Chapter 28
Multivariate Analysis

1 Scope of the Chapter

This chapter provides procedures for studying multivariate data.

2 Available Modules

Module 28.1: `nag_fac_analysis` — **Factor analysis and principal component analysis**

This module contains a procedure for computing principal component analysis from an n by m data matrix.

Module 28.2: `nag_canon_analysis` — **Canonical Analysis**

This module contains a procedure for computing canonical variate analysis for data on m variables from g groups.

Module 28.3: `nag_mv_rotation` — **Rotations**

This module contains a procedure for computing orthogonal factor rotations.

3 Background

3.1 Introduction

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 in multivariate analysis seek to examine the relationships between the p variables with the aim of reducing the dimensionality of the problem as compared with individual-directed methods which look the relationships between the individuals forming the data matrix. There are different variable-directed 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. All three 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 are any suitable rotations of these variables that lead to a simple interpretation of the new variables in terms of the original variables.

3.2 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 (with $a_1^T a_1 = 1$) of length p is found to maximise the variance $a_1^T S a_1$. The resulting 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 has the maximum variance. Further principal components can be derived such that each has maximum variance given that it is orthogonal to the previous components. Thus the original p correlated variables can be transformed to p orthogonal variables with decreasing variance. In practice the original variables will often be adequately represented by $k < p$ principal components in that the total variance of the k principal components represents a high proportion of the variance of the original p variables.

It can be shown that 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. An alternative way of approaching the problem is to consider the singular value decomposition of the scaled, mean-centred data matrix, X_s, (where the scaling is used to give results for the variance-covariance matrix rather than the sum of squares matrix). Computing the SVD of X_s gives

$$X_s = U\Gamma A^T$$

where A is the matrix containing the vectors a_i and Γ is the diagonal matrix containing the γ_i. The ratio $\gamma_i^2/\Sigma\gamma_i^2$ gives the proportion of variance explained by the ith principal component and is useful in selecting the number of principal components required to adequately represent the data. The values of the principal component variables for the individuals are known as the principal component scores and are given by U. These can be standardised so that the variance of these scores for each principal component is 1.0 or equal to the corresponding eigenvalue.

3.3 Factor Analysis

In some ways factor analysis is similar to principal component analysis but there are important differences in the underlying model.

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 $k < p$ hypothetical variables or factors, $f_1, f_2, \ldots, f_k$, while the variances are accounted for by unique components in addition to the factors. The factors 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 $i = 1, 2, \ldots, p$, $j = 1, 2, \ldots, k$, are the factor loadings and e_i, for $i = 1, 2, \ldots, p$, are independent random variables with variances ψ_i which represent the unique component of the variation of each of the p variables. 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

$$\Sigma = \Lambda\Lambda^T + \Psi$$

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

For a given Ψ the matrix Λ can be computed from the SVD of

$$X_s\Psi^{-1/2}$$

where X_s is the scaled, mean-centred data matrix. Thus Ψ can be seen as representing a variable weighting such that the higher the unique component variation the less weight the variable has in determining the common factors. Ψ can be estimated in several ways either fixed or iterating with the estimates of Λ. The simplest estimate of Ψ is given by the inverse of the diagonal elements of S^{-1}, where S is the sample variance-covariance matrix. This is known as principal factor analysis.

3.4 Canonical Variate Analysis

If the individuals can be classified into one of g groups then the total variation can be seen as the combination of between-group variation and within-group variation. The best discrimination between groups will be obtained by maximizing the ratio of the between-group variation to the within-group variation. Canonical variate analysis finds the linear combinations of the p variables which maximize this ratio. 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 ratio of the between-group sum of squares and cross-products matrix, B, to the within-group sums of squares and cross-products matrix, W. Alternatively they can be computed from the p by $(g-1)$ matrix

$$V = Q_x^T Q_g$$

where Q_g is an orthogonal matrix that defines the contrasts between 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, X_s,

$$X_s = Q_x R.$$

The within-group and between-group sums of squares and cross-products matrices can be written as

$$W = R^T(I - VV^T)R \quad \text{and} \quad B = R^T(VV^T)R.$$

Computing the SVD of V as

$$V = U_x \Delta U_g^T$$

gives the canonical correlations as the non-zero elements ($\delta_i > 0$) of the diagonal matrix Δ. 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 $W^{-1}B$ are given by

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

and the value of $\pi_i = \gamma_i^2 / \Sigma\gamma_i^2$ 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.

3.5 Rotations

Given a representation of p variables in $k < p$ dimensions, rotations can be used to simplify the relationship between the original variables and the variables chosen to define the k dimensions.

The most common type of rotations used 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

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

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