MOL 410/510: Introduction to Biological Dynamics Fall 2012
Problem Set #2, PCA (due 10/5/2012) 4 Questions, all are **MUST DO**

Clearly explain your answers, show all your work, and include any figures where requested. Don't forget that you can use Matlab's help function to get information on any of the functions introduced in this homework.

1. **PCA**
   Let's carry out a Principal Component Analysis by hand for a simple data set:

$$X = \begin{bmatrix} 3 & 2 & 4 & 0 & 6 & 3 & 1 & 5 & -1 & 7 \\ 1 & 3 & -1 & 7 & -5 & 1 & 0 & 2 & -1 & 3 \end{bmatrix}$$

   where each row corresponds to a gene (call them gene 1 and gene 2), and each column corresponds to an assay (an experimental condition). Let's say the first five columns are assays every 10 min following heat shock, and the second five columns are assays every hour following a shift from glucose to glycerol.

   (a) (2 pts) Plot the data points in the gene 1 - gene 2 plane.

   (b) (2 pts) What is the mean expression value for each gene?

   (c) (2 pts) What is the variance of the expression values for each gene? (Use var $= \frac{1}{N}\sum_i (x_i - \text{sample mean})^2$.)

   (d) Perform a Principal Component Analysis of the data in the matrix X:

      i. (2 pts) Create a new matrix $\tilde{X}$ by subtracting off the mean expression value for each gene from matrix entries for that gene.

      ii. (4 pts) Evaluate the 2×2 gene-covariance matrix C using the data in $\tilde{X}$.

      iii. (5 pts) Evaluate the eigenvalues of C.

      iv. (2 pts) What fraction of the total variance of the data is accounted for by the first principal component of C? (The total variance of the data is the sum of the variances of gene 1 and gene 2 that you evaluated earlier.)

      v. (6 pts) Find the principal component eigenvectors and plot their directions on the same plot as the data points. Don't forget to order your eigenvectors appropriately.

   (e) (5 pts) Re-express the gene-assay matrix $\tilde{X}$ as a principal component-assay matrix by projecting each data point (column) onto the PCs. (use Matlab, include code)

   (f) (4 pts) For each principal component (row) of the new matrix plot the data as a time series for each block of assays (the first five columns and the second five columns).

   (g) (2 pts) What can you say about the responses of cells to the two assays, heat shock and shift from glucose to glycerol?

2. **PCA on uncentered data.** Let's see the difference between finding the principal components of uncentered vs. centered data. You should use Matlab to solve this problem, include your code. Imagine you have the following data matrix X.

$$\mathrm{X}^{\mathrm{T}} = \begin{pmatrix} 12.1 & 6.9 \\ 6.6 & 7.5 \\ 11.5 & 6.1 \\ 8.7 & 8.1 \\ 15.4 & 9.0 \\ 6.9 & 6.0 \\ 8.8 & 8.5 \\ 10.1 & 8.3 \\ 19.1 & 11.2 \\ 15.6 & 7.6 \\ 22.2 & 9.3 \\ 10.2 & 7.5 \\ 15.7 & 9.2 \\ 10.0 & 6.4 \\ 7.2 & 6.3 \\ 11.8 & 7.7 \\ 11.3 & 8.3 \\ 12.5 & 7.5 \\ 11.4 & 9.5 \\ 12.9 & 9.3 \end{pmatrix}$$

Where each column of X^{T} is a dimension and each row of X^{T} is a sample (i.e. data point); that is, the matrix X has 20 points of 2-dimensional data.

(a) (4 pts) Without centering, i.e. mean subtracting, the data, find the covariance matrix of X.

(b) (4 pts) Find the eigenvalues and eigenvectors of the covariance matrix. Reorder your eigenvectors and eigenvalues so that the eigenvector with the highest eigenvalue is in the first column.

(c) (4 pts) Transform the data X into the principal component space and plot each point, include a printout of your graph.

(d) (12 pts) Now redo the previous three steps, but subtract the mean of each dimension from all of the data points. Include your code and the figure.

3. **PCA in higher dimensions.** So far, for instructive purposes, we have done PCA in two dimensions. PCA is much more useful for larger dimensions. Let's try one example. For this problem, you will need to download the data from the wiki:

   http://tglab.princeton.edu/mol510/homework/.

   Unzip the file, then load the data into Matlab by typing $\mathrm{load\ pca\_homeworkdata}$. There are two variables Y and time. time is 1x1001 vector that contains the time for each of the data points in Y. Y is a 10x1001 matrix, each row is a dimension and each column is a time point.

   (a) (4 pts) On the same figure, plot all 10 dimensions of Y vs time. Choose a different color for each line. (Tip: to draw a curve that will not erase previous curves, you can use $\mathrm{line(time,Y(1,:),'color',colormatrix(1,:))}$, where $\mathrm{colormatrix}$ is a 10x3 matrix of color values. You can obtain this using $\mathrm{colmatrix{=}jet(10)}$; this produces 10 colors that change gradually from blue to red.

   (b) (4 pts) On different figures, plot several (3 should suffice, do more if you are curious) dimensions versus one another, use dots not lines. For example, $\mathrm{plot(Y(1,:),Y(2,:),'.k')}$ will plot the second dimension versus the first dimension for all time points.

   (c) (4 pts) Calculate the covariance matrix of Y and image it using the function $\mathrm{imagesc}$. Add a color bar to the graph using $\mathrm{colorbar}$. This figure will show

   (d) (22 pts) Find the principal components of the data and transform the data into principal component space. Repeat the last three steps for the transformed data. How does the data differ now? How is the covariance matrix different now?

   (e) (3 pts) Plot the sorted eigenvalues vs principal component number. This is called a "scree" plot because of the way that the data usually falls like a cliff.

   (f) (3 pts) Plot the cumulative percent variance as a function of the principal component number. As an example, say the problem had three dimensions and the variances in principal component space were 60 30 and 10. Then the percent variance for the first principal component is 60, the percent variance for the first two PCs is 90, and for all three is 100. You can use the Matlab function cumsum.