Non-metric Multidimensional Scaling (NMDS)
**Objective:** Group data points into classes of similar points based on a series of variables

Lots of types of multidimensional scaling: PCA is aka Classic Multidimensional Scaling

The goal of NMDS is to represent the original position of data in multidimensional space as accurately as possible using a reduced number of dimensions that can be easily plotted and visualized (like PCA).

BUT (unlike PCA which uses Euclidian distances) NMDS relies on rank orders (distances) for ordination (i.e non-metric)

The use of distances omits some of the issues associated with using predictor variables alone (e.g., sensitivity to transformation)

Allows for much more flexible technique that accepts a variety of data types

Shepard 1962
Kruskal 1964
Tprgersen & Meuser 1962
Guttman 1968

Contributed to the development of multidimensional scaling
The math behind NMDS

NMDS is an iterative procedure which takes place over several steps:

1. Define the original data point positions in multidimensional space
2. Specify the number of reduced dimensions you want (typically 2)
3. Construct an initial configuration of the data in 2-dimensions
4. Compare distances in this initial 2D configuration against the calculated distances
5. Determine the stress on data points
6. Correct the position of the points in 2D to optimize the stress for all points
The math behind NMDS

Consider a 3 variable analysis with 4 data points

When we *compress* our 3D image to 2D we cannot accurately plot the true distances
E.g. the distances between AD and BC are too big in the image

<table>
<thead>
<tr>
<th>Data.ID</th>
<th>Variable1</th>
<th>Variable2</th>
<th>Variable3</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.9</td>
<td>1.9</td>
<td>1.5</td>
</tr>
<tr>
<td>B</td>
<td>1.7</td>
<td>0.5</td>
<td>1.6</td>
</tr>
<tr>
<td>C</td>
<td>3</td>
<td>2</td>
<td>3.1</td>
</tr>
<tr>
<td>D</td>
<td>1.9</td>
<td>3.5</td>
<td>3</td>
</tr>
</tbody>
</table>

The difference between the data point position in 2D (or # of dimensions we consider with NMDS) and the distance calculations (based on multivariate) is the **STRESS** we are trying to optimize
NMDS optimizing stress

**Stress** – value representing the difference between distance in the reduced dimension compared to the complete multidimensional space

NMDS tries to optimize the stress as much as possible

Think of optimizing stress as: “*Pulling on all points a little bit so no single point is completely wrong, all points are a little off compared to distances*”

Ideally we want as little stress as possible
To run NMDS you need to install the `ecodist` package.

```r
library(ecodist)
nmds(distMatrix, mindim=n, maxdim=n) (ecodist package)
```

- **Distance matrix of your data rows based on your predictor variables.**
  - You need to calculate this before running the NMDS analysis.

- **`mindim`** = minimum number of dimensions you want to use.
- **`maxdim`** = maximum number of dimensions you want to use.

- You can run NMDS with as many dimensions as you have predictor variables, but we are trying to reduce the dimensions so we can group data points.

- Typically, we want to set both of these values to 2 to simplify our output.
NMDS in R

Scores – these are the data point outputs that have been pulled to optimize the stress from multi dimensions in 2D (or the # of dimensions considered)

These are the values we plot to look at which data points group together

We can merge a class variable back into look if predetermined groups actually group out together or see what groups we could potentially combine

Distance matrix
Mahalanobis is good for correlated variables
NMDS in R

Stress – value representing the difference between distance in the reduced dimension compared to the complete multidimensional space

R will produce a list of values – one for each iteration it had to do – the more complex your dataset the more iterations (and time to run the analysis) are needed

The last value in the list is the final stress value which is uninformative by itself, but you should check to make sure the stress is stable when you consider more dimensions (modify maxdim)

> nmds_out$stress # the last value indicates the final stress
[1] 0.3094383 0.3169082 0.3035864 0.3091353 0.3120136 0.3134852 0.3087037
[8] 0.3086726 0.3071696 0.3011707
NMDS in R

Your data may NOT be able to be viewed in 2D due to high stress

Use the rationale: “Include dimensions until I don’t gain a significant reduction in my stress value”

If stress is too high for 2D or 3D NMDS might not be the best method

i.e. Visualizing your data in fewer dimensions compromises the data too much
NMDS - Biplot

Data points considering scores in 2D

Direction of the arrows +/- indicate the trend of points (towards the arrow indicates more of the variable)

The closeness of points will indicate how similar they are

It is up to you to determine where groupings should be made
NMDS - Biplot

Once you decide on groups you can then use graphics to simply distinguish them.

We cover this in Lab 5.