**ECE 3522: Stochastic Processes in Signals and Systems**

# Computer Assignment (CA) No. 12: Principal Components Analysis

For this assignment, we will use one of the data sets available in MATLAB known as the “Cities data set.” Instructions for how to load this data set can be found here: <http://www.mathworks.com/help/stats/feature-transformation.html>. We will focus on processing the matrix “ratings”. The vector “categories” tells you what each column of the matrix represents.

1. Compute the covariance matrix for the raw data (contained in the data matrix “ratings”).
2. Do an eigenvalue/eigenvector analysis of the covariance matrix, and display both these matrices. Discuss what you observe.
3. Compute the whitening transform we discussed in class, and compute a matrix of transformed data.
4. Demonstrate that the covariance of the transformed data is an identity matrix by computing the covariance of the transformed data matrix.
5. Plot the percent of the variance accounted for by the first n eigenvalues for n = [1,9]. Make sure the eigenvalues are sorted from largest to smallest.
6. Analyze the eigenvector corresponding to the two largest eigenvalues, and describe what variables in the ratings matrix (e.g., “climate”) are most influential in explaining the variance of the data. Hint: look for eigenvector coefficients that have the largest positive or negative values.
7. Using the original data, find the three most similar cities using the following algorithm:
   1. Find the two cities that are the closest together using a Euclidean distance between their ratings.
   2. Find the third city that is the closest to the average of the ratings vectors for (1).

Does your result make sense?

1. Repeat (7) using the decorrelated, or transformed data. Did your ranking of the top three cities change? Does your new ranking make more sense? Explain your findings.
2. Repeat (8) using only the first 3 features from the transformed feature matrix (the first three columns in the transformed feature matrix). Explain any differences you observe from the result in (7) and (8).

Principal Components Analysis is a very useful tool for understanding correlation in data and understanding how various random variables are related.