Unsupervised Learning III: Principal Component Analysis

Yudong Chen
School of ORIE, Cornell University
ORIE 4740 Lec 16–17

Suppose that $X \in \mathbb{R}^{n \times p}$ is a data matrix with n observations and p predictors.

Which of the following is true?

A. $X^\top X$ is a square matrix.

B. $X^\top X$ is a symmetric matrix.

C. All the eigen values of $X^\top X$ are non-negative.

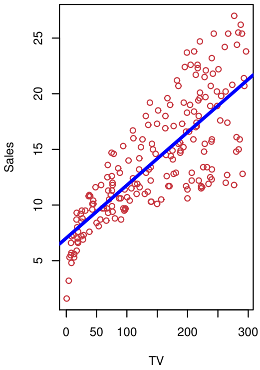
D. If u^1 is an eigen vector of $X^\top X$ with eigen value λ_1, then $u^{1\top} X^\top X u^1 = \lambda_1$

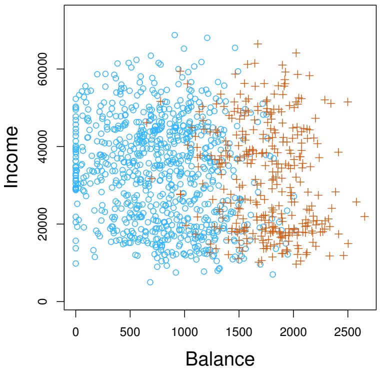
E. All of the above.

Supervised Learning

- Learn a rule for predicting an response variable based on some predictor variables.
- Have a set of training data, in which the predictors <u>and</u> response values are known for each samples.

$$(x_{11}, x_{12}, x_{13}, y_1), (x_{21}, x_{22}, x_{23}, y_2), \ldots\ldots, (x_{n1}, x_{n2}, x_{n3}, y_n)$$





Unsupervised Learning

- A set of p variables/features measured on n observations

$$(x_{11}, x_{12}, x_{13}), (x_{21}, x_{22}, x_{23}), \ldots\ldots, (x_{n1}, x_{n2}, x_{n3})$$

- No associated response y

- Goal: Discover interesting patterns about the data

Unsupervised Learning

Often more challenging than supervised learning.

Difficulties:

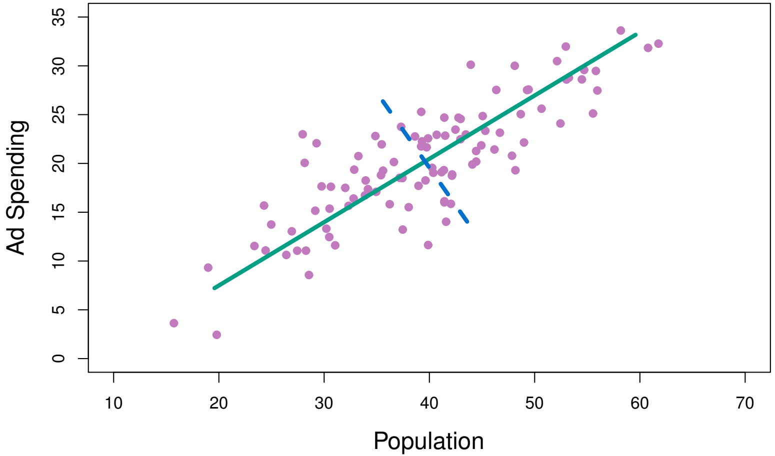
- No simple goal (want to find "interesting patterns")
- Contrast to supervised learning: predict the response
- No true answer (No Y)
- Difficult to assess model accuracy

Used in exploratory data analysis

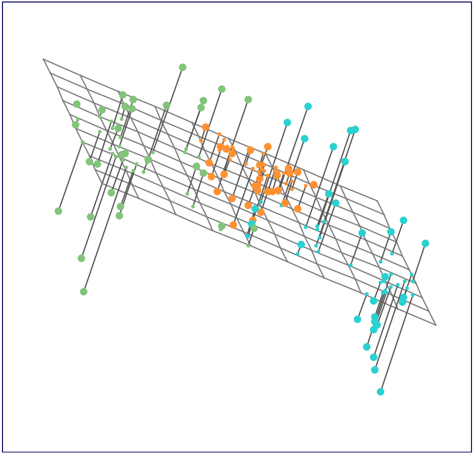
ISLR 10.2 (also cf. 6.3.1)

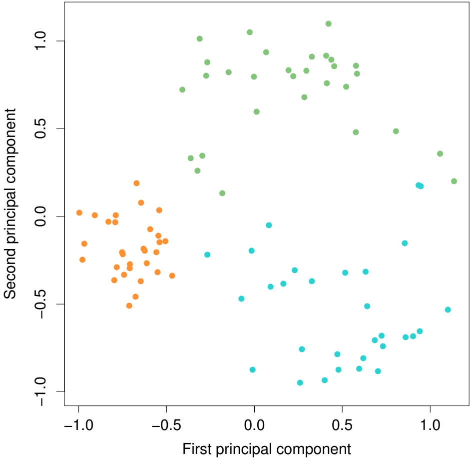
Find low-dimensional structures in the dataset

PCA in 2 dimensions



PCA in 3 dimensions





PCA

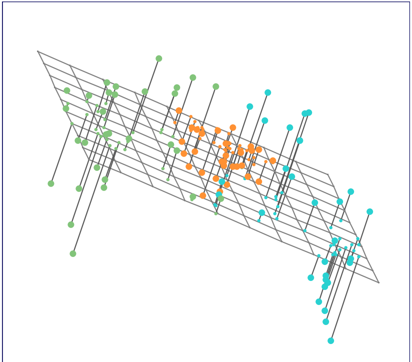
Idea: Find a succinct representation that best summarizes the data

Original data: p features/variables (p dimensional)

- Find a r-dim subspace on which the variance of the data is maximized
- (Equivalent) Find a r-dim subspace closest to the data

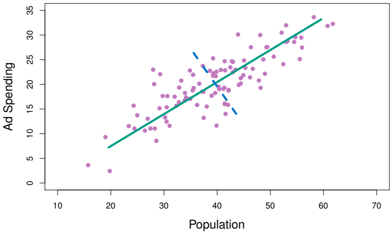
where $r < p$

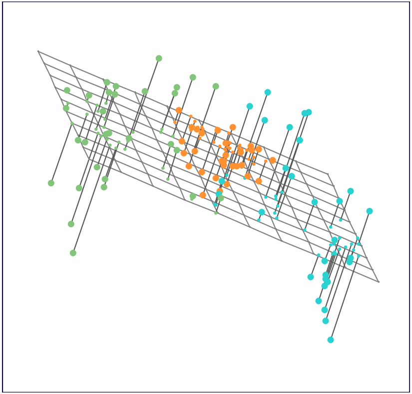
The first $r = 2$ principal components:



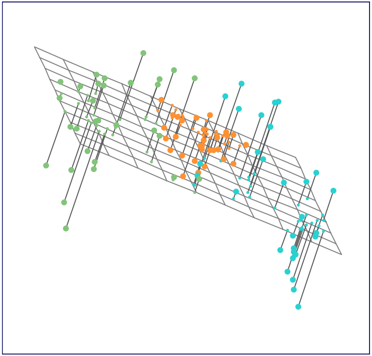
Principal Components

- $r = 1$: the first PC — a straight line
- $r = 2$: the first two PCs — a plane
- In general: the first r PCs — an r-dimensional subspace





Principal Components



A dimension reduction technique:

Reduce the original p-dim data to r-dim, such that (hopefully)

- Most of the information is kept
- Most of the noise is dropped
- Easier to store, manipulate and visualize

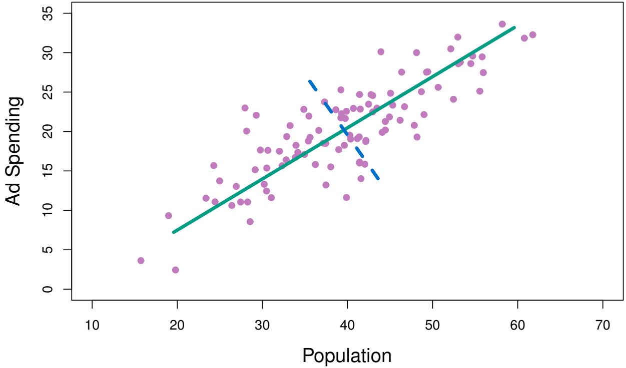
The First Principal Component: Mathematical Definitions

The 1st PC of features $X_1, \ldots, X_p$:

Linearly combine features in a way that retains the largest variance $\mathrm{Var}(Z_1)$

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

where $\phi_{11}^2 + \phi_{21}^2 + \cdots + \phi_{p1}^2 = 1$



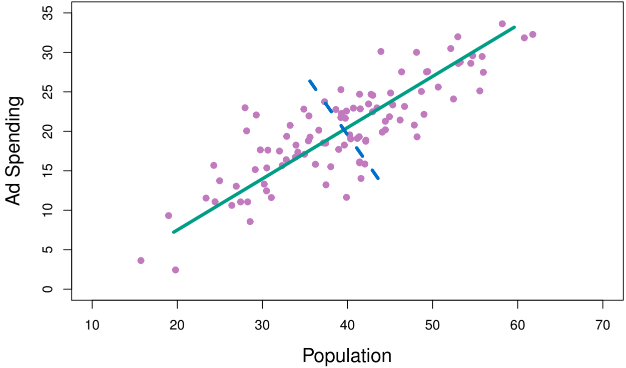
The First Principal Component: Loadings and Scores

Loadings: The direction of the line

- specified by p numbers $(\phi_{11}, \phi_{21}, \ldots, \phi_{p1})$

Scores: The projection of each observation onto this line

- specified by n numbers z_{i1}, $i = 1, 2, \ldots, n$



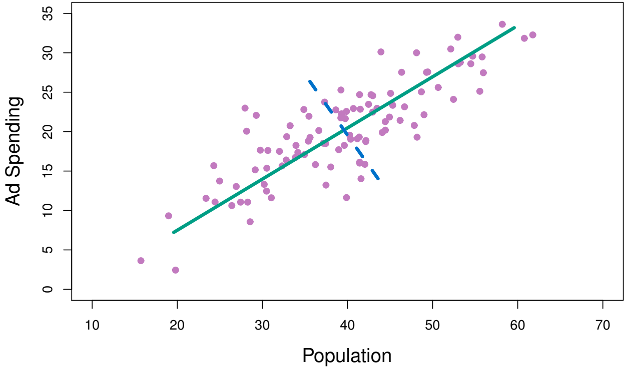
The Second Principal Component

The 2nd PC of features $X_1, \ldots, X_p$:

Linear combination of the features

$$Z_2 = \phi_{12} X_1 + \phi_{22} X_2 + \cdots + \phi_{p2} X_p,$$

with maximal variance, out of all combinations that are uncorrelated with Z_1

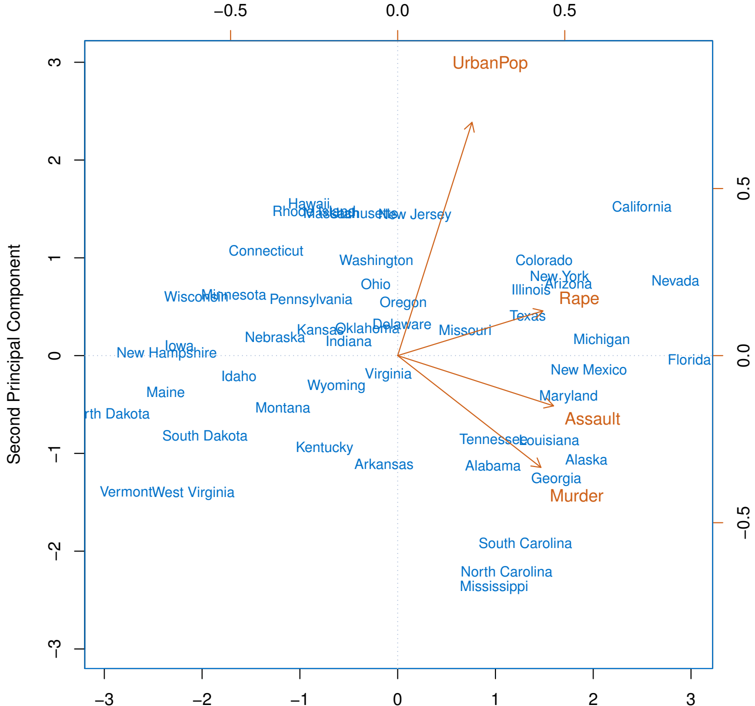


`USArrests` example

$p = 4$ features: `Assault, Murder, Rape, UrbanPop`

Find the loadings of the first 2 PCs:

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


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

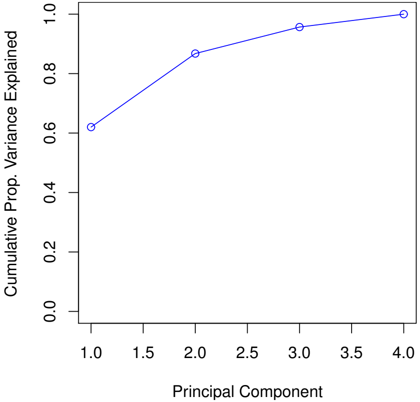
Choosing the number of principal components

In PCA, want to find directions that retains the most variance

Proportion of Variance Explained (PVE) of the m-th PC direction

$$= \frac{\text{Var. explained by } m\text{-th PC}}{\text{total Var.}}$$

```
> pr.var = pr.out$sdev^2
> pve = pr.var / sum(pr.var)
> plot( cumsum(pve) )
```



PCA: Math and Computation (Optional)

Recall: For first PC, want to find the normalized loadings $\phi_{j1}, j = 1, 2, \ldots, p$ that maximizes the variance of the linear combination

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

Mathematically: (assume x_{ij}'s centered)
Want to solve

$$\max_{\phi_{j1}, j=1,\ldots,p} \quad \mathrm{Var}(\{z_{i1}\}) = \frac{1}{n} \sum_{i=1}^{n} \Big(\sum_{j=1}^{p} \phi_{j1} x_{ij} \Big)^2$$

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

PCA: Math and Computation (Optional)

- Want to solve

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

- In matrix/vector notation:

$$\frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{p}\phi_{j1}x_{ij}\right)^2 = \frac{1}{n}\sum_{i=1}^{n}\langle\vec{\phi}_1, \vec{x}_i\rangle^2$$

$$= (\vec{\phi}_1)^\top \underbrace{\frac{X^\top X}{n}}_{\text{Cov}(X)} \vec{\phi}_1 \qquad \text{(Lecture 25 Sec 5.1)}$$

- From notes on linear algebra:

This is maximized when $\vec{\phi}_1 =$ the first eigenvector of covariance matrix $\frac{X^\top X}{n}$

PCA: Math and Computation (Optional)

1st PC $\vec{\phi}_1$ = the first eigenvector of covariance matrix $\frac{1}{n}X^\top X$

Similarly......

m-th PC $\vec{\phi}_k$ = the m-th eigenvector of covariance matrix $\frac{1}{n}X^\top X$