

Chapter 10: Unsupervised Learning

Thiago G. Martins | NTNU & Yahoo
Spring 2022

Introduction

Supervised vs. Unsupervised learning

- Supervised Learning definition
 - n observations.
 - Each containing features X_1, X_2, \dots, X_p and responses Y .
 - Regression and classification are widely known examples.
- Unsupervised Learning definition
 - n observations.
 - Each containing features X_1, X_2, \dots, X_p .
 - Objective: Discover interesting properties about the data.
 - Better data visualization
 - Reduce computational complexity
 - Discover groups among data points

Usefulness of Unsupervised Learning (Examples)

- Cancer research: Look for subgroups within the patients or within the genes in order to better understand the disease
- Online shopping site: Identify groups of shoppers as well as groups of items within each of those shoppers groups.
- Search engine: Search only a subset of the documents in order to find the best one for retrieval.

+++

General Challenges of Unsupervised Learning

- In general, unsupervised learning methods are
 - more subjective
 - hard to assess results
- There is usually no obvious ground-truth to compare to
- Remedy:
 - Unsupervised methods are usually part of a bigger goal
 - Evaluate them as how they contribute to such bigger goal
- Examples:
 - How clustering shoppers improved your recommendation algorithm?
 - How clustering documents reduced computational complexity and what was the cost involved?

Unsupervised Learning techniques

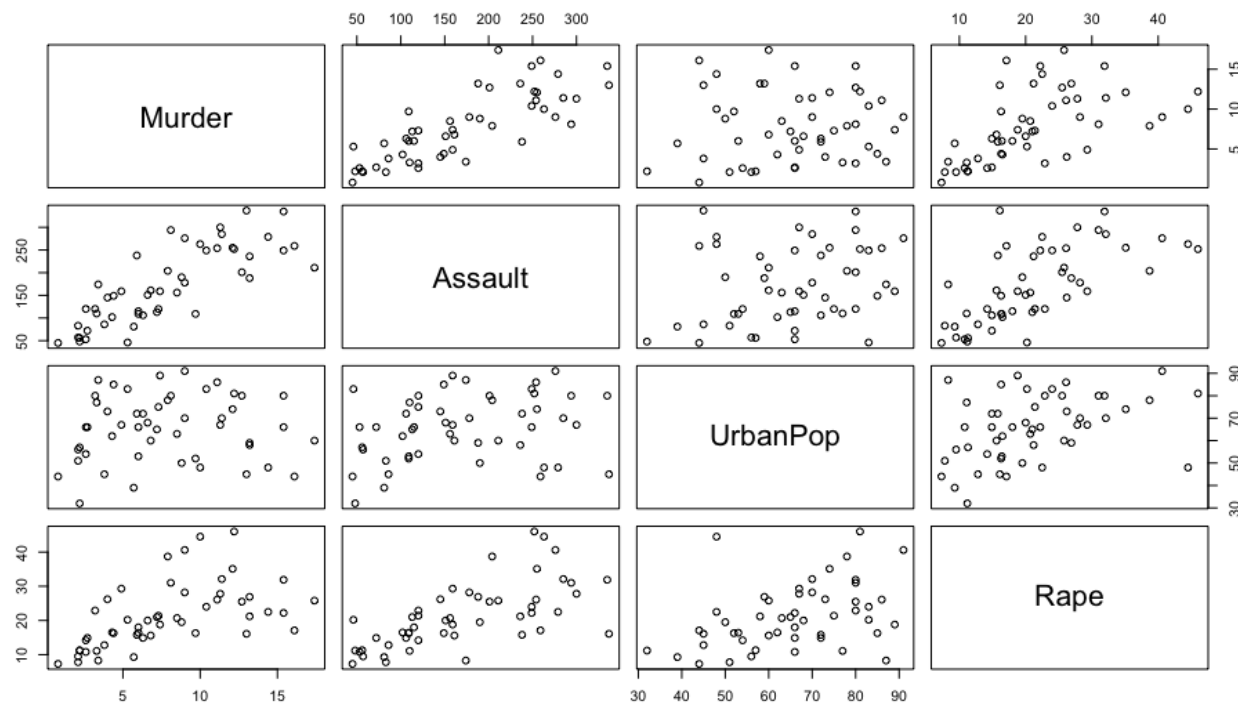
Covered in this module:

- PCA (Principal Component Analysis)
 - Data Visualization
 - Data pre-processing
- Clustering
 - Discovering unknown subgroups in the data
 - k-means clustering
 - Hierarchical clustering

Data Visualization

Data Visualization

- We want to visualize n observations with p features
- Two-dimensional scatterplots of data



Data Visualization

- Two-dimensional scatterplots of data
 - $p(p - 1)/2$ such scatterplots
 - each contain small fraction of the total information present in the dataset
- We want to find low dimensional representation of the data that captures most of the info as possible
 - Perfect scenario: 2 or 3 dimensions.
- **PCA: finds low dimension that captures most of the variability of the data**

Principal Component Analysis (PCA)

Principal Component Analysis (PCA)

- Discussed before in the context of Principal Components Regression
 - Turn large set of correlated variables into smaller set of orthogonal ones.
- This module focuses on PCA as a tool for data exploration

PCA - Recap

Principal Component Analysis (PCA)

- We want to create a $n \times M$ matrix Z , with $M < p$.
- The column Z_m of Z is the m -th principal component.

$$Z_m = \sum_{j=1}^p \phi_{jm} X_j \quad \text{subject to} \quad \sum_{j=1}^p \phi_{jm}^2 = 1$$

- We want Z_1 to have the highest possible variance.
 - That is, take the direction of the data where the observations vary the most.
 - Without the constrain we could get higher variance by increasing ϕ_j

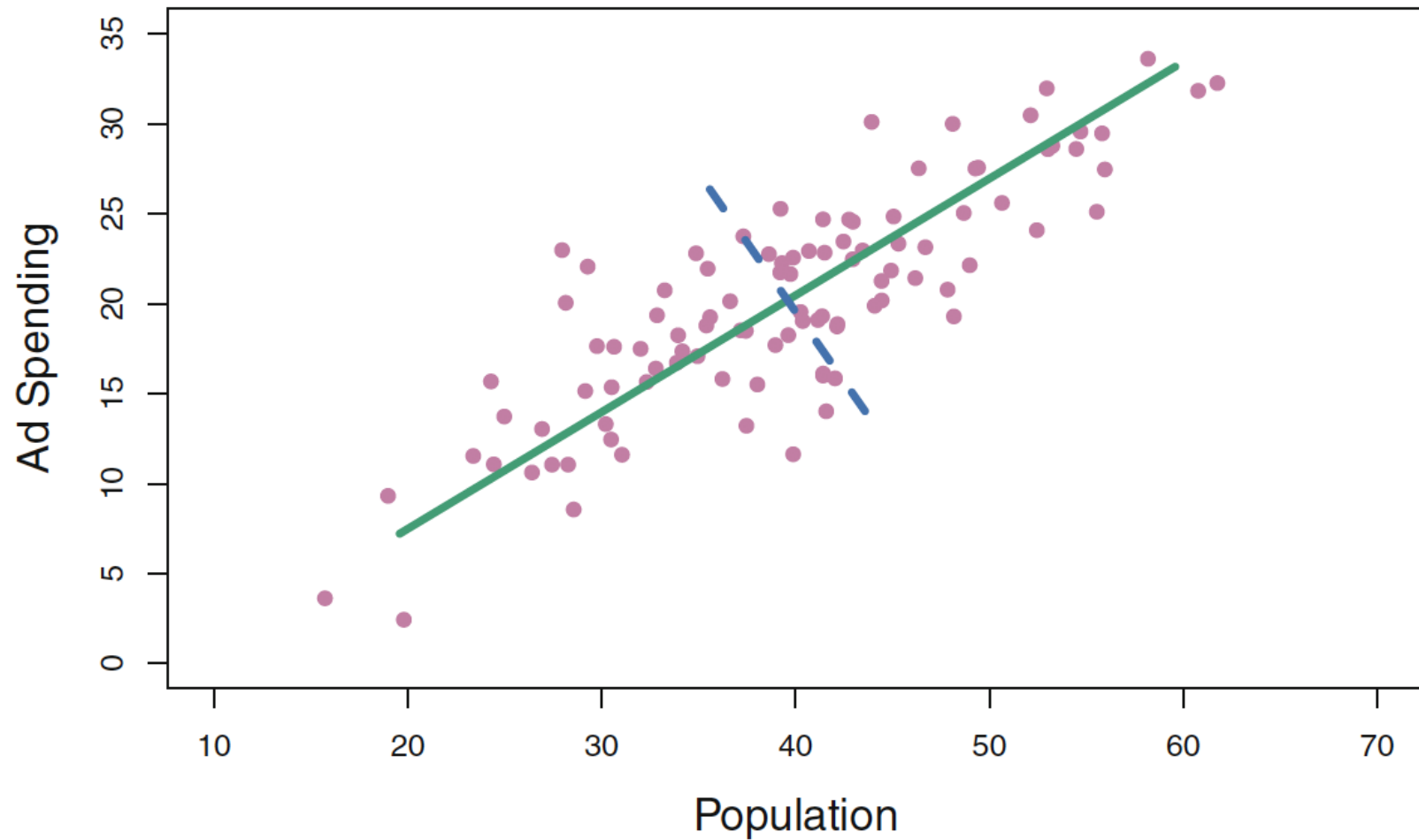
Principal Component Analysis (PCA)

- Z_2 should be uncorrelated to Z_1 , and have the highest variance, subject to this constrain.
 - The direction of Z_1 must be perpendicular (or orthogonal) to the direction of Z_2
- And so on ...
- We can construct up to p PCs that way.
 - In which case we have captured all the variability contained in the data
 - We have created a set of orthogonal predictors
 - But have **not** accomplished dimensionality reduction

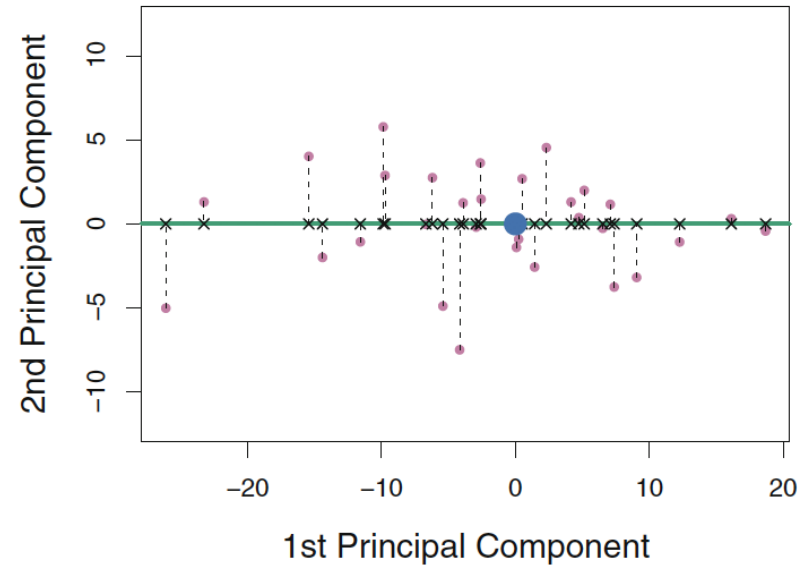
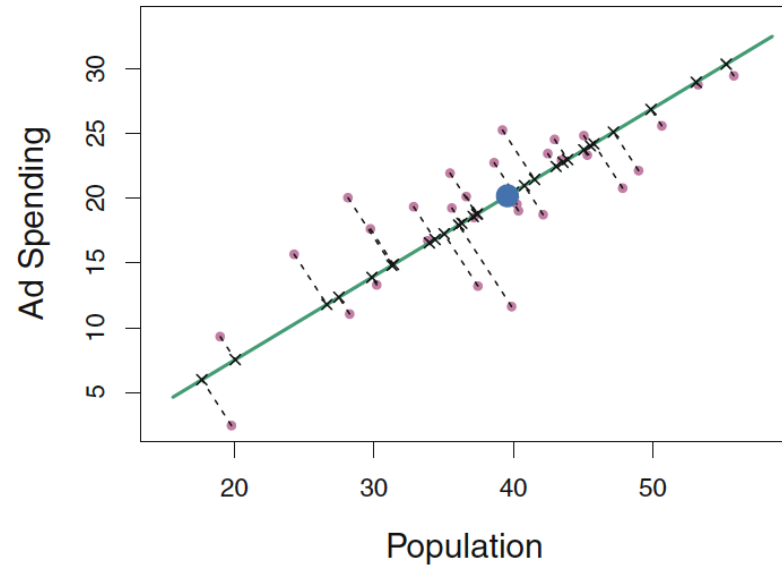
PCA Example: Interpretations

- M-dimension that capture most of the variability contained in the data
- M-dimension that is closest to the data points (average squared euclidean distances)

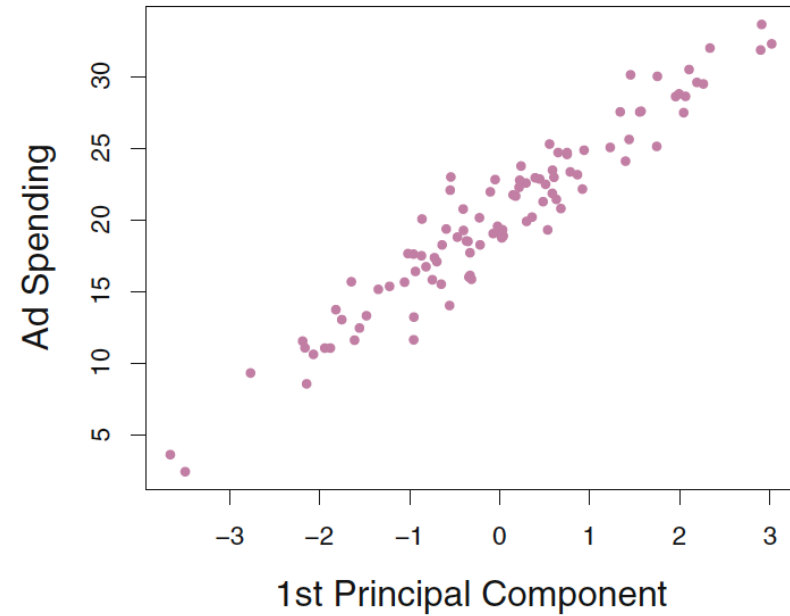
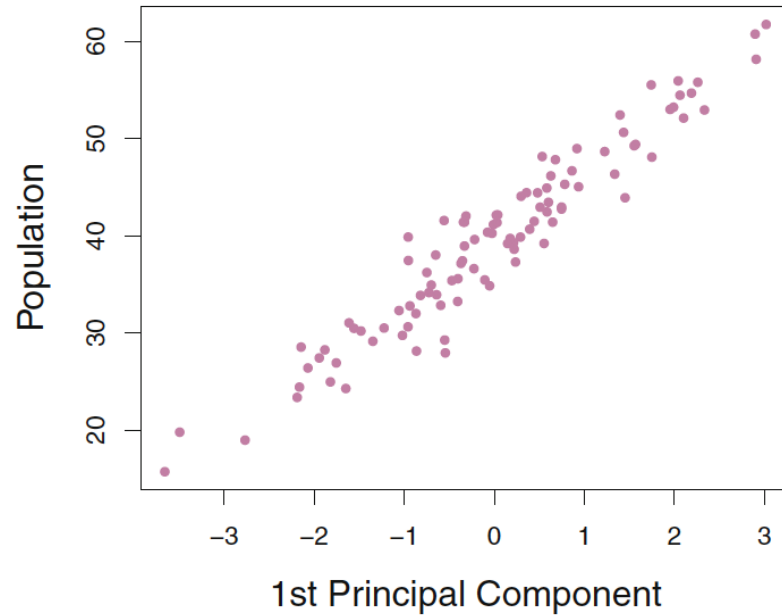
PCA Example - Ad spending



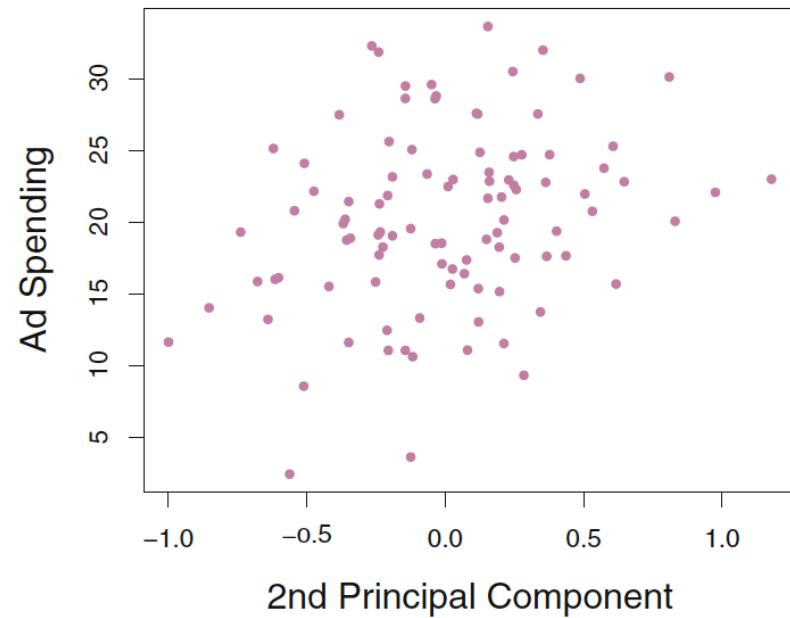
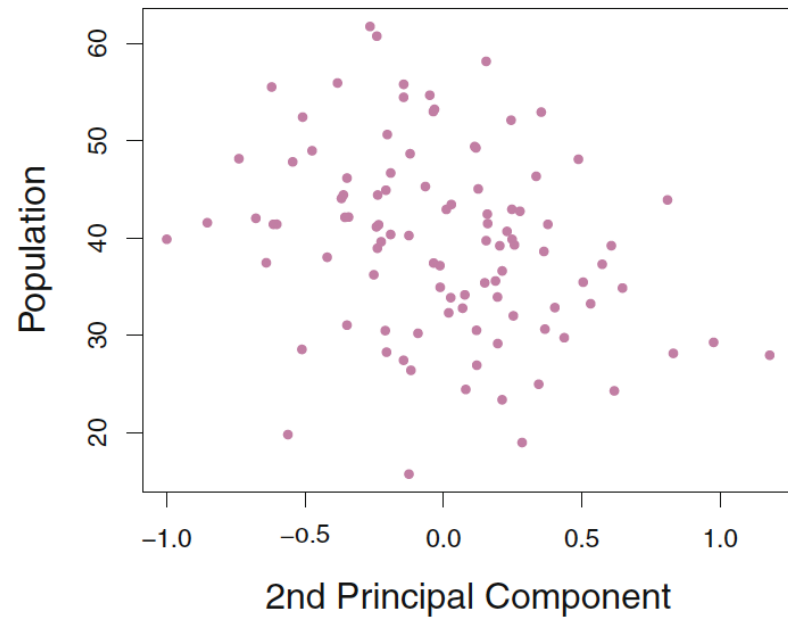
PCA Example - Ad spending (II)



PCA Example - Ad spending (III)



PCA Example - Ad spending (IV)



PCA - General setup

- Let \mathbf{X} be a matrix with dimension $n \times p$.
- Each column represent a vector of predictors.
- Assume $\mathbf{\Sigma}$ to be the covariance matrix associated with \mathbf{X} .
- Since $\mathbf{\Sigma}$ is a non-negative definite matrix, it has an eigen-decomposition

$$\mathbf{\Sigma} = \mathbf{C}\mathbf{\Lambda}\mathbf{C}^{-1}$$

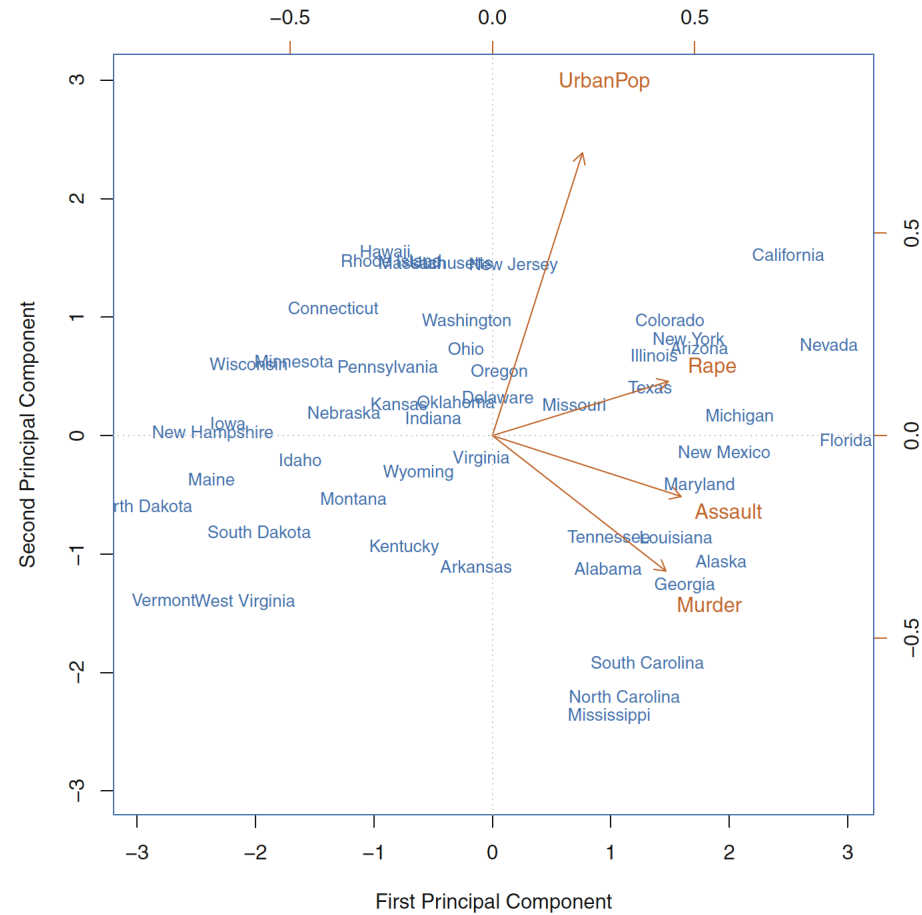
- $\mathbf{\Lambda} = \text{diag}(\lambda_1, \dots, \lambda_p)$ is a diagonal matrix of (non-negative) eigenvalues in decreasing order,
- \mathbf{C} is a matrix where its columns are formed by the eigenvectors of $\mathbf{\Sigma}$.

PCA - General setup (II)

- We want $\mathbf{Z}_1 = \boldsymbol{\phi}_1^T \mathbf{X}$, subject to $\|\boldsymbol{\phi}_1\|_2 = 1$
- We want \mathbf{Z}_1 to have the highest possible variance, $V(\mathbf{Z}_1) = \boldsymbol{\phi}_1^T \boldsymbol{\Sigma} \boldsymbol{\phi}_1$
- $\boldsymbol{\phi}_1$ equals the column eigenvector corresponding with the largest eigenvalue of $\boldsymbol{\Sigma}$
- The fraction of the original variance kept by the M principal component

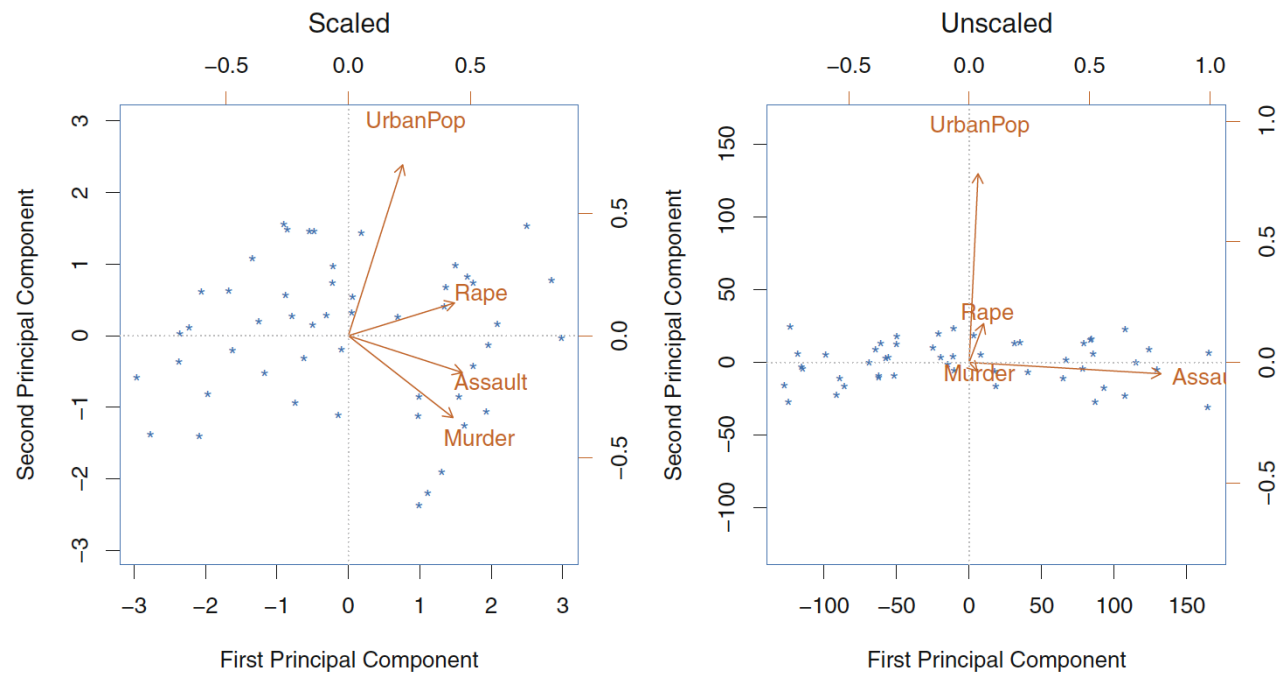
$$R^2 = \frac{\sum_{i=1}^M \lambda_i}{\sum_{j=1}^p \lambda_j}$$

Visualizing PC and loading



Scaling the variables

- Not all methodology needs scaling, e.g. linear regression
- PCA **usually** does



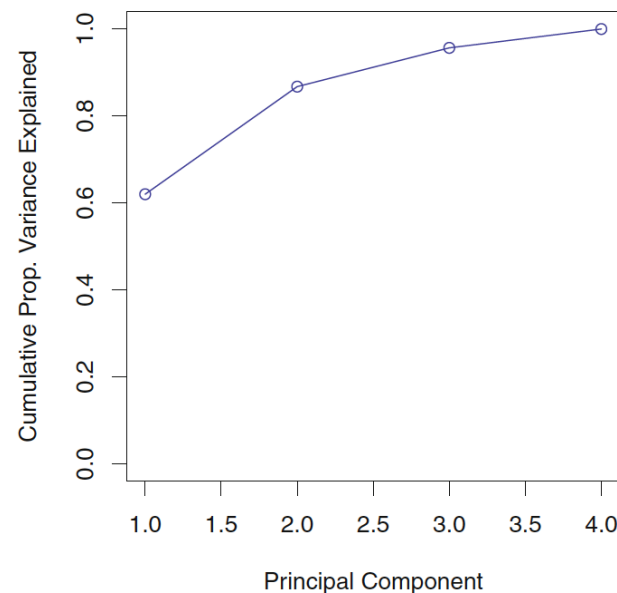
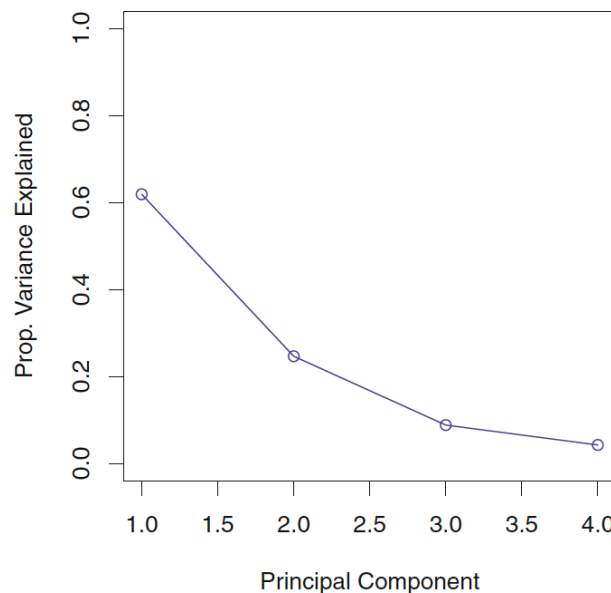
Uniqueness of PCs

- Each Principal Component loading vector is unique, up to a sign flip.
- Flipping the sign has no effect as the direction of the PC does not change.
- The approximation below will not change because the score vector sign will compensate the flip on the loading vector

$$x_{ij} \approx \sum_{m=1}^M z_{im} \phi_{jm}$$

Deciding how many PCs to use

- There is no objective answer
- Adhoc, by looking at the PVE graph



- Cast the selection based on the usage of the PCs in a supervised learning setting of interest (bigger goal)

PCA - Examples

- Lab 1: Principal component analysis applied to the `USArrests` dataset.
- Extra: PCA on the New York Times stories

Recommended Exercise 1

- For the New York Times stories dataset:
 - Create a biplot and explain the type of information that you can extract from the plot.
 - Create plots for the PVE and Cumulative PVE. Describe what type of information you can extract from the plots.

The `pca-examples.rdata` can be downloaded from the Blackboard.