Lecture 14 – Principle Component Analysis

Hamed Hasheminia

Agenda

- Introduction of PCA method
- Computation of PCA
- Geometry of PCA
- Proportion of Variance explained

Principle Component Analysis

- PCA is an unsupervised learning technique.
- PCA produces a low-dimensional representation of the variables that have maximal variance, and are mutually uncorrelated
- Apart from producing derived variables for use in supervised learning problems, PCA also serves as a tool for data visualization.

Principle Component Analysis: details

- The *first principal component* of a set of features $X_1,\ldots,X_p$ is the normalized linear combination of the features:

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

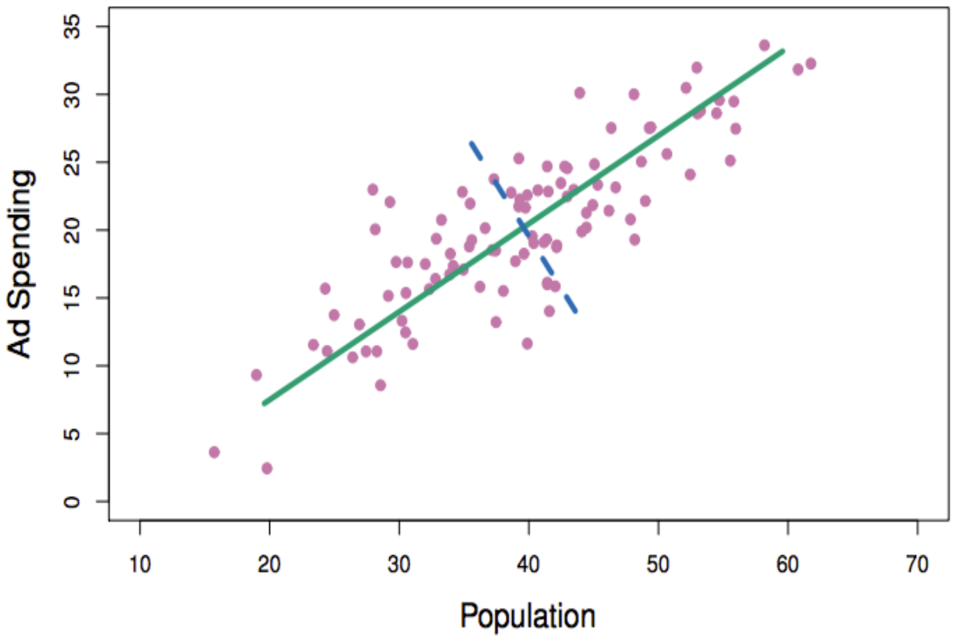
That has the largest variance. By normalized, we mean that

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

- We refer to elements $\phi_{11}, \ldots, \phi_{p1}$ as the loadings of the first principal component; together, the loadings make up the principal component loading vector,

$$\phi_1 = (\phi_{11}\ \phi_{21}\ \ldots\ \phi_{p1})^T.$$

PCA: Example



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 direction, and the blue dashed line indicates the second principal component direction.

Computation of Principle Components

- Suppose we have a n x p data set X. Since we are only interested in variance, we assume that each of the variables in X has been centered to have mean zero (this is, the column means of X are zero).
- We then look for the linear combination of the sample feature values of the form

$$z_{i1} = \phi_{11}x_{i1} + \phi_{21}x_{i2} + \ldots + \phi_{p1}x_{ip} \qquad (1)$$

- For $i = 1,\ldots,n$ that has the largest sample variance, subject to the constraint 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} Hence the sample variance of the z_{i1} can be written as $\frac{1}{n}\sum_{i=1}^{n} z_{i1}^2$

Computation: continued

- Plugging in (1) the first principle component loading vector solves the optimization problem:

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

- This problem can be solved via a singular-value decomposition of the matrix X, a standard technique in linear algebra.
- We refer to Z_1 as the first principal component, with realized values $z_{11}, \ldots, z_{n1}$

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 n data points $x_1,\ldots,x_n$ onto this direction, the projected values are the principal component scores $z_{11}, \ldots, z_{n1}$
- 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 $z_{12}, z_{22}, \ldots, z_{n2}$

take the form

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

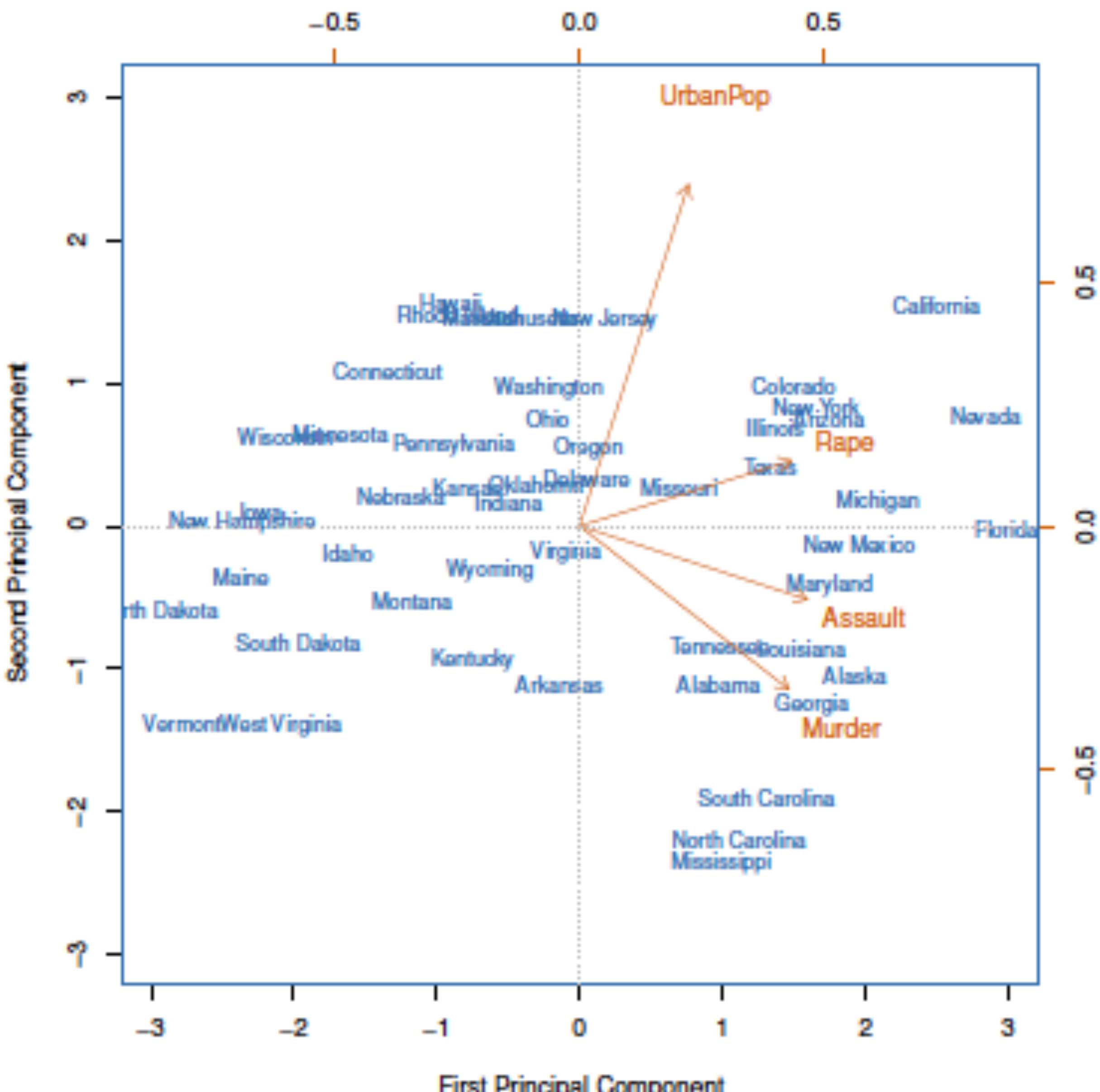
Where ϕ_2 is the second principal component loading vector, with elements $\phi_{12}, \phi_{22}, \ldots, \phi_{p2}$

Further principal components: continued

- 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 the direction ϕ_1. And so on.
- There are at most min(n-1,p) principal components.

Illustration

- `USAarrests` data: For each of the fifty states in the United States, the data set contains the number of arrests per $100,000$ residents for each of three crimes: `Assault`, `Murder`, and `Rape`. We also record `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 zero and standard deviation one.



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

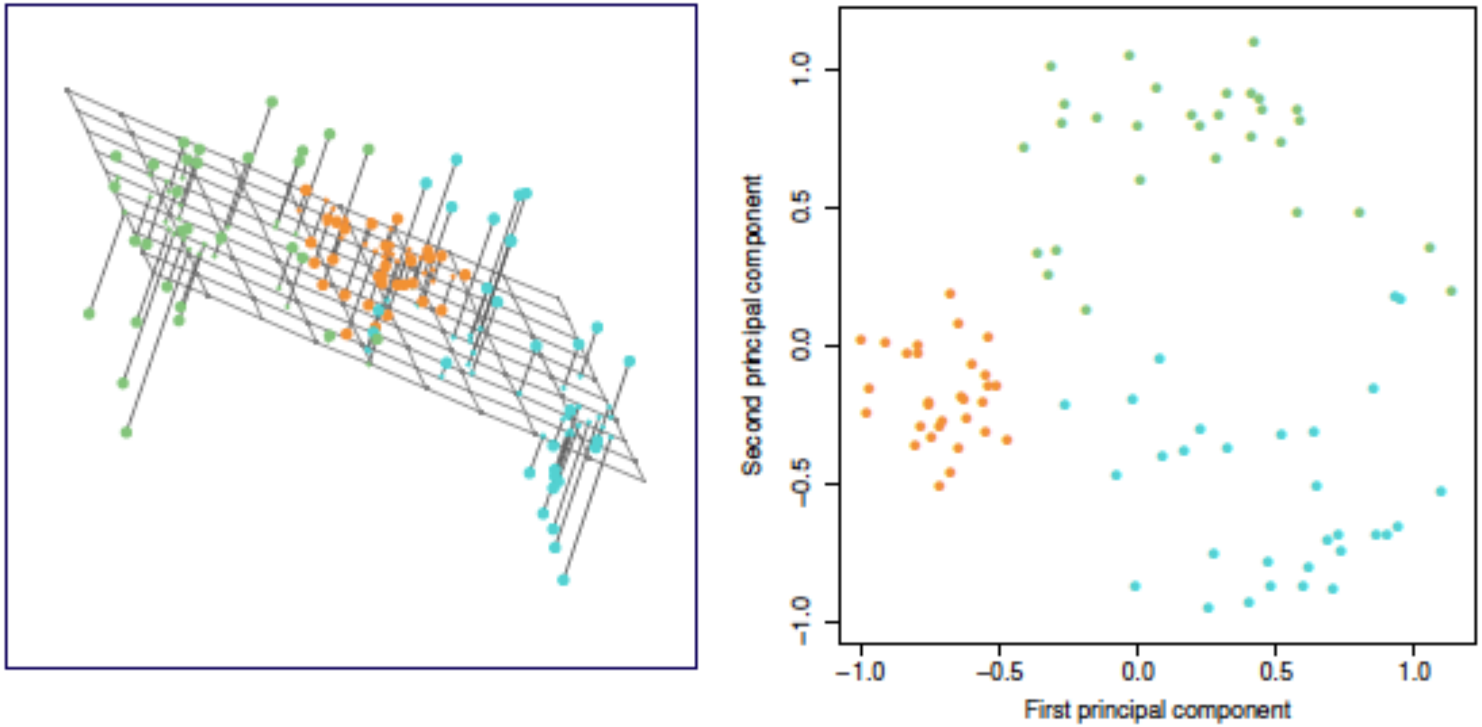
PCA - Loadings

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

PCA find the hyper-plane closest to the observations

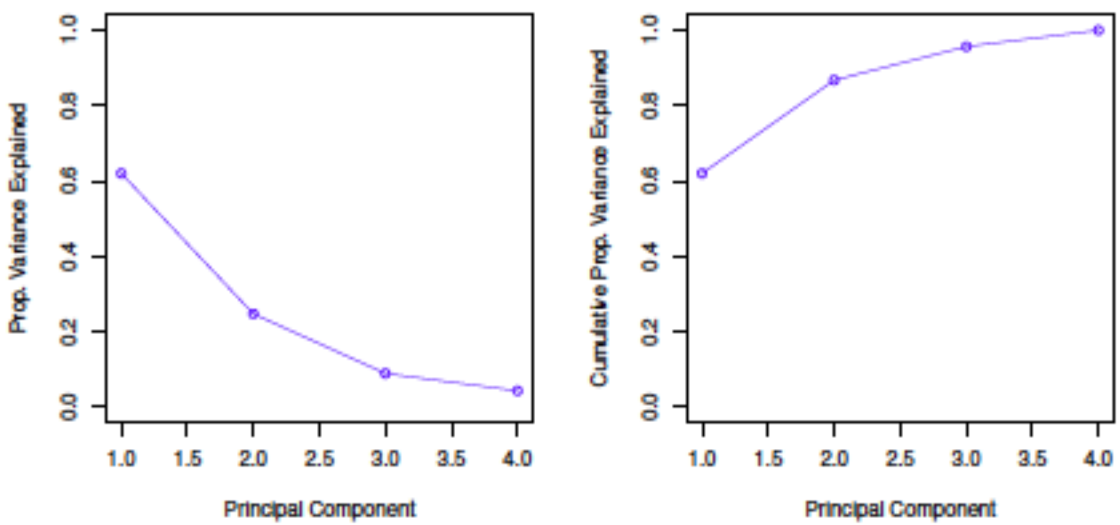
- The first principal component loading vector has a very special property: it defines the line in p-dimensional space that is closest to the n observations (using average squared Euclidian distance as a measure of closeness)
- The notion of principal components as the dimensions that are closest to the n observations extends beyond just the first component.
- For instance, the first two principal components of a data set space the plane that is closest to the n observations, in terms of average squared Euclidean distance.
  - Isn't it the same as regression lines?
    - No! Can you tell me why?

Another interpretation of Principal Components



Proportion Variance Explained

- To understand the strength of each component, we are interested in knowing the proportion of variance explained (PVE) by each one.



How many principal components should we use?

- If we use principal components as a summary of our data, how many components are sufficient?
  - 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” on the previous slide can be used as a guide: we look for an “elbow”.

Using principal components

- Visualizing Data
- We can adopt many statistical techniques, such as regression, classification, and classification to using n x M matrix whose columns are the first M << p principal components.
  - This can lead to less noisy results, since it is often the case that the signal (as opposed to the noise) in a data set is concentrated in its first few principal components.

Summary

- We learned how PCAs are being computed
- We learned about graphical representations of PCAs.
- We conceptually explored Geometrical properties of Principal Components.
- Proportion of Variance explained.