Low-dimensional Representation

J.H. Maindonald[*]

August 24, 2007

1 Ordination

Ordination is a generic name for methods for providing a low-dimensionaL view of points in multi-dimensional space, such that "similar" objects are near each other and dissimilar objects are separated. The plot(s) from an ordination in 2 or 3 dimensions may provide useful visual clues on clusters in the data and on outliers.

If data can be separated into known classes that should be reflected in any ordination, then the scores from classification using `lda()` may be a good basis for an ordination. Plots in 2 or perhaps 3 dimensions may then reveal additional classes and/or identify points that may be misclassified and/or are in some sense outliers. It may indicate whether the classes that formed the basis for the ordination seem real and/or the effectiveness of the discrimination method in choosing the boundaries between classes.

Here, the discussion will turn to multi-dimensional scaling (MDS) methods where distances are given, or that start by calculating distances between points, then using the distances as the starting point for an ordination. Similarities can be transformed into distances, though often with some arbitrainess in the way that this is done.

Examples are:

1. From Australian road travel distances between cities and larger towns, can we derive a plausible "map" showing the relative geographic locations?

2. Starting with genomic data, various methods are available for calculating genomic "distances " between, e.g., different insect species. The distance measures are based on evolutionary models that aim to give distances between pairs of species that are a monotone function of the time since the two species separated.

3. Given a matrix **X** of n observations by p variables, a low-dimensional representation is required, i.e., the hope is that a major part of the information content in the data can be summarized in a small number of constructed variables. There is typically no good model, equivalent to the evolutionary models used by molecular biologists, that can be used to motivate distance calculations. There is then a large element of arbritariness in the distance measure used.

In general, given a matrix **X** of n observations by p variables, results will depend strongly on the distance measure used. If Euclidean distances are used, what relative weight should be given to the different columns of **X**? Should logarithms of values be used, or should some other transformation be applied? A logarithmic scale makes sense for biological morphometric data, and for other data that has similar characteristics.

Various non-euclidean distances can be used. See the help page for the function `dist()`. The function `daisy()` in the *cluster* package offers a wider range of possibilities. Irrespective of the method used for the calculation of the distance measure, ordination methods typically yield a representation in Euclidean space. A particular issue is the calculalation of distances when some or all variables are not on an ordinal scale.

[*] Centre for Mathematics & Its Applications, Australian National University, Canberra ACT 0200, Australia.
mailto:john.maindonald@anu.edu.au

1.1 Distances

The starting point for discussion will be the calculation of Euclidean distances between points. The point of this subsection is to show how an **X**-matrix like representation can be recovered from a matrix of pairwise distances between points.

Given **X**, the squared Euclidean distance between points i and j is

$$\begin{aligned} d_{ij}^2 &= \sum_{k=1}^{p}(x_{ik}-x_{jk})^2 \\ &= \sum_{k=1}^{p} x_{ik}^2 + \sum_{k=1}^{p} x_{jk}^2 - 2\sum_{k=1}^{p} x_{ik}x_{jk} \end{aligned}$$

Thus

$$d_{ij}^2 = q_{ii} + q_{jj} - 2q_{ij} \tag{1}$$

where $q_{ii} = \sum_{k=1}^{p} x_{ik}^2; \quad q_{ij} = \sum_{k=1}^{p} x_{ik}x_{jk}$.

Observe that q_{ii} is the (i,j)th element of the matrix $\mathbf{Q} = \mathbf{XX}'$. Thus, the matrix $\mathbf{XX}'$ has all the information needed to construct distances.

It is convenient to assume a version of **X** in which columns have been centered, i.e.

$$\sum_{i=1}^{n} x_{ik} = 0, i = 1, \ldots p$$

This implies that

$$\begin{aligned} \sum_{i=1}^{n} q_{ij} &= \sum_{i=1}^{n}(\sum_{k=1}^{p} x_{ik}x_{jk}) \\ &= \sum_{k=1}^{p}(\sum_{i=1}^{n} x_{ik}x_{jk}) \\ &= \sum_{k=1}^{p}(x_{jk}\sum_{i=1}^{n} x_{ik}) \\ &= 0 \end{aligned}$$

i.e., that the rows and columns of **Q** sum to zero.

Given distances, find a low-dimension representation

It will now be shown that given distances d_{ij}, then equation 1 uniquely determines a matrix **Q** whose rows and columns sum to zero. This does not of course define a matrix **X** columns sum to zero uniquely. In particular, if **P** is an n by n orthogonal matrix, then **Q** is unchanged if we replace **X** by **XP**.

Set $A = \sum_{i=1}^{n} q_{ii}$. Summing $d_{ij} = q_{ii} + q_{jj} - 2q_{ij}$ over i, it follows that

$$\sum_{i=1}^{n} d_{ij}^2 = A + nq_{jj} \tag{2}$$

$$\sum_{j=1}^{n} d_{ij}^2 = A + nq_{ii} \tag{3}$$

$$\sum_{i=1}^{n}\sum_{j=1}^{n} d_{ij}^2 = 2nA \tag{4}$$

From equation 4

$$A = \frac{1}{2n}\sum_{i=1}^{n}\sum_{j=1}^{n} d_{ij}^2 \tag{5}$$

Adding equations 2 and 3

$$q_{ii} + q_{jj} = \frac{1}{n}(\sum_{i=1}^{n} d_{ij}^2 + \sum_{j=1}^{n} d_{ij}^2 - 2A)$$

Hence from equation 1

$$q_{ij} = -\frac{1}{2}d_{ij}^2 + \frac{1}{n}(\sum_{i=1}^{n} d_{ij}^2 + \sum_{j=1}^{n} d_{ij}^2 - 2A)$$

where A is given by equation 5.

Having thus recovered the symmetric matrix $\mathbf{Q}$, then the spectral decomposition yields

$$\mathbf{Q} = \mathbf{U\Lambda U}'$$

If the d_{ij} are genuine Euclidean distances, then the diagonal elements of $\mathbf{\Lambda}$ satisfy $\lambda_i \geq 0$. Thus, choose $XB = \mathbf{U\Lambda}^{\frac{1}{2}}$.

1.2 Ordination derived from a distance matrix – general

More generally, the matrix $\mathbf{Q}$ that is calculated as above will be positive semidefinite providing that the d_{ij} satisfy the triangle inequality, i.e.

$$d_{ij} \leq d_{ik} + d_{kj}$$

If $\mathbf{Q}$ is not positive semidefinite, the ordination can still proceed. However one or more eigenvalues λ_i will now be negative. If relatively small, it may be safe to ignore dimensions that correspond to negative eigenvalues. It is then more than otherwise desirable to check that the ordination reproduces the distances with acceptable accuracy.

Non-metric scaling

These methods all start from “distances”, but allow greater flexibility in their use to create an ordination. The aim is to represent the “distances” in as few dimensions as possible.

Often, it makes sense to give greater weight to small distances than to large distances. The distance scale should perhaps not be regarded as rigid. Larger distances may not be measured on the same Euclidean scale as shorter distances. The ordination should perhaps preserve relative rather than absolute distances.