Index of 8-L

Practical Information

Course schedule:

- 3 presentations next Friday: Teri, Ibrahim, John,
- an (informal) course evaluation form will be posted and distributed.

Today's session:

- brief review of Quiz 2,
- general course wrap-up:
  - Q & A on course material,
  - discussion of course content and organization,
- multidimensional scaling (Manly: Chapter 10(11)[1]),
  - data example: dogs,
  - new data example: road distances between towns on South Island, NZ.

Webpage changes:

- lecture 7 on canonical correlation analysis revised (twice); recommended to download latest version,
- road distance data (from Manly) added.

[1] Manly 3rd ed. includes a section in Chapter 12 on Principal Coordinates Analysis, which is a particular version of MDS.

Multidimensional Scaling: First Steps

Objective of MDS — to:

- seek a configuration in d-dimensional space[2] such that distances between points best match a distance matrix,[3]
- construct a diagram showing the relationships between a number of objects, given only a distance matrix between the objects (Manly),
- visualize the level of similarity of individual cases of a dataset.

Inputs for MDS may be:

i) distance (dissimilarity) matrix between observations, or

ii) variables measured for all observations from which distances are to be computed, e.g. using Euclidean distance (Session 5).

Versions/Algorithms of MDS:

1) "classical" MDS: translates the problem into a matrix analysis problem with eigenvalues and eigenvectors $\Rightarrow$ some links with other methods covered in course,

2) "modern" MDS: involves iterative optimization a particular function over possible configurations; includes the classical method as a special case.

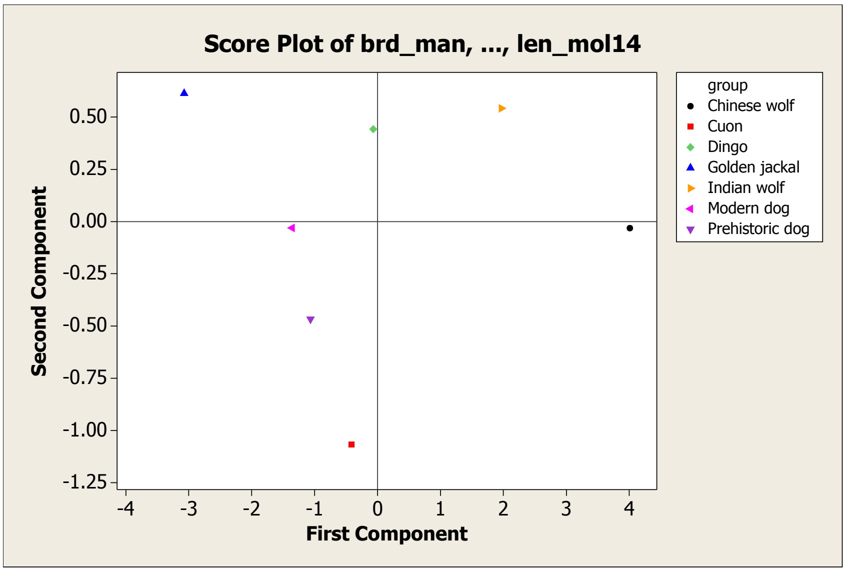
[2] Intuitively, n observations can be represented perfectly in $(n-1)$-dimensional space, provided the distances are "sensible", but for spaces of lower dimensions some loss of fit is expected.

[3] Venables & Ripley, *Modern Applied Statistics*, 3rd ed., p. 333.

MDS Example: Dogs

<u>Summary (dogs)</u>: 6 jaw measurements for each of 7 dog-related species; interest is in closeness between species, using Euclidean distance for standardized variables (5L—3):

- smallest distance between prehistoric and modern dog, largest distance between Chinese wolf and golden jackal,
- CA depicts distances between clusters, not observations,
- PCA can extract two components, whose scores are then plotted against each other:



— turns out to be exactly the same as classical MDS based on Euclidean distance for standardized variables,

  * same decomposition of variance by eigenvalues, [4]
  * same eigenvectors up to scaling and reflection,

- <u>summary of results</u>: very good separation of points by first component only.

[4] The eigenvalues are multiplied by $(n-1)$, where $n =$ no. of obs.

Classical MDS Algorithms

First idea: translate the problem into a matrix analysis framework for a suitably defined matrix B (details in Manly) whose eigenvalues/eigenvectors are computed,

* isses with matrix B: may not be positive definitive ($\Rightarrow$ negative eigenvalues) when distance matrix is not Euclidean,
* for Euclidean distance, also called *principal coordinates analysis*,
* for Euclidean distance on *standardized* variables, equivalent to PCA for correlation matrix,
* equivalent to modern MDS with specific ("strain") loss function (below).

Second idea: a better representation of fit is achieved by "adapting" the distances (or *dissimilarities*) between observations (δ_{ij}) to the point configuration distances (d_{ij}), e.g. by a simple linear regression,

$$d_{ij} = \beta_0 + \beta_1 \cdot \delta_{ij} + \varepsilon_{ij}, \quad 1 \leq i < j \leq n, \tag{1}$$

in which case the fitted values, ($\hat{d}_{ij}$), minimize Kruskal's stress measure, a so-called *loss function*,

$$\text{"Stress}-1\text{"}^2 = \sum_{ij} (d_{ij} - \hat{d}_{ij})^2 / d_{ij}^2.$$

* ($\hat{d}_{ij}$) are adjusted/transformed dissimilarities, or *disparities*,
* "Stress$-$1" measures goodness-of-fit, and small values (close to 0) are preferable; e.g. one may make decisions about reducing the number of dimensions on the resulting "Stress$-$1" value.[5]

[5] Stata displays "Mardia fit measures", which are simply proportions of cumulative eigenvalues, or squared eigenvalues.

Modern MDS Algorithms

Why is there a need to improve/extend the method?
— may be difficult to get good low-dimensional representations of large datasets (n large).

New idea: seek optimal configuration of points (in d-dim. space) to minimize goodness-of-fit criterion, for larger classes of solutions:

- metric scaling: extend linear relation in Eq. (1) to power or polynomial relations,
- nonmetric scaling: extend linear relation in Eq. (1) to general monotonic[6] functions $f : \delta_{ij} \mapsto f(\delta_{ij})$,
- other loss functions (e.g., "strain" loss) are possible and may work better with generalizations of Eq. (1).

Outline of modern algorithmic approach (following Manly):

1) set up start configuration Y for n objects in d-dim. space (where d is fixed for the algorithm),
2) compute the Euclidean distances $d_{ij} = d_{ij}(Y)$,
3) regress by Eq. (1), or extensions hereof, the distances d_{ij} on the true distances (dissimilarities) δ_{ij}, to get the $\hat{d}_{ij}$ (disparities),[7]
4) compute a stress (goodness-of-fit) statistic based on (d_{ij}) and $(\hat{d}_{ij})$,
5) move the configuration Y in a direction to reduce stress,

— loop through 1)−5) until no further improvement is possible.[8]

[6] Regression for monotonic functions also go under the name *isotonic* regression.
[7] Thus, "disparities are scaled to match the configuration distances (d_{ij}) as closely as possible". (Manly)
[8] In the iterative minimization process, care must be taken to avoid local minima; see Stata manual for details.

MDS Example: Road Distances

Data: matrix of road distances (in miles) between 13 towns on the South Island of New Zealand; interest is in (for demonstration purposes) reconstructing the physical (2-D) map, up to rotation and scaling,

- data in form of distance matrix $\Rightarrow$
  need ways to input/specify such data in software...

Results for different MDS algorithms:

- classical: 2 eigenvalues explain majority of var.; "Stress$-$1" $= 0.09$ (small), config. close to Manly's Fig. 11.4,
- modern, metric: power transform does not appear very useful (very small transformation power, convergence problems),
- modern, nonmetric: some improvement in loss function; no obvious difference/improvement in config. plot; some improvement in Shepard diagram:
  - plot of observed dissimilarities against disparities,[9]
  - observations close to line $\sim$ good fit.
- overall conclusions:
  - 2-D reconstruction successful,
  - no striking gains in modern over classical MDS.

[9] For simple regression, the plot is equivalent to plotting observations against fitted values.