Multi-dimensional scaling in R

Johannes Karl

2018-07-20

# MDS basics

This post focuses on MDS (Also sometimes called smallest space analysis (Guttman, 1968) or multidimensional similarity structure analysis (Borg & Lingoes, 1987)) as exploratory tool and a large amount of the information is lifted from Borg and Gronen (2005). Exploratory MDS might seem at first glance like tea-leaf reading, but over the years several rules have been developed that aid in determining structure from the MDS plots.

In the case I will talk about during this post we were interested in discovering how different measures of mindfulness, personality, and approach/ avoidance motivation are related to each other. To this extent participants answered statements about themselves on Likert-scales related to these concepts. An important concept in MDS is distance. The idea behind distance is that individuals reproduce mental distances between concepts when asked about dissimilarities between the concepts. To cite directly form Borg and Gronen (2005): " The most common approach is to hypothesize that a person, when asked about the dissimilarity of pairs of objects from a set of objects, acts as if he or she computes a distance in his or her “psychological space” of these objects."

In our case participants were not asked to rate dissimilarities between the measures, but rather rate themselves on these measures. Nevertheless, if we think about it these self-ratings can also represent psychological distances. A participant that scores high on avoidance behaviour might also score high on anxiety, but score low on approach behaviour. These differences can be represented as psychological distances with avoidance behaviour and anxiety being in close proximity and approach behaviour being distal from those two concepts. ## Correlations and MDS Data obtained from answers to Likert scales can not directly be considered as distances, but the correlation coefficients between multiple columns can. Correlations are appropriate for MDS.

### Obtaining a distance matrix from a correlation matrix.

The *psych* package provides a handy function (*cor2dist*) to transform a correlation matrix into a distance matrix. In the background this simple function is performed:

Pipeline provide a handy way of reducing the number of objects stored in the work space if they are not used in later analysis. In my current case I used the different measures of interested (Mindfulness facets, personality, etc.) and labeled the facets. The first line produces the correlation between the variables, the second line selects only the correlation coefficients from the output, and the second line converts them into a distance matrix which is returned as an object labelled distance.

distance <- psych::corr.test(mnd\_test[,facets]) %>%  
 .$r %>%  
 psych::cor2dist()

So why even bother with transforming correlation coefficients into distances? The reeason for this can be found in the assumptions about a geometrical plane. The first assumption is called non-negativty and can be expressed as:

Simply put on plane the the distance between any two points and is greater than 0 or equal to 0 (if ). This presents the first reason why correlation coefficients can not directly used as input for a MDS, because they can be negative. The second assumption called symmetry is self-explanatory:

For an MDS it is necessary that the distance between and is identical to the distance between and .

Last is the triangle inequality:

This triangle inequality says that going directly from to will never be farther than going from to via an intermediate point . If happens to be on the way, then the function is an equality. ## Evaluating stress Rather then goodness-of-fit indicators MDS uses a badness-of-fit indicator, *stress*. Stress is the normed sum of squares aggregating the representation errors of the model compared to the undrelying data. In an applied context we rarely examine the raw stress scores as it is dependent on the scale used. Rather we use a value to judge badness-of-fit that is called Stress-1 or 1 (Kruskal, 1964a). If you are interested, below is the formula for . **IT IS CURRENTLY NOT HERE :)**

One important property of is that missing data is skipped in the process of summing up the representation errors. This is probably less a problem if you are working with correlations derived from underlying data, but can be a problem if you are working with values obtained from other sources.

From the presented formula we can derive that if we perfectly represent the underlying data will be 0 and the greater the deviation gets the greater becomes.

### The math

## Metric and Non-metric MDS

**Include why ratio or ordinal**

## Interpretation of a MDS plot

In their introduction to Multidimensional Scaling Kruskal and Wish recommend that a MDS plot should be interpreted by applying the following rule (generalised from their example of Morse code): " Pick some point which is peripheral, that is, which lies at the outermost edge of the configuration. Ask yourself what is common to this point and its nearest neighbors, and how they differ from the points at the opposite edge of the configuration. Then repeat this process, using other peripheral points."

In a two dimensional plot it can be beneficial to first examine the x and y axis. This can yield important insight into the structure of the points.