**Classification of Cancerous Breast Tissue via Principal Component Analysis**

Jaceline Preval

December 16, 2016

**INTRODUCTION**

Principal Component Analysis (PCA) is a procedure done to high dimensional data to find the patterns and highlighting the similarities and differences. The main idea is once those patterns are found, the data can be compressed by reducing the numbers of dimension without losing the information. Since patterns can be hard to find in high dimensional data, PCA can be a powerful tool to analyze data.

PCA works first by normalizing the data matrix for each variable then taking the covariance matrix. The eigenvalue decomposition is applied to the covariance matrix to find the eigenvectors. The eigenvectors, which is a *p × p* matric, and the eigenvectors are paired so that the *i*th eigenvector corresponds to the *i*th eigenvalue. It turns out the eigenvectors are all the principal components for the covariance matrix and the first principal component is the eigenvector with the highest eigenvalue. The first principal component is also the linear combination with maximal variance which is essentially a dimension where the observations are maximally spread out or separated.

Assumptions should be made when using PCA for analysis. The data should be compromised of multiple variable that can be measure at the continuous level. Although ordinal variables can also be used. The variable should be linearly related to each other. This can be checked by looking at scatterplots of pair of variables. The variables should be at least moderately to each other or else PCA will be pointless.

**DESCRIPTION**

The Electrical Impedance Spectroscopy (EIS) data analyzed here was taken from a study of freshly excised breast tissue samples. EIS was performed in the range of 488 Hz to 1MHz. This data which is available publicly at the UCI Machine Learning Repository is composed 106 samples or instances. It also contains 10 variables or features of which 1 classifies each tissue sample as either carcinoma, fibro-adenoma, mastopathy, glandular, connective, or adipose tissue.

The 9 variables are defined as:

**I0:** Impedance (ohm) at 0Hz

**PA500:** Phase angle at 500kHz

**HFS:** High-frequency slope of the phase angle

**DA:** Distance between ends of the impedance spectrum

**AREA:** Area under the impedance spectrum

**A/DA:** AREA normalized by DA

**MAX IP:** Maximum of the spectrum

**DR:** Distance between I0 and the real component of the maximum frequency point

**P:** Length of the spectral curve

For this project, I want to determine if PCA can improve my classification rate by looking at my apparent error rate.

The first step will begin by creating randomly a training and test data. I will then train my data and test it to get my classification rate.

The second step will be to perform PCA on the overall data and repeat the first step to get a new classification rate.

Third step will be to test three different methods to decide how many of my 9 components should be retained to effectively summarize my data. For each method test, I will perform 2-Nearest Neighbor Classifier to see if my error rate was improved. I will also look if my error rate was improved for any k from 1 to 10.

These three methods:

1. Retain the component with the highest variance.
2. Retain the components whose eigenvalues are greater than the average of the eigenvalues.
3. Use the scree graph to assess how many components can be retained.

**SUMMARY**

For this project, I wanted to improve my error rate in classifying the cancerous tissue. I first start with a 6% error rate for a 2-Nearest Neighbor Classifier. But performing a principal component analysis on this data failed to lower my error rate on this data. I believe that further analysis could help improve the rate for this data.