330:525 – Optimum Signal Processing
Computer Experiment 7 — Due April 7, 2011

1. Please read the attached PDF paper on applying PCA to the 1984 Olympic track records. The attached files `olymp1.dat` and `olymp2.dat` contain the women's and men's track data in a form that can be read by the function `loadfile.m`.

   Read the data files into the 55×7 and 55×8 data matrices Y_1 and Y_2 and remove their column means using the function `zmean`.

   (a) For the women's data matrix Y_1, plot the scatterplot of the 100-meter and 200-meter columns. Notice that they lie mostly along a one-dimensional subspace. Perform a PCA on these two columns and determine the percentage variances carried by the two principal components. On the scatterplot, place the two straight lines representing the two principal components, as was done in Fig.16.1 of the SVD notes.

   (b) Repeat part (a) for the following track pairs: (100m, 800m), (100m, 3000m), (100m, Marathon), (3000m, Marathon). Comment on the observed clustering of the data points along one-dimensional directions. Does it make intuitive sense?

   (c) Next, consider the full data matrix Y_1. Working with the SVD of Y_1, perform a PCA on it and determine the percentage variances of the principal components. Determine the PCA coefficients of the first two principal components and compare them with those given in the attached paper. Based on the first component determine the countries that correspond to the top 15 scores. (*Hint:* use the MATLAB function `sort`.)

   (d) Repeat part (c) using the men's data matrix Y_2.

   (e) Next, combine the women's and men's data matrices into a single matrix by concatenating their columns, that is, $Y = [Y_1, Y_2]$. Carry out a PCA on Y and determine the percentage variances. Determine the PCA coefficients of the first principal component. Then, determine the top 15 countries.

2. *Southern Oscillation Index.* It has been observed that in the southern Pacific there occurs regularly an upwelling of large masses of lower-level colder water which has important implications for marine life and coastal weather. This effect, which is variable on a monthly and yearly basis, has been termed *El Niño*. It has been held responsible for many strange global weather effects in the past decades.

   One measure of the variability of this effect is the so-called *southern oscillation index* (SOI) which is the atmospheric pressure difference at sea level between two standard locations in the Pacific, namely, Tahiti and Darwin, Australia. This data exhibits a strong 40–50 month cycle and a weaker 10–12 month cycle.

   The SOI data, spanning the years 1920–1992, are in the included file `soi2.dat`. The monthly data must be concatenated, resulting into a long one-dimensional time series $y(n)$ and the mean must be removed. (The concatenation can be done by the following MATLAB commands: assuming that `Y` is the data matrix whose rows are the monthly data for each year, then redefine `Y=Y';` and set `y = Y(:);`)

   (a) It is desired to fit an AR model to this data, plot the AR spectrum, and identify the spectral peaks. Starting with model order $M = 15$, calculate the ordinary Burg estimate of the prediction-error filter , say $\mathbf{a}_b$.

   (b) Form the order-M autocorrelation and forward/backward data matrices Y, perform an SVD, and plot the principal component variances as percentages of the total variance. You will observe that beyond the 5th principal component, the variances flatten out, indicating that the dimension of the signal subspace can be taken to be of the order of $r = 5$–9.

   Start with the choice $r = 8$ and perform $K = 1$ and $K = 3$ rank-r enhancement operations on the data matrix, as expressed symbolically in MATLAB language:

```
Y = datamat(y,M,type)    % type = 0 or 2
Ye = Y;                  % initialize SVD iterations
for i=1:K,
   Ye = sigsub(Ye,r)     % rank-r signal subspace
   Ye = toepl(Ye,type)   % type = 0 or 2
end
ye = datasig(Ye,type)    % extract enhanced signal from Ye
```

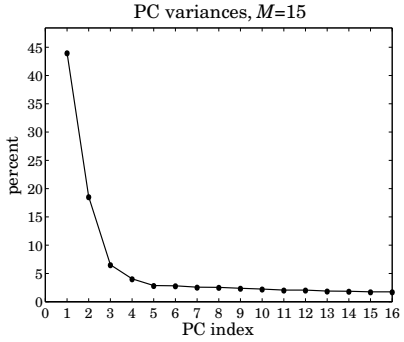
   (c) Using the *enhanced* data matrix Y_e, calculate the least-squares prediction error filter, $\mathbf{a}_{LS}$, by solving $Y_e\mathbf{a} = 0$.

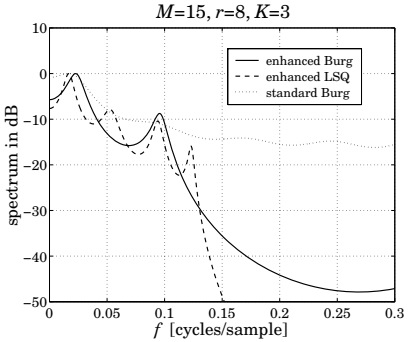
   (d) From the extracted *enhanced* signal $y_e(n)$, calculate the corresponding order-r Burg estimate of the prediction-error filter, say $\mathbf{a}_e$. (You could also do an order-M Burg estimate from $y_e(n)$, but the order-r choice is more appropriate since r is the assumed dimension of the signal subspace.)

   (e) Calculate and plot in dB the AR spectra of the three prediction filters, $\mathbf{a}_b$, $\mathbf{a}_{LS}$, $\mathbf{a}_e$. Normalize each spectrum to unity maximum.

   Identify the frequency of the highest peak in each spectrum and determine the corresponding period of the cycle in months. Identify also the frequency and period of the secondary peak that would represent the 10–12 month cycle.

   (f) Repeat the steps (a)–(e) for the following values of the parameters: For $M = 4$, $r = 3$, $K = 1, 3$. And then, for $M = 15$, $r = 5, 6, 7, 9$, and $K = 1, 3$. Moreover, do both the autocorrelation and forward-backward versions of the data matrices.



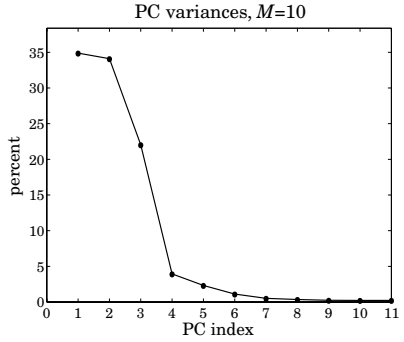


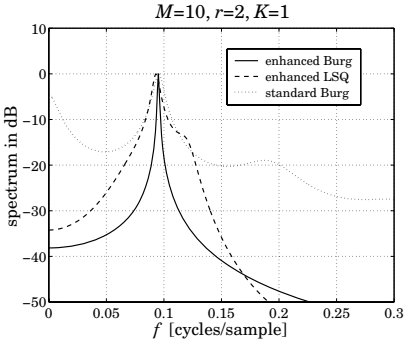
3. *Sunspot Numbers.* The Wolf sunspot numbers are of great historical importance in the development of spectral analysis methods (periodogram and parametric). Sunspot activity is cyclical and variation in the sunspot numbers has been correlated with weather and other terrestrial phenomena of economic significance. There is a strong 10-11 year cycle.

   The observed yearly number of sunspots over the period 1700–2004 can be obtained from the course's web page. The mean of this data must be removed.

   (a) It is desired to fit an AR model to this data, plot the AR spectrum, and identify the dominant peak corresponding to the 10–11 year cycle.

   (b) Perform the steps (a)–(e) as described in the previous experiment for the following values of the parameters: $M = 10$, $r = 2, 3, 4$, $K = 1, 3$. Try also the simpler case $M = 3$, $r = 2$, $K = 1, 3$.





4. *Extracting Sinusoids in Noise.* The file `sine1.dat` on the course's web page contains 25 samples of a signal consisting of two sinusoids of frequencies $f_1 = 0.20$, $f_2 = 0.25$ cycles/sample in additive zero-mean gaussian white noise. The SNRof both sinusoids was 0 dB. This signal was generated by

$$y(n) = \cos(2\pi f_1 n) + \cos(2\pi f_2 n) + 0.7071 v(n), \qquad n = 0, 1, \ldots, 24$$

   where the noise $v(n)$ was generated by MATLAB's `randn` function (with an initial state of 111, in case you want to regenerate the data.)

   In this experiment, because there are four complex sinusoids, the dimension of the signal subspace will be taken to be $r = 4$. (In the noiseless case, the rank of the data matrices would be 4.)

   Using the autocorrelation and the forward/backward methods, perform steps (a)–(e) of Question-1, for the following values of the parameters: $M = 10, 15, 20$, $r = 4$, $K = 1, 3$.

   Discuss your observations as to the appearance of false peaks, whether the value of M makes a big difference or not, and whether the autocorrelation or forward/backward method is better in any way.