12.2 Principal Components Analysis

Principal Components Analysis

- PCA produces a low-dimensional representation of a dataset. It finds a sequence of linear combinations of the variables that have maximal variance and are mutually uncorrelated.
- These can be used to produce derived variables for use in supervised learning problems (see Section 6.3).
- In the unsupervised learning context, PCA also serves as a tool for data visualization.

Principal Components: A Review

- The first principal component of a set of p features X is the normalized linear combination of the features

$$Z_1 = \sum_{j=1}^{p} \phi_{j1} X_j$$

  that has the largest variance.
- By normalized, we mean that $\sum_{j=1}^{p} \phi_{j1}^2 = 1$
  - The normalization prevents us from finding an arbitrarily large variance.
- The elements $\phi_{11}, \ldots, \phi_{p1}$ are called the *loadings* of the first principal component.
  - Together, they make up the first principal component loading vector.

Example

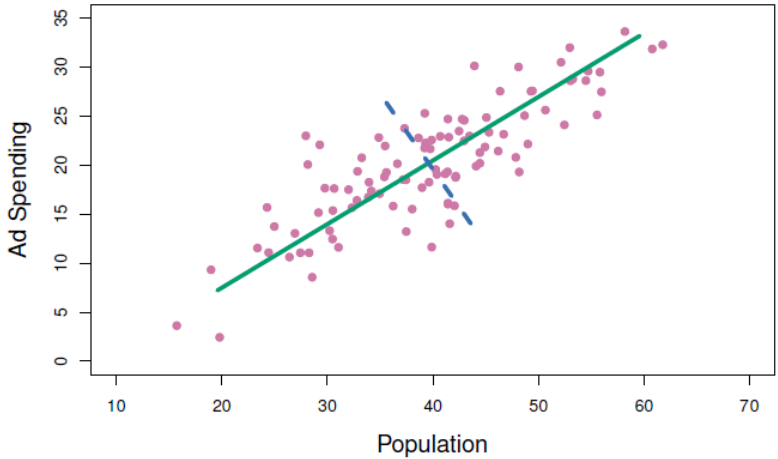


FIGURE 6.14. *The population size (pop) and ad spending (ad) for 100 different cities are shown as purple circles. The green solid line indicates the first principal component, and the blue dashed line indicates the second principal component.*

Computation of Principal Components

- Suppose we have a $n \times p$ data set X, centered such that each column has mean zero.
- We then look for the linear combination of the sample feature values of the form

$$z_{i1} = \sum_{j=1}^{p} \phi_{j1} x_{ij}$$

  such that $\sum_{j=1}^{p} \phi_{j1}^2 = 1$
- Since each of the x_{ij} has mean zero, then so does z_{i1} (for any values of ϕ_{j1}), and the sample variance can be written $\frac{1}{n} \sum_{i=1}^{n} z_{i1}^2$

Computation of Principal Components

- The first principal component solves the optimization problem

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

- This problem can be solved via a singular-value decomposition of the matrix X.

Geometry of PCA

- The loading vector ϕ_1 with elements $\phi_{11}, \phi_{21}, \ldots, \phi_{p1}$ defines a direction in feature space along which the data vary the most.
- If we project the n data points $x_1, \ldots, x_n$ onto this direction, the projected values are the principal component scores $z_{11}, \ldots, z_{n1}$ themselves.

Further Principal Components

- The second principal component is the linear combination of $X_1, \ldots, X_p$ that has maximal variance among all linear combinations that are uncorrelated with Z_1.
- The second principal component scores take the form

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

where and ϕ_2 is the second principal component loading vector.

Further Principal Components

- It turns out that constraining Z_2 to be uncorrelated with Z_1 is equivalent to constraining the direction ϕ_2 to be orthogonal to ϕ_1
  - ... and so on.
- The principal component directions $\phi_1, \phi_2, \phi_3, \ldots$ are the ordered sequence of right singular vectors of the matrix X
  - The variances of the components are $1/n$ times the squares of the singular values.
- There are are most $\min(n-1, p)$ principal components.

Example

- `USAarrests` data
  - For each of the 50 US states, the data contains the number of arrests per 100,000 residents for each of three violent crimes.
  - We also record the `UrbanPop`, the percent of the population in each state living in urban areas.
- The principal component score vectors have length $n = 50$ and the principal component loading vectors have length $p = 4$
- PCA was performed after standardizing each variable to have mean 0 and standard deviation 1.

Example

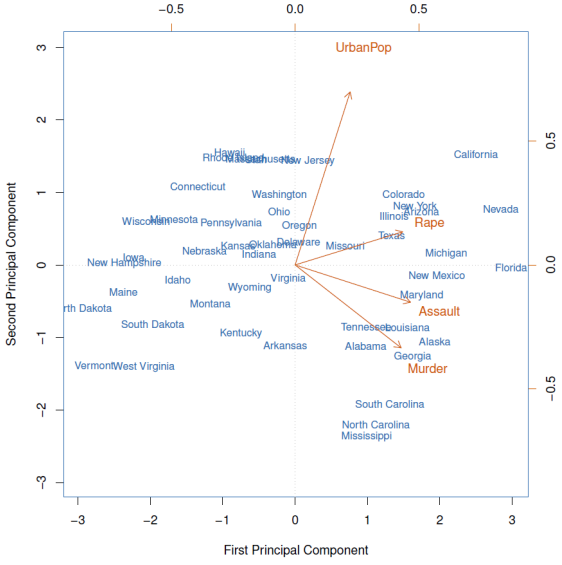

−0.5
0.0
0.5
UrbanPop
Hawaii
Rhode Island
Massachusetts
New Jersey
California
Connecticut
Washington
Colorado
New York
Michigan
Nevada
Minnesota
Ohio
Illinois
Rape
Wisconsin
Pennsylvania
Oregon
Texas
Nebraska
Kansas
Oklahoma
Delaware
Missouri
Indiana
Iowa
New Hampshire
Florida
Idaho
Virginia
New Mexico
Maine
Wyoming
Maryland
Montana
Assault
South Dakota
Kentucky
Tennessee
Louisiana
Alaska
Arkansas
Alabama
Georgia
Vermont
West Virginia
Murder
South Carolina
North Carolina
Mississippi
Second Principal Component
First Principal Component
3
2
1
0
−1
−2
−3
0.5
0.0
−0.5

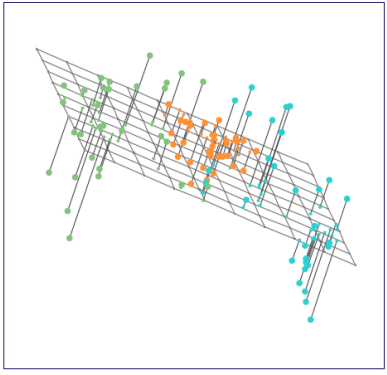
Figure Details

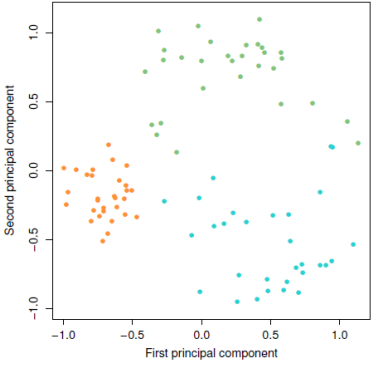
- The blue state names represent the scores for the first two principal components.
- The orange arrows indicate the first two principal component loading vectors (axes on top/right).
- This figure is known as a *biplot* because it displays both the principal component scores and the principal component loadings.

PCA Loadings

	PC1	PC2
Murder	0.5359	-0.4182
Assault	0.5832	-0.1880
Rape	0.5434	0.1673
UrbanPop	0.2782	0.8728

Another Interpretation



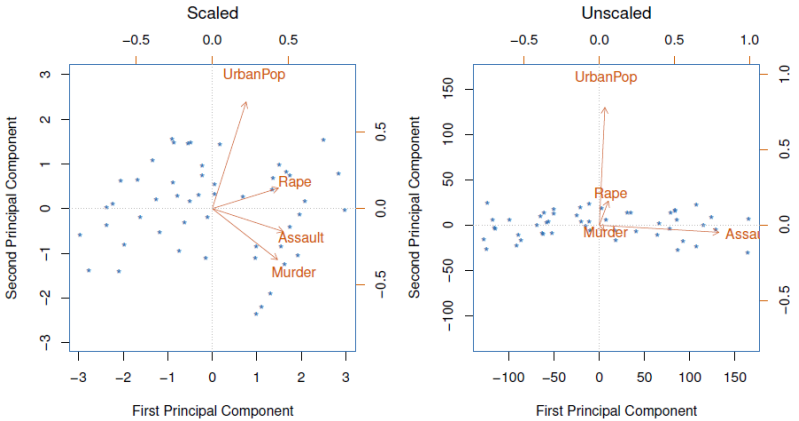


PCA Finds Hyperplanes

- The first principal component loading vector has a special property: it defines the line in p-dimensional space that is *closest* to the n observations.
- This idea extends before the first principal component.
  - For instance, the first *two* principal components span the *plane* that is closest to the n observations.

Scaling Matters

- If the variables are in different units, we typically want to scale to a standard deviation of 1.
  - . . . unless they are all in the same units/on the same scale to begin with.



Proportion of Variance Explained

- To understand the strength of each component, we examine the proportion of variance explained by each one.
- The *totl variance* in a data set with variables centered at zero is

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

  and the variance explained by the mth principal component is

$$\text{Var}(Z_m) = \frac{1}{n} \sum_{i=1}^{n} z_{im}^2$$

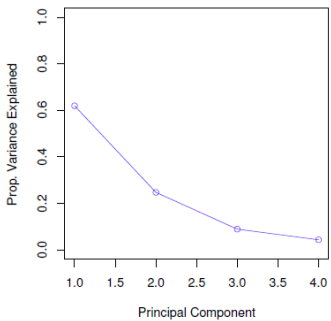
- It can be shown that $\sum_{j=1}^{p} \text{Var}(X_j) = \sum_{m=1}^{M} \text{Var}(Z_m)$ with $M = \min(n-1, p)$

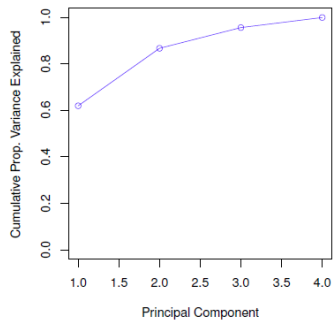
Proportion of Variance Explained

- So the PVE of the mth principal component is given by the positive quantity between 0 and 1,
$$\frac{\sum_{i=1}^{n} z_{im}^2}{\sum_{j=1}^{p} \sum_{i=1}^{n} x_{ij}^2}$$
- The PVEs sum to one.

Proportion of Variance Explained

We sometimes display the cumulative PVEs.





How Many Principal Components?

- If we use principal components to summarize our data, how many should we use?
- There is no simple answer to this question, as cross-validation is not available for this purpose.
  - Why not?
  - When could we use cross-validation to select the number of components?
- The *scree plot* from the previous slide can be used as a guide: we look for an "elbow".

Principal Components Analysis

There is a great overview of principal components analysis, going from high level conceptual to more mathematically detailed, here