Principal components analysis

The data for this exercise consist of rates of various crimes: murder, rape, robbery, assault, burglary, larceny, auto theft, for each of the US states. Each measure is expressed as number of crimes/100,000 population. We would like to determine if there are some linear combinations of these variables that account for a reasonable amount of the variance among states on all crime measures.

1. Read the data, **`crime.sas`** into SAS. (It is in **n: \data**).

   ```
   %include data(crime);
   proc print data=crime(obs=20); run;
   ```

2. Carry out a (default) principal components analysis of these data, and obtain an output data set containing scores for each state on the principal components.

   ```
   proc princomp data=crime out=prin;
   run;
   ```

   - How much variance is accounted for by 2 (or 3) components?
   - Can you find some interpretation for the component weights (eigenvectors) of the first 2 (or 3) components?

3. It is often helpful to plot the scores on several components to aid in interpreting the components. PROC PRINCOMP now does nice plots.

   ```
   *-- plot principal component scores;
   proc princomp data=crime out=prin plots(ncomp=3) =(pattern(vector) score);
   id st; run;
   ```

4. More useful is a biplot, that shows the principal component scores together with vectors representing the component weights (pattern or loadings)

   ```
   %biplot(data=crime,
         var=Murder Rape Robbery Assault Burglary Larceny Auto,
         id=ST);
   ```

   You can plot other combinations of components using the options DIM= and PLOTREQ=. E.g.,

   ```
   %biplot(data=crime,
         var=Murder Rape Robbery Assault Burglary Larceny Auto,
         dim=3, plotreq=Dim3*Dim1,
         id=ST);
   ```

5. From the biplot of components 1 & 2, try to estimate visually:
   - The correlation between auto theft and larceny; between auto theft and murder. Compare your guesses with the actual correlations.

- Which states are particularly high and particularly low on auto theft? Which are particularly high and low on crimes of personal violence?

6. You can also carry out a principal component analysis using PROC FACTOR, as follows:

```
proc factor data=crime
      scree rotate=varimax plot;
run;
```

- Can you think of any interpretation of the rotated 'factors' (components)?
- There are many other options for choosing the number of components/factors, as well as what is displayed. Use: `help factor` in the command bar.

7. Principal component analysis is distinctly different from any form of factor analysis (why?) The simplest form of (exploratory) factor analysis is "iterated principal factor analysis," followed by a varimax rotation, sometimes called "little jiffy."

```
proc factor data=crime method=prinit
      scree rotate=varimax plot;
run;
```

PCA in R

The same data is available for R in the file **n:\data\crime.csv**. In R, `princomp()` is the basic function for PCA. It returns an object, with standard methods, `print()`, `summary()`, `loadings()`. A `plot()` method gives a screeplot of variance of the components, and a `biplot()` method gives, well – a biplot. The following script will get you started.

```
crime <- read.csv("n:/data/crime.csv", row.names=9)
crime <- crime[,-1]  # drop state long names
str(crime)

# PCA on correlation matrix
(pca <- princomp(crime, cor=TRUE))
loadings(pca)
plot(pca)

# biplots to visualize variables and states
biplot(pca)
biplot(pca, choices=c(1,3))  # dim 1 & 3
```