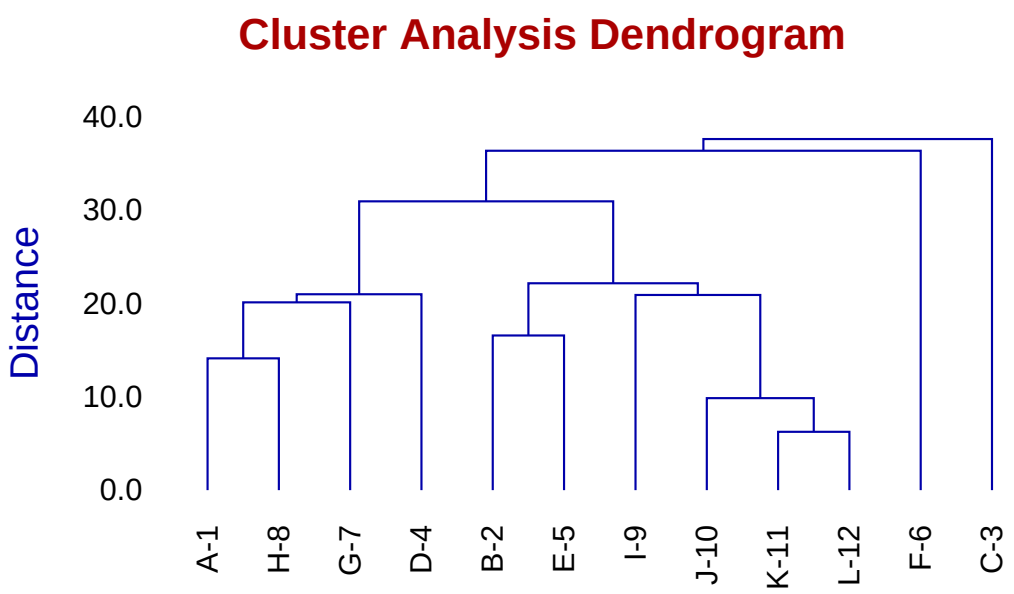
Dendrograms can be plotted after a distance matrix has been calculated and a linkage technique has been selected in order to build up a picture as to how merging can be used to partition samples into subgroups as defined by distance thresholds.

For example, open the main SIMFIT menu choose [Statistics], [Multivariate], then [Dendrograms] and read in the test file `cluster.tf1`, which should also be examined to see how to provide labels, as illustrated when this figure is displayed.




Of course the precise shape of such a figure depends on the metric and weights, etc. used to calculate the distance matrix and the linkage assumed when building up the groups. Further details about using SIMFIT to construct dendrograms will now be discussed.

## Partial clustering

An important application of distance matrices and dendrograms is in partial clustering. Unlike the situation with full clustering where we start with $n$ groups, each containing a single case, and finish with just one group containing all the cases, in partial clustering the clustering process is not allowed to be completed. There are two distinct ways to arrest the clustering procedure.

1. A number, $K$, between 1 and $n-1$ is chosen, and clustering is allowed to proceed until just $K$ subgroups have been formed. It may not always be possible to satisfy this requirement, e.g. if there are ties in the data.

2. A threshold, $D$, is set somewhere between the first clustering distance and the last clustering distance, and clustering terminates when this threshold is reached. The position of such clustering thresholds will be plotted on the dendrogram, unless $D$ is set equal to zero.

As an example of this technique consider the results in this table

Group assignments for Fisher Iris data
Data file: `iris.tf1`, 3 groups, variables included: 1 2 3 4
Transformation: Untransformed, Distance: Euclidean, Scaling: Unscaled,
Linkage: Group average, [weights not used], sub-clusters for K = 3
Odd rows: data ... Even rows: corresponding group number

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 |
| 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 25 | 26 | 27 | 28 | 29 | 30 | 31 | 32 | 33 | 34 | 35 | 36 |
| 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 37 | 38 | 39 | 40 | 41 | 42 | 43 | 44 | 45 | 46 | 47 | 48 |
| 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 49 | 50 | 51 | 52 | 53 | 54 | 55 | 56 | 57 | 58 | 59 | 60 |
| 1 | 1 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| 61 | 62 | 63 | 64 | 65 | 66 | 67 | 68 | 69 | 70 | 71 | 72 |
| 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| 73 | 74 | 75 | 76 | 77 | 78 | 79 | 80 | 81 | 82 | 83 | 84 |
| 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| 85 | 86 | 87 | 88 | 89 | 90 | 91 | 92 | 93 | 94 | 95 | 96 |
| 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| 97 | 98 | 99 | 100 | 101 | 102 | 103 | 104 | 105 | 106 | 107 | 108 |
| 2 | 2 | 2 | 2 | 2* | 2* | 3 | 2* | 2* | 3 | 2* | 3 |
| 109 | 110 | 111 | 112 | 113 | 114 | 115 | 116 | 117 | 118 | 119 | 120 |
| 2* | 3 | 2* | 2* | 2* | 2* | 2* | 2* | 2* | 3 | 3 | 2* |
| 121 | 122 | 123 | 124 | 125 | 126 | 127 | 128 | 129 | 130 | 131 | 132 |
| 2* | 2* | 3 | 2* | 2* | 3 | 2* | 2* | 2* | 3 | 3 | 3 |
| 133 | 134 | 135 | 136 | 137 | 138 | 139 | 140 | 141 | 142 | 143 | 144 |
| 2* | 2* | 2* | 3 | 2* | 2* | 2* | 2* | 2* | 2* | 2* | 2* |
| 145 | 146 | 147 | 148 | 149 | 150 | | | | | | |
| 2* | 2* | 2* | 2* | 2* | 2* | | | | | | |

This resulted from analysis of the famous Fisher iris data set in `iris.tf1` when $K = 3$ subgroups were requested.

We note that groups 1 (setosa) and 2 (versicolor) contained the all the cases from the known classification, but most of the known group 3 (virginica) cases (those identified by asterisks) were also assigned to subgroup 2. This table should also be compared to a table resulting from $K$-means clustering analysis of the same data set.

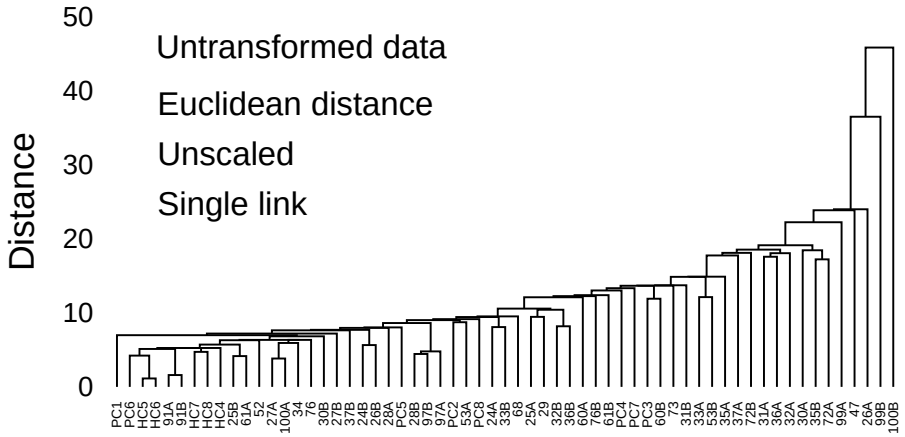
From the SimFIT dendrogram partial clustering procedure it is also possible to create a SimFIT MANOVA type file for any type of subsequent MANOVA analysis and, to aid in the use of dendrogram clusters as training sets for allocating new observations to groups, the subgroup centroids are also appended to such files. Alternatively a file ready for K-means cluster analysis can be saved, with group centroids appended to serve as starting estimates.

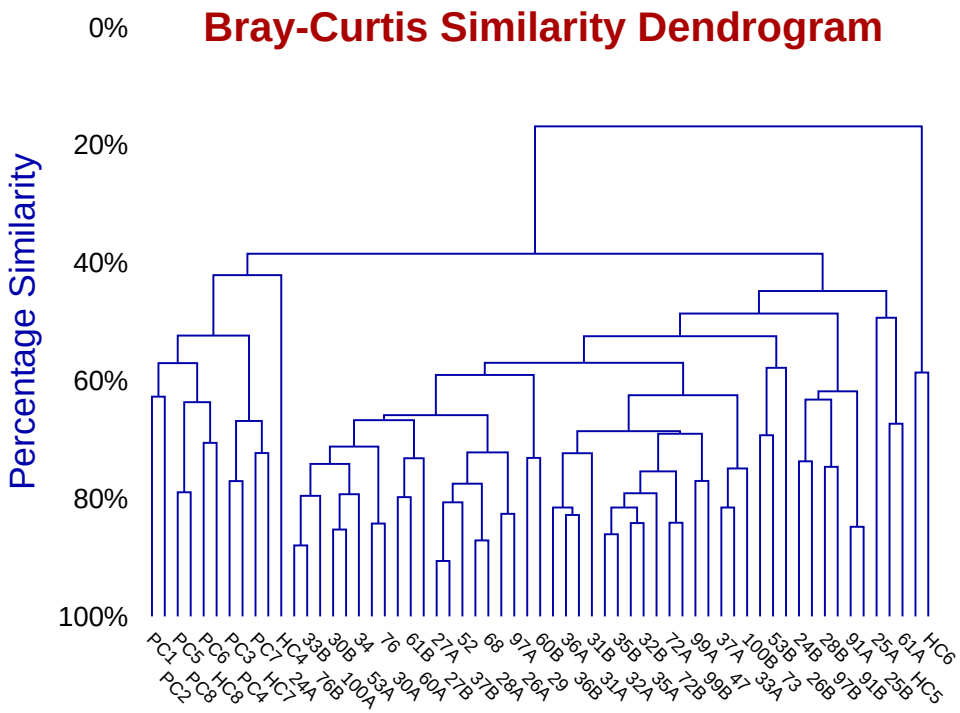
Finally, attention should be drawn to the advanced techniques provided by SimFIT for plotting dendrogram thresholds and subgroups illustrated next.

## Plotting dendrograms: standard format

Dendrogram shape is arbitrary in two ways; the $x$ axis order is arbitrary as clusters can be rotated around any clustering distance leading to $2^{n-1}$ different orders, and the distance matrix depends on the settings used. For instance, using a square root transformation, Bray-Curtis similarity, and a group average link generates the second dendrogram in this figure from the first. The data were contained in `cluster.tf2`, $y$ plotted are dissimilarities, while labels are $100 - y$, which should be remembered when changing the $y$ axis range.

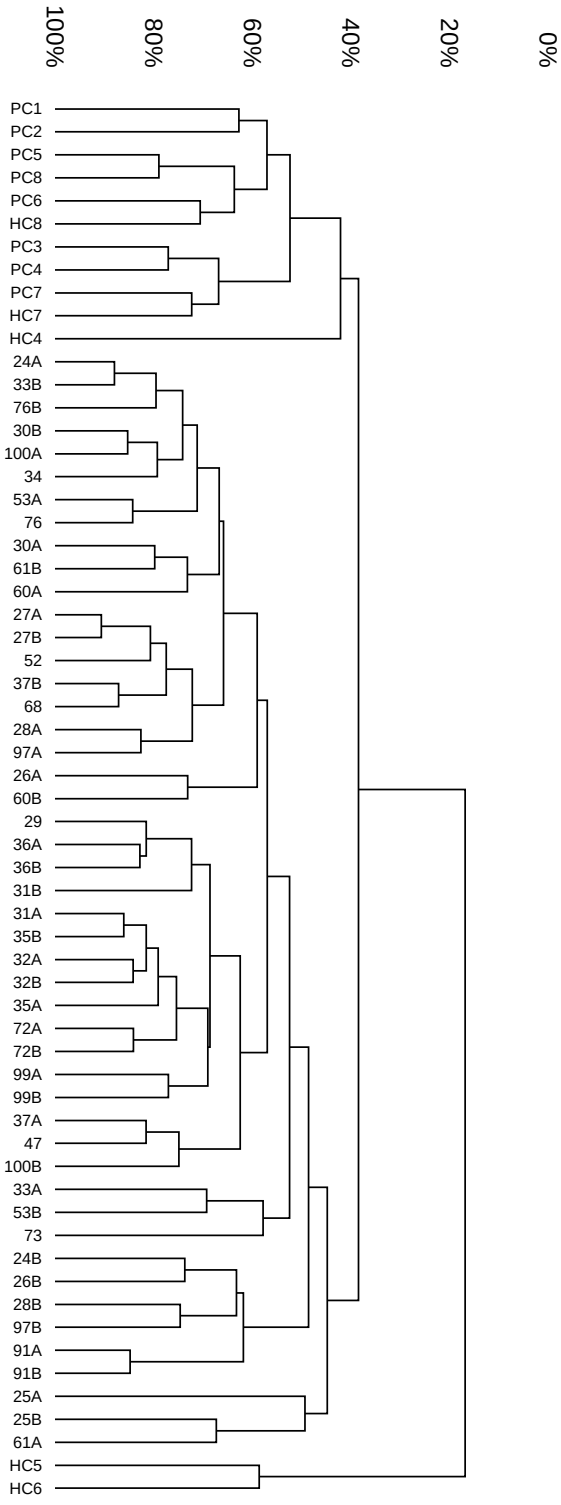
Users should not manipulate dendrogram parameters to create a dendrogram supporting some preconceived clustering scheme. You can set a label threshold and translation distance from the [$X$-axis] menu so that, if the number of labels exceeds the threshold, even numbered labels are translated, and font size is decreased.



## Plotting dendrograms: stretched format

Sometimes dendrograms are more readable if the white space is stretched without distorting the labels. So SimFIT PostScript graphs have a very useful feature: you can stretch or compress the white space between plotted lines and symbols without changing the line thickness, symbol size, or font size and aspect ratio. For instance, stretching, clipping and sliding procedures are valuable in graphs which are crowded due to overlapping symbols or labels, as in previous figures. If such dendrograms are stretched retrospectively using **editps**, the labels will not separate as the fonts will also be stretched so letters become ugly due to altered aspect ratios. SimFIT can increase white space between symbols and labels while maintaining correct aspect ratios for the fonts in PostScript hardcopy and, to explain this, the creation of this figure using the data in `cluster.tf2` will be described.

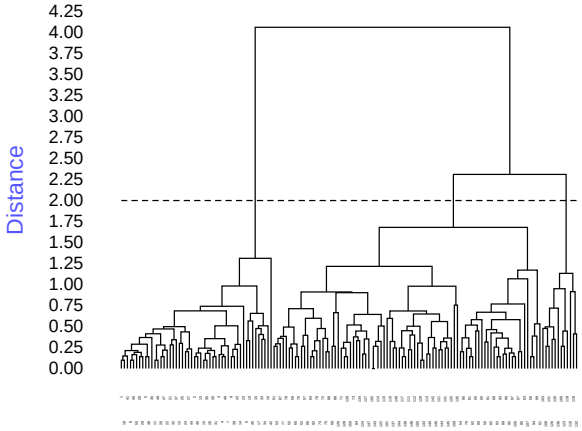



The title, legend and double $x$ labeling were suppressed, and landscape mode with stretching, clipping and sliding was selected from the PostScript control using the [Shape] then [Landscape +] options, with an $x$ stretching factor of two. Stretching increases the space between each symbol, or the start of each character string, arrow or other graphical object, but does not turn circles into ellipses or distort letters. As graphs are often stretched to print on several sheets of paper, sub-sections of the graph can be clipped out, then the clipped sub-sections can be slid to the start of the original coordinate system to facilitate printing.

If stretch factors greater than two are used, legends tend to become detached from axes, and empty white space round the graph increases. To remedy the former complication, the default legends should be suppressed or replaced by more closely positioned legends while, to cure the later effect, GSview can be used to calculate new BoundingBox coordinates (by transforming .ps to .eps). If you select the option to plot an opaque background even when white (by mistake), you may then find it necessary to edit the resulting .eps file in a text editor to adjust the clipping coordinates (identified by %#clip in the .eps file) and background polygon filling coordinates (identified by %#pf in the .ps file) to trim away unwanted white background borders that are ignored by GSview when calculating BoundingBox coordinates. Another example of this technique is with meta analysis plots, where it is also pointed out that creating transparent backgrounds by suppressing the painting of a white background obviates the need to clip away extraneous white space.

## Plotting dendrograms: subgroups

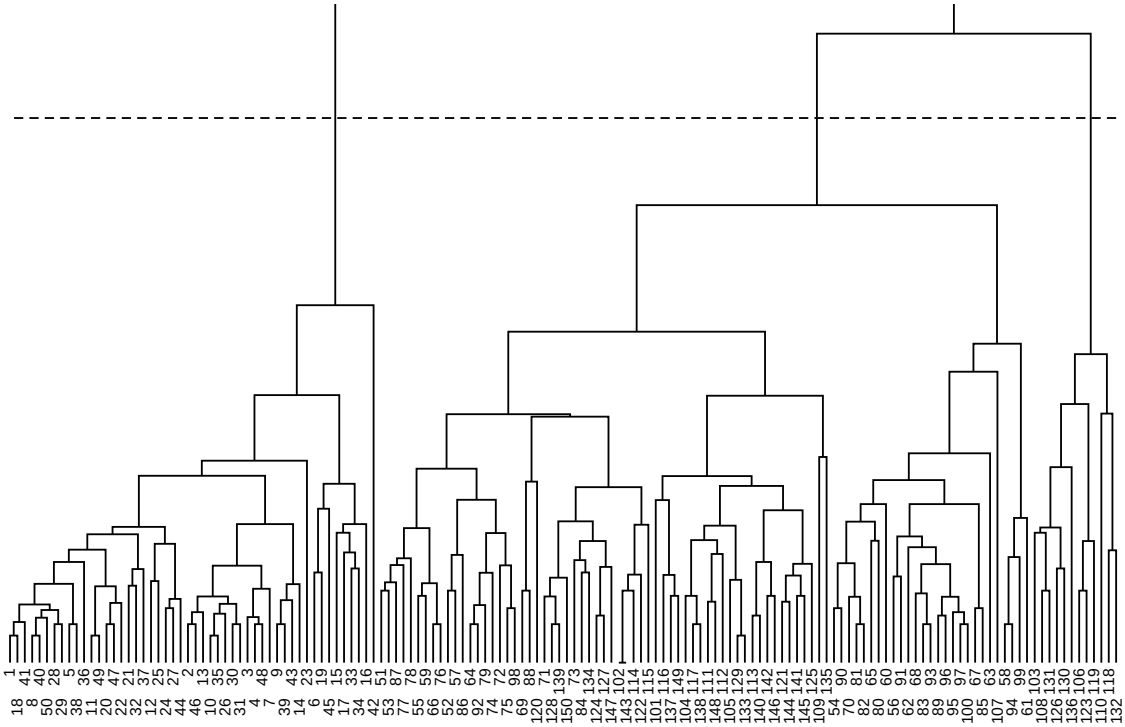
The procedures described can also be used to improve the readability of dendrograms where subgroups have been assigned by partial clustering. The next figure shows a graph from `iris.tf1` when three subgroups are requested, or a threshold is set corresponding to the horizontal dotted line. The figure was created by these steps.




First the title was suppressed, the $y$-axis range was changed to $(0, 4.25)$ with 18 tick marks, the $(x, y)$ offset was canceled as this suppresses axis moving, the label font size was increased from 1 to 3, and the $x$-axis was translated to 0.8.

Then the PostScript stretch/slide/clip procedure was used with these parameters

$$
\begin{aligned}
x_{\text{stretch}} &= 1.5 \\
y_{\text{stretch}} &= 2.0 \\
x_{\text{clip}} &= 0.15, 0.95 \\
y_{\text{clip}} &= 0.10, 0.60.
\end{aligned}
$$



Windows users without PostScript printing facilities must create a `*.eps` file using this technique, then use the SimFIT procedures to create a graphics file they can use, e.g. `*.jpg`. Use of a larger font and increased $x$-stretching would be required to read the labels, of course.