14) Principal Components Analysis

Vitor Kamada

March 2019

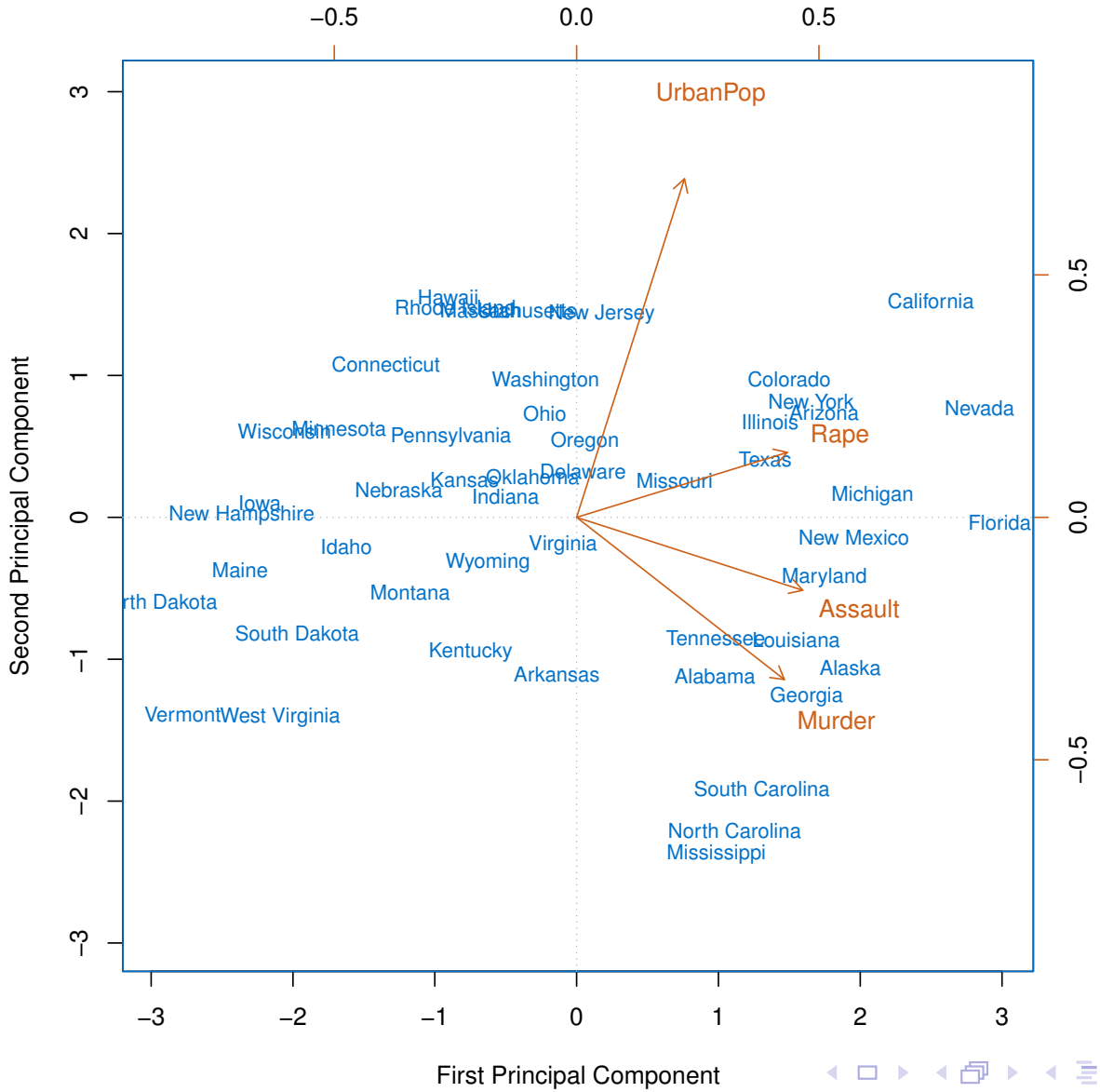
Reference

Tables, Graphics, and Figures from

James et al. (2017): Ch 10.2, and 10.4

Hastie et al. (2017): Ch 14.5

USArrests Data



Principal Component Analysis (PCA)

$$Z_1 = \phi_{11} X_1 + \phi_{21} X_2 + \ldots + \phi_{p1} X_p$$

$$z_{i1} = \phi_{11} x_{i1} + \phi_{21} x_{i2} + \ldots + \phi_{p1} x_{ip}$$

$$\underset{\phi_{11},\ldots,\phi_{p1}}{max} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j=1}^{p} \phi_{j1} x_{ij} \right)^2 \right\}$$

$$\text{subject to } \sum_{j=1}^{p} \phi_{j1}^2 = 1$$

$$z_{i2} = \phi_{12} x_{i1} + \phi_{22} x_{i2} + \ldots + \phi_{p2} x_{ip}$$

Singular Value Decomposition (SVD)

$$X_{n\times p} = U_{n\times p} D_{p\times p} V^T_{p\times p}$$

U and V are Orthogonal

$$U^T U = I_{n\times n} \text{ and } V^T V = I_{p\times p}$$

$$S = X^T X = V D^2 V^T$$

$$X X^T = U D^2 U^T$$

$$(S - \delta I) v = 0$$

Maximum Variance

$$z_1 = Xv_1 = u_1 d_1$$

$$Var(z_1) = \frac{d_1^2}{n}$$

Subsequent Principal Components z_j have maximum variance $\frac{d_j^2}{n}$, subject to being orthogonal to the earlier ones

OLS and Ridge Fitted Vector

$$X\hat{\beta}^{ls} = X(X^TX)^{-1}X^Ty$$

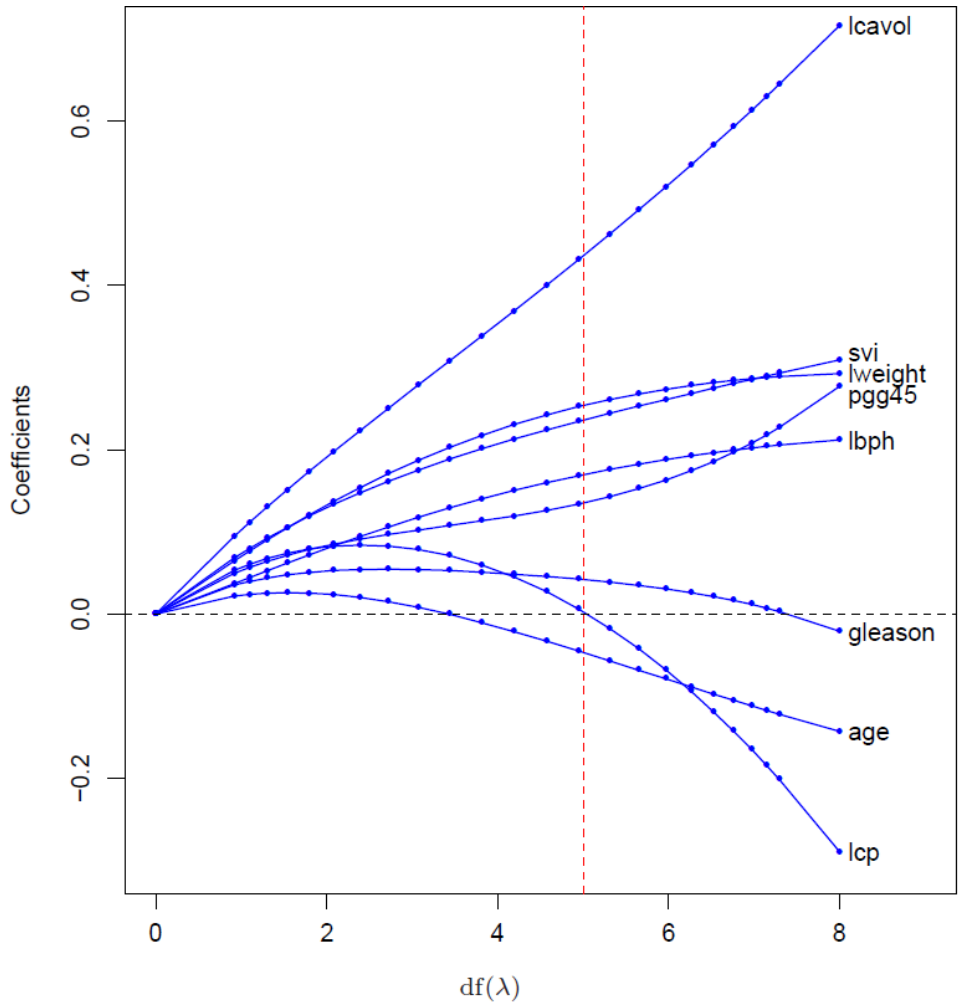
$$= UU^Ty$$

$$X\hat{\beta}^{ridge} = X(X^TX + \lambda I)^{-1}X^Ty$$

$$= UD(D^2 + \lambda I)^{-1}DU^Ty$$

$$= \sum_{j=1}^{p} u_j \frac{d_j^2}{d_j^2+\lambda} u_j^Ty$$

$$df(\lambda) = \sum_{j=1}^{p} \frac{d_j^2}{d_j^2+\lambda} = tr[X(X^TX + \lambda I)^{-1}X^T]$$



Effective Degrees of Freedom

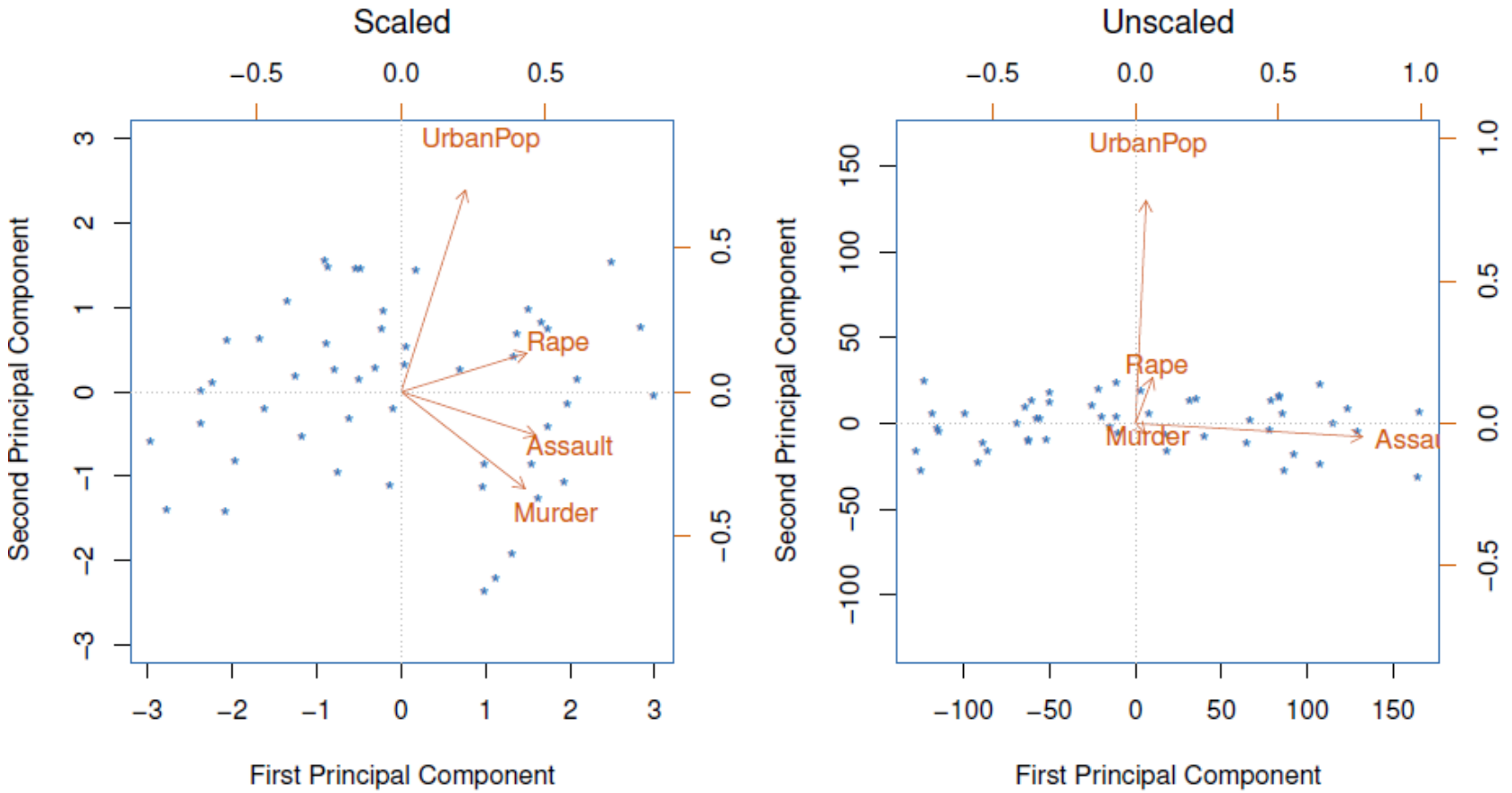
First and Second Principal Component

	PC1	PC2
Murder	0.5358995	−0.4181809
Assault	0.5831836	−0.1879856
UrbanPop	0.2781909	0.8728062
Rape	0.5434321	0.1673186

Scaling the Variables

Assault per 100 people rather per 100,00 people

Variance for Murder, Rape, Assault, and UrbanPop: 18.97, 87.73, 6945.16, and 209.5



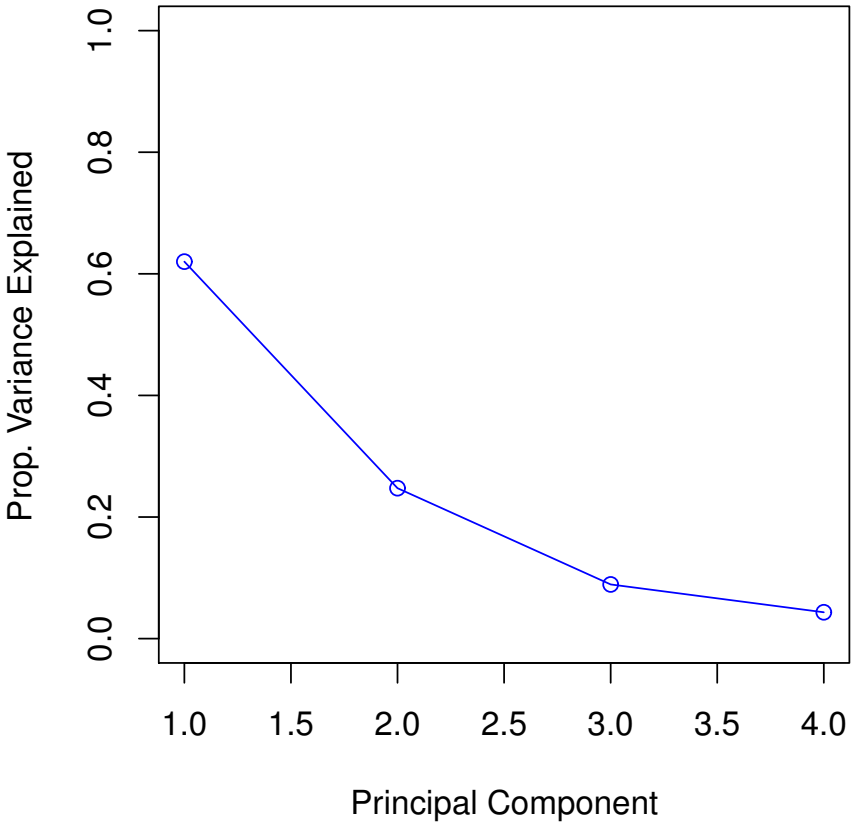
Proportion of Variance Explained (PVE)

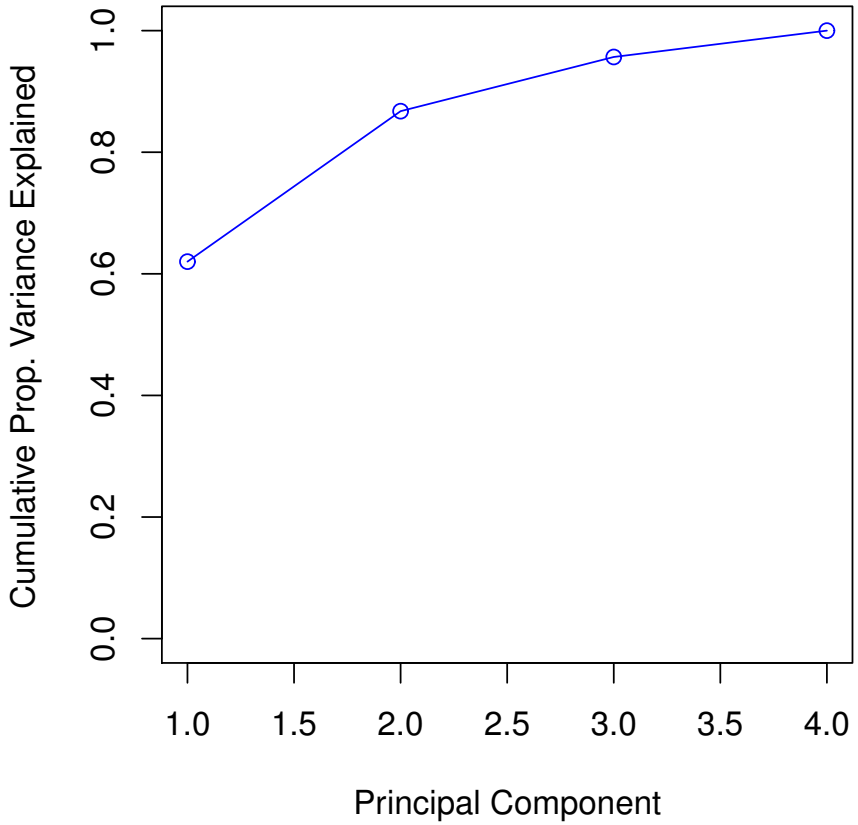
$$PVE = \frac{\frac{1}{n}\sum_{i=1}^{n} z_{im}^2}{\sum_{j=1}^{p} Var(X_j)}$$

$$\sum_{j=1}^{p} Var(X_j) = \sum_{j=1}^{p} \frac{1}{n} \sum_{i=1}^{n} x_{ij}^2$$

$$\frac{1}{n} \sum_{i=1}^{n} z_{im}^2 = \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j=1}^{p} \phi_{jm} x_{ij} \right)^2$$

Cumulative Proportion of Variance Explained





Means and Standard Deviations

```
pr.out=prcomp(USArrests, scale=TRUE)
```

```
pr.out$center
```

```
  Murder  Assault UrbanPop     Rape
   7.788  170.760   65.540   21.232
```

```
pr.out$scale
```

```
   Murder   Assault  UrbanPop      Rape
4.355510 83.337661 14.474763  9.366385
```

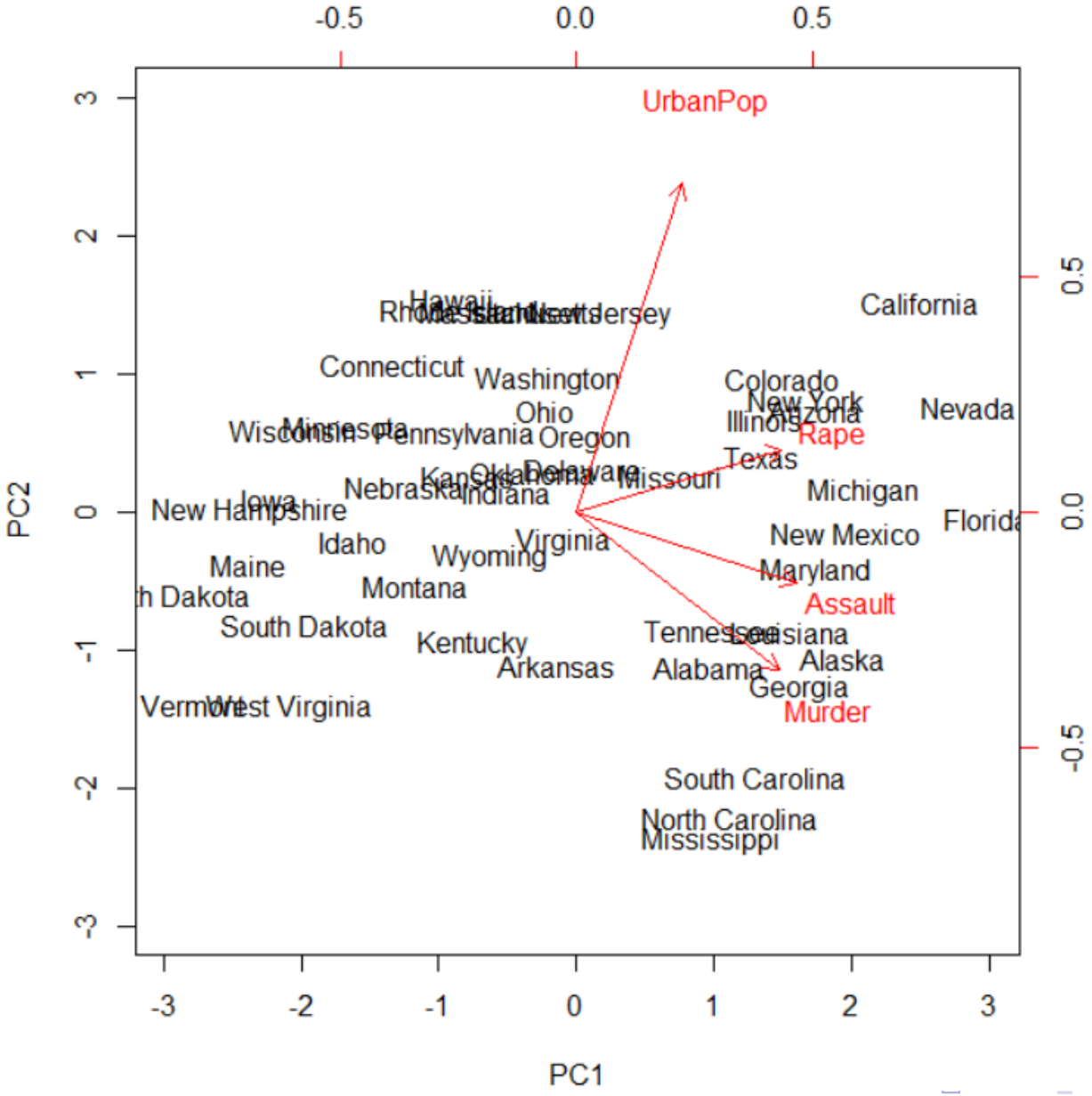
pr.out$rotation=-pr.out$rotation

pr.out$x=-pr.out$x

pr.out$rotation

	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

biplot(pr.out, scale=0)



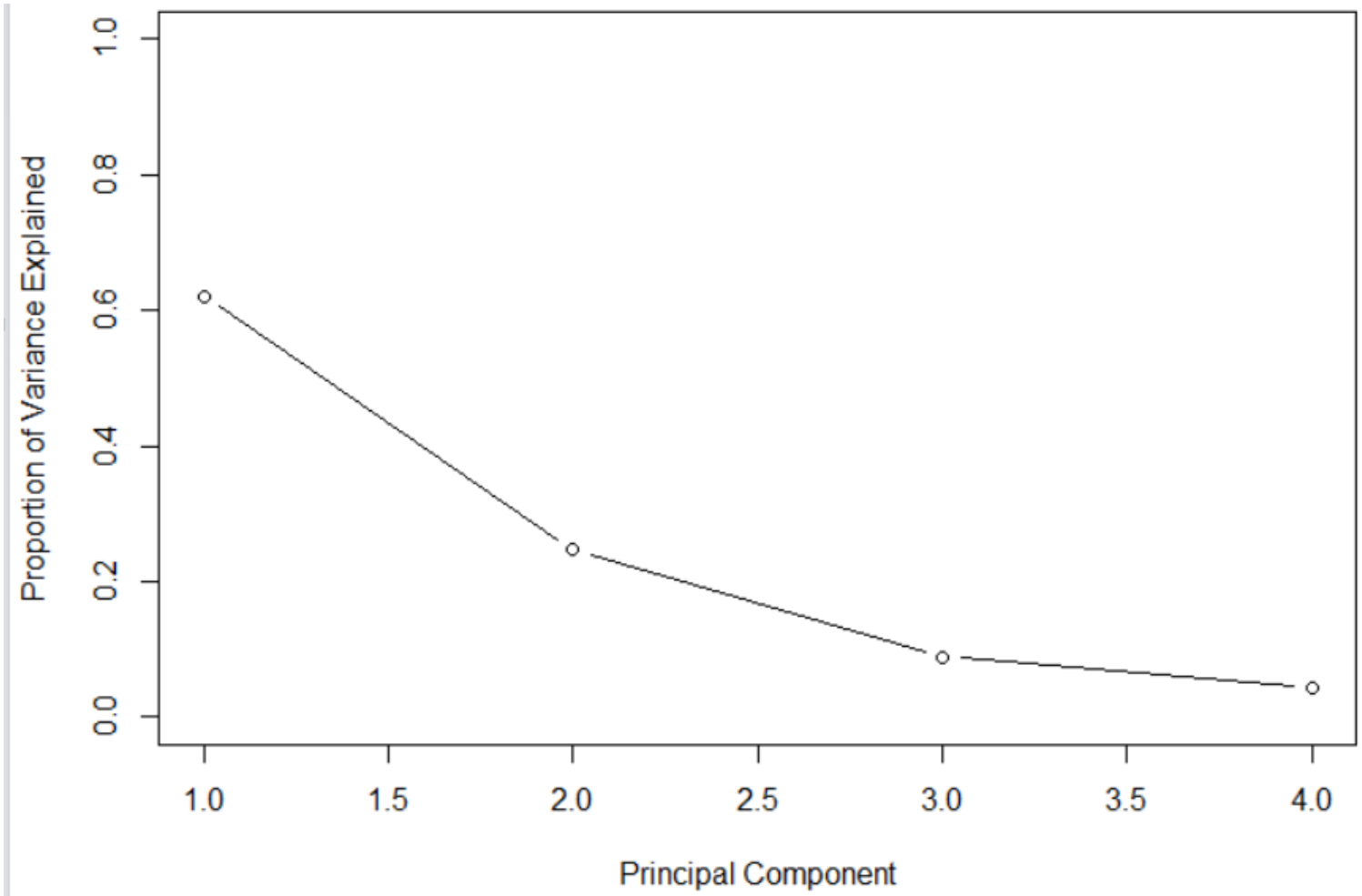
pr.var=pr.out$sdev^2; pr.var

```
2.4802416 0.9897652 0.3565632 0.1734301
```

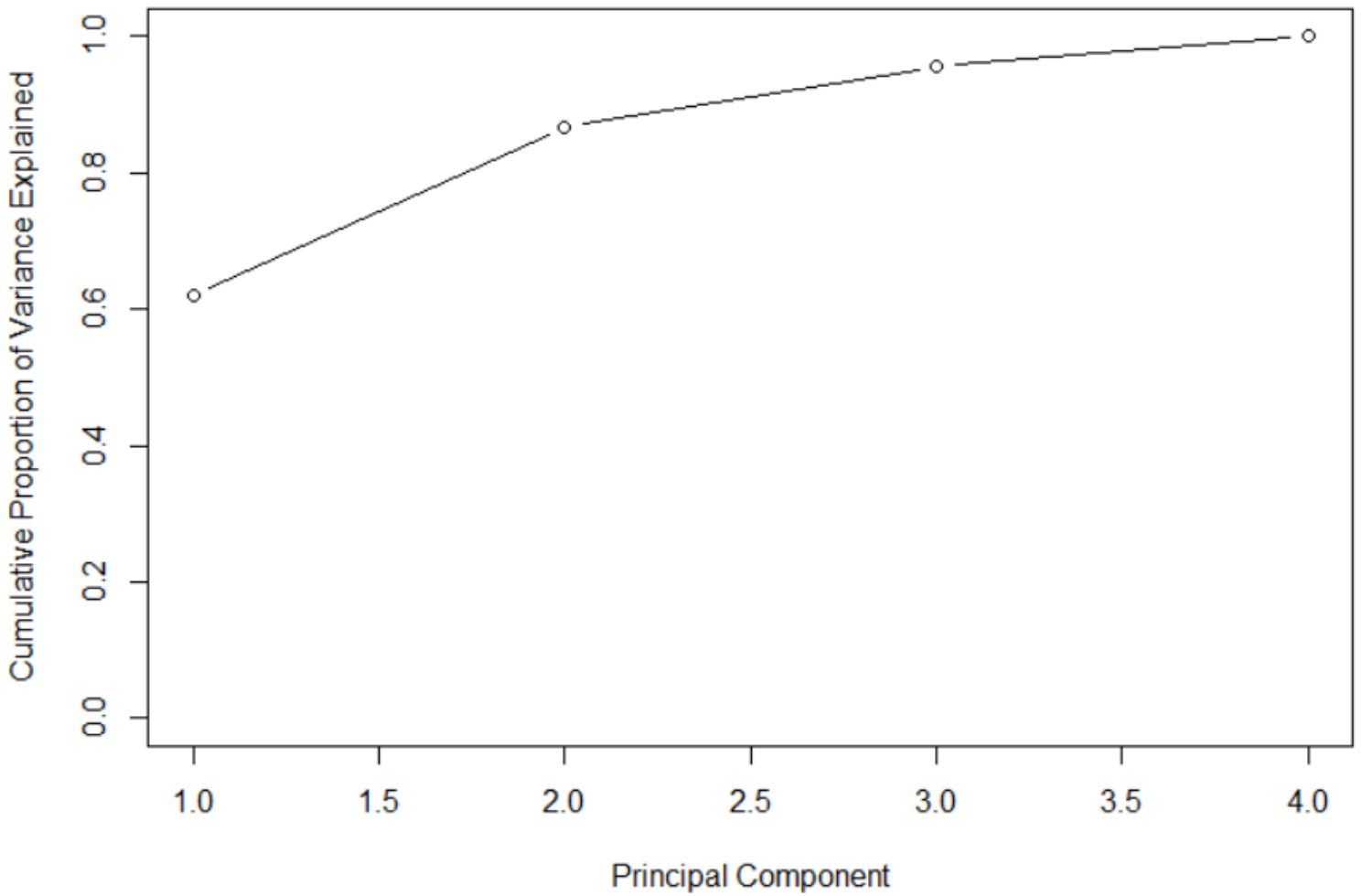
pve=pr.var/sum(pr.var); pve

```
0.62006039 0.24744129 0.08914080 0.04335752
```

plot(pve, xlab="Principal Component", ylab="Proportion of Variance Explained", ylim=c(0,1),type='b')



plot(cumsum(pve), xlab="Principal Component", ylab="Cumulative Proportion of Variance Explained", ylim=c(0,1),type='b')



6,830 Gene Expression on 64 Cancer Cell Lines

```
library(ISLR)
nci.labs=NCI60$labs; nci.data=NCI60$data
dim(nci.data)
```

64 6830

```
table(nci.labs)
```

```
     BREAST         CNS       COLON K562A-repro K562B-repro
          7           5           7           1           1
   LEUKEMIA MCF7A-repro MCF7D-repro    MELANOMA       NSCLC
          6           1           1           8           9
    OVARIAN    PROSTATE       RENAL     UNKNOWN
          6           2           9           1
```

pr.out=prcomp(nci.data, scale=TRUE)

summary (pr.out)

```
Importance of components:
                            PC1      PC2      PC3      PC4
Standard deviation     27.8535 21.48136 19.82046 17.03256
Proportion of Variance  0.1136  0.06756  0.05752  0.04248
Cumulative Proportion   0.1136  0.18115  0.23867  0.28115
                            PC5      PC6      PC7      PC8
Standard deviation     15.97181 15.72108 14.47145 13.54427
Proportion of Variance  0.03735  0.03619  0.03066  0.02686
Cumulative Proportion   0.31850  0.35468  0.38534  0.41220
                            PC9     PC10     PC11     PC12
Standard deviation     13.14400 12.73860 12.68672 12.15769
Proportion of Variance  0.02529  0.02376  0.02357  0.02164
Cumulative Proportion   0.43750  0.46126  0.48482  0.50646
```

Assign a color to each of the 64 cell lines

```
Cols=function (vec ){
  cols=rainbow (length (unique (vec )))
  return (cols[as.numeric (as.factor (vec))])
}
par(mfrow =c(1,2))
plot(pr.out$x [,1:2], col =Cols(nci.labs), pch =19,
  xlab ="Z1",ylab="Z2")
plot(pr.out$x[,c(1,3) ], col =Cols(nci.labs),
  pch =19, xlab ="Z1",ylab="Z3")
```

Projections of the NCI60 cancer cell lines onto the first three principal components

$\binom{6,830}{2}$ possible scatterplots