Multidimensional Scaling (MDS)

Multidimensional Scaling (MDS)

Multidimensional scaling (MDS) is an extended family of techniques that try to reproduce the relative positions of a set of points in a reduced space given, not the points themselves, but only a matrix with interpoint distances (dissimilarities). These distances might be measured with error, or even be non-Euclidean.

Metric Multidimensional Scaling (mMDS)

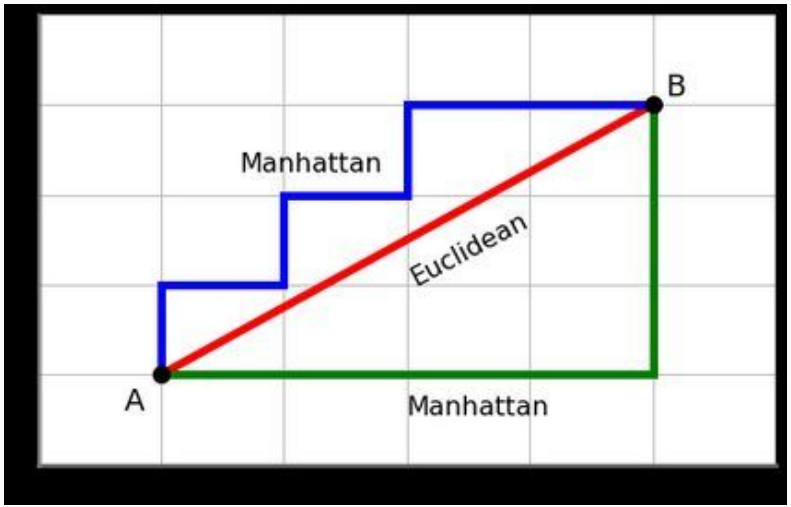
Principal Coordinates Analysis (PCO/PCoA) is one metric scaling technique (it is sometimes called classical or Torgerson scaling).

Not to be confused with Principal Components Analysis (PCA), where the “components” are new, uncorrelated variables that are constructed as *linear combinations* of the initial variables.

Principal Coordinate Analysis (PCO/PCoA)

Similar to clustering - finds what is similar/dissimilar - but tries to extract two dimensions to make a 2D map

Can use any real distances (i.e., Euclidean, Manhattan, etc.)



Principal Coordinate Analysis (PCO): also known as classical multidimensional scaling (classical MDS).

Here we start with distance matrix (compared with data matrix for PCA) and typically a transformation is applied to the distance matrix. We then essentially proceed as per PCA, i.e. calculate eigenvalues and eigenvectors (If the distance matrix is based on Euclidean distances, PCO and PCA are essentially equivalent!)

Principal Coordinate Analysis (PCO): also known as classical multidimensional scaling (classical MDS).

PCO is most often used when only distances are available (distances are not required to be Euclidean). If there are n dimensions (e.g. cities) in the distance matrix, there will be at most n−1 principal coordinates (compare with PCA, i.e. as many PCs as there are variables in the data matrix).

[MDS North Island example]

Correspondence Analysis (CA)

CA is a special case of mMDS where the distance measure is the chi-square distance.

It is conceptually similar to PCA but where the data are categorical, counts, rather than continuous.

CA is traditionally applied to contingency tables where rows and columns are treated equivalently.

Correspondence analysis is usually the best way to follow up on a significant chi-square test.

[CA HairEyeColor example]

Non-metric Multidimensional Scaling

NMDS relies on rank orders (distances) for ordination based on their relative similarities (or dissimilarities)

Any dissimilarity coefficient or distance measure may be used to build the distance matrix used as input

NMDS makes very few assumptions about the data

Non-metric Multidimensional Scaling

NMDS attempts to embed the points in a new space such that the order of the reconstructed distances in the new map is the same as the ordering of the original distance matrix.

Same aim as metric MDS but fewer assumptions!

[Eckmans colour perception example]

Method selection (choices…)

PCO and metric MDS will usually provide almost identical answers if Kruskal's STRESS 1 is used

Metric or non-metric?

Metric has few advantages over Principal Coordinates Analysis (unless many negative eigenvalues)

Non-metric does better with fewer dimensions but can be more prone to suboptimal solutions

How many dimensions?

STRESS <10% is "good representation"
Scree diagram (make your choice)
Two (or three) dimensions for visual ease

Conclusions

- Distance methods are useful when data consist of associations (similarities/distances) among observations
- There are many measures of distance, and their choice, much like transformations, is important in the outcome of the analysis
- Many multivariate methods are just special cases of one another, with special names to match
- PCO is the analytical equivalent of a PCA on the covariance matrix (and allows us to reconstruct the reduced space plot when raw data are not available)
- MDS is an iterative regression-based method using STRESS
- Non-metric MDS is a non-parametric version of MDS