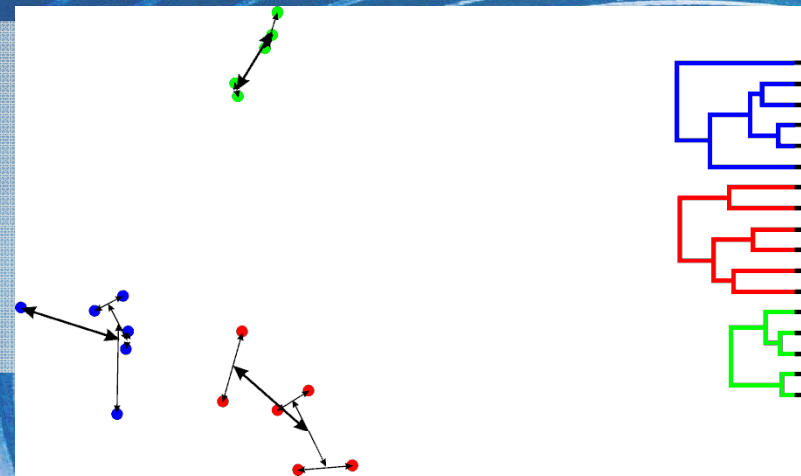


Hierarchical Clustering

Basics



Please read the introduction to principal component analysis first. There, we explain how spectra can be treated as data points in a multi-dimensional space, which is required knowledge for this presentation.

Hierarchical Clustering



We have a number of datapoints in an n-dimensional space, and want to evaluate which data points cluster together.

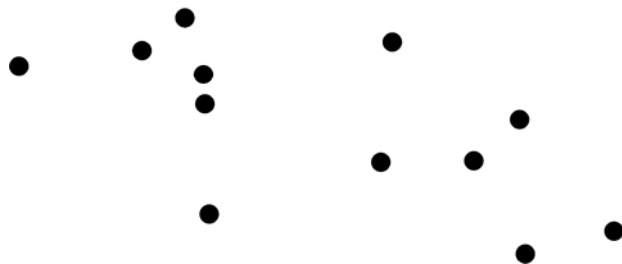
This can be done with a hierarchical clustering approach

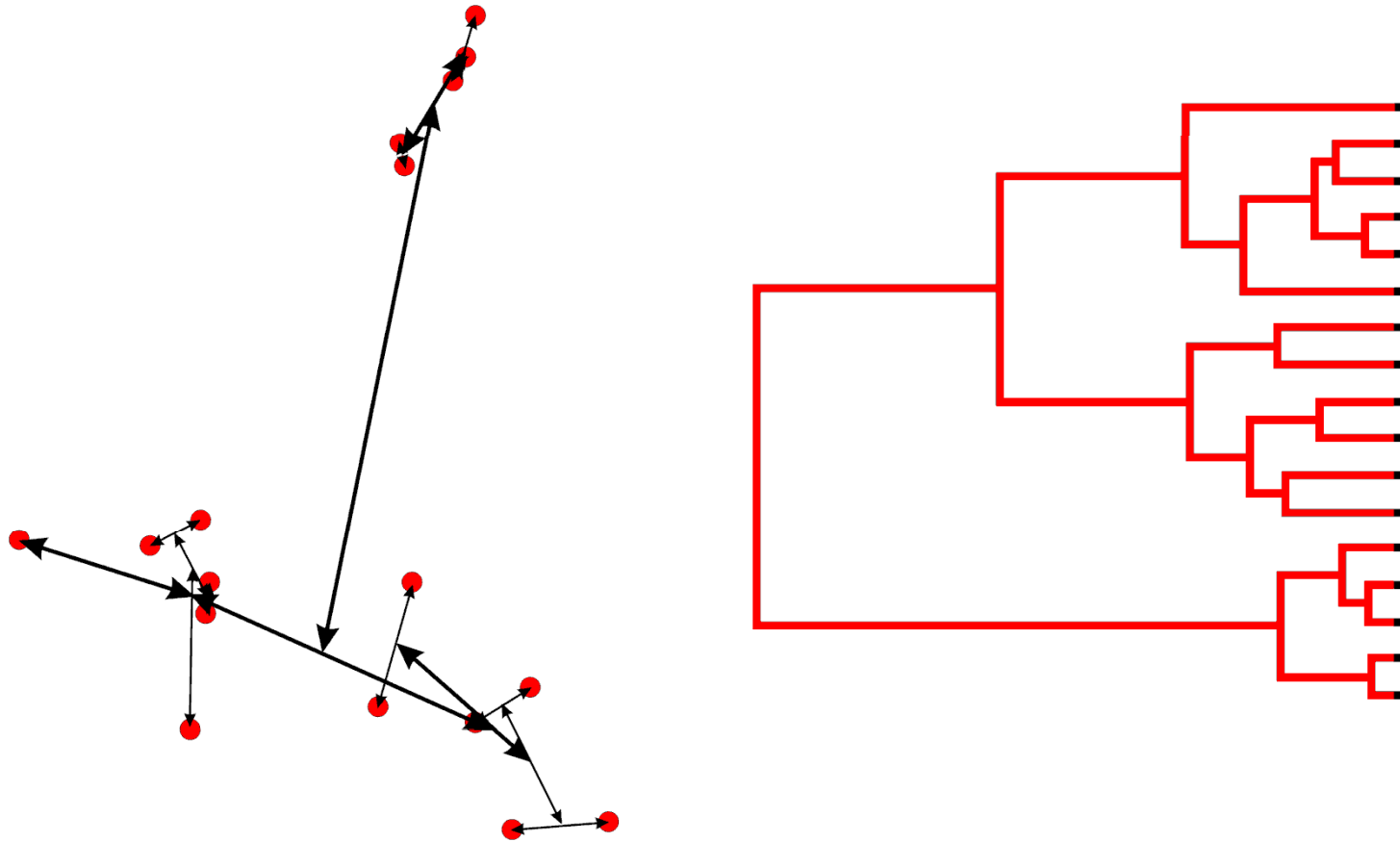
It is done as follows:

1) Find the two elements with the smallest distance (that means the most similar elements)

2) These two elements will be clustered together. The cluster becomes a new element

3) Repeat until all elements are clustered





Unsupervised Clustering

Normalize Data

Use PCA Data

Reduce Dimensions

70.0 % Sum Explained Variance

Create Full Tree Show Paths 40 Max. Path Length

70 Number of Classes Show ClinProtClustering.xml

OK Cancel Help Advanced <<

Cosine Distance Method 1.5 Minkowski Exponent

Average Linkage Method



Important parameters in hierarchical clustering are:

The distance method

This measure defines how the distance between two datapoints is measured in general

Available options: Euclidean (default), Minkowski, Cosine, Correlation, Chebychev, Spearman

The linkage method

This defines how the distance between two clusters is measured

Available options: Average (default) and Ward

Use PCA data:

Determines if the data is pretreated with a PCA.

Only reasonable if used with

Euclidean metric and with a reduction of explained variance



Which parameters to select for MALDI imaging data?

Recommended settings:

„Use PCA data“, „reduce dimensions to 70%-95% of explained variance“

Distance Method: „Euclidean“

Linkage Method: „Ward“

(do not use PCA data with non-euclidean distances!)

If you are interested in what the distance and linkage methods are, read on, otherwise quit here.

Distance - Euklidian

If there are two points $P = (p_1, p_2, \dots, p_n)$ and $Q = (q_1, q_2, \dots, q_n)$ in n -dimensional space, then the euclidean distance is defined as:

$$\sqrt{|p_1 - q_1|^2 + |p_2 - q_2|^2 + \dots + |p_n - q_n|^2}$$

If we speak of „distance“ in common language, the euclidean distance is implied. Example:

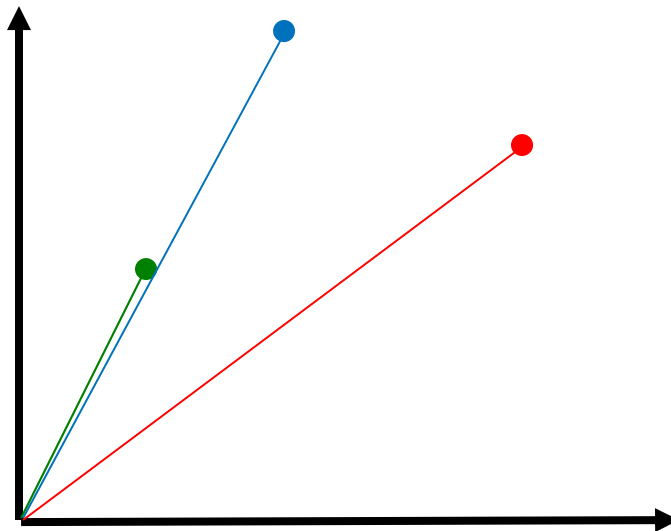


Euclidean distance is invariant against transformations of the coordinates.

The clustering results will improve if PCA-data are used and the data are reduced to a certain percentage of explained variance

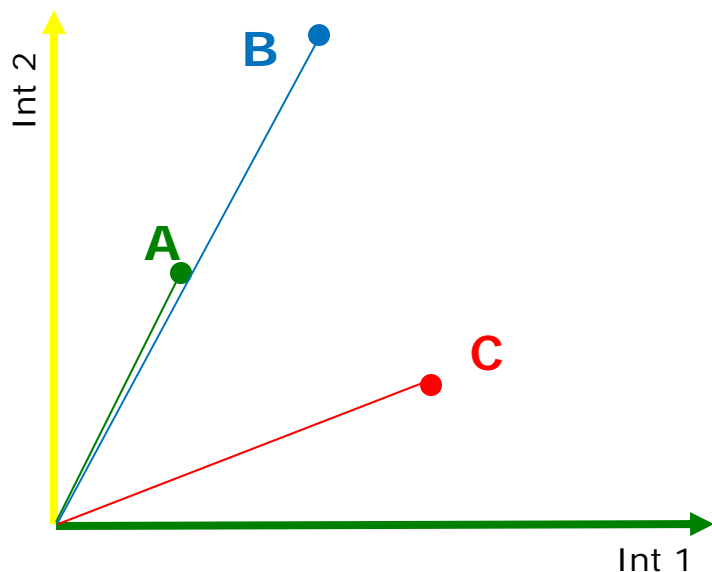
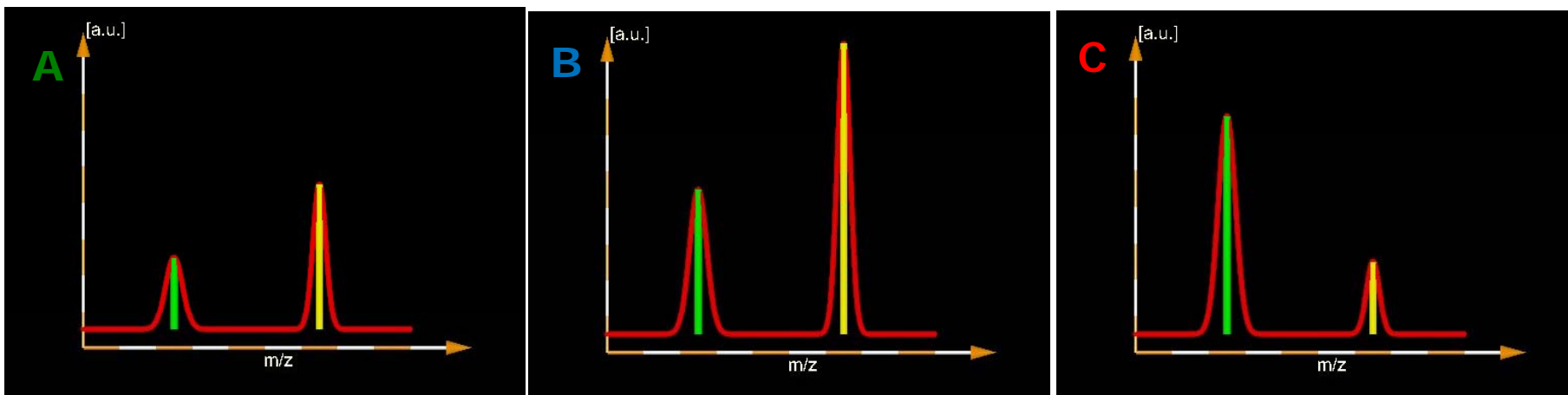
Distance – Cosine 1

In the cosine distance, the distance between two datapoints is defined by the 1-cosine of the angle between the vectors from the origin to the datapoints



The angle between the green and the blue line is very small, so the two datapoints have a small distance in the cosine distance

Distance – Cosine 2



The euclidean distance between spectra A and B is equal to the euclidean distance between A and C.

Spectra A and B are similar in cosine distance.

Note: These two spectra are only different in absolute intensity. Cosine distance does an intrinsic normalization.

Since spectra in ClinProTools are always normalized, cosine and euclidean distance behave rather similar.

Cosine is not reasonable with PCA data

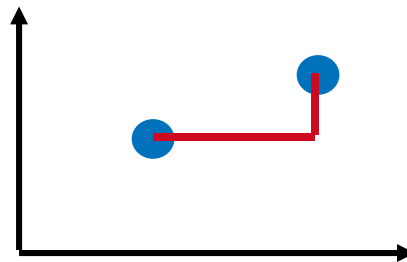
Distance - Minkowski

If there are two points $P = (p_1, p_2, \dots, p_n)$ and $Q = (q_1, q_2, \dots, q_n)$ in n -dimensional space, then the Minkowski distance is defined as:

$$\sqrt[a]{|p_1 - q_1|^a + |p_2 - q_2|^a + \dots + |p_n - q_n|^a}$$

Euclidean distance is a special case of the Minkowski metric ($a=2$)

One special case is the so called „City-block-metric“ ($a=1$):



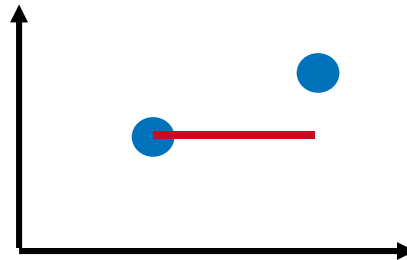
Clustering results will be different with unprocessed and with PCA data

Distance - Chebychev

If there are two points $P = (p_1, p_2, \dots, p_n)$ and $Q = (q_1, q_2, \dots, q_n)$ in n -dimensional space, then the Minkowski distance is defined as

$$\max(|p_1 - q_1|, |p_2 - q_2|, \dots, |p_n - q_n|)$$

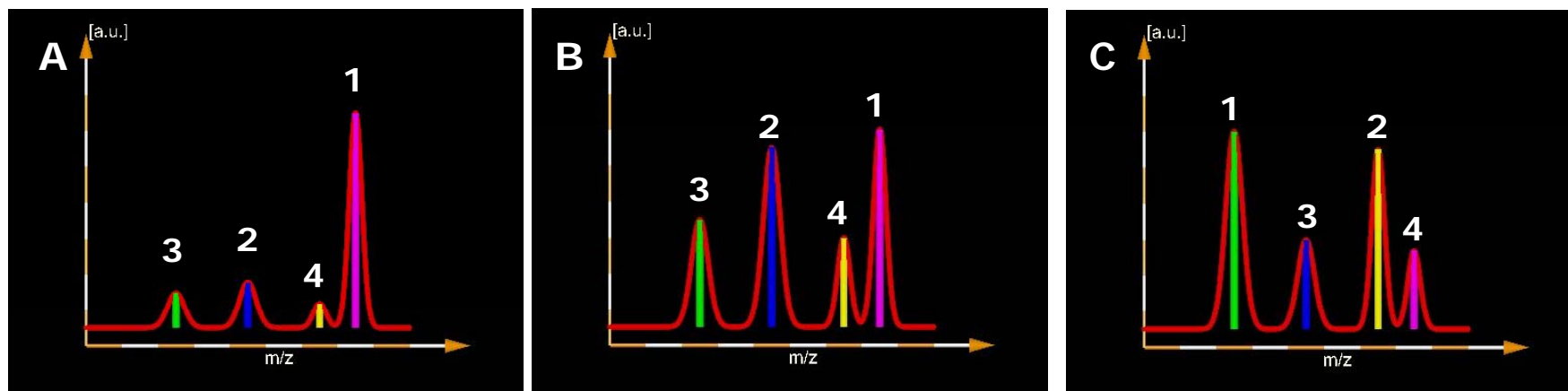
The Chebychev distance is also a special case of the Minkowski distance ($a \rightarrow \infty$). The distance is defined by the maximum distance in any coordinate:



Clustering results will be different with unprocessed and with PCA data

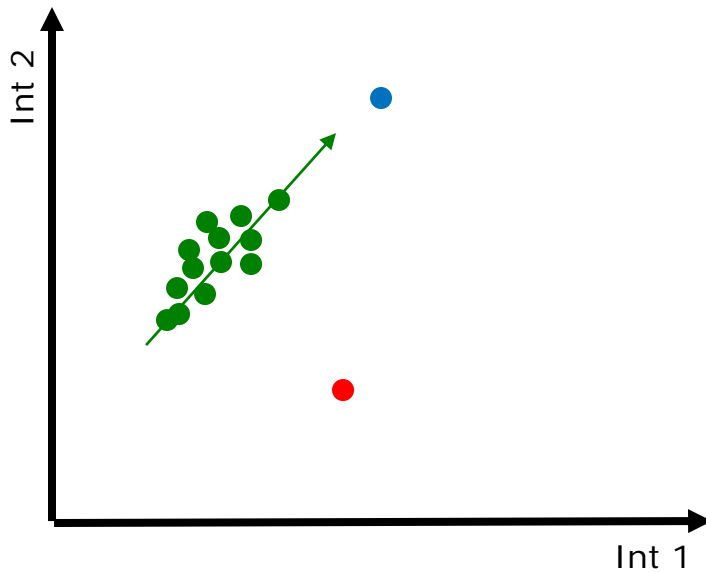
Distance - Spearman

In the Spearman distance the absolute intensities of the peaks are unimportant, two spectra will be close together if the relative peak-patterns are similar. Each peak will be assigned a rank in order of the intensity, and the ranks will be compared



In this example, spectra A and B are identical in the Spearman distance, while spectrum C has a big distance to A and B

Distance – Correlation



In the correlation distance method, the blue data point is closer to the green cluster than the red one, because the blue one correlates better with the data in the cluster.

Linkage - Single

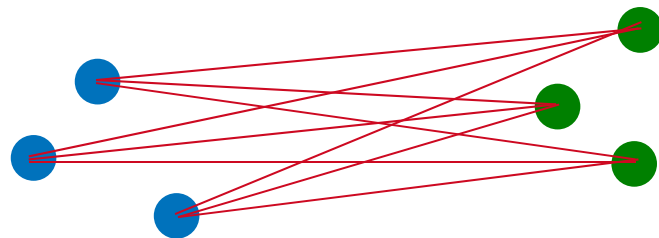
The single linkage is easiest to understand. The distance between two clusters is equal to the distance of the closest elements from the two clusters



The single linkage has practical disadvantages and is therefore not selectable in ClinProTools

Linkage - Average

In the average linkage the distance of two clusters is calculated as the average of the distances of each element of the cluster with each element of the other cluster



In this example the distance between the green and the blue cluster is the average length of the red lines

**Average linkage is the default setting in ClinProTools.
However, on imaging data the „Ward“ linkage gives usually
better results**

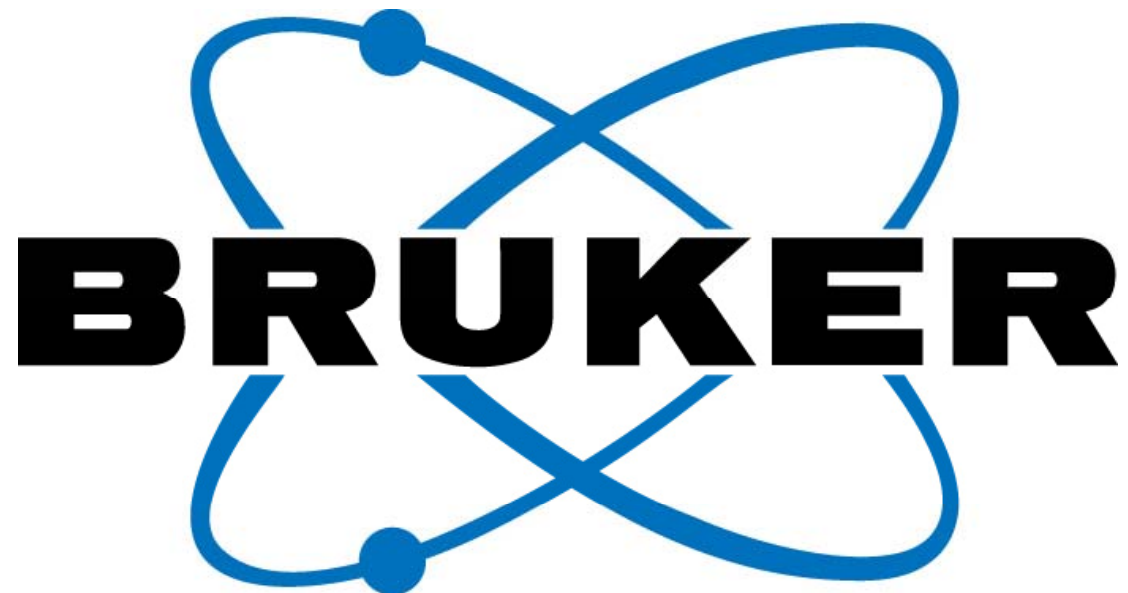
Linkage - Ward

In the ward linkage for each cluster a error function is defined. This error function is the average (RMS) distance of each datapoint in a cluster to the center of gravity in the cluster

$$D = \left[\begin{array}{c} \text{Diagram of unified cluster with center of gravity and distances} \\ \text{Error function of unified cluster} \end{array} \right] - \left[\begin{array}{c} \text{Diagram of two separate clusters with their own centers of gravity and distances} \\ \text{Error function for each cluster} \end{array} \right]$$

The distance (D) between to clusters is defined as the error function of the unified cluster minus the error functions of the individual clusters

The ward linkage usually gives the nicest clustering results with MALDI imaging data



www.bdal.com