MSc II: Further Statistical Analysis (Autumn Term)
Advanced Multivariate Analysis

1 Multidimensional Scaling

In general, multidimensional scaling techniques are used for *exploratory data analysis*. They are not normally based on any statistical model for the data, and so there are no associated tests of significance. Their basic purpose is *visualization*, to enable us to 'look' at the data and so form hypotheses about the relationships between the sample members, or the relationships between different samples. In addition we may hope to find a few dominant and *interpretable* dimensions among the possibly many dimensions required to represent the data exactly.

These techniques are especially useful when:

(a) data arise directly as an $n \times n$ matrix of 'distances' (dissimilarities) or similarities between pairs of data units, rather than as an $n \times p$ data matrix.

(b) data consist of a, possibly large, number of measurements that are quantitative, qualitative or a mixture of the two and it is convenient to combine these to form similarity or distance measures between the data units.

So the simplest form of multidimensional scaling starts with a given $n \times n$ distance matrix $D = (d_{rs})$ and seeks to find a configuration of n points in a, preferably low-dimensional, Euclidean space such that the inter-point distances reflect 'as closely as possible' the given dissimilarities or similarities.

If we can find a good representation in just 2 or 3 dimensions then we can plot the objects (data units) and obtain a good visual impression of the relationships existing between members of the sample. If we are fortunate we may also be able to *interpret* these dimensions. This is called reification. (See Arnold and Collins, Applied Statistics 42 (2) 1993.)
However we first need to consider what we mean by matching the given matrix $D = (d_{rs})$ 'as closely as possible'.

Some questions and Issues

(a) Can the distances (dissimilarities) be reproduced exactly?

The answer to this is yes - provided that the given distances are Euclidean in the first place. We shall develop a criterion for determining whether the given distances are Euclidean, and if they are then we shall show that they can be reproduced exactly in, at most, $n - 1$ dimensions. eg. Consider a 3×3 symmetric matrix of distances between 3 points:

$$\begin{array}{c} 1 \\ 2 \\ 3 \\ \\ \end{array} \begin{array}{c} \left[\begin{array}{ccc} 0 & & \\ 3 & 0 & \\ 5 & 4 & 0 \end{array}\right] \\ \begin{array}{ccc} 1 & 2 & 3 \end{array} \end{array}.$$

By simple geometry we can construct the equivalent triangle, thus representing the 3 points, *exactly* in 2 dimensions. Clearly we can do this here because the given distances satisfy the *metric inequality*. That is: $d_{13} \leq d_{12} + d_{23}$.

(b) Is an exact representation a *unique* representation?

The answer to this question is no. Clearly there can be no unique representation, $\{(x_1, y_1), (x_2, y_2), (x_3, y_3)\}$, of the 3 points in the above example. Such a coordinate representation in 2-dimensional Euclidean space is unique only up to a translation, rigid (orthogonal) rotation and reflection of the axes.

In general:

(i) The *position* of the origin is free. To fix its position we place the origin at the centroid.

(ii) The *orientation* of the axes is free within any orthogonal rotation. We shall choose to use the *principal axes* - so that the first dimension accounts for the maximum possible amount of the squared inter-point distances, and the remaining axes successively account for the maximum amount subject to orthogonality with the preceding axes. This is a sensible choice because, even if an exact representation exists in a far higher number of dimensions, we obtain the 'best' 1-dimensional, 2-dimensional, 3-dimensional ... representation by restricting to just the first one, first two, first three ... coordinates respectively.

(iii) The representation may be *reflected* without altering the representation of the distances. There is no standard way of selecting a particular reflection.

We shall consider two types of scaling.

(a) <u>Classical Scaling</u>
This method attempts to reproduce the given values of the distances (d_{rs}) as exactly as possible.

(b) <u>Ordinal Scaling</u>
This method attempts to reproduce just the *rank ordering* of the m given distances (d_{rs}) (where $m = \frac{1}{2}n(n-1)$) as closely as possible.

1.1 Classical Scaling (CS)

We start with an $n \times n$ distance matrix $D = (d_{rs})$.

<u>*Definition*</u>: An $n \times n$ matrix $D = (d_{rs})$ is called a *distance matrix* if it is symmetric and

$$d_{rr} = 0 \text{ for all } r = 1, \ldots, n \text{ and } d_{rs} \geq 0 \text{ whenever } r \neq s$$

<u>*Definition*</u>: An $n \times n$ distance matrix $D = (d_{rs})$ is said to be *Euclidean* if there exists a configuration of n points, $\mathbf{x}_1$, $\mathbf{x}_2$, ..., $\mathbf{x}_n$ in $I\!R^p$, for some p, such that

$$d_{rs}^2 = \sum_{k=1}^{p} (x_{rk} - x_{sk})^2 \;=\; (\mathbf{x}_r - \mathbf{x}_s)^T (\mathbf{x}_r - \mathbf{x}_s) \quad \text{(Pythagorean distance)}$$

[Aside: This definition may seem a little circular. What actually characterizes Euclidean distances? It is the transformations of the coordinate axes that leave the distances unaltered. ie translation, orthogonal rotation and reflection. These invariant transformations imply the Pythagorean formula in the definition.]

So, given any $n \times n$ distance matrix D:

(a) Is it Euclidean? We need a criterion to determine this.

(b) If it is Euclidean, how can we find a configuration of points reproducing these given distances?

(c) If it is <u>not</u> Euclidean, can we find a configuration of points whose Euclidean inter-point distances match the given (non-Euclidean) distances as closely as possible in some sense?

[Note: if distances are Euclidean then they satisfy the metric inequality (or triangle inequality)

$$d_{rs} \leq d_{rt} + d_{ts} \quad \text{for all points } r,\ s \text{ and } t$$

However the converse is <u>not</u> true. Satisfying the metric inequality is <u>not</u> a sufficient condition for distances to be Euclidean. We need some other criterion.]

We begin by <u>assuming</u> that D is Euclidean.

Thus there exists a configuration X, an $n \times p$ matrix, that represents the n points in terms of their coordinates in p dimensions such that the calculated Pythagorean distances

reproduce the given distance matrix D exactly.

$$\underset{n \times p}{X} = \begin{bmatrix} x_{11} & \dots & x_{1p} \\ x_{21} & \dots & x_{2p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \dots & x_{np} \end{bmatrix} = \begin{bmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_n^T \end{bmatrix}$$

Assume now that we have placed the origin at the centroid of this configuration so that column sums are all zero. That is

$$\sum_{r=1}^{n} x_{rj} = 0, \quad j = 1, \dots, p$$

So that

$$\mathbf{1}^T X = \mathbf{0}^T \text{ and } X^T \mathbf{1} = \mathbf{0} \tag{1}$$

[Note: If the configuration X does not have the origin as its centroid, we may use the centred configuration

$$X_c = H X$$

where H is the $n \times n$ centreing matrix $H = I - n^{-1}\mathbf{1}\mathbf{1}^T$. Check this and also check what happens when X is post-multiplied by the $p \times p$ matrix $H = p^{-1}\mathbf{1}\mathbf{1}^T$.]

The distances d_{rs} can be calculated directly from the coordinates given in X or we can do so via the $n \times n$ inner product matrix

$$XX^T = B = (b_{rs})$$

so that

$$b_{rs} = \mathbf{x}_r^T \mathbf{x}_s = \sum_{j=1}^{p} x_{rj} x_{sj} \tag{2}$$

Note that $B = XX^T$ is positive semi-definite (psd) by result (B.2) in the Notes for MSc Students. Also note that, because of (1)

$$\sum_{s=1}^{n} b_{rs} = \sum_{r=1}^{n} b_{rs} = 0 \quad \text{for all } r, s \tag{3}$$

or

$$B\mathbf{1} = \mathbf{0} \text{ and } \mathbf{1}^T B = \mathbf{0}^T \tag{3'}$$

(ie row and column sums of B are all equal to zero. We shall see shortly that B is independent of the particular coordinate representation of D - apart from the requirement that the origin is at the centroid.)

We can now calculate the distances d_{rs} via B.

By the definition of Euclidean distance

$$\begin{aligned} d_{rs}^2 &= (\mathbf{x}_r - \mathbf{x}_s)^T(\mathbf{x}_r - \mathbf{x}_s) \\ &= \mathbf{x}_r^T\mathbf{x}_r + \mathbf{x}_s^T\mathbf{x}_s - 2\mathbf{x}_r^T\mathbf{x}_s \qquad (4) \\ &= b_{rr} + b_{ss} - 2b_{rs} \qquad \text{from (2)} \end{aligned}$$

Thus from X we can calculate B and hence D. We shall now show that, given D, we can calculate B, which is therefore independent of the coordinate representation, X, of D. We assume only that the origin is at the centroid so that the elements in any row or column of B must sum to zero. (ie $(3')$ holds.)

Using (3), and summing (4) over r, over s, and then both r and s we obtain

$$\begin{aligned} \sum_{r=1}^{n} d_{rs}^2 &= T + nb_{ss} \quad \text{for all } s \quad \text{(column sums)} \qquad (5) \\ \sum_{s=1}^{n} d_{rs}^2 &= nb_{rr} + T \quad \text{for all } r \quad \text{(row sums)} \qquad (6) \\ \sum_{r=1}^{n}\sum_{s=1}^{n} d_{rs}^2 &= 2nT \quad \text{(total sum)} \qquad (7) \\ \text{where} \quad T &= \sum_{r=1}^{n} b_{rr} \quad \text{(trace of } B\text{)} \qquad (8) \end{aligned}$$

Note that, from (7) and (8)

$$\text{trace}(B) \propto \sum_{r=1}^{n}\sum_{s=1}^{n} d_{rs}^2 \quad \text{the sum of all squared inter-point distances}$$

Solving $(5)-(7)$ we obtain b_{ss} and b_{rr} in terms of the d_{rs}^2. Substituting into (4) we then get:

$$\begin{aligned} b_{rs} &= -\frac{1}{2}\left[d_{rs}^2 - \frac{1}{n}\sum_{s=1}^{n} d_{rs}^2 - \frac{1}{n}\sum_{r=1}^{n} d_{rs}^2 + \frac{1}{n^2}\sum_{r=1}^{n}\sum_{s=1}^{n} d_{rs}^2\right] \\ &= -\tfrac{1}{2}\left[d_{rs}^2 - \overline{d}_{r.}^2 - \overline{d}_{.s}^2 + \overline{d}_{..}^2\right] \end{aligned} \qquad (9)$$

Thus we can obtain the matrix $B(= XX^T)$ *directly* from the matrix of Euclidean distances $D = (D_{rs})$ in the following way:-

Set $A = (-\frac{1}{2}d_{rs}^2)$

Then B may be obtained by double-centreing A - subtracting the row average from each element and then subtracting the *new* column average from each element. (Or vice versa).

$$B = HAH \quad \text{where } H = I - n^{-1}\mathbf{1}\mathbf{1}^T \qquad (10)$$

giving $B = (b_{rs})$ as in (9). (Check.)

Clearly B, constructed in this way, has zero row and column sums since $H\mathbf{1} = \mathbf{0}$ and $\mathbf{1}^T H = \mathbf{0}^T$. (We also know this must be true from the definition of B as XX^T, with $X^T\mathbf{1} = \mathbf{0}$, which we used to obtain (9).)

<u>Summarizing</u> If $D = (d_{rs})$ is a given matrix of Euclidean distances, then, by definition, there exists a corresponding configuration X, in some finite dimensional space, say $\mathbb{R}^p$, with the origin at the centroid. Defining $B = XX^T$, so that B is psd and has row and column sums equal to zero, we find that B may be obtained directly from D by setting

$$\left.\begin{aligned} A &= (-\tfrac{1}{2}d_{rs}^2) \\ \text{and} \quad B &= HAH \end{aligned}\right\} \tag{11}$$

so that $b_{rs} = a_{rs} - \overline{a}_{r.} - \overline{a}_{.s} + \overline{a}_{..}$.

Thus, if D is a matrix of Euclidean distances, B, defined by (11), is psd. This is a *necessary* condition for D to be Euclidean. We shall show, by construction, that this is also a *sufficient* condition.

Suppose that B, calculated from D via (11), is psd. Then, from the spectral decomposition theorem part(ii) (Theorem 1 in Section B) we can factorize B into $B = XX^T$, where X is an $n \times k$ matrix with $k = \text{rank}(B)$.

That is, if $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_k > 0$ are the nonzero eigenvalues of B and $\mathbf{e}_i$, $(i = 1, \ldots, k)$ are the corresponding mutually orthogonal eigenvectors, scaled so that $\mathbf{e}_i^T\mathbf{e}_i = 1$, then Theorem 1 part(ii) tells us that

$$B = \sum_{i=1}^{k} \lambda_i \mathbf{e}_i \mathbf{e}_i^T = E\Lambda E^T$$

where $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_k)$ and $E = (\mathbf{e}_1, \ldots, \mathbf{e}_k)$.

[Note that $E^T E = I$, since the $\mathbf{e}_i$ are mutually orthogonal.]
Let

$$X = E\Lambda^{\frac{1}{2}} = (\sqrt{\lambda_1}\mathbf{e}_1, \ldots, \sqrt{\lambda_k}\mathbf{e}_k) \tag{12}$$

Then $B = XX^T$, and we have factorized B. The columns of X are the mutually orthogonal eigenvectors of B, scaled so that the sum of squared elements in each eigenvector is equal to the corresponding eigenvalue.

It remains to show that the configuration of n points in k-dimensional space, represented by the $n \times k$ matrix X, does indeed reproduce the inter-point distances given in $D = (d_{rs})$, and has its origin at the centroid.

We have:

$$X = (\sqrt{\lambda_1}\mathbf{e}_1, \ldots, \sqrt{\lambda_k}\mathbf{e}_k) = \begin{bmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_n^T \end{bmatrix},$$

Thus the distance squared between points (rows) r and s in the *constructed* space is given by $(\mathbf{x}_r - \mathbf{x}_s)^T(\mathbf{x}_r - \mathbf{x}_s)$, and

$$\begin{aligned}(\mathbf{x}_r - \mathbf{x}_s)^T(\mathbf{x}_r - \mathbf{x}_s) &= \mathbf{x}_r^T\mathbf{x}_r + \mathbf{x}_s^T\mathbf{x}_s - 2\mathbf{x}_r^T\mathbf{x}_s \\ &= b_{rr} + b_{ss} - 2b_{rs},\end{aligned} \tag{13}$$

since $B = XX^T$.

However B was obtained from D by setting

$$\left.\begin{aligned} A &= (-\tfrac{1}{2}d_{rs}^2) \\ \text{and} \quad B &= HAH \end{aligned}\right\}$$

so that

$$b_{rs} = a_{rs} - \overline{a}_{r.} - \overline{a}_{.s} + \overline{a}_{..} \tag{14}$$

Substituting (14) into (13), and recalling that A is symmetric, it is easy to show that

$$b_{rr} + b_{ss} - 2b_{rs} = a_{rr} + a_{ss} - 2a_{rs}$$

and hence, from (14)

$$\begin{aligned}(\mathbf{x}_r - \mathbf{x}_s)^T(\mathbf{x}_r - \mathbf{x}_s) &= a_{rr} + a_{ss} - 2a_{rs} \\ &= -2a_{rs} \\ &= d_{rs}^2\end{aligned}$$

as required, since $a_{rr} = -\frac{1}{2}d_{rr}^2 = 0$ for all r, and $-2a_{rs} = d_{rs}^2$.

<u>Finally</u> - to check that the origin is at the centroid we need to show that

$$\mathbf{1}^T X = \mathbf{0}^T,$$

ie that the column sums of X are zero. Now

$$B\mathbf{1} = HAH\mathbf{1} = \mathbf{0}$$

since $H\mathbf{1} = \mathbf{0}$. (Check this.)

It follows that the vector $\mathbf{1}$ is an eigenvector of B corresponding to the eigenvalue 0. (Note that this implies $\text{rank}(B) \leq n - 1$ since at least one eigenvalue of B is equal to zero.)

Thus $\mathbf{1}$ is orthogonal to all the eigenvectors corresponding to the non-zero eigenvalues $\lambda_1, \ldots, \lambda_k$, and is therefore orthogonal to all the columns of $X = (\sqrt{\lambda_1}\mathbf{e}_1, \ldots, \sqrt{\lambda_k}\mathbf{e}_k)$. Thus

$$\mathbf{1}^T X = \mathbf{0}^T,$$

as required, and the centre of gravity (or centroid) of configuration X lies at the origin.

This proves the sufficiency.

<u>Thus</u> Given a distance matrix D, the distances are Euclidean $\Longleftrightarrow$ B, as calculated in (11), is positive semi-definite.

The sufficiency proof shows how to calculate a corresponding configuation in k-dimensional space, where $k = \text{rank}(B)$, so that the distances are reproduced exactly and the origin is at the centroid. There are clearly infinitely many possible factorizations $B = XX^T$ corresponding to infinitely many possible axis orientations, but the method of construction given in the sufficiency proof refers the configuration to principal axes and has certain optimal properties.

Recall, from (7) and (8), that

$$\text{trace}(B) \;=\; \lambda_1 + \ldots + \lambda_k \;=\; \frac{1}{2n}\sum_{r=1}^{n}\sum_{s=1}^{n} d_{rs}^2$$

Now take the first ℓ coordinates $(\sqrt{\lambda_1}\mathbf{e}_1, \ldots, \sqrt{\lambda_\ell}\mathbf{e}_\ell)$. It can be shown (see Mardia, Kent and Bibby, Section 14.4) that

$$\lambda_1 + \ldots + \lambda_\ell \;=\; \frac{1}{2n}\sum_{r=1}^{n}\sum_{s=1}^{n} d_{rs}^{(\ell)2}$$

where $d_{rs}^{(\ell)2}$ is the squared Euclidean distance between points r and s based on the first ℓ coordinates. This implies that

$$\lambda_{\ell+1} + \ldots + \lambda_k \;=\; \frac{1}{2n}\sum_{r=1}^{n}\sum_{s=1}^{n} (d_{rs}^2 - d_{rs}^{(\ell)2}).$$

Thus the percentage of the total sum of squared distances accounted for in the first ℓ dimensions is

$$\frac{\lambda_1 + \ldots + \lambda_\ell}{\lambda_1 + \ldots + \lambda_k} \times 100,$$

and this is the <u>maximum possible sum of the squared distances that can be accounted for in ℓ dimensions</u>.

Thus for any ℓ, $1 \le \ell \le k$, we have the best possible ℓ-dimensional representation.

Now suppose that the distances are <u>not</u> Euclidean. Then B is not psd. However if we choose, say, the first ℓ eigenvalues, where these are 'large positive', and set

$$X = (\sqrt{\lambda_1}\mathbf{e}_1, \ldots, \sqrt{\lambda_k}\mathbf{e}_\ell)$$

as before, then it can be shown (again see Mardia, Kent and Bibby) that this is the optimal ℓ-dimensional Euclidean representation of the given distances.

Similarity Matrices

Suppose that what we are given is a matrix of *similarities* between n objects. (We will discuss later some of the commonly used similarity measures.)

Definition: An $n \times n$ matrix $C = (c_{rs})$ is called a *similarity matrix* if it is symmetric and if

$$c_{rs} \leq c_{rr} \quad \text{for all } r,\, s \tag{15}$$

To use the classical scaling (CS) technique just described we need to transform the similarities to distances. The usual transformation is given by:

$$d_{rs}^2 \;=\; c_{rr} + c_{ss} - 2c_{rs} \tag{16}$$

so that the resulting matrix D is symmetric, with $d_{rr} = 0$ for all r and $d_{rs}^2 \geq 0$ for all r and s. (Compare with equation (4)).

It is easily shown (see Mardia, Kent and Bibby, Section 14.2.3) that the B matrix calculated from the above distance matrix is just HCH, the double-centred version of the similarity matrix C. Thus the distances given by (16) are exactly Euclidean $\Longleftrightarrow$ the similarity matrix C is psd. Fortunately many of the standard similarity measures are psd.

We now look at some of the distance (dissimilarity) and similarity measures used in practice.

1.2 Distance/Dissimilarity Matrices

Assume first that we do not observe the distance matrix *directly*, but that we observe an $n \times p$ data matrix X in which the p variables are quantitative and continuous. Assume further that X has been means corrected (ie centred) so that the column sums are all zero.

A. Pythagorean (Euclidean) Distance and the Duality between Classical Scaling and Principal Components Analysis

Suppose that D is the $n \times n$ matrix of Pythagorean distances between the n points in p dimensions calculated from the $n \times p$ data matrix X. Let S be the maximum likelihood estimate of the population covariance matrix, so that $nS = X^T X$

CS approach

To carry out a classical scaling analysis on these data we would calculate D from X, form the matrix $A = (-\frac{1}{2}d_{rs}^2)$ and double-centre it to obtain

$$B = HAH$$

From the necessity proof earlier, we know in fact that

$$B = XX^T, \tag{17}$$

where X is the original, means corrected $n \times p$ data matrix. Clearly B is psd. In a Classical Scaling analysis we would then proceed to factorize B by finding its nonzero (> 0) eigenvalues $\lambda_1, \ldots, \lambda_k$ and the associated eigenvectors $\mathbf{e}_1, \ldots, \mathbf{e}_k$, and forming

$$X^* = (\sqrt{\lambda_1}\mathbf{e}_1, \ldots, \sqrt{\lambda_k}\mathbf{e}_k)$$

so that

$$B = X^* X^{*T}. \tag{18}$$

Thus the CS procedure just factorizes $B = XX^T$ differently (and is actually equivalent to an orthogonal rotation of the original p axes). The CS factorization has the optimal properties already noted.

[Note that

$$k = \text{rank}(B) \leq \min\{n-1, p\} \quad (\text{assuming } n > p)$$

since we know that B has at least one zero eigenvalue.]

<u>Principal Components (PCA) approach</u>

To carry out a Principal Components Analysis we form the $p \times p$ sums-of-squares-and-products (SSP) matrix

$$nS = X^T X.$$

This is the outer product matrix.

In a Principal Components Analysis we would then proceed to find the eigenvalues and corresponding eigenvectors of $X^T X = nS$ to obtain principal component sample variances, coefficients and hence scores. [Recall that the eigenvalues of S are the principal component sample variances and so the eigenvalues of $X^T X$ are just n times the component variances. The eigenvectors of S and $X^T X$ are identical.]

Thus CS carries out an eigenanalysis on the $n \times n$ matrix $B = XX^T$ and PCA carries out an eigenanalysis on the $p \times p$ matrix $nS = X^T X$. The duality between CS and PCA is a consequence of the relationship between the eigenstructure of these two matrices.

More precisely: From Theorem 3 in Section B of the Notes (see also Mardia, Kent and Bibby, Theorem A.6.2) it follows that the non-zero eigenvalues of XX^T and $X^T X$ are identical and the corresponding eigenvectors are related by a simple linear transformation.

To see this, let λ_i, $\mathbf{f}_i$, $i = 1, \ldots k$ be the non-zero eigenvalues and corresponding standardized eigenvectors of $X^T X = nS$, so that $\mathbf{f}_i^T \mathbf{f}_i = 1$ for all i. Then, for each i, $\lambda_i = n \times \text{Variance(component } i)$, $X\mathbf{f}_i = \mathbf{c}_i$, the vector of scores on component i, and of course

$$X^T X \mathbf{f}_i = \lambda_i \, \mathbf{f}_i. \tag{19}$$

Premultiplying equation (19) by X gives

$$(XX^T)\, X\mathbf{f}_i = \lambda_i \, X\mathbf{f}_i$$

so that the λ_i are the non-zero eigenvalues of XX^T as well as X^TX (hence the notation), and the $X\mathbf{f}_i$ are eigenvectors of XX^T corresponding to the λ_i. Thus for each i,

$$X\mathbf{f}_i \propto \mathbf{e}_i,$$

where $\mathbf{e}_i$ is the standardized eigenvector of XX^T corresponding to λ_i. We need to identify the scaling of $X\mathbf{f}_i$.

Since X is means corrected, $X\mathbf{f}_i = \mathbf{c}_i$, the vector of component scores, has a sample mean of zero, and so the sum of the squared elements of $X\mathbf{f}_i$ is just n times the sample variance of component i, which is equal to λ_i. Thus

$$X\mathbf{f}_i \;=\; \sqrt{\lambda_i}\,\mathbf{e}_i,$$

and the $n \times 1$ vector of scores on Principal Component i is just the $n \times 1$ vector of coodinates in the i'th dimension of the Classical Scaling solution. The results of PCA are exactly equivalent to the results of CS if we carry out the latter procedure on the matrix of Pythagorean distances calculated from X. This also confirms (from our knowledge of PCA) that the CS axes account successively for the largest amount of the total variance, which is proportional to the total sum of squared distances.

Analysing the $p \times p$ matrix X^TX using PCA is an example of an 'R-mode' analysis

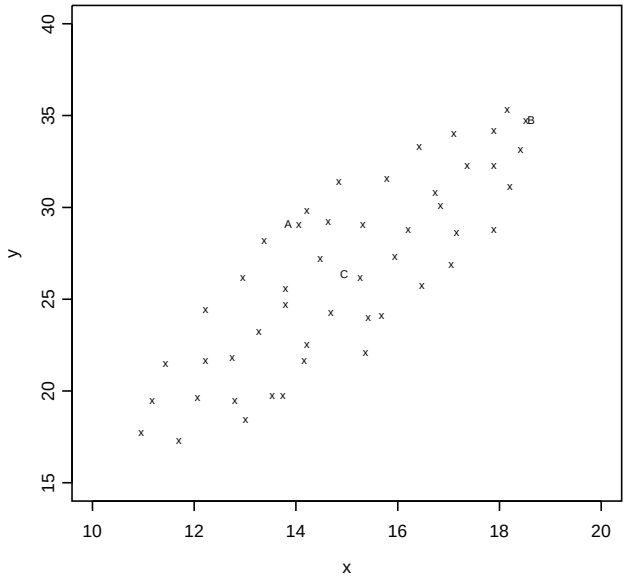
Analysing the $n \times n$ matrix XX^T using CS is an example of a 'Q-mode' analysis

In this case we should use PCA rather than CS since the $p \times p$ matrix X^TX is likely to be smaller than the $n \times n$ matrix XX^T.

B. <u>Mahalanobis Distance</u>

A problem with Principal Component Analysis (PCA) is that the components are not invariant under changes in the units of measurement. One method of overcoming this is to use the correlation matrix rather than the covariance matrix, that is to re-scale each variable so that it has unit sample variance. Then Classical Scaling (CS), using the Euclidean distances based on the re-scaled variables, is equivalent to PCA on the correlation matrix. The effect is to equalize the importance of each variable in the sample. However no account is taken of the correlation between the variables. A measure that takes account of both differing variances and the correlation between variables is the Mahalanobis distance. Mahalanobis distance is scale invariant.

Consider a sample of n units measured on 2 variables, x_1 and x_2, and consider a plot of x_2 versus x_1 that looks like this:



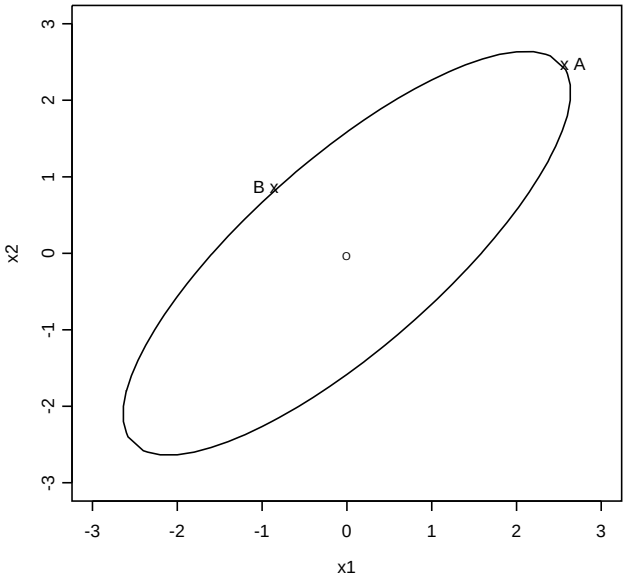
where C indicates the centroid and A and B represent two particular data points. Clearly x_1 and x_2 are highly correlated. Looking at the plot we 'see' that the distance BC is greater that the distance AC. However we may conjecture that, in probabilistic terms, BC and AC are comparable. Suppose that the observations $\mathbf{x} = (x_1, x_2)^T$ follow a bivariate normal distribution $\mathbf{x} \sim \text{BVN}(\boldsymbol{\mu}, \Sigma)$, with probability density function:

$$f(\mathbf{x}) = \frac{1}{2\pi|\Sigma|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

Contours of equal probability density are given by:

$$\{\mathbf{x} : \ (\mathbf{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}) = \text{constant}\},$$

as in the plot overleaf.



Points on the same elliptical contour are equally likely and should therefore appear equally distant from $\boldsymbol{\mu}$, denoted by O. We need to transform this elliptical contour into a circular contour centred on O $= \boldsymbol{\mu}$. The Mahalanobis distance achieves this.

Sample Mahalanobis distance $\mathrm{D}(\mathbf{x}_r, \mathbf{x}_s)$ between data units r and s is given by:

$$D^2(\mathbf{x}_r, \mathbf{x}_s) \;=\; (\mathbf{x}_r - \mathbf{x}_s)^T\, S^{-1}\, (\mathbf{x}_r - \mathbf{x}_s),$$

where S is the $p \times p$ sample covariance matrix (with divisor n). If S is of full rank then, from Theorem 2 in Section B, we can find the symmetric square root $S^{-\frac{1}{2}}$ of S^{-1}, so that

$$S^{-\frac{1}{2}}\, S^{-\frac{1}{2}} \;=\; S^{-1}$$

This can be used to find coordinates of n points in p-dimensional space whose Pythagorean inter-point distances reproduce the original Mahalanobis distances exactly and where elliptical contours of equal probability density are transformed to circular contours in the case of multivariate normality. Set

$$\mathbf{z} \;=\; S^{-\frac{1}{2}}\, \mathbf{x} \quad r = 1, \ldots, n$$

so that we have a new $n \times p$ matrix Z where

$$Z \;=\; \begin{bmatrix} \mathbf{z}_1^T \\ \mathbf{z}_2^T \\ \vdots \\ \mathbf{z}_n^T \end{bmatrix} \;=\; X\, S^{-\frac{1}{2}}$$

Now

$$S_{\mathbf{z}} \;=\; S^{-\frac{1}{2}}\, S\, S^{-\frac{1}{2}} \;=\; I \quad \text{(from C.8)}$$

so that the new variables, the elements of the general observation vector $\mathbf{z}$, have sample variances equal to one and sample correlations all equal to zero. [Since the data matrix X is means corrected, so is the new matrix Z. Z is the *standardized* data matrix] Clearly, in the case of multivariate normality, contours of equal probability density are given by:

$$\{\mathbf{z} :\; \mathbf{z}^T\, \mathbf{z} = \text{constant}\},$$

i.e they are circular.

We now have

$$\begin{aligned}(\mathbf{z}_r - \mathbf{z}_s)^T(\mathbf{z}_r - \mathbf{z}_s) &= (\mathbf{x}_r - \mathbf{x}_s)^T\, S^{-1}\, (\mathbf{x}_r - \mathbf{x}_s) \\ &= D^2(\mathbf{x}_r, \mathbf{x}_s)\end{aligned}$$

so that the Euclidean distances between the uncorrelated $\mathbf{z}$'s are exactly the Mahalanobis distances between the corresponding $\mathbf{x}$'s. If we calculate the matrix of Mahalanobis distances and carry out a CS analysis, then we obtain new coordinate axes that account successively for the maximum possible amount of squared Mahalanobis distances.

A more common use of Mahalanobis distance is to measure distances between *samples* that are assumed to come from populations with different mean vectors but with a common covariance matrix Σ.

Then if $\overline{\mathbf{x}}_r$ is the $p \times 1$ vector of sample means for the sample from population r, and $\overline{\mathbf{x}}_s$ is the $p \times 1$ vector of sample means for the sample from population s, the Mahalanobis distance between the two samples is defined as

$$D^2(\overline{\mathbf{x}}_r, \overline{\mathbf{x}}_s) \;=\; (\overline{\mathbf{x}}_r - \overline{\mathbf{x}}_s)^T\, S^{-1}\, (\overline{\mathbf{x}}_r - \overline{\mathbf{x}}_s),$$

where S is the pooled within-sample covariance matrix across all samples. (As used here, S is the maximum likelihood estimate of Σ.)

Note: For 2 samples (comparing 2 mean vectors), recall that the Hotelling T^2 statistic is proportional to $D^2(\overline{\mathbf{x}}_1, \overline{\mathbf{x}}_2)$ and is related to Fisher's linear discriminant function. Recall also that extending to $K > 2$ samples gives us 1-way MANOVA and $r \leq \min\{K-1, p\}$ corresponding canonical variates that are analogous to Fisher's discriminant function. We shall see that there is a duality between CS carried out on the matrix of between-sample Mahalanobis distances (Q-mode analysis) and Canonical Variate Analysis, an R-mode method based on between and within sample covariance matrices and closely related to 1-way MANOVA. (cf last year's notes.)

C. <u>The Minkowski metrics</u>

We assume still that we have a means-centred $n \times p$ data matrix X measuring n units on p quantitative, continuous variables. Then, for each $R = 1, 2, \ldots$ we can define a distance d_{rs} between units r and s as

$$d_{rs} = \left\{ \sum_{j=1}^{p} |x_{rj} - x_{sj}|^R \right\}^{\frac{1}{R}}$$

As indicated by the name, all these distances satisfy the metric or triangle inequality. However only $R = 2$ (Pythagorean distance) gives us Euclidean distances.

Setting $R = 1$ gives us the city block metric (or Manhattan city block metric)

$$d_{rs} = \sum_{j=1}^{p} |x_{rj} - x_{sj}|.$$

Clearly this is *not* Euclidean. In particular it is not invariant under an orthogonal rotation of axes.

$$R = \infty \implies d_{rs} = \max_j |x_{rj} - x_{sj}|$$

D. <u>The Wilkinson metric for $(0, 1)$ data</u>

This metric is defined for a particular type of binary data presented as an initial $n \times n$ data matrix X. It is best described using an example. The one used here is taken from Chatfield and Collins.

<u>Example</u> To assess relationships in a class of 24 primary school children, each child was asked to select 3 other children that they would like to sit next to. Thus the initial data consist of an 24×24 matrix of binary data, with both rows and columns indexed by the 24 children. It is not a similarity or distance matrix however. Each row r, for example, contains exactly three 1's in the columns that correspond to the children chosen by child r. All other elements in the row are coded 0. (See data handout.) We choose as the distance measure from child r to child s, the number of steps in a shortest route from child r to child s, via the choices of other children where necessary. Thus if child r chose child s initially, $d_{rs} = 1$. In this example $d_{1\,5} = d_{1\,6} = d_{1\,14} = 1$. However $d_{1\,7} = 2$, since we can go from child 1 to child 7, via child 5. If there is no route from child r to child s, then d_{rs} is set to $n+1 = 25$.

Note: This distance (or dissimilarity) matrix need not be symmetric. The shortest route from s to r is not necessarily the same length as the shortest route from r to s. However a version of the triangle inequality holds.

$$d_{rs} \leq d_{rt} + d_{ts} \quad \text{for all } r,\ s \text{ and } t,$$

where the order of the subscripts is important.

This measure is often used to investigate relationships among individuals. Eg dominance relationships in groups of animals.

It was also used by D. G. Kendall to measure distances between French Départements, administrative areas of France. Where 2 Départements have a common boundary, the corresponding data element is set to 1. Otherwise it is set to 0. The resulting distance matrix calculated using the Wilkinson metric is symmetric. Using CS it is possible to obtain a good reconstruction of the map of France.

Calculation of this metric is nontrivial. It can be done:

(i) by matrix methods, looking at successive powers of the data matrix X, X^2, X^3, ..., where $d_{rs} = k$ if the (r, s)'th element becomes nonzero for the first time in X^k;

(ii) by graph theoretic methods, where children are represented as nodes in a graph, the 1's in the data matrix indicated directed links of length 1, and the Wilkinson distances are the lengths of shortest routes from each node to every other node.

1.3 Similarity Measures for Presence/Absence Data

Our $n \times p$ data matrix X is now assumed to contain binary data indicating the presence (x_{ri}=1) or absence (x_{ri}=0) of each characteristic or feature i, $i = 1, \ldots, p$, in each individual r, $r = 1, \ldots, n$. For example, we may indicate the presence or absence of qualitative features on a fossil or skull, of types of artefact in a grave, of plant species in a given area etc. If we have p possible attributes, then for each pair of data units r, s we let

$$
\begin{aligned}
a &= \sum_{i=1}^{p} x_{ri}x_{si} &&= \text{No of } (1,1) \text{ occurrences} \\
b &= \sum_{i=1}^{p} (1 - x_{ri})x_{si} &&= \text{No of } (0,1) \text{ occurrences} \\
c &= \sum_{i=1}^{p} x_{ri}(1 - x_{si}) &&= \text{No of } (1,0) \text{ occurrences} \\
d &= \sum_{i=1}^{p} (1 - x_{ri})(1 - x_{si}) &&= \text{No of } (0,0) \text{ occurrences}
\end{aligned}
$$

Then $p = a + b + c + d$, the total number of characteristics considered. We consider 3 different similarity measures.

A. The Simple Matching Coefficient (Similarity matrix C)

Define the similarity c_{rs} between r and s by setting

$$c_{rs} = \frac{a+d}{p}.$$

This similarity measure has some unsatisfactory features however. Often the *presence* of the same attribute in 2 individuals says more about their likeness than the *absence* of the same attribute. Moreover d can be made as large as we choose by considering more and more attributes that are *not* present in either individual. And if d becomes large, then so does the similarity c_{rs}. For example. in Ecology, when comparing plant communities, the occurrence of the same species in two areas says more about the similarity of the areas than the absence of a particular species from both. There could be completely different species present in two areas, but an artificially high similarity because of the many species not present in *either* area.

B. The Ecological Coefficient (Similarity matrix C')

Set

$$c'_{rs} = \frac{a}{p},$$

so that c'_{rs} is the number of attributes present in both r and s as a proportion of *all* the attributes considered.

C. Jaccard's Coefficient (Similarity matrix C'')

Set

$$c''_{rs} = \frac{a}{p-d},$$

so that c''_{rs} is the number of attributes present in both r and s as a proportion of the number of attributes present in either r, s or both.

Note: Both C (simple matching coefficient) and C' (ecological coefficient) are positive semi-definite since:

$$\begin{aligned} C' &= \tfrac{1}{p} X X^T \\ \text{and} \quad C &= \tfrac{1}{p} \left[X X^T + (J-X)(J-X)^T \right] \end{aligned},$$

where J is the $n \times p$ matrix consisting entirely of 1's.

We saw earlier that the usual transformation from similarities to distances is

$$d_{rs}^2 = c_{rr} + c_{ss} - 2c_{rs},$$

so that C, when double-centred, is just the B matrix that would be obtained by setting $A = (-\frac{1}{2}d_{rs}^2)$ and double-centreing it. Thus if C is psd, then the distances defined above are Euclidean, and a Euclidean representation in a low-dimensional space is a reasonable goal. CS is carried out by double-centreing C to form B and then factorizing B via its eigen-decomposition to obtain $B = X^* X^{*T}$, where X^* is the required coordinate representation of the original similarity matrix.

Consider a well-known example of the analysis of presence/absence data. On an archaeological site n graves have been excavated and the presence or absence of p different varieties of pottery has been recorded. We can use either the simple matching coefficient or the ecological coefficient to calculate an $n \times n$ similarity matrix between the graves. It is of interest to determine the chronological ordering of the graves. This is an example of seriation, in which the objective is to pick out a one-dimensional structure or *ordering*. That is, we expect the data to be parameterized by a single axis or scale - in this case time. Other orderings might relate to attitude, performance, degree of satisfaction etc.

If we are looking to see whether the data are essentially one-dimensional, it is essential to represent (and plot) the data in at least 2 dimensions. This is because, although the data may really be one-dimensional, the points, as represented in a Euclidean space, can sometimes lie along a *curve* rather than A straight line. This is the famous 'horseshoe' effect, which occurs because it is often possible to measure accurately the distances between objects that are close together, but to be much less accurate in assessing distances between objects that are far apart. Thus large and moderate distances may appear to be of a similar magnitude. For example, using the grave data and the ecological coefficient c'_{rs}, we have $c'_{rs} = 0$ whenever the graves r and s have no artefacts in common. This implies that:

$$d_{rs}^2 = c'_{rr} + c'_{ss} = \frac{\text{No of artefact types in } r \text{ plus no in } s}{\text{Total No of artefact types}}$$

A plot in 2 dimensions might exhibit the 'horseshoe' effect, with the furthest points pulled together, but a clear ordering along the curve. A one-dimensional Euclidean representation would fail to detect the obvious one-dimensional ordering. (See Mardia, Kent and Bibby, p412, for an artificial example of this phenomenon).

Gower's Similarity Coefficient

Suppose now that we measure the n data units on a (possibly large) mixture of quantitative and qualitative variables, where some measurements may be missing.

In such cases the similarity coefficient devised by Gower can be used, and if there are no missing values then Gower's similarity coefficient is psd and the corresponding distances are Euclidean. Gower's similarity measure copes with a mix of qualitative and quantitative variables, for each of which a different method of calculating similarity may be appropriate. The method calculates an average of similarities, each based on one variable and each scaled to be between 0 and 1. See Gower (1971), and the handout, for details.

1.4 Ordinal Scaling (CS)

As in the case of Classical Scaling, we start with a given $n \times n$ matrix of dissimilarities between objects, which we shall now denote by (δ_{rs}). It is required to find a configuration of n points in, say, k dimensions, whose calculated Euclidean inter-point distances are in 'good agreement' with the given dissimilarities (δ_{rs}).

In Ordinal Scaling (OS), we attempt to match only the *rank order* of the given dissimilarities. This is reasonable, since it is often the case that the exact numerical values observed have little meaning. At the very least, they will be subject to measurement error. Moreover, it is often the case that with psychological or social science data, the dissimilarities may be directly observed as purely ordinal in character.

As with CS, there is no unique location, axis orientation or reflection. In addition, with OS there is no implied scale, since only the ordinal properties of the dissimilarities have meaning.

As before, we place the origin at the centroid of any Euclidean representation. The scale is fixed by requiring that the average squared distance from the origin is equal to one. That is, in the Euclidean representation,

$$\frac{1}{n}\sum_{i=1}^{n} d_{i0}^2 \;=\; 1.$$

In Ordinal Scaling, no orientation of the axes is set *a priori*, and we need to decide at the outset the dimension k of the required representation. To carry out Ordinal Scaling, there is no analytical procedure analogous to the eigen-decomposition of B in Classical Scaling. An iterative optimization procedure is required.

Given the dissimilarities (δ_{rs}), and the dimension k of the required Euclidean representation, we proceed as follows. Begin by ordering the δ_{rs} in ascending order

$$\delta_{r_1,s_1} \;\leq\; \delta_{r_2,s_2} \;\leq\; \ldots \;\leq\; \delta_{r_m,s_m},$$

where m is generally equal to $\frac{1}{2}n(n-1)$.

Suppose now that we have a proposed configuration X, of n points in k dimensions, with the origin at the centroid. Let the corresponding inter-point distances be given by (d_{rs}). We need some quantitative criterion to measure how well X represents the given dissimilarities. We can then try to optimize the criterion by making changes in X.

Consider first what would be regarded as a perfect match between the d_{rs} calculated from X and the given δ_{rs}.

1. The Monotone Requirement: We shall require that

$$\delta_{rs} < \delta_{tu} \Longrightarrow d_{rs} \leq d_{tu}$$

This is referred to as the primary monotone requirement. [The secondary monotone requirement requires that $d_{rs} < d_{tu}$ whenever $\delta_{rs} < \delta_{tu}$, and is generally regarded as too restrictive.]

2. Treatment of Ties: What if $\delta_{rs} = \delta_{tu}$? We could require that $d_{rs} = d_{tu}$. This is the secondary approach ties and is also too restrictive in general. We shall take the primary approach and assume no restrictions on the way ties in the δ_{rs} are broken. [ie if $\delta_{rs} = \delta_{tu}$ then we allow either $d_{rs} \leq d_{tu}$ or $d_{rs} > d_{tu}$.]

Any numbers d^*_{rs} satisfying just the primary monotone requirement

$$\delta_{rs} < \delta_{tu} \Longrightarrow d^*_{rs} \leq d^*_{tu} \tag{20}$$

are said to be *monotonically related* to the δ_{rs} and we shall regard this as sufficient for the purpose of reproducing the rank ordering of the δ_{rs}.

If we now have a set of distances d_{rs} calculated from a proposed configuration X, we need a way of measuring how far the d_{rs} are from reproducing the rank ordering of the δ_{rs}; ie how far are they from satisfying (20). Define the (squared) *stress* of X by:

$$S^2(X) = \min \sum_{r<s} (d^*_{rs} - d_{rs})^2 \Big/ \sum_{r<s} d^2_{rs}, \tag{21}$$

where the minimum is taken over all sets of numbers d^*_{rs} that are monotonically related to the δ_{rs}. This is a Least Squares criterion. The d^*_{rs} that minimize (21) represent the *least squares monotone regression* of d_{rs} on δ_{rs}, and $S^2(X)$, given by (21), represents the extent to which the rank order of the d_{rs} calculated from X differs from the rank order of the δ_{rs}. The denominator in (21) standardizes the stress and makes it invariant under any uniform expansion or contraction of the configuration X with its origin at the centroid. It is also invariant under translation and orthogonal rotation of the axes.

There are efficient routines for carrying out LS monotone regression. Essentially it is required to identify the optimal partition of the ordered δ_{rs} into consecutive blocks for which the corresponding d_{rs} are averaged to give a common fitted d^* for the block. The fitted d^*'s are then a strictly increasing function of the block number. (See Kruskal (1965) in Psychometrika 19, pp 115-129.) The algorithm is incorporated into many statistical packages.

Thus, given the matrix of dissimilarities (δ_{rs}), it is required to find a configuration X of n points in k dimensions such that $S(X)$, calculated from a LS monotone regression of the corresponding inter-point distances d_{rs} on the δ_{rs} (as in (21)), is minimized.

Outline of the Algorithm

1. Choose an initial configuration X^1 and set n=1.

2. Calculate $S(X^n)$, which is a function of the elements of X^n.

3. Move a certain distance in the steepest descent direction by making changes in X^n so as to decrease $S(X^n)$. (To carry out this step it may be noted that $S(X^n)$ is explicitly differentiable with respect to the nk elements of X^n, and thus the direction of steepest descent can be identified.) Set the new configuration equal to X^{n+1}, set n to $n+1$ and return to Step 2 - unless a convergence criterion is satisfied.

Notes:
1. The steepest descent algorithm is not normally a very efficient optimization algorithm and in general many better algorithms exist. However the objective function $S(X)$ is not 'nicely shaped' - in particular it is not locally quadratic and more sophisticated optimization methods do not seem to work well.

2. The initial configuration could be the classical scaling solution, or could be selected randomly. However $S(X)$ usually has many local minima and in practice it is not possible to distinguish whether the minimum found is local or global. Moreover, with fewer than around 30 points, it is possible for the points in a configuration X to move position by quite an amount without changing $S(X)$. For these reasons it is often recommended that a number of different starting configurations are used.

3. It is not possible to get the best k-dimensional solution directly from the best $(k+1)$-dimensional solution as in classical scaling. That is to say, the best k-dimensional solution is *not* a k-dimensional subspace of the best $(k+1)$-dimensional solution. For this reason, although the origin and scale are standardized, fixing a particular orientation of the axes is not really an issue. However we can always refer a configuration to principal axes if required.

4. For each proposed dimensionality k, the configuration that has the smallest stress is called the *best fitting configuration in k dimensions.* Set

$$S_k = \min_{X(n\times k)} S(X)$$

To choose the correct dimension, start with a dimension j that is higher than expected (say dimension 5 - or dimension $k+3$, where k is the expected dimensionality) and find S_j, S_{j-1}, ..., S_1. Going from any dimension $k+1$ to k, we can project the best $k+1$ dimensional configuration into k dimensions to provide a starting configuration for the next stage, and this turns out to be a good way of homing in on the global minimum in the

lower dimensions, say in 2 and 3 dimensions. Recall also that the 'horseshoe' effect suggests starting in a higher-dimensional space than the dimensionality expected. Plotting S_k versus k can suggest the correct dimension to use, perhaps where an 'elbow' is observable and the stress is 'low'.

5. No significance tests are possible using the stress criterion $S(X)$. However researchers have sometimes developed rules of thumb to assess the values of stress obtained. For example, one rule suggests: $S_k \geq 0.2$, poor; $0.10 \leq S_k \leq 0.15$, acceptable; $0.05 \leq S_k < 0.10$, satisfactory; $S_k < 0.05$, excellent. However it is now known that these criteria are much too simplistic. Stress will inevitably increase with increasing n and decrease with increasing k, and judging the quality of the representation is inevitably subjective to a large extent.

Some Advantages of OS over CS

1. Ordinal Scaling can easily cope with missing values in the dissimilarity ,matrix (δ_{rs}). They are just omitted from the monotone regression and hence the stress calculation. Note that, as the example with road distances shows, reproducing the rank order of the δ_{rs} generally imposes very tight constraints on the positions of the data points, and having some missing values may affect the optimal configuration very little.

2. Ordinal Scaling can be adjusted to cope with asymmetric matrices. (Cf the distance/dissimilarity matrix calculated from children's preference data.) All dissimilarities can be included in the LS monotone regression and in the value of stress $S(X)$, although the value of stress will obviously be increased since different dissimilarities correspond to the same distance.
(See the accompanying handout for the OS results on the children data.)

3. There is a better chance of representing the dissimilarities in a low number of dimensions using Ordinal Scaling since the OS criterion is less restrictive.

4. Ordinal Scaling is a more realistic procedure if the 'distances' are not believable as precise measurements.