Ch11-A Principal Component Analysis

Contents

11A Subsections
A.1 Principal Component Analysis
A.2 Princial Component
A.3 Computing 1st PC
A.4 Example:
A.5 Scaling
A.6 Another Interpretation
A.7 PVE

Textbook: James et al. ISLR 2ed.

11A Subsections

[ToC]

A.1 Principal Component Analysis

- Unsupervised Learning
- Can be part of exploratory data analysis
- No way to check your work
- Dimensionality reduction technique
- Ex.
  - Cancer researcher might look at gene expression levels in 100 patients
  - Online Shopping site

A.2 Princial Component

- Low-dimentional representation of data
- Can't see all paired scaterplot
- Which low-dimentional plot will make the observations most distinctive?
- The first principal component of ith observation is

$$Z_{i1} = \phi_{11}X_{i1} + \phi_{21}X_{i2} + \cdots + \phi_{p1}X_{ip}$$

- normalized: $\sum_{j=1}^{p} \phi_{j1}^2 = 1$.
- loadings: $\phi_{11}, \ldots, \phi_{p1}$
- orthogonal: $\sum_{j=1}^{p} \phi_{j1}\phi_{j2} = 0$.

A.3 Computing 1st PC

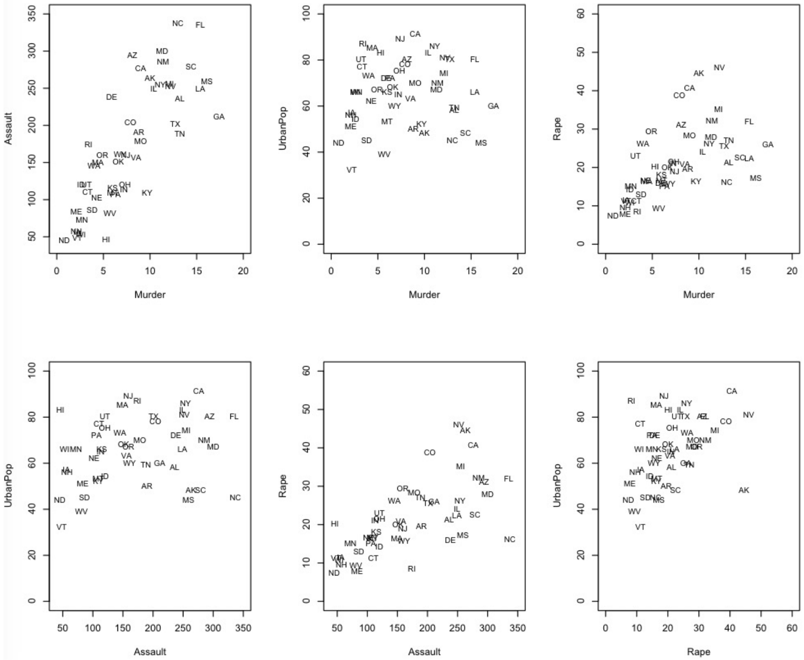
- Low-dimentional representation of data
- Can't see all paired scaterplot
- Which low-dimentional plot will make the observations most distinctive?
- The first principal component

$$\underset{\phi_{11},\ldots,\phi_{p1}}{\text{Maximize}}\left\{\frac{1}{n}\sum_{i=1}^{n}\left\{\sum_{j=1}^{p}\phi_{j1}x_{ij}\right\}^2\right\} \quad \text{subject to} \quad \sum_{h=1}^{p}\phi_{j1}^2=1.$$

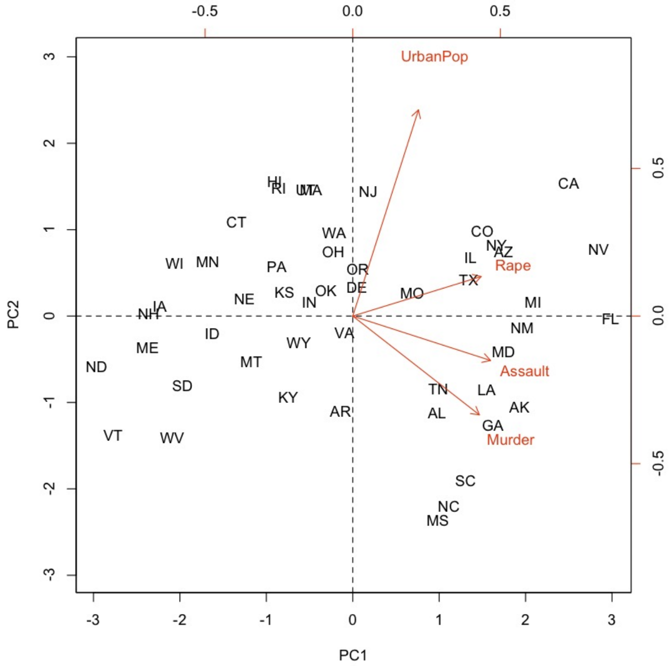
- It's maximizing average of Z_{i1}^2.
- Same as maximizing variance of Z_{i1}
- PC1 is the direction that the data varies the most.

- Second PC is same, but maximizing variance of Z_{i2} that is uncorrelated with Z_{i1}.
- PC2 will be orthogonal to PC1.

A.4 Example:



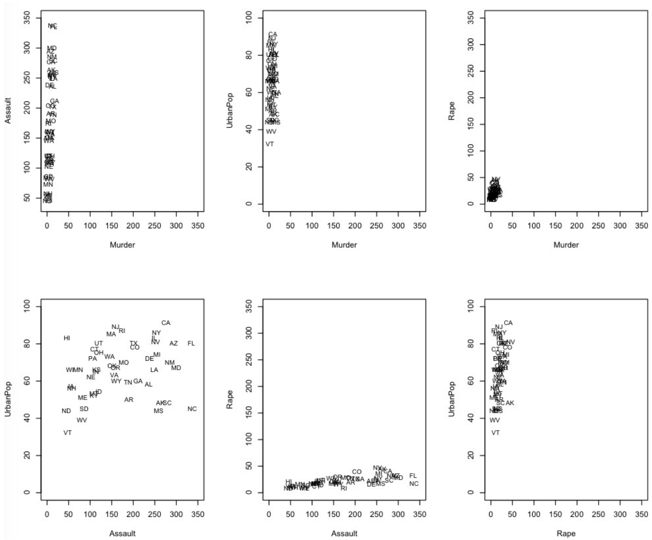
Assault
Murder
UrbanPop
Murder
Rape
Murder
UrbanPop
Assault
Rape
Assault
UrbanPop
Rape



PC1
PC2
-3
-2
-1
0
1
2
3
-0.5
0.0
0.5
UrbanPop
Rape
Assault
Murder
HI
RI
UT
MA
NJ
CA
CT
WA
CO
OH
NY
AZ
NV
WI
MN
PA
IL
OR
TX
KS
OK
DE
MO
NE
IN
MI
IA
NH
FL
ID
VA
NM
ME
WY
MD
MT
ND
SD
TN
LA
KY
AK
AR
AL
GA
VT
WV
SC
NC
MS

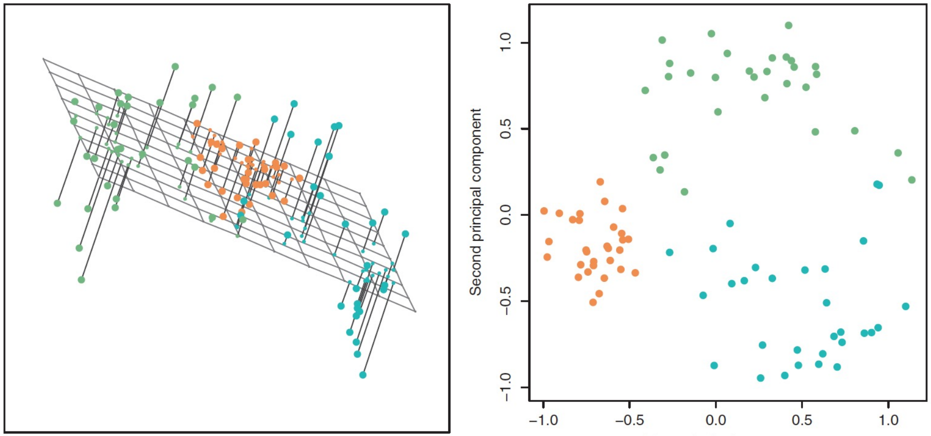
A.5 Scaling

- Scaling is important



A.6 Another Interpretation

- First two PC makes up a linear surface that is CLOSEST to obs.



-

A.7 PVE

- Proportion of Variation Explained
- Total variance in the data

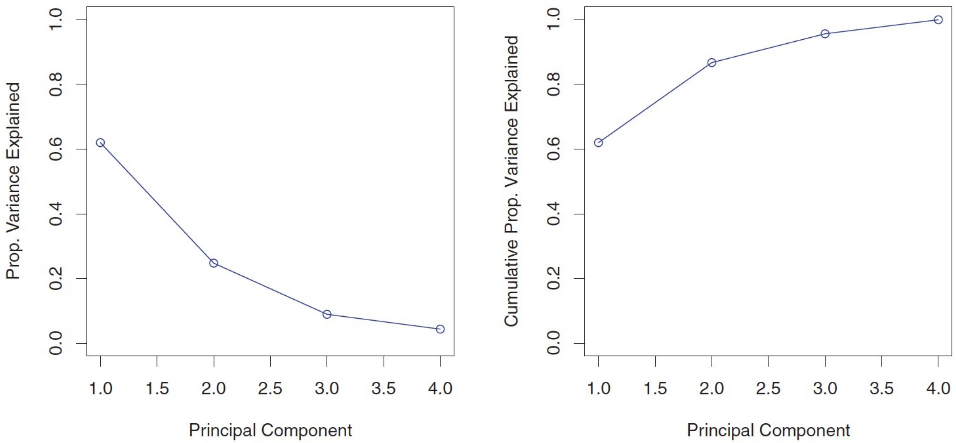
$$\sum_{j=1}^{p} Var(X_j) = \sum_{j=1}^{p} \frac{1}{n} \sum_{i=1}^{n} x_{ij}^2$$

- Variance explained by the mth PC is

$$\frac{1}{n} \sum_{i=1}^{n} Z_{im}^2$$

- PVE

$$\frac{\sum_{i=1}^{n} Z_{im}^2}{\sum_{j=1}^{p} \sum_{i=1}^{n} x_{ij}^2}$$



Prop. Variance Explained
0.0
0.2
0.4
0.6
0.8
1.0
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Principal Component
Cumulative Prop. Variance Explained
0.0
0.2
0.4
0.6
0.8
1.0
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Principal Component