Multivariate Verfahren – Summer Semester 2024 **Exercise sheet 6**

Hannah Schulz-Kümpel 11.07.2024

Multidimensional Scaling

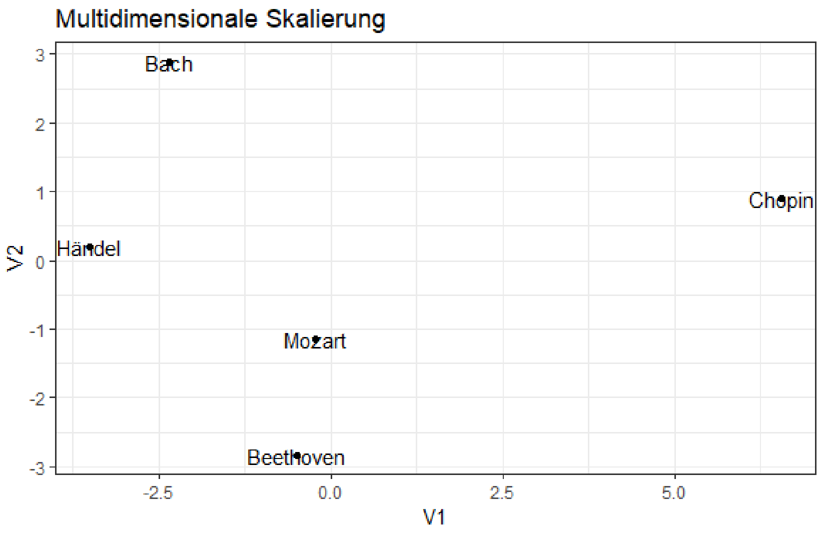
Question 1:

A study was conducted to investigate how the works of famous composers are judged by listeners. To do this, a study participant was asked to rank all pairs of the five composers *Bach, Mozart, Beethoven, Chopin* and *Händel*. The most similar pair was given a 1, the second most similar pair a 2, etc. A total of 10 pairs were compared. Then, the data was analyzed using multidimensional scaling (MDS).

a) What is the general aim of using an MDS? In addition, please state the practical advantage of using an MDS over a principal component analysis to graphically display the results of a cluster analysis graphically.

b) Briefly describe in your own words (no formulas!) which calculation steps are undertaken in a metric MDS. As data, you are given an $n \times p$ data matrix of the examined objects with all variables already standardized .

c) (i) Applying MDS in R yields the following output:



```
$points
                [,1]       [,2]
Händel    -3.4986510  0.1877229
```

```
Bach      -2.3515949  2.8932054
Beethoven -0.4864350 -2.8317511
Chopin     6.5536394  0.8904531
Mozart    -0.2169585 -1.1396304


$eig
[1]  61.00  18.52  4.03  -6.55  0.00
```

How would you interpret this output?

d) To obtain the "most suitable" dimension to reduce the data to for MDS, one may proceed the same way as for PCA - using eigenvalues. In this specific case, how many dimensions would you reduce to?

Multidimensional Scaling:
Exploratory Method for dimension reduction, mostly to visualize data in 2D/3D.
Data basis: $n \times n$-distance matrix of the n observations
Two popular variants:

1. Classical (metric) MDS: use of metric distances
2. Non-metric MDS: use of ordinal ranking only

Solution:

a) General goal of MDS:
- Dimension reduction,
- to graphically display the information from the data in 2D or 3D

Advantage over PCA: Since MDS works on general distance matrices, categorical variables can also be included.

b) Calculation steps of a metric MDS:
1. Calculate distance matrix from data
2. Calculate (scalar product) matrix B from the distances
3. Perform eigenvalue decomposition of the matrix B
4. Calculate new coordinates of the observations in the lower dimension

c) Interpretation:

Similar musical periods are close together (e.g. Bach and Handel, Mozart and Beethoven)
- Chopin somewhat remote (possibly due to focus on piano?)
- X-axis from left to right: Baroque, Classical, to Romantic
- Y-axis: Possibly linked to the age structure of the listeners?

d) With a threshold value of 0.8, two dimensions are necessary:

- Criterion:

$$\frac{\sum_{j=1}^{k} \lambda_j}{\sum_{j=1}^{n-1} |\lambda_j|}$$

Alternatively:

$$\frac{\sum_{j=1}^{k} \lambda_j^2}{\sum_{j=1}^{n-1} \lambda_j^2}$$

- Calculation:

$$\frac{\lambda_1 + \lambda_2}{\sum_{j=1}^{4} |\lambda_j|} = \frac{61.00 + 18.52}{61.00 + 18.52 + 4.03 + 6.55} \approx 0.88$$

Question 2: MDS in R

In this task, you are to perform both a classic (metric) and a non-metric MDS in `R`.
We use the data set `gardenflowers` from the package `HSAUR3` as a basis. This represents a distance matrix for 18 types of garden flowers, which was calculated on the basis of various characteristics. Since it is unclear whether the distances represent accurate objective distance measurements or are based on subjective, more *ordinal* assessments, we apply both types of MDS discussed.

a) Load the data set into `R`. Get an initial overview of the distance matrix by plotting it with `levelplot()` (package `lattice`). Can you already recognize the first similar groups?

b) Perform a classic MDS using the function `cmdscale()` (`stats`). Is a restriction to two dimensions justifiable? Plot the data on two dimensions and interpret the result.

c) Perform a non-metric MDS using the function `isoMDS()` (`MASS`). Assess again whether a restriction to two dimensions is justifiable and interpret the result using a two-dimensional plot.

d) Assuming you have the raw data as a data set:

(i) How could you check whether the content of the two dimensions used for the plots can be interpreted?

(ii) Additionally, assume that the distance matrix contains Euclidean distances. How would the results change if you were to perform a principal component analysis (on the original, standardized variables) instead of a metric MDS (on the pairwise distances) in subtask b)?

Solution:

a) *See* R*-Code!!*

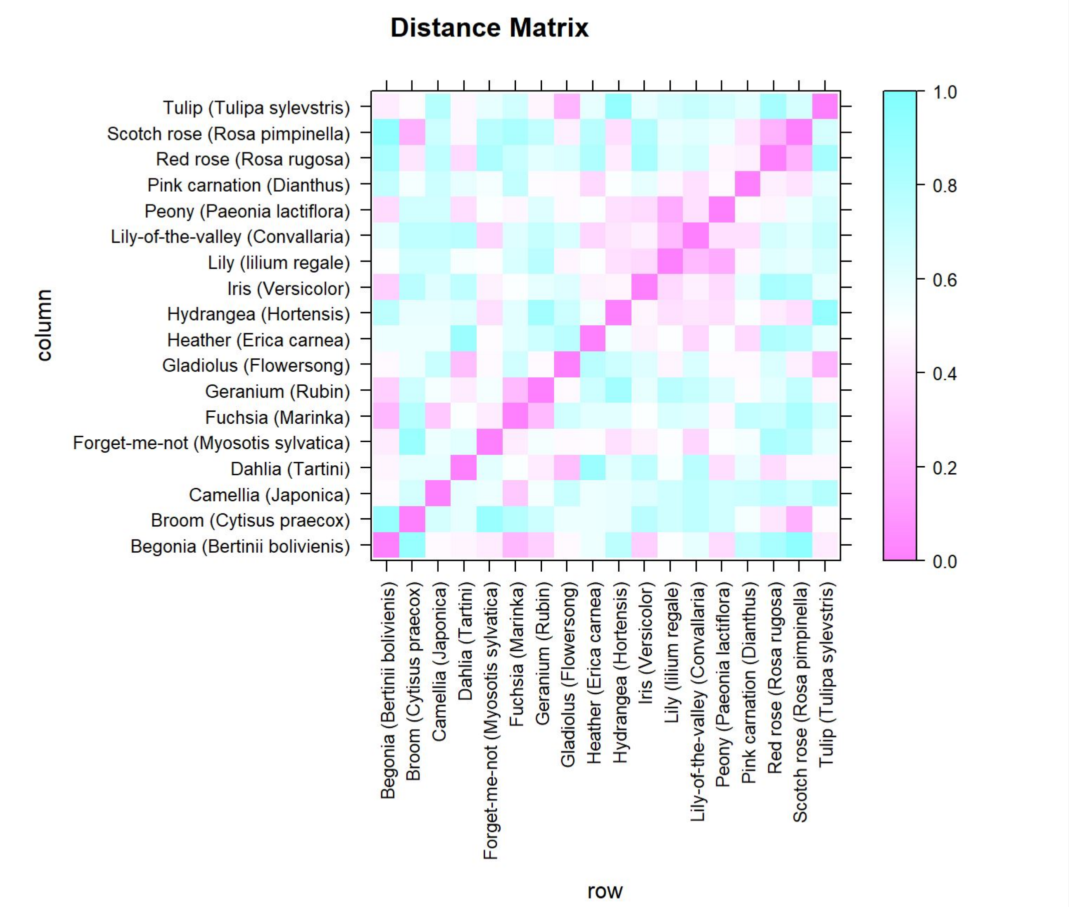


Abbildung 1: Distanzmatrix für 18 Arten von Gartenblumen (Datensatz `gardenflowers`)

You can't really see much from the distance matrix. Only the group from *'Pink carnation'* to *'Heather'* looks somewhat similar, i.e. the respective pairwise distance between these flower species seems to be lower.

b) *See* R*-Code!!*

The analysis of the eigenvalues shows that we have a value of $\alpha = 0.75$ for the first two eigenvalues. In practice, we would therefore not continue to work with two dimensions, but for reasons of simplicity and visualization we will do so anyway in the following.

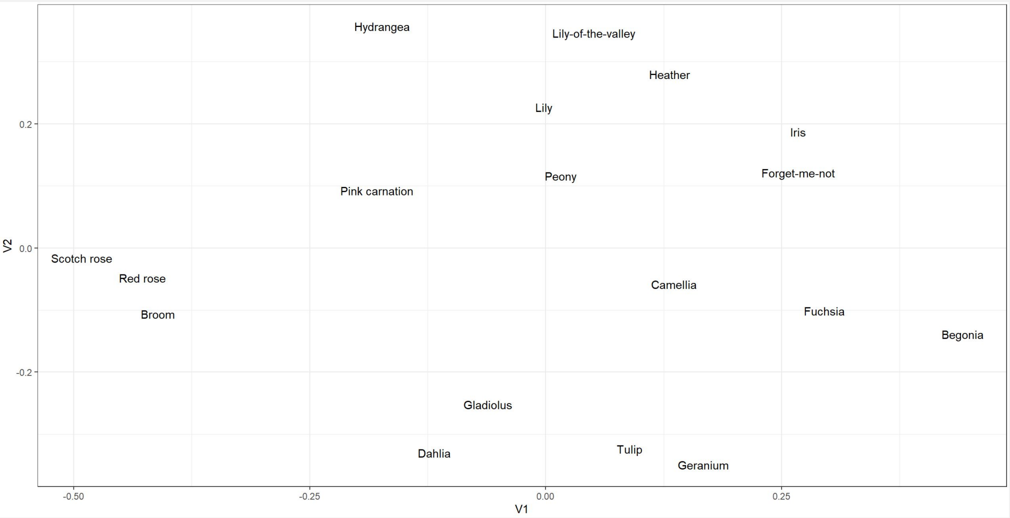


Abbildung 2: Two-dimensional visualization of flower species using classical MDS

Interpretation: This visualization indicates 4-5 groups, each consisting of at least three species.

c) *See* R*-Code!!*

Consideration of STRESS as a quality criterion for determining the dimensions: The deviation of just under 20% is very high. It can therefore also be seen here that the restriction to two dimensions is not optimal.

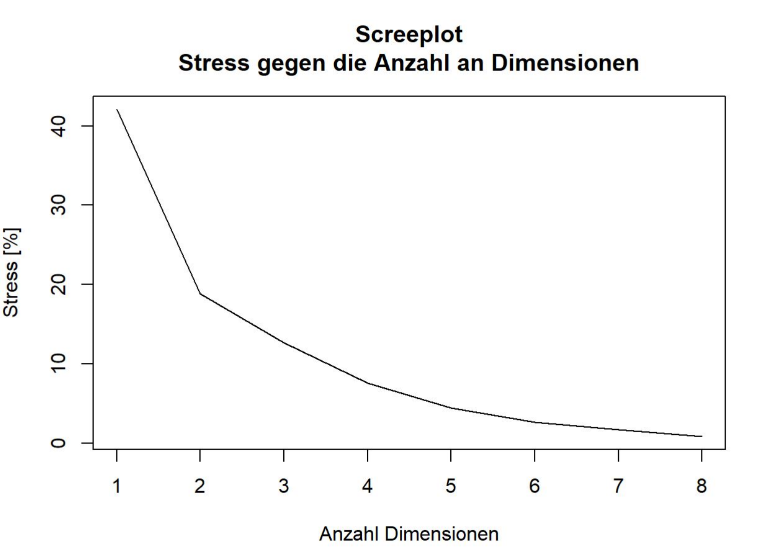


Abbildung 3: Screeplot of stress for different dimensions

Unfortunately, the screeplot only decreases very slowly, but at least a three-dimensional in three dimensions reduces the stress by 1/3 to approx. 12% and thus to 'satisfactory'.

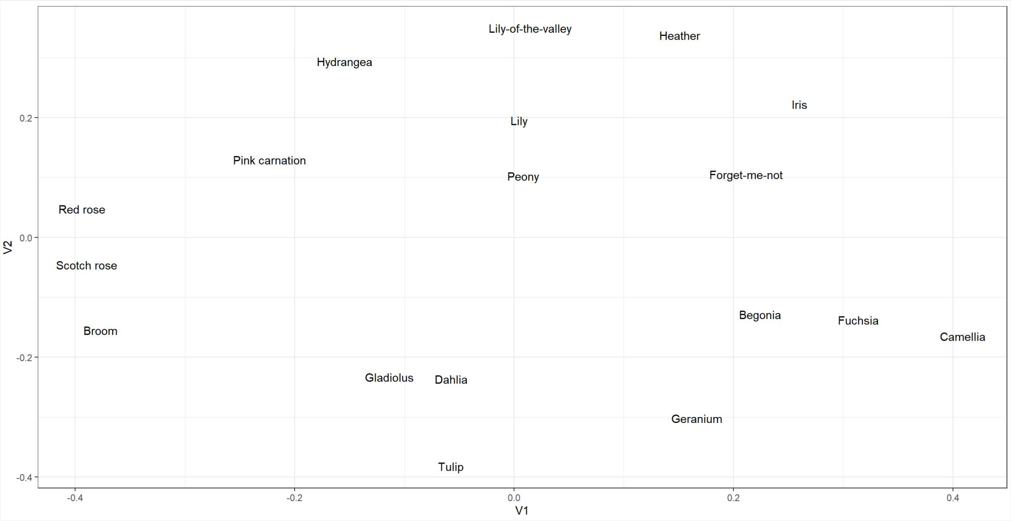


Abbildung 4: Two-dimensional visualization of flower species using non-metric MDS

Interpretation: Relatively similar results to the metric MDS. However, the species are somewhat more scattered here, which means that the use of this method would be less indicative of groups that are very different from one another.

d) *See* `R`*-Code!!*

i) Each of the two extracted dimensions/scales can be correlated with the original variables on the basis of which the distances were calculated.

   Ideally, there are then only a few very high correlations and many very very low correlations, so that it becomes clear which variables are included in which of the dimensions. The interpretation is similar to the principal component analysis.

ii) Not at all! the reason for this:

   PCA performs an eigenvalue decomposition of the covariance/correlation matrix. MDS performs an eigenvalue decomposition of the Gram matrix (scalar product matrix) B. In the case of using Euclidean distances, the two methods lead to identical results (see lecture slides).

   Please note: When using a different distance measure, a metric MDS leads to different results!