Principal Components Analysis - Worksheet

Part One

1. Suppose we have a dataset with 8 variables and we use *standardized* data (i.e. correlation PCA). What is the total amount of variance in our data?

2. Suppose I have a dataset with 3 variables and the eigenvalues of the covariance matrix are

$$\lambda_1 = 3, \lambda_2 = 2, \lambda_3 = 1.$$

   a. What proportion of variance is explained by the first principal component?

   b. What is the variance of the second principal component?

   c. What proportion of variance is captured by using both the first and second principal components?

3. The following output is produced in SAS after running PCA on the iris dataset:

Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	2.91849782	2.00446735	0.7296	0.7296
2	0.91403047	0.76727360	0.2285	0.9581
3	0.14675688	0.12604204	0.0367	0.9948
4	0.02071484		0.0052	1.0000

Eigenvectors

	Prin1	Prin2	Prin3	Prin4
Sepal_Length	0.521066	0.377418	-.719566	-.261286
Sepal_Width	-.269347	0.923296	0.244382	0.123510
Petal_Length	0.580413	0.024492	0.142126	0.801449
Petal_Width	0.564857	0.066942	0.634273	-.523597

- How much variance in the data is captured by a projection onto the span of the first three principal components?

- Which variable is most closely associated with PC 2?

- Observations that have larger scores on PC3 are somewhat likely to have larger/smaller than average sepal lengths? *(circle one)*

- If you had to reduce the dimensions of this data down to 2 variables, which variables would you choose?

- What is the total amount of variance for this example? How do you know?

List of Key Words/Phrases.

eigenvalue
eigenvector
principal components
directional variance
proportion of variance
correlation vs covariance PCA
orthogonal projection onto PCs
PCA loadings
PCA coordinates
biplot
zero eigenvalues
small eigenvalues