**STA 3180 Fall 2024**

**LAB 6: Principal Components Analysis**

**(Total 50 point)**

# Crime Data

**Set-Up:** Crime is anissue in cities in the U.S. Sixteen U.S. cities were selected to analyze their crime profile.

**Dataset:** This dataset contains rates of seven different types of crimes per 100,000 population in sixteen U.S. cities.

# Analysis: Conduct a full principal components analysis to include the following

**i)** Generate descriptive statistics on the variables. Remember to include the standard

deviations of the variables to help determine if you should use the correlation or covariance matrix for your analysis. *Hint: You may want to put these summary statistics in a table*.

A correlation matrix should be used because the variances range greatly from 24.104 to 93,003.133. Moreover, there isn’t one variable that seems to be more important than the other, as all variables relate to crime levels (measured in the same units).

**ii)** Generate a correlation matrix for these variables. Remember to check to see if the signs are all positive or negative and if these correlations justify using PCA.

Most of the signs are positive, except for the correlation between the variables of murder and larceny. As a result, the correlation matrix was used.

**iii)** Use R to conduct a PCA analysis using the method chosen. Provide summary output.

**iv)** Construct a Scree Plot. What information does it convey?

The Scree Plot shows the eigenvalues of the variables and suggests which principal components are the most important. The number of principal components to use is determined by the elbow of the plot. In this case, the *m* would be 2, as the plot forms a sharp angle at vertex 2.

1. Construct a Bi-Plot.
2. Construct a Score Plot labeling by city.
3. What proportion of the variation in the data is explained by the first two principal components?

68.28232% of the variation in the data is explained by the first two principal components, as shown in part iii.

1. What proportion of the variation in the murder rate is explained by the first two principal components?

45.52% of the variation in the murder rate is explained by the first two principal components.

1. What proportion of the variation of burglary is explained by the first two principal components?

30.265% of the variation of burglary is explained by the first two principal components.

1. Use all this information to write a brief report on your PCA. Your report should include an attempt to label the first two principal components. It should also include comments on the placing of various cities in the space of the first two principal components.

Principal component analysis (PCA) is used to describe variance in a data set while also trying to use the least number of variables possible. This is done by transforming the original variables into uncorrelated variables (principal components) depending on the amount of variance they capture. Concerning the lab, the points (cities) that are close together have similar scores on the principal components in the plot, specifically regarding crime rates.

The first principal component accounts for most of the variance (49.2%), as all the crime variables have positive values on the Component 1 axis. The second principal component only accounts for 19.1 % of the variance. Murder and larceny are the two most highly correlated variables, with positively correlated variables being mostly violent, while negatively correlated variables are mostly non-violent.