Data Analysis and Visualization in R (IN2339)

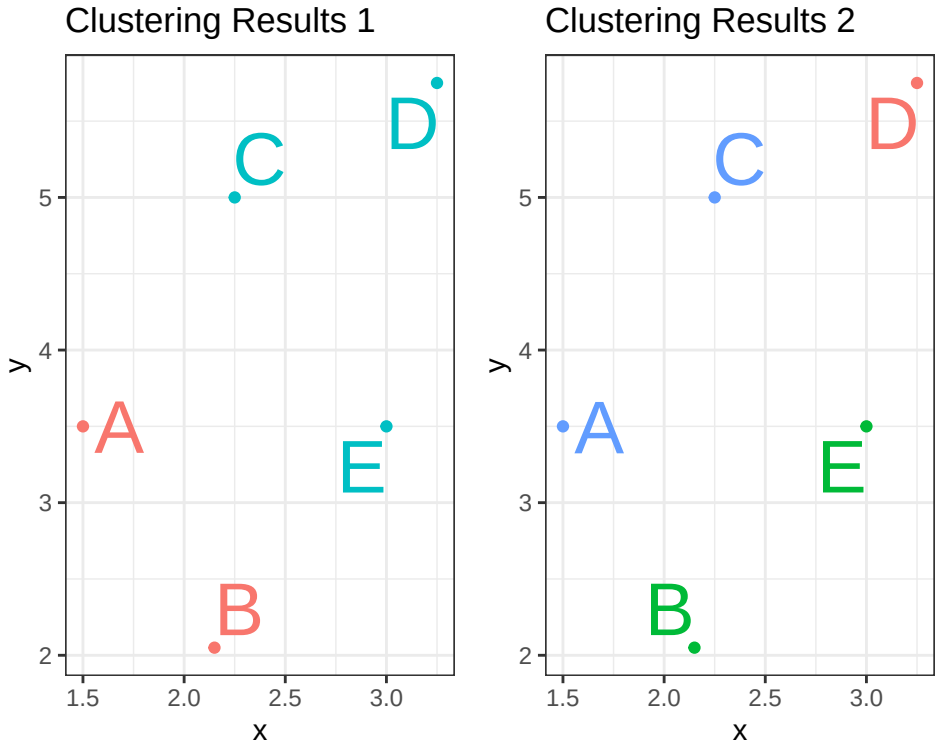
Exercise Session 5 - High dimensional visualization

Daniela Klaproth-Andrade, Felix Brechtmann, Julien Gagneur

Quizzes (from the lecture)

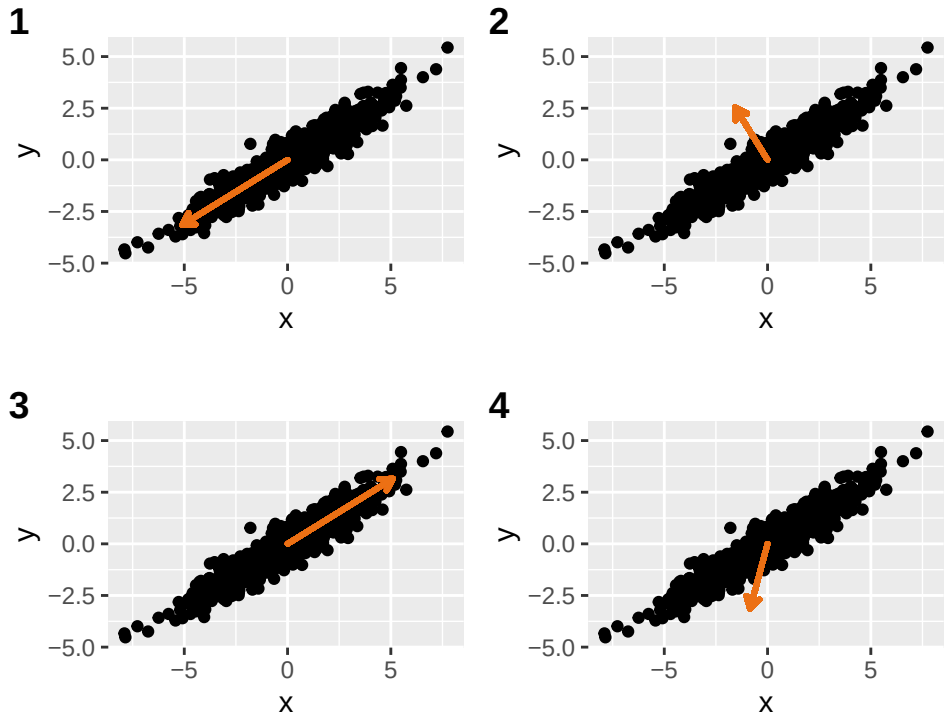
The following quizzes will be solved orally by the students and the professor during the lecture.

1. What could make k-means clustering fail?
    1. When clusters are isotropically distributed.
    2. When clusters have similar sizes.
    3. When data are not normalized.
    4. When clusters have similar variances.
2. What is the Rand Index when comparing these two different clusterings?



    1. 0
    2. 0.2
    3. 0.4
    4. 1

3. Which of the following vectors corresponds to PC1?



4. Each entry A, B, C, D in the table shows hypothetical explained variance values attributed to PC1 and PC2 of a 4-dimensional dataset. Which entr(ies) A, B, C or D are possible?

	PC1	PC2
A	60 %	50 %
B	35 %	40 %
C	60 %	40 %
D	40 %	25 %

Tutorial

The following exercises will be solved during the tutorial sessions.

Section 00 - Getting ready

1. Make sure you have already installed and loaded the following libraries:

```
library(ggplot2)
library(data.table)
library(magrittr)   # Needed for %>% operator
library(tidyr)
library(GGally)

library(pheatmap)
library(mclust)
```

Section 01 - Visualizing multiple variables

In this exercise, we will revisit how to do a correlation analysis and potential pitfalls of correlation analysis. We will use the gene expression data in `cancer_data.rds` which stores the expression of 20 genes across 30 different tumor samples. Gene expression measures the abundance of RNAs per gene and is indicative of how active a gene is in a sample. If the expression of a gene differs a lot between conditions (e.g. healthy versus cancer or different cancer types) it could hint that this gene plays an important role in this context.

Load the gene expression data in `cancer_data.rds` as a `data.table` with the following line of code:

```
expr <- readRDS("extdata/cancer_data.rds") %>% as.data.table(keep.rownames="tumor_type")
head(expr[, 1:6])
```

```
##     tumor_type        MYC         SRM       GBE1         FUK        UGP2
## 1:       DOHH2 -0.4950154  0.20907327 -0.4366726 100.0000000 100.0000000
## 2:      FARAGE -0.4913630 -0.35590874  0.5322076   0.1037538  -0.6494896
## 3:          HT -0.4559935 -0.23852493  0.6714241  -1.0136358  -0.9651291
## 4:    Kapas231  1.6152666  0.92666297  1.0168147  -0.9978958   1.0064845
## 5:     OCI-LY1  0.4942733  2.40062582 -1.4255166   1.0676067   0.9862178
## 6: OCI-LY1-B50  0.1940777 -0.04941387 -0.4810955   0.2423189   0.7046970
```

1. We are interested in the correlations between genes. Plot the pairwise correlations of the variables in the dataset. Which pair of genes has the highest correlation? *Hint:* remember that you can exclude a column `"colA"` from a data table `DT` with `DT[, -"colA"]`.

2. Visualize the raw data in a heatmap with `pheatmap`. *Hint:* `pheatmap` does not work well with `data.tables`, you should therefore convert it to a matrix before plotting with `as.matrix()`

3. Does the latter plot suggest some outliers? Could they have affected the correlations? Check by using an appropriate plot the impact of these outliers on the correlations in question 1. Substitute them with missing values (NA) and redo the previous questions 1 and 2.

Section 02 - Heatmaps and Hierarchical clustering

1. Consider the full `iris` data set without the `Species` column for clustering. Create a pretty heatmap with the library `pheatmap` of the data without clustering.

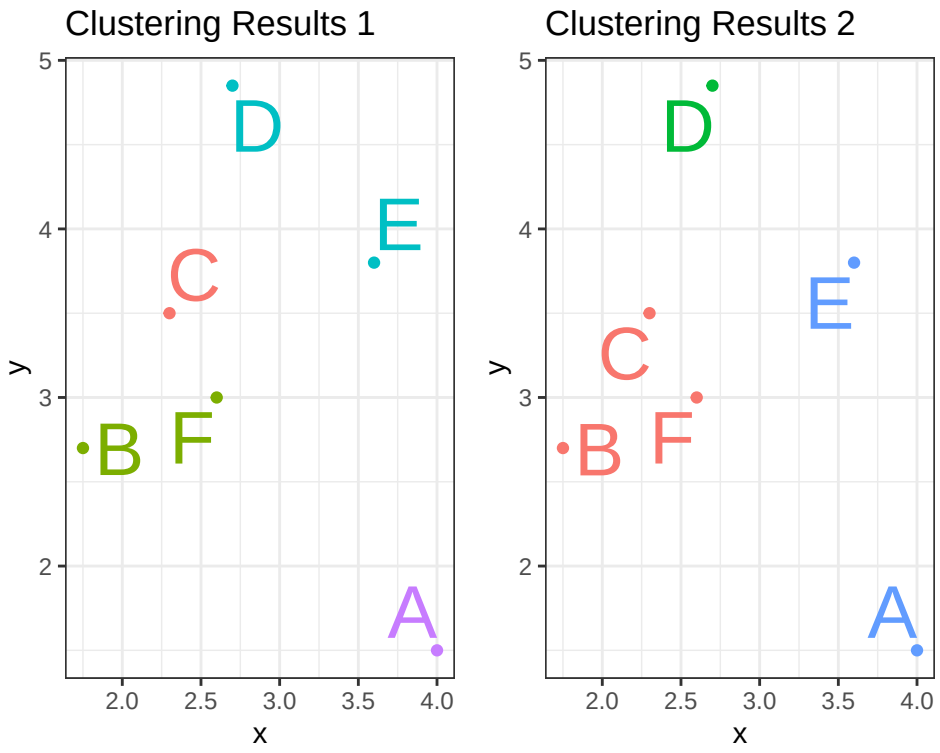
2. Now, create a pretty heatmap using `complete linkage` clustering of the rows of the data set. *Hint:* You can specify a clustering method with the `clustering_method` argument in `pheatmap`

3. Obtain the dendrogram of the row clustering using `complete linkage` clustering and partition the data into 3 clusters. *Hint:* You can use `cutree` for cutting the tree.

4. Annotate the rows of the heatmap with the `Species` column of the `iris` dataset and the three clusters from complete linkage clustering. What do you observe when you compare the clustering and the species labels?

Section 03 - Cluster comparison

1. Compute the Rand index between the two following clustering results from two different clustering algorithms. You can solve this exercise with pen and paper.



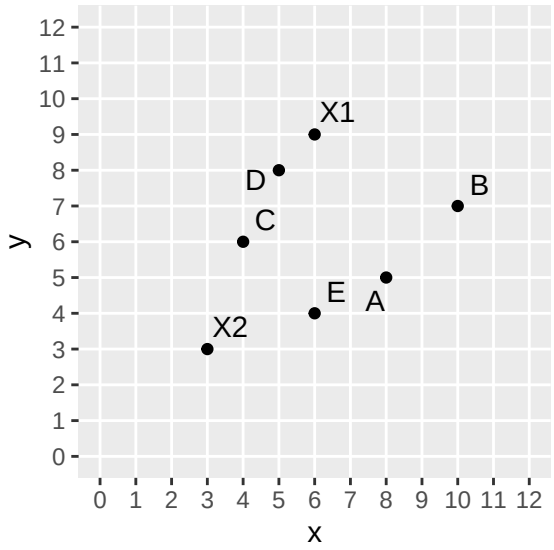
Homework

Please solve the exercises below at home. The solutions will be discussed in the central exercise.

Section 04 - Clustering and Heatmaps

In this section, we aim to compare two different clustering algorithms - hierarchical clustering and k-means clustering.

1. This plot represents a random initialization of a k-means algorithm with k=2. X1, X2 are the randomly positioned centroids and A to E are the points of the 2-dimensional dataset. Calculate the new positions of the centroids after the first iteration using the euclidean distance. *Hint:* You can solve this exercise either with pen and paper or using R.



2. Perform k-means clustering on the iris data set with $k = 3$.

3. Reproduce section 2 from the tutorial. Use the `table` function to compare the partitions from the complete linkage and k-means clustering.

4. Create a pretty heatmap using `complete` clustering of the rows annotated with the species and both clustering results - complete linkage clustering and the k-means clustering. What do you observe when you compare the two different clustering algorithms and the species labels?

Section 05 - Cluster Comparison

1. Compute the pairwise Rand indices between the clustering results from the previous sections (complete, average and k-means) and species label. *Hint:* `rand.index()` from the library `fossil`.

2. Visualize the pair wise Rand indices with a pretty heatmap. What is the best clustering in this scenario according to the computed Rand indices?

Section 06 - Dimensionality reduction with PCA

1. Let X be the `iris` data set without the `Species` column and only for the species setosa. Perform PCA on X. Make sure that you scale and center the data before performing PCA.

2. Which proportion of the variance is explained by each principle component?

3. Compute the projection of X from the PCA result and plot the projection on the first two principle components. *Hint:* `predict()`. Additionally, look at the biplot and come up with an interpretation of the first principal component.

4. Plot the first principal component against the other variables in the dataset and discuss whether this supports your previously stated interpretation.

5. Repeat the steps 1 - 4 for all species jointly (not only setosa). Discuss whether your original interpretation of the first principal component changed when performing the PCA for all species jointly. Use color to differentiate between the species in your plots.