Principal Component Analysis and Clustering

PSC 8185: Machine Learning for Social Science

Iris Malone

April 4, 2022

Announcements

- Problem Set 6 Released: Due April 18
- Problem Set 7 Released April 17 (Due April 27)
- Next Week: Jupyter Notebook (Python)

Recap

Where We've Been:

- Supervised learning used for prediction problems
- Parametric and non-parametric approaches navigate bias-variance tradeoff
- Deep learning often performs better than traditional ML algorithms

Recap

Where We've Been:

- Supervised learning used for prediction problems
- Parametric and non-parametric approaches navigate bias-variance tradeoff
- Deep learning often performs better than traditional ML algorithms

New Terminology:

- Stochastic Gradient Descent
- Neurons
- Backpropagation

Agenda

1. Unsupervised Learning
2. PCA
3. Clustering

Unsupervised Learning

Supervised vs Unsupervised Learning

1. Supervised Learning
    1.1 Regression and Classification
    1.2 Random Forests, Boosting, Bagging
    1.3 SVM
    1.4 Deep Learning and Neural Nets
2. Unsupervised Learning
    2.1 Principal Component Analysis and Clustering (This Week)
    2.2 Sentiment Analysis (April 18)
    2.3 Topic Modeling (April 25)

Supervised vs Unsupervised Learning

Supervised Learning:

- Predicts a given outcome
- Data includes x and y
- Evaluate accuracy

Unsupervised Learning:

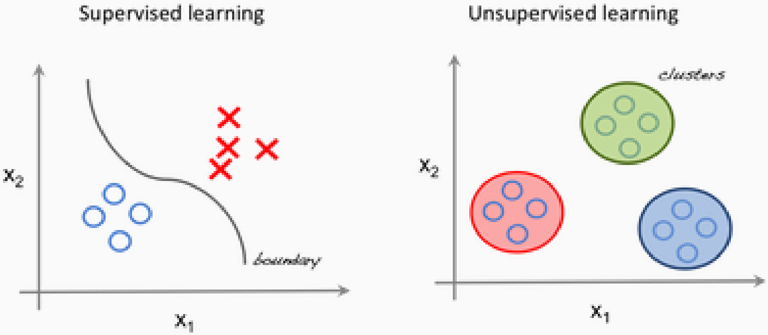
- Descriptive data analysis
- Data includes x
- No standard evaluation

Supervised vs Unsupervised Learning

Main Idea: Supervised learning includes information about an outcome of interest; unsupervised learning does not.

Supervised vs Unsupervised Learning

Main Idea: Supervised learning includes information about an outcome of interest; unsupervised learning does not.



Unsupervised Learning

- **Main Idea:** Describe relationship between observations in a data matrix
- Common Objectives:
  - Simplify high-dimensional data to explain variation in as few dimensions as possible
  - Identify meaningful groupings of the data

Real-World Applications:

- Describe sociological trends across groups
- Anomaly Detection
- Hand-Writing Analysis
- Defining ‘Nationalism’ or ‘State Capacity’

Types of Unsupervised Learning

- Principal Component Analysis (PCA)
- Clustering
    - K-Means Clustering
    - Hierarchical Clustering
- Sentiment Analysis
- Topic Modeling

PCA

Recall: High-Dimensional Data

- $p >> n$ is pretty common due to experimental advances and cheaper computers
- Need method to summarize high-dimensional data

Principal Component Analysis

- This is the most popular unsupervised procedure ever
- First theorized by Karl Pearson (1901)
- Developed by Harold Hotelling (1933)
- Provides a way to visualize and summarize information about high-dimensional data

Principal Component Analysis

Main Idea: Define a small set of M dimensions which summarize the information in all p predictors.

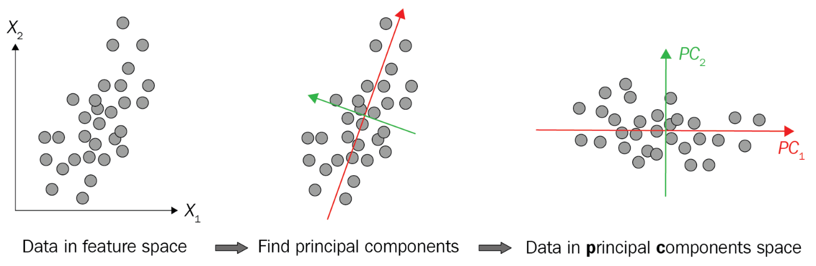


Figure 1: Example Dimensionality Reduction

Principal Components

- Each of the n observations live in p-dimensional space, but not all dimensions equally interesting.
- PCA seeks to find the most interesting dimensions, meaning the dimensions with the largest amount of variation among observations

The Most Interesting Man in the World

Principal Components Procedure

Procedure:

- Pre-process the data
- Identify similarities between groups of predictors $X_1, X_2, \ldots, X_p$
- Transform groups of predictors into M linear combination known as **principal component** Z

$$Z_m = \sum_{i=1}^{p} \phi_{jm} X_j$$

Pre-Processing the Data: Centering and Scaling

- Centering is key to ensure dimensions look at variance and not mean of predictors

Pre-Processing the Data: Centering and Scaling

- Centering is key to ensure dimensions look at variance and not mean of predictors
- Scale the variance to have mean zero and look for the linear combination with the largest sample variance

Pre-Processing the Data: Centering and Scaling

- Centering is key to ensure dimensions look at variance and not mean of predictors
- Scale the variance to have mean zero and look for the linear combination with the largest sample variance
- Scaling is key to good interpretation
- Unscaled data means the PCA loading vector will have a very large loading for the variable with the highest variance

Transform Groups Predictors

Procedure:

- Pre-process the data
- Identify similarities between groups of predictors $X_1, X_2, \ldots, X_p$
- Transform groups of predictors into M linear combination known as **principal component** Z

$$Z_m = \sum_{i=1}^{p} \phi_{jm} X_j$$

Principal Component Characteristics

- Solution to an optimization problem where the first two principal components span a plane which is closest to the data
- First and second principal components must be orthogonal (i.e., they don't explain the same types of variation)

PCA Loss Function

Let X be the data matrix with n samples and p variables. From each variable, we subtract the mean of the column (i.e. we center the variables).

PCA Loss Function

Let X be the data matrix with n samples and p variables. From each variable, we subtract the mean of the column (i.e. we center the variables).

To find the first principal component we minimize variance of the n samples projected onto ϕ_1:

$$max_{\phi_{11}\ldots\phi_{p1}} \frac{1}{n}\sum_{i=1}^{n}(\sum_{j=1}^{p}\phi_{j1}x_{ij})^2 \qquad \text{subject to} \sum_{j=1}^{p}\phi_{j1}^2 = 1$$

PCA Loss Function

Let X be the data matrix with n samples and p variables. From each variable, we subtract the mean of the column (i.e. we center the variables).

To find the first principal component we minimize variance of the n samples projected onto ϕ_1:

$$max_{\phi_{11} \ldots \phi_{p1}} \frac{1}{n} \sum_{i=1}^{n} (\sum_{j=1}^{p} \phi_{j1} x_{ij})^2 \qquad \text{subject to} \sum_{j=1}^{p} \phi_{j1}^2 = 1$$

Projection of the i^{th} sample onto ϕ_1 is the score Z_{i1}

Finding the second principal component

Let X be the data matrix with n samples and p variables. From each variable, we subtract the mean of the column (i.e. we center the variables).

To find the second principal component we solve:

$$max_{\phi_{11} \dots \phi_{p2}} \frac{1}{n} \sum_{i=1}^{n} (\sum_{j=1}^{p} \phi_{j2} x_{ij})^2 \qquad \text{subject to} \sum_{j=1}^{p} \phi_{j2}^2 = 1 \& \sum_{j=1}^{p} \phi_{j1} \phi_{j2} = 1$$

Loadings

- **Loadings**: Aspects of the vector which passes the closest to a cloud of observations in terms of squared Euclidean distance

Loadings

- **Loadings**: Aspects of the vector which passes the closest to a cloud of observations in terms of squared Euclidean distance
- The loading make up an element of the principal component loading vector ($\phi_1 = (\phi_{11}, \phi_{21}, \ldots)$)

$$Z_m = \sum_{i=1}^{p} \phi_{jm} X_j$$

Loadings

- **Loadings**: Aspects of the vector which passes the closest to a cloud of observations in terms of squared Euclidean distance
- The loading make up an element of the principal component loading vector ($\phi_1 = (\phi_{11}, \phi_{21}, \ldots)$)

$$Z_m = \sum_{i=1}^{p} \phi_{jm} X_j$$

- Loadings:
  - Size describes how much a variable contributes to a particular principal component
  - Sign explains correlation between elements

What is the first principal component?

The first principal component (the most interesting dimension) is the vector which passes the closest to a cloud of samples in terms of squared **Euclidean distance**.

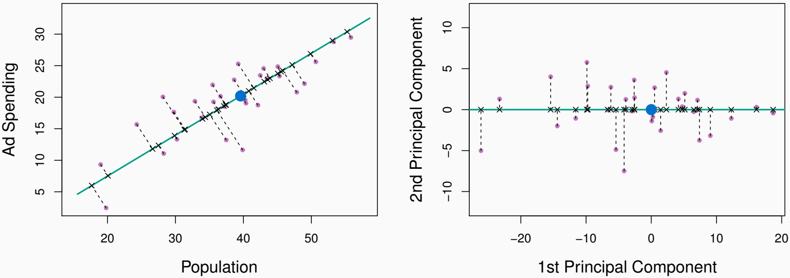
$$d\left(x_{i'j}, x_{ij}\right) = \sqrt{\sum_{i=1}^{n} \left(x_{ij} - x_{i'j}\right)^2}$$

Interpretation:

- We expect some observations to be 'closer' (more similar) to each other
- When distance is small, observation pairs are more similar
- When distance is larger, observation pairs are more dissimilar

Example Euclidean Distance

Vector which passes the closest to a cloud of samples in terms of squared **Euclidean distance**, i.e. the green line minimizing the average squared length of the dotted lines



What does this look like with 3 variables?

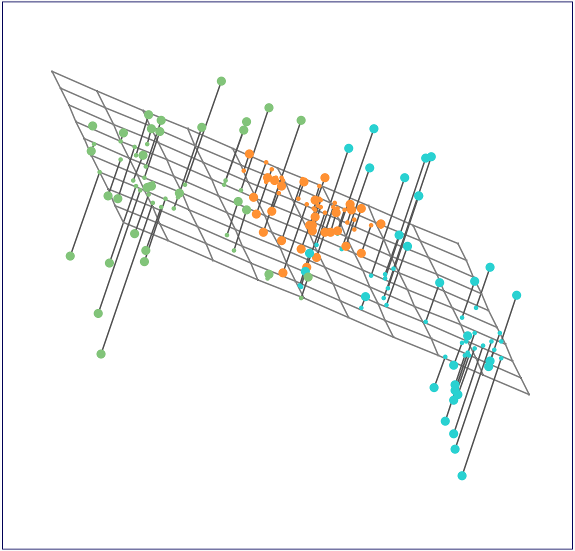


Figure 2: PC spans the plane that best fits the data (like SVM hyperplane)

What is PCA Good For?

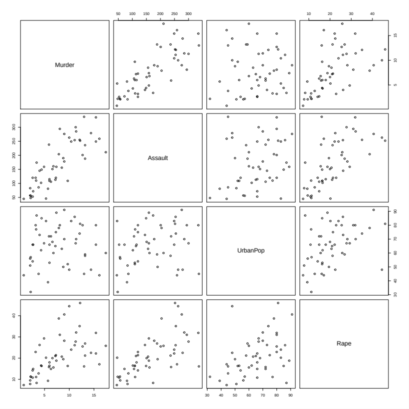
Example: US Arrests data contains info on 3 crime statistics (assault, murder, rape) and population ($p = 3$) for 50 states ($n = 50$).

Potential Research Questions:

- Do crimes correlate with each other?
- Do states with larger urban populations see more crime?

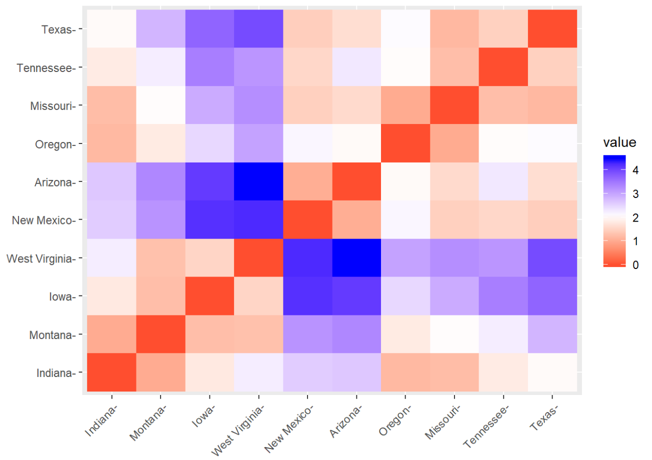
What is PCA Good For?

US Arrests data contains info on 3 crime statistics (assault, murder, rape) and population ($p = 3$) for 50 states ($n = 50$).



Example Euclidean Distance

- Compare how similar states are based on crime statistics
- When distance is small, observation pairs are more similar (red)
- When distance is larger, observation pairs are more dissimilar (blue)



Interp: PC pairs states with similar crime rates

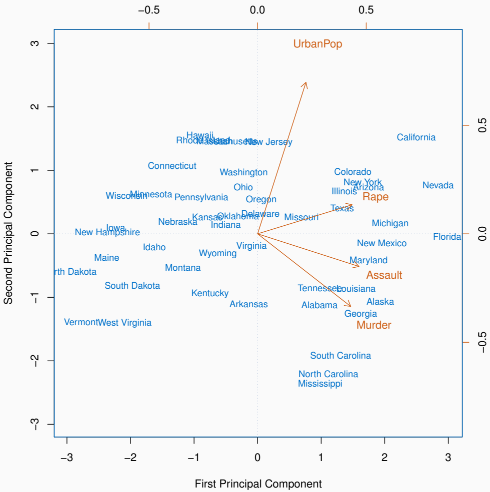
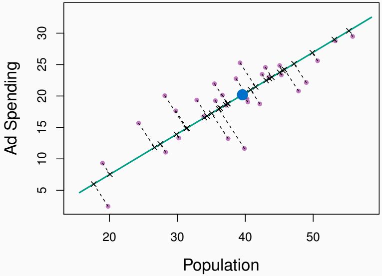
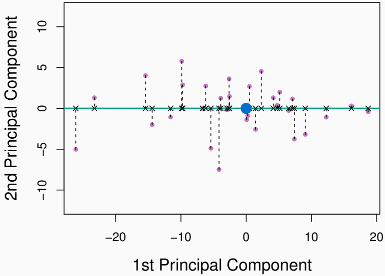


Figure 3: A **biplot** of the first two principal component scores and loading vectors.

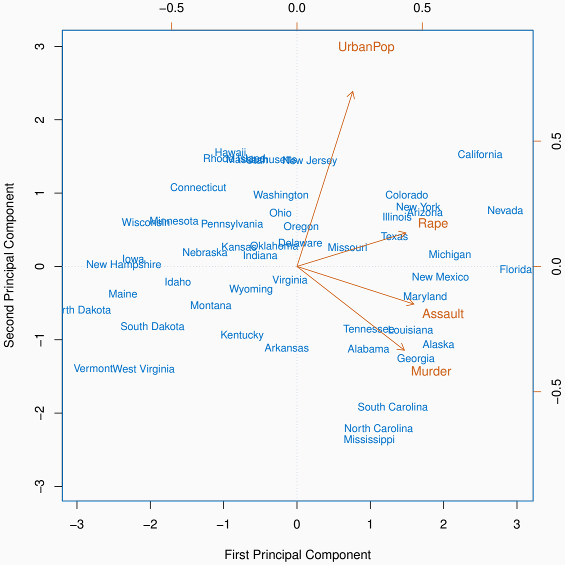
A second interpretation

Another way to explain the first principal component is that it is the dimension with the highest variance between variables



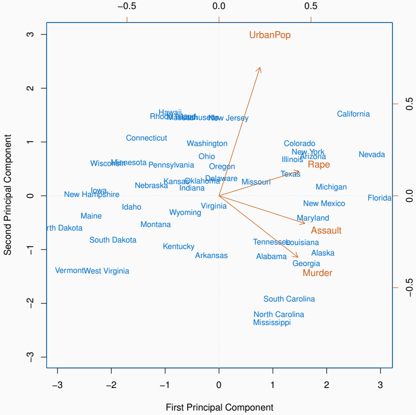


Interp: PC explains variation in crime levels



Example: USArrests Interpretation

- States with high levels of rape also have high levels of assault and murder
- Urban population is orthogonal (unrelated) to crime rates
- Different types of states have different crime rates



How many principal components are enough?

Rule of Thumb: 2 Principal Components capture most of the relevant information.

How many principal components are enough?

Rule of Thumb: 2 Principal Components capture most of the relevant information.

More Precise Answer:

- **Proportion of Variance Explained** (PVE): tells us sum of the variance explained by the m-th principal component over the total variance
- Can assess how much variation in data: low PVE = noisy data; high PVE = highly separable data

Proportion of Variance Explained

- First principal component explains the direction in space in which the data vary the most

Proportion of Variance Explained

- First principal component explains the direction in space in which the data vary the most
- Second principal component explains the direction in space in which the data vary the second most, etc.

Proportion of Variance Explained

- First principal component explains the direction in space in which the data vary the most
- Second principal component explains the direction in space in which the data vary the second most, etc.
- Total variance of the score vectors is the same as the total variacne of the original variables:

$$\sum_{i=1}^{p} \frac{1}{n} \sum_{j=1}^{n} z_{ji}^2 = \sum_{k=1}^{p} Var(x_k)$$

Proportion of Variance Explained

The variance of the m^{th} score variable is:

$$\frac{1}{n}\sum_{i=1}^{n} z_{im}^2 = \frac{1}{n}\sum_{i=1}^{n}(\sum_{j=1}^{p}\phi_{jm}x_{ij})^2$$

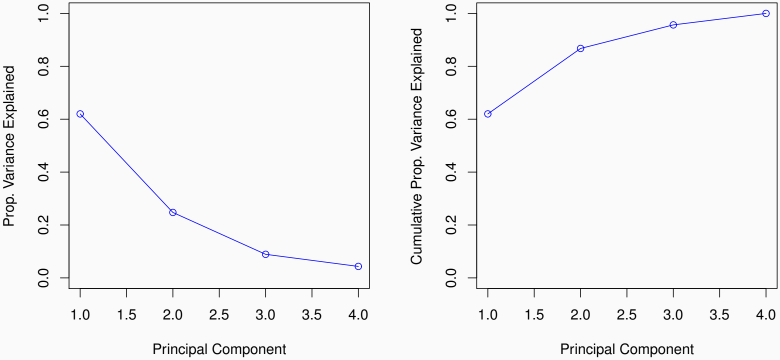
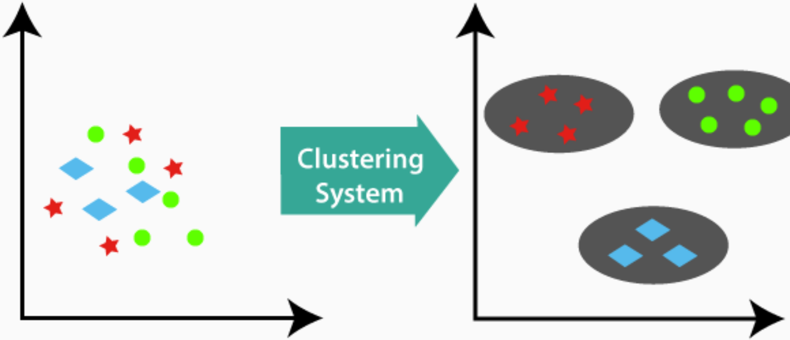


Figure 5: **Scree plot** showing proportion variance explained by each principal component

Clustering

Clustering

Main Idea: Partition the data into distinct groups based on intra-group similarities (or inter-group differences)



PCA vs Clustering

- **PCA:** Simplify multiple predictors into small number of principal components to explain variance
- **Clustering:** Find subgroups among observations based on individual or combination of predictors

Types of Clustering

1. **K-Means Clustering:** Partition observations into pre-set number of clusters

Types of Clustering

1. **K-Means Clustering:** Partition observations into pre-set number of clusters
2. **Hierarchical Clustering:** Partition observations, but with no pre-set number of clusters

K-Means Clustering

Main Idea: Partition observations into pre-set number of clusters

K-Means Clustering

Main Idea: Partition observations into pre-set number of clusters

- Goal:
  - Maximize the similarity of samples within each cluster
  - Minimize within-cluster variation
- Loss Function:

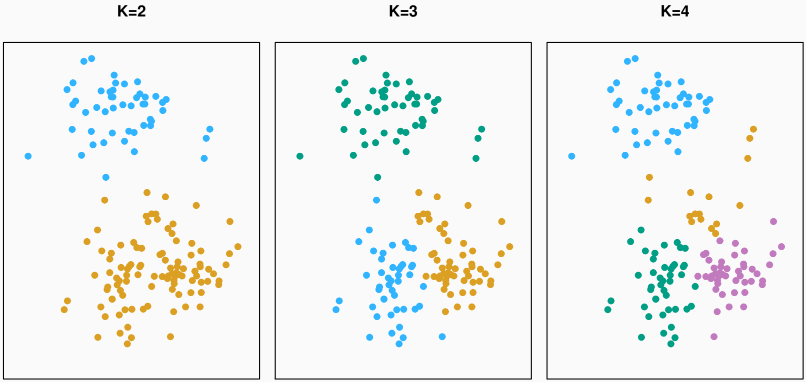
$$min_{C_1,\ldots,C_k} \sum_{l=1}^{K} W(C_l)$$

- $W(C_l)$ is measure of similarity between pairs of observations

$$W(C_l) = \frac{1}{|C_l|} \sum_{i,j \in C_l} \text{Distance}^2(x_i, x_j)$$

- K is the number of clusters and must be fixed in advance
- $\text{Distance}^2(x_i, x_j)$ is Euclidean distance between observations

K-Means Clustering



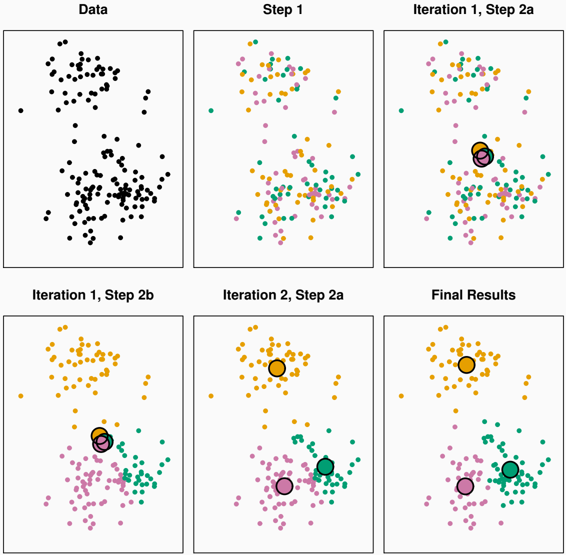
K-Means Procedure

- Assign each sample to a cluster from 1 to K arbitrarily, e.g. at random
- Assign obs. to closest centroid
    - Find the centroid of each cluster , i.e. the average $\bar{x}$ of all the samples in the cluster

$$x_{l,j} = \frac{1}{|C_l|} \sum_{i,j \in C_l} x_{i,j} \text{for} j = 1, \ldots, p$$

    - Reassign each sample to the nearest centroid
- Reposition centroids to new center
- Iterate until centroid position becomes static

K-Means Clustering



Limits to K-Means Procedure

- As iterations increase, clustering will improve until a local optimum has been reached

Limits to K-Means Procedure

- As iterations increase, clustering will improve until a local optimum has been reached
- **Problem:** Algorithm focuses on minimizing local differences rather than global ones (like CART greedy algorithm)

Limits to K-Means Procedure

- As iterations increase, clustering will improve until a local optimum has been reached
- **Problem:** Algorithm focuses on minimizing local differences rather than global ones (like CART greedy algorithm)
- Significance: Different initializations → different clusters

Example: K-Means Output with Different Initializations

Figure 6: Value above each plot is minimum; 3 different values, 1 common minimum

Example: K-Means Output with Different Initializations

Figure 7: In practice we start from many random initializations and choose the output which minimizes the loss function

Advantages and Disadvantages to K-Means Clustering

Advantages

Disadvantages

Advantages and Disadvantages to K-Means Clustering

Advantages

- Guaranteed convergence
- Good scalability for large-p multi-dimensional data
- If large p, then faster than hierarchical clustering

Disadvantages

Advantages and Disadvantages to K-Means Clustering

Advantages

- Guaranteed convergence
- Good scalability for large-p multi-dimensional data
- If large p, then faster than hierarchical clustering

Disadvantages

- Need to specify number of clusters
- Different initializations $\rightarrow$ different results

Hierarchical Clustering

Main Idea: Partition observations, but with no pre-set number of clusters

Hierarchical Clustering

Main Idea: Partition observations, but with no pre-set number of clusters

Types of Hierarchical Clustering:

- **Agglomerative:** Bottom-up approach; cluster starting from leaves and build up to trunk (most common method)
- **Divisive:** Top-down approach; cluster starting from root node and build down to leaves

Agglomerative Clustering

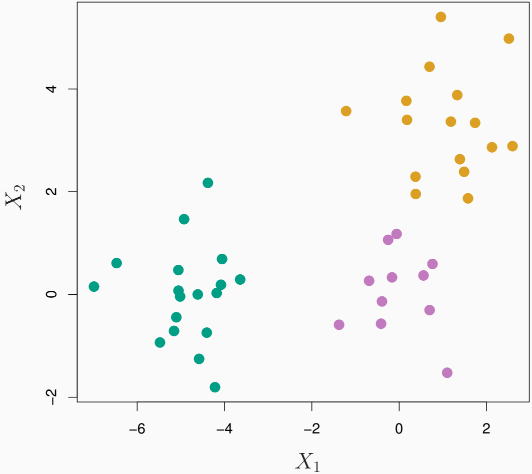
- **Intuition:** Look at Euclidean distance between observations and cluster similar observations (small distance)

Agglomerative Clustering

- **Intuition:** Look at Euclidean distance between observations and cluster similar observations (small distance)
- Find families of cluster based on links (distance lengths) between observations
- Iterate until all data organized and nested

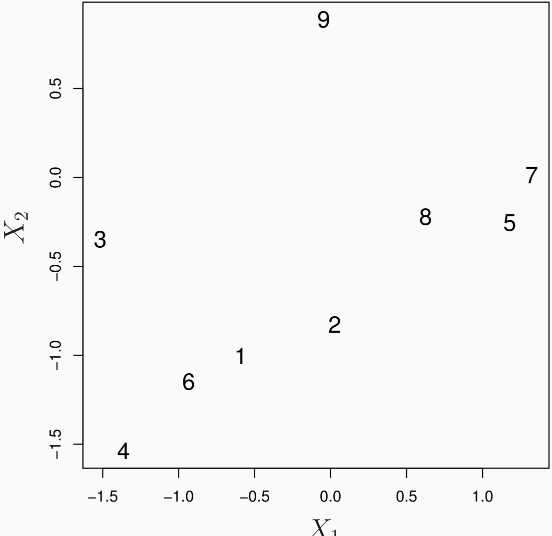
Agglomerative Clustering

Intuition: Look at Euclidean distance between observations and cluster similar observations



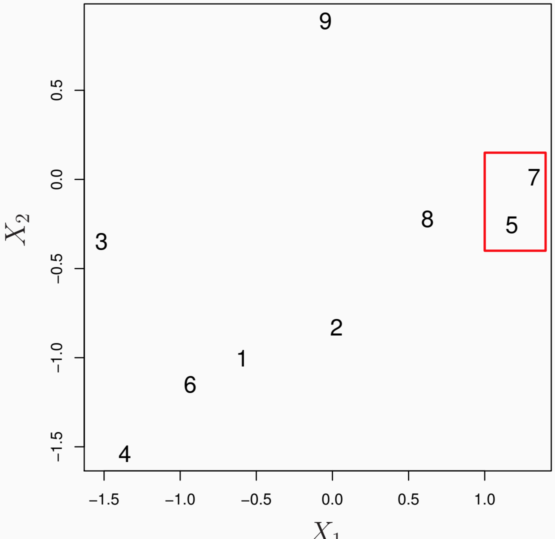
Agglomerative Clustering

Intuition: Look at Euclidean distance between observations and cluster similar observations



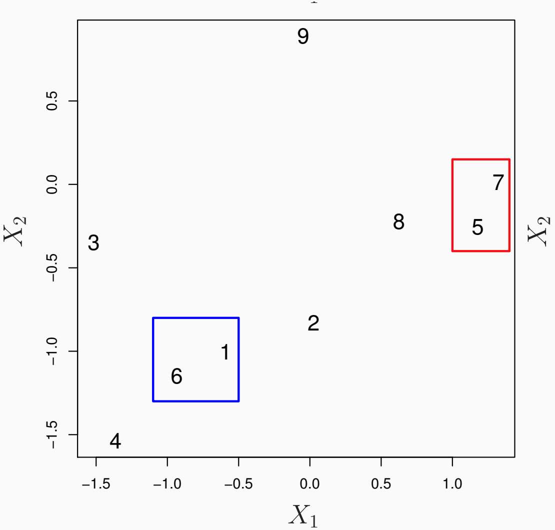
Agglomerative Clustering

Intuition: Look at Euclidean distance between observations and cluster similar observations



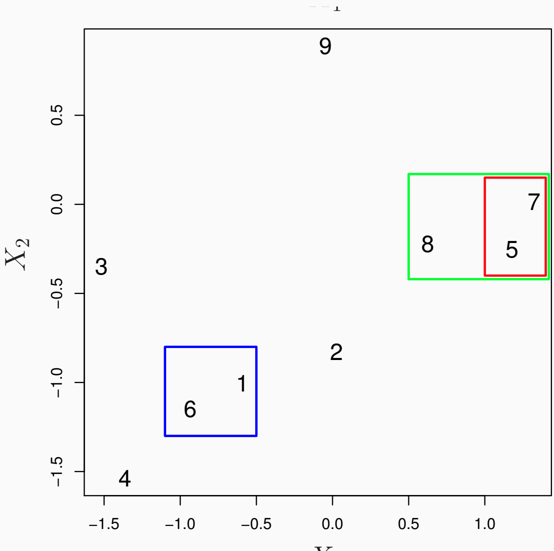
Agglomerative Clustering

Intuition: Look at Euclidean distance between observations and cluster similar observations



Agglomerative Clustering

Intuition: Look at Euclidean distance between observations and cluster most similar observations



Dendogram

We visualize the clusters using a **dendogram**.

Dendogram

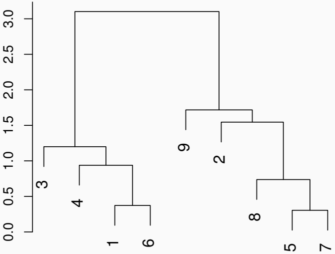
We visualize the clusters using a **dendogram**.

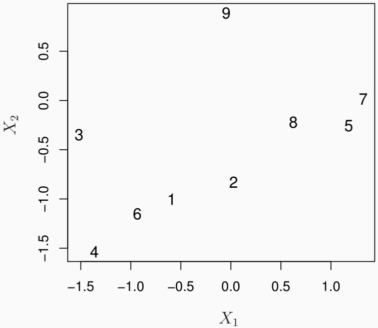
- Dendogram clusters data according to given linkage algorithm
- Clusters obtained by cutting dendogram at given height
- Cutting data exploits nested structure data

Dendogram

We visualize the clusters using a **dendogram**.

- Dendogram clusters data according to given linkage algorithm
- Clusters obtained by cutting dendogram at given height
- Cutting data exploits nested structure data





Reading a Dendogram

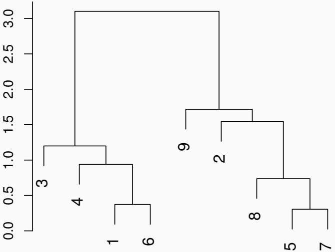
- Dendogram has **leaves** and **branches**
- Branches measure dissimilarity (larger Eucl. distance)
    - Y-axis measures degree of dissimilarity
    - The height of the branches shows how different observations are.
- Each leaf represents an observation.
    - Leaves fuse into branches around other observations that are similar to them
    - Re-ordering leaves on a branch doesn't affect their meaning

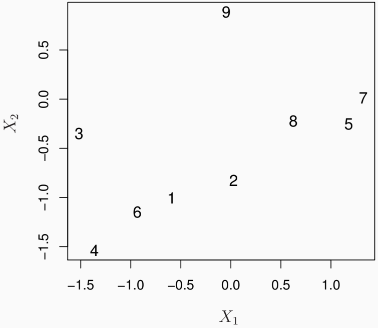
Reading a Dendogram: Similarities

- Read as moving from leaves → branches
- Fusions lower in the tree indicate greater similarity
- Fusions higher in the tree indicate less similarity
- Cannnot draw inferences about similarity based on horizontal proximity

Reading a Dendogram: Similarities

- Observations 5 and 7 are quite similar
- Observation 9 is *not* necessarily similar to observation 2
- Observation 9 is as similar to obs. 2 as it is to obs. 8, 5, 7



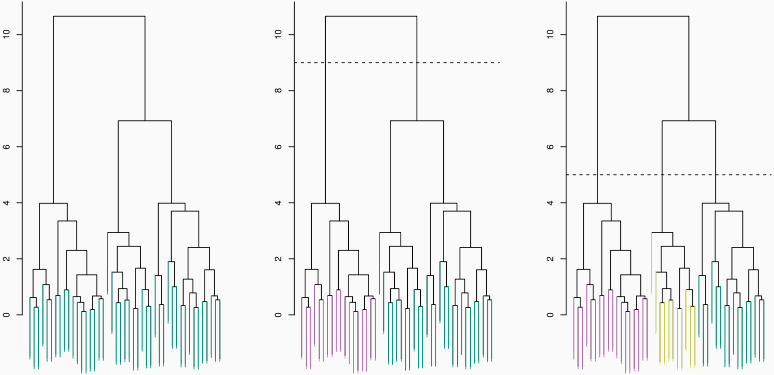


Reading a Dendogram: Number Clusters

- Identify clusters by making horizontal cut in data
- Distinct set observations below cut are each a **cluster**
- Different cut positions lead to differnt cuts

Reading a Dendogram: Number Clusters

- Far Left: No cut → one cluster
- Center: Cut at height 9 (dashed line) → two clusters
- Far Right: Cut at height 5 → 3 clusters



Hierarchical Clustering Procedure

- Begin with n observations and a measure (normally Euclidean distance) of all $\binom{n}{2}$ pairwise dissimilarities

Hierarchical Clustering Procedure

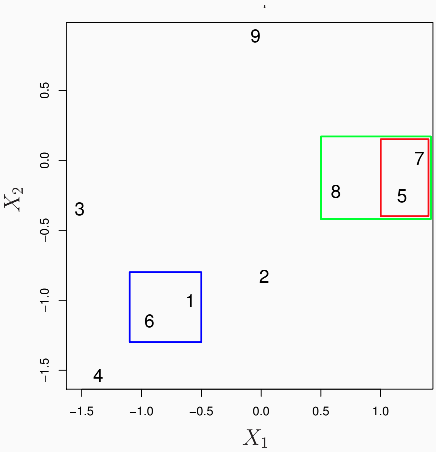
- Begin with n observations and a measure (normally Euclidean distance) of all $\binom{n}{2}$ pairwise dissimilarities
- For $i = 1, n, n-1, \ldots, 2:$
    - Examine all pairwise inter-cluster dissimilarities among the i clusters and identify pair of clusters that are least dissimilar (most similar)
    - Calculate height dendogram based on dissimilarity (less dissimilar lower on tree)
    - Fuse least dissimilar clusters together
    - Compute the new pairwise inter-cluster dissimilarities among the $i-1$ remaining clusters

Hierarchical Clustering Procedure

- Sometimes we have pair of *groups* of observations (instead of pair of observations),
- Describe dissimilarity between groups as **linkage**.
- Types of Linkages
  - Complete
  - Average
  - Single
  - Centroid

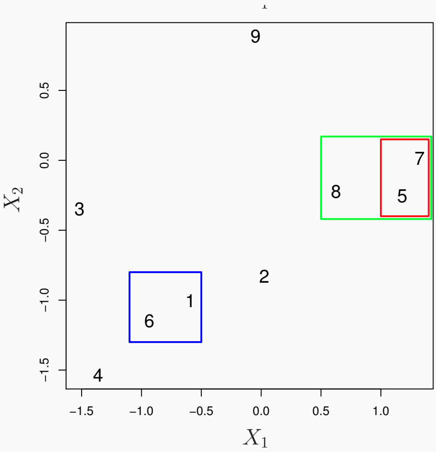
Hierarchical Clustering Algorithms

- **Complete Linkage**: Maximal inter-cluster dissimilarity. Record the largest of the dissimilarities.



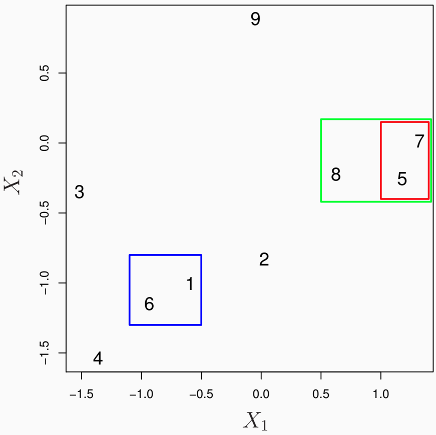
Hierarchical Clustering Algorithms

- **Single Linkage:** Minimal inter-cluster dissimilarity. Record the smallest of these dissimilarities.

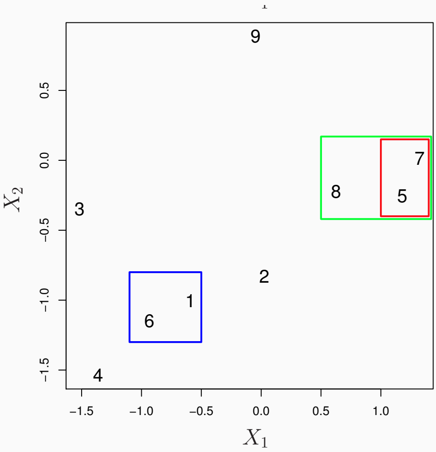


Hierarchical Clustering Algorithms

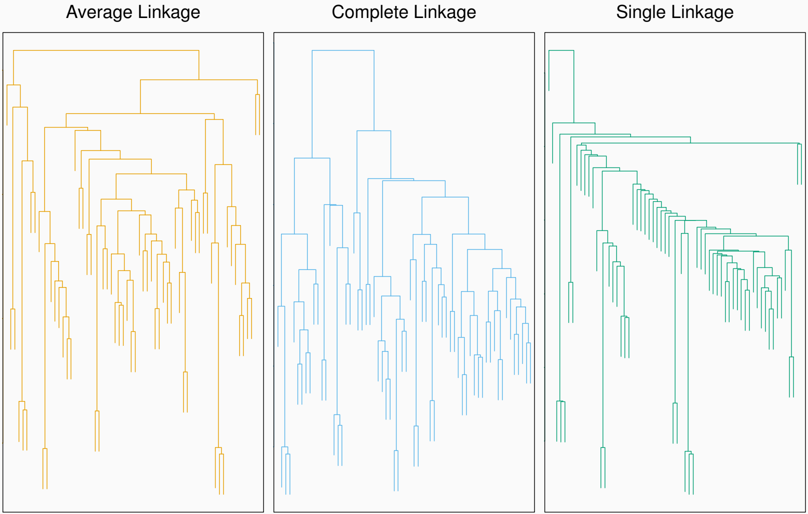
- **Average Linkage:** Mean inter cluster dissimilarity. Record the average of dissimilarities across all pairwise dissimilarities.



- **Centroid Linkage**: Dissimilarity between the centroid for cluster A and the centroid for cluster B.



Different Linkages



Correlation Distance

Rule of Thumb: Cluster using Euclidean Distance to measure similarities and dissimilarities

Alternative Approach: Correlation Distance

- Considers 2 observations similar if their features are highly correlated

Correlation Distance

Example: Suppose that we want to cluster customers at a store for market segmentation.

- Samples are customers
- Each variable corresponds to a specific product and measures the number of items bought by the customer during the years

Approaches:

- Euclidean distance would cluster all customers who purchase few things (quantity)
- But we might want to cluster customers who purchase similar things (category)
- Here, correlation distance may be more appropriate dissimilarity between samples

Correlation Distance

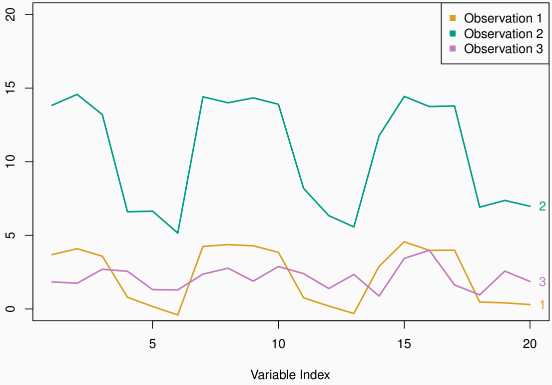


Figure 13: Observations 1 and 3 have similar values so there is small Euclidean distance, but large correlation; observations 1 and have different values so large Euclidean distance, but highly correlated

Advantages and Disadvantages to Hierarchical Clustering

Advantages

Disadvantages

Advantages and Disadvantages to Hierarchical Clustering

Advantages

- Attractive visualization of clusters
- Captures highly flexible relationships

Disadvantages

Advantages and Disadvantages to Hierarchical Clustering

Advantages

- Attractive visualization of clusters
- Captures highly flexible relationships

Disadvantages

- Not suitable for large data
- If large p, performs worse than k-means clustering
- Many different ways to cut the cluster → different results

Problems in Clustering

1. Is clustering appropriate? Could a sample belong to more than one cluster?
2. How many clusters are appropriate?
3. Are the clusters robust?

Conclusion

- Unsupervised learning effective tool for data analysis
- PCA good at exploring trends across multi-dimensional data
- Clustering good at identifying similarities between groups of observations
- Clustering effective, but subjective to different user inputs