# Topological Mapping and Multidimensional Scaling

Brian D. Ripley

*Professor of Applied Statistics*
*University of Oxford*

## References

BDR (1996) *Pattern Recognition and Neural Networks.* Cambridge University Press. *Chapter 9.*

Cox, M.A.A. & Cox, T.F. (1994) *Multidimensional Scaling.* Chapman & Hall.

Kohonen, T. (1995) *Self-Organizing Maps.* Springer.

## Motivation

Kohonen describes his own motivation as:

> 'I just wanted an algorithm that would effectively map similar patterns (pattern vectors close to each other in the input signal space) onto contiguous locations in the output space.'
>
> Kohonen (1995, p. VI)

He interpreted 'contiguous' via a rectangular or hexagonal 2-D lattice; multidimensional scaling had much earlier used a 2-D (Euclidean) space and interpreted 'contiguous' as 'nearby'.

Multidimensional scaling (MDS) is a family of techniques to map points in feature space into $\mathbb{R}^k$ (usually for $k = 2$) to approximate desirable aspects of the distances between points.

Which aspects are thought desirable varies between methods. The best known outside psychometrics and statistics is *Sammon mapping*, which is one of the crudest methods.

## Two Examples

Both from BDR (1996), not quite used for their original purposes.

### Viruses

This is a dataset on 61 viruses with rod-shaped particles affecting various crops (tobacco, tomato, cucumber and others) described by Fauquet *et al.* (1988). There are 18 measurements on each virus,the number of amino acid residues per molecule of coat protein; the data come from a total of 26 sources. We consider the 38 *Tobamoviruses* for subgroups.

## *Leptograpsus* Crabs

Campbell & Mahon (1974) studied rock crabs of the genus *Leptograpsus*. One species, *L. variegatus*, had been split into two new species, previously grouped by colour form, orange and blue. Preserved specimens lose their colour, so it was hoped that morphological differences would enable museum material to be classified. Carapaces (shells) may need to be sexed.

Data are available on 50 specimens of each sex of each species, collected on sight at Freemantle, Western Australia. Each specimen has measurements on

- the width of the frontal lip of the carapace
- the rear width,
- length along the midline,
- the maximum width of the carapace,
- body depth

A $\log$ transformation was used for all the variables.

# Dissimilarities

The Kohonen algorithms lie on finding the closest 'reference vector' $\mathbf{m}_i$ in feature space. How 'closest' is measured is not discussed, but the code uses Euclidean distance. May need more general measures if the features are not purely continuous (and if they are highly correlated).

*Dissimilarities* generalize distances; they are non-negative and symmetric, but need not satisfy the triangle inequality. Lots are known, e.g. Jaccard matching coefficient, Mahalanobis distance, Manhattan distance and combinations too.

We will assume that we are given a matrix $(d_{rs})$ of dissimilarities between pairs of objects, and $d_{rr} = 0$ for all $r$.

# MDS Methods

Seek a configuration of points in $\mathbb{R}^k$ with Euclidean distances $(\widetilde{d}_{rs})$ close to $(d_{rs})$ in some sense.

## Classical scaling

Based on recovering exactly a Euclidean configuration from $(d_{rs})$ in high dimensions, then projecting to minimize $\sum_{r,s}\left[d_{rs}^2 - \widetilde{d}_{rs}^2\right]$. From 1930s and 1960s.

## Sammon mapping

Minimize

$$E(d, \widetilde{d}) = \sum_{r \neq s} \frac{(d_{rs} - \widetilde{d}_{rs})^2}{d_{rs}}$$

so emphasis on preserving short distances. Sammon's (1969) algorithm is a poor minimizer, and there are often lots of local minima.

## Ordinal ('non-metric') methods

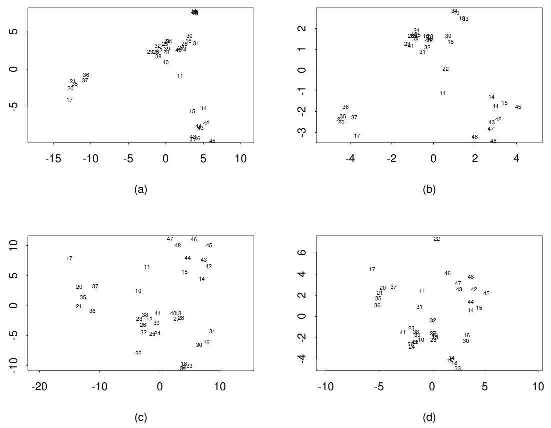
Choose increasing function $\theta$ and configuration with distances $(\delta_{rs})$ so that ordering of distances is similar, by minimizing

$$\frac{\sum_{r,s}[\theta(d_{rs}) - \delta_{rs}]^2}{\sum_{r,s}\delta_{rs}^2}.$$

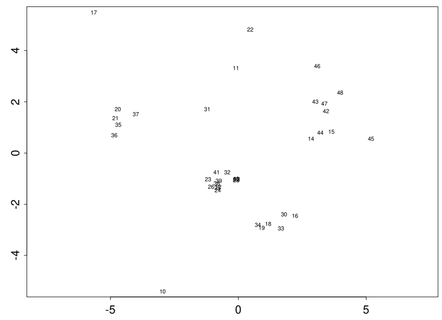
Idea due to psychologist R. N. Shepard (1962); practical method by Joe Kruskal (1964).

# MDS Examples

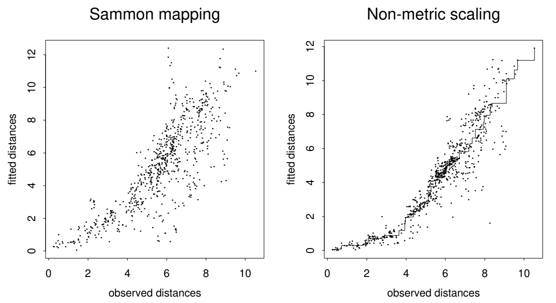
## Viruses



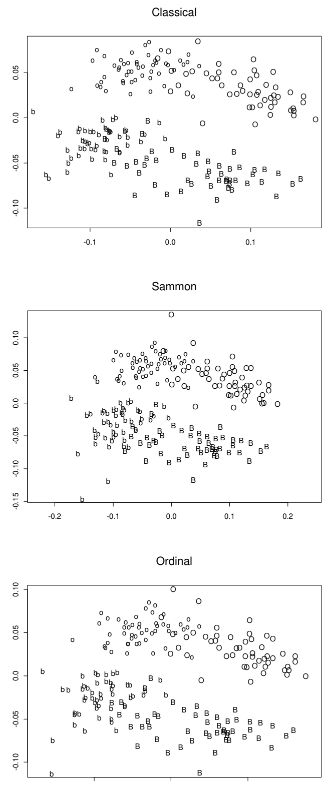

Principal component (top row) and Sammon mapping (bottom row) plots of the viruses. The plots in the left column are of the raw data, those in the right column with variables rescaled to have unit variance.



Non-metric multidimensional scaling plot of the *Tobamovirus* group of the viruses example. The variables were scaled before Euclidean distance was used.



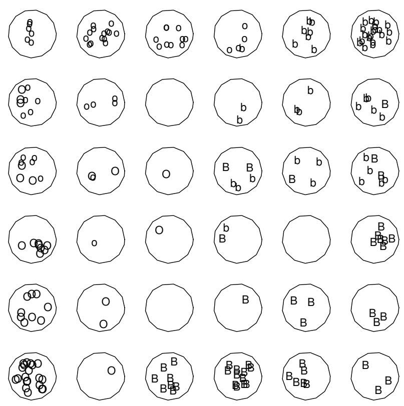

Distortion plots of Sammon mapping and non-metric multidimensional scaling for the viruses data. For the right-hand plot the fitted isotonic regression is shown as a step function.



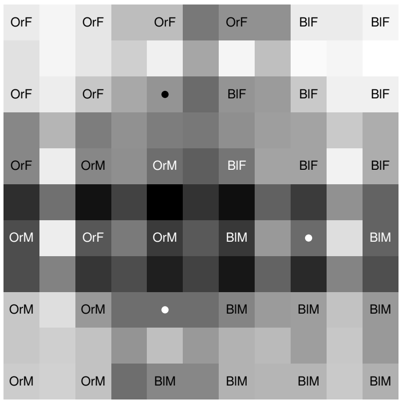

Classical, Sammon and non-metric multi-dimensional scaling plots for the crabs data.

# SOM Examples

## Crabs

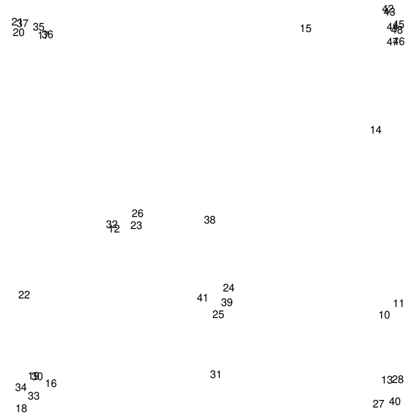


SOM mapping of the crabs data to a $6 \times 6$ grid. The labels of those examples mapped to each cluster are distributed randomly within the circle representing the cluster. As before the coding is upper case for males, lower case for females, 'B' for blue and 'O' for orange.




Ultsch representation of the SOM representatives for previous figure. The rows and columns between the units represent the magnitude of the gradient (black being high); the greylevel at each unit represents the median of the surrounding gradients. The label is the most common class mapped to that representative.

## Viruses



SOM mapping of the viruses data to a $5 \times 5$ grid.

# Automated Sleep EEG Analysis

From

Roberts, S. & Tarassenko, L. (1995) Automated sleep EEG analysis using an RBF network. In *Neural Network Applications*, ed. A. F. Murray, pp. 305–322. Kluwer.

Coefficients of AR(10) fit to 128-sample frames of EEG produced every second.

Populated cells of $10 \times 10$ SOM mapping of AR coefficients.