# Principal Component Analysis

Sam Bowyer

2023-01-23

## Task 1

First we shall load the `USArrests` dataset and extract the three features of interest.

```
data(USArrests)
X0 = USArrests[, c("Murder", "Assault", "Rape")]
head(X0)
```

```
##            Murder Assault Rape
## Alabama      13.2     236 21.2
## Alaska       10.0     263 44.5
## Arizona       8.1     294 31.0
## Arkansas      8.8     190 19.5
## California    9.0     276 40.6
## Colorado      7.9     204 38.7
```

Next we must center the data.

```
X0 = as.matrix(X0)
n = nrow(X0)
p = ncol(X0)
C = diag(n) - matrix(rep(1/n,n*n), n)
X = C %*% X0
```

We can quickly check that indeed each feature now has zero mean (or at least very close to zero because of unavoidable floating point errors).

```
mean(X[,"Murder"]); mean(X[,"Assault"]); mean(X[,"Rape"])
```

```
## [1] 1.213438e-15
```

```
## [1] -2.685581e-14
```

```
## [1] -1.674771e-15
```

Now we may find the principal components by calculating the covariance matrix and performing a spectral decomposition to obtain the eigenvalues and eigenvectors.

```
# Covariance matrix
S = t(X) %*% X

# Spectral decomposition
decomp = eigen(S)
decomp
```

```
## eigen() decomposition
## $values
## [1] 342827.5561   2384.2733    329.5721
```

```
##
## $vectors
##             [,1]        [,2]        [,3]
## [1,] -0.04180743  0.02555358  0.99879886
## [2,] -0.99630506 -0.07612980 -0.03975532
## [3,] -0.07502247  0.99677042 -0.02864195
```

```
A = decomp$vectors
lambda = decomp$values

Y = X %*% A
```

Now we can see that the principal components are given by the columns of `Y` (in order from left to right).

```
Y
```

```
##              [,1]        [,2]       [,3]
##  [1,]  -65.222803  -4.8603090  2.8127790
##  [2,]  -93.737279  16.2271656 -2.1241283
##  [3,] -123.530499   0.3621893 -4.8675948
##  [4,]  -19.081280  -3.1652836  0.2955000
##  [5,] -106.354850  11.3245200 -3.5280427
##  [6,]  -34.432355  14.8838930 -1.7099189
##  [7,]   61.483255  -5.5883155 -1.7768759
##  [8,]  -66.505098 -10.5816700 -4.4032968
##  [9,] -164.751721  -1.6754982  0.7678911
## [10,]  -40.835871   1.7354050  7.8698642
## [11,]  124.480459   8.4057098  2.5044205
## [12,]   51.316900  -3.2775127 -2.9623783
## [13,]  -78.267771  -3.1305893 -0.5808744
## [14,]   57.588568   4.1509812  1.7156184
## [15,]  115.314712  -1.3060610 -0.7344958
## [16,]   55.871194   0.9777460  0.5234750
## [17,]   61.821875  -0.1654366  4.5062540
## [18,]  -78.341768  -4.7970082  4.4646754
## [19,]   88.681234  -6.8528175 -1.8075225
## [20,] -129.402041  -3.2024834 -1.8183161
## [21,]   22.191252  -3.3460627 -2.3775927
## [22,]  -85.149423   7.5202246  0.5606261
## [23,]   99.082846   1.0770124 -0.9742925
## [24,]  -87.951469 -10.6239479  4.9123553
## [25,]   -7.786676   6.4252874  0.7231386
## [26,]   61.969061  -0.1603078  0.8078340
## [27,]   69.006766   0.4288367 -0.6147010
## [28,]  -82.982434  18.6159669  0.4675747
## [29,]  114.457628  -3.1789329 -0.8225755
## [30,]   11.915223  -1.5387740  0.1496458
## [31,] -114.784242   2.2281318 -1.2452668
## [32,]  -83.436109  -1.4001329 -0.1406399
## [33,] -165.458638 -17.6380590 -1.2561940
## [34,]  126.632687  -4.4914899 -1.5809379
## [35,]   50.580243   4.0193361  1.5257543
## [36,]   19.829083   0.2459461 -0.3657211
## [37,]   11.232006   8.8634315 -2.6480918
## [38,]   65.057967  -1.4194080  1.2697025
## [39,]   -2.074387 -13.2490247 -4.1411389
```

```
## [40,] -108.211619  -6.8074247  2.2646244
## [41,]   85.246134  -2.0539137 -0.3720401
## [42,]  -17.827788   4.4755129  4.5577752
## [43,]  -30.653819   2.0775701  3.5816553
## [44,]   50.639120   5.4097220 -2.6122840
## [45,]  123.292654  -0.7966996 -0.4135891
## [46,]   14.715608   0.6115882  1.3131708
## [47,]   25.450473   6.8162622 -2.9016463
## [48,]   90.410804  -5.1134094  1.8247011
## [49,]  118.324415  -1.5658354 -0.2013894
## [50,]   10.187770  -4.8960311 -0.4374899
```

To plot the two-dimensional reduction of the dataset we can simply multiply `X` by `Y'` which only contains the first two columns of `Y`, i.e. the first two principle components of `X`.

```
X_reduction = X %*% A[,c(1,2)]
plot(X_reduction[,1], X_reduction[,2])
```