Day 8: Unsupervised Learning and Dimensionality Reduction

ME314: Introduction to Data Science and Machine Learning

Jack Blumenau

20th July 2023

Day 8 Outline

Unsupervised Learning

Principal Components Analysis

Principal Component Regression

Clustering

Unsupervised Learning

Unsupervised Learning

Unsupervised vs Supervised Learning:

- Most of this course focuses on supervised learning methods such as regression and classification.
- In that setting we observe both a set of features $X_1, X_2, \ldots, X_p$ for each object, as well as a response or outcome variable Y. The goal is then to predict Y using $X_1, X_2, \ldots, X_p$.
- Here we instead focus on unsupervised learning, where we observe only the features $X_1, X_2, \ldots, X_p$. We are not interested in prediction, because we do not have an associated response variable Y.

The Challenge and Opportunity of Unsupervised Learning

Challenge:

- Unsupervised learning is more subjective than supervised learning, as there is no simple goal for the analysis, such as prediction of a response.
- The goal is more amorphous. We want to discover interesting things about our data: is there an informative way to visualize the data? Can we discover subgroups among the variables or among the observations?

The Challenge and Opportunity of Unsupervised Learning

Opportunity:

- It is often easier to obtain unlabeled data – from a lab instrument or a computer – than labeled data, which can require human intervention.
- Techniques for unsupervised learning are of growing importance in a number of fields:
  - cancer patients grouped by their gene measurements
  - shoppers characterized by their browsing and purchase histories
  - voters clustered by their social media activity

Today

- Principal components analysis, a tool used for data visualization or data pre-processing before supervised techniques are applied, and
- Clustering, a broad class of methods for discovering unknown subgroups in data.

Principal Components Analysis

Principal Components Analysis

- Principal components analysis aims to summarize the variation in a matrix of indicators, $X_{i,j}$
- The idea of principal components analysis is to re-describe the $X_{i,j}$ in terms of a smaller number of uncorrelated new variables that capture as much as possible of the total variance in the data
- These new uncorrelated variables are the **principal components**
- Apart from producing derived variables for use in supervised learning problems, PCA also serves as a tool for data visualization and data summary

Running example: views about representation

What do people want from their representatives?

In a recent survey, we asked people for their views on their political representatives. In particular, we asked them the degree to which they wanted their political representatives to:

- Pursue policy goals that they preferred (**substantive** representation)
- Have descriptive characteristics (gender; sexuality; race/ethnicity; education) that they shared (**descriptive** representation)

They are complicated concepts, however, so rather than asking survey respondents a single question on their views, we asked two batteries of survey items.

Running example: views about representation

Think about what you would want from **someone who acts and speaks for you in politics**. Do you agree or disagree that **this person should share your...**

	Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree
Education	○	○	○	○	○
Sexual orientation	○	○	○	○	○
Class background	○	○	○	○	○
Ethnicity	○	○	○	○	○
Gender	○	○	○	○	○

Running example: views about representation

Think about what you would want from **someone who acts and speaks for you in politics.** To what extent do you agree or disagree with the following statements? **This person should**...

	Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree
...speak in favour of the views and opinions that I hold on different political issues	○	○	○	○	○
...raise issues that are important to me	○	○	○	○	○
...promote policies that would benefit me, even on issues I am unfamiliar with	○	○	○	○	○
...promote the policy views that I hold	○	○	○	○	○

Running example: views about representation

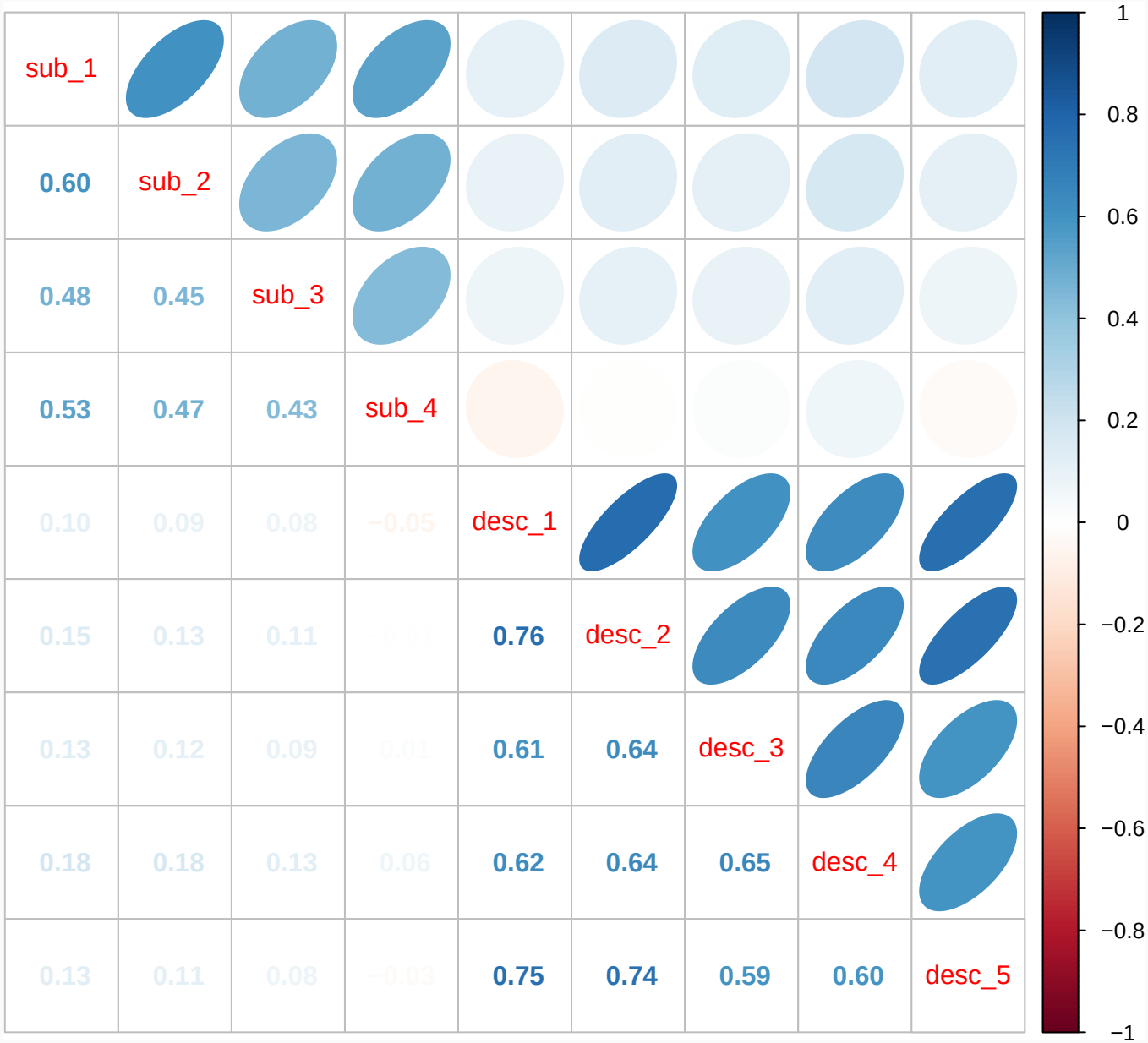
```
##   sub_1 sub_2 sub_3 sub_4 desc_1 desc_2 desc_3 desc_4 desc_5
## 1     3     3     3     4      3      3      3      3      3
## 2     5     5     5     5      5      5      5      5      5
## 3     3     3     3     4      3      3      2      3      2
## 4     4     3     2     4      1      1      1      2      1
## 5     4     4     4     4      3      3      3      3      3
## 6     3     4     3     4      3      3      3      3      3
```

5 = strongly agree; 1 = strongly disagree

Running example: views about representation

```
##        sub_1 sub_2 sub_3 sub_4 desc_1 desc_2 desc_3 desc_4 desc_5
## sub_1   1.00  0.60  0.48  0.53   0.10   0.15   0.13   0.18   0.13
## sub_2   0.60  1.00  0.45  0.47   0.09   0.13   0.12   0.18   0.11
## sub_3   0.48  0.45  1.00  0.43   0.08   0.11   0.09   0.13   0.08
## sub_4   0.53  0.47  0.43  1.00  -0.05  -0.01   0.01   0.06  -0.03
## desc_1  0.10  0.09  0.08 -0.05   1.00   0.76   0.61   0.62   0.75
## desc_2  0.15  0.13  0.11 -0.01   0.76   1.00   0.64   0.64   0.74
## desc_3  0.13  0.12  0.09  0.01   0.61   0.64   1.00   0.65   0.59
## desc_4  0.18  0.18  0.13  0.06   0.62   0.64   0.65   1.00   0.60
## desc_5  0.13  0.11  0.08 -0.03   0.75   0.74   0.59   0.60   1.00
```

Running example: views about representation



Running example: views about representation

- We are clearly measuring related attitudes on these dimensions
- It is somewhat inconvenient, however, to have multiple highly correlated measures
    - Harder to summarise the 'position' of any given individual on each dimension
    - Potentially harder to use these concepts in down-stream prediction tasks (i.e. are people who want more substantive representation likely to vote for different political parties, etc)
- It would be helpful to have a technique to reduce the complexity of the data but still capture the main dimensions of interest

Principal Components Analysis: details

- Goal: transform a set of features $X_{i,1}, X_{i,2}, \ldots, X_{i,P}$ into a smaller number of uncorrelated variables that capture as much variance in the original set as possible
- PCA finds linear combinations of the X_p variables that are uncorrelated with one another. For the first two principal components:

$$Z_{i,1} = \phi_{1,1}X_{i,1} + \phi_{2,1}X_{i,2} + \cdots + \phi_{p,1}X_{i,p} \quad (1)$$

$$Z_{i,2} = \phi_{1,2}X_{i,1} + \phi_{2,2}X_{i,2} + \cdots + \phi_{p,2}X_{i,p} \quad (2)$$

- $Z_{i,1}$ (the principal component scores) are a weighted sum of the original features, where the weights (ϕ) are coefficients that we estimate
- We refer to the elements $\phi_{1,1}, \ldots, \phi_{p1}$ as the loadings of the principal component
- Note that there are different loadings for the different principal components (i.e. $\phi_{1,1} \neq \phi_{1,2}$)

Computation of Principal Components

- We center each of the variables in $\mathbf{X}$ to have mean zero (that is, the column means of $\mathbf{X}$ are zero). This prevents PCA from being sensitive to the scales of the input variables
- We then look for the linear combination of the sample feature values of the form

$$Z_{i,1} = \phi_{1,1} X_{i,1} + \phi_{2,1} X_{i,2} + \cdots + \phi_{p,1} X_{i,p} \tag{3}$$

  for $i = 1, \ldots, n$ that has largest sample variance
- We do this subject to the constraint that $\sum_{j=1}^{p} \phi_{j1}^2 = 1$. (Why? Because we can get arbitrarily large variance for $Z_{i,1}$ if ϕ can be arbitrarily large)
- Since each of the $X_{i,p}$ has mean zero, then so does $Z_{i,1}$.

Computation of Principal Components

- More formally, the first principal component loading vector solves the optimization problem

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

  where, because $\frac{1}{n}\sum_{i=1}^{n} x_{i,j} = 0$, the first term is the sample variance of $Z_{i,1}$.

- 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 $Z_{11}, \cdots, Z_{n1}$.

Geometry of PCA

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

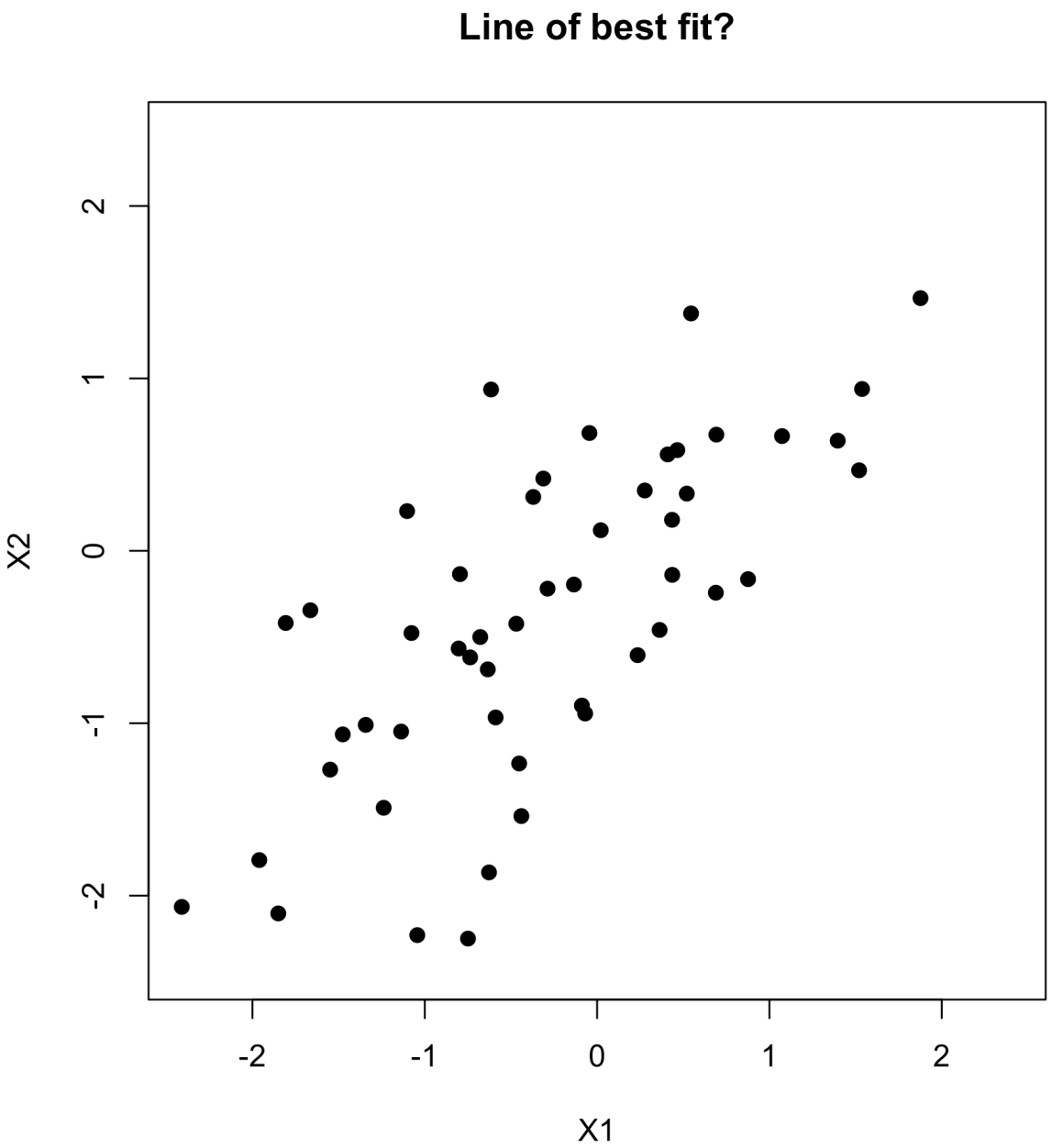
PCA finds the hyperplane closest to the observations

- 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 Euclidean 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 principal component.
- For instance, the first two principal components of a data set span the plane that is closest to the n observations, in terms of average squared Euclidean distance.

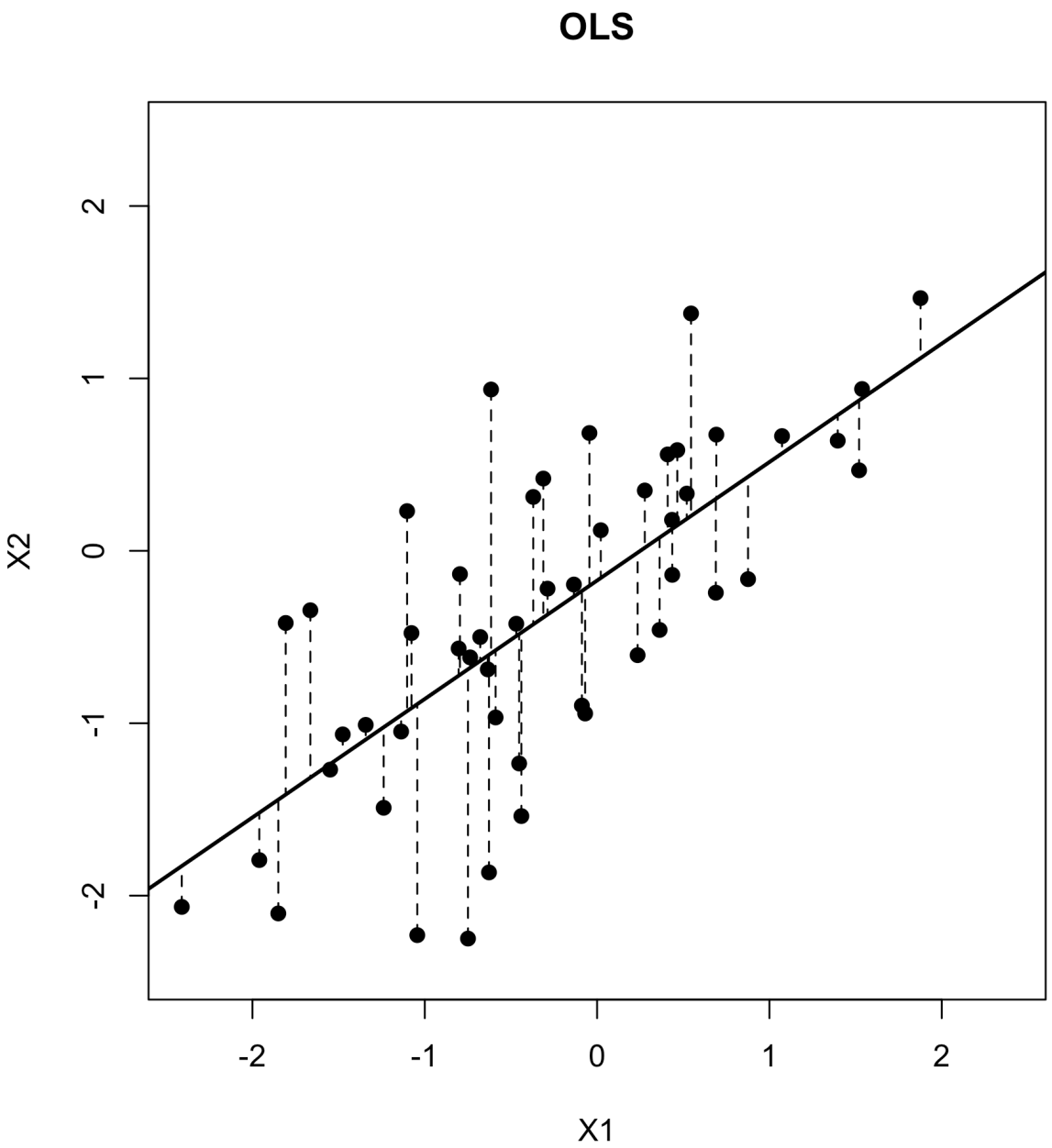
What is going on here?

- PCA is concerned with finding a linear relationship between our X variables that best summarises the variation within them
- We find the first principal component by finding the line that 'best' predicts our data, X
- But this sounds a lot like regression!
- In OLS, we find the line that 'best' describes variation in Y
- What is the difference?

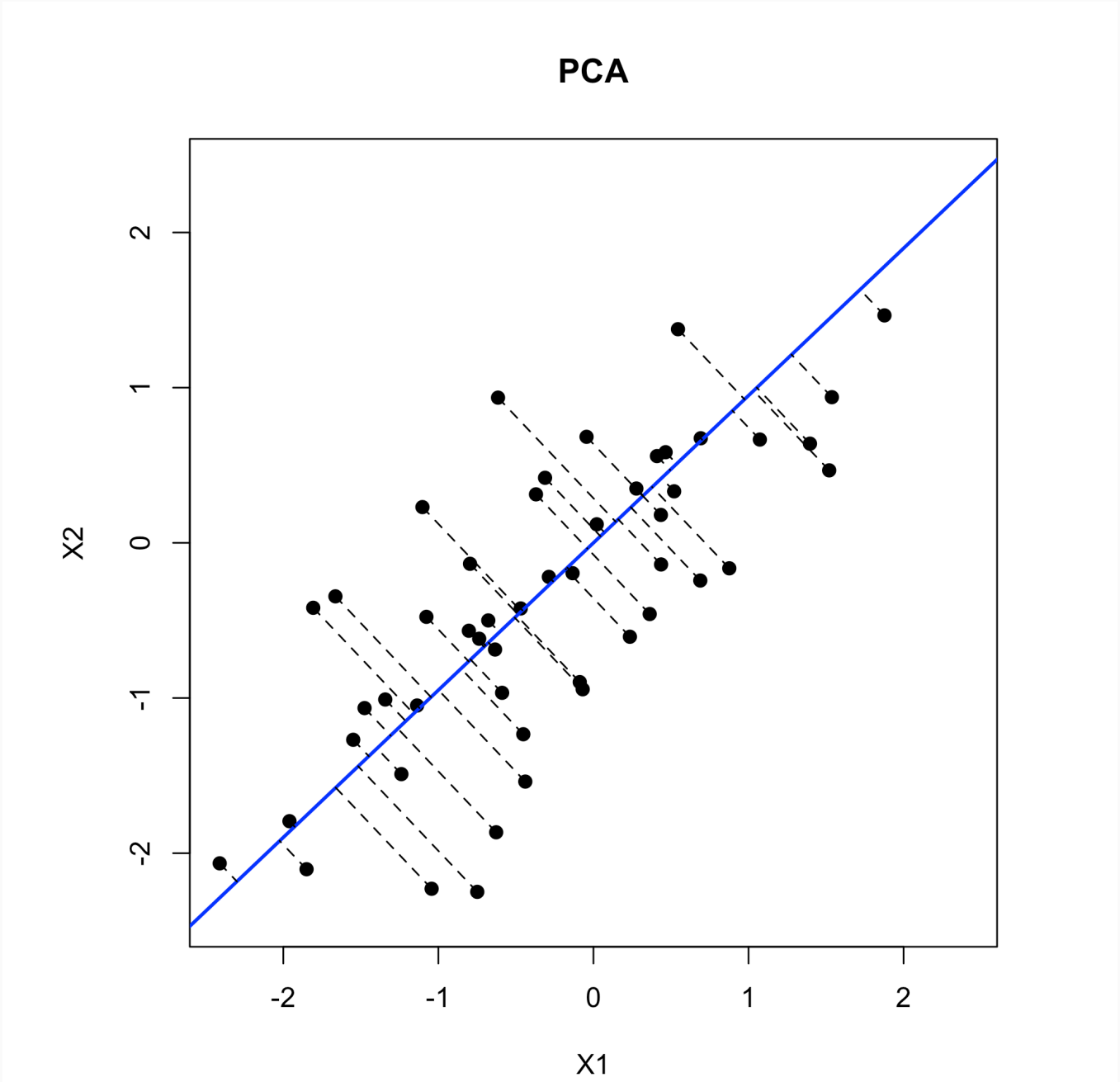
PCA: illustration



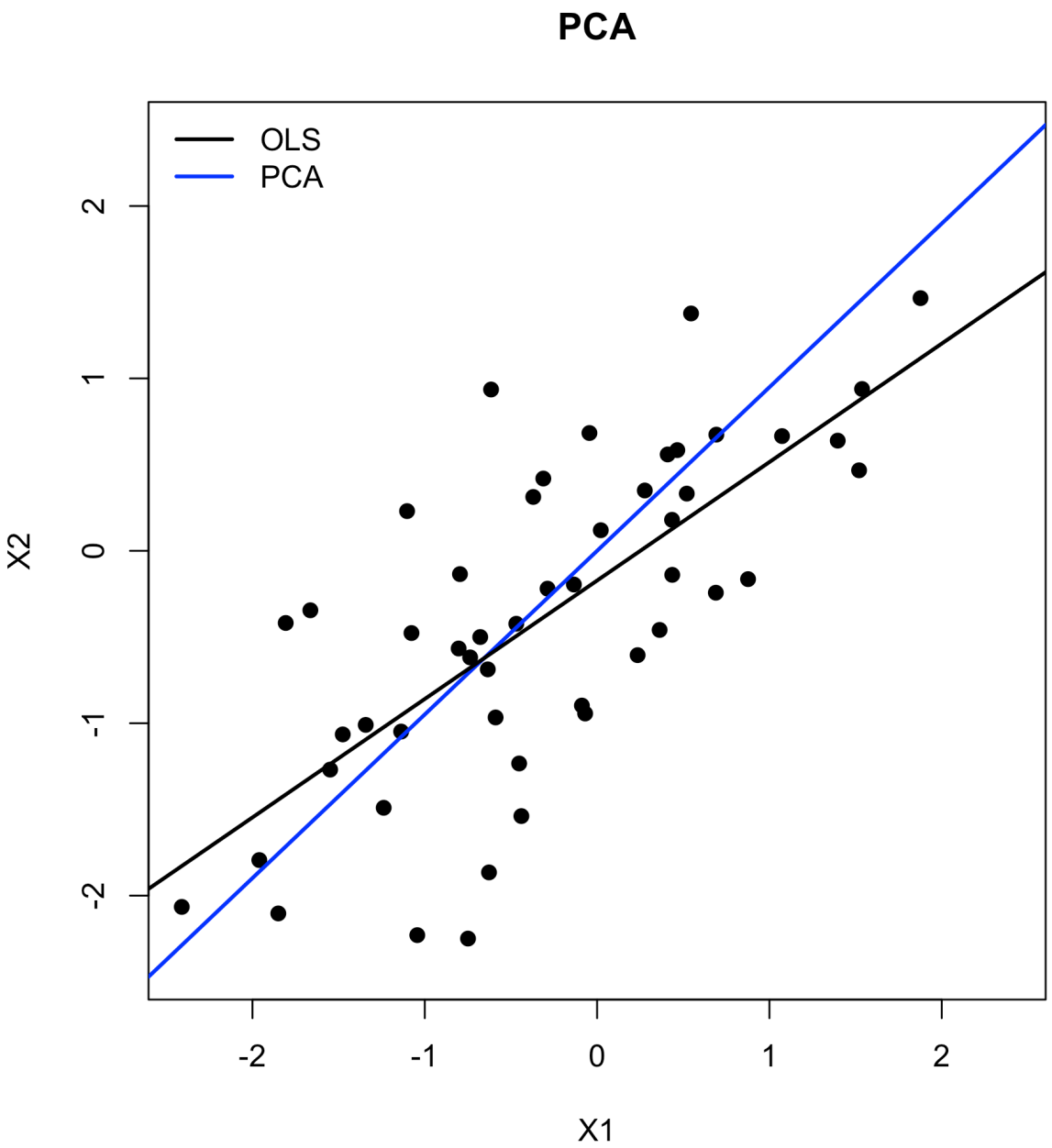
PCA vs OLS



PCA vs OLS



PCA vs OLS



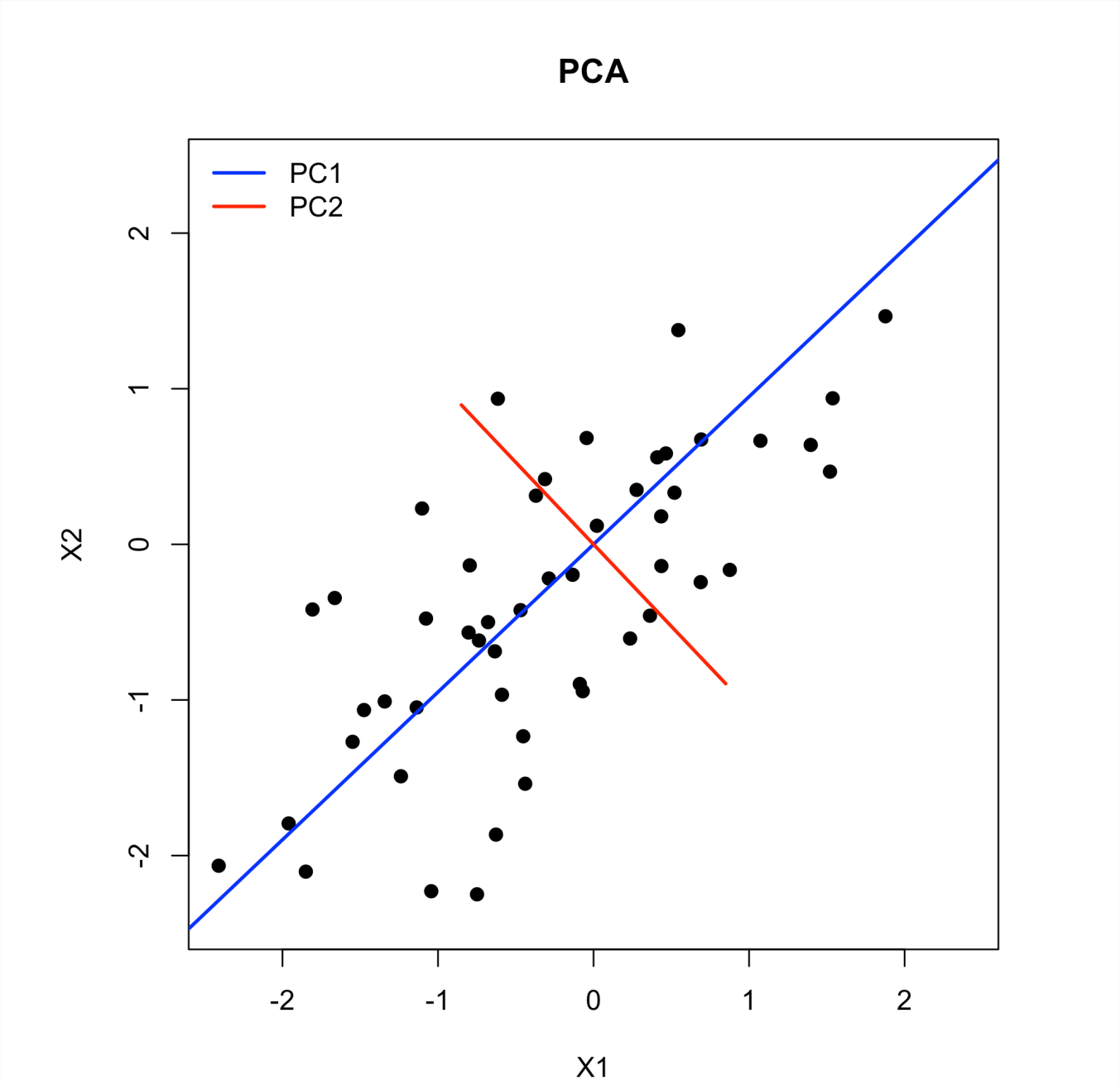
Further principal components

- 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 $z_{12}, z_{22}, \ldots, z_{n2}$ take the form

$$Z_{i2} = \phi_{12} X_{i1} + \phi_{22} X_{i2} + \cdots + \phi_{p2} X_{ip},$$

  where ϕ_2 is the second principal component loading vector, with elements $\phi_{12}, \phi_{22}, \cdots, \phi_{p2}$.
- The ϕ are estimated in the same way as for PC1, but with the added constraint that Z_2 should be uncorrelated with Z_1.
- Constraining Z_2 to be uncorrelated with Z_1 is equivalent to constraining the direction ϕ_2 to be orthogonal (perpendicular) to the direction ϕ_1.

PC1 and PC2



Example - Representation

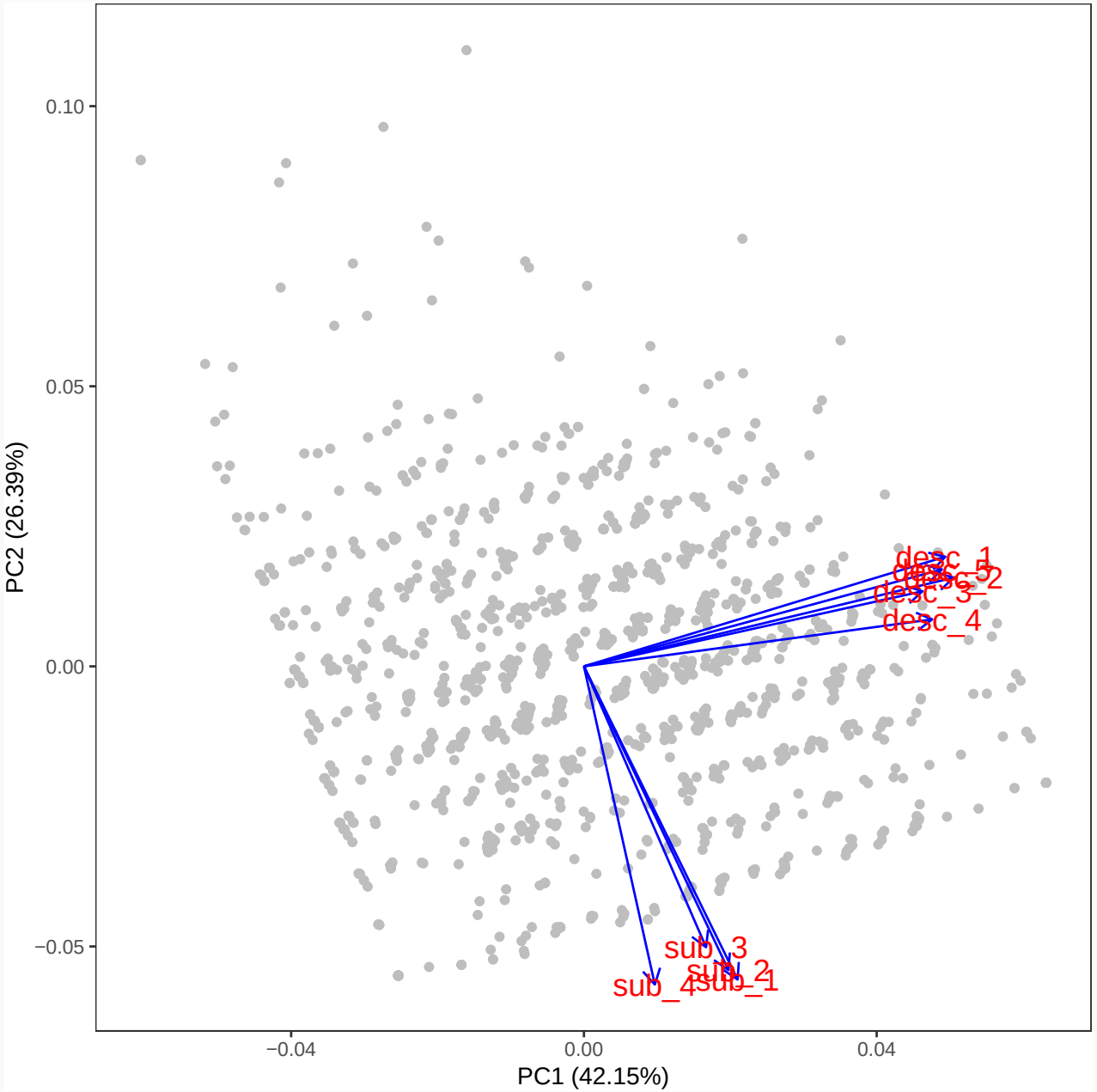
What happens when we apply PCA to our matrix of survey responses?

- Data: Includes 2091 observations of responses to 9 survey items
- The principal component score vectors therefore have length $n = 2091$
- The principal component loading vectors therefore have length $p = 9$
- We standardize each variable to have mean zero and standard deviation one before running PCA

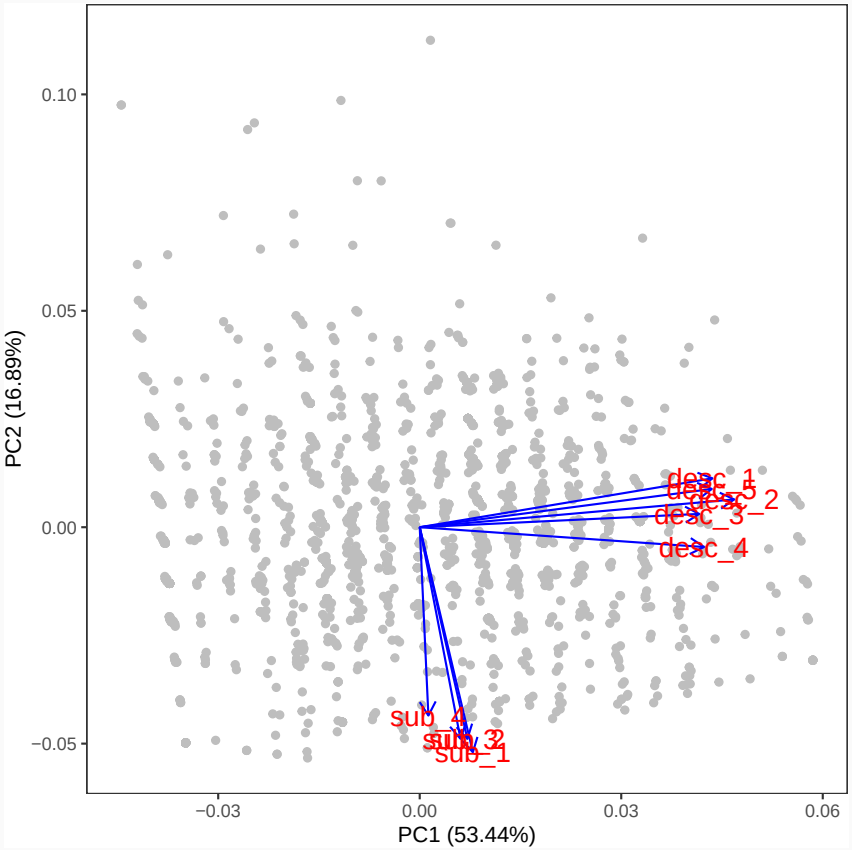
Code:

```
uk_pca <- prcomp(uk)
```

Example - Representation Biplot



Example - Representation Biplot



- Grey points represent the 1st and 2nd PCA scores for each observation
- Blue arrows represent the first 2 PCA loading vectors
- The first principal component mostly describes variation in the responses to the descriptive questions
- The second principal component mostly describes variation in the responses to the substantive questions

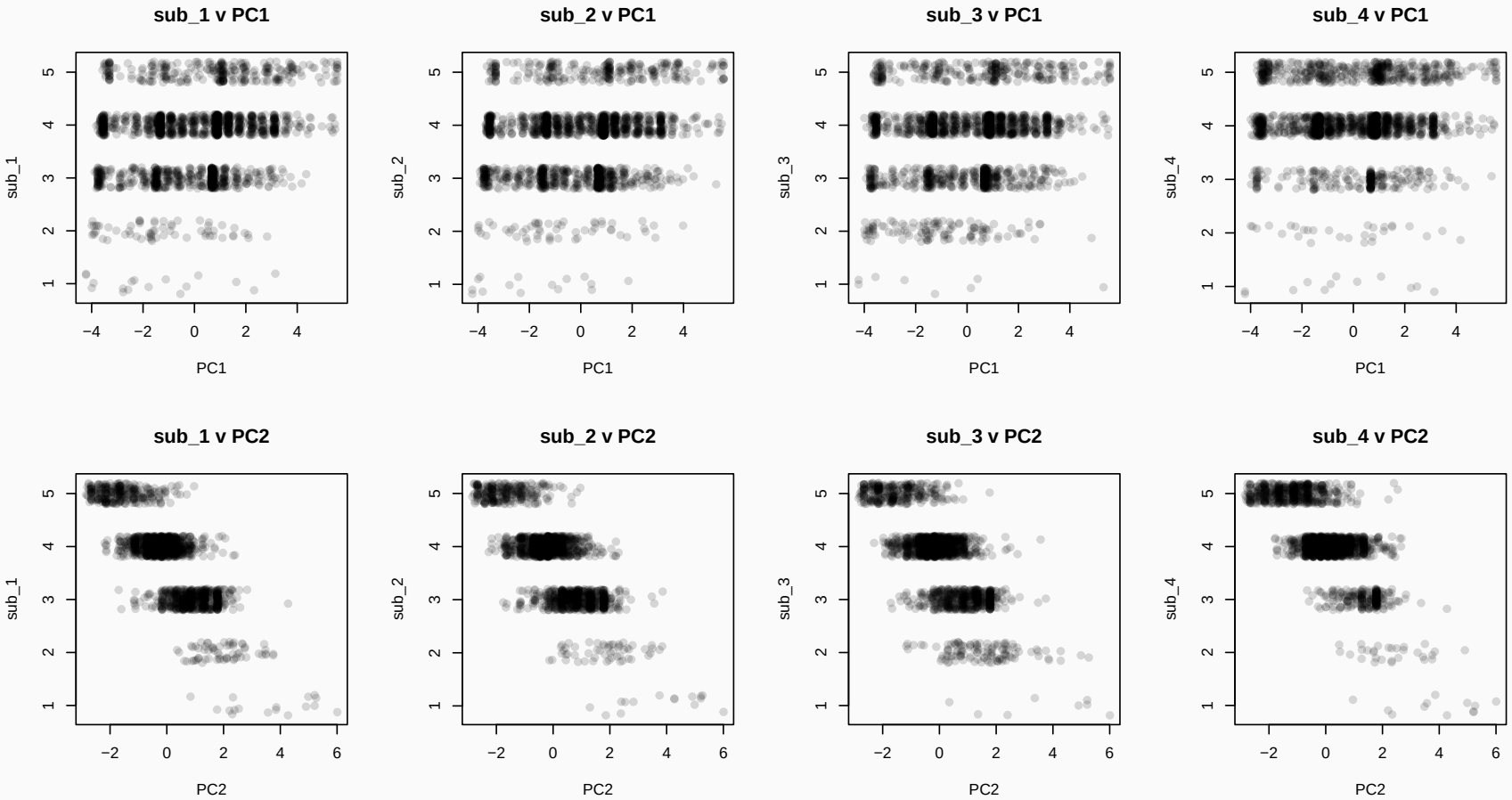
Example - Representation Biplot

We can also directly examine the loading vectors

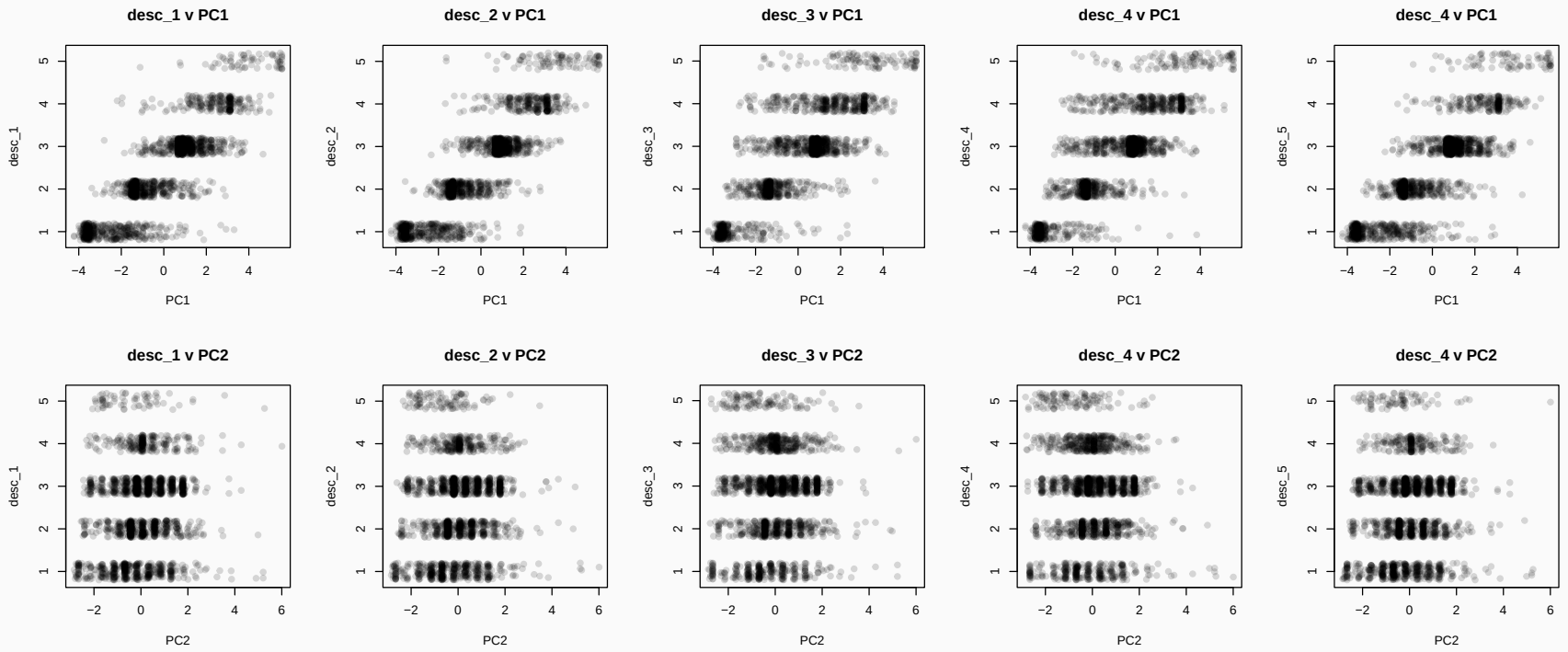
```
##           PC1    PC2
## sub_1   0.081 -0.529
## sub_2   0.072 -0.498
## sub_3   0.062 -0.497
## sub_4   0.013 -0.443
## desc_1 0.443  0.115
## desc_2 0.477  0.065
## desc_3 0.423  0.031
## desc_4 0.431 -0.047
## desc_5 0.443  0.090
```

- PC1 puts roughly equal weight on the five ”descriptive” variables
- PC2 puts roughly equal weight on the four ”substantive” variables

Example - PC correlations

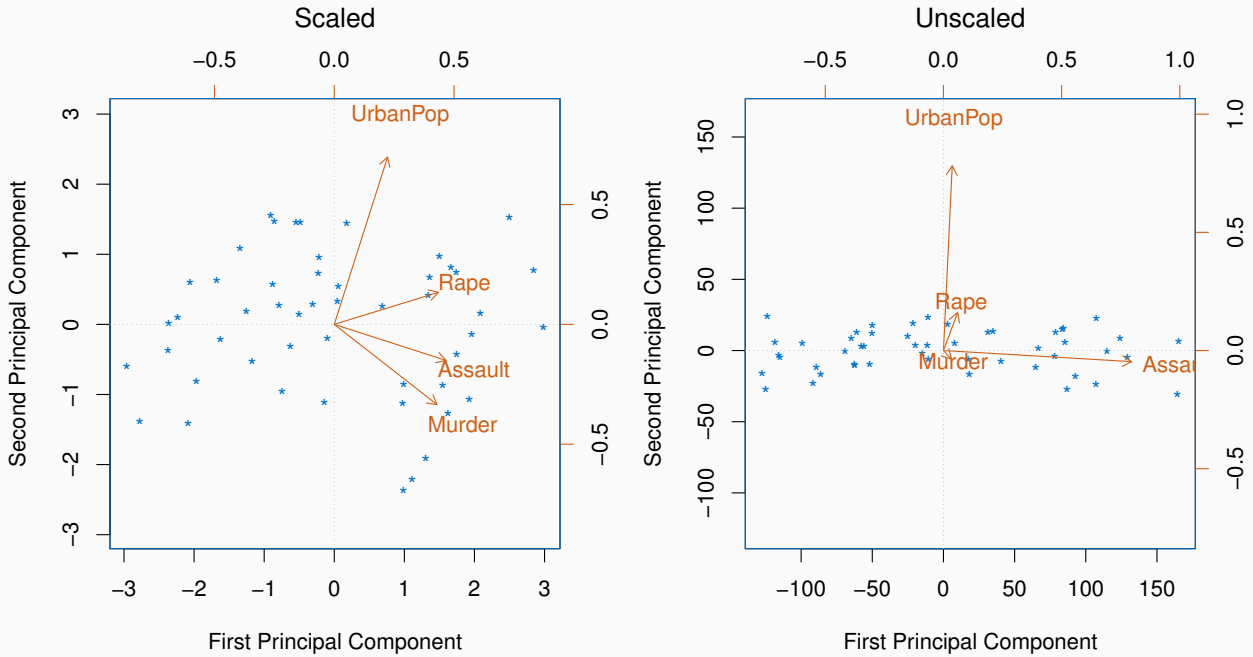


Example - PC correlations



Scaling of the variables matters!

If the variables are in different units, scaling each to have standard deviation equal to one is recommended.



How many principal components should we use?

- We have only been focusing on the first two principal components of our data
- In general, we can calculate p principal components
- Clearly, transforming X into a matrix of PCA scores with the same number of dimensions is not very helpful
- Question: How many principal components do we want/need?
- Answer: It depends what we want them for!

How many principal components should we use?

For data summary/description:

- When we are trying to summarise complicated data, we typically want the smallest number of principal components that allows us to explain a large amount of variation in the original data
- This is typically achieved by looking at a scree plot, which describes the fraction of variance explained by increasing numbers of principal components
- In practice, we look for an 'elbow' in the plot which suggests declining marginal returns of further components

Proportion Variance Explained

- To understand the strength of each component, we calculate the proportion of variance explained (PVE) by each one.
- The total variance present in a data set (assuming that the variables have been centered to have mean zero) is defined as

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

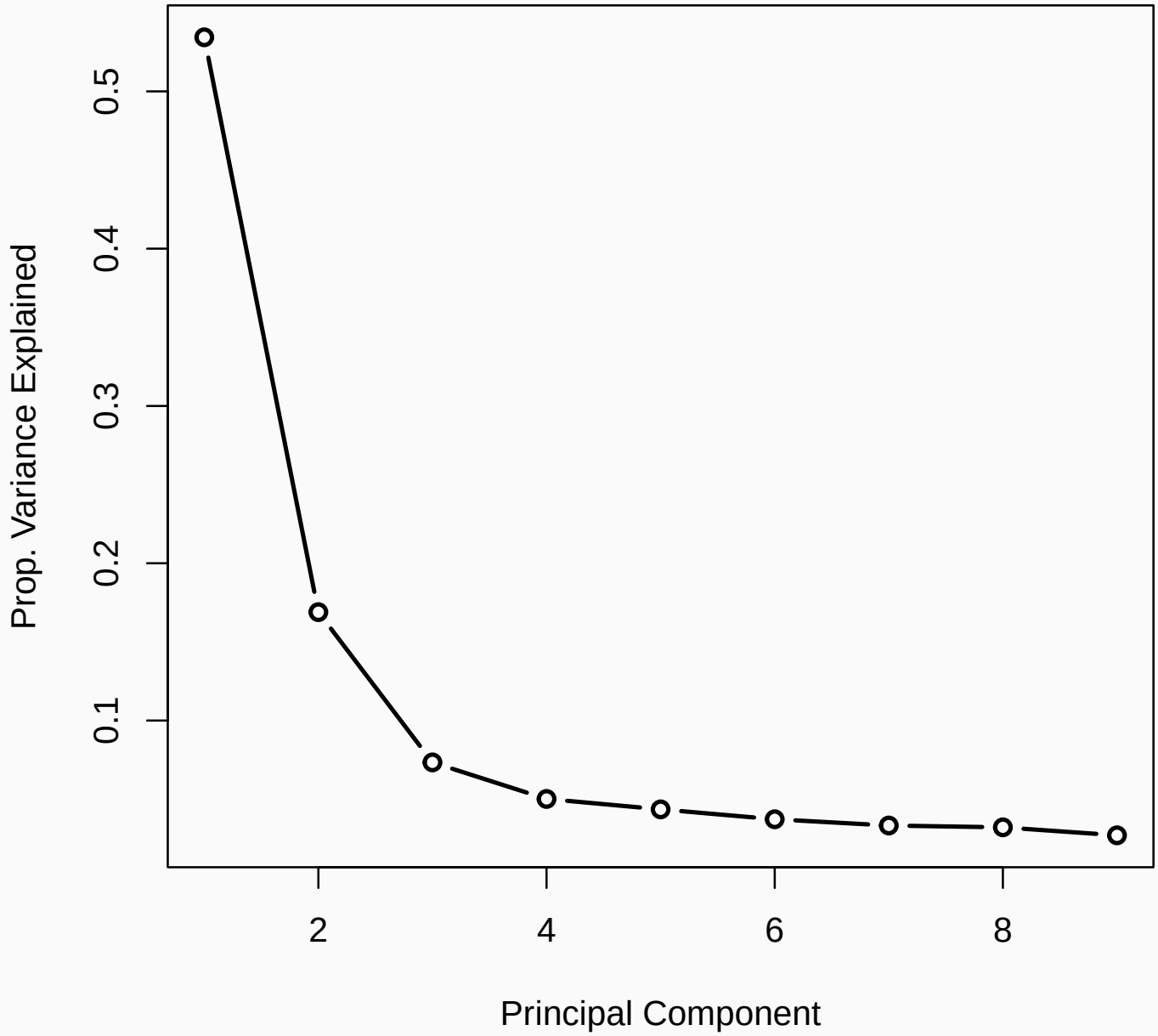
  and the variance explained by the mth principal component is

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

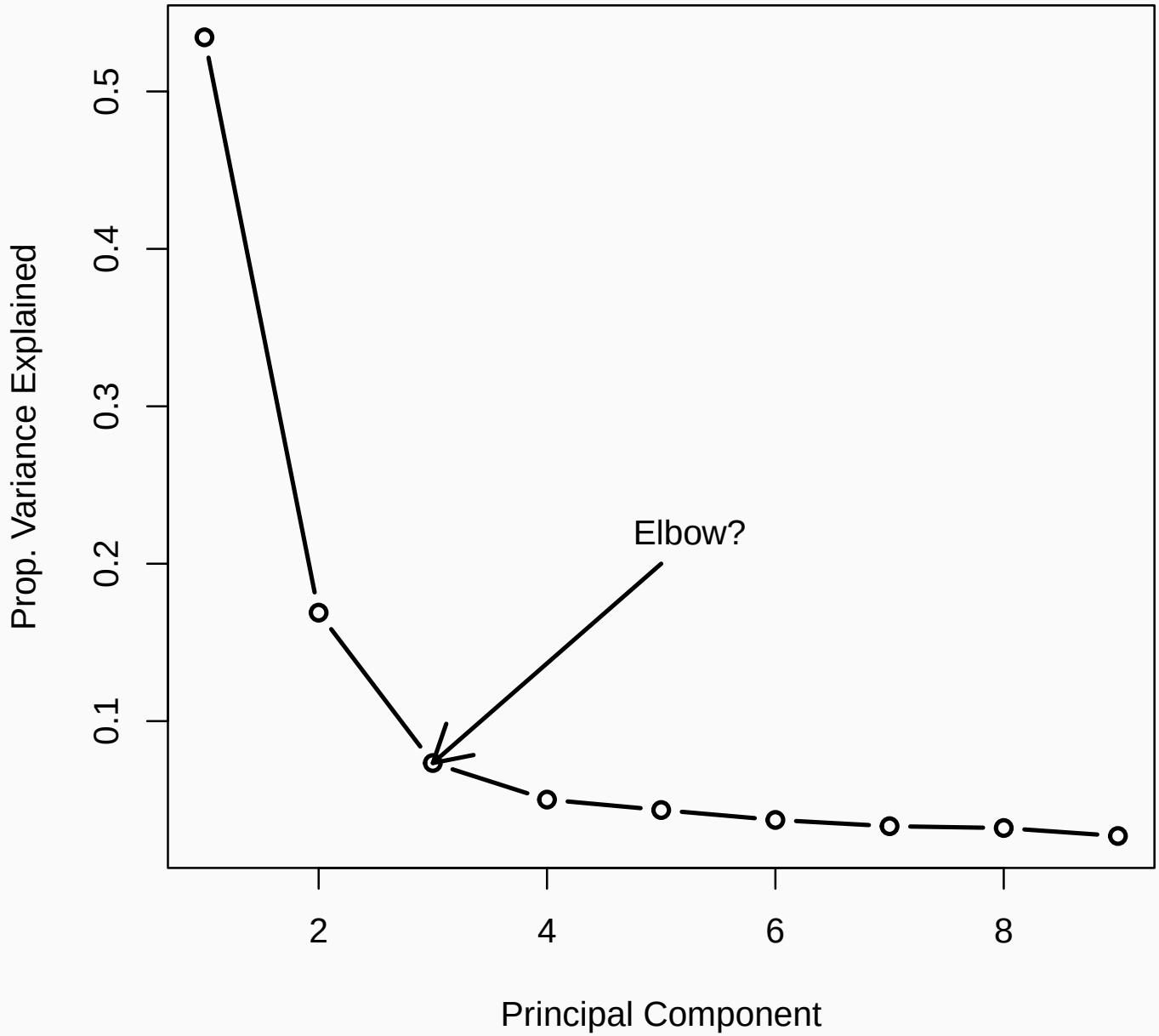
- Therefore, the PVE of the mth principal component is given by the positive quantity between 0 and 1

$$\frac{\sum_{i=1}^{n} z_{im}^2}{\sum_{j=1}^{p} \sum_{i=1}^{n} x_{ij}^2}.$$

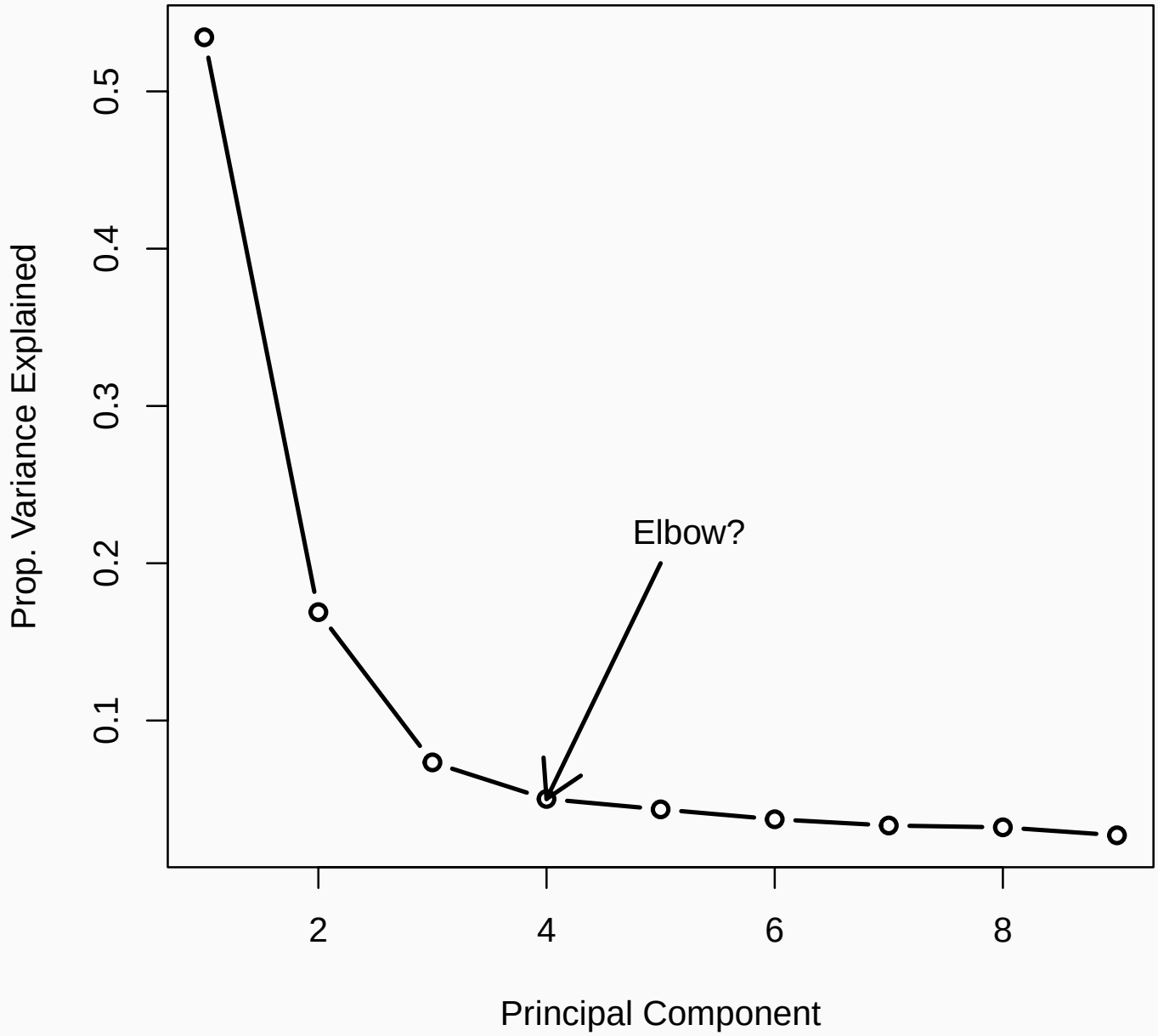
Scree plot - example



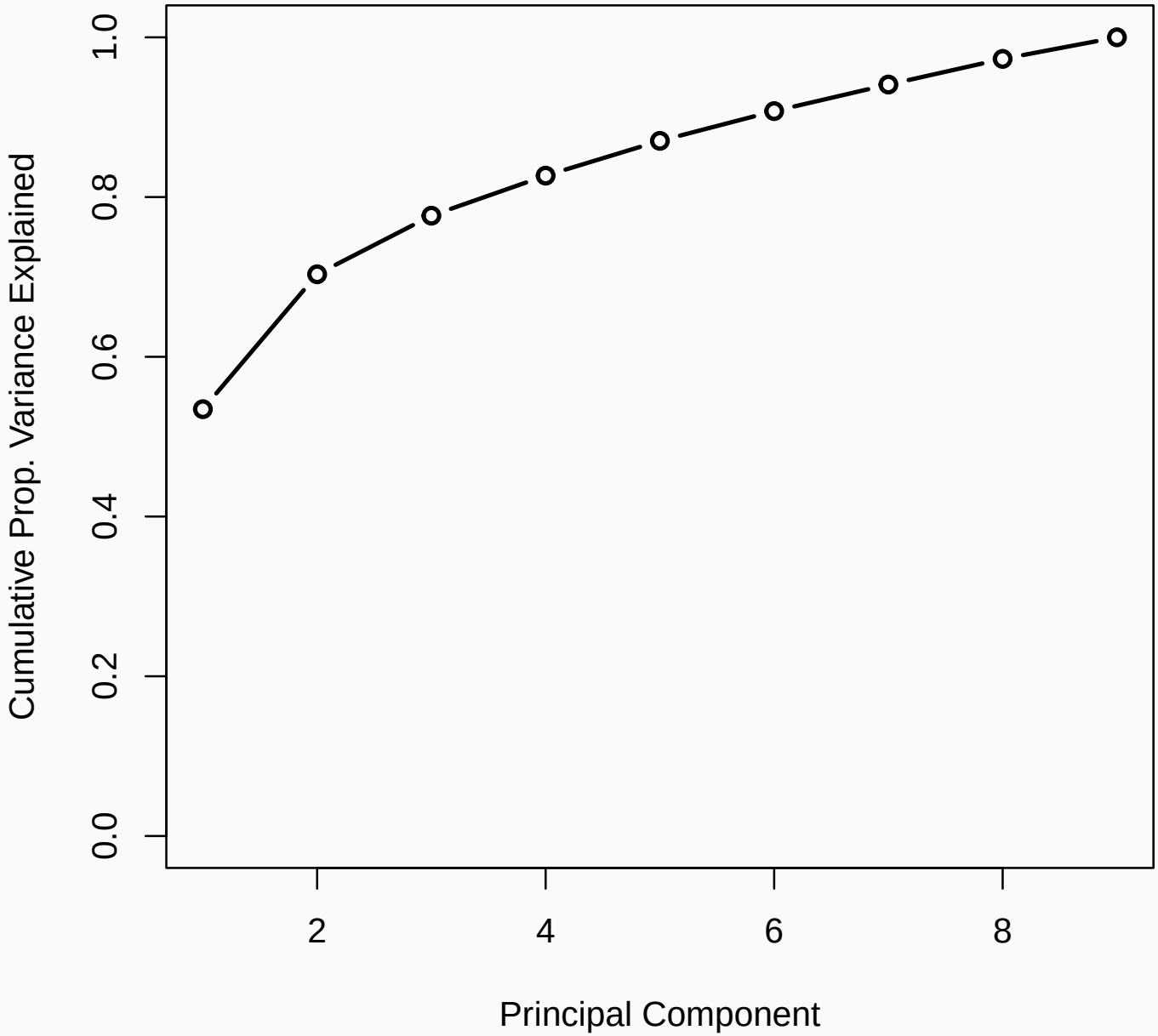
Scree plot - example



Scree plot - example



Scree plot - example



What is going on here?

- The first M principal component score and loading vectors provide the best M directional approximation to our Xs

$$X_{i,j} \approx \sum_{m=1}^{M} Z_{i,m}\phi_{j,m} \tag{4}$$

- We can therefore "reconstruct" the X values from our estimated principal components and loadings
- If we do this sequentially, first just using the first principal component, then adding more (i.e. increasing M), we will more and more closely reconstruct the relative values of X
- We are explaining more total variance in the data with each principal component, but the additional amount explained decreases as we increase M

Principal Component Regression

Principal Component Regression

- Yesterday we discussed involved fitting linear regression models with some kind of shrinkage (lasso; ridge) using the original predictors, $X_1, X_2, \dots, X_p$.
- We can also use methods like PCA to transform the predictors and then fit a least squares model using the transformed variables.
- These methods can have similar predictive benefits to the shrinkage models, and can outperform OLS applied to the original predictors

Principal Component Regression

- Recall the definition of the first principal component as a linear combination of P original X variables

$$Z_{i,1} = \phi_{1,1}X_{i,1} + \phi_{2,1}X_{i,2} + \cdots + \phi_{p,1}X_{i,p} \tag{5}$$

- With this definition, and a Y we wish to predict, rather than estimating the regression:

$$Y_i = \sum_{p=1}^{P} \beta_p X_{i,p} + \epsilon_i \tag{6}$$

- We can instead estimate (using ordinary least squares):

$$Y_i = \sum_{m=1}^{M} \theta_m Z_{i,m} + \epsilon_i \tag{7}$$

where $M << P$ and $Z_{i,m}$ are the scores for the mth principal component and $\theta_0, \theta_1, \ldots, \theta_M$ are the regression coefficients.

Principal Component Regression - example

- We can use our representation example to illustrate PC regression
- Our goal is to predict the political interest of our respondents using the items on descriptive and substantive representation
- Model 1: $Y_i = \sum_{p=1}^{P} \beta_p X_{i,p} + \epsilon_i$
- Model 2: $Y_i = \sum_{m=1}^{M} \theta_m Z_{i,m} + \epsilon_i$

Principal Component Regression - example

```
##
## ====================================
##              Model 1      Model 2
## ------------------------------------
## (Intercept)    -0.25         2.54 ***
##                (0.24)       (0.03)
## sub_1           0.22 ***
##                (0.06)
## sub_2           0.25 ***
##                (0.06)
## sub_3           0.18 ***
##                (0.05)
## sub_4           0.12
##                (0.06)
## desc_1         -0.04
##                (0.06)
## desc_2         -0.02
##                (0.05)
## desc_3          0.04
##                (0.04)
## desc_4          0.01
##                (0.05)
## desc_5         -0.05
##                (0.05)
## pc1                          0.02
##                             (0.02)
## pc2                         -0.39 ***
##                             (0.03)
## ------------------------------------
## R^2             0.08         0.08
## Adj. R^2        0.08         0.08
## Num. obs.    2090         2090
## ====================================
## *** p < 0.001; ** p < 0.01; * p < 0.05
```

Principal Component Regression - example

Note that:

- The PC scores capture the most salient patterns (political interest is unrelated to views about descriptive representation, and positively associated with views about substantive representation)
- The standard errors on the PC scores are smaller $\longrightarrow$ the signal from the item batteries is captured by the PC scores, while the noise is not
- The predictive accuracy of the models (as measured by adj. R^2) are roughly comparable

Principal Component Regression - example

We can use cross-validation to determine which is the better fit (and how many principal components to use):

```
##          Model   MSE
## 1 Original Xs 2.354
## 2     PC1:PC2 2.324
## 3     PC1:PC3 2.338
## 4     PC1:PC4 2.341
## 5     PC1:PC5 2.342
```

$\longrightarrow$ In this case, the principal component regression with PC1 and PC2 has the lowest MSE.

Clustering

Clustering

- Clustering refers to a very broad set of techniques for finding subgroups, or clusters, in a data set.
- We seek a partition of the data into distinct groups so that the observations within each group are quite similar to each other.
- To make this concrete, we must define what it means for two or more observations to be similar or different.
- Indeed, this is often a domain-specific consideration that must be made based on knowledge of the data being studied.

PCA vs Clustering

- PCA looks for a low-dimensional representation of the observations that explains a good fraction of the variance.
- Clustering looks for homogeneous subgroups among the observations.

Applications of Clustering

Clustering methods have numerous applications across various fields, from finance to healthcare, and from computer vision to natural language processing.

1. **Image Segmentation and Compression**
   - Used in digital image processing for object recognition and image file size reduction.
2. **Customer Segmentation in Marketing**
   - Businesses segment customers into different groups based on their behaviors or characteristics
3. **Document Clustering in Information Retrieval**
   - Improves the efficiency and relevance of search results in information retrieval systems such as search engines
4. **Bioinformatics**
   - Utilized in genomics to cluster gene expression data, aiding in the identification of similar behaviors across different conditions

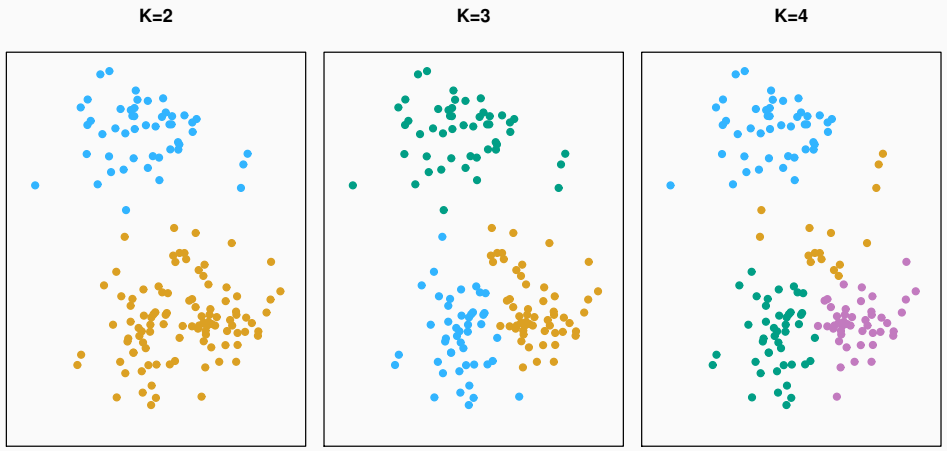
Clustering for Segmentation

- Suppose we have access to a large number of measurements (e.g. median household income, occupation, distance from nearest urban area, and so forth) for a large number of people.
- Our goal is to perform segmentation by identifying subgroups of people who might be more receptive to a particular form of advertising, or more likely to purchase a particular product.
- The task of performing segmentation amounts to clustering the people in the data set.

Two clustering methods

- In K-means clustering, we seek to partition the observations into a pre-specified number of clusters.
- In hierarchical clustering, we do not know in advance how many clusters we want; in fact, we end up with a tree-like visual representation of the observations, called a dendrogram, that allows us to view at once the clusterings obtained for each possible number of clusters, from 1 to n.

K-means clustering



- A simulated data set with 150 observations in 2-dimensional space.
- Panels show the results of applying K-means clustering with different values of K, the number of clusters.
- The color of each observation indicates the cluster to which it was assigned using the K-means clustering algorithm.
- Note that there is no ordering of the clusters, so the cluster coloring is arbitrary. These cluster labels were not used in clustering; instead, they are the outputs of the clustering procedure.

Details of K-means clustering

- Let $C_1, \ldots, C_K$ denote sets containing the indices of the observations in each cluster. These sets satisfy two properties:

    1. Each observation belongs to at least one of the K clusters
    2. Clusters are non-overlapping (no observation belongs to more than one cluster)

- For instance, if the ith observation is in the kth cluster, then $i \in C_k$.

Details of K-means clustering: continued

- The idea behind K-means clustering is that a good clustering is one for which the within-cluster variation is as small as possible.
- The within-cluster variation (WCV) for cluster C_k is a measure of the amount by which the observations within a cluster differ from each other.
- We aim to solve:

$$\underset{C_1,\ldots,C_K}{\text{minimize}} \left\{ \sum_{k=1}^{K} \text{WCV}(C_k) \right\}. \tag{8}$$

- In words, this formula says that we want to partition the observations into K clusters such that the total within-cluster variation, summed over all K clusters, is as small as possible.

How to define within-cluster variation?

- Typically we use Euclidean distance

$$\mathrm{WCV}(C_k) = \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2, \tag{9}$$

  where $|C_k|$ denotes the number of observations in the kth cluster.

- This is the sum of all the pairwise (squared) distances between the observations in the cluster, divided by the number of observations in that cluster

- Note that trying to directly minimize this function is hard! There are $\approx K^N$ ways of partitioning N observations into K clusters.

K-Means Clustering Algorithm

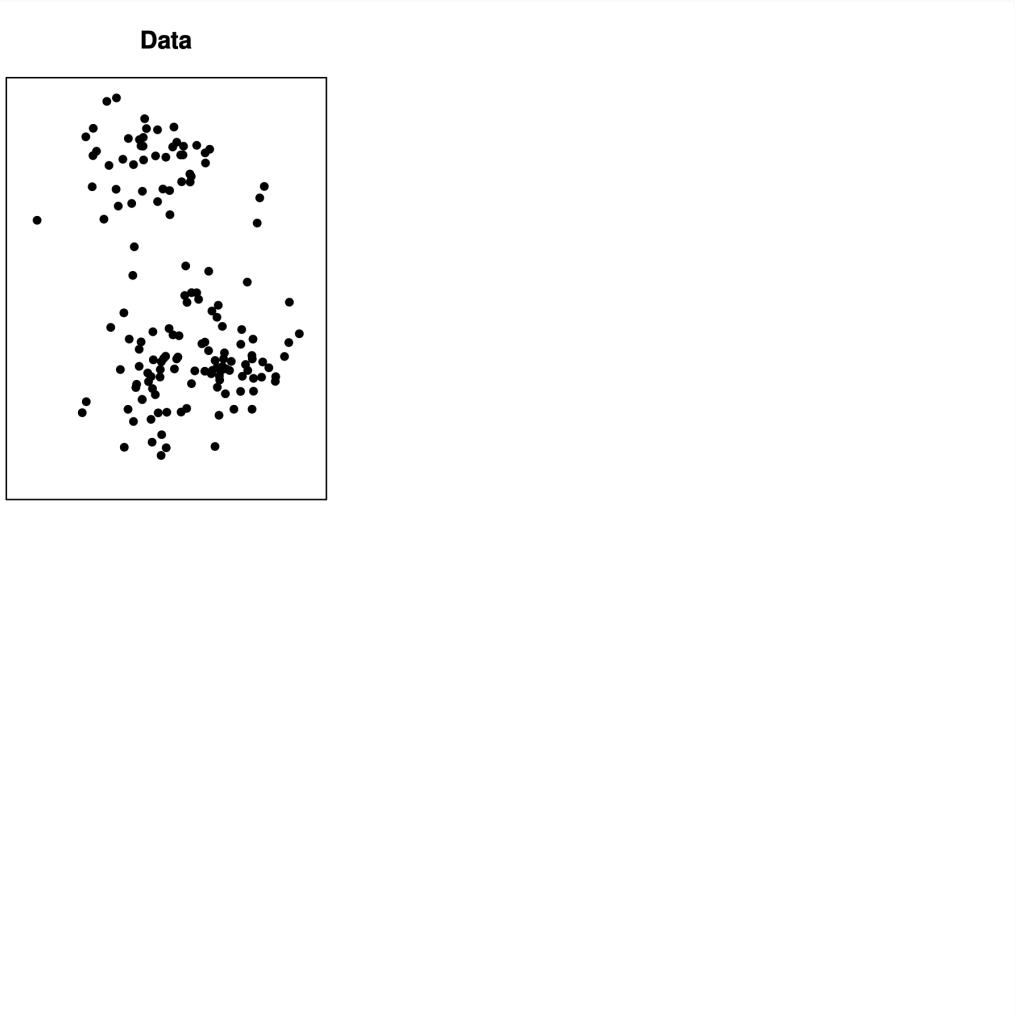
An exact solution is therefore very hard, but we can find a local optimum using a very simple algorithm:

1. Randomly assign a number, from 1 to K, to each of the observations. These serve as initial cluster assignments for the observations.
2. Iterate until the cluster assignments stop changing:
    2.1 For each of the K clusters, compute the cluster centroid. The kth cluster centroid is the vector of the p feature means for the observations in the kth cluster.
    2.2 Assign each observation to the cluster whose centroid is closest (where closest is defined using Euclidean distance).

Properties of the Algorithm

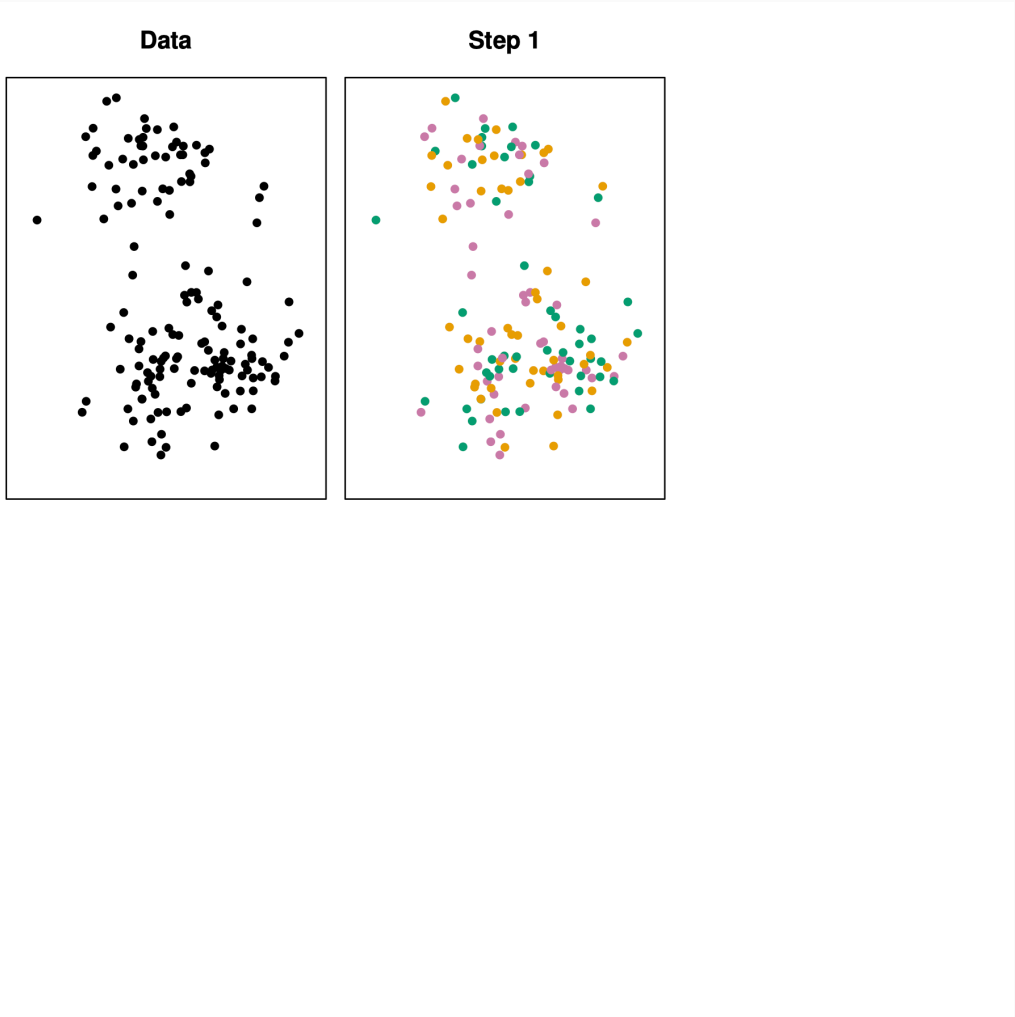
- This algorithm is guaranteed to decrease the value of the objective at each step.
- That is, the within-cluster variation will decrease with each iteration of the algorithm
- When the WCV can no longer decrease by moving any observations into new clusters, the algorithm terminates and we will have reached a local optimum
- However it is not guaranteed that the termination point will result in the global minimum within-cluster variation because the final cluster assignments will depend on the initial random allocation of points to clusters

K-means clustering



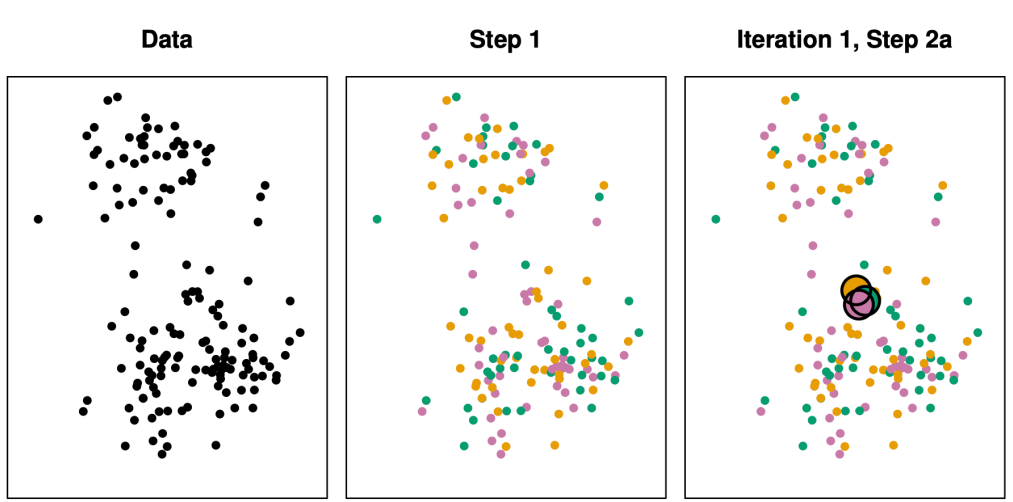
Start with raw data in multiple dimensions.

K-means clustering



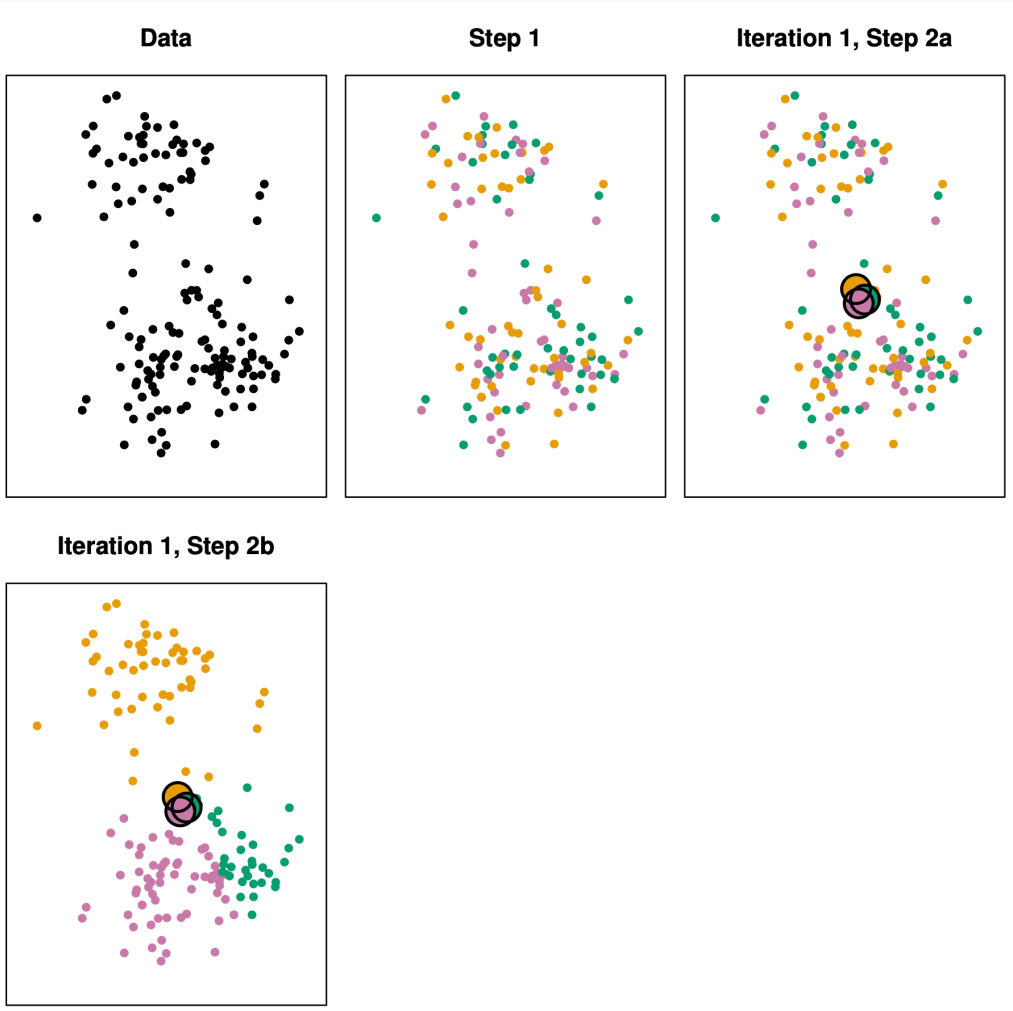
Step 1: each observation is randomly assigned to a cluster.

K-means clustering



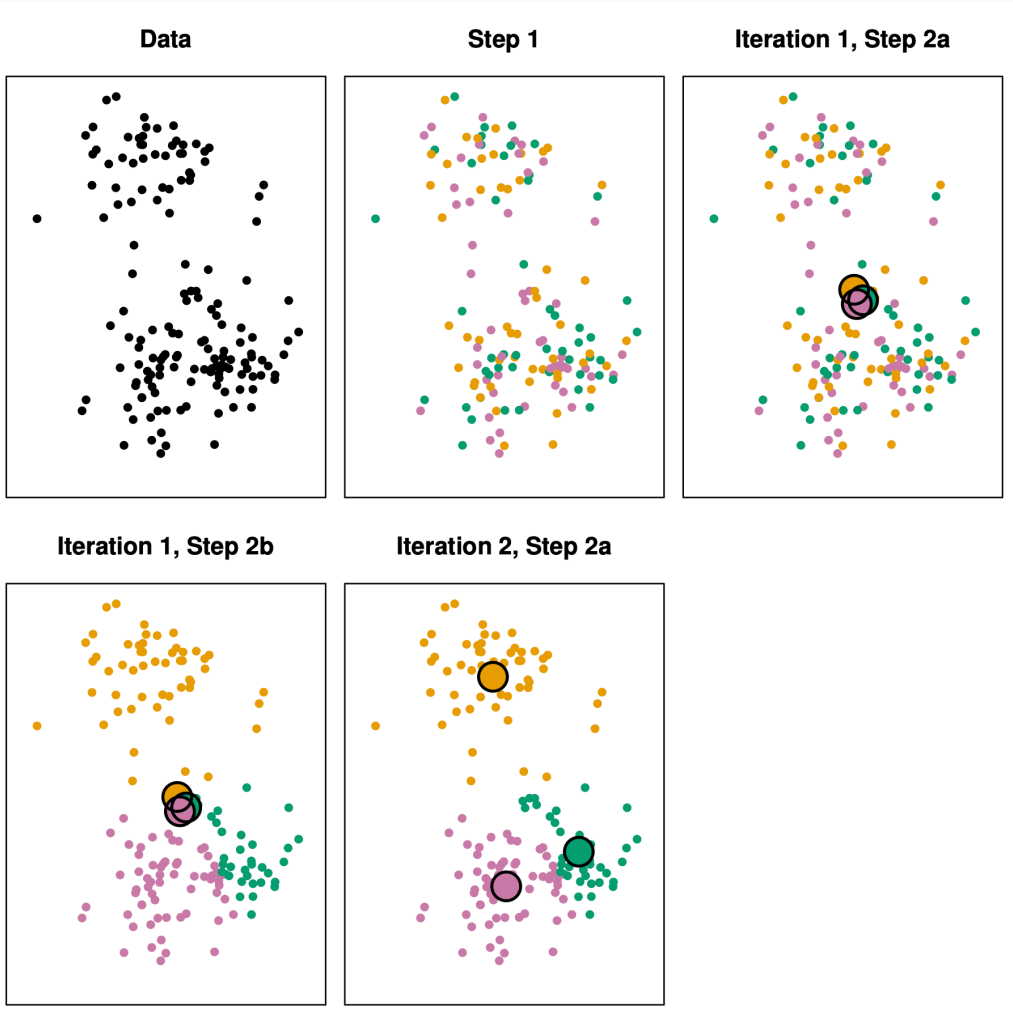
Step 2(a): the cluster centroids are computed. Initial centroids are overlapping because initial assignments were random.

K-means clustering



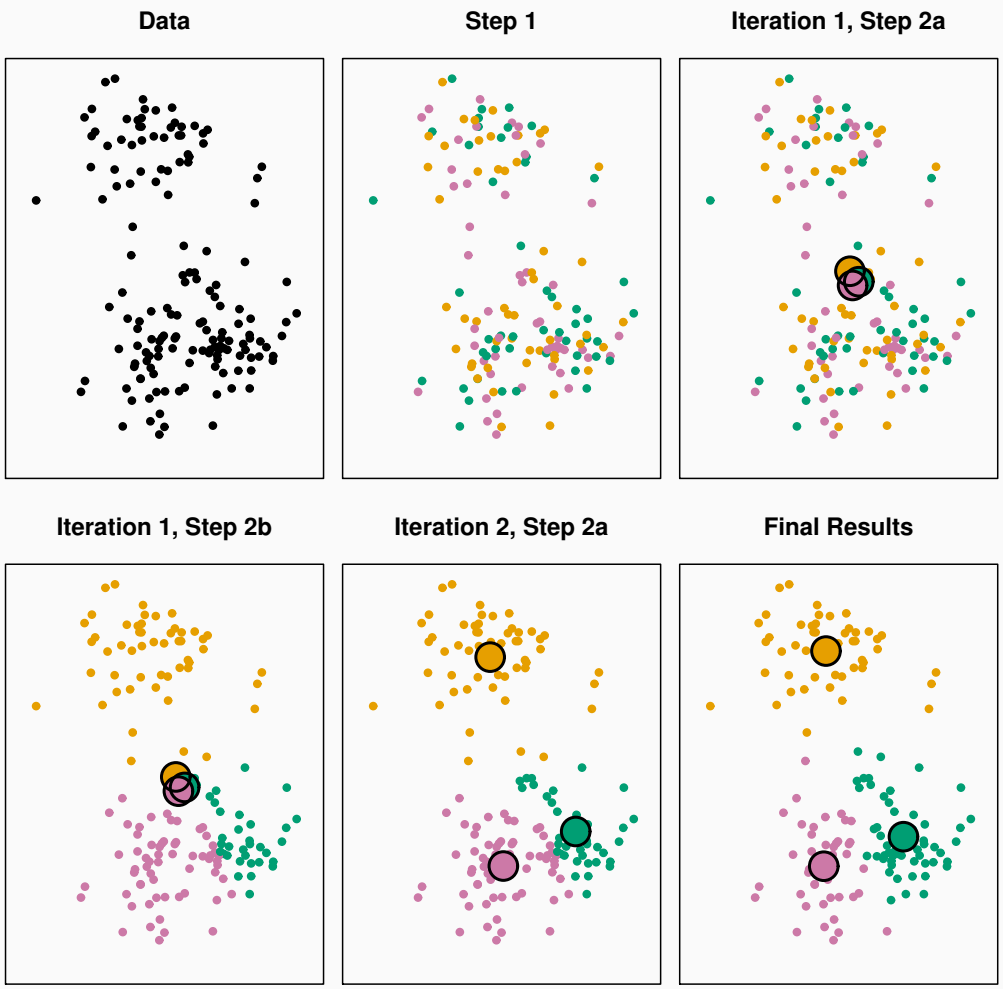
Step 2(b): each observation is assigned to the nearest centroid.

K-means clustering



Perform Step 2(a) again, leading to new cluster centroids.

K-means clustering



Results obtained after 10 iterations.

Starting Values

- The initial random starting values can be consequential for the clusters that form
- This is because there are different local optima that are reached depending on the starting point of the algorithm
- For instance, we can perform K-means clustering six times on the data from previous figure with $K = 3$, each time
- Each iteration uses a different random assignment of the observations in Step 1 of the algorithm...

Example: different starting values

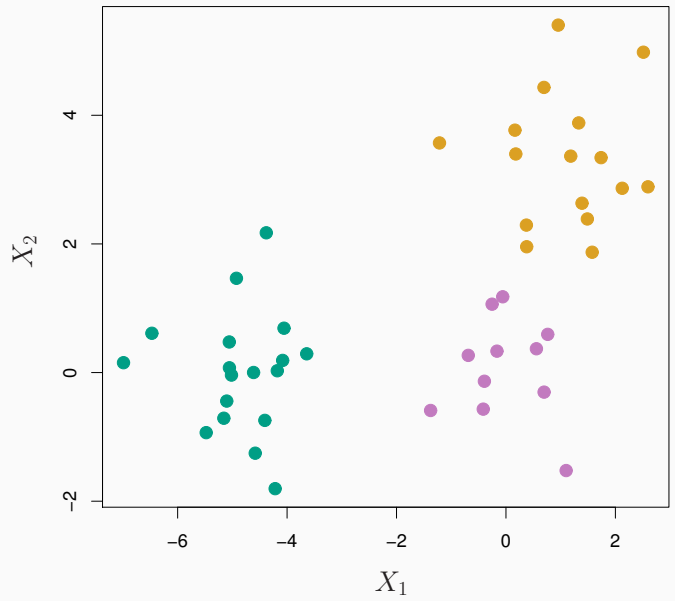
Hierarchical Clustering

- K-means clustering requires us to pre-specify the number of clusters K. This can be a disadvantage.
- Hierarchical clustering is an alternative approach which does not require that we commit to a particular choice of K.
- Here, we describe bottom-up or agglomerative clustering. This is the most common type of hierarchical clustering, and refers to the fact that a dendrogram is built starting from the leaves and combining clusters up to the trunk.

Hierarchical Clustering Algorithm

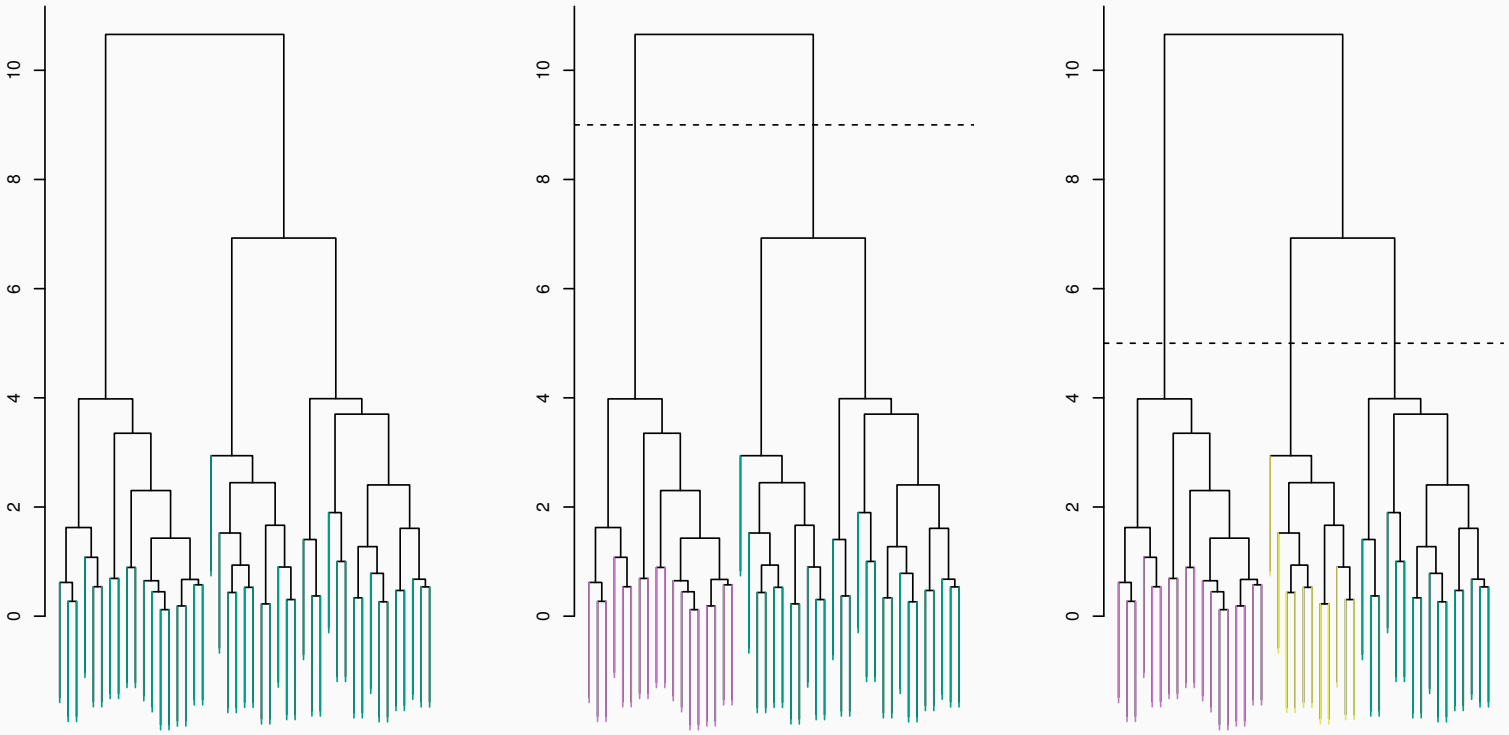
- Start with each point in its own cluster.
- Identify the closest two clusters and merge them.
- Repeat.
- Ends when all points are in a single cluster.

An Example



- 45 observations generated in 2-dimensional space.
- In reality there are three distinct classes, shown in separate colors.
- However, we will treat these class labels as unknown and will seek to cluster the observations in order to discover the classes from the data.

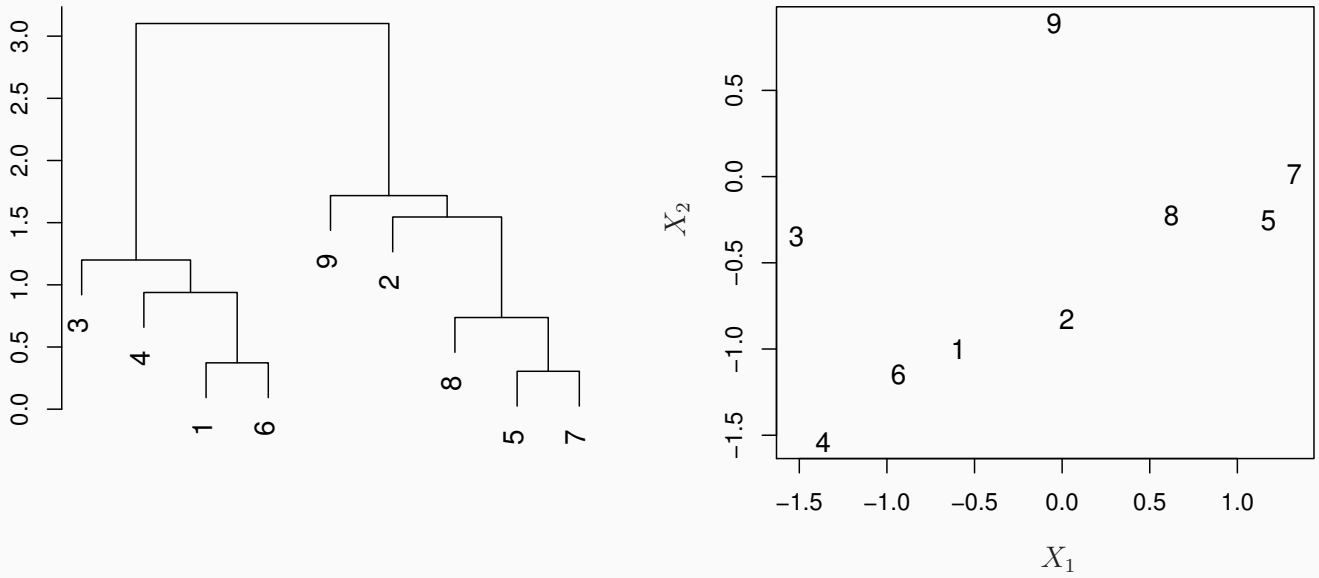
Application of hierarchical clustering



Details of previous figure

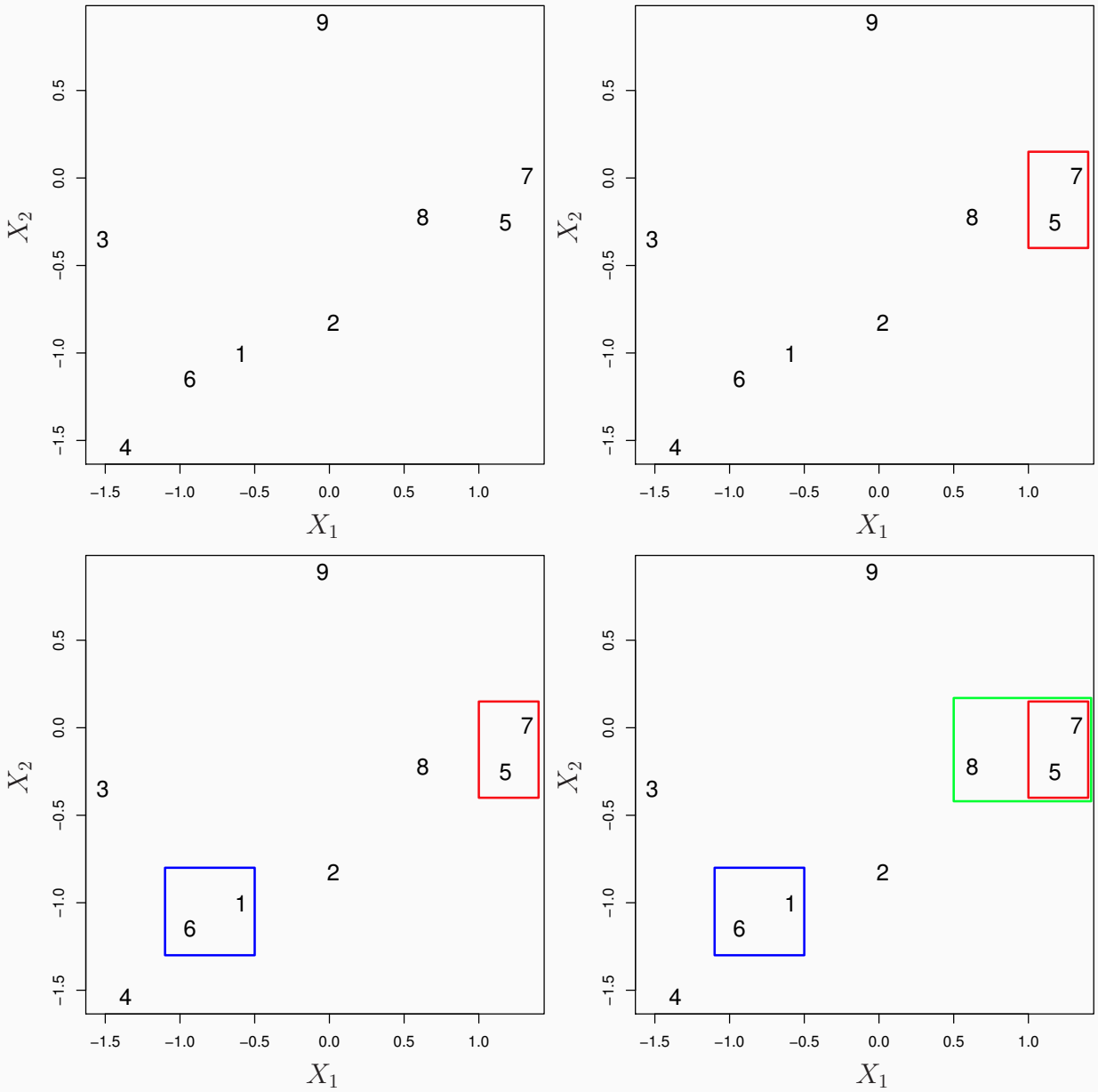
- Left: Dendrogram obtained from hierarchically clustering the data from previous slide, with complete linkage and Euclidean distance.
- Center: The dendrogram from the left-hand panel, cut at a height of 9 (indicated by the dashed line). This cut results in two distinct clusters, shown in different colors.
- Right: The dendrogram from the left-hand panel, now cut at a height of 5. This cut results in three distinct clusters, shown in different colors.

Interpreting dendrograms



- The raw data on the right was used to generate the dendrogram on the left.
- Observations 5 and 7 are quite similar to each other, as are observations 1 and 6.
- However, observation 9 is no more similar to observation 2 than it is to observations 8, 5, and 7, even though observations 9 and 2 are close together in terms of horizontal distance.
- This is because observations 2, 8, 5, and 7 all fuse with observation 9 at the same height, approximately 1.8.

Merges in previous example



Types of Linkage

Linkage	Description
Complete	Maximal inter-cluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the largest of these dissimilarities.
Single	Minimal inter-cluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the smallest of these dissimilarities.
Average	Mean inter-cluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the average of these dissimilarities.
Centroid	Dissimilarity between the centroid for cluster A (a mean vector of length p) and the centroid for cluster B.

Practical issues

- Should the observations or features first be standardized in some way? For instance, maybe the variables should be centered to have mean zero and scaled to have standard deviation one.
- In the case of hierarchical clustering,
    - What dissimilarity measure should be used?
    - What type of linkage should be used?
- How many clusters to choose? (in both K-means or hierarchical clustering). This is a difficult problem! Typically, as with PCA, this is an aspect of researcher judgment.

Example - Clustering Policy Regimes

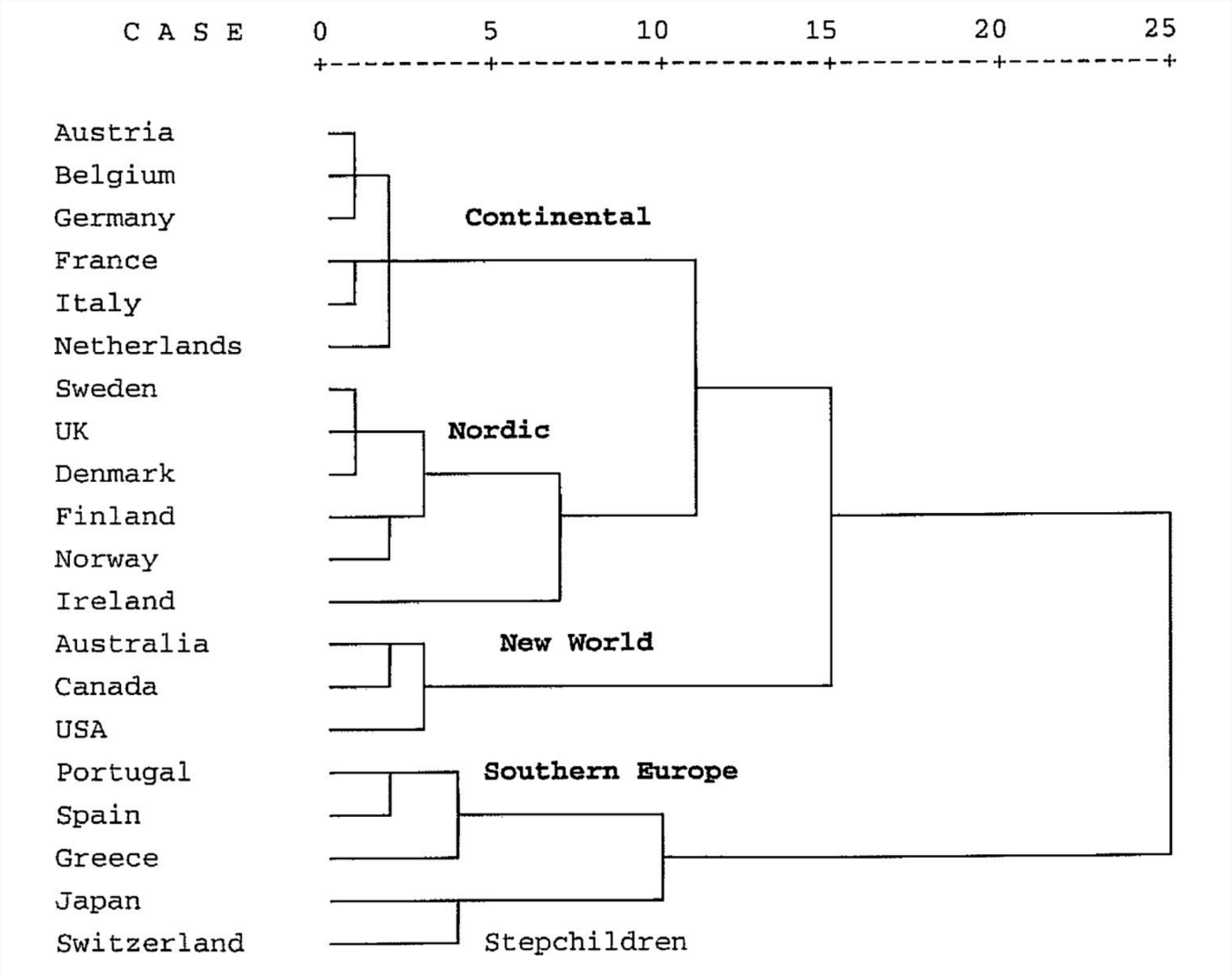
How have political clusters around the world changed over time? (Castles and Obinger, 2008, WEP)

The idea that the politics and policies of states are distinctively clustered in terms of enduring affinities is an old one. Collective proper nouns with territorially specified designations are frequently used in politics – 'English-speaking', 'Southern European', 'The West', and so on – and indicate supposed differences in the political character different groups of states. But do the *actual* political policies and outcomes of different countries cluster according to these heuristics? And how has this clustering changed over time?

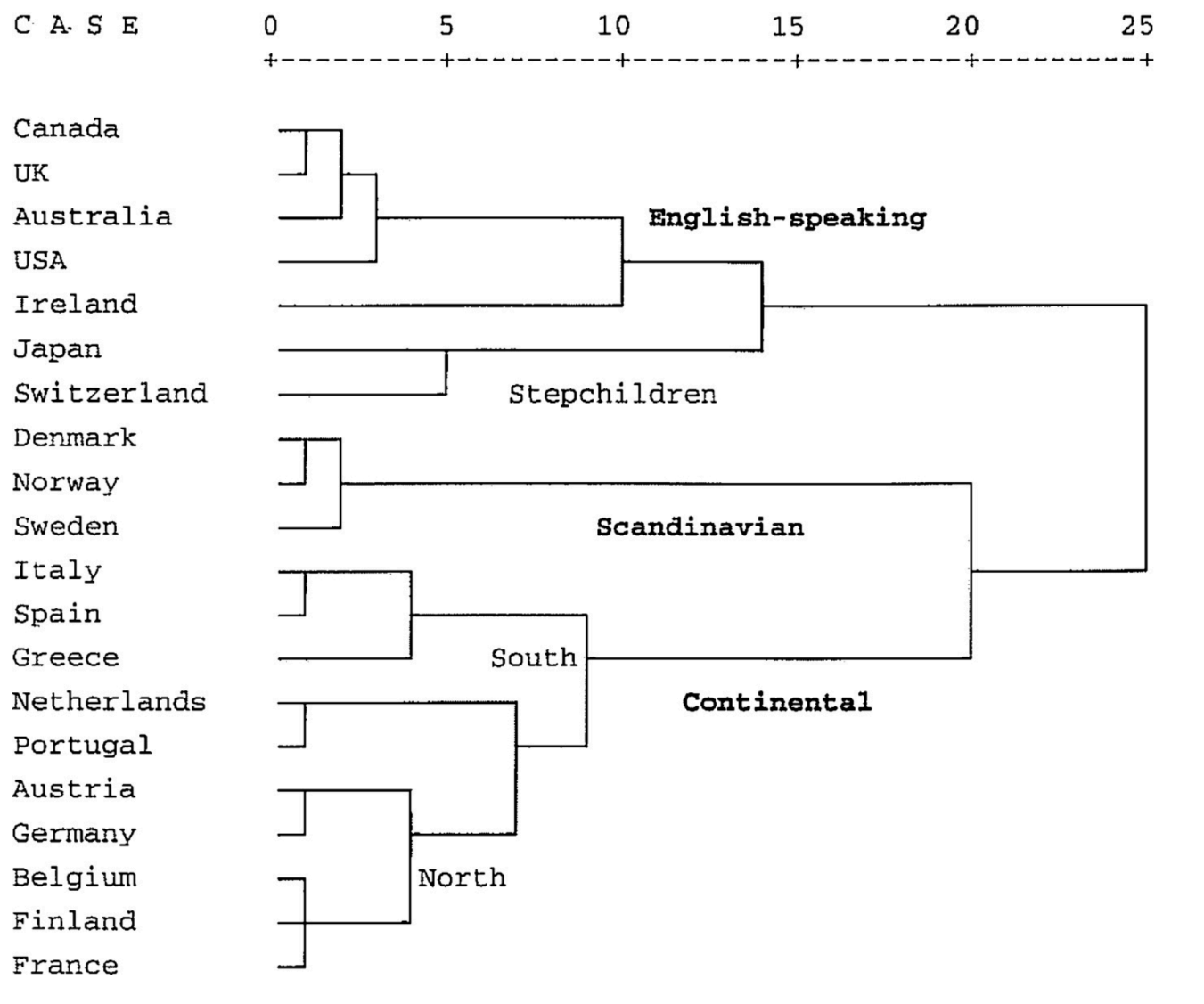
Example - Clustering Policy Regimes

- $N = 20$ OECD countries
- $X =$ a set of covariates, including:
  - size of government
  - spending priorities
  - tax rates
  - economic and labour market performance
  - gender-related policies
  - demographics
- Strategy: hierarchical clustering for 1960-75 and 2000-2004

Example - Clustering Policy Regimes (1960-75)



Example - Clustering Policy Regimes (2000-2004)



Conclusions

- Unsupervised learning is important for understanding the variation and grouping structure of a set of unlabeled data, and can be a useful pre-processor for supervised learning.
- It is intrinsically more difficult than supervised learning because there is no gold standard (like an outcome variable) and no single objective (like test set accuracy).
- It is an active field of research, with many recently developed tools .
- We will speak about more unsupervised models when we discuss topic models and word-embedding models for text data.