The MDS-GUI: A Graphical User Interface for Comprehensive Multidimensional Scaling Applications

Andrew Timm[1,2] , Sugnet Gardner-Lubbe[1]
[1]University of Cape Town, Cape Town, South Africa: Department of Statistical Sciences
[2]Corresponding author: Andrew Timm, e-mail: timmand@gmail.com

Abstract

The **MDS-GUI** is an *R* based graphical user interface for performing numerous Multidimensional Scaling (MDS) methods. The intention of its design is that it be user friendly and uncomplicated as well as comprehensive and effective. This paper will discuss its capabilities and demonstrate its use with a well known MDS example data set.

1. Introduction

The MDS-GUI has been developed to provide the user, even with no theoretical background on the subject, with the opportunity to perform a number of MDS methods and output a host of relevant details and graphics. In broad terms, the GUI allows the user to simply and efficiently input their desired data, choose the type of MDS they would like to perform as well as select the type of output they would like to achieve by the analysis. The use of sub-menus and property tabs gives the user the option to fine tune specific parameters of the desired MDS procedure as well as provide options to alter the way in which the resulting plots are displayed The graphical outputs are of an interactive nature and allow the user to make adjustments to the output with a cursor to observe any difference in results. Multidimensional Scaling is usually an iterative technique, which is a quality preserved by the graphics of the software. The user is thus able to have a visual display of the processes at work and observe the moving ordination configuration.

2. Multidimensional Scaling

Multidimensional Scaling (MDS) is a multivariate method of ordination. What MDS does is find a set of vectors in p dimensional space (where p has been predefined) such that the matrix of Euclidean distances among them corresponds as closely as possible to some function of the input dissimilarity matrix. The result allows the researcher to observe a configuration in two or three dimensional space such that each of the n objects of their data is represented by a point. The dimensions of the plane onto which the configuration is plotted is not limited to two (or three dimensional hyperplane plotting), however $p > 3$ produces obvious difficulties in visual analysis of the coordinates. Interpretation of the configuration relies on the fact that objects positioned 'close' to one another are supposedly similar, and those 'far' apart are less similar. While there are a wide range of types of MDS, there are two categories under which the majority of the methods fall. These two categories are Metric and Non-Metric Multidimensional Scaling. Methods classified as Metric MDS make the assumption that there are metric qualities in the measurement of the proximities, while Non-Metric methods only make use of the ordering of the proximities in the derivation of the MDS configuration.

In order to optimally match the distances in the MDS configuration to the matrix of input dissimilarities each method of MDS is usually based, at least in part, by a specific loss function that is minimized, usually called "stress". Three versions of stress are available in the MDS-GUI, being Kruskal's Stress (Kruskal, 1964) or STRESS1, STRESS2 and Normalised Raw Stress.

Six methods of Multidimensional Scaling are included in the current version of the MDS-GUI. These methods are: Classical Scaling (Principal Coordinate Analysis), Metric SMACOF (Scaling by Majorizing a Complicated Function), Metric Least Squares Scaling, Non-Metric SMACOF, Sammon Mapping and Kruskals Analysis. For more information on MDS in general and for full descriptions of individual MDS methods and versions of stress, refer to Cox and Cox (2001) and Borg and Groenen (2005).

3. The MDS-GUI

The software was developed using the *R* (R-Development-Core-Team, 2011) statistical programming language and the *R*-wrapped version of the GUI building language, *tcltk*. In addition to these, packages affiliated with the **tcltk** *R* package, such as **tcltk2** (Grosjean, 2011) and **tkrplot** (Tierney, 2011) were utilised extensively during the construction of the user interface. The majority of methods relating to Multidimensional Scaling and optimisation used by the MDS-GUI were derived from the functions found in the **MASS** (Venables and Ripley, 2002) package. Functions relating to all SMACOF algorithms were based on the code developed by LeRoux (2012). The **MDSGUI** is available from the R-Forge and CRAN websites.

3.1 Data Handling

The majority of MDS methods require input of the $n \times n$ proximity matrix, Δ. The MDS-GUI allows for importing a matrix of dissimilarities directly. Alternatively, in the cases where the given data is not already in this format, Δ is derived automatically by the GUI. When the data is in similarity format, Δ=max(**S**)-**S** or 1-**S**. When a **Z**:($n \times m$) matrix is provided (where m is the number of variables of the data), Δ must be calculated with one of a number of distance calculation methods. The MDS-GUI currently provides for the following metrics: Euclidean, Weighted-Euclidean, City-Block, Mahalanobis, Minkowski, Canberra, Divergence, Bray-Curtis, Soergel, Bhattacharyya, Wave-Hedges, Angular Seperation and Correlation.

3.2 Layout

The frontend of the MDS-GUI is shown in Figure 1. The layout of the GUI was set out in such a way that it resembles a look and feel common to many Microsoft Windows based programs, to which most users are accustomed.

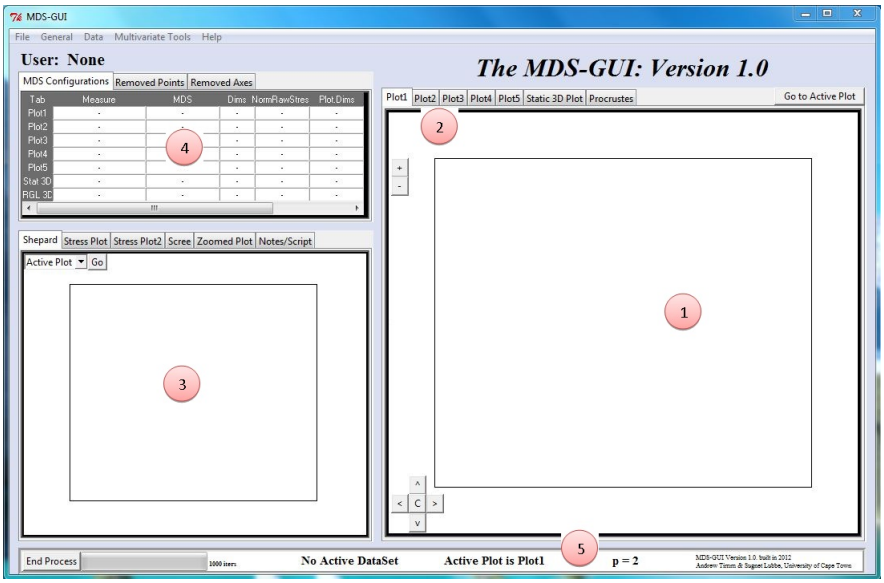


Figure 1: The MDS-GUI

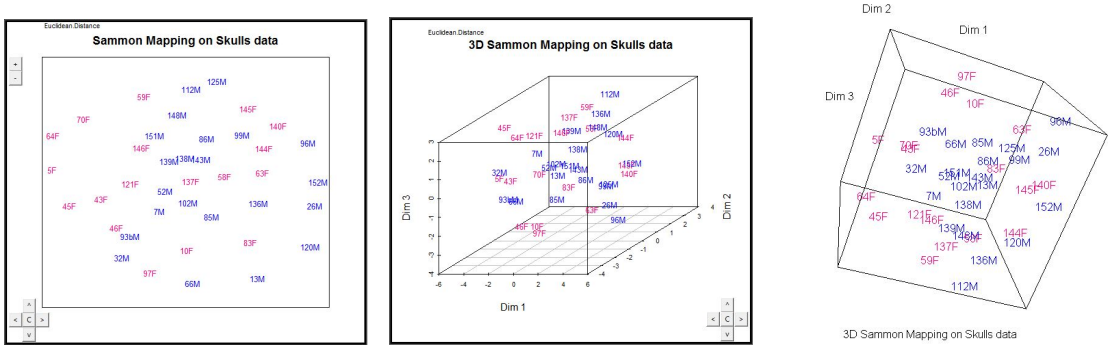
There are five major areas that make up the frontend, as indicated by the numbered markers in Figure 1. Each of these areas are described as follows: **(1) Main Plotting Area:** Location for two dimensional output of any MDS procedure. The output configuration represents Euclidean distances between points from **X** (Where **X** is the $n \times p$ MDS coordinate matrix). The aspect ratio of this area is strictly one as the relative distance must be equal over all plotting dimensions. **(2) Plotting Tabs:** Five plotting environments are available. The user is thus able to perform separate analyses in one instance of the program.**(3) Secondary Plotting Areas:** Houses multiple diagnostic plots for the MDS output. These include: Shepard Diagram, Scree Plot, Stress Plot, Logged-Difference Stress Plot, Zoomed Area and a Notes/Scripting area. **(4) Table Section:** Holds three tables, each providing information for all plotting areas. This allows a direct numerical comparison between all current plots. **(5) Information Panel:** This area displays information relevant to the applications of the user. The pane includes information regarding the data set being used, the current plotting area and the software developer details.

3.3 Features

The MDS-GUI offers a wide range of analytical features, allowing the user to produce detailed results and make thorough assessments of them. Most of these features will be mentioned here, and a small number graphically demonstrated. Demonstration of these features will use the *Skulls* data(Fawcett, 1901).

The first features to be mentioned are those relating to how the MDS configurations are displayed by the MDS-GUI. The software makes visual provisions for configurations when $p \in \{1,2,3\}$. Figure 2 demonstrates the two and three dimensional results. Plotting in three dimensions provides the user with two possible options. The first is a static plot using the **scatterplot3d** package by Ligges and Mächler (2003) and the second is producing a dynamically controlled plot using the **rgl** package by Adler and Murdoch (2011). In the cases of the two dimensional and static three dimensional plots, the output is displayed in area one of Figure 1. When $p > 3$, the user is given the option to either display a subset of

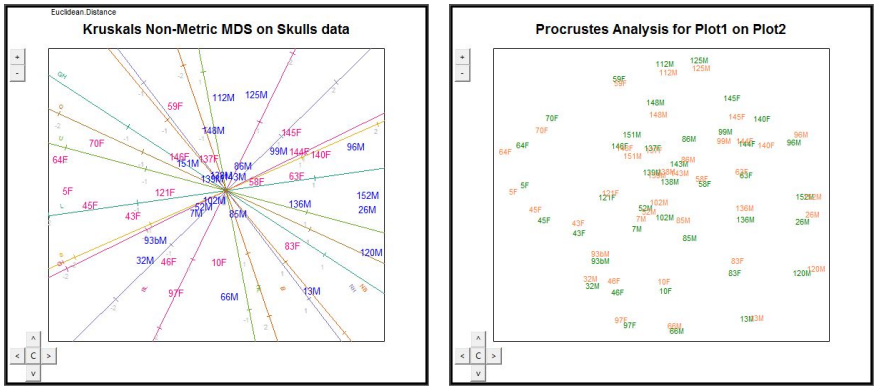
the dimensions in a two or three dimensional space, or simply observe only the coordinate matrix, $\mathbf{X}$. All three plots in Figure 2 distinguish between the male and female skulls by use of colour coding. In order to define such categories, a categorical variable column is added to the data and identified upon uploading the data to the GUI. Once a configuration, or number of configurations, has been achieved, a host of



(a) 2D Configuration (b) Static 3D Configuration (c) RGL 3D Configuration

Figure 2: Configuration Output

features are available to adjust its display. These include: Manual point location alterations with the mouse cursor; relocation of a group of points with relative positions maintained; comprehensive zoom capabilities, using frontend buttons, the keyboard '+' and '-' keys, or advanced zooming through a menu; Rotation and Reflection; Point Colour coding, either manually or through a menu where categories of the data may be coded; and Point Labeling. In addition to this, full control over the standard base plotting settings of *R* are controllable via the GUI, including 'col', 'cex' and other 'par' parameters. Two further features relating to the configuration are the options to display the variable axes of the data and to perform Procrustes Analysis on two different configurations. Examples of these are shown in Figures 3(a) and 3(b) respectively. Displaying variable axes, as shown in Figure 3(a), is useful in



(a) Variable Axes (b) Procrustes Analysis

Figure 3: Analytical Features

interpreting the association between the variables of the data on each point in the configuration. This feature is only available if the input data is in the form of a samples $\times$ variables data matrix $\mathbf{Z}$. Each axis relates to a single variable in the data and runs through the origin of the plot. The regression prediction biplot axes method is used to construct these axes. For more information see Gower and Hand (1996), section 3.3.2. These axes have positive and negative ends, and the correlation of variables can be assessed according to how they run in comparison to each another.

The Procrustes Analysis, Figure 3(b), is found in the *Procrustes* Tab in area two of Figure 1. The analysis is used to assess the degree of similarity or dissimilarity between two active configurations in *Plots 1-5*. If the selected configurations are represented by the matrices $\mathbf{X}_i : n \times 2$ and $\mathbf{X}_j : n \times 2$, then the Procrustes plot finds an orthogonal matrix $\mathbf{Q} : 2 \times 2$ to minimise the criterion $tr[(\mathbf{X}_i - \mathbf{X}_j\mathbf{Q})(\mathbf{X}_i - \mathbf{X}_j\mathbf{Q})']$. In the Procrustes tab $\mathbf{X}_i$ and $\mathbf{X}_j\mathbf{Q}$ are plotted simultaneously. This is useful in observing how two MDS methods differ in the assigning of the configuration.

While many other features of the MDS-GUI exist, they will only be listed here and not be described in detail. They are: Shepard Plot; Scree Plot; Stress Plots; Iterations Observable; Comprehensive Menu Structure; Settings Fully User Definable; Export Results to PDF; Save and Load Workspaces; Object Categories; Removing Points and Axes; Copy to Clipboard; Function Code Display; Notes and Scripting

Tab; Full Brushing Capabilities on both configuration and Shepard Plot; Real time changes of stress value on table during configuration alterations; altered configurations used as starting configurations. For full descriptions of all these features, the user is urged to consult the user manual of the MDS-GUI.

4. An Example: Morse-Code

Morse-Code is a universal, non spoken, means of transmitting messages. The code uses a series of long and short 'beeps' where every letter and number has its own sequence. These long and short signals are described as 'dashes' and 'dots' respectively. Rothkopf (1957) set out to determine the level of perceived similarity among the various coded sequences. The data he gathered has proved to be ideal for analysis by Multidimensional Scaling and has served as example data in Borg and Groenen (2005) and Buja et al. (2004), among others.

4.1 Morse-Code Data

The study done by Rothkopf (1957) involved the collection of confusion data from 598 subjects identifying the audio similarity between 36 Morse code signals (26 letters, 10 numbers). The result of this was a 36×36 asymmetric matrix. As with many MDS programs, the functions of the MDS-GUI require any dissimilarity/similarity matrix be symmetric. The adapted symmetric version of the square similarity matrix (also provided by Rothkopf) is therefore used. Each element of the matrix represents the percentage of respondents that determined the signal pairing to be 'the same'.

4.2 Method and Results

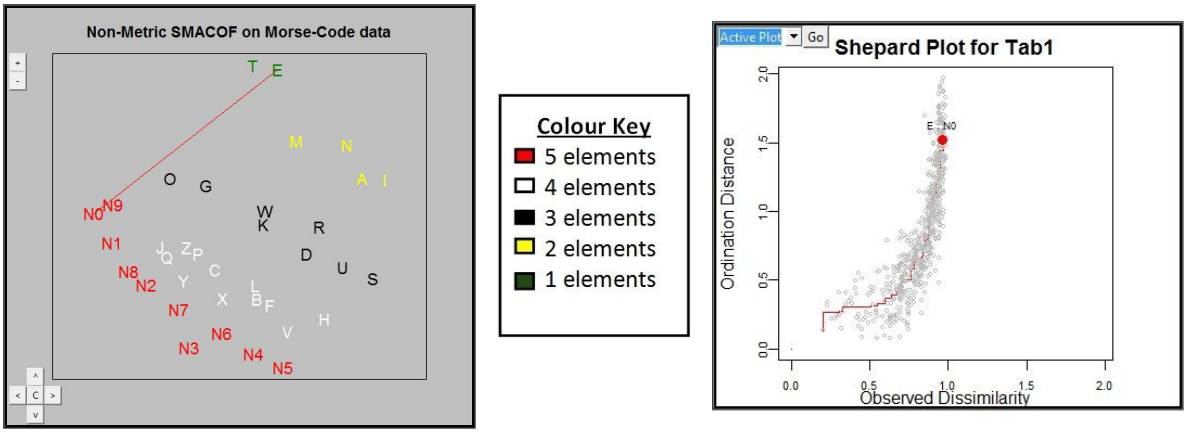
Analysis of the Morse-Code data using the MDS-GUI first requires the data be uploaded into the program. Most regular *R*-users will be aware of the data structure requirements of *R* when uploading data, and these standards also apply to the MDS-GUI. Seeing that the data already comes in the form of the $n \times n$ similarity matrix, the data is loaded through the *Load Similarity Matrix* command in the *Data* menu. The user is then prompted to name their data appropriately. As the data is already in the form of the proximity matrix, all features of the program relating to variables of the data are automatically deactivated and therefore unavailable to the user for the analysis of the Morse-Code data. All processes within the MDS-GUI require input of a proximity matrix in the form of dissimilarity measurement: the Morse-Code data is therefore automatically converted, by the software, into the appropriate format.

Tab	Measure	MDS	Dims	NormRawStres	Plot.Dims	Tolerence	Iterations
Plot1	NA	ClasScal	2	0.741	1&2	-	-
Plot2	NA	M.Smac.Sym	2	0.099	1&2	1e-05	139
Plot3	NA	Met.LeastSqs.	2	0.647	1&2	1e-05	139
Plot4	NA	NM.Smac.Sym	2	0.039	1&2	1e-05	96
Plot5	NA	Kruskal	2	0.04	1&2	1e-05	-
Plot1	NA	Sammon	2	0.09	1&2	1e-05	-

Figure 4: Morse-Code Data: MDS-GUI Table

A starting point of many analyses using Multidimensional Scaling is reviewing the results of multiple MDS methods in order to select that which is the most appropriate. The MDS-GUI allows for this process to be performed swiftly and simply, as with multiple plotting areas, the user is able to compare stress values, in different formats, between the various results directly. Figure 4 (A screen-shot of the table area of the MDS-GUI) shows the Normalised Raw Stress (NRS) values achieved by the various methods for $p = 2$. In all instances where a starting configuration is required, the Classical Scaling result is used as is standard practice. The MDS-GUI also allows the user to specify a starting configuration other than the default, either a completely random configuration, or a configuration already found in any of plotting areas. Inspection of the table reveals that, the Non-Metric SMACOF method produced the lowest stress value, and will therefore be used in the analysis. The stress value of 0.039 is considered a very good result and the configuration produced will be considered satisfactory.

Figures 5(a) and 5(c) show the configuration obtained from performing Non-Metric SMACOF on the Morse-Code data when $p = 2$, and the Shepard Diagram relating to the configuration, respectively. The Shepard Diagram is a very useful diagnostic tool when using MDS. Each of the points in the Shepard Diagram represents a pairing of points in the configuration, thus in this case, there are $\binom{36}{2} = 630$ points in the plot. The horizontal axis of the plot represents the observed dissimilarities, δ, while the vertical axis depicts the ordination based distances from the configuration, d. The researcher may then observe the accuracy of each pairing in the MDS configuration in terms of their observed value. When Metric MDS is used, the extent of deviation of the configuration is measured by the extent the points deviate



(a) MDS Configuration (b) Colour Key (c) Shepard Diagram

Figure 5: Morse-Code Data: Non-Metric SMACOF Results

vertically from the diagonal line, as an ideal configuration will have all $d_{ij} = \delta_{ij}$. However, in all Non-Metric cases such as this, only the ordering of points is relevant, therefore, the extent of deviation is measured by the extent that the points vertically deviate from the isotonic regression function, shown by the red line in Figure 5(c). i.e. the ideal configuration will have $d_{ij} = \hat{d}_{ij}$. Interpretation of this Shepard Diagram reveals that the Non-Metric SMACOF method has overstated the pairings with higher observed distances considerably. It is clear to see the effect of relaxing the metric condition, as a metric configuration producing a similar Shepard Diagram would incur a very high stress value and would be considered a poor result. Since the method was non-metric however, overstating distances is non influential, and the excellent stress value was achieved through the MDS maintaining the ordering of original proximities.

The configuration itself, as shown in Figure 5(a), reveals some interesting observations. In order to conveniently analyze the points, the use of categorising the data is appropriate. The category definition found to be most effective with the Morse-Code data was determined by identifying the fact that the sequences making up the code for the symbols range from length one to length five. A valid hypothesis therefore might be that subjects taking part in the experiment may be found to be more likely to incorrectly identify sequences of similar lengths as being similar. Each object in the data is therefore defined according to its Morse-Code sequence length in a categorical column added to the data. The MDS-GUI is then prompted to identify that column as categorical information. Displaying of all configurations therefore distinguishes each point according to its sequence length by means of colour-coding. The colour Key for the configuration is provided in Figure 5(b). Investigation of Figure 5(a) suggests that this hypothesis is likely to be justifiable. It can be observed that each grouping is clearly defined, for the most part, according to the length of the sequence, i.e. all objects of sequence length five have been grouped together, etc. Furthermore, the ordering of the groupings follows the natural progression of the sequence length, such that sequence length four grouping is between sequence length three grouping and sequence length five grouping, and so on.

Another useful feature of the MDS-GUI is the indexing link between the Shepard Plot and the Configuration Plot which are found in areas, 3 and 4 of Figure 1. Selecting any point of the Shepard Diagram with the mouse cursor will not only identify the associated object pairing on the Shepard Plot itself, but will also plot a line between the points on the configuration plot. Each highlighted point on the Shepard Plot is assigned its own colour, which corresponds to the line colour incorporated into the configuration. This allows for convenient differentiation and interpretation of the results. The red line in Figure 5(a) highlights the length between the two points ‘N0’ (Number zero) and ‘E’. This particular point pairing is of interest as the two sequences are intuitively the least similar in the set, with ‘E’ defined as ‘dot’ and ‘N0’ defined as ‘dash dash dash dash dash’. This is confirmed by it’s very low value of similarity in the data. It is however noted from the Shepard Plot that the pairs ordination distance is far from the highest in the MDS configuration (this is despite having its ordination distance far greater than its own observed dissimilarity). Individual cases such as these can be assessed in turn by interested parties in order to understand the composition of the stress value. This observation is merely pointed out in the interest of diagnostic completion. In practice, since the stress value is adequately low, such misrepresentations may be overlooked.

4.3 Conclusion

The study of the Morse-Code data using Multidimensional Scaling techniques with the MDS-GUI revealed the following. The MDS method found to be most effective for the data was Non-Metric SMACOF scaling, which produced a Normalised Raw Stress value of 0.039. The non-metric nature of the procedure allowed for a relaxation of the metric assumptions and therefore the produced configuration achieved an optimum result by overstating the majority of the paired object distances. The configuration itself showed clear distinction between groupings of the data based on the length of the Morse-Code sequence of each object. This result strongly suggests that subjects involved in the experiment conducted by Rothkopf (1957) were more inclined to incorrectly identify two sequences as 'the same' when they were of an equal sequence length.

5. Discussion

The MDS-GUI was primarily designed to aid those interested in performing Multidimensional Scaling and who do not necessarily have the *R* programming expertise or the time to learn the relevant *R* packages. Multidimensional Scaling was found to have its origins in applications of psychometrics and the social sciences. Since then, relevant fields have included ecology, marketing and biometrics. Researchers in these specific fields may not have the required *R* skills or statistical knowledge to perform Multidimensional Scaling effectively, and the MDS-GUI is intended to be a useful tool in these situations.

This paper set out to provide an introductory look at the visual applications available in the MDS-GUI. While a non-detailed list of the features of the software was provided, interested parties are urged to seek out the journal paper "MDSGUI: A package for comprehensive Multidimensional Scaling Analysis in R" (Timm and Gardner-Lubbe, 2013). The journal paper provides a more detailed analysis of the work covered in this paper and extends the demonstration with further case studies.

References

Adler, D. and Murdoch, D. (2011). *rgl: 3D visualization device system (OpenGL)*. R package version 0.92.798.
URL: *http://CRAN.R-project.org/package=rgl*

Borg, I. and Groenen, P. J. F. (2005). *Modern Multidimensional Scaling: Theory and Applications Second Edition*, Springer, New York.

Buja, A., Swayne, D. F., Littman, M., Dean, N. and Hormann, H. (2004). *Interactive Data Visualization with Multidimensional Scaling*. University of Pennsylvania.

Cox, T. F. and Cox, M. A. (2001). *Multidimensional Scaling: Second Edition*, Chapman and Hall, Boca Raton.

Fawcett, C. D. (1901). A second study of the variation and correlation of the human skull, with special reference to the naqada crania., *Biometrika* **1**: 408–467.

Gower, J. C. and Hand, D. J. (1996). *Biplots*, Chapman and Hall, London.

Grosjean, P. (2011). *SciViews-R: A GUI API for R*, UMONS, Mons, Belgium.
URL: *http://www.sciviews.org/SciViews-R*

Kruskal, J. B. (1964). Multidimensional scaling by optimising goodness of fit to a nonmetric hypothesis, *Psychometrika* **29**: 1–27,115–129.

LeRoux, N. J. (2012). SMACOF R code for metric and non-metric algorithms. Personal Communication.

Ligges, U. and Mächler, M. (2003). Scatterplot3d - an R package for visualizing multivariate data, *Journal of Statistical Software* **8**(11): 1–20.
URL: *http://www.jstatsoft.org*

R-Development-Core-Team (2011). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
URL: *http://www.R-project.org/*

Rothkopf, E. Z. (1957). A measure of stimulus similarity and errors in some pairedassociate learning, *Journal of Experimental Psychology* **53**: 94–101.

Tierney, L. (2011). *tkrplot: TK Rplot*. R package version 0.0-20.
URL: *http://CRAN.R-project.org/package=tkrplot*

Timm, A. W. and Gardner-Lubbe, S. (2013). MDSGUI: A package for comprehensive multidimensional scaling in R, *In press* .

Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S: Fourth Edition*, Springer, New York.
URL: *http://www.stats.ox.ac.uk/pub/MASS4*