PCA Homework Problems

Due date: Wednesday, Sept. 12.
All data required for these problems can be found in the dropbox folder for the class.

1. Write your own MATLAB or Python script to create the PCT, $\mathbf{y} = f(\mathbf{x})$, of the random vector $\mathbf{x}$ with samples generate using the following synthetic data sets:

   (a) Samples from multi-variate zero-mean Gaussians with random covariance matrices and dimensions B = 2, 3, 10, 100, 400, and 1000. Scatter plot the samples of $\mathbf{y}$ in 3 dimensions (of course, for B = 2, you should scatter plot the result in 2 dimensions). Plot the eigenvalues of the covariance matrices. In the case of B = 2 and 3, whiten $\mathbf{y}$ by multiplying by $\Lambda^{-\frac{1}{2}}$ and scatter plot the results. What observations can you make?

   (b) In this excercise, you will using 5 spectra from the NASA Aster spectral library. They are in the dropbox folder stored in the file `Aster5Things.mat` together with the wavelengths, and the names of the materials that produced the spectra. Plot the 5 spectra. I will hereafter refer to them as $\mathbf{e}_m$ for $m = 1, \ldots, 5$. The notation $\mathcal{I}_3 = \{1,2,3\}$, $)I_4 = \{1,2,3,4\}$, and $\mathcal{I}_5 = \{1,2,3,4,5\}$ and $M_i = \max(\mathcal{I}_i)$ is also used in this problem.

   Use the code, DirichletSample.m (available in the dropbox folder), or a Python equivalent to create data sets X_i consisting 5000 convex combinations of the spectra using spectra $\mathbf{e}_m$ for $m \in \mathcal{I}_i$ for $i = 3, \ldots, 5$.

   For example, for $m \in \{1,2,3\}$, you will use DirichletSample to generate 5000 triples, $p_{n,1}, p_{n,2}$, and $p_{n,3}$ with the property that $p_{n,1} + p_{n,2} + p_{n,3} = 1$ and $0 \leq p_{n,1}, p_{n,2}, p_{n,3} \leq 1$, for $n = 1, \ldots, 5000$. You will the use triples to create 5000 spectra by $x_n = p_{n,1}\mathbf{e}_1 + p_{n,2}\mathbf{e}_2 + p_{n,3}\mathbf{e}_3$. Create other data sets, $X_{\mathcal{N},m}$ by adding noise from a multivariate Gaussian distribution to the samples in X. That is, write $x_n = p_{n,1}\mathbf{e}_1 + p_{n,2}\mathbf{e}_2 + p_{n,3}\mathbf{e}_3 + \mathbf{n}$ where $\mathbf{n}$ is a sample from a multi-variate Gaussian, which is often written $\mathbf{n} \sim \mathcal{G}(\mu, \mathbf{\Sigma})$. Calculate the PCT of these data sets, plot the eigenvalues, and scatter plot the results in 3D. What do you observe?

2. Verify empirically that the diagonal elements of $\Lambda = V\bar{\mathbf{C}}_X V^t = diag\left(\lambda_1^2, \lambda_2^2, \ldots, \lambda_B^2\right)$ are the variances of $y_1, y_2, \ldots, y_B$.

3. Calculate the PCT of the spectra in the images stored in the files `SanBarThings.mat` and `GulfPortCampusThings.mat`. Each file contains several data. They each contain the hyperspectral image, the wavelengths, and an RGB version of the hyperspectral image.

   In addition, `GulfPortCampusThings.mat` contains `GulfPortOceanMask` which is 1 for pixels that are not in the ocean and 0 otherwise. The ocean pixels should not be included in the calculation of the PCT. How would you refer to these hyperspectral images if you want to convey the wavelengths using acronyms? Scatter plot the results in 3 dimensions. What do the extreme points represent? Plot the eigenvalues of the covariance matrices. What do you observe?