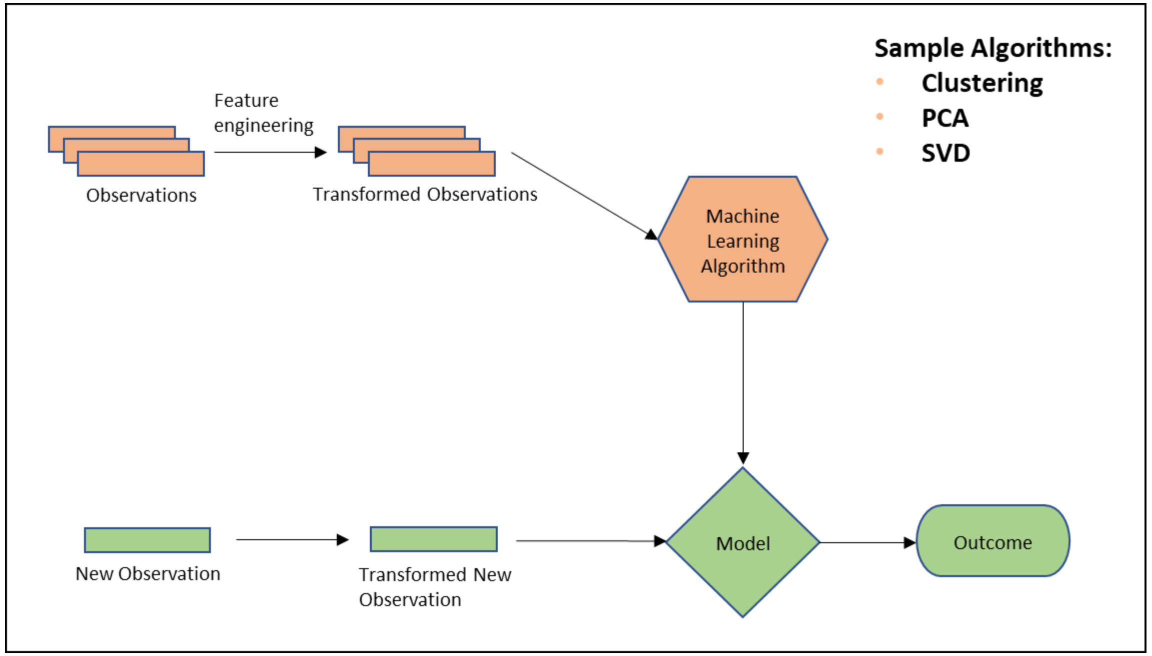
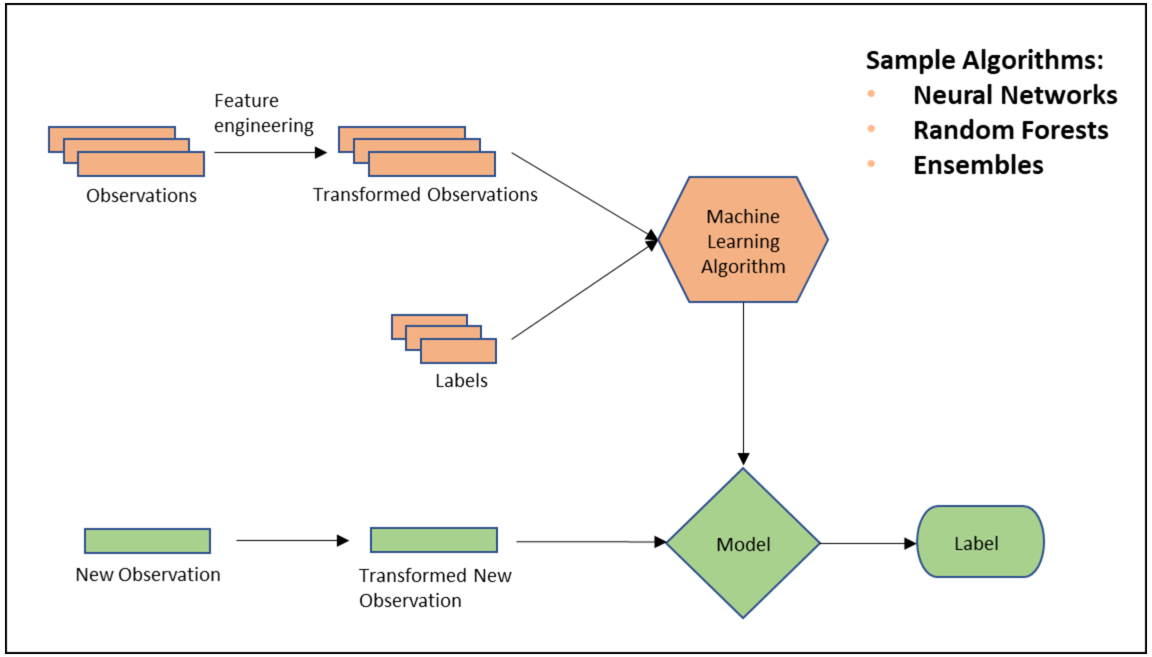
Machine Learning Live Session #7

Supervised vs. Unsupervised Learning

- Unsupervised learning: no labels associated with observations
  - Try to infer relationships between the observations or between the features
  - Useful for data visualization and dimension reduction
- Supervised learning: each observation is associated with a label
  - Try to infer a relationship between the features and labels
    - The label acts as a teacher that supervises the learning process
  - Use the relationship to predict label for a new observation

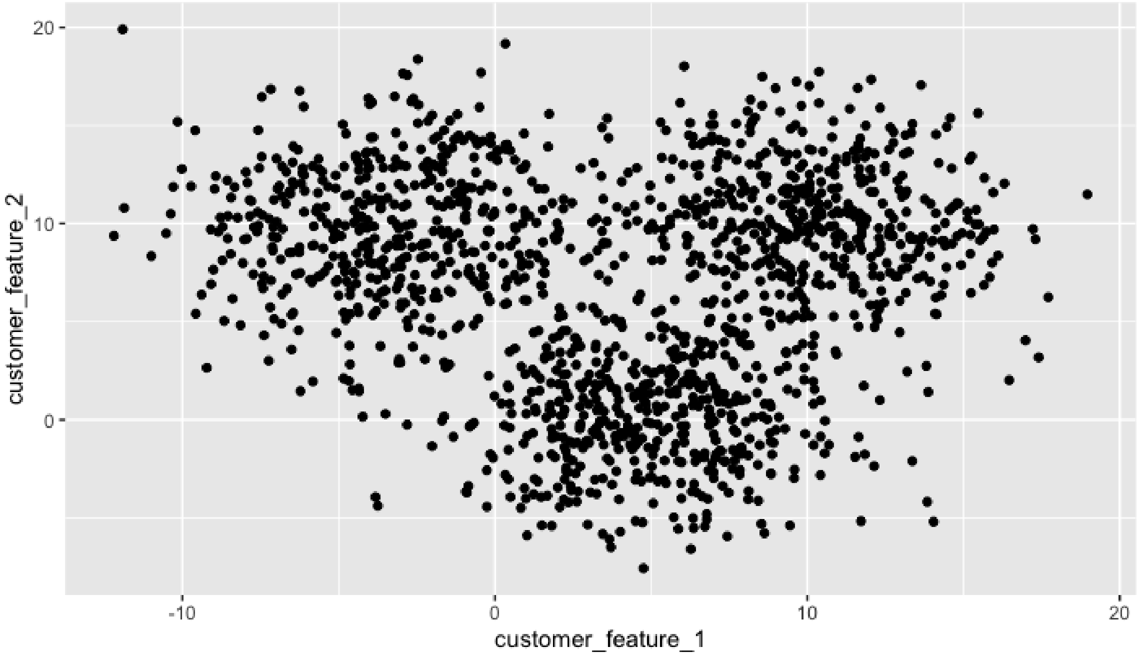


Unsupervised Learning



Supervised Learning

Supervised vs. Unsupervised Learning



Unsupervised Learning

- Observations (here, customers) have no labels
- Use unsupervised learning to explore and learn about customers
  - E.g., do sub-groups of customers exist, with each sub-group exhibiting similar characteristics?
  - This is called customer segmentation

Supervised Learning

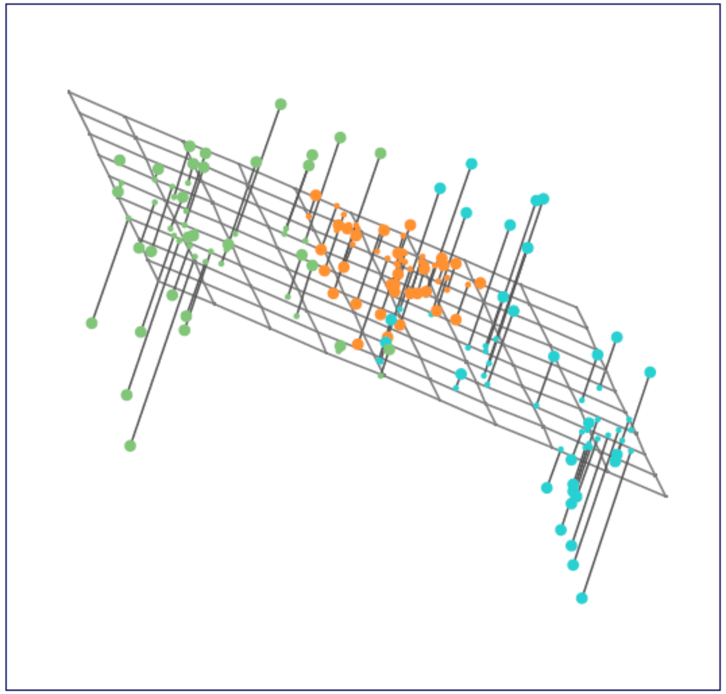
- Each observation (here, customer) is associated with a label
  - E.g., whether the customer left or stayed
- Use supervised learning to predict the label for new customers

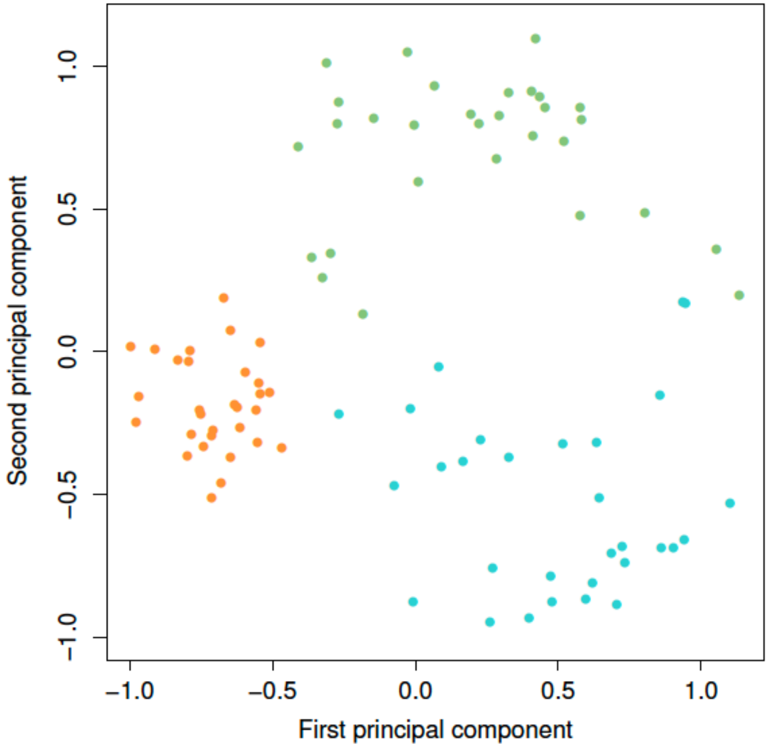
Unsupervised Learning

- Unsupervised learning methods try to infer relationships between features or between observations
  - More subjective than supervised learning, since there is no clear objective
- Unsupervised learning is an important step in the machine learning process
  - Exploring and visualizing the data
  - Dimension reduction
- Unsupervised learning methods discussed in this course are:
  - Principal component analysis (PCA)
  - k-means clustering
  - Hierarchical clustering
- **Important: these methods are intended for numerical features only**

Principal Component Analysis

- In PCA, we try to find a low-dimensional representation of the data
  - If this low-dimensional representation is close enough to the data:
    - Use the low-dimensional representation to decrease the dimension of the problem
    - Visualize the data (not the focus here)





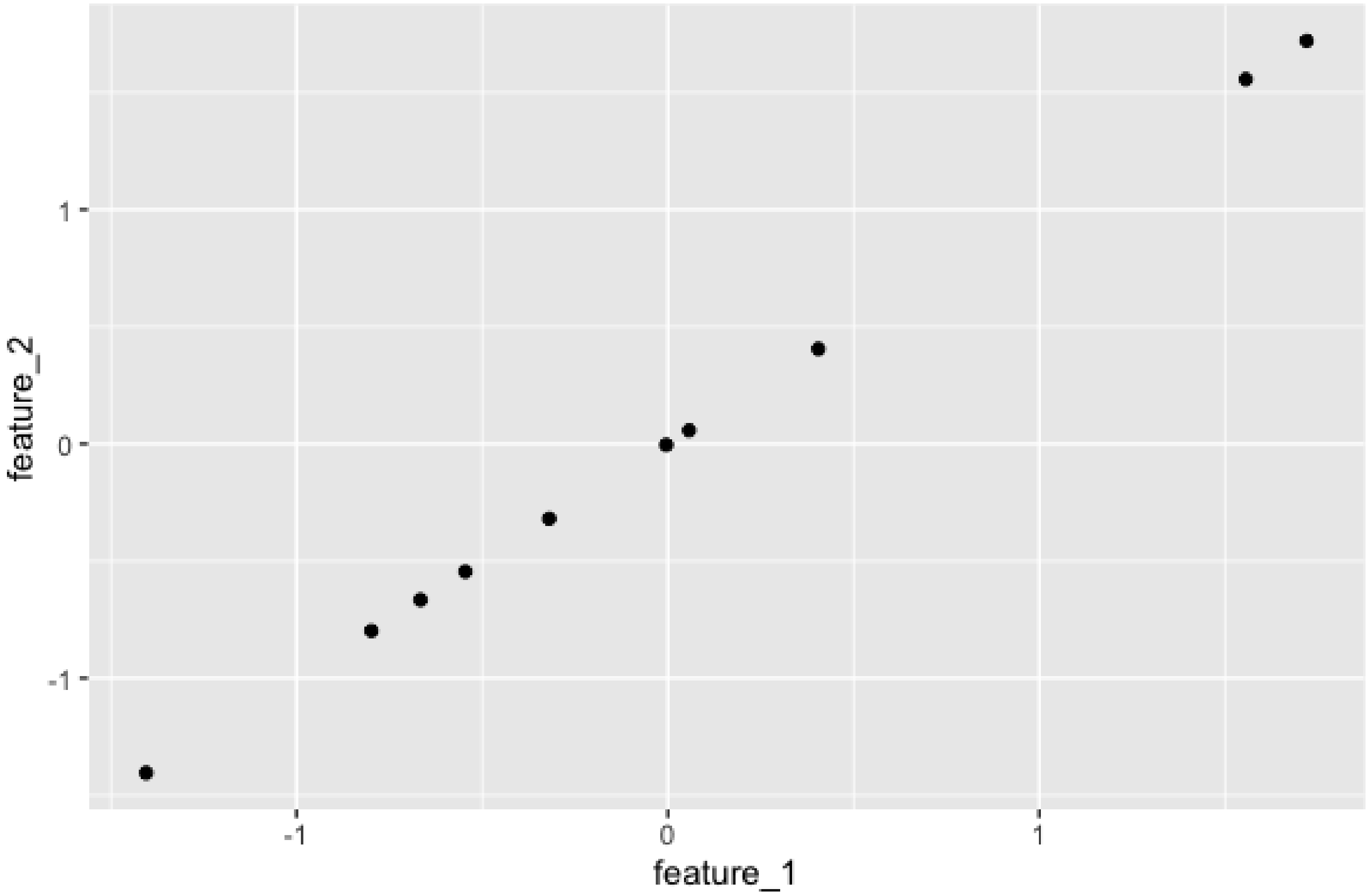
James, G., Witten, D., Hastie, T., & Tibshirani, R. (2017) *An Introduction to Statistical Learning with Applications in R*. Springer

Principal Component Analysis

- What does it mean to "find a low-dimensional representation of the data"?

feature_1	feature_2
-0.67	-0.67
-0.32	-0.32
1.56	1.56
0.0	0.0
0.06	0.06
1.72	1.72
0.41	0.41
-1.4	-1.4
-0.8	-0.8
-0.55	-0.55

Observations

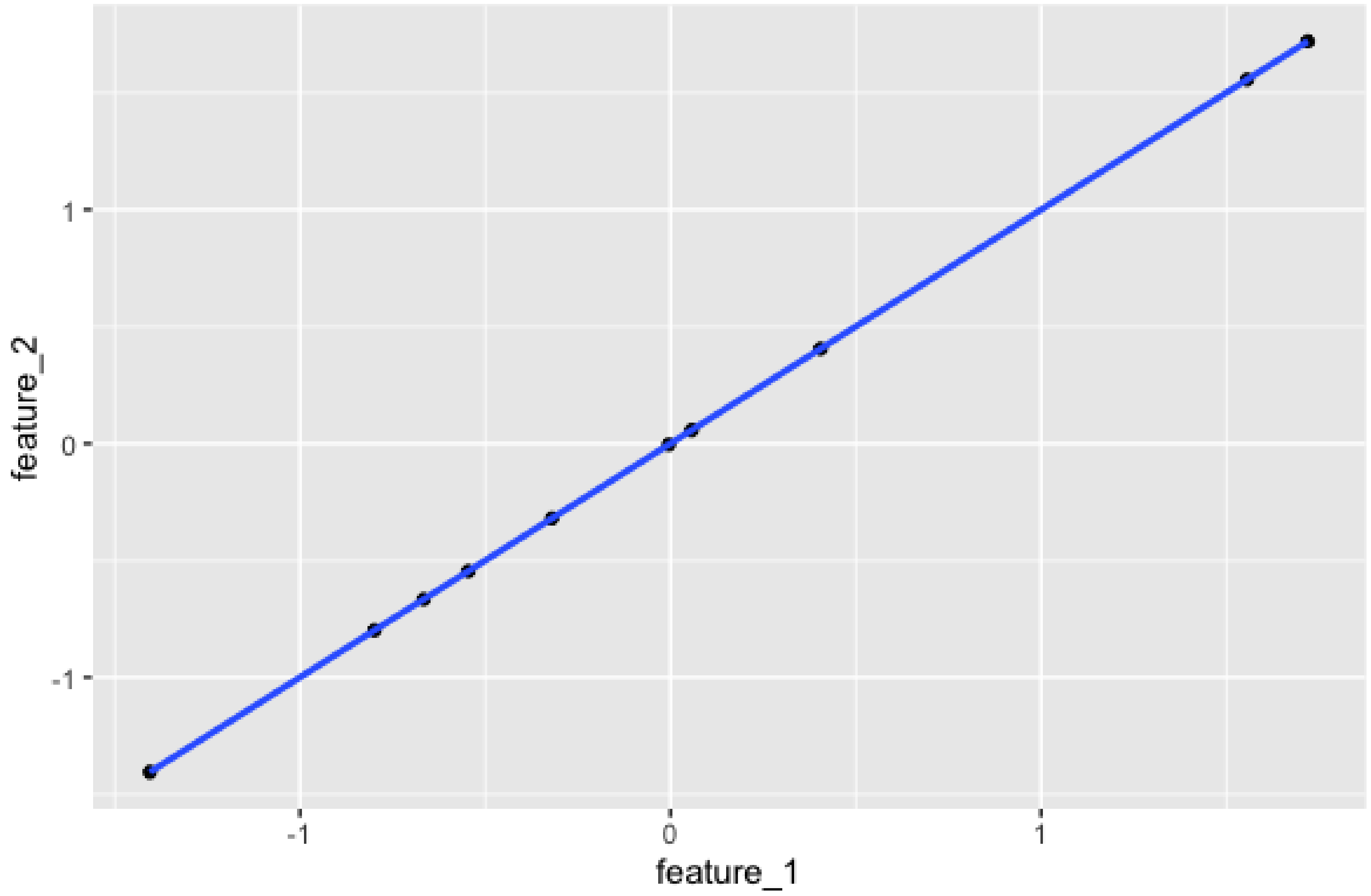


Principal Component Analysis

- Consider an extreme example: the observations fall along a one-dimensional line

feature_1	feature_2
-0.67	-0.67
-0.32	-0.32
1.56	1.56
0.0	0.0
0.06	0.06
1.72	1.72
0.41	0.41
-1.4	-1.4
-0.8	-0.8
-0.55	-0.55

Observations



Principal Component Analysis

- Write the observations in terms of the unit vector in the direction of the one-dimensional line
  - The unit vector is called a principal component (PC)

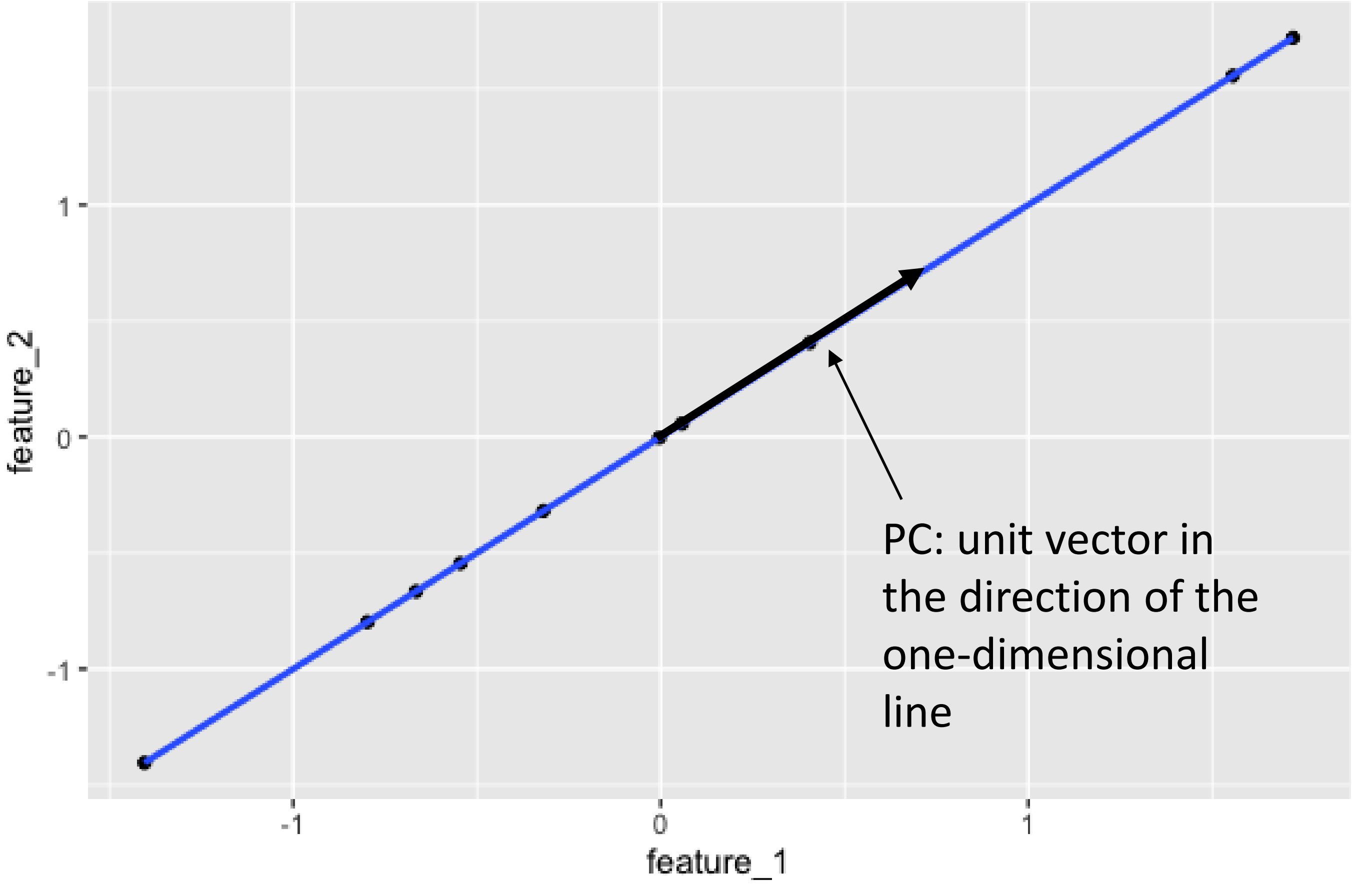
feature_1	feature_2
-0.67	-0.67
-0.32	-0.32
1.56	1.56
0.0	0.0
0.06	0.06
1.72	1.72
0.41	0.41
-1.4	-1.4
-0.8	-0.8
-0.55	-0.55

Observations

→

PC

Observations in terms of PC



Principal Component Analysis

- Write the observations in terms of the unit vector in the direction of the one-dimensional line
  - The unit vector is called a principal component (PC)

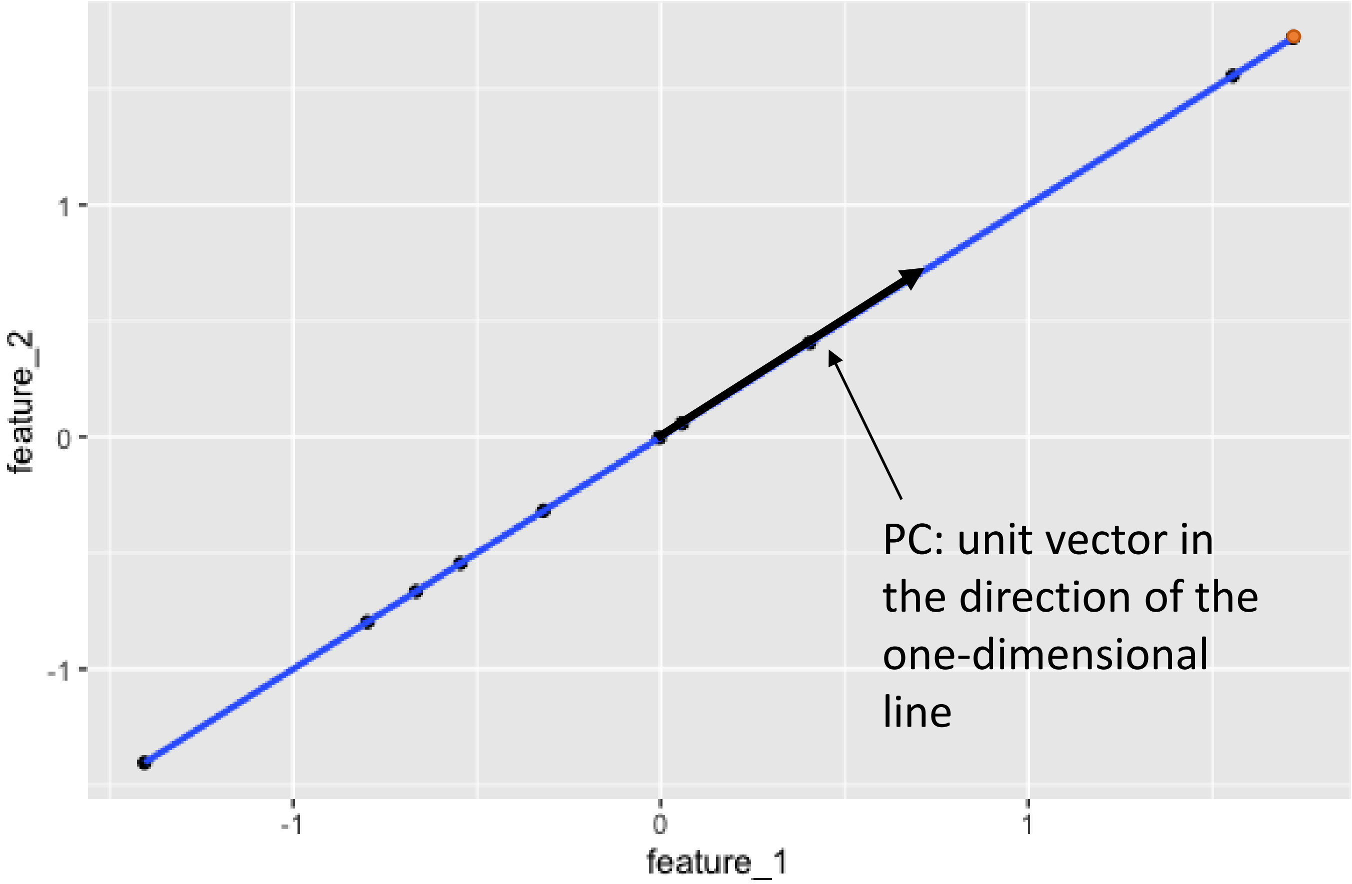
feature_1	feature_2
-0.67	-0.67
-0.32	-0.32
1.56	1.56
0.0	0.0
0.06	0.06
1.72	1.72
0.41	0.41
-1.4	-1.4
-0.8	-0.8
-0.55	-0.55

Observations

→

PC

Observations in terms of PC



Principal Component Analysis

- Write the observations in terms of the unit vector in the direction of the one-dimensional line
  - The unit vector is called a principal component (PC)

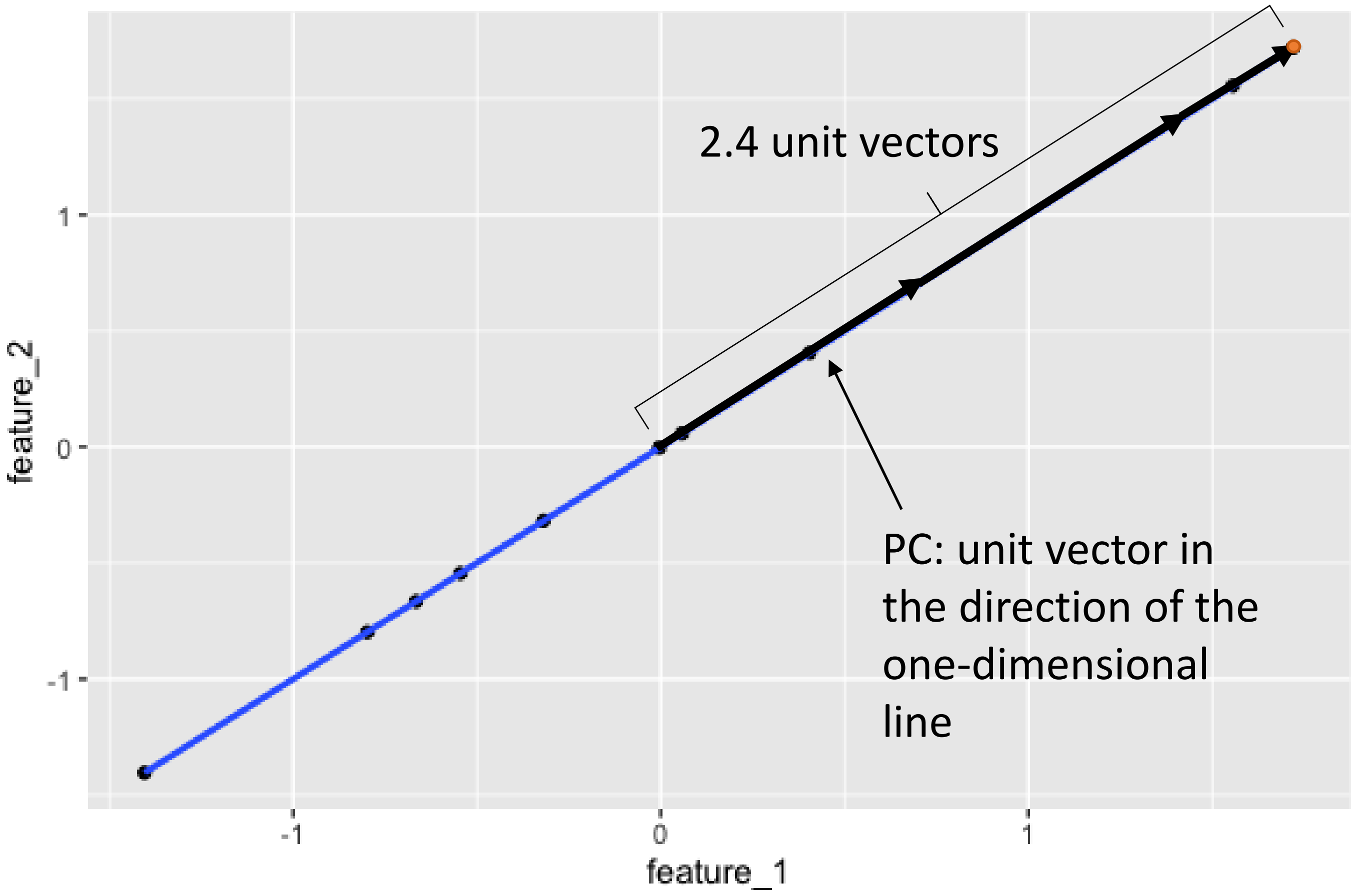
feature_1	feature_2
-0.67	-0.67
-0.32	-0.32
1.56	1.56
0.0	0.0
0.06	0.06
1.72	1.72
0.41	0.41
-1.4	-1.4
-0.8	-0.8
-0.55	-0.55

Observations

→

PC
2.4

Observations in terms of PC



Principal Component Analysis

- Write the observations in terms of the unit vector in the direction of the one-dimensional line
  - The unit vector is called a principal component (PC)

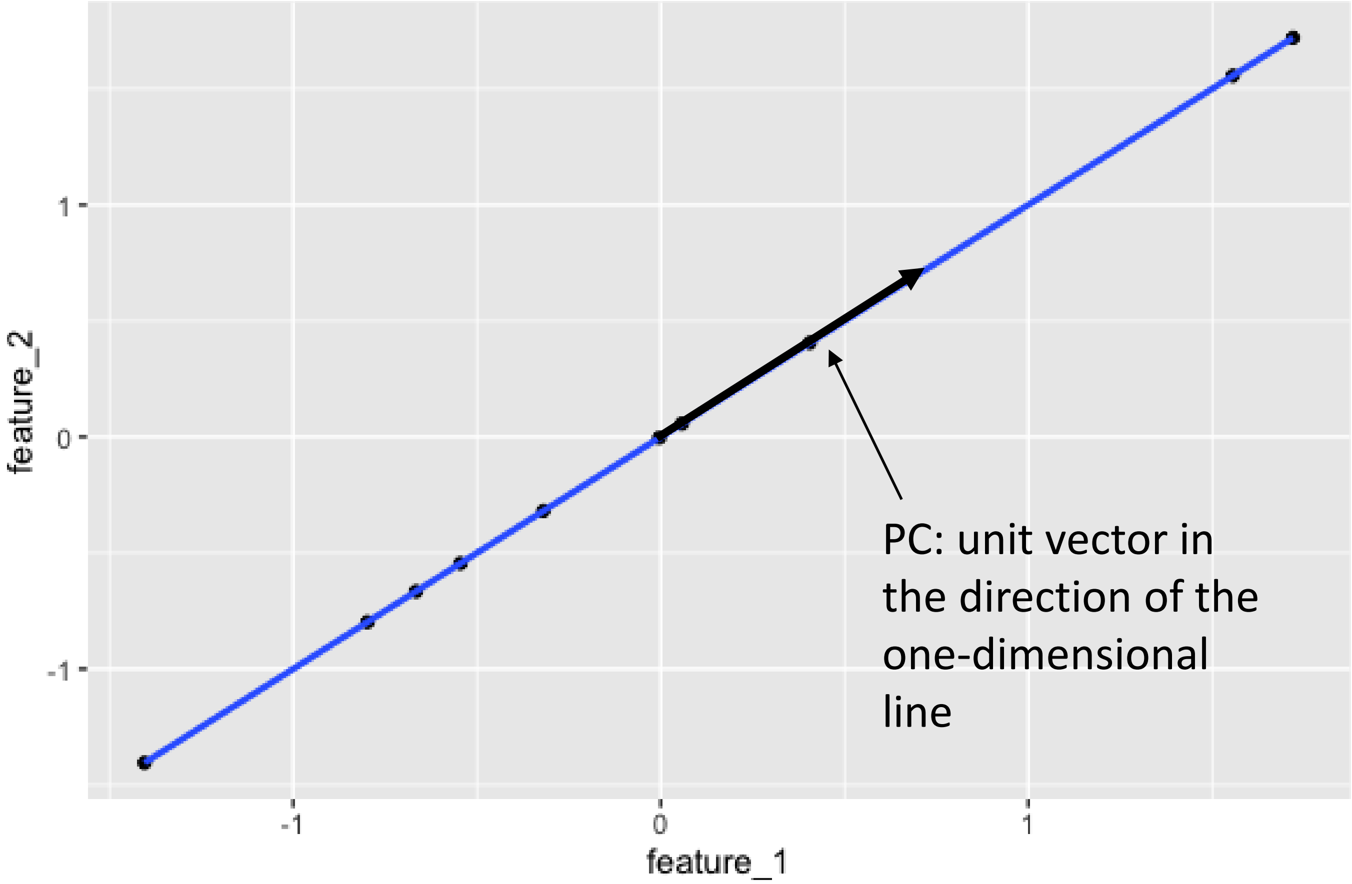
feature_1	feature_2
-0.67	-0.67
-0.32	-0.32
1.56	1.56
0.0	0.0
0.06	0.06
1.72	1.72
0.41	0.41
-1.4	-1.4
-0.8	-0.8
-0.55	-0.55

Observations

→

PC
-0.9
-0.5
2.2
0.0
0.1
2.4
0.6
-2.0
-1.1
-0.8

Observations in terms of PC

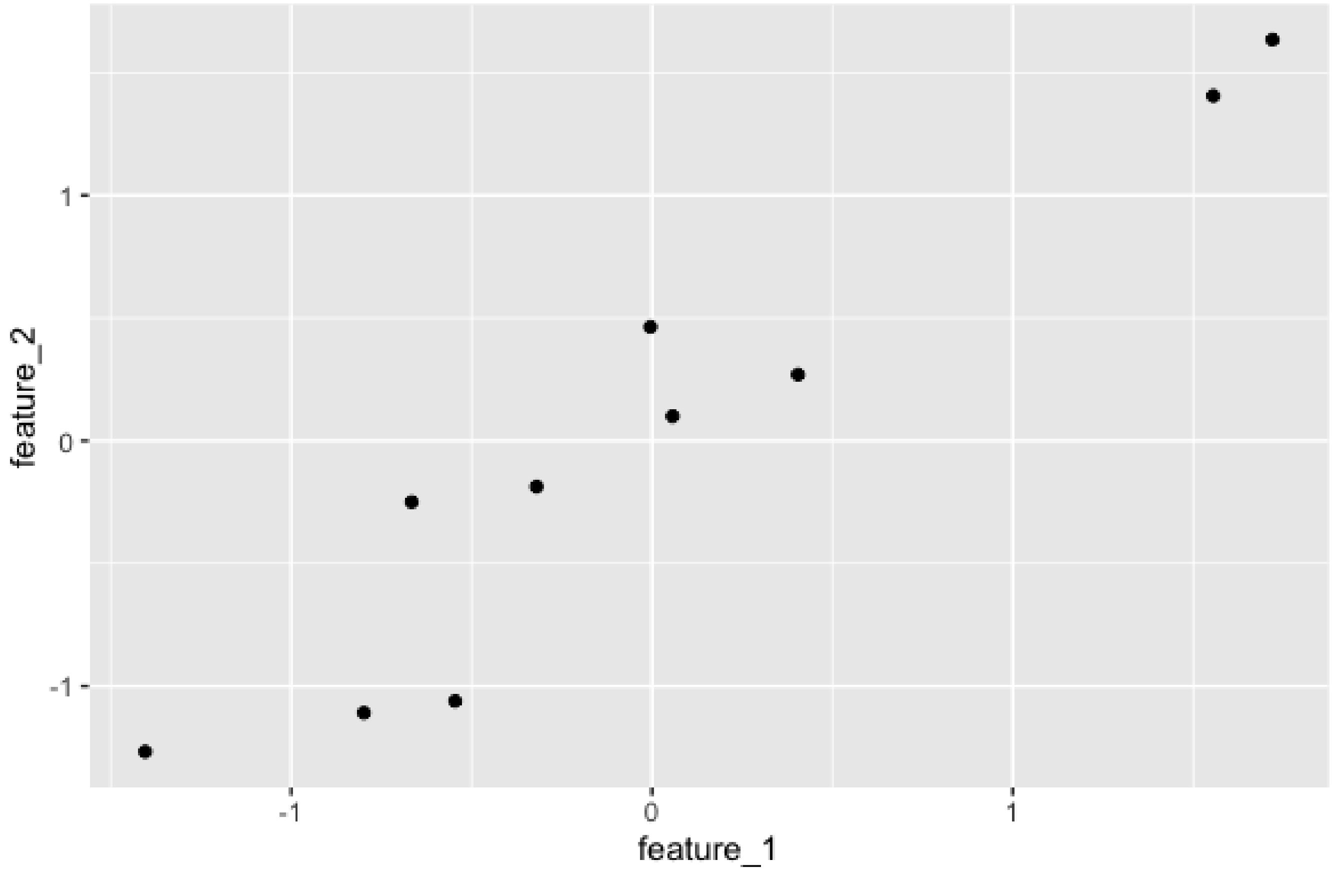


Principal Component Analysis

- What if the observations don't necessarily fall along a one-dimensional line, but are close?

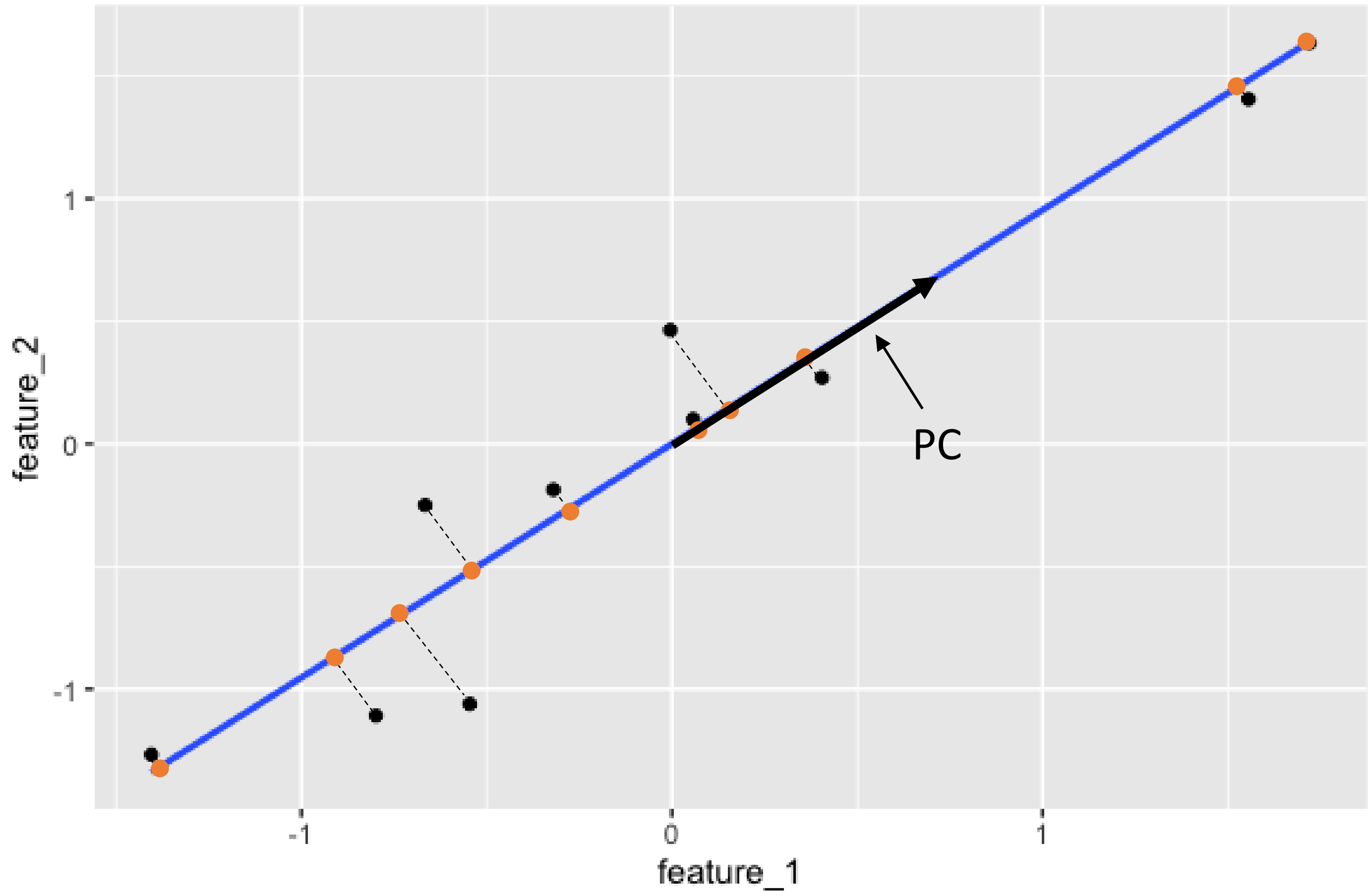
feature_1	feature_2
-0.67	-0.25
-0.32	-0.19
1.56	1.41
0.0	0.46
0.06	0.1
1.72	1.63
0.41	0.27
-1.4	-1.27
-0.8	-1.11
-0.55	-1.06

Observations



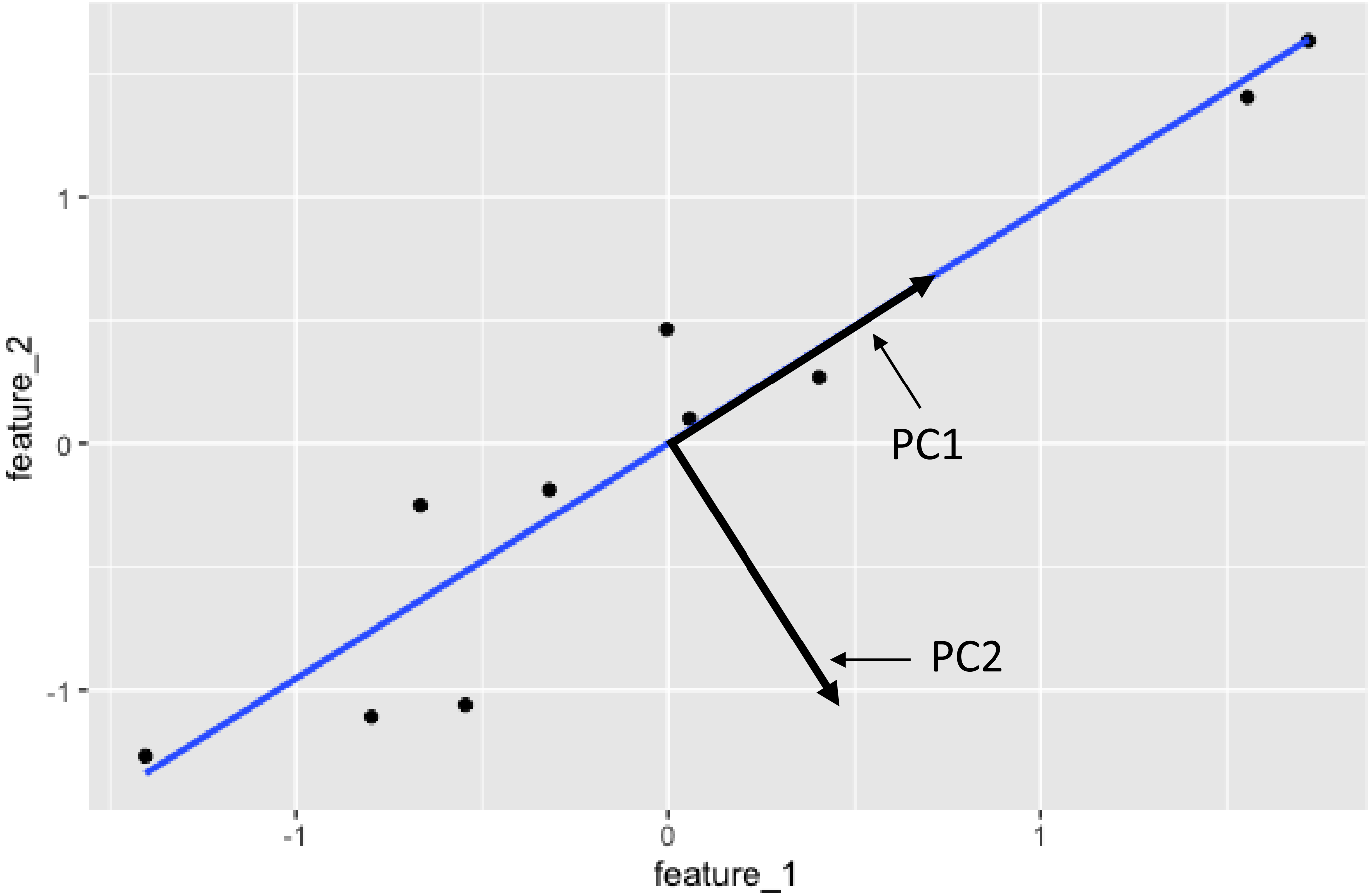
Principal Component Analysis

- Let PC be the unit vector in the direction the data varies most
- Project each observation onto the one-dimensional line spanned by PC
- Not able to completely express each observation only using PC
  - There is some error, which is the distance between the projection and the original observation
  - Need another unit vector in the direction orthogonal to PC



Principal Component Analysis

- Now, let PC1 be the unit vector in the direction the data varies most and PC2 be the unit vector in the direction orthogonal to PC1
- Write the observations in terms of PC1 and PC2



Principal Component Analysis

- Now, let PC1 be the unit vector in the direction the data varies most and PC2 be the unit vector in the direction orthogonal to PC1
- Write the observations in terms of PC1 and PC2

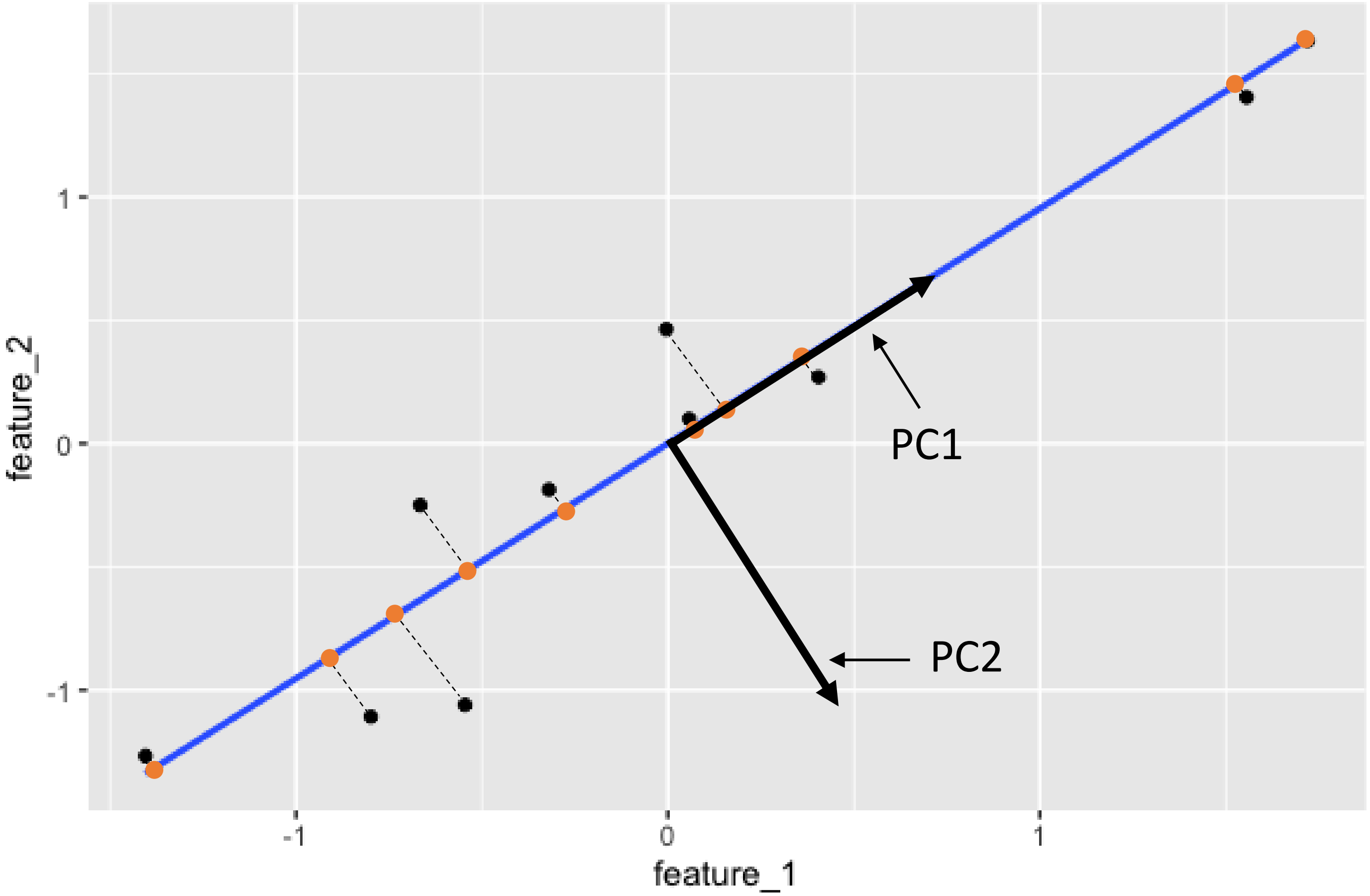
feature_1	feature_2
-0.67	-0.25
-0.32	-0.19
1.56	1.41
0.0	0.46
0.06	0.1
1.72	1.63
0.41	0.27
-1.4	-1.27
-0.8	-1.11
-0.55	-1.06

Observations

→

PC1	PC2
-0.65	-0.29
-0.36	-0.09
2.09	0.11
0.32	-0.33
0.11	-0.03
2.37	0.06
0.48	0.1
-1.89	-0.1
-1.35	0.22
-1.14	0.36

Observations in terms of PC1 and PC2



Principal Component Analysis

- If PC1 is a good enough approximation to the observations → try using only PC1 to represent the data

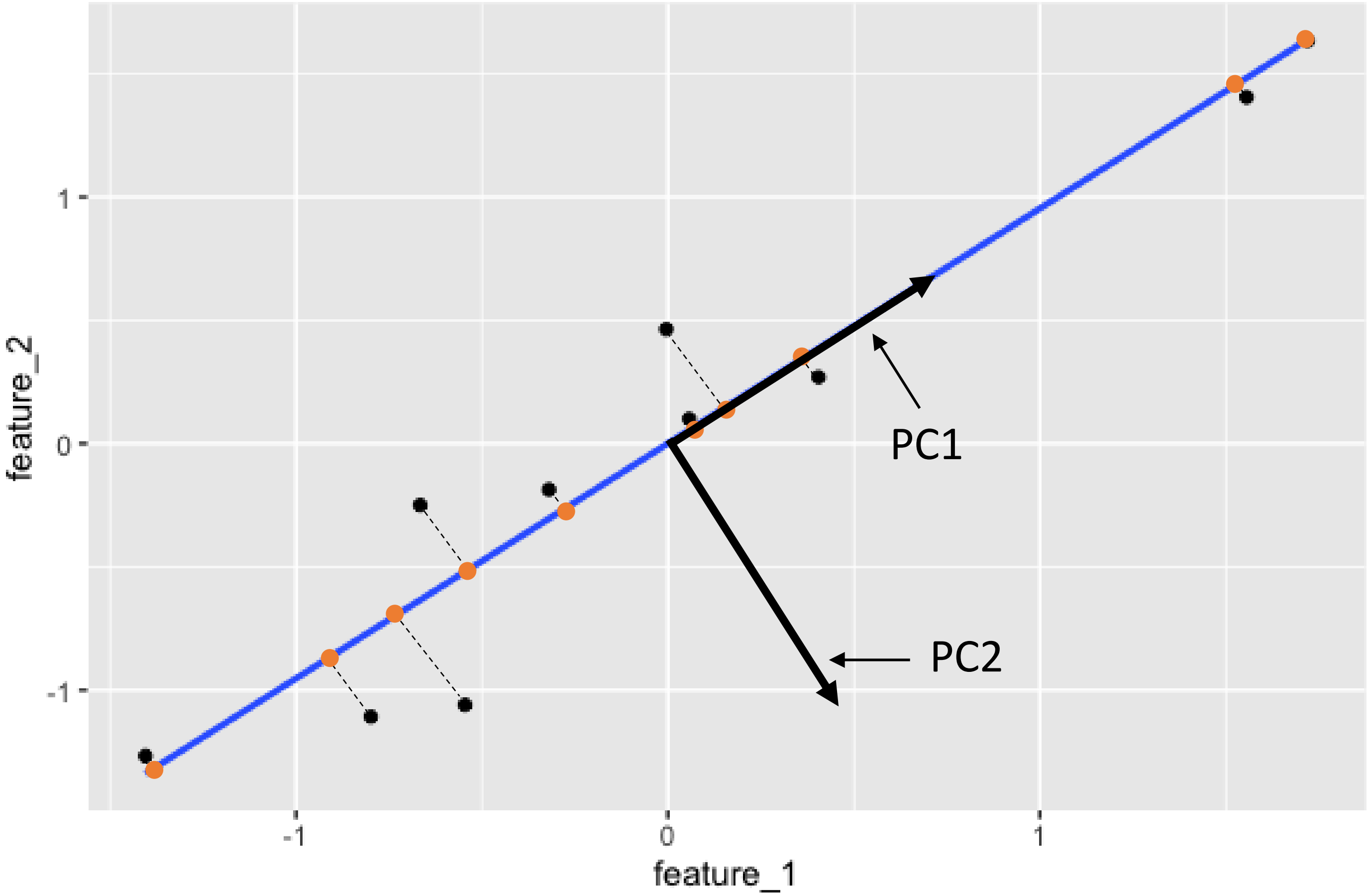
feature_1	feature_2
-0.67	-0.25
-0.32	-0.19
1.56	1.41
0.0	0.46
0.06	0.1
1.72	1.63
0.41	0.27
-1.4	-1.27
-0.8	-1.11
-0.55	-1.06

Observations

$\approx$

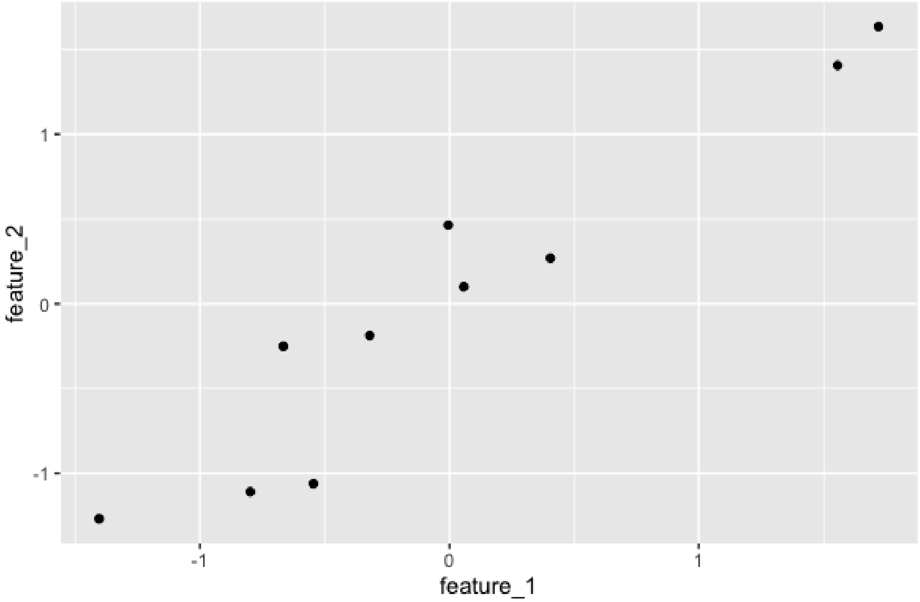
PC1
-0.65
-0.36
2.09
0.32
0.11
2.37
0.48
-1.89
-1.35
-1.14

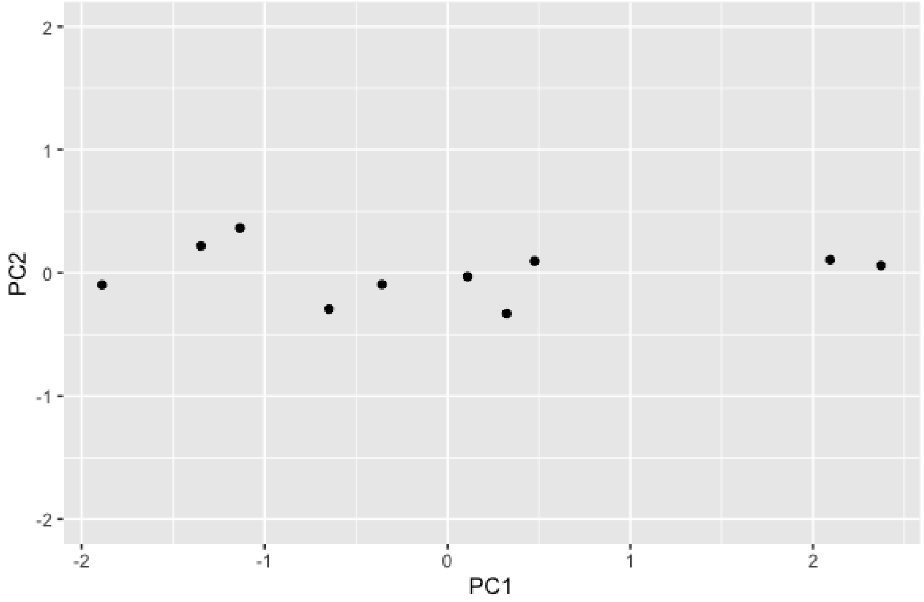
Observations in terms of PC1 only



Principal Component Analysis

- Writing the original observations in terms of the PCs is equivalent to a rotation of the observations





The rotation component of the PCA object is used to perform this rotation in R (see the Module #7 Lab)

```
> pca_results$rotation
                PC1        PC2
feature_1 0.7071068  0.7071068
feature_2 0.7071068 -0.7071068
```

The rotation component of the PCA object also tells us which features have the largest influence on each PC

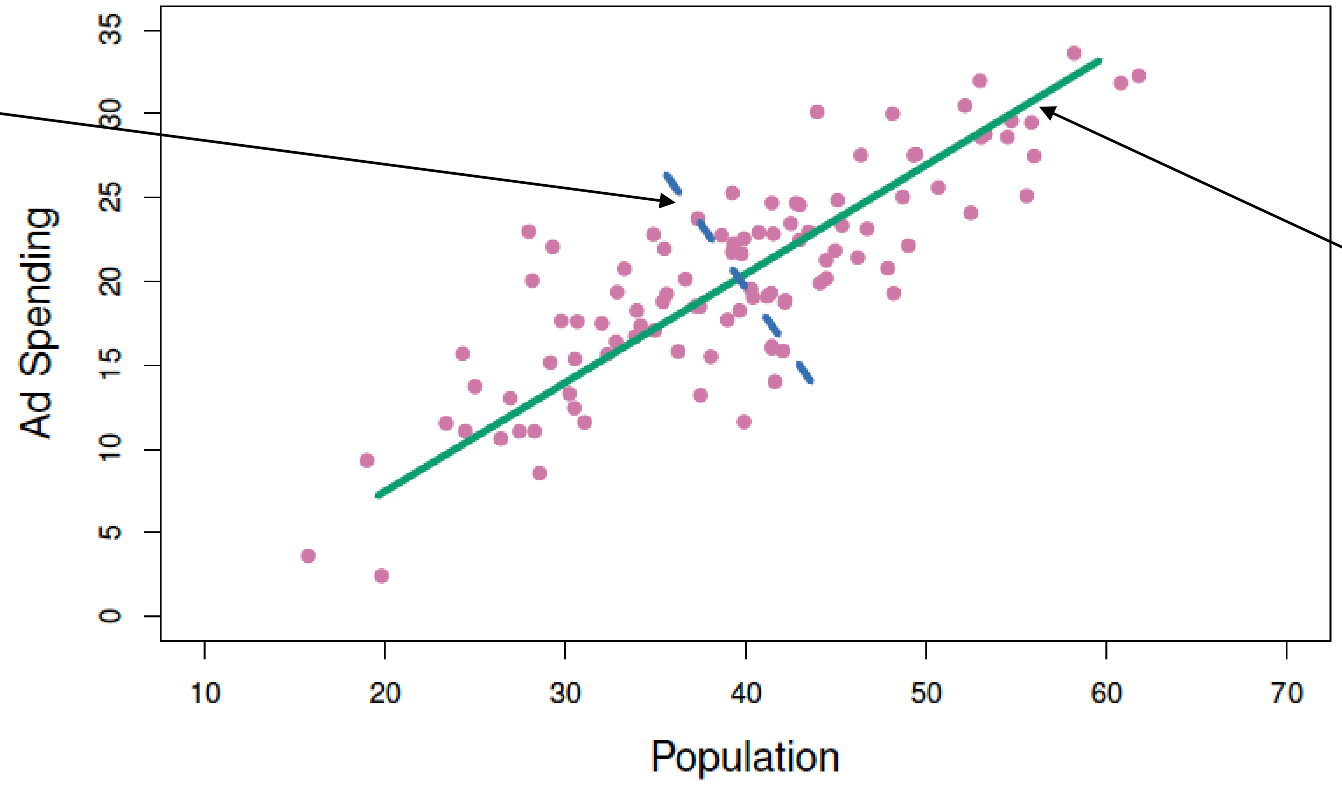
Within each PC column, the absolute value is the amount of influence the feature had on the PC. Here, each feature has the same amount of influence on PC1 and PC2

Principal Component Analysis

- Find the directions (unit vectors) in which the data varies most
  - Directions are found sequentially, with each direction being orthogonal to all previous directions
  - Directions are the PCs

Second direction, called the second PC.

If we project the data onto this direction, it has the largest variance out of all possible directions that are orthogonal to the first direction



First direction, called the first PC.

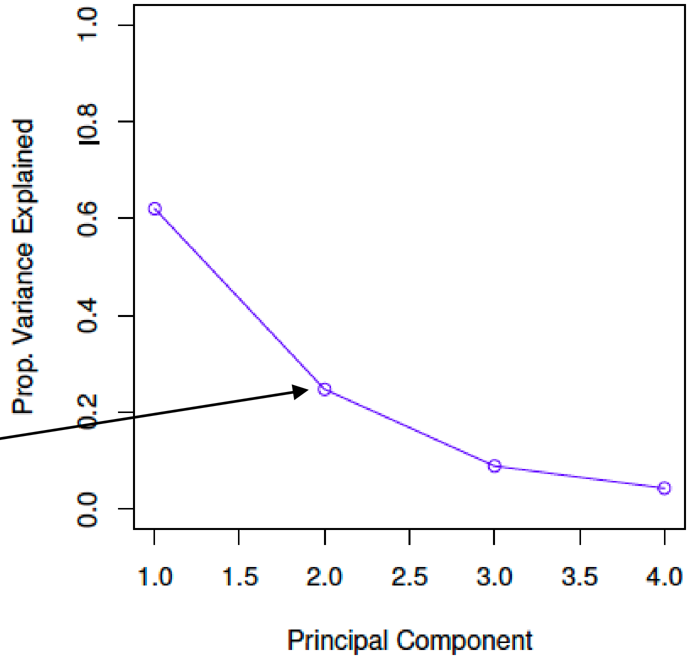
If we project the data onto this direction, it has the largest variance out of all possible directions

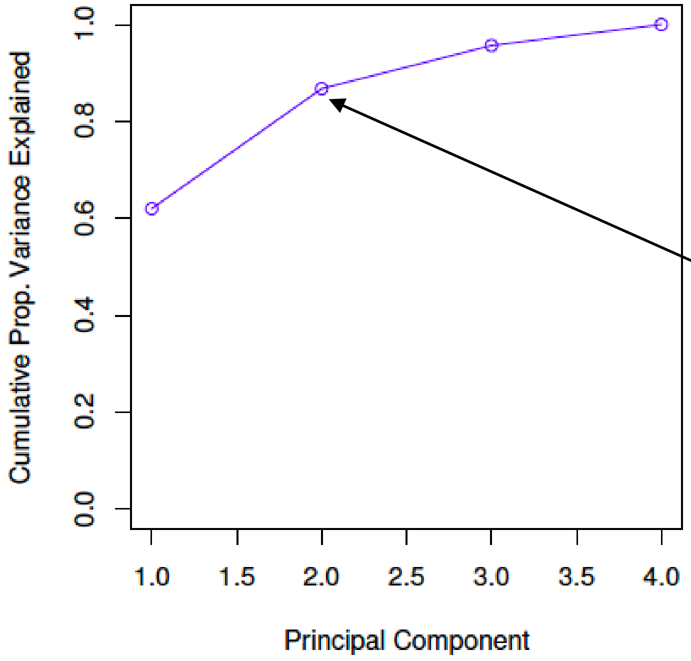
James, G., Witten, D., Hastie, T., & Tibshirani, R. (2017) *An Introduction to Statistical Learning with Applications in R*. Springer

Principal Component Analysis

- For a dataset with d numerical features, there are d PCs
- Amount of variation explained by the PCs is decreasing, i.e., the first PC explains the most variation in the data and the last PC explains the least
- How do we decide how many PCs are enough?

Plot of the proportion of variance explained vs the principal component number. Look for the elbow.





Plot of the cumulative proportion of variance explained vs the number of principal components. Look for the elbow.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2017) *An Introduction to Statistical Learning with Applications in R*. Springer

Principal Component Analysis

- PCA can be used for dimension reduction
  - If the first p PCs explain a large amount of the variation in the data
    - Project the data onto the subspace spanned by these p PCs
    - Use this p-dimensional data instead of the original numerical data
- Numerical features should be centered and scaled before applying PCA
  - Centering is **ALWAYS** required since PCs are unit vectors originating from the origin
  - Scaling is **ALMOST ALWAYS** required to ensure each numerical feature has equal weight
    - Otherwise, PCA results in finding the features with the most variance
    - Special case: if all the numerical features are measured on similar scales AND we want features with larger variances to have more importance, then scaling may not be needed