**R: Principal Component Analysis**

*Goal:* Analyze and plot the principal components of a dataset using principal component analysis (PCA).

*Data:* **USArrests** from the base R package.

**Table of Contents**

1 --- Calculating Principal Components

2 --- Evaluation and Plotting

4 --- Summary Table

**Calculating Principal Components**

First, let us load and examine the data. We see this dataset has four variables. *Murder*, *Assault*, and *Rape* measure the rates of each respective crime in the given state, while *UrbanPop* represents the population of the given state living in the city.

Before we input the data to PCA, we must standardize using the **scale** function – this means transforming the data to have a mean of 0 and a standard deviation of 1.

data("USArrests")

head(USArrests)

# Standardize the data

X <- scale(USArrests)

Next we run the PCA using the **prcomp** function. This automatically calculates the loading matrix (aka eigenvectors) and transforms the data into the PCA space to be plotted. The loading matrix is output below.

# Run PCA, which performs singular-value decomposition

pc <- prcomp(X)

pc$rotation <- pc$rotation \* -1 # Reflect loadings matrix for positive values

pc

Standard deviations (1, .., p=4):

[1] 1.5748783 0.9948694 0.5971291 0.4164494

Rotation (n x k) = (4 x 4):

PC1 PC2 PC3 PC4

Murder 0.5358995 -0.4181809 0.3412327 -0.64922780

Assault 0.5831836 -0.1879856 0.2681484 0.74340748

UrbanPop 0.2781909 0.8728062 0.3780158 -0.13387773

Rape 0.5434321 0.1673186 -0.8177779 -0.08902432

**Evaluation and Plotting**

The **biplot** of a PCA plots the transformed points against the two principal components with the highest variance. Then, each original variable is plotted as a vector to demonstrate the relationship between the principal components and the original variables.

Based on the vector directions, we see that PC1 mainly captures variance in the three crimes – Murder, Assault, and Rape – while PC2 mainly captures variance in UrbanPop.

# View variances

pc$sdev^2

[1] 2.4802416 0.9897652 0.3565632 0.1734301

# Plot the biplot

biplot(pc, cex=0.6, xlab="First Principal Component", ylab="Second Principal Component", panel.first = c(abline(h = 0, v = 0, col="lightgray", lty="dotted")))

Chart

Description automatically generated

The **scree plot** of a PCA displays the variance for each principal component. There are two variants to the scree plot. The first, coded below, displays the variance captured by each principal component.

# Scree plots

layout(matrix(1:2, ncol=2)) # Place next to each other

screeplot(pc, main="", col.lab="white") # Plot as bars

title(xlab="Principal Components", ylab="Variance")

axis(1, at=c(0,0.7,1.9,3.1,4.3), labels=c(0,1,2,3,4), pos=c(0,0))

screeplot(pc, type="lines", main="", col.lab="white") # Plot as lines

title(xlab="Principal Components", ylab="Variance")

Chart, histogram

Description automatically generated

The second type of scree plot, rather than displaying the total variance, reveals the **proportion of variance explained** (PVE) by each principal component. Below, the left plot displays PVE for each individual component, 1 through 4. The right plot displays the cumulative PVE – the proportion of variance explained by just PC1, then the combination of PC1 and PC2, and so on.

# Calculate PVE

pve <- (pc$sdev^2) / sum(pc$sdev^2)

pve

[1] 0.62006039 0.24744129 0.08914080 0.04335752

# Plot PVE

plot(pve, type="b", pch=19, xlab="Principal Components", ylab="Proportion of Variance Explained (PVE)", xaxt="n", ylim=c(0,1))

axis(1, at=c(1,2,3,4), labels=c(1,2,3,4), pos=c(-0.05,1))

cum\_pve <- numeric(length(pve))

cum\_pve[1] <- pve[1]

for (i in 2:length(cum\_pve)){

cum\_pve[i] <- pve[i] + cum\_pve[i-1]

}

cum\_pve

plot(cum\_pve, type="b", pch=19, xlab="Combined Principal Components", ylab="Cumulative PVE", xaxt="n", ylim=c(0,1))

axis(1, at=c(1,2,3,4), labels=c(1,2,3,4), pos=c(-0.05,1))

Chart, line chart

Description automatically generated

As expected in PCA, we observe that the first principal component captures a large amount of variance in the dataset, about 62%, while subsequent components captures progressively less variance. Below, we include a summary table of the variance explained by each component.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Total Variance | Proportion of Variance Explained | Cumulative Proportion of Variance Explained |
| **PC1** | **2.480** | **62.0 %** | **62.0 %** |
| **PC2** | **0.990** | **24.7 %** | **86.7 %** |
| PC3 | 0.357 | 8.9 % | 95.7 % |
| PC4 | 0.173 | 4.3 % | 100 % |