cluster dendrogram — Dendrograms for hierarchical cluster analysis

Description

cluster dendrogram produces dendrograms (also called cluster trees) for a hierarchical clustering. See [MV] cluster for a discussion of cluster analysis, hierarchical clustering, and the available cluster commands.

Dendrograms graphically present the information concerning which observations are grouped together at various levels of (dis)similarity. At the bottom of the dendrogram, each observation is considered its own cluster. Vertical lines extend up for each observation, and at various (dis)similarity values, these lines are connected to the lines from other observations with a horizontal line. The observations continue to combine until, at the top of the dendrogram, all observations are grouped together.

The height of the vertical lines and the range of the (dis)similarity axis give visual clues about the strength of the clustering. Long vertical lines indicate more distinct separation between the groups. Long vertical lines at the top of the dendrogram indicate that the groups represented by those lines are well separated from one another. Shorter lines indicate groups that are not as distinct.

Quick start

Dendrogram of most recent cluster analysis

  cluster dendrogram

Same as above

  cluster tree

As above, but orient horizontally instead of vertically

  cluster tree, horizontal

Dendrogram of cluster analysis named myclus

  cluster tree myclus

As above, and apply leaf labels from variable mylabels instead of observation numbers

  cluster tree myclus, labels(mylabels)

As above, but rotate leaf labels 90 degrees and reduce text size by half

  cluster tree myclus, labels(mylabels) ///
    xlabel(, angle(90) labsize(*.5))

Show top 20 branches and associated frequencies from most recent cluster analysis

  cluster tree, cutnumber(20) showcount
Menu

Statistics > Multivariate analysis > Cluster analysis > Postclustering > Dendrograms

Syntax

```
cluster dendrogram [clname] [if] [in] [, options]
```

<table>
<thead>
<tr>
<th>option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>quick</td>
<td>do not center parent branches</td>
</tr>
<tr>
<td>labels(varname)</td>
<td>name of variable containing leaf labels</td>
</tr>
<tr>
<td>cutnumber(#)</td>
<td>display top # branches only</td>
</tr>
<tr>
<td>cutvalue(#)</td>
<td>display branches above # (dis)similarity measure only</td>
</tr>
<tr>
<td>showcount</td>
<td>display number of observations for each branch</td>
</tr>
<tr>
<td>countprefix(string)</td>
<td>prefix the branch count with \textit{string}; default is \texttt{n=}</td>
</tr>
<tr>
<td>countsuffix(string)</td>
<td>suffix the branch count with \textit{string}; default is empty string</td>
</tr>
<tr>
<td>countinline</td>
<td>put branch count in line with branch label</td>
</tr>
<tr>
<td>vertical</td>
<td>orient dendrogram vertically (default)</td>
</tr>
<tr>
<td>horizontal</td>
<td>orient dendrogram horizontally</td>
</tr>
</tbody>
</table>

Plot

```
line_options
```

affect rendition of the plotted lines

Add plots

```
addplot(plot)
```

add other plots to the dendrogram

Y axis, X axis, Titles, Legend, Overall

```
twoway_options
```

any options other than by() documented in [G-3] twoway_options

Note: cluster tree is a synonym for cluster dendrogram.

In addition to the restrictions imposed by if and in, the observations are automatically restricted to those that were used in the cluster analysis.

Options

**Main**

quick switches to a different style of dendrogram in which the vertical lines go straight up from the observations instead of the default action of being recentered after each merge of observations in the dendrogram hierarchy. Some people prefer this representation, and it is quicker to render.

labels(varname) specifies that varname be used in place of observation numbers for labeling the observations at the bottom of the dendrogram.

cutnumber(#) displays only the top # branches of the dendrogram. With large dendrograms, the lower levels of the tree can become too crowded. With cutnumber(), you can limit your view to the upper portion of the dendrogram. Also see the cutvalue() option.

cutvalue(#) displays only those branches of the dendrogram that are above the # (dis)similarity measure. With large dendrograms, the lower levels of the tree can become too crowded. With cutvalue(), you can limit your view to the upper portion of the dendrogram. Also see the cutnumber() option.
showcount requests that the number of observations associated with each branch be displayed below the branches. showcount is most useful with cutnumber() and cutvalue() because, otherwise, the number of observations for each branch is one. When this option is specified, a label for each branch is constructed by using a prefix string, the branch count, and a suffix string.

countprefix(string) specifies the prefix string for the branch count label. The default is countprefix(n=). This option implies the use of the showcount option.

countsuffix(string) specifies the suffix string for the branch count label. The default is an empty string. This option implies the use of the showcount option.

countinline requests that the branch count be put in line with the corresponding branch label. The branch count is placed below the branch label by default. This option implies the use of the showcount option.

vertical and horizontal specify whether the $x$ and $y$ coordinates are to be swapped before plotting—vertical (the default) does not swap the coordinates, whereas horizontal does.

Plot line options affect the rendition of the lines; see [G-3] line_options.

Add plots addplot(plot) allows adding more graph twoway plots to the graph; see [G-3] addplot_option.

twoway_options are any of the options documented in [G-3] twoway_options, excluding by(). These include options for titling the graph (see [G-3] title_options) and for saving the graph to disk (see [G-3] saving_option).

Remarks and examples stata.com

Examples of the cluster dendrogram command can be found in [MV] cluster linkage, [MV] clustermat, [MV] cluster stop, and [MV] cluster generate. Here we illustrate some of the additional options available with cluster dendrogram.
Example 1 of [MV] cluster linkage introduces a dataset with 50 observations on four variables. Here we show the dendrogram for a complete-linkage analysis:

```
use https://www.stata-press.com/data/r16/labtech
cluster completelinkage x1 x2 x3 x4, name(L2clnk)
cluster dendrogram L2clnk, labels(labtech) xlabel(, angle(90) labsize(*.75))
```

The same dendrogram can be rendered in a slightly different format by using the quick option:

```
cluster dendrogram L2clnk, quick labels(labtech)
xlabel(, angle(90) labsize(*.75))
```

Some people prefer this style of dendrogram. As the name implies, this style of dendrogram is quicker to render.

You can use the if and in conditions to restrict the dendrogram to the observations for one subgroup. This task is usually accomplished with the cluster generate command, which creates a grouping variable; see [MV] cluster generate.
Here we show the third of three groups in the dendrogram by first generating the grouping variable for three groups and then using `if` in the command for `cluster dendrogram` to restrict it to the third of those three groups.

```
. cluster generate g3 = group(3)
. cluster tree if g3==3
```

Because we find it easier to type, we used the synonym `tree` instead of `dendrogram`. We did not specify the cluster name, allowing it to default to the most recently performed cluster analysis. We also omitted the `labels()` and `xlabel()` options, which brings us back to the default action of showing, horizontally, the observation numbers.

This example has only 50 observations. When there are many observations, the dendrogram can become too crowded. You will need to limit which part of the dendrogram you display. One way to view only part of the dendrogram is to use `if` and `in` to limit to one particular group, as we did above.

The other way to limit your view of the dendrogram is to specify that you wish to view only the top portion of the tree. The `cutnumber()` and `cutvalue()` options allow you to do this:
We limited our view to the top 15 branches of the dendrogram with `cutn(15)`. By default, the 15 branches were labeled $G_1$–$G_{15}$. The `showcount` option provided, below these branch labels, the number of observations in each of the 15 groups.

The `cutvalue()` option provides another way to limit the view to the top branches of the dendrogram. With this option, you specify the similarity or dissimilarity value at which to trim the tree.

```
. cluster tree, cutvalue(75.3)
    countprefix("(") countsuffix(" obs") countinline
    ylabel(, angle(0)) horizontal
```

This time, we limited the dendrogram to those branches with dissimilarity greater than 75.3 by using the `cutvalue(75.3)` option. There were 16 branches (groups) that met that restriction. We used the `countprefix()` and `countsuffix()` options to display the number of observations in each branch as “(# obs)” instead of “n=#”. The `countinline` option puts the branch counts in line with...
the branch labels. We specified the horizontal option and the angle(0) suboption of ylabel() to get a horizontal dendrogram with horizontal branch labels.

Technical note

Programmers can control the graphical procedure executed when cluster dendrogram is called. This ability will be helpful to programmers adding new hierarchical clustering methods that require a different dendrogram algorithm. See [MV] cluster programming subroutines for details.

Reference


Also see

[MV] cluster  — Introduction to cluster-analysis commands

[MV] clustermat  — Introduction to clustermat commands