NAG Library Chapter Introduction

g03 – Multivariate Methods

Contents

1 Scope of the Chapter

This chapter is concerned with methods for studying multivariate data. A multivariate dataset consists of several variables recorded on a number of objects or individuals. Multivariate methods can be classified as those that seek to examine the relationships between the variables (e.g., principal components), known as variable-directed methods, and those that seek to examine the relationships between the objects (e.g., cluster analysis), known as individual-directed methods.

Multiple regression is not included in this chapter as it involves the relationship of a single variable, known as the response variable, to the other variables in the dataset, the explanatory variables. Routines for multiple regression are provided in Chapter g02.

2 Background to the Problems

2.1 Variable-directed Methods

Let the n by p data matrix consist of p variables, $x_1, x_2, \ldots, x_p$, observed on n objects or individuals. Variable-directed methods seek to examine the linear relationships between the p variables with the aim of reducing the dimensionality of the problem. There are different methods depending on the structure of the problem. **Principal component analysis** and **factor analysis** examine the relationships between all the variables. If the individuals are classified into groups, then **canonical variate analysis** examines the between group structure. If the variables can be considered as coming from two sets, then **canonical correlation analysis** examines the relationships between the two sets of variables. All four methods are based on an eigenvalue decomposition or a singular value decomposition (SVD) of an appropriate matrix.

The above methods may reduce the dimensionality of the data from the original p variables to a smaller number, k, of derived variables that adequately represent the data. In general, these k derived variables will be unique only up to an **orthogonal rotation**. Therefore, it may be useful to see if there exists suitable rotations of these variables that lead to a simple interpretation of the new variables in terms of the original variables.

2.1.1 Principal component analysis

Principal component analysis finds new variables which are linear combinations of the p observed variables so that they have maximum variation and are orthogonal (uncorrelated).

Let S be the p by p variance-covariance matrix of the n by p data matrix. A vector a_1 of length p is found such that

$$a_1^{\mathrm{T}} S a_1 \text{ is maximized subject to } a_1^{\mathrm{T}} a_1 = 1.$$

The variable $z_1 = \sum_{i=1}^{p} a_{1i} x_i$ is known as the first principal component and gives the linear combination of the variables that gives the maximum variation. A second principal component, $z_2 = \sum_{i=1}^{p} a_{2i} x_i$, is found such that

$$a_2^{\mathrm{T}} S a_2 \text{ is maximized subject to } a_2^{\mathrm{T}} a_2 = 1 \text{ and } a_2^{\mathrm{T}} a_1 = 0.$$

This gives the linear combination of variables, orthogonal to the first principal component, that gives the maximum variation. Further principal components are derived in a similar way.

The vectors a_i, for $i = 1, 2, \ldots, p$, are the eigenvectors of the matrix S and associated with each eigenvector is the eigenvalue, γ_i^2. The value of $\gamma_i^2 / \sum \gamma_i^2$ gives the proportion of variation explained by the ith principal component. Alternatively, the a_i can be considered as the right singular vectors in a SVD of a scaled mean-centred data matrix. The singular values of the SVD are the γ_i-values.

Often fewer than p dimensions (principal components) are needed to represent most of the variation in the data. A test on the smaller eigenvalues can be used to investigate the number of dimensions needed.

The values of the principal component variables for the individuals are known as the principal component scores. These can be standardized so that the variance of these scores for each principal component is 1.0 or equal to the corresponding eigenvalue. The principal component scores correspond to the left-hand singular vectors in the SVD.

2.1.2 Factor analysis

Let the p variables have variance-covariance matrix Σ. The aim of factor analysis is to account for the covariances in these p variables in terms of a smaller number, k, of hypothetical variables or factors, $f_1, f_2, \ldots, f_k$. These are assumed to be independent and to have unit variance. The relationship between the observed variables and the factors is given by the model

$$x_i = \sum_{j=1}^{k} \lambda_{ij} f_j + e_i, \quad i = 1, 2, \ldots, p$$

where λ_{ij}, for $i = 1, 2, \ldots, p$ and $j = 1, 2, \ldots, k$, are the factor loadings and e_i, for $i = 1, 2, \ldots, p$, are independent random variables with variances ψ_i. These represent the unique component of the variation of each observed variable. The proportion of variation for each variable accounted for by the factors is known as the communality.

The model for the variance-covariance matrix, Σ, can then be written as

$$\Sigma = \Lambda\Lambda^{\mathrm{T}} + \Psi,$$

where Λ is the matrix of the factor loadings, λ_{ij}, and Ψ is a diagonal matrix of the unique variances ψ_i.

If it is assumed that both the k factors and the e_i follow independent Normal distributions then the parameters of the model, Λ and Ψ, can be estimated by maximum likelihood, as described by Lawley and Maxwell (1971). The computation of the maximum likelihood estimates is an iterative procedure which involves computing the eigenvalues and eigenvectors of the matrix

$$S^* = \Psi^{-1/2} S \Psi^{-1/2},$$

where S is the sample variance-covariance matrix. Alternatively, the SVD of the matrix $R\Psi^{-1/2}$ can be used, where $R^{\mathrm{T}}R = S$. When convergence has been achieved, the estimates $\hat{\Lambda}$, of Λ, are obtained by scaling the eigenvectors of S^*. The use of maximum likelihood estimation means that likelihood ratio tests can be constructed to test for the number of factors required.

Having found the estimates of the parameters of the model, the estimates of the values of the factors for the individuals, the **factor scores**, can be computed. These involve the calculation of the **factor score coefficients**. Two common methods of computing factor score coefficients are the regression method and Bartlett's method. Bartlett's method gives unbiased estimates of the factor scores while the estimates from the regression method are biased but have smaller variance; see Lawley and Maxwell (1971).

2.1.3 Canonical variate analysis

If the individuals can be classified into one of g groups, then canonical variate analysis finds the linear combinations of the p variables that maximize the ratio of the between-group variation to the within-group variation. These variables are known as canonical variates. As the canonical variates provide discrimination between the groups, the method is also known as **canonical discrimination**.

The canonical variates can be calculated from the eigenvectors of the within-group sums of squares and cross-products matrix or from the SVD of the matrix

$$V = Q_x^{\mathrm{T}} Q_g,$$

where Q_g is an orthogonal matrix that defines the groups and Q_x is the first p columns of the orthogonal matrix Q from the QR decomposition of the data matrix with the variable means subtracted. If the data matrix is not of full rank, the Q_x matrix can be obtained from a SVD. If the SVD of V is

$$V = U_x \Delta U_g^{\mathrm{T}},$$

then the nonzero elements ($\delta_i > 0$) of the diagonal matrix Δ are the canonical correlations. The largest δ_i is called the **first canonical correlation** and associated with it is the first canonical variate.

The eigenvalues, γ_i^2, of the within-group sums of squares matrix are given by

$$\gamma_i^2 = \frac{\delta_i^2}{1-\delta_i^2}.$$

The value of $\pi_i = \gamma_i^2 / \sum \gamma_i^2$ gives the proportion of variation explained by the ith canonical variate. The values of the π_i give an indication as to how many canonical variates are needed to adequately describe the data, i.e., the dimensionality of the problem. The number of dimensions can be investigated by means of a test on the smaller canonical correlations.

The canonical variate loadings and the relationship between the original variables and the canonical variates are calculated from the matrix U_x. This matrix is scaled so that the canonical variates have unit variance.

2.1.4 Canonical correlation analysis

If the p variables can be considered as coming from two sets then canonical correlation analysis finds linear combinations of the variables in each set, known as canonical variates, such that the correlations between corresponding canonical variates for the two sets are maximized. Let the two sets of variables be denoted by x and y, with p_x and p_y variables in each set respectively. Let the variance-covariance of the dataset be

$$S = \begin{bmatrix} S_{xx} & S_{xy} \\ S_{yx} & S_{yy} \end{bmatrix}$$

and let

$$\Sigma = S_{yy}^{-1} S_{yx} S_{xx}^{-1} S_{xy},$$

then the canonical correlations can be calculated from the eigenvalues of the matrix Σ. Alternatively, the canonical correlations can be calculated by means of a SVD of the matrix

$$V = Q_x^{\mathrm{T}} Q_y,$$

where Q_x is the first p_x columns of the orthogonal matrix Q from the QR decomposition of the x-variables in the data matrix, and Q_y is the first p_y columns of the Q matrix of the QR decomposition of the y-variables in the data matrix. In both cases, the variable means are subtracted before the QR decomposition is computed. If either set of variables is not of full rank, an SVD can be used instead of the QR decomposition. If the SVD of V is

$$V = U_x \Delta U_y^{\mathrm{T}},$$

then the nonzero elements ($\delta_i > 0$) of the diagonal matrix Δ are the canonical correlations. The largest δ_i is called the **first canonical correlation** and associated with it is the first canonical variate. The eigenvalues, γ_i^2, of the matrix Σ are given by

$$\gamma_i^2 = \frac{\delta_i^2}{1+\delta_i^2}.$$

The value of $\pi_i = \gamma_i^2 / \sum \gamma_i^2$ gives the proportion of variation explained by the ith canonical variate. The values of the π_i give an indication as to how many canonical variates are needed to adequately describe the data, i.e., the dimensionality of the problem; this can also be investigated by means of a test on the smaller values of the γ_i^2.

The relationship between the canonical variables and the original variables, the canonical variate loadings, can be computed from the U_x and U_y matrices.

2.1.5 Rotations

There are two principal reasons for using rotations: either

(a) simplifying the structure to aid interpretation of derived variables, or

(b) comparing two or more datasets or sets of derived variables.

The most common type of rotations used for (a) are **orthogonal rotations**. If Λ is the p by k loading matrix from a variable-directed multivariate method, then the rotations are selected such that the elements, λ_{ij}^*, of the rotated loading matrix, Λ^*, are either relatively large or small. The rotations may be found by minimizing the criterion

$$V = \sum_{j=1}^{k}\sum_{i=1}^{p}\left(\lambda_{ij}^*\right)^4 - \frac{\gamma}{p}\sum_{j=1}^{k}\left(\sum_{i=1}^{p}\left(\lambda_{ij}^*\right)^2\right)^2$$

where the constant, γ, gives a family of rotations, with $\gamma = 1$ giving **varimax rotations** and $\gamma = 0$ giving **quartimax** rotations.

Given an orthogonal rotation matrix X, a solution may be further simplified by removing the orthogonality restriction with an oblique **ProMax** rotation. Let Y denote the matrix defined by a power transformation of X, designed to increase high values in X and decrease low values. Then the ProMax solution is based on a least squares fit of X to Y.

For (b) **Procrustes** rotations are used. Let A and B be two l by m matrices, which can be considered as representing l points in m dimensions. One example is if A is the loading matrix from a variable-directed multivariate method and B is a hypothesised pattern matrix. In order to try to match the points in A and B there are three steps:

(i) translate so that centroids of both matrices are at the origin,

(ii) find a rotation that minimizes the sum of squared distances between corresponding points of the matrices,

(iii) scale the matrices.

For a more detailed description, see Krzanowski (1990).

2.2 Individual-directed Methods

While dealing with the same n by p data matrix as variable-directed methods, the emphasis is the n objects or individuals rather than the p variables. The methods are generally based on an n by n distance or dissimilarity matrix such that the (k, j)th element gives a measure of how 'far apart' the individuals k and j are. Alternatively, a similarity matrix can be used which measures how 'close' individuals are. The form of the measure of distance or similarity will depend upon the form of the p variables. For continuous variables it is usually assumed that some form of Euclidean distance is suitable. That is, for x_{ki} and x_{ji} measured for individuals k and j on variable i respectively, the contribution to distance between individuals k and j from variable i is given by

$$\left(x_{ki} - x_{ji}\right)^2.$$

Often there will be a need to scale the variables to produce satisfactory distances. For discrete variables, there are various measures of similarity or distance that can easily be computed. For example, for binary data a measure of similarity could be

1 – if the individuals take the same value,

0 – otherwise.

Given a measure of distance between individuals, there are three basic tasks that can be performed.

(i) *Group the individuals;* that is, collect the individuals into groups so that those within a group are closer to each other than they are to members of another group.

(ii) *Classify individuals;* that is, if some individuals are known to come from certain groups, allocate individuals whose group membership is unknown, to the nearest group.

(iii) *Map the individuals;* that is, produce a multidimensional diagram in which the distances on the diagram represent the distances between the individuals.

In the above, (i) leads to cluster analysis, (ii) leads to discriminant analysis and (iii) leads to scaling methods.

2.2.1 Hierarchical cluster analysis

Approaches for cluster analysis can be classified into two types: hierarchical and non-hierarchical. Hierarchical cluster analysis produces a series of overlapping groups or clusters ranging from separate individuals to one single cluster. For example, five individuals could be hierarchically clustered as follows.

$$\begin{array}{lccccc} \text{Step 1} & (1) & (2) & (3) & (4) & (5) \\ \text{Step 2} & (1,2) & & (3) & (4) & (5) \\ \text{Step 3} & (1,2) & & & (3,4) & (5) \\ \text{Step 4} & (1,2) & & & (3,4,5) & \\ \text{Step 5} & & & (1,2,3,4,5) & & \end{array}$$

The clusters at a level are constructed from the clusters at a previous level. There are two basic approaches to hierarchical cluster analysis: agglomerative methods which build up clusters starting from individuals until there is only one cluster, or divisive methods which start with a single cluster and split clusters until the individual level is reached. This chapter contains the more common agglomerative methods.

The stages in a hierarchical cluster analysis are usually as follows.

(i) form a distance matrix;

(ii) use selected criterion to form hierarchy;

(iii) print cluster information in the form of a dendrogram or use information to form a set of clusters.

These three stages will be considered in turn.

(i) Form a distance matrix

For the n by p data matrix X, a general measure of the distance between object j and object k, d_{jk}, is

$$d_{jk} = \left(\sum_{i=1}^{p} D\left(x_{ji}/s_i, x_{ki}/s_i\right) \right)^{\alpha},$$

where x_{ji} and x_{ki} are the (j,i)th and (k,i)th elements of X, s_i is a standardization for the ith variable and $D(u,v)$ is a suitable function. Three common distances for continuous variables are:

(a) Euclidean distance: $D(u,v) = (u-v)^2$ and $\alpha = \frac{1}{2}$.

(b) Euclidean squared distance: $D(u,v) = (u-v)^2$ and $\alpha = 1$.

(c) Absolute distance (city block metric): $D(u,v) = |u-v|$ and $\alpha = 1$.

The common standardizations are the standard deviation and the range. For dichotomous variables there are a number of different measures (see Krzanowski (1990) and Everitt (1974)); these are usually easy to compute. If the individuals in a cluster analysis are themselves variables, then a suitable distance measure will be based on the correlation coefficient for continuous variables and contingency table statistics for discrete data.

(ii) Form Hierarchy

Given a distance matrix for the n individuals, an agglomerative clustering method produces a hierarchical tree by starting with n clusters, each with a single individual and then at each of $n-1$ stages, merging two clusters to form a larger cluster until all individuals are in a single cluster. At each stage, the two clusters that are nearest are merged to form a new cluster and a new distance matrix is computed for the reduced number of clusters.

Methods differ as to how the distances between the new cluster and other clusters are computed. For three clusters i, j and k, let n_i, n_j and n_k be the number of objects in each cluster, and let d_{ij},

d_{ik} and d_{jk} be the distances between the clusters. If clusters j and k, are to be merged to give cluster jk, then the distance from cluster i to cluster jk, $d_{i.jk}$, can be computed in the following ways.

(a) Single link or nearest neighbour: $d_{i.jk} = \min\big(d_{ij}, d_{ik}\big)$.

(b) Complete link or furthest neighbour: $d_{i.jk} = \max\big(d_{ij}, d_{ik}\big)$.

(c) Group average: $d_{i.jk} = \frac{n_j}{n_j+n_k}d_{ij} + \frac{n_k}{n_j+n_k}d_{ik}$.

(d) Centroid: $d_{i.jk} = \frac{n_j}{n_j+n_k}d_{ij} + \frac{n_k}{n_j+n_k}d_{ik} - \frac{n_j n_k}{\big(n_j+n_k\big)^2}d_{jk}$.

(e) Median: $d_{i.jk} = \frac{1}{2}d_{ij} + \frac{1}{2}d_{ik} - \frac{1}{4}d_{jk}$.

(f) Minimum variance: $d_{i.jk} = \big[\big(n_i + n_j\big)d_{ij} + (n_i + n_k)d_{ik} - n_i d_{jk}\big] / \big(n_i + n_j + n_k\big)$.

For further details, see Everitt (1974) or Krzanowski (1990).

(iii) Produce Dendrogram and Clusters

Hierarchical cluster analysis can be represented by a tree that shows at which distance the clusters merge. Such a tree is known as a dendrogram; see Everitt (1974) and Krzanowski (1990).

A simple example is

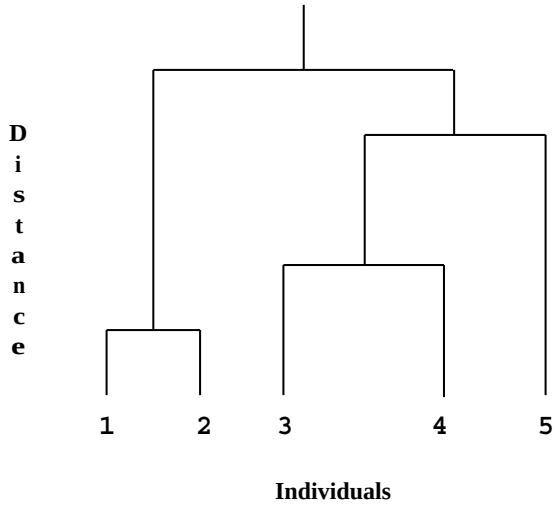

Figure 1

The end points of the dendrogram represent the individuals that have been clustered.

Alternatively, the information from the tree can be used to produce either a chosen number of clusters or the clusters that exist at a given distance. The latter is equivalent to taking the dendrogram and drawing a line across at a given distance to produce clusters.

2.2.2 Non-hierarchical clustering

Non-hierarchical cluster analysis usually forms a given number of clusters from the data. There is no requirement that if first $k-1$ and then k clusters were requested then the $k-1$ clusters would be formed from the k clusters.

Most non-hierarchical methods of cluster analysis seek to partition the set of individuals into a number of clusters so as to optimize a criterion. The number of clusters is usually specified prior to the analysis. One commonly used criterion is the within-cluster sum of squares. Given n individuals with p variables measured on each individual, x_{ij}, for $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, p$, the within-cluster sum of squares for K clusters is

$$\mathrm{SS}_c = \sum_{k=1}^{K}\sum_{i\in S_k}\sum_{j=1}^{p}\big(x_{ij} - \bar{x}_{kj}\big)^2,$$

where S_k is the set of objects in the kth cluster and $\bar{x}_{kj}$ is the mean for the variable j over cluster k. Starting with an initial allocation of individuals to clusters, the method then seeks to minimize SS_c by a series of re-allocations. This is often known as K-means clustering.

In the K-means case individuals belong to a single cluster and are excluded from all remaining clusters. Alternatively, probabilities of cluster membership can be estimated and each cluster can have its own distributional properties. For example, given an initial set of probabilities, the Normal (Gaussian) mixture model uses the E–M method of Dempster *et al.* (1977) to maximize the sum of log-likelihoods over K clusters for a given covariance model ranging from pooled variance to individual covariance matrices.

2.2.3 Discriminant analysis

Discriminant analysis is concerned with the **allocation** of objects to n_g groups on the basis of observations on those objects using an allocation rule. This rule is computed from observations coming from a **training set** in which group membership is known. The allocation rule is based on the distance between the object and an estimate of the location of the groups. If p variables are observed and the vector of means for the jth group in the training set are $\bar{x}_j$ then the usual measure of the distance of an observation, x_k, from the jth group mean is given by Mahalanobis squared distance

$$D_{kj}^2 = \big(x_k - \bar{x}_j\big)^{\mathrm{T}} S_*^{-1} \big(x_k - \bar{x}_j\big),$$

where S_* is either the within-group variance-covariance matrix, S_j, for the n_j objects in the jth group, or a pooled variance-covariance matrix, S, computed from all n objects from all groups where

$$S = \frac{\displaystyle\sum_{j=1}^{n_g} \big(n_j - 1\big) S_j}{\big(n - n_g\big)}.$$

If the within-group variance-covariance matrices can be assumed to be equal then the pooled variance-covariance matrix can be used. This assumption can be tested using the test statistic

$$G = C\left(\big(n - n_g\big)\log|S| - \sum_{j=1}^{n_g} \big(n_j - 1\big)\log|S_j| \right),$$

where

$$C = 1 - \frac{2p^2 + 3p - 1}{6(p+1)\big(n_g - 1\big)} \left(\sum_{j=1}^{n_g} \frac{1}{\big(n_j - 1\big)} - \frac{1}{\big(n - n_g\big)} \right).$$

For large n, G is approximately distributed as a χ^2 variable with $\frac{1}{2}p(p+1)\big(n_g - 1\big)$ degrees of freedom; see Morrison (1967).

In addition to the distances, a set of prior probabilities of group membership, π_j, for $j = 1, 2, \ldots, n_g$, may be used. The prior probabilities reflect your view as to the likelihood of the objects coming from the different groups.

It is generally assumed that the p variables follow a multivariate Normal distribution with, for the jth group, mean μ_j and variance-covariance matrix Σ_j. If $p\big(x_k \mid \mu_j, \Sigma_j\big)$ is the probability of observing the observation x_k from group j, then the posterior probability of belonging to group j is

$$p\big(j \mid x_k, \mu_j, \Sigma_j\big) \propto p\big(x_k \mid \mu_j, \Sigma_j\big)\pi_j.$$

An observation is allocated to the group with the highest posterior probability.

In the **estimative** approach to discrimination, the parameters μ_j and Σ_j in $p\big(j \mid x_k, \mu_j, \Sigma_j\big)$ are replaced by their estimates calculated from the training set. If it is assumed that the within-group variance-covariance matrices are equal then the **linear discriminant function** is obtained; otherwise if it is assumed that the variance-covariance matrices are unequal then the **quadratic discriminant function** is obtained.

In the Bayesian **predictive** approach, a non-informative prior distribution is used for the parameters giving the posterior distribution for the parameters from the training set, X_t, of, $p(\mu_j, \Sigma_j \mid X_t)$. A predictive distribution is then obtained by integrating $p(j \mid x_k, \mu_j, \Sigma_j)p(\mu_j, \Sigma_j \mid X)$ over the parameter space. This predictive distribution, $p(x_k \mid X_t)$, then replaces $p(x_k \mid \mu_j, \Sigma_j)$ to give

$$p(j \mid x_k, \mu_j, \Sigma_j) \propto p(x_k \mid X_t)\pi_j.$$

In addition to allocating the objects to groups, an atypicality index for each object and for each group can be computed. This represents the probability of obtaining an observation more typical of the group than that observed. A high value of the atypicality index for all groups indicates that the observation may in fact come from a group not represented in the training set.

Alternative approaches to discrimination are the use of canonical variates and logistic discrimination. Canonical variate analysis is described above and as it seeks to find the directions that best discriminate between groups these directions can also be used to allocate further observations. This can be viewed as an extension of **Fisher's linear discriminant function**. This approach does not assume that the data is Normally distributed, but Fisher's linear discriminant function may not perform well on non-Normal data. In the case of two groups, logistic regression can be performed with the response variable indicating the group allocation and the variables in the discriminant analysis being the explanatory variables. Allocation can then be made on the basis of the fitted response value. This is known as **logistic discrimination** and can be shown to be valid for a wide range of distributional assumptions.

2.2.4 Scaling methods

Scaling methods seek to represent the observed dissimilarities or distances between objects as distances between points in Euclidean space. For example if the distances between objects A, B and C were 3, 4 and 5, the distances could be represented exactly by three points in two-dimensional space. Only their relative positions would be important, the whole configuration of points could be rotated or shifted without effecting the distances between the points. If a one-dimensional representation was required, the 'best' representation might give distances of $2\frac{1}{3}, 3\frac{1}{3}$ and $5\frac{2}{3}$, which may be an adequate representation. If the distances were 3, 4 and 8 then these distances could not be exactly represented in Euclidean space, even in two dimensions, the best representation being the three points in a straight line giving distances 3, 4 and 7.

In practice, the use of scaling methods has to decide upon the number of dimensions in which the data is to be represented. The smaller the number the easier it will be to assimilate the information. The chosen number of dimensions needs to give an adequate representation of the data but will often not give an exact representation because either the number of chosen dimensions is too small or the data cannot be represented in Euclidean space.

Two basic methods are available depending on the nature of the dissimilarities or distances being analysed. If the distances can be assumed to satisfy the metric inequality

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

then the distances can be represented exactly by points in Euclidean space and the technique known as metric scaling, classical scaling or principal coordinate analysis can be used. This technique involves the computing of the eigenvalues of a matrix derived from the distance matrix. The eigenvectors corresponding to the k largest positive eigenvalues gives the best k dimensions in which to represent the objects. If there are negative eigenvalues then the distance matrix cannot be represented in Euclidean space.

Instead of the above approach of requiring the distances from the points to match the distances from the objects as closely as possible, sometimes only a rank order equivalence is required. That is, the ith largest distance between objects should, as far as possible, be represented by the ith largest distance between points. This would be appropriate when the dissimilarities are based on subjective rankings. For example, if the objects were foods then a number of judges rank the foods for different qualities such as taste and texture, the resulting distances would not necessarily obey the metric inequality, but the rank order would be significant. Alternatively, by relaxing the requirement from matching distances to rank order equivalence only, the number of dimensions required to represent the distance matrix may be decreased. The requirement of rank order equivalence leads to non-metric or ordinal multi-dimensional scaling. The criterion used to measure the closeness of the fitted distance matrix to the

observed distance matrix is known as STRESS, which is given by

$$\sqrt{\frac{\sum_{i=1}^{n}\sum_{j=1}^{i-1}\left(\hat{d}_{ij}-\tilde{d}_{ij}\right)^2}{\sum_{i=1}^{n}\sum_{j=1}^{i-1}\hat{d}_{ij}^2}},$$

where $\hat{d}_{ij}^2$ is the Euclidean squared distance between the computed points i and j, and $\tilde{d}_{ij}$ is the fitted distance obtained when $\hat{d}_{ij}$ is monotonically regressed on the observed distances d_{ij}; that is, $\tilde{d}_{ij}$ is monotonic relative to d_{ij} and is obtained from $\hat{d}_{ij}$ with the smallest number of changes. Thus STRESS is a measure of by how much the set of points preserve the order of the distances in the original distance matrix, and non-metric multidimensional scaling seeks to find the set of points that minimize the STRESS.

3 Recommendations on Choice and Use of Available Functions

See Section 4 for a list of functions available in this Chapter.

Note also that nag_glm_binomial (g02gbc) will fit a logistic regression model and can be used for logistic discrimination.

4 Functionality Index

Canonical correlation analysis ..... nag_mv_canon_corr (g03adc)

Canonical variate analysis ..... nag_mv_canon_var (g03acc)

Cluster Analysis,
- compute distance matrix ..... nag_mv_distance_mat (g03eac)
- construct clusters following nag_mv_hierar_cluster_analysis (g03ecc) ..... nag_mv_cluster_indicator (g03ejc)
- construct dendrogram following nag_mv_hierar_cluster_analysis (g03ecc) ..... nag_mv_dendrogram (g03ehc)
- frees memory following nag_mv_dendrogram (g03ehc) ..... nag_mv_dend_free (g03xzc)
- Gaussian mixture model ..... nag_mv_gaussian_mixture (g03gac)
- hierarchical ..... nag_mv_hierar_cluster_analysis (g03ecc)
- K-means ..... nag_mv_kmeans_cluster_analysis (g03efc)

Discriminant Analysis,
- allocation of observations to groups, following nag_mv_discrim (g03dac) ..... nag_mv_discrim_group (g03dcc)
- Mahalanobis squared distances, following nag_mv_discrim (g03dac) ..... nag_mv_discrim_mahaldist (g03dbc)
- test for equality of within-group covariance matrices ..... nag_mv_discrim (g03dac)

Factor Analysis,
- factor score coefficients, following nag_mv_factor (g03cac) ..... nag_mv_fac_score (g03ccc)
- maximum likelihood estimates of parameters ..... nag_mv_factor (g03cac)

Principal component analysis ..... nag_mv_prin_comp (g03aac)

Rotations,
- orthogonal rotations for loading matrix ..... nag_mv_orthomax (g03bac)
- Procustes rotations ..... nag_mv_procustes (g03bcc)
- ProMax rotations ..... nag_mv_promax (g03bdc)

Scaling Methods,
- multidimensional scaling ..... nag_mv_ordinal_multidimscale (g03fcc)

principal coordinate analysis ..... nag_mv_prin_coord_analysis (g03fac)

Standardize values of a data matrix ..... nag_mv_z_scores (g03zac)

5 Auxiliary Functions Associated with Library Function Arguments

None.

6 Functions Withdrawn or Scheduled for Withdrawal

None.

7 References

Chatfield C and Collins A J (1980) *Introduction to Multivariate Analysis* Chapman and Hall

Dempster A P, Laird N M and Rubin D B (1977) Maximum likelihood from incomplete data via the *EM* algorithm (with discussion) *J. Roy. Statist. Soc. Ser. B* **39** 1–38

Everitt B S (1974) *Cluster Analysis* Heinemann

Gnanadesikan R (1977) *Methods for Statistical Data Analysis of Multivariate Observations* Wiley

Hammarling S (1985) The singular value decomposition in multivariate statistics *SIGNUM Newsl.* **20(3)** 2–25

Kendall M G and Stuart A (1976) *The Advanced Theory of Statistics (Volume 3)* (3rd Edition) Griffin

Krzanowski W J (1990) *Principles of Multivariate Analysis* Oxford University Press

Lawley D N and Maxwell A E (1971) *Factor Analysis as a Statistical Method* (2nd Edition) Butterworths

Morrison D F (1967) *Multivariate Statistical Methods* McGraw–Hill