22) Principal Components Analysis

Vitor Kamada

January 2020

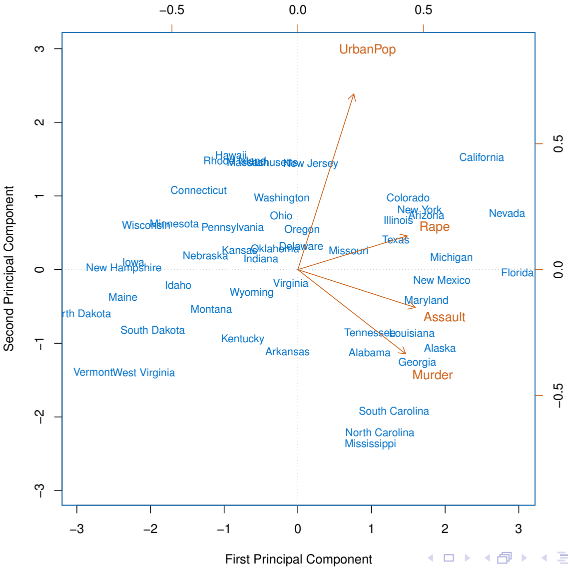
Reference

Tables, Graphics, and Figures from

James et al. (2017): Ch 10.2

Hastie et al. (2017): Ch 14.5

USArrests Data



Principal Component Analysis (PCA)

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

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

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

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

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

Singular Value Decomposition (SVD)

$$X_{n\times p} = U_{n\times p} D_{p\times p} V_{p\times p}^T$$

U and V are Orthogonal

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

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

$$XX^T = UD^2 U^T$$

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

Maximum Variance

$$z_1 = Xv_1 = u_1 d_1$$

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

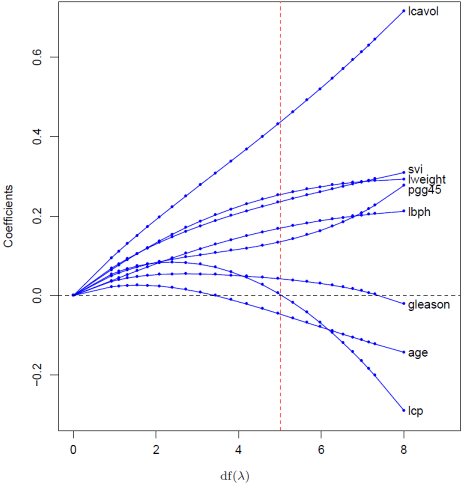
Subsequent Principal Components z_j have maximum variance $\frac{d_j^2}{n}$, subject to being orthogonal to the earlier ones

OLS and Ridge Fitted Vector

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

$$X\hat{\beta}^{ridge} = X(X^TX + \lambda I)^{-1}X^Ty$$
$$= UD(D^2 + \lambda I)^{-1}DU^Ty$$
$$= \sum_{j=1}^{p} u_j \frac{d_j^2}{d_j^2 + \lambda} u_j^T y$$

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



Effective Degrees of Freedom

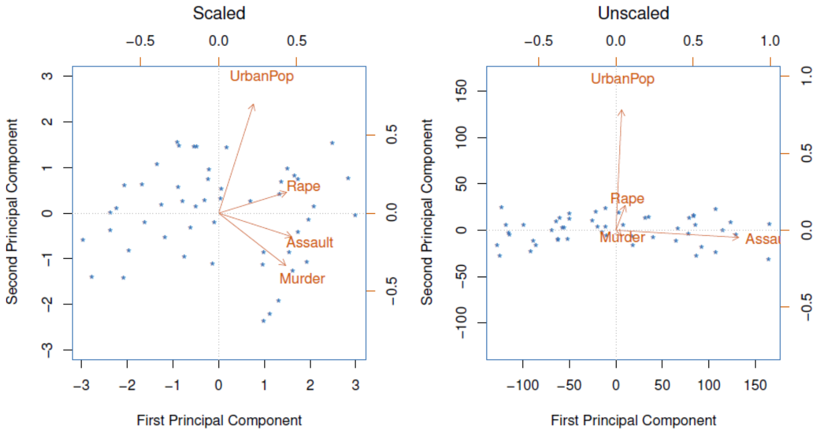
First and Second Principal Component

	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