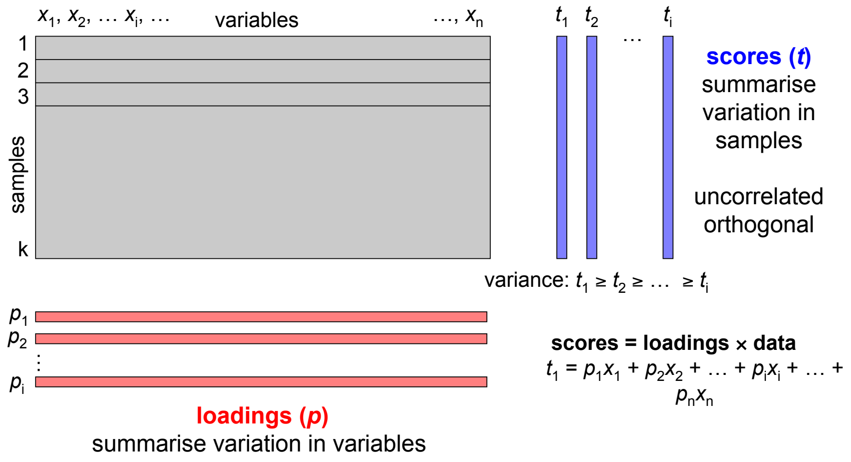
Multivariate data

Multivariate data consist of the results of observations of many different characters (variables) for a number of individuals (objects). Each variable may be regarded as constituting a different dimension, such that if there are *n* variables each object may be said to reside at a unique position in an abstract entity referred to as *n*-dimensional hyperspace. This hyperspace is necessarily difficult to visualise (!), and the underlying theme of multivariate analysis (MVA) is thus simplification or what is termed dimensionality reduction, which usually means that we want to summarise a large body of data by means of relatively few parameters, preferably the two or three which lend themselves to graphical display (biplots or 3D plots), with minimal loss of information.

Principal components analysis (PCA)

Conventionally the reduction of the multivariate data generated by metabolomics has normally been carried out using principal components analysis (PCA). PCA is a well-known technique for reducing the dimensionality of multivariate data whilst preserving most of the variance, and whilst it does not take account of any groupings in the data, neither does it require that the populations be normally distributed, i.e. it is a non-parametric method.


$x_1, x_2, \ldots x_i, \ldots$
variables
$\ldots, x_n$
1
2
3
samples
k
t_1 t_2 … t_i
scores (t)
summarise variation in samples
uncorrelated
orthogonal
variance: $t_1 \geq t_2 \geq \ldots \geq t_i$
p_1
p_2
p_i
loadings (p)
summarise variation in variables
scores = loadings × data
$t_1 = p_1x_1 + p_2x_2 + \ldots + p_ix_i + \ldots + p_nx_n$

PCA is described as <u>unsupervised</u> because no *a priori* information (e.g. diseased patients *vs*. healthy individuals) is used to aid the analysis. PCA is used to transform a number of (possibly) correlated variables into a (smaller) number of uncorrelated variables called principal components (PCs) that represent the natural variance in the data. The first principal component (PC1) accounts for the most variability in the data, and each succeeding PC accounts for as much of the remaining variability as possible, and so on. Therefore, analysis of the data should focus on the first few PCs since these account for the majority of the variation. These PCs scores can then be plotted and 'clusters' in the data visualized, and these plots may also be used to detect outliers. The information in terms of what is important for the scores plots can be determined by plotting the loadings plots (a weighted vector the same size as the original data).

Discriminant analysis (DA)

The closely-related discriminant analysis (DA; sometimes referred to as discriminant function analysis (DFA) or canonical variates analysis (CVA)) is often used to separate the objects (samples) into groups on the basis of the retained PCs and the *a priori* knowledge of the appropriate number of groupings; this is achieved by minimising the within-group variance and maximising the between-group variance. Provided that the data set contains "standards" (i.e. known things) it is evident that one can establish the closeness of any unknown samples to a standard, and thus effect the identification of the former, a technique termed 'guilt by association'.

An important thing to note is that DA is not usually performed on the original feature space (metabolite data) because one can not feed co-linear variables or too many variables into this algorithm. The starting point for DA is the inverse of the pooled variance-covariance matrix within *a priori* groups. This inverse can only exist when the matrix is non-singular, i.e., its determinant is other than zero, which implies that it is of full rank; i.e.

Generally if: $(N_s - N_g - 1) > N_v$ where N_s = Number of samples (*e.g.* patients)
N_g = Number of groups (*e.g.* diseased *vs*. healthy)
N_v = Number of inputs (metabolites).