*Tutorials and worked examples for simulation,
curve fitting, statistical analysis, and plotting.*
https://simfit.org.uk
https://simfit.silverfrost.com

Multivariate analysis is used to study n by m data matrices where the n rows represent subjects while the m columns are values for variables observed for the n subjects.

To be precise, consider the possible outcome from testing eight people exposed to mosquito attacks with five different types of clothing as follows, where a 1 indicates attacked by mosquitos and a 0 indicates freedom from attack.

Blocks	Groups (Clothing Type)				
(Subjects)	Light-loose	Light-tight	Dark-long	Dark-short	None
1	0	0	0	1	0
2	1	1	1	1	1
3	0	0	0	1	1
4	1	1	0	1	0
5	0	1	1	1	1
6	0	1	0	0	1
7	0	0	1	1	1
8	0	0	1	1	0

Here there are 8 subjects and 5 variables, but in addition there is a first column identifying the subjects. So the actual data matrix used for analysis would have dimensions $n = 8$ and $m = 5$, whereas the above table has $n = 8$ and $m = 6$ because, with some multivariate techniques provided by SIMFIT, the additional first column can be used to identify subjects if the data are rearranged into groups, or if some subjects are excluded from analysis. Note also that it is often useful to exclude selected variables from an analysis and so, if this is done, the remaining columns will be re-numbered.

So, from now on, we shall consider a matrix X with elements x_{ij} as follows

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix}$$

where all the values are to be used in a current analysis.

Almost all multivariate techniques require that the vector of column means and the covariance matrix should be estimated from X and, in addition, subsequent analysis will usually require a singular value decomposition (SVD) because it is the most reliable method for determining the rank of a matrix. SIMFIT provides the ability to check the rank of any matrix in this way, and this should be done with any data matrices that prove problematical to analyze.

It should be obvious that the units of measurements for the variables should lead to similar values for the x_{ij} so that all m variables have comparable means and variances, otherwise columns with large values will dominate columns with small values. This can be achieved by centralizing the matrix by subtracting the column means, and then normalizing by dividing by the column standard deviation. If this is required then the SIMFIT program **editmt** can be used to pre-process data matrices, or it can be done interactively before the data are submitted for analysis, or performed automatically by the routine. However, care must be exercised when centralizing and normalizing because some techniques can give biased results if this is done uncritically. For instance, partial least squares will give biased predictions if data sets for calibration and prediction are pre-processed uncritically before analysis.