices.

CHALLENGE 1 *This problem uses the data in* `dsig.m`*, posted on the website. The data vector gives the stock price of John Deere & Co. each month for 50 years.*

First, **detrend** *the data: Given the price data (call it $\hat{p}(t)$ for $t = 1, \ldots, N$), determine the straight line $\ell(t)$ that passes through the points $\hat{p}(1)$ and $\hat{p}(N)$. We will use the original data and the detrended data $p(t) = \hat{p}(t) - \ell(t)$.*

- *1a. Apply the discrete sine transform (DST) to the detrended data and plot the result.*

  Make a table of the coefficients and the periods of the 5 largest (in magnitude) coefficients.

- *1b. Apply the discrete cosine transform (DCT) to the original data and plot the result.*

  Make a table of the coefficients and the periods of the 5 largest (in magnitude) coefficients.

- *1c. Apply the* `db3` *wavelet transform to the original data, and to the detrended data. Repeat with another wavelet of your own choosing.*

  Plot the results, as in the demonstration program `waveletsigdemo.m`*.*

- *1d. Discuss any patterns in the data that become evident using the different transforms.*

In addition to data analysis, data compression is quite important. Let's see how well we can represent this data with a smaller number of transform coefficients.

CHALLENGE 2 *Now consider compressing the approximations to k coefficients:*

- *Save the first k coefficients for the DST and DCT.*
- *Save the largest k coefficients (in magnitude) for the wavelet approximations.*

Plot the resulting 4 signals when 420 coefficients are saved. (Don't do the wavelet transforms on the detrended data.)

Make a plot with 4 curves, for DST, DCT, and the two wavelets. The horizontal axis is the number of coefficients k, and the vertical axis is the relative error in the approximation: (2-norm of signal minus approximation) divided by 2-norm of signal.

Repeat, making the plots for the signal defined in `zsig.m`*, which consists of data extracted from two pixel rows of the* `elk.jpg` *photo.*

Discuss the results. Which methods worked best, and what characteristics of the data allow greater compression?

The practical use of wavelets depends on being able to easily compute the coefficients. One formula from the supplementary notes is particularly useful, and in the following challenge, you establish its correctness.

CHALLENGE 3 *Recall, from the supplementary notes, that*

$$a_{m-1}[k] = 2^{m-1} \int_{-\infty}^{\infty} f(x)\phi(2^{m-1}x - k)\, dx.$$

Using this definition, prove for the Haar wavelet that

$$a_{m-1}[k] \quad = \quad \frac{1}{2}a_m[2k] + \frac{1}{2}a_m[2k+1]).$$
