STATISTICAL SUPPORT
SWISS CENTER FOR AFFECTIVE SCIENCES
CISA

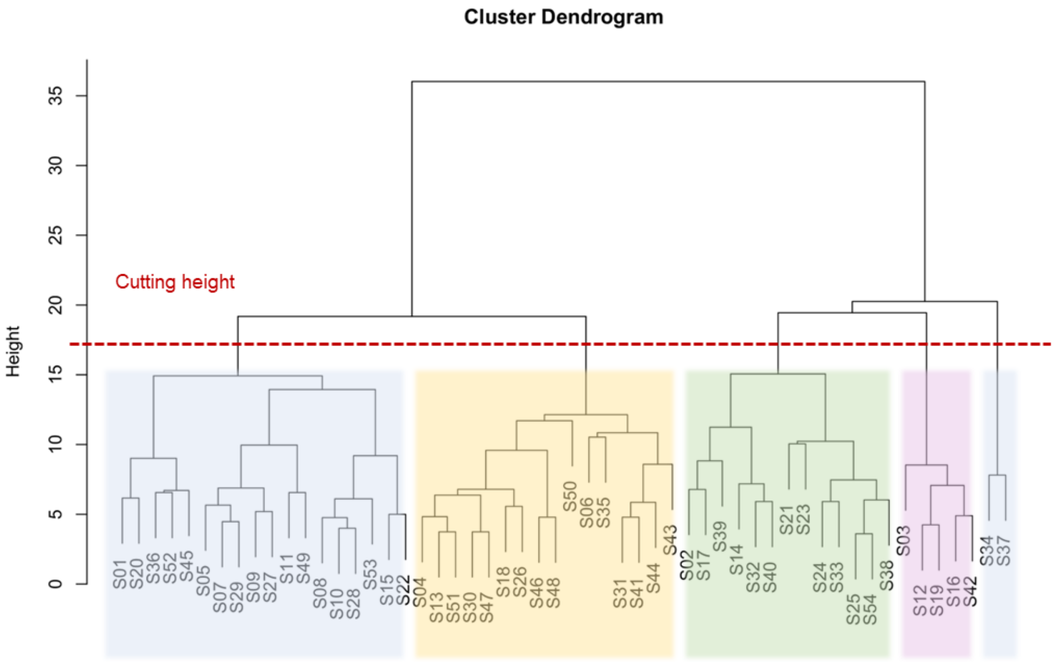
Hierarchical clustering

E-mail distributed on 12-01-2023

Dear all,

A popular method for finding groups in multivariate data is clustering. Clustering can be applied to the columns of a data set (variables) or its rows (observations), with the latter being more common. The goal is usually to find groupings of subjects with similar patterns of values on the variables. This is especially appealing for emotion data—survey or experiment—where it may be useful to distinguish qualitatively different responders (e.g., fearful, angry), as well as non-responders.

Although many clustering methods have been developed, one of the simplest and most useful is **hierarchical clustering**, specifically **agglomerative nesting** (AGNES). AGNES builds a tree diagram that progressively links up the most similar observations and groups of observations, until all are in a single cluster. This tree is called a dendrogram, and quantifies the distance between the observations. The higher a branch that separates an observation (or group of observations) from its neighbors, the more dissimilar the groupings are (see attached example).



Formally, a hierarchical cluster analysis with AGNES proceeds as follows:

- Compute the *N* × *N* distance matrix of all observations (Euclidean distance is default)
- Run agglomerative nesting on the distance matrix, using a linkage criterion (Ward's criterion usually performs best)
- Plot the dendrogram and identify a plausible cluster number
- Label the observations according to the identified clusters

Visual identification of clusters may seem subjective and inexact but this method frequently outperforms more quantitative criteria (e.g., *k*-means, model-based clustering) because it immediately emphasizes **interpretability.** In fact, a good general rule is that clusters should not be retained if they cannot be interpreted, even if quantitative criteria suggest they are meaningful. Likewise, inferential tests for deciding on cluster numbers should be avoided. In that sense, AGNES is quite similar to other visual/descriptive methods that outperform inferential methods, such as the QQ-plot for checking normality of residuals, and the scree plot for determining the number of principal components in PCA.

However, once clusters have been identified, it can make sense to compare them inferentially on their mean values, using a classical MANOVA (=*K* independent groups, *Q* outcome variables). Pairwise contrasts will further reveal on which variables the clusters differ. Finally, it is worthwhile to investigate which other variables in your data set might predict cluster membership (e.g., demographics, conditions), using ANOVA for continuous variables and chi-square analysis for categorical variables.

The attached R script gives an example code for how to run hierarchical clustering, using the `cluster` library, and functions `dist`, `agnes`, and `cutree`. The basic dendrogram is quite primitive as a visualization, but it can be extended in many appealing ways, some of which are available in the package `dendextend`, which also allows graphical comparison of different cluster solutions.

Best,
Ben

Ben Meuleman, Ph.D.
Statistician
Swiss Center for Affective Sciences
University of Geneva | Campus Biotech
Chemin des Mines 9 | CH-1202 Genève
ben.meuleman@unige.ch | +41 (0)22 379 09 79

```
########################################
## HIERARCHICAL CLUSTERING WITH AGNES
########################################

library(cluster)
## TOY DATA FOR EMOTION SURVEY
emotion <- read.csv("https://drive.switch.ch/index.php/s/Uyvv3V3Kg2OmJfv/download")
head(emotion)

## CLUSTERING
distances <- dist(emotion[,-c(1:4)],method="euclidean") #DISCARD FIRST FOUR COLUMNS
```

```
attr(distances,"Labels") <- emotion$ID #LABELS FOR DENDROGRAM LEAVES
hclus <- agnes(distances,method="ward")
hclus <- as.hclust(hclus)
plot(hclus)
emotion$clusters <- as.factor(cutree(hclus,k=4))

## GROUP DESCRIPTIVES
xtabs(~clusters,data=emotion)
xtabs(~VR_demo+clusters, data=emotion)
aggregate(as.matrix(emotion[,-c(1:4,21)])~clusters,data=emotion,FUN=mean,na.rm=TRUE)

## GROUP DIFFERENCES
mlm <- lm(as.matrix(emotion[,-c(1:4,21)])~clusters,data=emotion)
MAOV <- manova(mlm)
summary(MAOV,test="Pillai")
```